POLITECNICO DI MILANO
Corso di Laurea in Ingegneria Informatica
Dipartimento di Elettronica e Informazione

INTRODUCING TRANSFORMATIONS OF VARIABLES IN ESTIMATION OF DISTRIBUTION ALGORITHMS

AI & R Lab
Laboratorio di Intelligenza Artificiale
e Robotica del Politecnico di Milano

Relatore: Ing. Matteo Matteucci
Correlatore: Ing. Luigi Malagó

Tesi di Laurea di:
Emanuele Corsano, matricola 725342
Davide Antonio Cucci, matricola 724767

Anno Accademico 2010-2011

Abstract

Evolutionary Algorithms have become popular in the literature because of their effectiveness as heuristic search strategies for complex combinatorial optimization problems. This work focuses on the analysis of the Estimation of Distribution Algorithms, often presented in the literature as an evolution of Genetic Algorithms, where classical genetic operator have been replaced by statistical operators, such as model selection, estimation and sampling. The behaviour of this class of algorithms can be interpreted as a stochastic walk on the densities of the statistical model employed, i.e. subset of the probability simplex. In this work we propose to map the variables in the optimization problem to a new set of variables. By introducing an independence model over the transformed variables we implicitly define low dimensional models in the original probability simplex. We introduce a criterion based on information theory to chose among such models that can be interpreted from the point of view of Information Geometry. We propose to use such approach as a models selection strategy in the context of EDAs. This leads to the proposal of a novel class of Estimation of Distribution Algorithms which is able to solve multivariate problems by employing sequences of low dimensional statistical models.

Sommario

Il calcolo evolutivo e in particolare gli algoritmi evolutivi sono stati protagonisti di notevole interesse per la loro efficacia nel campo della risoluzione di problemi combinatori di ottimizzazione. Questo lavoro si focalizza sull'analisi di un'evoluzione degli algoritmi genetici basata sull'utilizzo di distribuzioni e modelli di probabilitá: gli algoritmi EDA (Estimation of Distribution Algorithms). Il comportamento di tali algoritmi puó essere interpretato come l'evoluzione, guidata da metodi di ricerca stocastica, di una distribuzione all'interno di un particolare spazio di probabilitá. L'approcio introdotto implica l'utilizzo di modelli parametrici che identifichino tale distribuzione. La scelta del modello, il numero di parametri usati e le caratteristiche della funzione obiettivo del problema di ottimizzazione (funzione fitness), risultano critici per una efficace convergenza verso soluzioni globalmente ottime. Partendo da una generale analisi delle problematiche legate al comportamento del metodo della discesa del gradiente esatto del valor medio della funzione fitness, si cerca di collegare tali criticitá alle difficoltá di convergenza degli algoritmi EDA basati su modelli di indipendenza a n parametri. Questo lavoro introduce un nuovo metodo di trasformazioni delle variabili collegato all'uso di mappe, che permette di ottenere un insieme di problemi equivalenti caratterizzati da criticitá diverse. Tale metodo puó essere visto anche come un modo per generare, dal modello di indipendenza, nuovi modelli a n parametri, che possono essere impiegati per la risoluzione del medesimo problema. Viene quindi presentato un algoritmo per la produzione e la scelta dei modelli, che permette agli algoritmi EDA di essere caratterizzati da migliori proprietÃ rispetto al problema della convergenza. In particolare, sfruttando i risultati della teoria della geometria dell'informazione, viene proposto un criterio per scelta dei modelli basato sulla minimizzazione della distanza di Kullback-Leibler. L'algoritmo presentato ha permesso di ottenere buoni e incoraggianti risultati in termini di numero di valutazioni della funzione fitness e di capacitá nel raggiungimento della soluzione ottima, in modo particolare se confrontato con altri algoritmi EDA a modello

univariato. Tale lavoro si pone quindi come punto di partenza per lo studio di tecniche di trasformazioni delle variabili per la scelta dei modelli, allo scopo di incrementare le prestazioni degli algoritmi EDA.

Contents

Chapter 1

Introduction

Various interesting real-world problems regard the optimization of pseudo-boolean fitness functions, i.e., real valued functions defined over a set of binary variables. No polynomial time algorithm is known which is able to find the solution for a general instance of these problems and it is common belief that such an algorithm is impossible to design [1]. Evolutionary Algorithms have become popular in the literature because of their effectiveness as heuristic search strategies for these problems.

Genetic Algorithms and Estimation of Distribution Algorithms, two families of stochastic search strategies belonging to the Evolutionary Algorithms class, work by evolving a population of candidate solutions which is used to iteratively generate new individuals by means of stochastic operators. This work focuses on the analysis of the Estimation of Distribution Algorithms, often presented in the literature as an evolution of Genetic Algorithms, where classical genetic operator have been replaced by statistical operators, such as model selection, estimation and sampling. EDAs explicitly employ a probability model which is used to learn the features in the candidate solutions that are responsible of good fitness values. At every iteration the better fitness individuals in the population are selected and the density belonging to the model that better fits this sample is chosen. A new population is then generated by sampling from this distribution. The main question which arises here is under which conditions the sequence of densities considered leads to sample the global optimum for the fitness function with high probability.

In EDAs, at every iteration a new density belonging to a the probability model employed is chosen to generate new candidate solutions. This behaviour can be interpreted as a stochastic walk on the manifold of the probability distributions belonging to the model. Experiments have shown that the mean fitness of the individuals in the populations increases almost monotonically but often the best candidate solution found by these algorithms is a local optimum for the fitness function. This leads to conjecture that the probability of EDAs to find the global optimum for the fitness function are strictly related to the choice of the model and in particular to the shape of the expected value of the fitness function. We start discussing a variety of example to validate this conjecture. In particular, we show that in these examples the probability for an EDA to find the global optimum for the fitness function is related with the number and the position of the critical points in the expected value of the fitness with respect to the probability model employed.

It is known in that the most effective EDAs are those which employ a complex and expressive probability model which is able to enclose and reproduce the interactions among variable values in the selected, high fitness individuals. One example is Bayesian Optimization Algorithm [29], which uses Bayesian Networks as probability models. Another example is Distribution Estimation using Markov Random Fields [10]. The main issue here is the computational complexity of the model selection, estimation and sampling operators employed by these algorithms.

In this work we propose to map the variables in the optimization problem to a new set of variables by means of one-to-one maps defined on the space of the candidate solutions. By introducing an independence model over the transformed variables we implicitly define low dimensional models in the original probability simplex. Each one of these models encloses a different structure of interactions among random variables. We introduce a criterion based on information theory to chose among such models that can be interpreted from the point of view of Information Geometry.

We propose to use such approach as a models selection strategy in the context of EDAs. In particular we perform the model selection and the estimation steps at the same time: we choose the low dimensional model among the ones defined which contains the density that causes the minimal information loss when the selected individual sample is represented with the chosen distribution. This leads to the proposal of a novel class of Estimation of Distribution Algorithms which is able to solve multivariate problems by employing sequences of low dimensional statistical models.

This work is organized as follows:

In Chapter 2 the Genetic Algorithms and the Estimation of Distribution Algorithms are introduced along with the main ideas and issues.

In Chapter 3 the mathematical framework used throughout the rest of this work is introduced. First we define the optimization problems the evolutionary algorithms are designed to solve. Then we five an overview of the main results from Information Geometry.

In Chapter 4 a number of examples are studied and we show the correlation between the probability of EDAs to find the global optimum for the fitness function and the shape of the expected value of the fitness function w.r.t. the statistical model employed.

In Chapter 5 we introduce the idea of mapping the variables in the optimization problem to a new set of variables by means of one-to-one maps defined on the space of the candidate solutions. In this chapter we show how an independence model can be introduced over the transformed variables and how this leads to the definition of new low dimensional models in the original probability simplex.

In Chapter 6 a novel EDA is proposed: Function Composition Algorithm. It uses an heuristic search strategy based on results of Information Geometry to perform the model selection and estimation steps. The algorithm is discussed in detail along with a theoretical analysis of its behaviour.

In Chapter 7 the experimental results of the application of FCA on some known test function are presented. FCA is compared with two algorithms known in the literature: Population Based Incremental Learning [6] and Stochastic Gradient Descent [22].

In Chapter 8 conclusion and further work are discussed.

Chapter 2

Genetic Algorithms and Estimation of Distribution Algorithms

Evolutionary algorithms have become popular in the literature because of their effectiveness as heuristic search strategies for complex combinatorial optimization problems. Informally, most of this search strategies try to mimic what happens in nature: the survival abilities of living beings are constantly measured by the environment and only the most suited are allowed to survive. This selection pressure is what drives the evolutionary process and makes species change and adapt to the natural environment over generations. In Evolutionary Algorithms the end user specifies a function which measure the quality of the solutions for his application and the search strategy maintain a population of candidate solutions on which selection is applied by means of the fitness function. In this chapter we review some techniques ad approaches introduced in evolutionary computation, starting from the analysis of the Genetic Algorithms (GAs) and the Estimation of Distribution Algorithms (EDAs). In particular in the first section, we present the structure of the Genetic Algorithm, the genetic operators and their effects. In the second section the Estimation of Distribution Algorithms are discussed, as an evolution of GAs, with an analysis of their properties and the main advantages they have introduced.

where a population (a multi-set) of candidate feasible solutions is evolved

from one to another and

2.1 Genetic Algorithms

Genetic Algorithms are a particular type of Evolutionary Algorithms introduced by John Holland in 1975 [19]. They are inspired by natural genetics and evolutionary theories of Charles Darwin [14]. With GA it is possible to solve an optimization problem by a stochastic search, where the optimization problem is defined as the minimization (maximization) of a function f. GAs evolves a population (a multi-set) of candidate feasible solutions generating new individuals by means of the application of stochastic operators to the current population. GAs are typically used in a black-box contests, when the analytic form of the function f to be optimized is unknown or, in general, when an exact algorithm is intractable in terms of computational complexity. Various types of GAs have been proposed, here we introduce the Simple Genetic Algorithm SGA [16, 23], well known and studied in the literature.

2.1.1 Algorithm

A GA, basically, evolves a population P of m individuals, that represents a state of evolution (current solution). Every individual, also called chromosome, is a string x of n genes. A gene, in the simplest case, is a boolean variable $\{0, 1\}$ and its value is called allele. During one iteration of the algorithm, there are two main steps: selection and reproduction. The selection allows to choose the best individuals, measured by the fitness function $f(x)$, to be used to generate the next population. Reproduction is the process of generating new individuals starting from the current population. It is a combination of two operators, crossover and mutation, which are applied in sequence. Crossover generates two new individuals (children) by recombination of genes of two strings (parents). Mutation is an operator that randomly changes the individual alleles. It is possible to summarize the algorithm in the following steps:

1. Create the initial population
2. Evaluate the fitness of the entire population with $f(x)$
3. Select the best individuals for reproduction
4. Create the new individuals with crossover and mutation

5. Evaluate the fitness of the new population

6. Verify if one of the termination conditions is reached, else go to step 3

The algorithm starts from a population of individuals (usually random), generates a new population at each iteration, and evaluates every individual of a population by a fitness function $f(x)$, that measures a quality of every string. Here two cases arise: in the first case, called white-box contest, the analytic expression of the function f is known. In the second case, called black-box contest, few informations are known about the structure of the fitness function and only a procedure, e.g. an algorithm, is given to evaluate the fitness of candidate solutions. It is important to notice that the single fitness function evaluation, in some contexts, can be time consuming. This aspect can be critical in a GA, when the number of $f(x)$ evaluations is generally high.

In the third step, k individuals in the current population are selected by selection operator, according to the selection rate α (selected portion of population). The probability $p(x)$ of an individual to be selected depends on the type of the selection applied and on $f(x)$. From the selected population P_s, two individuals are randomly chosen and recombined by Crossover operator, that exchanges some alleles between individuals, with a certain schema. In this way, is possible to preserve the alleles of parents in the next population. The reproduction process is completed by mutation, that changes a genes of the new individuals with a certain probability.

At the end of each iteration the algorithm ends if the optimal solution is found. In some cases, e.g., in a black-box contexts, the fitness value of the optimum is unknown, thus the algorithm ends if one of the following termination conditions is verified:

- Suboptimal solution is found and satisfies minimum criteria
- Maximum number of generations is reached
- Maximum computational time is reached

The behaviour of the algorithm depends mainly on the type of genetic operators used. In the next section they are discussed in depth.

2.1.2 Genetic Operators

Here we present the basic concepts and the most common genetic operators of selection, mutation and crossover.

Selection

Selection represents one of the most important operator in the GAs. As happens in nature, it allows to reveal the individuals with best fitness. The degree to which the best individuals are favoured in the selection is usually called Selection pressure. A comprehensive analysis of selection in Evolutionary Algorithms can be found for example in [24] and [27].

One type of selection is Proportional Selection. This operator assigns the probability to be selected to an individual, proportionately to the fitness value, as follows:

$$p(x) = \frac{f(x)}{\sum_P f(x)}$$

where P is the population set, $x \in P$ is an individual of the population and $f(x) > 0 \ \forall \, x$. Note that, in general, the Proportional Selection can be also used with negative function with an appropriate offset. Let consider an $n = 3$ boolean variables example problem defined as the maximization of the function $f(x) = \{number\ of\ ones\ of\ x\}$, where the starting population is $P = \{110, 000, 111, 010, 110\}$. The probability to select the individual $x_1 = 110$ is

$$p(x_1) = \frac{f(x_1)}{\sum_P f(x)} = \frac{2}{8}$$

Another type of selection is Tournament Selection. This method consists in the creation of several tournaments, each composed of ts randomly selected individuals, where ts is the tournament size. In every tournament $k < ts$ individual are selected by means of Proportional Selection. The main advantage of Tournament Selection is the tuning of the selection pressure through the modification of the tournament size ts.

The less sophisticate type of selection method is Truncation Selection. In this method, the individuals are ordered by fitness values and the k highest rated are selected. Let consider the example above, with $k = 3$. Truncation selection order the population P by the fitness function, obtaining $P_o = \{111, 110, 110, 010, 000\}$, and select the first k individual. The selected population is $P_s = \{111, 110, 110\}$.

Finally, a further selection method is Ranking Selection. In this method the population is ordered by fitness and a value depending on its rank is assigned to each individual. The probability of an individual to be selected is:

$$p(x) = \frac{r(x)}{\sum_P r(x)}$$

where $r(x)$ is a rank function. Let reconsider the example above. Ranking Selection order the population P by the fitness function, obtaining $P_o =$

$\{111, 110, 110, 010, 000\}$. The rank values associated to the individual of P_o are: $r(111) = 1$, $r(110) = 2 + 3$, $r(010) = 4$ and $r(000) = 5$. The selection probability of $x_1 = 010$, for example, it is

$$p(x_1) = \frac{r(x_1)}{\sum_P r(x)} = \frac{4}{15}$$

There also other selection schemas, for example see Boltzman Seletion [25].

Crossover

Crossover is the main operator in the process of reproduction. With crossover it is possible to maintain certain sequences of alleles of selected individuals in the next generations. The basic purpose of this operator is to increment the probability of generating new individuals with high fitness, by recombining the genetic material of good and selected individuals.

The first crossover operator that was introduced is One-point Crossover. It generates a random value sp from 1 to n, that represents the splitting point, where n is the length of the string. One-point Crossover selects two random parent individuals and makes two children individuals exchanging the first part of the strings, according to the selected splitting point. This method can be extended by adding more splitting points. Let consider an example with two parent individuals

$$x_{p1} = 110|11000 \qquad x_{p2} = 100|01011$$

and splitting point $sp = 3$. Applying One-point Crossover we obtain two child individuals

$$x_{c1} = 100|11000 \qquad x_{c2} = 110|01011$$

Another common type of crossover is Uniform Crossover. In this case, the operator makes two child individuals as function of a random crossover mask. Crossover mask is a string of length n and each bit is set to one or zero according to the uniform distribution. Uniform Crossover generates two children as a copy of parent individuals, each bit of the first child is exchanged with the corresponding bit of the second child when the associated bit in the crossover mask is one, and remain the same otherwise. Let consider an example with two parent individuals

$$x_{p1} = 11011000 \qquad x_{p2} = 10001011$$

and the crossover mask $c_m = 11000010$. Applying Uniform Crossover we obtain two child individuals

$$x_{c1} = 10011011 \quad x_{c2} = 11001000$$

Mutation

Mutation is an operator, analogous to the biological mutation, which is applied in order to maintain a genetic diversity in the population. The main effect of mutation is to increase the exploration power of the reproduction process, i.e the ability of the algorithm to generate different individual in terms of alleles. In fact, in absence of this operator, for example, is impossible to include new alleles in the population that do not exist in the previous one. The mutation operator is typically implemented flipping every bit with a fixed small probability μ, that it is called mutation rate. Let consider for instance the population $P = \{100, 110, 010, 000, 110\}$. In this case crossover can not generate any individual with the last gene equals to 1, because there is no individual in P with this allele. Otherwise, mutation can insert it changing the last bit of any individual in P.

In a GA it may be important to keep in each iteration the best individual at the previous iteration, because it represents the current best solution for the problem. It can be lost during reproduction, and even during selection with certain non-deterministic schemas. In order to avoid this problem, at each iteration the e best fitness individuals are preserved in the new generation. This technique is called Elitism.

2.1.3 Schema Theorem and Building Block Hypothesis

From the GA behaviour, Holland observed that in some problems the best individuals found in a population share some sequences of adjacent alleles of length $j \ll n$. This sequences, called Building Block, if composed in a strings of length n, generate, with high probability, new individuals with higher fitness value. However, Building Block can be destructed during the recombination of the individuals by Crossover operator. The Holland's original theoretical explanation of this observations is the Schema theory, that it is based on the concept of schema. For individuals of length n, a schema is a subset of the space Ω in which all the chromosomes share a particular set of defined values. Schemata are represented as strings extended with a wildcard symbol $*$, e.g. $\Omega = 0, 1^N$the schema $(1 * *)$ represents the chromosomes $(100), (101), (110), (111)$. A schema S is characterized by the order $o(S)$, which is the the number of defined (non-wildcard) positions,

and by the length $\delta(S)$, which is the distance between its first and last defined positions. Now we can express the probability of selecting schema S at generation t, under the proportional selection assumption, as:

$$P(S,t) = \frac{m(S,t)f(S,t)}{M\bar{f}(t)} = \mathbb{E}[m(S,t+1)]M$$

where M is the population size, $m(S,t)$ is the number of instances of schema S in t, $f(S,t)$ is the mean fitness of the individuals that matching the schema in t and $\bar{f}(t)$ is the mean fitness of the entire population in t.
The crossover and mutation operators change the individuals, thus modify the distribution of schema in the population. Under single point crossover, the (lower bound) probability that the schema S survive at generation t is:

$$P(S\ survives) = 1 - P(S\ does\ not\ survive) = 1 - \frac{\delta(S)}{n-1}P_{diff}(S,t)$$

where $P_{diff}(S,t)$ is the probability that the second parent individual does not match schema S.
The probability of not changing all $o(S)$ non $*$ genes by mutation, where p_m represents the mutation probability, is:

$$(1-p_m)^{o(S)}$$

Usually $p_m \ll 1$, thus we can approximate the last expression as:

$$(1-p_m)^{o(S)} \approx 1 - o(S)p_m$$

It represent the (lower bound) probability of an order o(S) schema S, which survive at generation t.

Theorem 1. *(Schema Theorem) The expected number of schema S at generation $t+1$ when using a canonical GA with Proportional Selection, single point crossover with rate p_c and gene wise mutation with rate p_m is:*

$$\mathbb{E}[m(S,t+1)] \geq \frac{m(S,t)f(S,t)}{\bar{f}(t)}\{1 - p_c\frac{\delta(S)}{1-n}P_{diff}(S,t) - o(S)p_m\}$$

A more generic form for the Schema Theorem might take the form:

$$\mathbb{E}[m(S,t+1)] \geq m(S,t)\alpha(S,t)\{1-\beta(S,t)\}$$

Specifically, the schema S survives when $\alpha(S,t) \geq \{1-\beta(S,t)\}$. This is the basis for the observation that short (defining length), low order schema of above average population fitness are favoured by GAs. This is known as the Building Block Hypothesis.

2.2 Estimation of Distribution Algorithms

The Estimation of Distribution Algorithms, sometimes called Probabilistic Model-Building Genetic Algorithms PMBGA, are a particular evolution of GAs. The term EDA was introduced in 1996 by Muhlenbein and Paaß[26]. The approach proposed with EDAs consist in replacing the GAs reproduction operators with new ones based on probability distributions.

More precisely, a subset of all the possible probability distributions over n binary variables is chosen as the search space and the distribution among those which better fits the selected fraction of the population is chosen. Then a new population is obtained by sampling from this distribution. This paradigm was mainly introduced in order to perform an effective and efficient search. In GAs, indeed, the individuals of population can be seen as a multiple solutions that evolve at the same time. In EDAs, the current solution can be seen as a model distribution that represents multiple individuals that share certain characteristics. This approach permits to learn the dependencies among the problem variables and reach better solutions (correlated exploration). In the next section we introduce this concepts in depth.

2.2.1 Probabilistic Model

Every function can be characterized by the type and the order of interactions among variables. We say that there is an interaction between two or more variables $x_1, x_2, ...x_n$ if the effects on the fitness of the value of x_1 depends on the values of the other $x_2, ..., x_n$ variables. Let us consider, for example, a boolean 3-variable function $f(x) = -10x_1x_2 + x_3$. In this case, we have high fitness values when $x_1 = -x_2$. This can be seen as an interaction of order 2 among x_1 and x_2. In a population, the interactions among the variables of the f function can be revealed by the occurrences of the alleles.

In EDAs the interactions are expressed explicitly through the joint probability distribution associated with the variables of the individuals selected at each generation. The estimation of the model, in EDAs, allows to learn the interactions among the variables, in order to generate, by sampling from the model probability distributions, new individuals with similar correlations. EDAs are typically classified in univariate, bivariate and multivariate, by the order of variable interaction that theirs probabilistic model includes.

There are many types of models used in EDAs. PBIL [6] is a univariate algorithm and its model is represented by the vector of marginal probabilities. In PBIL, as in other univariate algorithms, such as Univariate Marginal

Distribution Algorithm UMDA [26] and Compact Genetic Algorithm cGA [17], the joint probability distribution, becomes the product of the marginal probabilities of n variables:

$$p(X) = \prod_{i=1}^{i} p(X_i)$$

where $X = (X_1, ..., X_n)$ is the vector of variables.
On the other side, Bayesian optimization algorithm BOA [29], is an example of multivariate EDA, based on a Bayesian network. Bayesian networks are composed by nodes that represent variables and edges that encode the joint probability distribution:

$$p(X) = \prod_{i=1}^{n} p(X_i|\Pi_{X_i})$$

where Π_{X_i} is the set of X_i parents nodes in the network and $p(X_i|\Pi_{X_i})$ is its conditional probability. This model can describe all the possible distributions and allows to do a full correlated exploration.

The characterization of the model is very critical in EDAs. A complex model, generally, permits to encode more variable interactions and obtain better performance. However, complex models require a high number of parameters and a difficult process of estimation and sampling.

2.2.2 PBIL

Population Based Incremental Learning, is one of the first EDAs that was proposed. It is based on a simple representation of the model, the vector of marginal probability. The n components of this vector are the expectation values of the associated variables. In each iteration, the population is selected, the vector of the marginal probability is estimated and new individuals are sampled from the current distribution according to a learning rate parameter γ.

Let us consider an example of a PBIL iteration, where $n = 3$, $f = x_1 + 2x_2 - x_3$, $m = 8$ and learning rate $\gamma = 1$. The algorithm starts and create a initial random population

$$P_1 = \{100, 101, 110, 111, 001, 000, 100, 010\}$$

The evaluation of the fitness and the selection are applied over the population (we consider Truncation Selection, with selection rate $s = 0.5$)

$$P_{1e} = \{f(100) = 1, f(101) = 0, f(110) = 3, f(111) = 2,$$

$$f(001) = -1, f(000) = 0, f(100) = 1, f(010) = 2\}$$

$$P_{1s} = \{110, 111, 010, 100\}$$

The vector of the marginal probability is estimated

$$p(X)_1 = (\frac{3}{4}, \frac{3}{4}, \frac{1}{4})$$

A new m individuals are sampled according to this distribution

$$P_2 = \{110, 010, 110, 101, 110, 100, 110, 011\}$$

The algorithm ends when the convergence is reached, i.e., the population is composed by m equal individuals, that correspond to the solution found.

2.3 Conclusions

In this chapter we have presented a compact review of the main ideas and concept of evolutionary algorithms. We presented Genetic Algorithms and the Estimation of Distribution Algorithms as examples of evolutionary search strategies.

Chapter 3

The Mathematical Framework

In this section we present the theories and the results which form the background for this work. Here we define precisely and present some results the optimization of pseudo-boolean functions and their stochastic relaxation. Later the main results from Information Geometry[5] are discussed. This theory regards the introduction of a geometrical structure on the manifold of the probability distributions

3.1 Pseudo-boolean Optimization

In this section the problem of optimizing a class of real-valued functions defined over binary variables, i.e., $f : \{-1, 1\}^n \to \mathbb{R}$ is introduced. These functions are called *pseudo-boolean functions*. A comprehensive review of the pseudo-boolean optimization problem can be found in [8].

3.1.1 The Problem

In the following, instead of the usual 0/1 encoding for variables, we use the following harmonic map: $y \in \{0, 1\}$, $x \in \{-1, 1\}$, $x = (-1)^y$. Moreover, the domain Ω of f is $\{-1, 1\}^n$. Sometimes we use "$-$" and "$+$" as short-cuts for -1 and $+1$.

Given a pseudo-boolean function f and a vector of binary variables

$(x_1, x_2, ..., x_n)$, we want to find the global minimum (maximum).

$$(P) \quad \min f(x), x \in \Omega$$

We call a point x, sometimes indicated with $\overline{x}$, a *candidate solution* for the problem (P). We also refer to elements in Ω as *individuals*, because of the population based approach to the problem (P) of the GAs and EDAs search strategies. $f(x)$ is often called the *fitness function* since it often measures the quality of $x \in \Omega$.

3.1.2 Pseudo-boolean Functions

It is well known that every pseudo-boolean function has a unique representation given by the square-free multi-linear polynomial

$$f(x) = \sum_{\alpha \in I} c_\alpha x^\alpha \tag{3.1}$$

The following multi-index notation used is here and throughout all this work: $I \in \{0,1\}^n$, $I^* = I \backslash \{0\}$, $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n) \in I$ and $x^\alpha = \prod_{i \in 1...n} x_i^{\alpha_i}$ where x_i is the i-th component of the x vector. In the following we use a lexicographic ordering for the vector indices α: $-1 \prec +1$ and $0 \prec 1$. Multi-indices with cardinality 1, i.e., there is a single one in the index, is often used to indicate binary variables. So for instance a three variables vector is written as $(x_{100}, x_{010}, x_{001})$, the coefficient vector c of a two variables pseudo-boolean function is $\left(c_{00}, c_{01}, c_{10}, c_{11}\right)$ and if $n = 2$ and $\alpha = 11$ we have that

$$x^\alpha = x^{(11)} = \prod_{i \in 1...n} x_i^{\alpha_i} = x_{10}^1 x_{01}^1 = x_{10} x_{01}$$

In the (3.1) expansion there are at most 2^n coefficients c_α. In the following we call L the set of all the indices of not-null coefficients.

It is always possible to calculate the c_α coefficients. Note that the complexity of this operation could is, in the general case, the same as solving the problem (P). One way to do this is to exploit the following relation between the values of f and its coefficients:

$$c_\alpha = \mathbb{E}_0[f(x)x^\alpha] = \frac{1}{2^n} \sum_{x \in \Omega} f(x)x^\alpha$$

where $\mathbb{E}_0[\cdot]$ is the expected value of the argument with respect to the uniform probability distribution

Proof.

$$\mathbb{E}_0[f(x)x^\alpha] = \mathbb{E}_0[\sum_{\beta\in L} c_\beta x^\beta x^\alpha] = \mathbb{E}_0[c_\alpha] = c_\alpha$$

since $E_0[x^\alpha x^\beta] = 1$ iff $\alpha = \beta$ and 0 otherwise. □

3.1.3 Stochastic Relaxation

Call $\mathcal{P}$ the manifold of all the probability distributions over a set of n random variables, i.e., all the distributions such as

$$p(x) : \Omega \to [0, 1]$$

$$\text{s.t.} \quad \sum_{x\in\Omega} p(x) = 1$$

and a parameters vector ξ that uniquely identifies a distribution in $\mathcal{P}$. We write p_ξ to indicate the distribution p with parameters ξ and $p_\xi(\cdot)$ for its joint probability function.

Consider now the map

$$\mathbb{E}_{p_\xi}[f] : \mathcal{P} \to [\min f, \max f] \in \mathbb{R}$$

where $\mathbb{E}_{p_\xi}$ is the expected value calculated w.r.t. the distribution p_ξ and reads as

$$\mathbb{E}_{p_\xi}[f] = \sum_{x\in\Omega} p_\xi(x) f(x)$$

It is easy to see that $\mathbb{E}_{p_\xi} = \max f$ if and only if the following holds for p

$$p(x) > 0 \Rightarrow f(x) = max f \tag{3.2}$$

that is, the probability of the vector x is greater than zero if $f(x)$ is a global maximum. The same considerations holds for the minima of f. If f is not constant over Ω the distributions p such as (3.2) holds are distribution *with reduced support*, i.e., there exist at least one $x \in \Omega$ such as $p(x) = 0$.

In this work we address the stochastic relaxation of the problem (P), that is, we look for the minima of the map previously defined:

$$(R) \quad \min \mathbb{E}_{p_\xi}[f], p \in \mathcal{P}$$

It is known that the problem (P) and (R) are equivalent. More precisely, the following theorem holds.

Theorem 2. *Given the unconstrained optimization problem* (P) *and its associated stochastic relaxation* (R)

1. *(P) and (R) are equivalent, i.e., given a solution to either one it is immediate to obtain the solution for the other one.*

2. *Call $\Omega' \subset \Omega$ the set of points where f reaches its global optimum. The solutions to (R) are distributions with reduced support included in Ω'.*

3. *There exists a sequence of parameters ξ_n such as $\lim_{n\to+\infty} s_{\xi_n} = s'$ and $\mathbb{E}_{s'}[f] = min f$.*

The two problems (P) and (R) have the same complexity, which is exponential in the number of variables in the general case. This mean that no search strategy can be designed to be both fast and correct in finding solutions for every instance of (P) or (R). The difference of considering the stochastic relaxation is that the domain of the new variables is a subset of $\mathbb{R}$, so it is continuous instead of discrete. This allows us to apply techniques that came from continuous optimization field.

In general $2^n - 1$ parameters are needed to represent a distribution in $\mathcal{P}$. We discuss this point in detail in later sections. If a subset $\mathcal{M} \subset \mathcal{P}$ of distribution is chosen, for example the set of all independent distributions over n variables, and the search is restricted to this set, under some conditions all the solution to (P) are implied by solutions to the problem (R')

$$(R') \quad \min \mathbb{E}_{p_\xi}[f], p \in \mathcal{M}$$

In particular we require that the topological closure of $\mathcal{M}$ includes distributions with reduced support included in Ω'.

3.2 Parametrizations for $\mathcal{P}$

In the previous section the stochastic relaxation of the combinatorial optimization problem (P) has been introduced. The search space for this problem is the probability simplex $\mathcal{P}$. In this section we present three possible parametrizations for $\mathcal{P}$, i.e., three ways for defining a coordinate system on this space.

3.2.1 The A matrix

We introduce here the linear transformation matrix A_n that is useful in later sections to specify the transformation, or coordinate change, between different parametrizations, i.e.,

$$A_n = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{\otimes n}.$$

Here the symbol $\otimes$ indicates the Kronecker Product, $A_2 = A_1 \otimes A_1$ and $A_n = A_{n-1} \otimes A_1$.

Due to the properties of the Kronecker product, A_n is always invertible, since A_1 is invertible. Moreover A_n^{-1} can be derived from A_n with:

$$A_n^{-1} = \frac{1}{2^n} A_n.$$

3.2.2 Raw Parameters ρ_α

We can identify generic probability distributions by specifying all the values of its joint probability function $p(x_1, x_2, ..., x_n)$. The cardinality of the domain of the joint probability function is 2^n, so we have 2^n different parameters ρ_α, called *raw parameters*. In particular,

$$\rho_\alpha = p\Big(x = \big((-1)^{\alpha_1}, (-1)^{\alpha_2}, ..., (-1)^{\alpha_n}\big)\Big).$$

Two constraints apply here:

$$\sum_{\alpha \in I} \rho_\alpha = 1$$

and

$$\forall \alpha \in I,\ 0 \leq \rho_\alpha \leq 1.$$

The first constraint can be used to obtain one parameter as a function of all the others, actually reducing the number of free parameters by 1.

The set of the ρ which satisfy the constraints above can be represented by a polytope call *probability simplex* which in two variable is the tetrahedron in Figure 3.1. On the border we have distributions whose support is not full, i.e., $p(x) = 0$ for some $x \in \Omega$. In particular, the vertices are the the $\delta(x)$ distributions for which $p(x) = 1$.

3.2.3 The Expectation Parameters η_α

Another way to specify a generic probability distribution $p \in \mathcal{P}$ is to determine all the α-moments η_α, known in the literature as the *expectation parameters*,

$$\eta_\alpha = \mathbb{E}_p[X^\alpha] = \sum_{x \in \Omega} x^\alpha p(x)$$

where $p(x)$ is the joint probability. It is easy to see that η_0 is always 1, so like in the previous case there are $2^n - 1$ free parameters.

Moreover, from the non-negativity constraints on ρ one can derive a set of inequalities $A\eta \geq 0$ that identify the domain of the η vector, called the

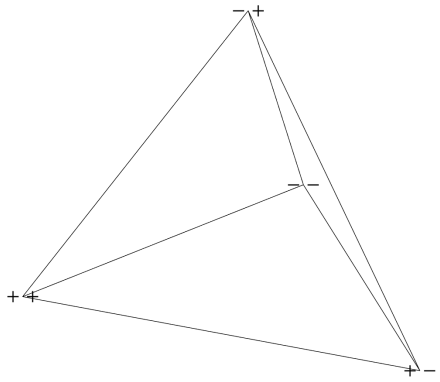

Figure 3.1: The probability simplex

marginal polytope. This can also be determined by the convex hull of the x^α monomials evaluated for each $x \in \Omega$. Details can be found in [31] and [7]. We introduce here an example.

Consider the two variable case, i.e., $x = (x_{10}, x_{01})$. Every probability distribution p can be parametrized with the vector $\eta = \{1, \eta_{01}, \eta_{10}, \eta_{11}\}$:

$$\eta_{01} = \mathbb{E}_s[x_{01}] \quad \eta_{10} = \mathbb{E}_s[x_{10}] \quad \eta_{11} = \mathbb{E}_s[x_{01}x_{10}]$$

The values of the three monomials x^α monomials are

x_{01}	x_{10}	$x_{01}x_{10}$
-1	-1	+1
-1	+1	-1
+1	-1	-1
+1	+1	+1

The expectation polytope, the domain of the η vector, is given by the convex hull of the four points $(-1, -1, +1)$, $(-1, +1, -1)$, $(+1, -1, -1)$ and $(+1, +1, +1)$, shown in Figure 3.2. This is a three-dimensional polytope and can be expressed as a linear transformation of the probability simplex.

Using the expectation parameters one can express the joint probability function in a compact way:

$$p(x;\, \eta) = 2^{-n} \sum_{\alpha \in I} \eta_\alpha x^\alpha. \tag{3.3}$$

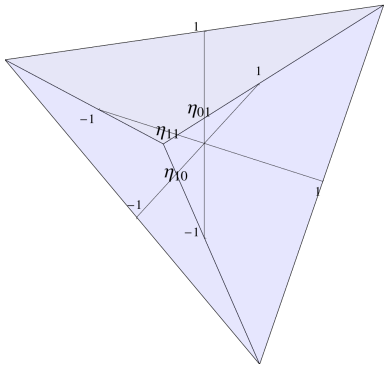

Figure 3.2: The expectation polytope

It is possible to convert the η vector in the equivalent ρ and vice-versa by means of the A_n matrix defined before, i.e.,:

$$\eta = A_n \rho,$$

$$\rho = 2^{-n} A_n \eta.$$

3.2.4 The Natural Parameters θ_α

A third possible parametrization is obtained by introducing an exponential map. Since the logarithm of the joint probability of a generic distribution is a pseudo-boolean function, we can expand it as:

$$\log p(x;\, \theta) = \sum_{\alpha \in I} \theta_\alpha x^\alpha = \sum_{\alpha \in I^*} \theta_\alpha x^\alpha - \theta_0,$$

or, in a more compact form, $\log \rho = A_n \theta$, where the log function is applied element-wise to the components of the vector ρ. The parameters θ_α are known in the literature as *natural parameters*.

As always, the constraint that all probabilities sum to one must hold and the normalizing factor θ_0, usually written as $-\psi(\theta)$, can be obtained from the other parameters:

$$\psi(\theta) = -\theta_0 = -\log \Big[\sum_{x \in \Omega} \exp\{ \sum_{\alpha \in I^*} \theta_\alpha x^\alpha \} \Big].$$

Since $e^x > 0$, $\forall x \in \mathbb{R}$, the components of the θ vector are free and θ_α can take any value in $[-\infty, +\infty]$.

This parametrization only covers $\mathcal{P}_>$, the set of strictly positive probability distributions, i.e., $p(x) \in \mathcal{P}$ s.t. $p(x) > 0 \, \forall x \in \Omega$.

3.2.5 From η to θ Parameters

Consider a distribution p and its η and θ parametrization. A way to convert one parameter vector into the other can be derived exploiting the matrix relationship given before and the fact that A_n is always invertible. It follows that

$$\eta = A_n \exp(A_n \theta) \tag{3.4}$$

$$\theta = 2^{-n} A_n \log(2^{-n} A_n \eta) \tag{3.5}$$

where the log and the exp functions are applied element-wise.

3.3 Probability Models

As it has been shown in the previous section, to represent a generic probability distribution $2^n - 1$ parameters are required. We are interested in defining statistical models $\mathcal{M} \subset \mathcal{P}$ such as every distribution in $\mathcal{M}$ can be uniquely identified with a lower number of parameters. Moreover, a desirable property for such model would be to be closed in topological sense, i.e., they should include all the distributions which are limits of sequences of distributions in the model itself. We use these models as search spaces for the stochastic relaxation problem (R).

3.3.1 The Exponential Family

The family of distributions which can be parametrized with the θ vector is called *the exponential family* [9]. Its joint probability reads as

$$p(x,\, \theta) = \exp\left(\sum_{i \in 1 \ldots n} \theta_i T_i(x) - \psi(\theta) \right),\ \ \theta_i \in \mathbb{R}, \tag{3.6}$$

where T_i are called *canonical* or *sufficient statistics*. Since we deal with binary domain the sufficient statistics are pseudo-boolean functions themselves, so T_i reduce to x^α and Equation (3.6) reads as:

$$p(x,\, \theta) = \exp\left(\sum_{\alpha \in I^*} \theta_\alpha T_\alpha(x) - \psi(\theta) \right). \tag{3.7}$$

Since $p(x,\, \theta) > 0$ for every θ, none of the distributions with reduced support is included in exponential family. It is possible to show that this model is a proper subset of $\mathcal{P}$ and it is not topologically closed, i.e., there exist distributions $q = \lim_{n \to +\infty} p(\cdot,\, \theta_n)$ that are not included in $\mathcal{M}$.

It is possible to define the *extended exponential family* [13] as the union of the exponential family and its topological closure and to give conditions that have to hold for the reduced support distributions to belong to the closure. For example, if all the sufficient statistics x^α with $|\alpha| = 1$ appear in the model then every $\delta(x)$ distribution with support equal to an element x in Ω belongs to the extended exponential family. As a consequence, a solution for (R) implies a solution for (P). For details see [22].

From the observation above and the fact that the θ parameters are free, follows that the exponential family seems an appropriate choice for defining models in $\mathcal{P}$. We can identify a statistical model by choosing a subset of the x^α identified by indices $\alpha \in L \subset I$. The choice of the corresponding monomials allows to determine which interaction among the variables in x are enclosed in the model.

We discuss here an example. We want to find the θ parametrization of all the independent distributions of two variables, i.e., the ones for which $p(x_1, x_2) = p_1(x_1)p_2(x_2)$. Since $p_i(x_i)$ reads as

$$p_i(x_i) = e^{\theta_{i,1}x_i + \theta i,0}$$

it follows that

$$p(x_1, x_2) = e^{(\theta_{1,1}x_1 + \theta_{1,0})}e^{(\theta_{2,1}x_2 + \theta_{2,0})}.$$

Since the canonical statistic x_1x_2 does not appear in the expansion it follows that a parametrization for these distribution is $(\theta_{00}, \theta_{01}, \theta_{10}, 0)$. Like in the general case θ_{00} can be derived from the other parameters θ_α, so that

$$\theta_{00} = -\log\Big[\sum_{x\in\{-1,1\}^2} \exp\{\sum_{\alpha\in\{01,10\}} \theta_\alpha x^\alpha\}\Big] =$$

$$= -\log\left(e^{-\theta_{01}-\theta_{10}}(1 + e^{2\theta_{01}})(1 + e^{2\theta_{10}})\right)$$

It follows that there are only two free parameters. We can see here that to exclude interactions between x_1 and x_2 in the model, it is enough to set to 0 the related θ_{11} coefficient.

The condition given before on sufficient statistics holds for the independence model, so every distribution with reduced support on one vertex of the probability simplex is included in the closure of this model. We show here an example of how this can be seen by direct calculations. Consider the reduced support distribution p such as $p(1, 1) = 1$. We have that

$$p(1, 1) = \frac{e^{2\theta_{01}+2\theta_{10}}}{(1 + e^{2\theta_{01}})(1 + e^{2\theta_{10}})}$$

$$\lim_{\theta_{01},\theta_{10}\to+\infty} p(1,1) = 1$$

This means that there is a sequence of parameter θ such as it admits as limit the desired reduced support distribution, thus it belongs to the closure of the model. Same considerations hold for the other reduced support distributions.

The Gibbs Distribution

We present here a well known model belonging to the exponential family, useful for the theoretical analysis of the stochastic relaxed problem (R). Consider a pseudo-boolean function f as defined in Section 3.1.2 and the probability distribution

$$p(x,\ \beta) = \frac{e^{-\beta f(x)}}{Z(\beta)}$$

$$Z(\beta) = \sum_{x\in\Omega} e^{-\beta f(x)}$$

with $\beta \geq 0$. In statistical physics literature this is known as the *Gibbs (or Boltzmann) distribution* [15], $f(x)$ is the *energy function*, the parameter β the *inverse temperature* and the $Z(\beta)$ the *partition function.*

Here the function $f(x)$ can be seen as the only sufficient statistic $T(x)$ of an exponential family model and it can be decomposed with the usual expansion. The monomials x^α appear and the joint probability can be expressed by (3.7), with $\theta = -\beta c$ and $-Z(\beta) = 1/\psi(\theta)$.

If we look at the limits of the parameter β we have that for $\beta \to 0$, $p(x,\beta)$ tends to the uniform distribution over Ω, while for $\beta \to +\infty$ the limit is the distribution with reduced support on the minima of $f(x)$. Moreover, $\nabla_\beta \mathbb{E}_\beta[f] = -Var_\beta[f]$. This means that the derivative of the expected value of f w.r.t. the distribution p_β is always negative, thus $\mathbb{E}_\beta$ decreases monotonically to its minimum value as β tends to $+\infty$.

In principle the Gibbs distribution seems a good candidate model for the stochastic relaxation problem since its limit is a global optimum for (R). Moreover, the probability of sampling the global optimum for f from $p(x,\beta)$ can be increased easily increasing β. The problem is that in order to employ the Gibbs distribution we need an explicit analytical expression for f and an efficient way to compute the partition function $Z(\beta)$, which involves a summation over 2^n components. The Gibbs distribution can be generalized in a way such that for $\beta \to 0$ $p(x,\beta)$ is a given distribution.

One result is that the curve described by the Gibbs for different β values in the space of distributions follows the gradient of the expected value of the fitness function w.r.t. distributions in $\mathcal{P}$ in the θ parametrization.

3.3.2 Sub-models in the η Parametrization

Let us first introduce the example of the previous section in the η parametrization. The constraint $p(x_1, x_2) = p_1(x_1)p_2(x_2)$ translates into

$$\mathbb{E}[x_1 x_2] = \sum_{\Omega} p(x_1, x_2) x_1 x_2 = \sum_{x_1 \in \{-1,1\}} p(x1) \sum_{x_2 \in \{-1,1\}} p(x2) =$$

$$= \mathbb{E}[x_1]\mathbb{E}[x_2] = \eta_{10}\eta_{01}.$$

So a parametrization for all the independent distributions over two variables is $\eta = \{1, \eta_{01}, \eta_{10}, \eta_{01}\eta_{10}\}$. Like in the θ parametrization, we have two free parameters out of four.

The problem with the η parametrization is that in the general case is not easy to obtain the domain of the parameters. In fact usually a number of inequalities exponential in the number of variables is involved in the definition of the marginal polytope. It is important to note that if an η_α is not a free parameter in the sub-model considered it is not zero but a non-linear function of the other η parameters.

3.4 The Expected Value of f

In this section we analyse the expected value of the fitness function f calculated with respect to a distribution p as a function of the parameters of p. This function is the subject of the optimization for the stochastic relaxation (R) and can be argued that its shape influences the abilities of search strategies for (R) to find the optimum for $\mathbb{E}_p[f]$.

Consider a distribution $p \in \mathcal{P}$ and a pseudo-boolean function f whose set of non-null coefficients is identified by the set of indices $L \subset I$. $\mathbb{E}_p[f(x)]$ is defined as

$$\mathbb{E}_p[f(x)] = \sum_{x \in \Omega} p(x) f(x) = \sum_{x \in \Omega} \rho_x \left(\sum_{\alpha \in L} c_\alpha x^\alpha \right)$$

Here one can exploit the η parametrization of p to write $\mathbb{E}_p$ in a somewhat simpler way.

$$\mathbb{E}_p[f(x)] = \mathbb{E}_s[\sum_{\alpha \in L} c_\alpha x^\alpha] = \sum_{\alpha \in L} c_\alpha \mathbb{E}_p[x^\alpha] = \sum_{\alpha \in L} c_\alpha \eta_\alpha$$

It is immediate to see that only the η_α parameters for which the corresponding coefficient c_α in f is not null appear in this expansion. Moreover, $\mathbb{E}_p$ is a linear function of the η parameters. Note that in the worst case there

are 2^n not null c_α coefficients and thus $\mathbb{E}_p$ can be a linear function of an exponential number of variables.

As a consequence, the following result holds for the the derivatives of the $\mathbb{E}_p$

$$\frac{\partial \mathbb{E}_p[f(x)]}{\partial \eta_\alpha} = c_\alpha, \quad \forall \alpha \in L$$

The gradient vector is constant and not-null for every point η in the domain.

Up to now we have evaluated the expected value of f on a generic distribution parametrized with a vector of $2^n - 1$ parameters. As explained before, one usually chooses a subset of $\mathcal{M} \subset \mathcal{P}$ as the search space for the stochastic relaxation problem. This usually means limiting the number of free parameters. We are interested in studying the expected value of the fitness function on the defined sub-model. In the following we write $\mathbb{E}_{\mathcal{M}}[f]$ to indicate the expected of f with respect to probability distributions $p \in \mathcal{M}$. We start presenting an example, then we state some general result.

Consider the independence model over two binary variables x_{10}, x_{01}. We have already shown the η parametrization for this class of probability distributions and we know that it is $\eta = (1, \eta_{01}, \eta_{10}, \eta_{01}\eta_{10})$. The expected value of the fitness function over this model reads as

$$\mathbb{E}[f(x)] = c_{00} + c_{01}\eta_{01} + c_{10}\eta_{10} + c_{11}\eta_{01}\eta_{10}.$$

The function is not linear, its gradient

$$\nabla \mathbb{E}[f(x)] = \begin{cases} c_{01} + c_{11}\eta_{10} \\ c_{10} + c_{11}\eta_{01} \end{cases}$$

could cancel for some values of η.

If we solve the system

$$\begin{cases} \nabla \mathbb{E}[f] = 0 \\ -1 \leq \eta_{10} \leq 1 \\ -1 \leq \eta_{01} \leq 1 \end{cases}$$

we can derive conditions on the c vector for the presence of critical points in the interior of the η domain. In particular for this example we obtain

$$|c_{11}| \geq |c_{01}| \wedge |c_{11}| \geq |c_{10}|.$$

If this condition on the c_α coefficients hold and thus we have a critical point inside the η domain, it is meaningful to evaluate the function Hessian matrix to gather informations about the nature of the critical point. We omit

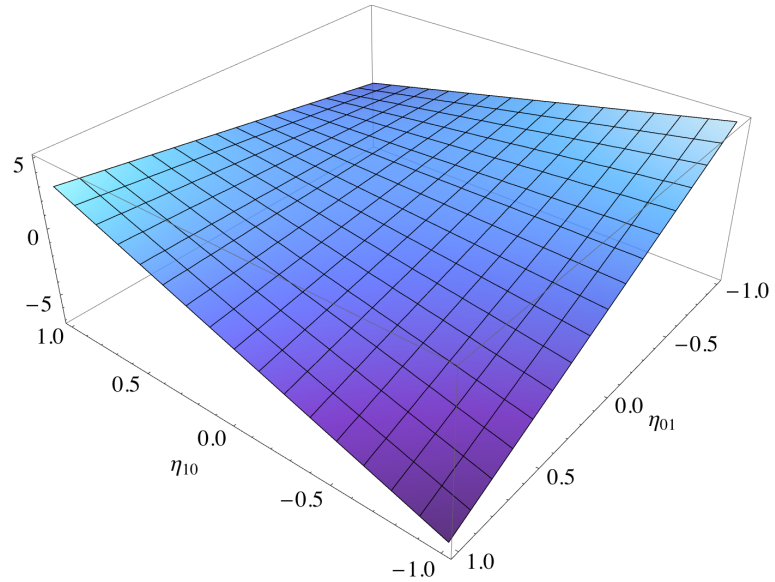

Figure 3.3: The expected value of $f(x) = x_{10} - 2x_{01} + 4x_{10}x_{01}$ over the independence model as a function of η_{01} and η_{10}.

calculations here but it is easy to see that the eigenvalues of the matrix are always $\lambda_{1,2} = \pm c_{11}$. They are not null and have opposite sign, thus the critical point is a saddle point.

In this example we have shown that the expected value of f calculated on a model $\mathcal{M}$ can have critical points. Here we give a condition for the linearity of the expected value as a function of the η parameters:

Theorem 3. *If every η_α such that c_α appears in the expansion of f is a free parameter of $\mathcal{M}$, then $\mathbb{E}_\mathcal{M}[f]$ has no stationary points.*

The proof follows immediately looking at the formulas derived at the beginning of the section. If c_α is not null then the monomial $c_\alpha \eta_\alpha$ appears in the expansion of $\mathbb{E}_\mathcal{M}[f]$. If η_α is free in the model η parametrization then it cannot be derived from the others (e.g. $\eta_{11} = \eta_{01}\eta_{10}$ in the previous example), thus the function is linear and its gradient constant.

This means that if the model is chosen to be expressive enough to capture all the interaction between variables present in the function f, then the expected value of f calculated on the model is linear in the η parameters.

If the coefficients of f are not known we can not be sure that the condition previously stated holds. However, in general we chose to work with low dimensional model for computational reasons and thus with models which do not include all the interactions among variables existing if f. This implies that the the expression of $\mathbb{E}[f]$ is non linear and its landscape could contain

critical points. However the following result holds, whose proof can be found in [21].

Theorem 4. *Consider a model $\mathcal{M}$. If $\mathbb{E}_{\mathcal{M}}[f]$ admits a stationary point for some distribution $p \in \mathcal{M}$ then p is a saddle point.*

To the best knowledge of the authors the weakest sufficient condition that guarantees that expected value of f on the chosen model has no saddle points is not known. Note that if the expected value of the fitness function in the η parametrization is not linear, in the general case its gradient has zeros (i.e. saddle points), inside or outside the η vector domain.

We end this section introducing a theorem that relates the gradient of $\mathbb{E}[f]$ with the covariance between the sufficient statistics and the fitness function in the θ parametrization. This theorem suggests a way to estimate the gradient vector in a point p given a sample from the distribution and the fitness evaluations for each element in the sample.

Theorem 5. *Consider a model $\mathcal{M}$ and its θ parametrization*

1. $\frac{\partial \mathbb{E}_\theta[f]}{\partial \theta_\alpha} = Cov_\theta(f, T_\alpha)$.
2. *$p_\theta \in \mathcal{M}$ is a stationary point for $\mathbb{E}_\theta[f]$ if and only if $Cov_\theta(f, T_\alpha) = 0$ $\forall \alpha$ such that the sufficient statistic T_α is included in the model.*

Further details can be found in [21].

3.5 Information Geometry

Information Geometry proposes to introduce a geometrical structure to study probability distributions in a statistical model $\mathcal{M} \subseteq \mathcal{P}$ of the probability distributions. The Riemannian geometric structure was introduced by Rao [30]. Csiszár studied the geometry of the f-divergence in detail and applied it to information theory [12], [11]. Nagaoka and Amari developed a theory of dual structures [28] and unified all of those theories in the dual differential-geometrical framework [2].

In this section we first recall some concepts from information theory and then we move to a brief presentation of the main Information Geometry results used in this work.

3.5.1 Entropy and Mutual Information

In information theory, Entropy is a measure of the uncertainty associated with a random variable, or equivalently, it is a measure of the average information content one is missing when he does not know the value of the

random variable. The concept was introduced by Claude E. Shannon in 1948 [33]. In the discrete case, the Entropy $H(x)$ of a random variable $x \in \Omega$ whose distribution is p reads as:

$$H(x) = \sum_{x \in \Omega} p(x) \log[p(x)] = \mathbb{E}_p\big[\log p(x)\big]$$

Another useful concept which comes from information theory is the Mutual Information between two random variables x and y. This quantity measures the mutual dependency of the two variables, or equivalently, how much uncertain one is about the value of y once it is known the value of x. For example, if x and y are independent then if we know the value of x we can not say anything about the value of y. Thus the Mutual Information between x and y is zero. Formally the Mutual Information $I(x, y)$ reads as:

$$I(x, y) = \sum_{x,y \in \Omega} p(x, y) \log \Big[\frac{p(x, y)}{p(x)p(y)}\Big],$$

where $p(x, y)$ is the joint probability and $p(x)$, $p(y)$ are the marginal probabilities of x and y.

3.5.2 The geometry of $\mathcal{P}$

Consider a family of probability distributions $\mathcal{M} = p(\cdot, \xi)$ over n variables where ξ is a vector of parameters that uniquely identifies a distribution (i.e. the η or the θ vector). $\mathcal{M}$ can be regarded as a k-dimensional manifold having ξ as a coordinate system, where k is the dimension of the ξ vector.

This manifold is Riemannian and the Fisher G information matrix plays the role of the Riemannian metric tensor.

$$g_{ij} = \mathbb{E}\left[\frac{\partial \log p(x,\, \xi)}{\partial \xi_i} \frac{\partial \log p(x,\, \xi)}{\partial \xi_j}\right].$$

The squared distance between two nearby distributions reads as

$$ds^2 = \sum g_{ij}(\xi) d\xi^i d\xi^j = 2KLD[p(x,\, \xi) : p(x,\, \xi + d\xi)]$$

where KLD is the *Kullback-Leibler Divergence* [20], also called *relative entropy*. This is a measure of the loss of information (in terms of entropy) when a *true* distribution p is approximated with a model distribution q. In infinitesimal neighbourhood the KLD becomes symmetric and is related to the metric defined on the manifold.

It turns out that there are two meaningful and not equivalent ways of defining straight paths, or *geodesics*, connecting two points p and q on this manifold, the linear mixture of the η or θ coordinates:

m-geodesic: $(1-\alpha)\eta_p + \alpha\eta_q$
e-geodesic: $(1-\alpha)\theta_p + \alpha\theta_q$

Here m stands for *mixture* and e for *exponential*. It turns out that the η coordinate system is m-flat while the θ one is e-flat.

These two coordinate systems have the following crucial property: the directions of small changes along different axes are mutually orthogonal. More precisely

$$\langle \partial_{\theta_\alpha}, \partial_{\eta_\beta} \rangle = \mathbb{E}\left[\frac{\partial \log p(x)}{\partial \theta_\alpha}\frac{\partial \log p(x)}{\partial \eta_\beta}\right] = \delta_{\alpha\beta}$$

where $\delta_{\alpha\beta}$ is the Kronecker delta.

This allow us to introduce an analogous of the Pythagoras theorem: consider three distributions p,q,r such as the e-geodesic connecting r and q is orthogonal to the m-geodesic connecting p and r. Then we have

$$D[p:q] = D[p:r] + D[r:q]$$

where D is the Kullback-Leibler Divergence.

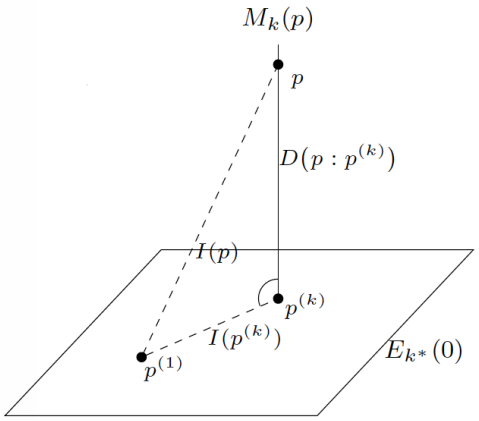

Figure 3.4: Orthogonality of e and m geodesics and Pythagoras theorem equivalent.

3.5.3 k-cut Mixed Coordinate System

The orthogonality property introduced in the previous section can be used to define a new coordinate system that enlightens the hierarchical structure of the correlations between variables in probability distributions. Details can be found in [3], [4].

Consider the following partitions of coordinates:

$$\theta = (\theta_k,\ \theta_{k^*})$$

$$\eta = (\eta_k,\ \eta_{k^*})$$

here k is a short-cut and indicates the set off all the multi-indexes such as $|\alpha| \geq k$ and k^* is its the complement, i.e. all the α such as $|\alpha| < k$. The θ_k* are called the *higher order interactions*.

A probability distribution can be equivalently parametrized with any of the partitions

$$(\eta_k,\ \theta_{k^*})$$

The point is that in this parametrization any change in the θ_{k^*} coordinates does not change the η_k part; the other way round holds as well. Keeping the marginals η_k constant defines an m-flat sub-manifold $M_k(\eta_k)$ of $\mathcal{P}$ which includes all the distributions with the same marginals and different higher order interactions. For different values of η_k the different sub-manifolds do not overlap and define a foliation of $\mathcal{P}$. Another foliation is defined by the e-flat sub-manifolds $E_{k^*}(\theta_{k^*})$ composed by the distributions which have the same higher order interaction θ_{k^*} but different marginals.

3.5.4 Projections

Given a generic distribution p we define

$$p^{(k)} = \underset{q \in E_{k^*}(0)}{\arg\min} D[p : q]$$

This is the closest distribution to p among the ones with no intrinsic interactions more than k variables, in terms of Kullback-Leilber divergence. The k-cut parametrization of $p^{(k)}$ has the same marginals η_k of p and $\theta_{k^*} = 0$. We are actually making an orthogonal projection of the distribution p on the sub-manifold $E_{k^*}(0)$. Because of the properties stated before, this means moving on the m-geodesic connecting the two distribution p and $p^{(k)}$ on the m-flat manifold $M_k(\eta_k)$.

Again, for the orthogonality of the involved sub-manifolds, $p^{(k)}$ can be equivalently written as the projection of the uniform distribution $p^{(0)}$ on the manifold $M_k(\eta_k(p))$. This means moving on the e-geodesic connecting $p^{(0)}$ and $p^{(k)}$ on the e-flat manifold $E_{k^*}(0)$, that is:

$$p^{(k)} = \underset{q \in M_k(\eta_k(p))}{\arg\min}\ D[q : p^{(0)}]$$

We discuss an example here. A generic distribution p over 3 variables is given along with its η parametrization. We are interested on the nearest distribution to p between the ones with no interactions of order $k > 1$, i.e. we want to find the distribution $q \in E_1(0)$ that is nearest to p in terms of Kullback-Leibler Divergence. This means to project p orthogonally on the sub-manifold $E_1(0)$. We introduce the k-cut parametrization of p, with $k = 1$. It reads as:

$$(\eta_{001}, \eta_{010}, \eta_{100};\ \theta_{011}, \theta_{110}, \theta_{101}, \theta_{111}).$$

For what has been said before, it is immediate to write the 1-cut parametrization for the distribution q. It is

$$(\eta_{001}, \eta_{010}, \eta_{100};\ 0, 0, 0, 0),$$

since it has the same marginals of p but no interactions of order $k > 1$. The problem with this parametrization is that in the general case we do not have an explicit expression for $p(x, (\eta_k, \theta_{k^*}))$. So we would like to determine the full θ parametrization of q. To do this we exploit the Legendre transformations which relate directly the parameters θ_α and η_α. We have that:

$$\theta_\alpha = \frac{\partial \varphi(\eta)}{\partial \eta_\alpha}$$
$$\eta_\alpha = \frac{\partial \psi(\theta)}{\partial \theta_\alpha}$$

where $\varphi(\eta)$ is the negative entropy of the distribution as a function of the η vector and $\psi(\theta)$ is the partition function. Since q is the independent distribution over three variables we have that its entropy decomposes in the sum of the entropies of the marginal distributions q_i

$$H_q(\eta) = \sum_{i \in \{1,2,3\}} H_{q_1}(\eta)$$

Since $p(x_i, \eta) = (1 + \eta_i x_i)/2$ we can write the entropy of the i-th variable as

$$H_{q_i} = \frac{1+\eta_i}{2} \log \frac{2}{1+\eta_i} + \frac{1-\eta_i}{2} \log \frac{2}{1-\eta_i}.$$

Going forward with the calculations we have

$$\theta_i = \frac{\partial \varphi(\eta)}{\partial \eta_i} = \frac{\partial}{\partial \eta_i}\big[-\sum H_{q_i}(\eta_i)\big] = \frac{\partial}{\partial \eta_i}\big[-H_{q_i}(\eta_i)\big] = \frac{1}{2} \log \Big[\frac{1-\eta_i}{1+\eta_i}\Big].$$

Here we used an intuitive short-cut with the indices of the vectors η, and θ, i.e. η_i is the marginal of the i-th variable. The usual multi-index notation

would have been η_{001} for the third variable, η_{010} for the second, and so on. Since the relevant parameters are only three, this simplification can be used without risk of confusion.

This procedure of decomposing a distribution in its k-order reductions is related to the concept of Max-Likelihood estimation and model fitting.

These results about the sub-manifold geometries can be used to give another expression of the mutual information between a set of random variables. Consider a generic distribution p. We have that

$$I(p) = D[p : p^{(1)}] = \sum_{k\in\{2,\dots,n\}} D[p^{(k)} : p^{(k-1)}] =$$

$$= D[p : p^{(n-1)}] + D[p^{(n-1)} : p^{(n-2)}] + \dots + D[p^{(2)} : p^{(1)}].$$

This allows us to distinguish precisely between k-order interactions between variables in a distribution. More precisely, we have that $D[p^{(k)} : p^{(k-1)}]$ measures precisely the amount of interactions of order k.

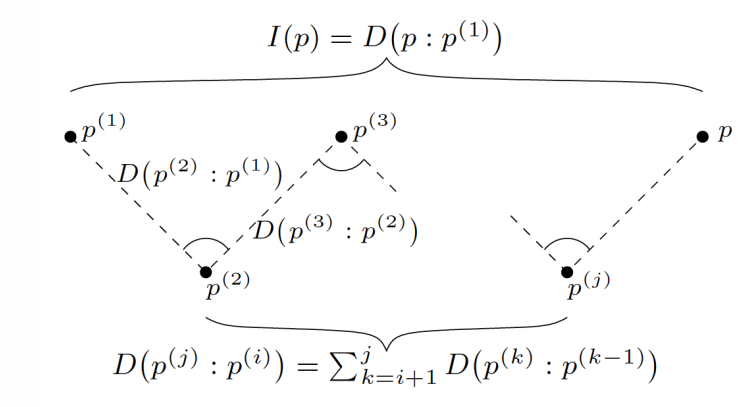

Figure 3.5: Probability distribution decomposition in terms of k-order interactions

Chapter 4

The Expected Value of the Fitness function

The stochastic relaxation problem (R) was defined in the previous chapter as the problem of finding the distribution s in $\mathcal{P}$ for which the expected value of the fitness function f calculated over the distribution p is optimum. Usually the search space for the problem (R) is limited to a sub-manifold $\mathcal{M} \subset \mathcal{P}$. In this chapter we concentrate on example fitness functions in two or three variables and we present an in depth analysis of $\mathbb{E}_{\mathcal{M}}[f]$ as a function of the η and θ parametrization of the distributions belonging to the independence model. This chapter is organized as follows: first motivations for this analysis are presented, along with the definition of the Exact Gradient Descent search strategy for (R). This strategy is used to investigate the relations between the expected value of the fitness function and the convergence proprieties of real search strategies for (R). In the following sections, we study analytically some two and three variables fitness functions when the model considered is the independence model. The last sections cover some results which allow to generalize the observation drawn from the studied examples.

4.1 Motivations

All population based search strategies iteratively produce new populations of candidate solutions applying different reproduction and selection operators to the current individuals till some termination criterion is satisfied.

There is an interpretation which holds for the great majority of evolutionary search strategies: their behaviour can be interpreted as a stochastic walk in the manifold of the probability distributions $\mathcal{P}$. For the Estimation of Distribution Algorithms this is straightforward since they employ distributions belonging to a fixed model $\mathcal{M} \subset \mathcal{P}$ to generate populations. For Genetic Algorithms this is more complex since new individuals are generated with operators such as crossover and mutation which work directly on candidate solutions and even if the populations can be interpreted as samples form probability distributions in $\mathcal{P}$ it is difficult to characterize the model they belong to. A comprehensive discussion of this point of view for GAs can be found in [38].

Experimental analysis on the behaviour of various evolutionary algorithms show that the mean value of the fitness function over the population improves almost monotonically during the optimization process. Some intuition about this fact can be gained observing that most of the search strategies are based on some kind of selection scheme that removes from the populations the poorest fitness individuals, thus explicitly improving the mean fitness $\overline{f}$. One way to interpret this observation is the following: evolutionary search strategies perform a local search in $\mathcal{M}$ such as the new distribution p' has an expected fitness greater than the current one. Thus it can be argued that the performances of evolutionary search strategies are influenced by the shape of the expected value of f calculated on the model $\mathcal{M}$. We show an example of this fact in the next section.

4.1.1 The PBIL Behaviour

The Population Based Incremental Learning algorithm was introduced in Chapter 2 to clarify the basic EDA ideas and techniques. Recall that at every iteration PBIL fits an independent distribution p using the selected individuals as sample and generates a new population sampling from p. Interpreting this behaviour as suggested in the previous section, PBIL performs a stochastic walk on the manifold $\mathcal{M}$ of the independent distributions over n variables.

It has been proven in [18] that PBIL, for big enough populations, is always able to find the global optimum for f in the case it is linear in the x variables, i.e., all of its c_α coefficients with $|\alpha| > 1$ are zero. This means that the distribution with reduced support on the global optimum for f is the only attractor for the PBIL search strategy. On the other hand, it is known that if f is non linear other attractors could exist.

Consider now two pseudo-boolean function f and g whose coefficients

vectors are $c_f = (0, 1, 2, -4)$ and $c_g = (0, 1, 2, -0.5)$. An approximation of the dynamics of PBIL under the hypothesis of infinite population has been determined experimentally by executing PBIL with a population four orders of magnitude bigger than the cardinality of the domain Ω.

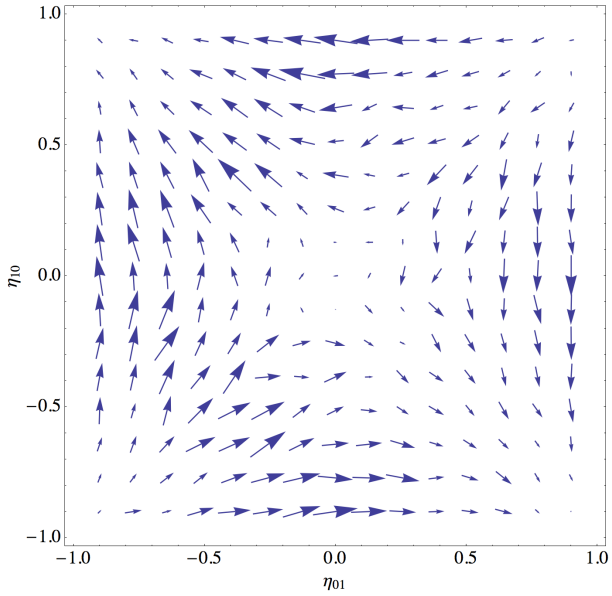

Figure 4.1: PBIL gradient field for f

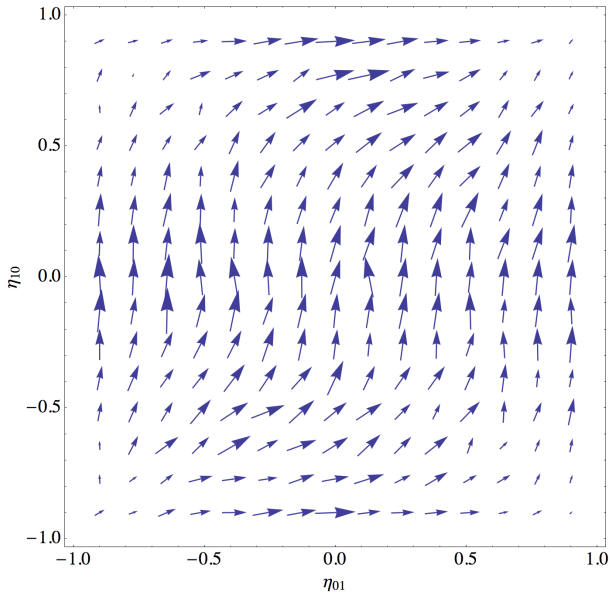

Figure 4.2: PBIL gradient field for g

In Figure 4.1 two different attractors can be seen in $\eta = (1, -1)$ and $\eta = (-1, 1)$ so there exist initial conditions starting from which the PBIL algorithm is not able to converge to the global optimum. The function f is non linear and this result was expected. Instead in the second figure only one attractor can be seen, corresponding to the global optimum for g, even tough the function is non linear. Note that $\mathbb{E}_{\mathcal{M}}[f]$ has a saddle point in (0.5, 0.25), i.e., in the interior of the η domain, while the saddle point for $\mathbb{E}_{\mathcal{M}}[g]$ lays in $(4, 2)$, outside the parameter domain. Here $\mathcal{M}$ is the independence model. This result generalizes for all two variable fitness functions: if the expected value of f calculated over the independence model has a saddle point inside the parameters domain then there exist two attractors for the PBIL search strategy. The proof is given in later sections. It is still an open issue if this connection with the expected fitness function shape generalizes to more than one variable.

4.1.2 The Exact Gradient Descent Strategy

In this section we present a local greedy search strategy that is explicitly connected with the structure of the expected fitness function and whose convergence properties are intimately related with the presence of critical points in the expected value of the function f.

It is well known that the gradient of a function is the vector that represents the direction of steepest ascent. This local information can be used to update the parameters of the current distribution, for example

$$\eta(t+1) = \eta(t) + \gamma \nabla \mathbb{E}_{\eta(t)}[f]$$

where γ is a learning rate.

We know from the Information Geometry theory, presented in the second chapter, that the geometry defined on the space of the probability distributions belonging to a model $\mathcal{M}$ is not Euclidean, points that have equal euclidean distance in the space of the parameters may represent distributions with different Kullback-Leibler distance in the manifold $\mathcal{P}$. Moreover, the axes of the parameters space are not orthogonal. Thus the gradient does not exactly indicate the direction of steepest ascent of $\mathbb{E}_{\mathcal{M}}[f]$. The exact approach would be to use the natural gradient instead:

$$\tilde{\nabla} \mathbb{E}_p[f] = I_p^{-1} \nabla \mathbb{E}_p[f]$$

where I_p is the Fisher information matrix evaluated on the distribution p.

Note that the Exact Gradient Descent search strategy can not be applied on real problem since calculating the exact gradient has a complexity

exponential in the number of variables. In fact this requires to know all the c_α coefficients or, equivalently, to evaluate f for every $x \in \Omega$. This can be seen remembering the expansions of the expected fitness function given in Chapter 3.

Given the interpretation of the behaviour of EDAs and GAs proposed in the previous section it seems reasonable to consider the performances of the Exact Gradient Descent search strategy as a lower bound for the performances of evolutionary algorithms which explicitly employ the same model $\mathcal{M}$ as EGD. One of the reasons is that EGD is greedy and deterministic and it converges in general to a local optimum for $\mathbb{E}_{\mathcal{M}}[f]$, depending on the shape of the expected fitness function on the considered model $\mathcal{M}$.

4.2 The shape of the Expected Fitness function

In this section we analyse the expected value of a fitness function f over the independence model. We are interested in characterizing its critical points and the attraction basins for the Exact Gradient Descent search strategy. Now we introduce some example cases in two and three variables.

4.2.1 Two Variables Case

Expected Fitness Function with Saddle Point

Let us consider a specific example of two variables case. The fitness function is:

$$f(x) = x_{10} - 2x_{01} + 4x_{10}x_{01}$$

and its expected value calculated on the independence model ($\eta_{11} = \eta_{01}\eta_{10}$) is:

$$\mathbb{E}[f(x)] = \eta_{10} - 2\eta_{01} + 4\eta_{10}\eta_{01}$$

We can now evaluate the gradient of the expected value as:

$$\nabla\mathbb{E}[f(x)] = \begin{cases} -2 + 4\eta_{10} \\ 1 + 4\eta_{01} \end{cases}$$

The condition $|c_{11}| \geq |c_{01}| \wedge |c_{11}| \geq |c_{10}|$ is satisfied, thus the expected fitness function has a saddle point in $\eta_{10} = -\frac{c_{01}}{c_{11}} = 0.5, \eta_{01} = -\frac{c_{10}}{c_{11}} = -0.25$. Note that the gradient vector components, in the two variable case, are always linear in η_α.

The trajectories of the gradient in the space η_{10}, η_{01} are shown in Figure 4.3. Is easy to see how the space is divided in two attraction basins; there exist a local optimum in $\eta = (1, 1)$ and the global optimum is in $\eta = (-1, -1)$.

In cases like this the ability of the Exact Gradient Descent search strategy to converge to the global optimum depends on the starting distribution. The dimensions of the attraction basins are defined by the position of the saddle point, thus, if the optimum distribution has a big attraction basin, an Exact Gradient Descent search strategy reaches the optimum distribution with high probability.

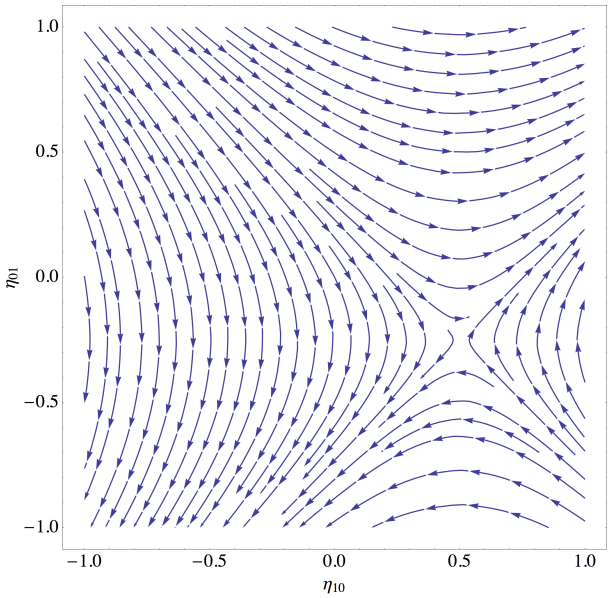

Figure 4.3: Trajectories of the Exact Gradient Descent search strategy for f in the η parametrization

Now we analyse the two variable case in the natural parametrization θ. The expected value of the fitness function read as:

$$\mathbb{E}[f(x)] = \frac{e^{\theta_{10}-\theta_{01}}(c_{00}+c_{10}-c_{01}-c_{11})}{e^{-\theta_{10}-\theta_{01}}+e^{\theta_{10}-\theta_{01}}+e^{-\theta_{10}+\theta_{01}}+e^{\theta_{10}+\theta_{01}}}+$$
$$+\frac{e^{-\theta_{10}+\theta_{01}}(c_{00}-c_{10}+c_{01}-c_{11})}{e^{-\theta_{10}-\theta_{01}}+e^{\theta_{10}-\theta_{01}}+e^{-\theta_{10}+\theta_{01}}+e^{\theta_{10}+\theta_{01}}}+$$
$$+\frac{e^{-\theta_{10}-\theta_{01}}(c_{00}-c_{10}-c_{01}+c_{11})}{e^{-\theta_{10}-\theta_{01}}+e^{\theta_{10}-\theta_{01}}+e^{-\theta_{10}+\theta_{01}}+e^{\theta_{10}+\theta_{01}}}+$$
$$+\frac{e^{\theta_{10}+\theta_{01}}(c_{00}+c_{10}+c_{01}+c_{11})}{e^{-\theta_{10}-\theta_{01}}+e^{\theta_{10}-\theta_{01}}+e^{-\theta_{10}+\theta_{01}}+e^{\theta_{10}+\theta_{01}}}$$

The gradient vector of the expected fitness function is:

$$\nabla\mathbb{E}[f(x)] = \begin{cases} \frac{4e^{2\theta_{01}}\left(\left(1+e^{2\theta_{10}}\right)c_{01}+\left(-1+e^{2\theta_{10}}\right)c_{11}\right)}{\left(1+e^{2\theta_{10}}\right)\left(1+e^{2\theta_{01}}\right)^2} \\ \frac{4e^{2\theta_{10}}\left(\left(1+e^{2\theta_{01}}\right)c_{10}+\left(-1+e^{2\theta_{01}}\right)c_{11}\right)}{\left(1+e^{2\theta_{10}}\right)^2\left(1+e^{2\theta_{01}}\right)} \end{cases}$$

The coordinate of the critical points for $\nabla\mathbb{E}_\theta[f]$ are

$$\bar{\theta} = \Bigg\{ (\log\left[-\frac{\sqrt{-c_{10}+c_{11}}}{\sqrt{c_{10}+c_{11}}}\right], \log\left[-\frac{\sqrt{-c_{01}+c_{11}}}{\sqrt{c_{01}+c_{11}}}\right]),$$

$$(\log\left[-\frac{\sqrt{-c_{10}+c_{11}}}{\sqrt{c_{10}+c_{11}}}\right], \log\left[\frac{\sqrt{-c_{01}+c_{11}}}{\sqrt{c_{01}+c_{11}}}\right]),$$

$$(\log\left[\frac{\sqrt{-c_{10}+c_{11}}}{\sqrt{c_{10}+c_{11}}}\right], \log\left[-\frac{\sqrt{-c_{01}+c_{11}}}{\sqrt{c_{01}+c_{11}}}\right]),$$

$$(\log\left[\frac{\sqrt{-c_{10}+c_{11}}}{\sqrt{c_{10}+c_{11}}}\right], \log\left[\frac{\sqrt{-c_{01}+c_{11}}}{\sqrt{c_{01}+c_{11}}}\right]) \Bigg\}.$$

At most one of these solutions can be real, thus belonging to the θ parameters domain. One reason for this comes from the fact that there can be at most one saddle point in the η parametrization and it is known that the critical points in $\mathbb{E}_\theta[f]$ are the same as in the η domain, once the coordinate conversion (3.5) has been applied.

Let us consider the function f of the previous example and its expected value w.r.t. the independence model in the θ parametrization, i.e., $\theta_{11} = 0$.

$$\mathbb{E}[f(x)] = \frac{e^{\theta_{10}+\theta_{01}}}{e^{-\theta_{10}-\theta_{01}} + e^{\theta_{10}-\theta_{01}} + e^{-\theta_{10}+\theta_{01}} + e^{\theta_{10}+\theta_{01}}} +$$

$$+\frac{e^{-\theta_{10}-7\theta_{01}}}{e^{-\theta_{10}-\theta_{01}} + e^{\theta_{10}-\theta_{01}} + e^{-\theta_{10}+\theta_{01}} + e^{\theta_{10}+\theta_{01}}} +$$

$$+\frac{e^{-\theta_{10}-5\theta_{01}}}{e^{-\theta_{10}-\theta_{01}} + e^{\theta_{10}-\theta_{01}} + e^{-\theta_{10}+\theta_{01}} + e^{\theta_{10}+\theta_{01}}} +$$

$$+\frac{e^{\theta_{10}+3\theta_{01}}}{e^{-\theta_{10}-\theta_{01}} + e^{\theta_{10}-\theta_{01}} + e^{-\theta_{10}+\theta_{01}} + e^{\theta_{10}+\theta_{01}}}.$$

Substituting the coefficient values in the previously expressions, we have that the saddle point for $\mathbb{E}_\theta[f]$ is at

$$\theta = \Big(\frac{1}{2}\log(3),\ -\frac{1}{2}\log(\frac{5}{3})\Big).$$

In the Figure 4.4 the expected value of the fitness function expresses in the natural parametrization θ is shown, while in Figure 4.5 can be seen some Exact Gradient Descent trajectories. Remember that the domain of the θ_α parameters is $(-\infty, \infty)$.

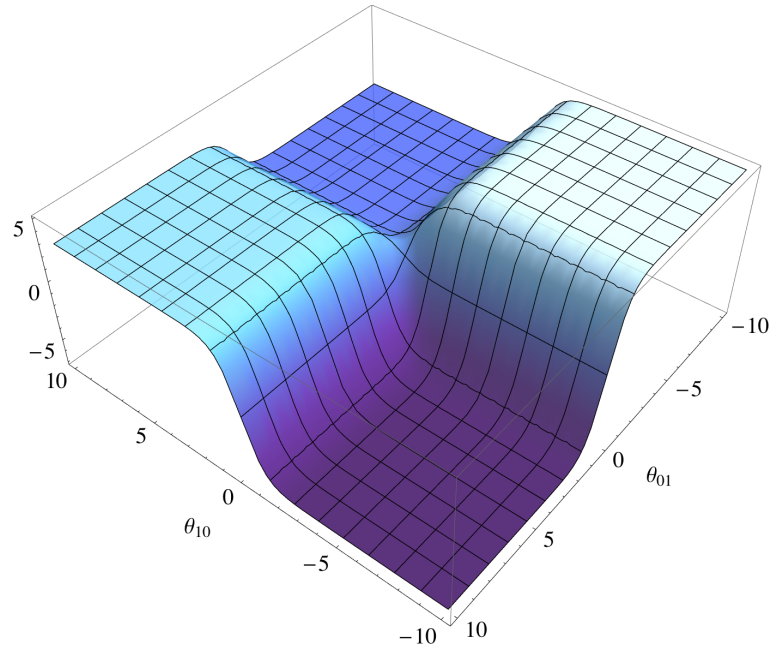

Figure 4.4: The expected value of f calculated over the independence model as a funciton of θ_{01} and θ_{10}

Expected Fitness Function without Saddle Point

We consider now another two variable fitness function f:

$$f(x) = 2x_{10} - 4x_{01} + x_{10}x_{01}$$

In this case, considering the independence model, the saddle point condition is not satisfied and the components of the gradient vector of the expected fitness function are never zero at the same time. The expected value of f, as a function of the η parameters, is shown in Figure 4.6. There exists only one attraction basin in the domain of the η parameters and an Exact Gradient Descent strategy always converges to the global optimum as it can be seen in the Figure 4.7.

Domain and Eigenvectors

The absence of the saddle point in the expected fitness function does not mean that the saddle point does not exists, but implies that is out of the parameters domain (note that, when c_{11} tends to 0, the modulus of the coordinates of the saddle point tends to ∞). The saddle point, although it is outside the domain, has an effect on the shape of the expected value of fitness function in the domain. In particular, in the two variables case,

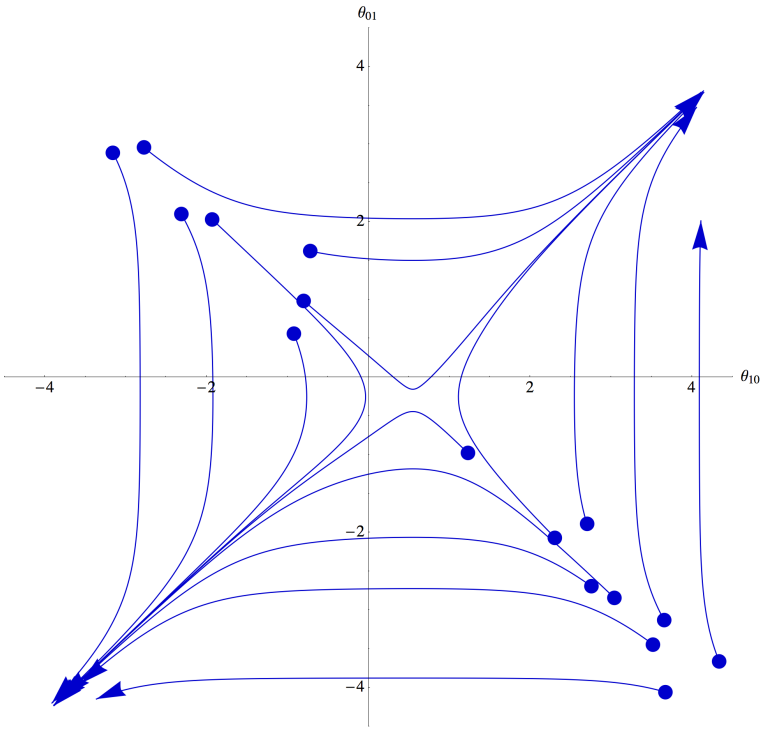

Figure 4.5: Trajectories of the Exact Gradient Descent search strategy for f in the θ parametrization

it can creates different basins, i.e., Exact Gradient Descent strategies can converge to sub-optimal solutions.

We consider now the example fitness function:

$$f(x) = -1.2x_{10} + x_{10}x_{01}$$

In this case the saddle point condition is not satisfied, thus the saddle point of the expected fitness function is out of the domain. Its position is

$$\nabla\mathbb{E}[f(x)] = \begin{cases} \eta_{10} = 0 \\ -1.2 + \eta_{01} = 0 \end{cases} \implies \begin{cases} \eta_{10} = 0 \\ \eta_{01} = 1.2 \end{cases}$$

It is possible to see that the eigenvectors associated to the saddle point of the expected fitness function are always $(1, -1)$ and $(-1, 1)$. We can see in Figure 4.8 that the eigenvector trajectories divide the domain in multiple basins. The gradient trajectories do not cross the eigenvector trajectories and can only end on the border.

Note that the gradient vector of the expected fitness function in the domain and on the border is never zero. This observation, in the domain,

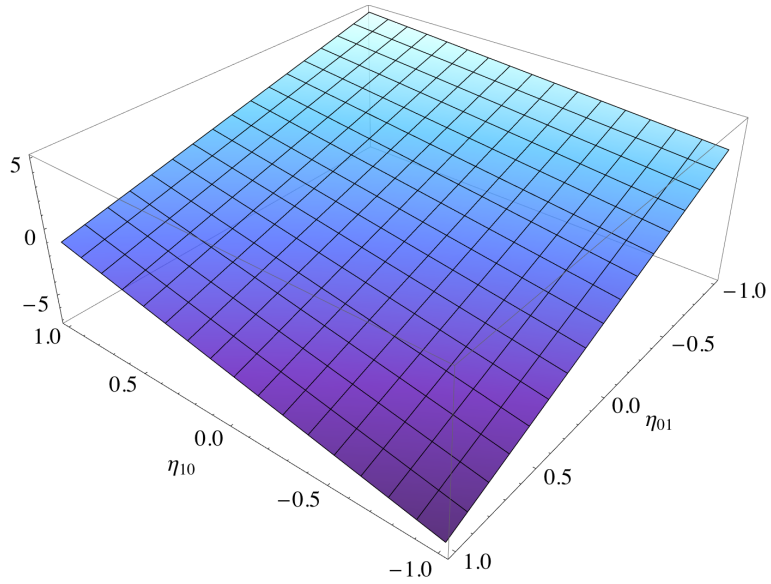

Figure 4.6: The expected value of f calculated over the independence model as a funciton of η_{10} and η_{01}

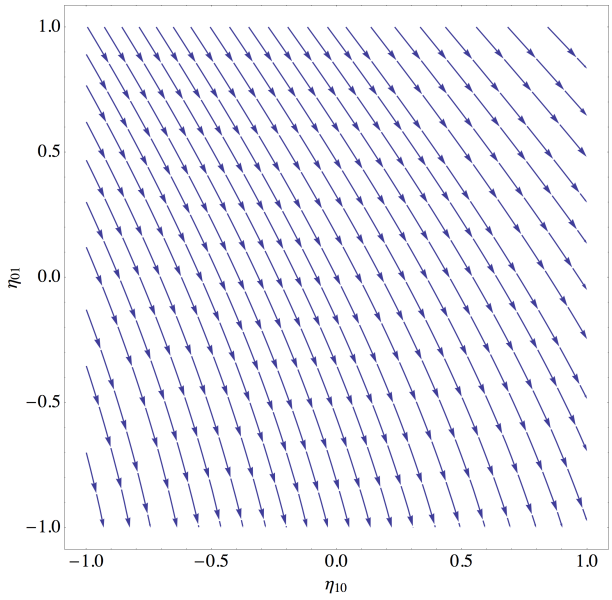

Figure 4.7: Trajectories of the Exact Gradient Descent search strategy for f in the η parametrization

derives from the absence of saddle points of the expected fitness function. On the border, we must consider the component of the gradient vector associated to the unconstrained coordinate. This component is the projection

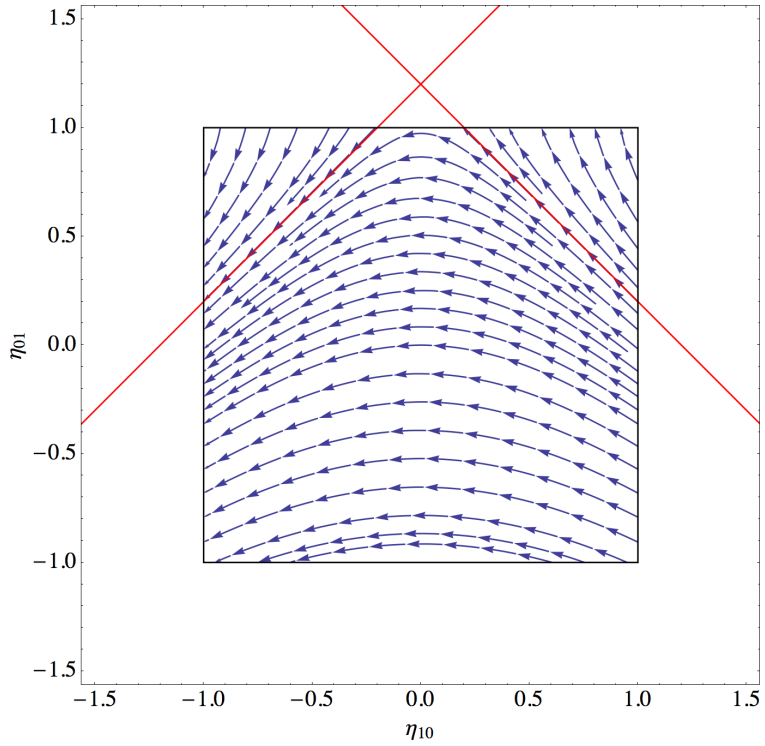

Figure 4.8: Trajectories of the Exact Gradient Descent search strategy for f in the η parametrization (The gradient trajectories on the border of the domain are not represented). In red, the eigenvector of the saddle point for $\mathbb{E}[f]$.

of the gradient on the border and, in general, can be zero only if the concerned edge links two vertices with the same fitness. This case is impossible if the edge is crossed by an eigenvector trajectory. This comes from the the values of the eigenvectors, which are always the same in the two variable case, as we have already seen. In particular, the eigenvector trajectories are straight lines, because the gradient is linear, and always cross the border with $|\frac{\pi}{4}|$ angle. This implies that the projection of the gradient on the border in correspondence of the intersection with the eigenvector trajectories is never zero, because the eigenvector trajectories never cross the border with perpendicular angle.

Now we consider the trajectories of the gradient between the basins. As we have seen in the example above, in the η parametrization, gradient trajectories of the expected fitness function are limited into their basin. However, the gradient trajectories reache always the border of the domain. Since the gradient can not be zero on the border if an eigenvector trajectory crosses it, then there exists always a gradient trajectory, driven by its projection on

the border, which crosses the eigenvector trajectory and reaches the basin of the optimum distribution.

Considering the last example, we can calculate that one of the eigenvector trajectories crosses the border in $\eta_{10} = 0.2, \eta_{01} = 1$. If we constrain the gradient vector of the expected fitness function on the border of the domain, where $\eta_{01} = 1$, we obtain:

$$\nabla \mathbb{E}[f(x)] = \begin{cases} 0 \\ -0.2 \end{cases}$$

The first component of the gradient, that corresponds to the projection of the gradient on the border, is not zero in $\eta_{10} = 0.2, \eta_{01} = 1$ and on all the edge, thus, as we have already seen, the gradient trajectory crosses the eigenvector trajectory on the border and reaches other basin.

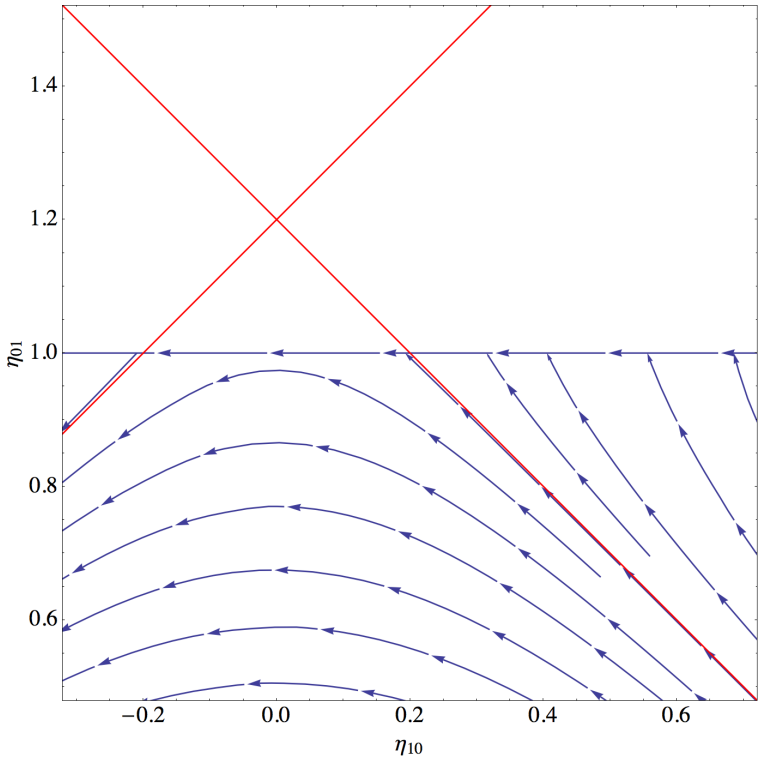

Figure 4.9: Projection of the gradient of the expected fitness function on the border of the domain in correspondence of $\eta_{10} = 0.2, \eta_{01} = 1$

Now, we reconsider the last example in the natural parametrization θ. It is possible to compute the θ coordinates for every point η belonging to the eigenvector trajectories previously found, applying the coordinate conversion (3.5) given in the previous chapter.

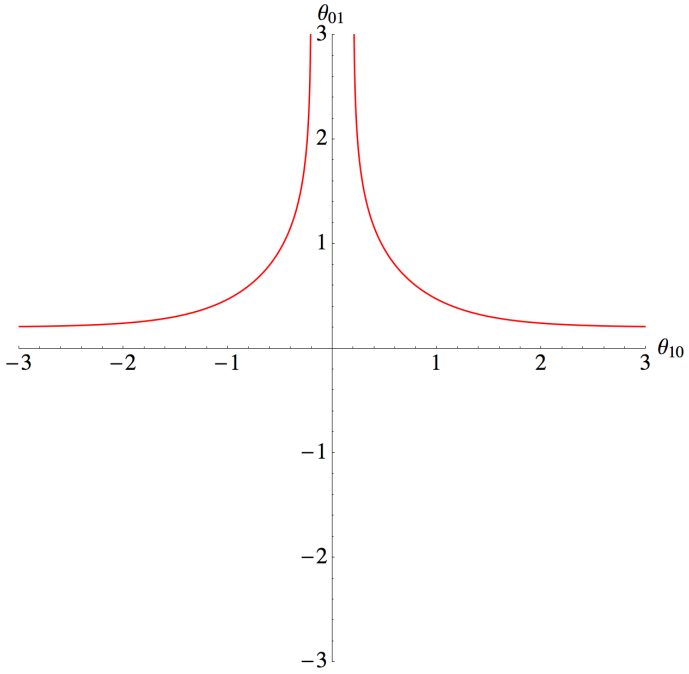

Figure 4.10: Eigenvector trajectories represented in θ parametrization by coordinate conversion

As we have already seen, the border, in the natural parametrization, is never reached, thus we expect to see that the gradient trajectories of the expected fitness function can not cross the eigenvector trajectories. However, in η and θ parametrizations the trajectories of the gradient of the expected fitness function are not the same, thus the eigenvector trajectories obtained in η parametrization and expressed in θ are not, in general, gradient trajectories in θ. The Figure 4.12 shows some gradient trajectories in θ found solving numerically the associated differential equations. As can be seen, the gradient trajectories can cross the η eigenvector trajectories. From an experimental analysis, results that the gradient trajectories can always reach the optimum basin in a finite time, starting from a finite values of θ coordinates. In addition, the time required to reach the optimum basin rises when the θ coordinates increase, because, in general, the module of the gradient vector decreases in the proximity of the reduced support distributions. Note that, from some starting points, the gradient trajectories reach the optimum basin covering all quadrants.

In conclusion, the effect of the eigenvector trajectories in the domain cause an increment of the time to reach the optimum basin from the gradient trajectories. We can define the general conditions on the position of the

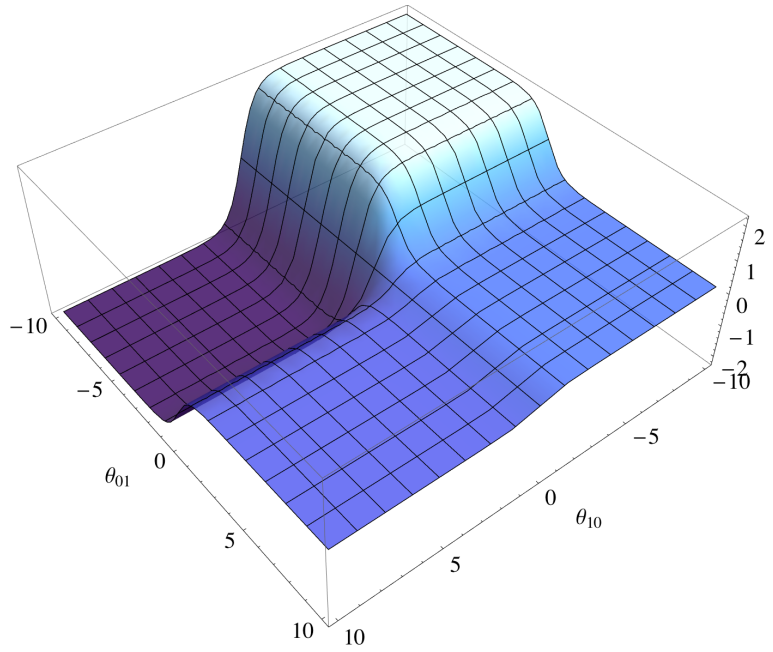

Figure 4.11: The expected value of f calculated over the independence model as a function of θ_{01} and θ_{10}

saddle point of the expected fitness function, in η parametrization, that guarantee a total absence of eigenvector trajectories in the domain, as:

$$\begin{cases} \eta_{01} < |\eta_{10}| - 2 \\ \eta_{01} > |\eta_{10}| + 2 \end{cases}$$

4.2.2 Three Variables Case

Expected Fitness Function with Saddle Points

Now we analyse the three variables case. The fitness function has the following form:

$$f(x) = c_{000} + c_{100}x_{100} + c_{010}x_{010} + c_{001}x_{001} + c_{110}x_{100}x_{010} +$$

$$+ c_{101}x_{100}x_{001} + c_{011}x_{010}x_{001} + c_{111}x_{100}x_{010}x_{001}$$

and the gradient vector of the expected fitness function corresponds to:

$$\nabla \mathbb{E}[f(x)] = \begin{cases} c_{001} + c_{101}\eta_{100} + c_{011}\eta_{010} + c_{111}\eta_{100}\eta_{010} \\ c_{010} + c_{110}\eta_{100} + c_{011}\eta_{001} + c_{111}\eta_{100}\eta_{001} \\ c_{100} + c_{110}\eta_{010} + c_{101}\eta_{001} + c_{111}\eta_{010}\eta_{001} \end{cases}$$

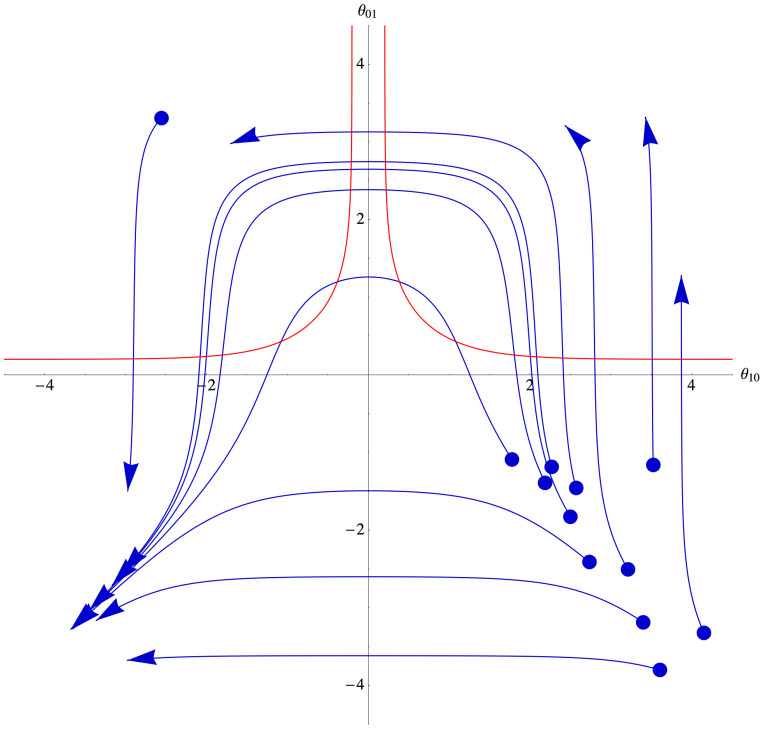

Figure 4.12: Gradient trajectories of the expected fitness function, represented in θ parametrization, when the saddle point is out of the domain

We take a specific example case with the following fitness function:

$$f(x) = -x_{100} - 2x_{010} - 4x_{001} + 16x_{100}x_{010}x_{001}$$

Note that it has no second order monomials. We can express the gradient vector as:

$$\nabla\mathbb{E}[f(x)] = \begin{cases} -4 + 16\eta_{100}\eta_{010} \\ -2 + 16\eta_{100}\eta_{001} \\ -1 + 16\eta_{010}\eta_{001} \end{cases}$$

We verified the presence of the saddle points, finding the zeros of the expected value of the fitness function. In this example there are the following two saddle points:

$$\left\{(\eta_{100} = -\frac{1}{\sqrt{2}}, \eta_{010} = -\frac{1}{2\sqrt{2}}, \eta_{001} = -\frac{1}{4\sqrt{2}}),\right.$$

$$\left.(\eta_{100} = \frac{1}{\sqrt{2}}, \eta_{010} = \frac{1}{2\sqrt{2}}, \eta_{001} = \frac{1}{4\sqrt{2}})\right\}$$

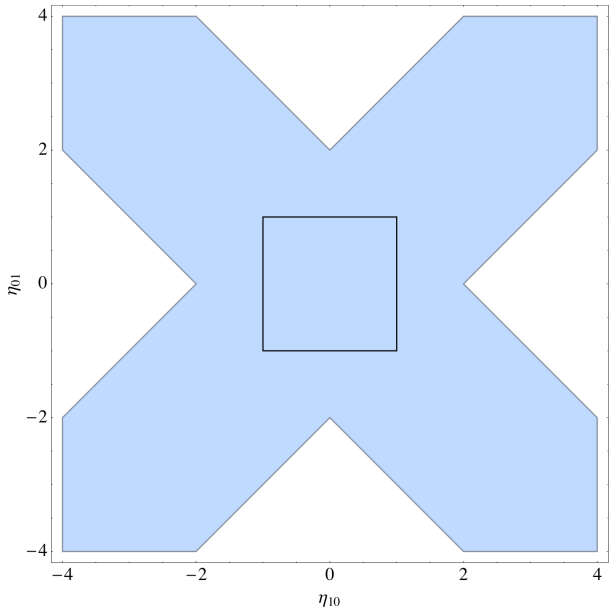

Figure 4.13: Saddle points in the highlighted region have eigenvectors not crossing the η parameters domain.

The Figure 4.14 represents some trajectories of the Exact Gradient Descent over the independence model that show the behaviour of this search strategy. This specific example shows that, when the number of variables is greater than two, the expected fitness function can have more than one saddle point. This observation is important, because the presence of multiple critical points implies, in general, a presence of multiple local minima. If the gradient vector of the expected fitness function is set to zero, in fact, we obtain a second order equation.

$$\nabla\mathbb{E}[f(x)] = \begin{cases} c_{001} + c_{101}\eta_{100} + c_{011}\eta_{010} + c_{111}\eta_{100}\eta_{010} = 0 \\ c_{010} + c_{110}\eta_{100} + c_{011}\eta_{001} + c_{111}\eta_{100}\eta_{001} = 0 \\ c_{100} + c_{110}\eta_{010} + c_{101}\eta_{001} + c_{111}\eta_{010}\eta_{001} = 0 \end{cases} \implies$$

$$\implies \begin{array}{c} (-c_{110}c_{101}c_{111} - c_{100}c_{111}^2)\eta_{10}^2 + (2c_{010}c_{101}c_{111} - 2c_{100}c_{011}c_{111})\eta_{10} + \\ +(-c_{001}c_{110}c_{011} + c_{010}c_{101}c_{011} - c_{100}c_{011}^2 - c_{010}c_{001}c_{111}) = 0 \end{array}$$

Saddle Points of Constrained Independence Submodels

We consider a three variables example with the following fitness function:

$$f(x) = x_{100} + 2x_{010} + 10x_{001} - 4x_{100}x_{010}$$

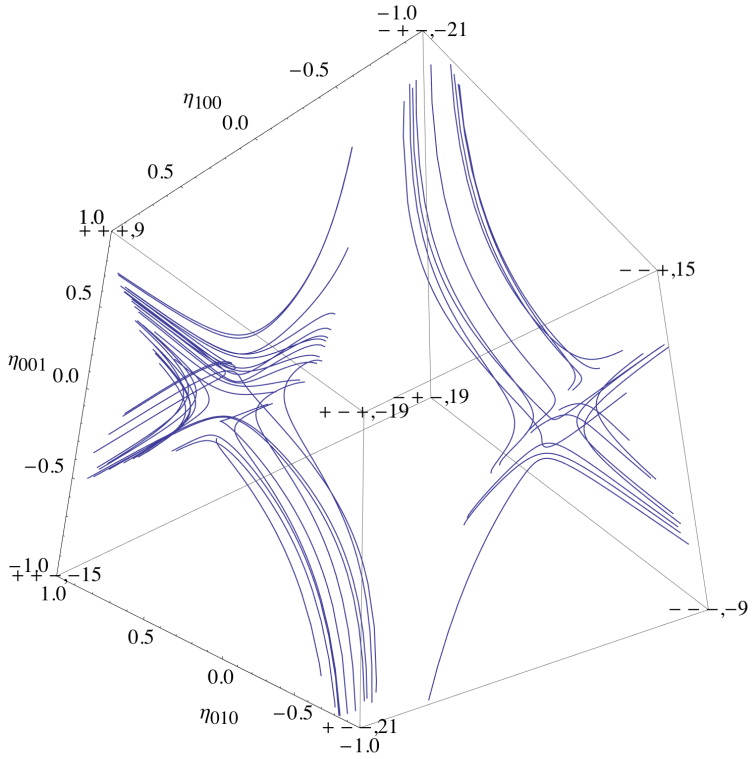

Figure 4.14: Trajectories of the Exact Gradient Descent search strategy for f in the η parametrization

Note that the function has few monomials and no third order interactions. The gradient vector of the expected fitness function is:

$$\nabla \mathbb{E}[f(x)] = \begin{cases} 10 \\ 2 - 4\eta_{100} \\ 1 - 4\eta_{010} \end{cases}$$

As it can be seen, the gradient vector is constant in relation with the third variable, thus the gradient module can never be zero and the expected fitness function has not saddle points. The Figure 4.15 shows the cube that represents the neighborhood relationships between the elements of the domain, based on the hamming distance. On the vertices of the cube, the corresponding fitness values are indicated. Note that in this case, every vertex on the cube face $\{(+-+), (++-), (-+-), (---)\}$ has a lower fitness value with respect to each vertex of the opposite face. This forces the gradient trajectories to be directed from the first face to the opposite, thus the first gradient component is always grater then zero (in this case constant).

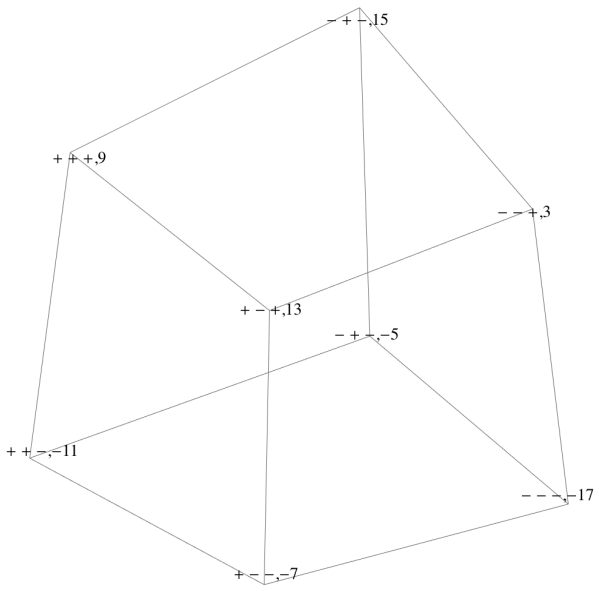

Figure 4.15: Representation of the neighborhood relationships between the elements of the domain, based on the hamming distance. (On the vertices, the fitness values are indicated

As it can be seen in the Figure 4.16, the trajectories that starts from any point of the domain end on the face $\{(+-+),(+++),(-++),(--+)\}$. Now we limit the analysis of the problem on the independence submodels

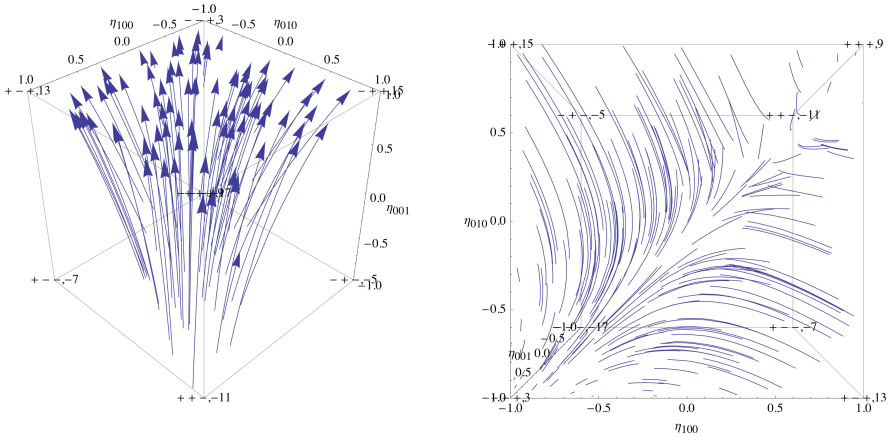

Figure 4.16: Two views of the Exact Gradient Descent trajectories in the η parameter space.

constrained on the cube face $\{(+-+), (+++), (-++), (--+)\}$, where the variable x_{001} is set to 1. This corresponds to consider, for some aspects, the two variable problem defined by the fitness value of the face. With x_{001} fixed to 1, the first component of the gradient vector is zero, thus we can limit our analysis to the last two components. We calculate the position of the saddle point of the expected fitness function on the face of the constrained independence submodel, as do in the two variable case, thus:

$$\nabla\mathbb{E}[f(x)] = \begin{cases} 2 - 4\eta_{100} = 0 \\ 1 - 4\eta_{010} = 0 \end{cases} \implies \begin{cases} \eta_{100} = \frac{2}{4} \\ \eta_{010} = \frac{1}{4} \end{cases}$$

The saddle point of the consider submodel is in the domain and, as we have previously seen, not represent a saddle point of the independence submodel, however it can create two different attraction basins as shown in the Figure 4.16. This example is very significant, because reveals that the saddle points of the expected fitness function of the constrained independence submodels, in absence of saddle points of the expected fitness function of the independence model, can create different attraction basins, thus an Exact Gradient Descent search strategy can converge to a local minima.

4.3 On the Position of the Saddle Points

In this chapter we present some simple proofs about the two variable fitness functions and the shape of their expected fitness over the independence model. We consider these considerations useful to get a deeper understanding of the relation between fitness functions, expected fitness over a submodel and convergence abilities of exact gradient descent search strategies.

4.3.1 High Order Interactions

We have seen that the coefficients c_α with $|\alpha| > 1$ represent interactions between variables in the fitness function. The more these coefficients are high the more the non-linearity of f influences the behaviour of a search strategy.

We have already introduced the conditions that have to hold for the saddle point to be inside the parameters domain in the two variable case. In the following figure the locus of the saddle point as a function of c_{11} has been plotted for the two variables function $f = x_{10} + 2x_{01} + c_{11}x_{10}x_{01}$. When the interaction between x_{10} and x_{01} is high enough then a saddle point appears in the domain and search strategies could fail to converge to the global optimum.

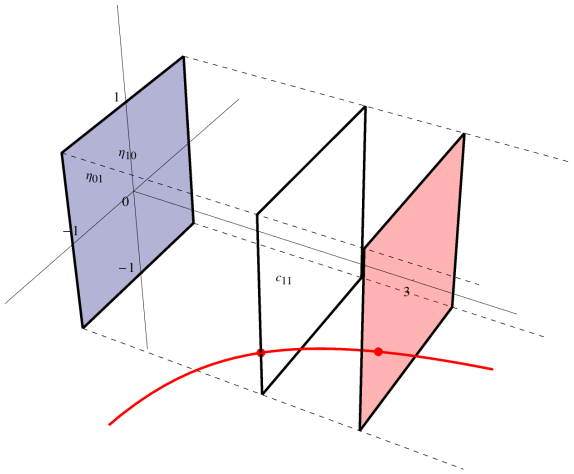

Figure 4.17: Locus of the saddle point as a function of c_{11}

For the three variables case the conditions on the coefficients for a saddle to exist inside the parameters domain become much more complex. In fact there are eight coefficients involved and the equations $\nabla\mathbb{E}[f] = 0$ are non linear. However we have seen in an example in previous sections how an high c_{111} coefficient could lead to the presence of even two saddle points inside the η domain. Thus we conjecture that the shape of the expected fitness calculated on a sub-model is strictly related to strength of the dependencies between variables in the fitness function that are not captured by the considered model.

4.3.2 Again on the Saddle Point Position

In this section we focus on the behaviour of a greedy search strategy when its search space is restricted to the border of the independence model and we show that it converge to the global optimum for f if and only if the saddle point for $\mathbb{E}_{\mathcal{M}}[f]$ is outside the domain of the η parameters.

As it is presented in detail in the following chapter, the independence model covers a two dimensional surface in the three dimensional space of the probability simplex. This surface includes four out of the six edges, in particular, are included in the model the edges for which the variables assignments have hamming distance 1. For example, the edge connecting $(1, 1)$ and $(1, -1)$ is included in the independence model, while the one from $(1, 1)$ to $(-1, -1)$ is not. It is possible to see that the expected value of f over these edges is a linear combination of the fitness of the two edges, depending on the position, and that the gradient is constant and directed towards the

vertex with higher fitness. Consider the function f whose fitness values are shown in the table below. A graph can be used to show the dynamics of an Exact Gradient Descent strategy when its search space is reduced to the borders of the independence model. As we can see in the Figure 4.18 EGD not converge to the global optimum for certain starting condition.

x_{10}	x_{01}	$f(x_{10}, x_{01})$
-1	-1	2
-1	1	3
1	-1	4
1	1	1

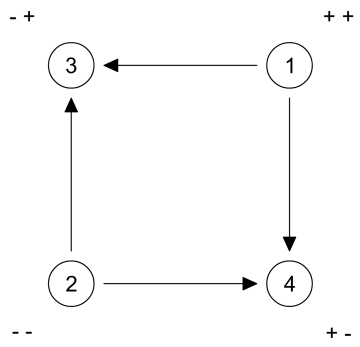

Figure 4.18: Graph representation of the dynamics on the border of the independence model

It is possible to compute the c coefficients for f and check for saddle presence in the interior of the η domain. The coefficients are $c = (\frac{5}{2}, 0, -\frac{1}{2}, -1)$ and the saddle points coordinates are $\eta = (-\frac{1}{2}, 0)$. As we have show in the previous sections, in this case EGD does not converge to the global optimum for f for certain starting conditions. We have also seen that this could hold even for more advanced strategies.

Now we show that since the expected value of f calculated on the independence model has a saddle point in the interior of the η domain the same holds for every function that has the same dynamics as f on the border of the model. This is done by showing that the saddle point for f does not move outside the η domain if the fitness values of f are perturbated preserving the directions of travel of EGD on the border of the independence model.

First consider a function f' for which $f'(x_{10}, x_{01}) = \gamma f(x_{10}, x_{01})$. It holds that $c_{\alpha,f'} = c_{\alpha,f}$. The γ factors cancel in the expression of the saddle

point position for f', thus it is the same as for f. This shows that scaling all the values of the fitness does not change the position of the saddle point.

Consider now the class F' of two variables functions with the same image values of f except for f_{--}. To preserve the dynamics on the border of the independence model f_{--} can assume all the values in the range $(f_{-+}, -\infty)$. Note that if $f_{--} = f_{-+}$ the dynamics on the border are not the same since on the edge $(-1,-1),(-1,1)$ the gradient would be null in every point. We can give the saddle point coordinates as a function of f_{--}:

$$\overline{\eta} = \left(\frac{-f_{++} + f_{+-} - f_{-+} + f_{--}}{f_{++} - f_{+-} - f_{-+} + f_{--}}, \frac{-f_{++} - f_{+-} + f_{-+} + f_{--}}{f_{++} - f_{+-} - f_{-+} + f_{--}}\right)$$

For functions belonging to F' we have

$$\overline{\eta} = \left(\frac{f_{--}}{-6 + f_{--}}, \frac{-2 + f_{--}}{-6 + f_{--}}\right)$$

The locus of the saddle point has been plotted as a function of f_{--} in Figure 4.19. It is always inside the η domain. Note that if f_{--} is eliminated and η_2 is expressed as a function of η_1 it holds that $\eta_2 = (1 + 2\eta_1)/3$.

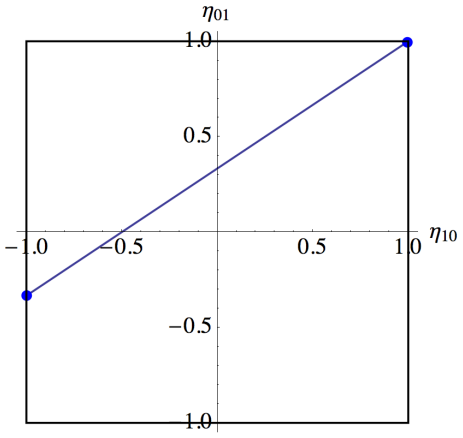

Figure 4.19: Locus of the saddle points of f' as a function of f'_{--}

This can be proven formally. One has to show that

$$\begin{cases} |\overline{n_1}| = |\dfrac{a + f_{--}}{b + f_{--}}| \leq 1 \\ |\overline{n_2}| = |\dfrac{c + f_{--}}{b + f_{--}}| \leq 1 \end{cases}$$

where $a = -f_{++} + f_{+-} - f_{-+}$, $b = f_{++} - f_{+-} - f_{-+}$ and $c = f_{++} - f_{+-} + f_{-+}$, for every f_{++}, f_{+-} and f_{-+} such as

$$f_{+-} > f_{-+} > f_{++} > f_{--} \tag{4.1}$$

We have gone through all the calculations and found out that a sufficient condition for this to hold is:

$$a > b < c \ \wedge \ f_{--} \leq \frac{1}{2}(-a-b) \ \wedge \ f_{--} \leq \frac{1}{2}(-b-c)$$

It is possible to see expanding a, b and c that this condition holds for every fitness assignment that preserves (4.1). Note that the strict inequalities translates immediately in the requirements on f. For example

$$c > b \Rightarrow f_{++} < f_{-+}$$

Similar results can be derived for f_{-+}, f_{+-} and f_{++}. This result, along with the global scaling property stated before, can be used to reconduct every function for which EGD behaves like in Figure 4.18, when its search space is restricted to the border of the independence model, to the case of f. Since $\mathbb{E}[f]$ calculated on the independence model has a saddle point in the interior of the η domain the same holds for every such a function. This shows that for the two variables case the presence of local minima for EGD on the border of the independence model implies the presence of a saddle point in $\mathbb{E}[f]$ and thus that EGD and other search strategies converge to a local optima for f when the entire independence model is given as the search space.

There are other possible dynamics that involve a saddle point in the expected fitness but it is possible to see that they are just rotations or symmetries of the f case.

4.3.3 The Attraction Basin Which Includes the Uniform Distribution

Consider an exact gradient descent search strategy with the uniform distribution as the initial condition. In this section we prove that this strategy converges to the global optimum for every two variables fitness function.

We consider a generic two variable problem, with the optimum in $(-1, 1)$. Note that the choice of optimum element is arbitrary, and the demonstration does not lose generality. The fitness values can be express by the sum of the coefficients c_α with the proper signs as follows:

$$\begin{aligned} f_{--} &= c_{00} - c_{10} - c_{01} + c_{11} \\ f_{-+} &= c_{00} - c_{10} + c_{01} - c_{11} \\ f_{+-} &= c_{00} + c_{10} - c_{01} - c_{11} \\ f_{++} &= c_{00} + c_{10} + c_{01} + c_{11} \end{aligned}$$

As already seen in the previous results, if the problem has a saddle point in the domain, the attraction basins are opposite, i.e., $x_{1\,opt} = -x_{1\,subopt}, x_{2\,opt} = -x_{2\,subopt}$. In our specific example, the problem has a local minima in $(1, -1)$. We can express some conditions on the fitness of the elements of the domain, based on the knowledge of the optimum position, as follows:

$$\begin{cases} f_{-+} > f_{+-} \\ f_{-+} > f_{--} \\ f_{+-} > f_{--} \\ f_{-+} > f_{++} \\ f_{+-} > f_{++} \end{cases} \implies$$

$$\implies \begin{cases} c_{00} - c_{10} + c_{01} - c_{11} > c_{00} + c_{10} - c_{01} - c_{11} \\ c_{00} - c_{10} + c_{01} - c_{11} > c_{00} - c_{10} - c_{01} + c_{11} \\ c_{00} + c_{10} - c_{01} - c_{11} > c_{00} - c_{10} - c_{01} + c_{11} \\ c_{00} - c_{10} + c_{01} - c_{11} > c_{00} + c_{10} + c_{01} + c_{11} \\ c_{00} + c_{10} - c_{01} - c_{11} > c_{00} + c_{10} + c_{01} + c_{11} \end{cases} \implies$$

$$\implies \begin{cases} c_{01} > c_{10} \\ c_{01} > c_{11} \\ c_{10} > c_{11} \\ -c_{10} > c_{11} \\ -c_{01} > c_{11} \end{cases} \implies \begin{cases} c_{01} > c_{10} > c_{11} \\ -c_{10} > c_{11} \\ -c_{01} > c_{11} \end{cases}$$

The last condition of the coefficients c_α permits to characterize the following three cases on the position of the saddle point (remembering that its coordinates are $\eta_{10} = -\frac{c_{01}}{c_{11}}$, $\eta_{01} = -\frac{c_{10}}{c_{11}}$):

1.

$$\begin{cases} c_{01} > 0 \\ c_{10} > 0 \\ c_{11} < 0 \\ c_{01} > c_{10} > c_{11} \\ -c_{10} > c_{11} \\ -c_{01} > c_{11} \end{cases} \implies \begin{cases} \frac{c_{01}}{c_{11}} < 0 \implies \eta_{10} > 0 \\ \frac{c_{10}}{c_{11}} < 0 \implies \eta_{01} > 0 \\ |c_{01}| > |c_{10}|, \frac{c_{01}}{c_{10}} > 1 \implies \frac{\eta_{10}}{\eta_{01}} > 1 \end{cases}$$

The coordinate η_{10} and η_{01} are in the first quadrant and the position of the saddle point is always under the bisector, thus the uniform distribution is in the optimum basin.

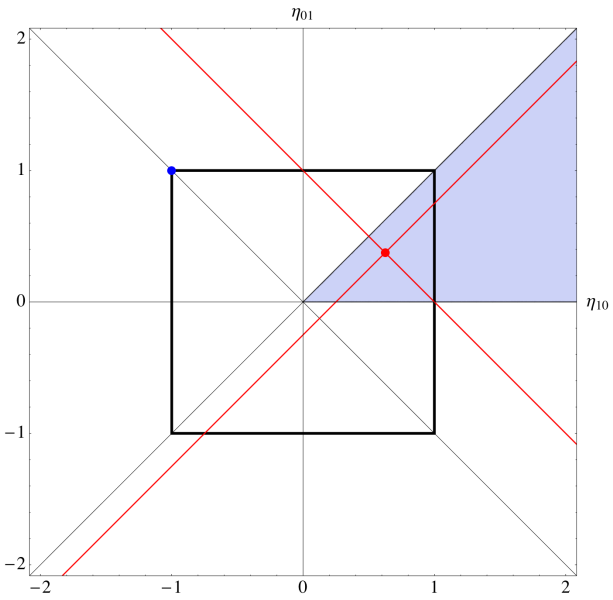

Figure 4.20: Locus of saddle point when $c_{01} > 0, c_{10} > 0, c_{10} < 0$

2.

$$\begin{cases} c_{01} > 0 \\ c_{10} < 0 \\ c_{11} < 0 \\ c_{01} > c_{10} > c_{11} \\ -c_{10} > c_{11} \\ -c_{01} > c_{11} \end{cases} \implies \begin{cases} \frac{c_{01}}{c_{11}} < 0 \implies \eta_{10} > 0 \\ |c_{10}| < |c_{11}|, \frac{c_{10}}{c_{11}} < 1 \implies 0 > \eta_{01} > -1 \end{cases}$$

The coordinate η_{10} and η_{01} are in the second quadrant, thus the uniform distribution is always in the optimum basin.

3.

$$\begin{cases} c_{01} < 0 \\ c_{10} < 0 \\ c_{11} < 0 \\ c_{01} > c_{10} > c_{11} \\ -c_{10} > c_{11} \\ -c_{01} > c_{11} \end{cases} \implies \begin{cases} |c_{01}| < |c_{11}|, \frac{c_{01}}{c_{11}} < 1 \implies \eta_{10} < 0, \eta_{10} > -1 \\ |c_{10}| < |c_{11}|, \frac{c_{10}}{c_{11}} < 1 \implies \eta_{01} < 0, \eta_{01} > -1 \\ |c_{01}| < |c_{10}|, \frac{c_{01}}{c_{10}} < 1 \implies 0 < \frac{\eta_{10}}{\eta_{01}} < 1 \end{cases}$$

The coordinate η_{10} and η_{01} are in the third quadrant and the position of the saddle point is always under the bisector, thus the uniform distribution is always in the optimum basin.

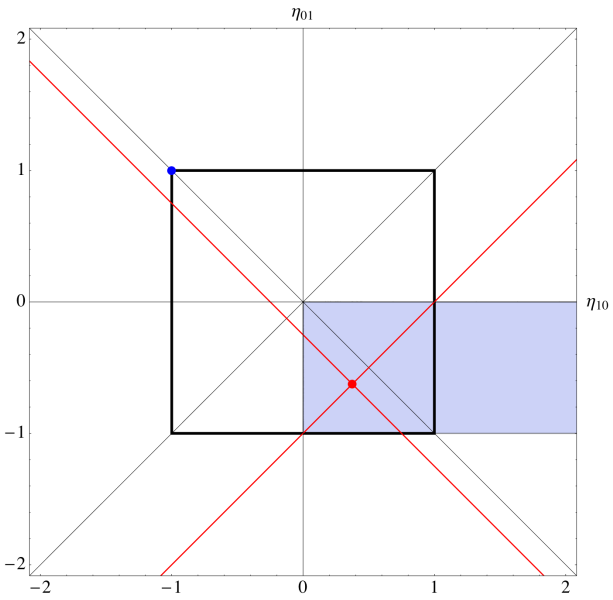

Figure 4.21: Locus of saddle point when $c_{01} > 0, c_{10} < 0, c_{10} < 0$

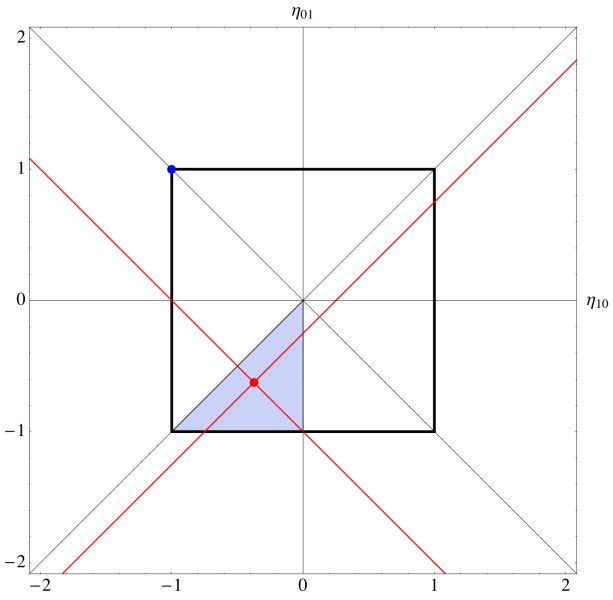

Figure 4.22: Locus of saddle point when $c_{01} < 0, c_{10} < 0, c_{10} < 0$

Note that this does not imply that we have found the perfect search strategy since calculating the exact gradient vector $\nabla \mathbb{E}_{\mathcal{M}}[f]$ has computational complexity exponential in the number of variables.

4.3.4 The Locus of the Saddle Point in PBIL

In this section we determine analytically the position of the saddle point for the dynamics of the PBIL algorithm under the infinite population assumption.

Consider a pseudo-boolean function f and suppose its optimum and sub-optimum are located, without loss of generality, in $(1,1)$ and $(-1,-1)$. We have shown in the previous sections that since these two value assignments have no variable value in common, $\mathbb{E}[f]$ has a saddle point.

Let s be a starting independent distribution characterized by its marginal probabilities. Here we use the following short-cut notation for the marginal probabilities: $p_1 = p(x_{10} = 1)$ and $p_2 = p(x_{01} = 1)$. Since the infinite population hypothesis holds, $R[(1,1)] = p_1 p_2$, $R[(1,1)] = (1-p_1)(1-p_2)$ and so on. Here with $R[\overline{x}]$ we mean the fraction of individuals $\overline{x}$ in the population.

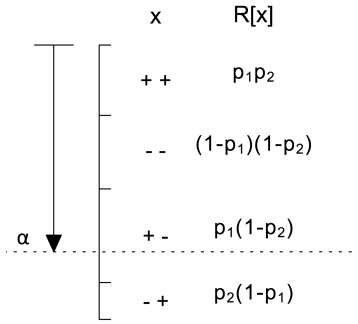

Figure 4.23: Representation of the population after sampling and before selection

The composition of the population after selection is shown in Figure 4.23 where we have assumed, again without loss of generality, that $f_{+-} > f_{-+}$. Selection discards the last $1-\alpha$ elements, where α is the selection rate. After selection the new distribution is estimated from the resulting population. Suppose it holds

$$p_1 p_2 < \alpha \leq p_1 p_2 + (1-p_1)(1-p_2) \tag{4.2}$$

This means that the selected population contains all the optimal individuals and part of the sub-optimal ones. All the individuals of the types $(1,-1)$ and $(-1,1)$ are discarded. The Max-Likelihood estimators of p_1 and p_2 thus are

$$\hat{p}_1 = \frac{p_1 p_2}{\alpha} \quad \hat{p}_2 = \frac{p_1 p_2}{\alpha}$$

We are looking for distributions such as $\hat{p}_1 = p_1$ and $\hat{p}_2 = p_2$, or, in other words, the initial distributions such as the process of sampling, selection and estimation returns the same marginal probabilities. This holds when $p_1 = p_2 = \alpha$. Substituting this in the condition (4.2) we have that

$$\alpha^2 < \alpha \leq \alpha^2 + (1-\alpha)^2$$

that holds for every $\alpha \in (0,\ 1/2]$. This means that for every alpha in this range the distribution with $p_1 = p_2 = \alpha$ is a saddle point for PBIL. Consider now another case, suppose it holds that

$$p_1 p_2 + (1-p_1)(1-p_2) \leq \alpha \leq p_1 p_2 + (1-p_1)(1-p_2) + p_1(1-p_2) \quad (4.3)$$

this means that the selection preserves all the $(1,1)$, $(-1,-1)$ individuals and part of the $(1,-1)$ ones. Call $k = \alpha - p_1 p_2 - (1-p_1)(1-p_2)$ the number of the $(1,-1)$ individuals. The Max-Likelihood estimator this time reads as

$$\hat{p}_1 = \frac{p_1 p_2 + k}{\alpha} \quad \hat{p}_2 = \frac{p_1 p_2}{\alpha}$$

The distributions such as $\hat{p}_1 = p_1$ and $\hat{p}_2 = p_2$ are the one for which $a = p_1 = 1 - p_2$. Substituting this into (4.3) we have that $1/2 \leq \alpha < 1$.

We have shown that if it holds that $f_{++} < f_{--} < f_{+-} < f_{-+}$ for every selection rate α there is a distribution s such as the gradient perceived by the PBIL algorithm in s is null. We have gone through similar calculations and derived similar results for the other possible fitness configuration that imply a saddle point in the expected fitness on the independence model. Moreover it can be shown with the same technique that the PBIL gradient is never null if the function f has no saddle point inside the parameters domain.

This example shows how the convergence properties of real search strategies can be influenced by the landscape of the expected value of the fitness function f.

4.4 Conclusions

In this section we have presented an in depth analysis of the expected value of a pseudo-boolean function f calculated over the independence model for some two and three variables example. In those cases we have derived the exact expressions of $\mathbb{E}[f]$ in the η and θ parametrization, we computed the positions of the critical points and in some cases we were able to give conditions on the coefficients c_α for these to exist in the parameters domain.

At the same time we analysed the behaviour of a theoretical search strategy based on the exact gradient descent of $\mathbb{E}[f]$, trying to relate its performance with the shape of this function. We observed that the presence of critical points in $\mathbb{E}[f]$ could prevent this strategy to converge to the global optimum for certain starting condition. We have seen that the same happens for a real search strategy, PBIL, which explicitly uses the same probability model $\nabla\mathbb{E}[f]$ was calculated on.

In the last section we proved, at least for the two variable case, the importance of analysing the behaviour of a greedy search strategy confined on the border of the independence model along with other minor results about the structure of $\mathbb{E}[f]$ and the position of the saddle point for the PBIL strategy in the two variable cases.

Even if the mathematical tools employed in this analysis were not powerful enough to gives results for higher dimensional cases, the study of these examples enabled us to achieve deeper understanding about the problem and form the basis for the development of the rest of this work.

Chapter 5

Transformations of Variables

We have seen in the previous chapter how the performances of evolutionary search strategies could be influenced by the critical points in the expected value of the function f to be optimized. We have related this fact to the presence of more than one attractors for the Exact Gradient Descent search strategy. In this chapter we introduce the class $\mathcal{T}$ of the one-to-one maps between elements of the search space Ω and its sub-class $\mathcal{L}^k$. We discuss the idea of composing the fitness function with maps in $\mathcal{T}$ or $\mathcal{L}^k$ and we analyse the effects of these compositions on the expected value of f. We show how improvements in the performance of EGD strategy can be achieved if the proper map is chosen. The concepts and the ideas discussed in this chapter forms the core of this work.

5.1 Concepts and Definitions

In this section the class $\mathcal{T}$ of one-to-one maps between elements in Ω is introduced.

5.1.1 The Idea

Consider a pseudo-boolean function f, defined over the domain $\Omega = \{-1, 1\}^n$, and the one-to-one function

$$T(x) : \Omega \to \Omega$$

This map associates a variable assignment y to every $x \in \Omega$. We are interested in the composition of the function f with T, i.e., in the pseudo-boolean function g

$$g : \Omega \to \mathbb{R}$$

$$g(y) = f \circ T = f(T(x))$$

Let us discuss an example.

$f(x_{10}, x_{01})$	x_{10}	x_{01}	$T[(x_{10}, x_{01})]$	y_{10}	y_{01}	$g(y_{10}, y_{01})$
2	-1	-1	$\Leftrightarrow$	-1	1	2
3	-1	1	$\Leftrightarrow$	1	-1	3
4	1	-1	$\Leftrightarrow$	1	1	4
1	1	1	$\Leftrightarrow$	-1	-1	1

The two functions f and g are different and thus is natural to expect that the same holds for their coefficients vectors c and d. In fact the two coefficients vector $c_\alpha = \left(\frac{5}{2},\ -\frac{1}{2},\ 0,\ -1\right)$ and $d_\alpha = \left(\frac{5}{2},\ \frac{1}{2},\ 1,\ 0\right)$.

The function g has the same multi-set of images as f. More precisely, consider the following set of variables assignments

$$N_f(a) = \{x \in \Omega | f(x) = a\}$$

It holds that

$$|N_f(a)| = |N_g(a)| \ \forall a \in \mathbb{R} \tag{5.1}$$

This comes from the fact that T is one-to-one. Obviously these sets have cardinality greater than zero for at most 2^n different fitness values a. Intuitively, T simply permutes the values of f over the domain Ω.

The condition (5.1) allow us to introduce a notion of equivalence between pseudo-boolean functions defined over n variables:

$$f \equiv_{\mathcal{T}} g \Leftrightarrow \forall a \in \mathbb{R} \ \ |N_f(a)| = |N_g(b)|$$

where $\mathcal{T}$ is the class of all the possible maps T. Essentially, we say that f is $\mathcal{T}$-equivalent to g if the values of g are a permutation of the f ones with respect to the domain Ω, that is, g can be written as the composition of f with an appropriate map T. Note that if $g = f \circ T$ then $f = g \circ T^{-1}$. The existence and the uniqueness of the inverse of T map is guaranteed by the fact that T has been defined to be one-to-one.

The introduction of this notion of equivalence is motivated by the fact that for every map T and every functions $f, g = f \circ T$ it holds that

$$\min_{x \in \Omega} f(x) = T^{-1}\big[\min_{y \in \Omega} g(y)\big]$$

so to the ends of finding a solution for the problem (P) it makes no difference if we consider the function f or any other $g \equiv_T f$. Let us show with an example how the composition of the function f to be optimized with a proper map T affects the stochastic relaxation of f. Consider again the functions introduced in the previous example. Their expected value on the independence model read, as always, as

$$\mathbb{E}[f] = \tfrac{5}{2} - \tfrac{1}{2}\eta_{01} - \eta_{01}\eta_{10} \quad \mathbb{E}[g] = \tfrac{5}{2} + \tfrac{1}{2}\eta_{01} + \eta_{10}$$

Remember the coefficients vector calculated before. The original function expected value has a saddle point in $\eta = (-\frac{1}{2},\, 0)$ while $\mathbb{E}[g]$ is linear in the η parametrization. This means that there are respectively two and one attractors for the Exact Gradient Descent strategy. Remembering the results of the previous chapter we can conclude that evolutionary search strategies perform better optimizing g than f.

Note that this improvement depends entirely on the specific map applied. The one we applied here produced good results but this is not true in general. Consider the inverse example, i.e., g is the function to optimize and T^{-1} is the proposed map. This time opposite effects are obtained: the expected value of $f = g \circ T^{-1}$ with respect to the independence model has a saddle point inside the η domain while the g one is linear in the η parametrization.

5.1.2 T as Vector-valued Pseudo-boolean Function

Consider again the map $T(x) : \Omega \to \Omega$. It is possible to express $T(x)$ as a vector-valued function

$$T(x) = \Big(t_1(x),\, t_2(x),\, ..., t_n(x)\Big) \tag{5.2}$$

In the following we use the usual multi-index notation for the sub-elements of T, which are ordered like the x variable vectors. So for example $y_{010} = t_{010}\big(x_{100},\, x_{010},\, x_{001}\big)$.

Every t_h is a function defined on the vector of binary variables x and has $\{-1, 1\}$ as image set. They are pseudo-boolean functions and the usual expansion holds:

$$t_h = \sum_{\alpha \in I} c_{h,\alpha} x^{\alpha}$$

Thus every possible map can be characterized by the n vectors of 2^n coefficients $c_{h,\alpha}$. In the following c_h is the coefficient vector of h-th component of T and $c_{h,\alpha}$ is the α coefficient in the multi-linear expansion of $t_h(x)$, so for example $c_{010,011}$ is the coefficient of the monomial $x_{010}x_{001}$ in the expansion

of t_{010}. We sometimes use the intuitive short-cut $100 = 1$, $010 = 2$ and so on for multi-indices with cardinality one.

Note that this expansion of T is very expressive. Only few assignments for the t_h coefficient vectors represent valid one-to-one maps belonging to class $\mathcal{T}$. We deal with this point later.

The expression (5.2) allow us to compute directly the expression of $f \circ T$ starting from the monomial expansion of f.

$$f\Big(T(x)\Big) = \sum_{\alpha \in I} c_\alpha \prod_{h \in 1 \ldots n} t_h(x)^{\alpha_h} \tag{5.3}$$

Intuitively, one has to substitute every instance of the variable x_h in the expansion of f with the k-th component of the vector-valued function $T(x)$. Consider the following two variables example:

$$T(x_{10}, x_{01}) = \Big(- x_{10}x_{01},\ -x_{01}\Big)$$

$$c_{10} = \Big(0, 0, 0, -1\Big) \quad c_{01} = \Big(0, -1, 0, 0\Big)$$

x		$T(x)$	
x_{10}	x_{01}	$-x_{10}x_{01}$	$-x_{01}$
-1	-1	-1	1
-1	1	1	-1
1	-1	1	1
1	1	-1	-1

It can be seen inspecting the table that this coefficients assignment makes T one to one. We now derive the expansion of $g = f \circ T$ using (5.3).

$$f = c_{00} + c_{10}x_{10} + c_{01}x_{01} + c_{11}x_{10}x_{01}$$

$$g = f \circ T = c_{00} + c_{10}\Big(- x_{01}x_{10}\Big) + c_{10}\Big(- x_{01}\Big) + c_{11}\Big(- x_{01}x_{10}\Big)\Big(- x_{01}\Big) =$$

$$= c_{00} - c_{10}x_{01} + c_{11}x_{10} - c_{01}x_{10}x_{01}$$

Remember that $x_\alpha^2 = 1$. So we have that the g coefficients vector is $d = (c_{00},\ c_{11},\ -c_{10},\ -c_{01})$.

In this section we have presented the class of the one-to-one variable maps $\mathcal{T} : \Omega \to \Omega$ and the idea that composing elements of this class with the fitness function to be optimized. Hints were given on how the composition with the proper map $T \in \mathcal{T}$ could lead to improvements in the performances of EGD.

5.2 Probability Models and the $\mathcal{T}$ Maps

In this section we show how the variable maps of the class $\mathcal{T}$ can be used to define probability models over the binary variables x.

Consider the binary variables vector x, a variable map $T \in \mathcal{T}$ and the vector $y = T(x)$. x and y both take values in Ω. Let now $\mathcal{M}_T$ be a probability model defined over the transformed variables y. It still defines a family of distribution with Ω as support, so for example sampling from the uniform distribution over the y variables we obtain, on average, every element of Ω with equal probability. We are interested in which is the probability model obtained if the map T^{-1} is applied on the support of distributions in $\mathcal{M}_T$, i.e., in the distributions over the x variables such as

$$p(x = T^{-1}[\overline{y}]) = p(y = \overline{y}) \quad \forall \overline{y} \in \Omega \tag{5.4}$$

In other words the joint probability $p(y = \overline{y})$ gives the probability that the y variables take the value $\overline{y} \in \Omega$. Every $\overline{y}$ can be mapped back on the x variables with T^{-1}, i.e., $\overline{x} = T^{-1}(\overline{y})$ and thus the equivalent joint probability on the x variables is specified completely.

Note that definition (5.4) completely specifies the distribution over the x variables since T has been defined to be one-to-one. This also implies that every distribution over the y variables is mapped on one and only one distribution on the x variables. This gives another interpretation of the maps $T \in \mathcal{T}$. We have seen that is defined to associate in a on-to-one way elements in Ω. Now we have extended this point. Since joint probabilities are functions whose domain is Ω we can apply maps in T to their domain obtaining new joint probability functions and thus defining an equivalence relation between probability distributions. This point becomes clearer in the following.

If $\mathcal{M}_T$ is specified in the η parametrization the following holds. Consider the $\eta_{x,\alpha}$ component of the η parameters vector

$$\eta_{x,\alpha} = \mathbb{E}[x_\alpha] = \mathbb{E}[\prod_{i \in 1 \ldots n} t_i^{-1}(y)] \tag{5.5}$$

The generic t_i^{-1} component of the inverse map T^{-1} is a pseudo-boolean function of the y variables. So once the products are expanded and the squared terms replaced with 1 the argument of the expected value is a pseudo-boolean function of the y variables. Since $\mathbb{E}[y_\alpha] = \eta_{y,\alpha}$ the expected value translates into a linear function of the $\eta_{y,\alpha}$ parameters. We deal with an example later to clarify this point.

A very important observation, related to the equivalence of (P) and (R), take place here. Note that a distribution with reduced support on one variable assignment, $!\exists\, \overline{y}$ such that $p(\overline{y}) = 1$, retain this property when the map T^{-1} is applied. In other words, the number of vertices of the probability simplex which are included in a model $\mathcal{M}_T$ is equal to the one of the corresponding $\mathcal{M}$ over the untransformed variables, even though in general the vertices are different in the two models.

In the following we examine in depth a two variable example that is simple enough to expose clearly and analytically a number of concept and observations.

5.2.1 A two Variable Example

Consider the independence model $\mathcal{M}_1$ defined over two binary variables x_{10} and x_{01}. Its η parametrization is

$$\eta_x = \Big(1,\ \eta_{x,01},\ \eta_{x,10},\ \eta_{x,10}\eta_{x,01}\Big)$$

This model has two free parameters and it covers a two dimensional surface in the three dimensional expectation polytope. The same surface is represented in the probability simplex in Figure 5.1.

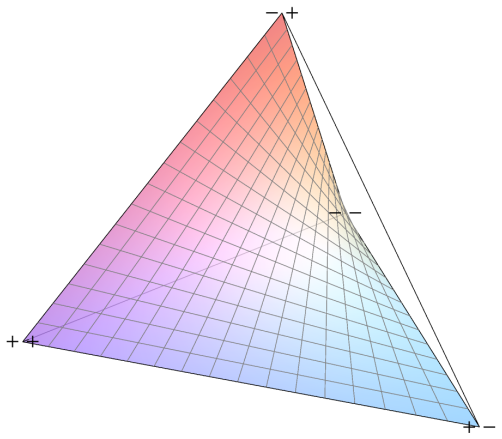

Figure 5.1: The two variables independence model in the probability simplex

Let now $T_1 \in \mathcal{T}$ be the variable map

$$y = (y_{10}, y_{01}) = T_1(x_{10}, x_{01}) = \Big(x_{10}x_{01}, x_{01}\Big)$$

Because of the definition of $\mathcal{T}$ y_{10} and y_{01} are again two binary variables whose domain is $\{-1, 1\}$ and thus an independence model $\mathcal{M}_2$ can be defined over them. Its parametrization in the η_y space is again

$$\eta_y = \Big(1,\ \eta_{y,01},\ \eta_{y,10},\ \eta_{y,10}\eta_{y,01}\Big)$$

We want to characterize the probability distributions obtained picking a distribution s from $\mathcal{M}_2$ an then applying T^{-1} over its support, i.e., given the parametrization η_y of s we want to find its corresponding η_x vector in the untransformed variable space. First note that $T_1^{-1} = T_1$ since $x_\alpha^{-1} = x_\alpha$ for binary variables with domain $\{-1,\ 1\}$, i.e.,

$$x = T_1^{-1}(y) = \Big(y_{10}y_{01},\ y_{01}\Big)$$

We employ (5.5) to obtain $\eta_{x,\alpha}$ as a function of the two free parameters of $\mathcal{M}_2$: $\eta_{y,01}$ and $\eta_{y,10}$:

$$\begin{aligned}
\eta_{x,01} &= \mathbb{E}[x_{01}] = \mathbb{E}[t_{01}^{-1}(y_{10},\ y_{01})] = \mathbb{E}[y_{01}] = \eta_{y,01} \\
\eta_{x,10} &= \mathbb{E}[x_{10}] = \mathbb{E}[t_{10}^{-1}(y_{10},\ y_{01})] = \mathbb{E}[y_{10}y_{01}] = \eta_{y,11} = \eta_{y,10}\eta_{y,01} \\
\eta_{x,11} &= \mathbb{E}[x_{01}x_{10}] = \mathbb{E}[t_{01}^{-1}(y_{10},\ y_{01})t_{10}^{-1}(y_{10},\ y_{01})] = \mathbb{E}[y_{01}^2 y_{10}] = \eta_{y,10}
\end{aligned}$$

Thus we have that

$$\eta_x = \Big(1,\ \eta_{y,01},\ \eta_{y,10}\eta_{y,01},\ \eta_{y,10}\Big)$$

One immediately sees that this vector does not represent an independent distribution over the two variables x_{10}, x_{01} since $\eta_{x,11} \neq \eta_{x,10}\eta_{x,01}$. So we have a model over the x variables with two free parameters for which it holds that $\eta_{10} = \eta_{01}\eta_{11}$. We have seen how every distribution in this model has a $\mathcal{T}$-equivalent distribution in the independence model defined over the variables $y = T(x)$.

This model again covers a two dimensional surface in the expectation polytope, different by the previous one, as it can be seen in Figure 5.2(a).

Another model $\mathcal{M}_3$ can be defined in a similar way, considering the map

$$z = T_2(x) = \Big(x_{10},\ x_{10}x_{01}\Big)$$

Its η parametrization with respect to the untransformed variables x_{10} and x_{01} can be determined directly T^{-1} in the expansions of $\eta_{x,\alpha}$. The result is

$$\eta_x = \Big(1,\ \eta_{z,10}\eta_{z,01},\ \eta_{z,10},\ \eta_{z,01}\Big)$$

The surface it covers in the probability simplex is shown in Figure 5.2(b).

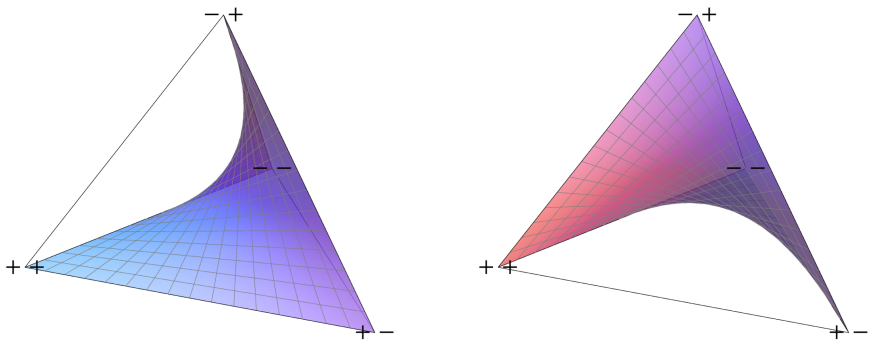

Figure 5.2: Models $\mathcal{M}_2$ (left) and $\mathcal{M}_3$ (right) represented as surfaces in the probability simplex.

5.2.2 The Borders of the Independence Models

Comparing Figures 5.1, 5.2(a) and 5.2(b) it is possible to see that the set of edges of the probability simplex tetrahedron included in the models are different while the vertices belongs to all the models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$. This happens because the maps $T \in \mathcal{T}$ are defined to be one to one, thus every vertex is mapped onto another and every vertex of the simplex belongs to the independence model. Consider for example the family of independent distributions over the y_{10} and y_{01} variables whose ρ parametrization is shown in the table below, where $\beta \in (0, 1)$. These distributions belong to $\mathcal{M}_2$.

	y_{10}	
y_{01}	-1	1
-1	β	0
1	$1-\beta$	0

It is clear that these distributions have reduced support and that they compose the edge connecting the vertices $(-1, -1)$ and $(-1, 1)$ of the probability simplex in the y variables. Consider now the distributions obtained by applying $T_1^{-1}(y)$ to the support of these distributions. This family is defined over the x_{10} and x_{01} variables and its ρ parametrization is shown in the table below.

	x_{10}	
x_{01}	-1	1
-1	0	β
1	$1-\beta$	0

These distributions again have reduced support and lay on the edge between $(-1, 1)$ and $(1, -1)$ in the probability simplex in the x variables. It is possible to see that these distributions do not belong to the independence model over x_{10} and x_{01}. One way to see this is to compute the marginal probabilities $p(x_{10} = 1) = \beta$ and $p(x_{01} = 1) = 1 - \beta$ and notice that $p(1, 1) = 0 \neq \beta(1 - \beta)$ for all values of $\beta \in (0, 1)$. In fact this edge does not belong to the independence model $\mathcal{M}_1$.

Suppose now that the function f to be optimized has two global optima in $x = (-1, 1)$ and $x = (1, -1)$. This means that all the reduced support distributions s over the x variables such as $p(x = (1, -1) = \beta$ and $p(x = (-1, 1)) = 1 - \beta$ are global optima for $\mathbb{E}[f]$ and thus solutions for (R). We have seen that the only distributions of this family that belong to the independence model over the x variables are the ones obtained for $\beta = 0$ and $\beta = 1$. This comes from the fact that such distributions lay on an edge of the probability simplex that is not included in the independence model. Note that this edge is included in the independence model defined over the y variables.

These observations allow us to get more insight in the examples in Section 5.1.1. We had seen that the function $g(y) = f(x) \circ T$ can present dynamics on the border of the independence model different from the f ones. Now this becomes clearer since the border of the independence model defined over the y variables is, in general, different from the one associated to the x variables.

5.2.3 The Mixed Parametrization

The Amari's mixed parametrization, introduced in Section 3.5.3, it is useful to get further understanding of the structure of the models introduced in the previous section. In this section we derive the mixed parametrization $(1,\ \eta_{x,01},\ \eta_{x,10},\ \theta_{x,11})$ as a function of the free η parameters of the independence models defined over the transformed variables y and z.

Consider the model $\mathcal{M}_2$. Its η_x parameters vector is

$$\eta_x = \Big(1,\ \eta_{y,01},\ \eta_{y,10}\eta_{y,01},\ \eta_{y,10}\Big)$$

So we already have three component of the mixed parameters vector. One way to compute the missing θ_{11} parameter is to apply the relation (3.5) and obtain the θ vector as a function of the η parameters. We have that the mixed parametrization of $\mathcal{M}_2$ is

$$\Big(1,\ \eta_{y,01},\ \eta_{y,10}\eta_{y,01},\ \frac{1}{4}\log\Big[\frac{(1+\eta_{y,10})^2}{(-1+\eta_{y,10})^2}\Big]\Big)$$

In Figure 5.3(a) the surface representing the model has been plotted in the three dimensional space of the mixed parametrization. The plane with mixed parametrization $(1,\ \eta_{x,01},\ \eta_{x,10},\ 0)$ is the independence model $\mathcal{M}_1$. Remember that $\theta_{x,11}$ ranges in $(-\infty,\ \infty)$.

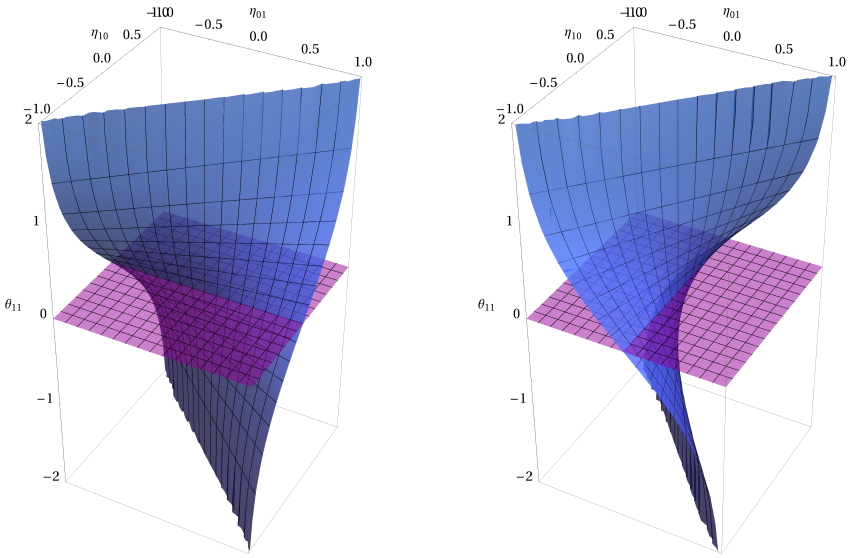

Figure 5.3: $\mathcal{M}_2$ and $\mathcal{M}_3$ in blue, the independence model over the x variables in purple

Note that when $\theta_{x,11} = +\infty$ the two model cover the diagonals of the upper and lower faces of the infinite parallelepiped which forms the parameters space. These diagonals corresponds to the edge of the probability simplex connecting the vertices $(1, 1)$, $(-1, -1)$, for the upper face, and $(-1, 1)$, $(1, -1)$ for the lower one. This can be seen intuitively. When $\theta_{x,11} = +\infty$ there is complete correlations between the variables values, so for example knowing x_{01} one can completely determine $x_{10} = x_{01}$. This means that a variable assignment $(1, -1)$ cannot come from distribution on the upper diagonal.

Again, remembering that $\theta_{x,11}$ is orthogonal to $\eta_{x,01}$, $\eta_{x,10}$ and that in the mixed parametrization it encodes pure correlation between the random variables x_{01} and x_{10}, it is possible to see that most of distributions belonging to $\mathcal{M}_2$ are not independent and the degree of correlation depends only on $\eta_{y,10}$. This seems straightforward since $y_{10} = t_{10}(x) = x_{01}x_{10}$.

Solving the equation $\theta_{x,11} = 0$ shows that the distributions of the independence model $\mathcal{M}_1$ such as their parametrization is $\eta_x = (1,\ \eta_{x,01},\ 0,\ 0)$ belong also to $\mathcal{M}_2$. A notable distribution included in this set is the uniform

distribution $\eta_x = (1,\ 0,\ 0,\ 0)$.

Similar results hold for the model $\mathcal{M}_3$. Its mixed parametrization is

$$\Big(1,\ \eta_{z,01}\eta_{z,10},\ \eta_{z,10},\ \frac{1}{4}\log\Big[\frac{(1+\eta_{z,01})^2}{(-1+\eta_{z,01})^2}\Big]\Big)$$

Again it is possible to see that the distributions $\eta_x = (1,\ 0,\ \eta_{x,10},\ 0)$ are included in both $\mathcal{M}_1$ and $\mathcal{M}_3$. Note that the uniform distribution is the only point, apart from the four vertices of the probability simplex, that belong to all the three models.

In this section we have presented the idea of defining a probability model $\mathcal{M}_y$ over binary variables and then to apply the inverse map T^{-1} over the support of the distributions belonging to $\mathcal{M}_y$, thus obtaining a new probability model $\mathcal{M}_x$. We have characterized some example models and shown how the reduced support distributions included in $\mathcal{M}_x$ and $\mathcal{M}_y$ can be different. In the next section we give some more results about the class $\mathcal{T}$.

5.3 Existance Theorem

In this function we show how, in principle, every optimization problem (R) for f can be reconduced to an equivalent one on $g = f \circ T \in \mathcal{T}$ for which the Exact Gradient Descent strategy always converge to the global optimum for $\mathbb{E}[f]$. The argument proceed as follows: first we introduce a class of pseudo-boolean functions for which $\nabla\mathbb{E}[f]$, calculated over the independence model, is always positive. This means that for this functions there exists only one attractor for the Exact Gradient Descent search strategy. Then we show that every pseudo boolean function f is $\mathcal{T}$-equivalent to the proper element of the previously defined class.

Note that the proof is not constructive so in this section we do not provide a procedure to obtain g whose complexity is not higher that the one of solving (R). A technique that goes in this direction is presented in later chapters.

5.3.1 A Particular Pseudo-boolean Function

Let us consider a pseudo-boolean function f defined over n variables such as it holds that

$$\forall a, b \in \Omega\ \ a >_L b \Rightarrow f(a) >= f(b) \tag{5.6}$$

where $>_L$ is the lexicographic ordering with $-1 \prec 1$. In other words, the values of f are increasing once the domain has been ordered lexicographically. An example of a function for which this holds is

x_{10}	x_{01}	$f(x_{10}, x_{01})$
-1	-1	1
-1	1	2
1	-1	3
1	1	4

Each component of the gradient of $\mathbb{E}[f]$ calculated over the independence model is always grater then zero. This is intuitive since we can always achieve an improvement of the value of $\mathbb{E}[f]$ increasing the marginal probabilities $p(x_\alpha = 1)$. We prove this more formally.

Consider the first variable x_1. It holds that $f(1, \cdot) \geq f(-1, \cdot)$. The expected value of f over the independence model reads as

$$\mathbb{E}[f(x)] = \sum_{i \in \{-1,1\}} \left[\sum_{J \in \{-1,1\}^{n-1}} p(i, J) f(i, J) \right]$$

We have that $p(i, J) = p(i)p(J)$. The marginal probability $p(J)$ is a probability distribution itself, the former reads as

$$\mathbb{E}[f(x)] = \sum_{i \in \{-1,1\}} p(i) \left[\sum_{J \in \{-1,1\}^{n-1}} p(J) f(i, J) \right] =$$

$$= p(x_1 = 1)\mathbb{E}[f(1, \cdot)] + p(x_1 = -1)\mathbb{E}[f(-1, \cdot)]$$

Remembering that the marginal $p(x_i) = \frac{1}{2}(1+\eta_i x_i)$ we can write the previous in the η parametrization. Let $a = \mathbb{E}[f(1, \cdot)] - \mathbb{E}[f(-1, \cdot)]$:

$$\mathbb{E}[f(x)] = \frac{a}{2}\eta_1 + \frac{\mathbb{E}[f(1, \cdot)] + \mathbb{E}[f(-1, \cdot)]}{2}$$

Note that a does not depend on η_1. It can be seen that if (5.6) holds then $\mathbb{E}[f(1, J)] \geq \mathbb{E}[f(-1, J)]$. In fact we have that $\mathbb{E}[f(1, J)]$ assumes his lowest value when the distribution $p(1, J)$ has reduced support Ω' included in the set of the minima of $f(1, J)$. Similar considerations hold for the maximum of $\mathbb{E}[f(-1, J)]$ and, because of the fitness ordering condition, it holds that $\min f(1, J) \geq \max f(-1, J)$ with equality if and only if $f(x) = c \, \forall x \in \Omega$. So we have that

$$\frac{\partial \mathbb{E}[f]}{\partial \eta_1} = \frac{a}{2} \geq 0$$

with equality if and only if the function is constant. Note that this already implies that for every non-constant function of the class we are examining $\mathbb{E}[f]$ over the independence model has no saddle points inside the η domain.

For the k-th variables the expansion shown before generalize as:

$$\mathbb{E}[f] = \sum_{H \in \{-1,1\}^{k-1}} p(H)\Big[\mathbb{E}[f(H,1,\cdot)]p(x_k = 1) + \mathbb{E}[f(H,-1,\cdot)]p(x_k = -1)\Big]$$

Similarly, let $a_H = \mathbb{E}[f(H,1,\cdot)] - \mathbb{E}[f(H,-1,\cdot)]$. We have that

$$\mathbb{E}[f] = \sum_{H \in \{-1,1\}^{k-1}} p(H)\Big[\frac{a_H}{2}\eta_k + \frac{\mathbb{E}[f(H,1,\cdot)] + \mathbb{E}[f(H,-1,\cdot)]}{2}\Big]$$

Again, $p(H)$ does not depend on η_k because of the independence assumption and it is possible to see that $\mathbb{E}[f(H,1,\cdot)] \geq \mathbb{E}[f(H,-1,\cdot)]$ for every H. This implies that the partial derivative of $\mathbb{E}[f]$ with respect to η_k is a sum of non-negative terms.

Note that for greedy search strategies confined on the border of the independence model there exist no local minima. In fact it is easy to see that for every value assignment that is not a global optimum there always exists another one with better fitness at hamming distance 1.

5.3.2 The Correct Map in $\mathcal{T}$

We have seen in the previous section that if the images of f monotonically increase once the domain Ω has been ordered lexicographically, then every component of $\nabla\mathbb{E}[f]$ calculated over the independence model is always greater than zero. This implies that the global optimum for $\mathbb{E}[f]$ is the only attractor for the Exact Gradient Descent strategy on the independence model. It is clear that there always exist a map $T \in \mathcal{T}$ which reorders the fitness values over the domain in a way that the property (5.6) holds for $f \circ T = g$. Thus the following theorem holds. Let f be a pseudo-boolean function over n variables:

Theorem 6. *There always exists a function $g = f \circ T$ with $T \in \mathcal{T}$ such as an Exact Gradient Descent search strategy always converges to the global optimum for $\mathbb{E}[g]$.*

Remember that

$$\min_{x \in \Omega} f(x) = T^{-1}\big[\min_{y \in \Omega} g(y)\big]$$

This result can also be interpreted with the ideas presented in the previous section. The following equivalent theorem holds. Consider a pseudo-boolean function f defined over the x variable vector, a variable map $T \in \mathcal{T}$ and an independence model defined over $y = T(x)$, $\mathcal{M}_T$.

Theorem 7. *There always exists a map T such as an Exact Gradient Descent search strategy on the model $\mathcal{M}_T$ always converges to the global optimum for $\mathbb{E}[f]$.*

Let us discuss a three variable example. Consider the pseudo-boolean function f with coefficients vector

$$c = \Big(0,\ 4,\ 2,\ 0,\ 1,\ 0,\ 0,\ -16\Big)$$

Remembering the lexicographic ordering on for the coefficient vectors, it is possible to see that this function has no interactions of order two between variables. However, there is a strong interaction of order three. We have analysed a similar function in Section 4.2.2 and we have already seen that it has two saddle points inside the parameters domain. Lets write explicitly all the eight fitness values of the function.

x_{100}	x_{010}	x_{001}	$f(x)$
-1	-1	-1	9
-1	-1	1	-15
-1	1	-1	-19
-1	1	1	21
1	-1	-1	-21
1	-1	1	19
1	1	-1	15
1	1	1	-9

It is possible to see looking at the disposition of fitness values over the domain that there are various local optima on the border of the independence model. For example $f(1,-1,1)$ is not a global optimum and all of its neighbours at hamming distance 1 have lower fitness: $f(-1,-1,1) = -15$, $f(1,1,1) = -9$ and $f(1,-1,-1) = -21$. We have already seen that since there are two saddle points in the expected value of f calculated on the independence model the EGD strategy in general converges to a local optimum for $\mathbb{E}[f]$. Consider now the following map T:

$$T(x_{100}, x_{010}, x_{001}) = \Big(-x_{100}x_{010}x_{001},\ x_{001},\ x_{010}\Big)$$

If f is composed with T we have the following function g

x_{100}	x_{010}	x_{001}	$g(x) = f \circ T$
-1	-1	-1	-21
-1	-1	1	-19
-1	1	-1	-15
-1	1	1	-9
1	-1	-1	9
1	-1	1	15
1	1	-1	19
1	1	1	21

The property (5.6) holds so the function g has no saddle points inside the parameters domain. This implies that the EGD strategy converges to the global optimum for $\mathbb{E}[f \circ T]$ from every starting distribution. This example is an application of Theorem 6. However, up to now no simple way to obtain the map $T \in \mathcal{T}$ which achieves these results has been given.

5.4 A Subclass of $\mathcal{T}$

In this section we introduce the class $\mathcal{L}^1 \subset \mathcal{T}$ and its extension $\mathcal{L}^k$ by means of the composition operator. This allows us to easily generate variable maps to be composed with f such that they are one-to-one and whose inverse is straightforward to compute.

5.4.1 Motivations

Since the effect of the $\mathcal{T}$ maps is essentially a permutation of the fitness values over the domain Ω, the cardinality of the class $\mathcal{T}$ is strictly related to the number of possible permutations of 2^n elements, where n is the number of variables, that is $(2^n)!$, higher than the cardinality of the search space Ω. Thus, Theorems 6 and 7 seem to be difficult to exploit in practice.

The observations and results of the previous sections suggest that the interesting maps are only those which are able to change the EGD performances. The class $\mathcal{T}$ includes these maps and many others that are not useful to this end.

The Role of the Negation

Consider the class of maps such as they only negate the values of some variables, i.e., $a = (1, -1)$, $T(a) = (1, 1)$. An example of map of this class is

$$T_a(x) = \Big(t_{100}(x),\, t_{010}(x),\, t_{001}(x)\Big) = \Big(-x_{001},\, x_{010},\, x_{100}\Big)$$

It is possible to see that these kind of maps do not change the behaviour of greedy search strategy on the border of the independence model. In fact if the variable assignments $x = a$ and $x = b$ are neighbours, i.e., their hamming distance is 1, $T(a)$ and $T(b)$ are still neighbours. This happens because the same sets of variables have been negated in each assignment.

The Role of the Variable Swap

Consider now the class of the variables maps such as they only perform a swap of the values of a variable couple, i.e., $a = (-1, 1)$, $T(a) = (1, -1)$. An example of a map of this class is

$$T_b(x) = \Big(x_{100},\ x_{001},\ x_{010}\Big)$$

These kind of transformations do not change the dynamics on the border of the independence model. Again this comes from the fact that if the variable assignments a and b are neighbours the same holds for $T(a)$ and $T(b)$.

Consider now the final example in Section 5.3.2. We considered a three variable function f whose expected value over the independence model had two saddle points. We found that the function $g = f \circ T_c$ has no saddle points in its expected value over the same model. Remember that

$$T_c(x) = \Big(-x_{100}x_{010}x_{001},\ x_{001},\ x_{010}\Big)$$

It is possible to see that

$$T_c = T_a \circ T_b \circ \overbrace{\Big(x_{100}x_{010}x_{001},\ x_{010},\ x_{001}\Big)}^{T_d}$$

If the properties of maps such as T_a and T_b where true it should be possible to discard T_a and T_b and employ only T_d, in which the negations and the variables swaps have been eliminated. It is possible to see that the coefficient vector d of the function $g = f \circ T_d$ is

$$d = \Big(0,\ 4,\ 2,\ -16,\ 0,\ 0,\ 0,\ 1\Big)$$

and that the two saddle points for $\mathbb{E}[g]$ calculated over the independence model lay outside the parameters domain.

On the Coefficients of t_α

Another problem with the class $\mathcal{T}$ raises when these maps have to be represented and applied. We have seen two ways of doing this. The first method

consist in giving all the 2^n associations $b = T(a)$, which can be immediately discarded because of excessive complexity. Alternatively, one can specify all the not null coefficients of the pseudo-boolean functions t_h composing T. The problem with this approach is that it seems difficult to specify and check constraints on the $c_{h,\alpha}$ coefficients to guarantee that each t_h co-domain is $\{-1, 1\}$ and that $T(x)$ is one-to-one.

5.4.2 The Single Product Maps $\mathcal{L}^1$

Here we define the class of the single product variables maps $\mathcal{L}^1$. This class is composed by the maps $T(x)$ for which at most one variable is replaced with the product of other two while the others retain their value. An example is the three variables map

$$T(x) = \Big(x_{100}x_{010},\, x_{010},\, x_{001}\Big)$$

More precisely, the following condition has to hold on the functions t composing a map $T \in \mathcal{L}^1$

$$!\exists\, \alpha \Big(t_\alpha = x_\alpha x_\beta \wedge \forall\gamma(\gamma \neq \alpha \Rightarrow t_\gamma = x_\gamma)\Big)$$

where the multi-indices α, β, γ considered by the quantifiers are limited to the ones with cardinality 1. Note that the operations such as minus and variable swap discussed in the previous section, considered useless to our ends, have been explicitly forbidden. In the following we call $L(x_\alpha,\, x_\beta)$ the single product map in which x_α is replaced with $x_\alpha x_\beta$.

The only t_h different from the identity map is $t_\alpha = x_\alpha x_\beta$. Its co-domain is $\{-1, 1\}$ and $|N_{t_h}(1)| = |N_{t_h}(-11)| = 2$. This implies that every $T \in \mathcal{L}^1$ is one-to-one. Moreover, since $x_\alpha^{-1} = x_\alpha$ we have that $x_\alpha = t_\alpha^{-1}(y) = y_\alpha y_\beta$, thus $T^{-1} = T$.

It is possible to see that the vector of coefficients d of the function $g = f \circ T \in \mathcal{L}^1$ it is always a permutation of the c one. This happens because every element of the n variables pseudo-boolean function basis is mapped into another one, eventually the same.

Note that the class $\mathcal{L}^1$ is composed by $n(n-1) \approx n^2$ non equivalent maps.

5.4.3 The Class $\mathcal{L}^k$

The class $\mathcal{L}^1$ includes only a small subset of the maps in $\mathcal{T}$. We can expand this class including all the maps obtained composing in every possible way

the elements of $\mathcal{L}$. More precisely, we have that

$$\mathcal{L}^k = \mathcal{L}^{k-1} \cup \Big(\bigcup_{l \in \mathcal{L},\, m \in \mathcal{L}^{k-1}} m \circ l \Big)$$

So $\mathcal{L}^k$ is composed by every possible composition of n maps belonging to $\mathcal{L}$, for example $\mathcal{L}^3 \ni L = L_1 \circ L_2 \circ L_3$ with $L_1, L_2, L_3 \in \mathcal{L}^1$. Since every element of the composition sequence is one-to-one the same holds for the map result of the composition. Remembering the proprieties of the inverse maps of elements in $\mathcal{L}^1$ it is easy to see that the inverse map $L^{-}1$ is obtained simply reversing the order of the composition sequence, so $L^{-1} = L_3 \circ L_2 \circ L_1$.

It holds also for $\mathcal{L}^k$ that the coefficients vector d of the function $g = f \circ T \in \mathcal{L}^k$ is a permutation of c. Note that there exist permutations of coefficients that are not achievable with a map in $\mathcal{L}^k$. It is easy to find a counter example. Consider the three variable function f whose coefficients are $c = (0, 3, 2, 1, 0, 0, 4, 0)$ and consider the function g such as its coefficients vector is c with c_{001} and c_{110} swapped. It is possible to see that the image set of g is not a permutation of the f one, so there not exists a map $L \in \mathcal{L}^k$ that achieves this inversion. Note that does not exist such a map either in $\mathcal{T}$. It is still an open issue whatever exist or not coefficient swaps which do not alter the image set of f and are not achievable with maps in $\mathcal{L}^k$.

Another observation is that there are few non trivial reordering of the image values of f that cannot be produced composing f with a proper map belonging to $\mathcal{L}^k$. Here is a counter example. Consider a three variable function f. We want to find the map T that achieves $f(T(a)) = f(a)$ if $a \notin \{(1, 1, -1),\ (1, 1, 1)\}$ and $f(T(1, 1, -1)) = f(1, 1, 1)$, i.e., we ask for a map T that swaps in $g = f \circ T$ the fitness of the two values assignments $(1, 1, 1)$ and $(1, 1, -1)$. The result is

$$T(x) = \Big(x_{001},\ x_{010},\ \frac{1}{2}(x_{001} - x_{100}x_{001} - x_{001}x_{010} - x_{100}x_{010}x_{001}) \Big)$$

This maps comprises sums of x monomials that are not achievable composing transformations of the class $\mathcal{L}^1$.

It is not easy to give results about the cardinality of the class $\mathcal{L}^k$, essentially because there exist sequences of compositions that represent the identity map T_{id}, i.e., the map for which $T(x) = x\ \forall x \in \Omega$. For example, it is possible to see with some verbose calculations that for $n = 3$

$$\Big(L(x_{100}, x_{010}) \circ L(x_{010}, x_{001}) \Big)^4 = T_{id}$$

Here we employed the intuitive notation $(a \circ a) = a^2$. This suggest that the class $\mathcal{L}^k$ could have regular algebraic properties with respect to the

composition operator. We conjecture that the cardinality of the class $\mathcal{L}^k$ is substantially smaller than $|\mathcal{L}^1|^k \approx n^{2k}$. Moreover, since the class $\mathcal{T}$ is finite, the same holds for $\mathcal{L}^\infty \subset \mathcal{T}$.

5.5 Conclusions

In this chapter we have introduced the class $\mathcal{T}$ of the one-to-one maps between elements in the domain Ω. We have seen that these maps allow the introduction of two notion of equivalence. The first regards pseudo-boolean functions: two functions f and g are said to be $\mathcal{T}$-equivalent if there exists a map $T \in \mathcal{T}$ such as $g = f \circ T$, or equivalently, $f = g \circ T^{-1}$, where T^{-1} is the inverse map of T, whose existence is guaranteed by the fact that T is one-to-one. We have seen that it makes no difference to the ends of finding a solution for the problem (P) if we consider the function f or any other $g \equiv_T f$. We have also proven that if the map T is properly chosen the Exact Gradient Descent search strategy over the independence model always converges to the optimum for $\mathbb{E}[f \circ T]$.

The second notion on equivalence regards probability models. We have seen how a map $T \in \mathcal{T}$ can be applied to the support of probability distributions and how this procedure can be applied to models in order to produce new ones, $\mathcal{T}$-equivalent to the original. We have developed an example in two variables and we have shown how a bundle of n free parameters probability models can be associated to the independence model. We have reinterpreted the previous result with this notion of equivalence and we have shown that for every function f there is a model $\mathcal{T}$-equivalent to the independence model for which the Exact Gradient Descent search strategy always converge to the global optimum for $\mathbb{E}[f]$.

However, we still lack for a strategy to find the map in $\mathcal{T}$ for which these convergence results hold. We have defined the class $\mathcal{L}^k$ and we have shown that this is a proper subset of $\mathcal{T}$. The class $\mathcal{L}^k$ is defined in a way such that its elements are easily enumerated and the inverse maps are easily computed. This properties are critical if a search in the space of the maps has to be performed. It is still an open issue if the same results for $\mathcal{T}$ hold also for $\mathcal{L}^k$.

Chapter 6

FCA: A new Algorithm

We have seen in the previous chapter how variable maps of the class $\mathcal{L}^k$ can be defined and the effects of the compositions of these maps with the fitness function to be optimized. In this chapter we propose a strategy to choose the proper map in $\mathcal{L}^k$ to compose with f, based on results from information geometry. We show how two mapping and un-mapping steps can be introduced in the Population Based Incremental Learning strategy and how these affect its performances. This chapter is organized as follows. First we review the basic EDAs principles and assumptions, then the novel FCA search strategy is presented and described in detail. Then the behaviour of FCA on the simple yet inspiring two dimensional case is studied analytically. In the last part of the chapter the observed behaviour of FCA with some two variable fitness functions is analysed to confirm the theoretical analysis.

6.1 The Kullback-Leibler Divergence and the EDAs Idea

In Chapter 2 we presented a brief review of the Estimation of Distribution Algorithms as an evolution of the Genetic Algorithms. The main difference between these two class of Evolutionary Algorithms lies in the reproduction operators. While the GAs employ operators such as crossover and mutations which produce new individuals working directly on the genetic material of the ones in the selected population, EDAs do this estimating the probability distribution which, according to an heuristic, better "fits" the selected pop-

ulation and then produce a new generation of candidate solutions sampling from this distribution.

The background assumption which has to hold for EDAs to work as expected is the following: the probability distribution which is fitted on the sample of the selected individuals is able to capture the features of these candidate solutions which are responsible of the high fitness. These features manifest themselves by means of correlations between variable values or, from an Information Theory point of view, by the presence of mutual information between variables. Thus it is critical for the performance of EDAs that the mutual information between variables in selected individual is not lost during the estimation and sampling phases. This depends heavily on the family of probability distribution chosen for the estimation process, i.e., the probability model employed. Let us clarify this point with an example.

Suppose that after sampling and truncation selection the population is composed by two individuals with equal fitness:

x						$f(x)$
$+1$	-1	$+1$	-1	$+1$	-1	10
-1	$+1$	-1	$+1$	-1	$+1$	10

The selection phase has favoured these two individuals which manifest full correlation between pairs of adjacent variables, i.e., $x_1 = \pm 1$ and $x_h = -x_{h-1}$. The desired behaviour of the EDAs estimation and sampling phases is the reproduction of this feature in the next generation. Suppose now that the model employed by the EDA of this example is the independence model. It is easy to see that the independent distribution for which this sample has maximum Likelihood is the uniform distribution and that a sample from this does not produce individuals preserving the correlations which leaded to high fitness values. This happens because the mutual information between two independent variables is always zero, thus the information on the relations among different variable values available in the selected population is all lost during the estimation step. For a discussion of the effects of reproduction operators on EDAs and GAs exploration abilities see [36]. This example shows how the choice of the probability model employed by the EDAs is crucial for them to work as expected and heavily depends on the structure of the interaction among variables in the fitness function f.

In Chapter 3 we have introduced the Kullback-Leibler Divergence as the metric defined over the manifold of the probability distributions $\mathcal{P}$. For two distributions over the discrete domain Ω it reads as

$$D[p:q] = \sum_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \Big[\log \frac{p(x)}{q(x)} \Big]$$

An interpretation in terms of information theory is the following. If p is a *true* distribution and q a model distribution, $D[p : q]$ is a measure of the loss of information when p is approximated by q. For instance, in case a bivariate distribution $p(x_{01}, x_{10})$ is approximated with the independent distribution having the same marginals $p(x_{01})$, $p(x_{10})$ we lose the correlations between variable values, i.e., the mutual information between x_{01} and x_{10} $I(x_{01}, x_{10}) = D[p(x_{01}, x_{10}) : p(x_{01})p(x_{10})]$.

This reasoning leads to conclude that a good strategy to perform the EDA estimation step is to chose the distribution $q \in \mathcal{M}$ such as

$$q = \arg\min_{s \in \mathcal{M}} D[s : p] \tag{6.1}$$

where p is the distribution representing the population after selection and $\mathcal{M}$ is the model employed. Note that this corresponds to choose the distribution $s \in \mathcal{M}$ for which the likelihood of the population after selection is maximum. This is well known in the EDAs literature, for example see [37].

6.2 FCA, a Novel Search Strategy

We have seen in the previous chapter how a n free parameter model can be associated with every variable map T of the class $\mathcal{T}$. We have shown that these models are $\mathcal{T}$-equivalent to the independence model defined over the transformed variables $y = T(x)$. The main result of the previous chapter is the fact that for every function f there exists a map T such that the expected value of f calculated over the model associated with T has no saddle points and its gradient is always non negative. We have seen that in general, if this condition holds for $\mathbb{E}_{\mathcal{M}}[f]$, the Exact Gradient Descent strategy always converges to the global optimum for $\mathbb{E}[f]$. What we lack is a strategy to chose the correct map T to apply among the others in $\mathcal{T}$, which we have shown to have a cardinality grater than the original search space Ω.

Informally, the FCA strategy tries to iteratively build the map L in $\mathcal{L}^k$ such that its associated model is the nearest in terms of Kullback-Leibler Divergence to the distribution of the selected individuals. Once the map L has been chosen, the distribution belonging to the independence model defined over the mapped variables $y = L(x)$ is estimated using the selected population as sample. This procedure ensures that the minimum amount of mutual information between variable values in the selected individuals is lost during the model estimation step. We discuss this point in detail in later sections.

6.2.1 An Iteration of FCA

The search strategy proposed in this chapter is a variation of the classical PBIL iteration. We introduce two new steps: one before the estimation step and the other after the new population has been sampled from the estimated distribution. In these steps we first apply a variable map T and then its inverse T^{-1}.

1. Evaluate fitness and perform truncation selection on the population
2. **Apply a map $T \in \mathcal{L}^k$ to the individuals in the selected fraction of the population**
3. Perform a Max Likelihood estimation of the independent distribution s using the mapped selected individuals
4. Sample a new population from s
5. **Apply the inverse map T^{-1} on the new population**

It is possible to see that the only difference between an iteration of the PBIL search strategy is the employ of the maps T and T^{-1} before and after the estimation step. This rather subtle steps makes the difference since even if the distributions estimated and sampled from belong to the independence model, this is done working with variables assignments which are transformed by means of the map T. This implies that the actual model we are working with is different from the independence model and it depends on the particular map applied. In other words FCA chooses at every iteration a map $T \in \mathcal{L}^k$ and makes a step in its stochastic walk in the manifold of the distributions $\mathcal{P}$ along a model T-equivalent to the independence model.

6.2.2 The Choice of $T \in \mathcal{L}^\infty$

In this section we address the problem of how the proper map T can be chosen efficiently. We propose to exploit the KLD minimization (6.1) to chose the variable map $\mathcal{L}^k$ to apply. Consider a map $T \in \mathcal{L}^\infty$ and a population after selection P. We have

- The distribution representing the selected individuals P
- The distribution P_T obtained mapping with T the support of P
- The independence model defined over the transformed variables $y = T(x)$, $\mathcal{M}_T$. Remember that this model is different from the independence model over the untransformed variables x.

- The projection of P_T on $\mathcal{M}_T$, $P_T^{(1)}$.

Here with the term *projection* we mean the Information Geometry concept presented in Section 3.5.4. In the following we use the Amari's notation to indicate projections, e.g. $P^{(1)}$ is the distribution with the same marginals as P but no interactions among variables, i.e., the independent distributions nearest to P in terms of Kullback Leibler Divergence.

We propose to choose the map $T \in \mathcal{L}^\infty$ such as

$$T = \underset{L \in \mathcal{L}^\infty \bigcup T_{id}}{\arg\min}\ D[P_L : P_L^{(1)}] \tag{6.2}$$

In other words, our goal is to apply the map T such as the selected individuals mapped with T best resemble an n variables independent sample. This implies that the projection of the distribution P_T on the independence model $\mathcal{M}_T$ over the transformed variables cause the minimal mutual information loss obtainable with models associated with the $\mathcal{L}^\infty$ family.

However, it is computationally infeasible to perform the minimization proposed in (6.2). We propose to employ the following greedy iterative strategy to obtain an approximation of (6.2). The initial map T_0 is set to be the identity map T_{id}. At iteration t the following map is considered

$$T_t = T_{t-1} \circ \underset{L \in \mathcal{L}^1}{\arg\min} D[P_{T_{t-1} \circ L} : P^{(1)}_{T_{t-1} \circ L}] \tag{6.3}$$

In other words at every iteration we compose the map found at previous iteration T_{t-1} with the element L in $\mathcal{L}^1$ such as the model associated with the resulting map $T_t = T_{t-1} \circ L$ is closer to the distribution P_T.

This process ends when one of the following conditions holds:

- Every element $L \in \mathcal{L}^1$ has been considered and for each of those $D[P_{T_{t-1} \circ L} : P^{(1)}_{T_{t-1} \circ L}] \geq D[P_{T_{t-1}} : P^{(1)}_{T_{t-1}}]$. In other words, no further KLD reduction is achievable composing T_{t-1} with more elements in $\mathcal{L}^1$.

- When the number of elements belonging to $\mathcal{L}^1$ which have been composed to form T is greater than the k parameter. This implies restricting the search space in (6.2) to $\mathcal{L}^k$.

- (Optional) Every element in L has been considered and for each of those the reduction of Kullback-Leibler Divergence with respect to the previous iteration map T_{t-1} is lower than a certain threshold λ. More precisely, for all $L \in \mathcal{L}^1$ it holds that $D[P_{T_{t-1}} : P^{(1)}_{T_{t-1}}] - D[P_{T_{t-1} \circ L} : P^{(1)}_{T_{t-1} \circ L}] \leq \lambda$.

The third ending condition has been introduced since it has been observed during experiments that the last iterations of the greedy minimization procedure (6.3) often lead to very small improvements in terms of KLD and thus it is not clear if they are meaningful or only dependant on the stochastic noise present in the selected population. It is difficult to characterize a good choice of λ since the Kullback-Leibler Divergence values encountered in the minimization step vary substantially and depend on a number of factors such as the number of variables n, the size of the selected population, the fitness function f and the parametrizations of the distributions considered. Remember for example how two pairs of distributions with equal euclidean distance in the η parameter space could have different KLD in the manifold of probability distributions $\mathcal{P}$. As we show in the next chapter, the choice of the KLD threshold λ is critical for the performances of the FCA search strategy.

The iterative greedy strategy (6.3) in general does not return the same result as (6.2) since in principle there can be distributions s for which there not exists an $L \in \mathcal{L}^1$ such as

$$D[s_L : s_L^{(1)}] < D[s : s^{(1)}]$$

but the map which achieves such KLD reduction lives in $\mathcal{L}^k$ with $k \geq 2$. In other words, two concatenation steps could be needed to achieve KLD reduction. We conjecture that the quality of the approximation obtained with the iterative greedy search depends heavily on the structure of the interaction between variables in P, and thus in f.

6.2.3 How the KLD is Computed

Here we address the problem of how, given $T \in \mathcal{L}^k$, $D[P_T : P_T^{(1)}]$ can be computed. First the map T is applied to every individual in the selected population obtaining their transformed counterpart y. Then the distribution P_T is computed counting the occurrences of each individual $y \in \Omega$

$$P_T(y) = \frac{N[T(x)]}{|P|}$$

The projection $P_T^{(1)}$ on the independence model over the y variables is determined computing the marginal probabilities $p(y_\alpha = 1)$ with the Max-Likelihood estimators

$$\hat{p}(y_\alpha = 1) = \frac{\sum_{y \in P_T} y_\alpha}{|P|}$$

Since P_T is independent, is joint probability factorize in the product of the marginals, thus for computing the KLD we simply have to calculate

$$D[P_T : P_T^{(1)}] = \sum_{j \in \Omega | P_T(j) > 0} P_T(j) \Big[\log[P_T(j)] - \sum_{\alpha} \log[p(y_\alpha = j_\alpha)] \Big] \quad (6.4)$$

where $|\alpha| = 1$.

The cardinality of Ω is exponential in the number of variables but the populations employed in EDAs in general do not exceed n^2 individuals, thus $P_T(y) = 0$ for most of the $y \in \Omega$.

6.2.4 Introducing a Learning Rate γ

Most of the Estimation of Distribution Algorithms do not allow that the probability distribution used to sample the new population changes "too much" between successive iterations. This is done to limit the misleading effects of "unlucky" samples on the convergence of EDAs search strategies.

Let us assume the independence model is employed. At iteration t a new probability distribution $\hat{s}$ is estimated from the selected population. Usually the distribution s_{t-1} from which the current population has been sampled is not discarded but is mixed with s. For instance, in the η parametrization

$$\eta_{t+1} = \gamma \eta_{\hat{s}} + (1 - \gamma)\eta_{t-1}$$

With $\gamma \in (0, 1]$. This corresponds to choose a point along the e-geodesic connecting $\hat{s}$ and s_{t-1} as the resulting probability distribution at iteration t. γ is called *learning rate* This slows down the velocity of parameters changes along the search strategy iteration and helps slowing down the convergence.

In the FCA search strategy it is difficult to characterize this geodesic since the distributions $\hat{s}$ and s_{t-1} in general belong to different models. So the same is achieved at populations level. At iteration t, a fraction $(1-\gamma)$ of individuals belonging to the population before selection at iteration $t-1$, and thus coming from s_{t-1} is added to the sample drawn from the distribution $\hat{s}$.

6.2.5 Computational Complexity

Here we evaluate the computational complexity of the search strategy proposed. This gives a measure of how the execution time depends on the number of variables in the fitness function and how it scales when the problem size increases. For each iteration, given the number of variables n, the class of maps $\mathcal{L}^k$ considered and the population size m we have that for each iteration step the worst case computational complexity is

1. Fitness evaluation: m times the complexity of evaluating $f(x)$
2. Truncation selection: $\mathcal{O}(m \log m)$
3. Choice of $T \in \mathcal{L}^k$ and Max-Likelihood estimation: $\mathcal{O}(kn^2m)$
4. Sampling: $\mathcal{O}(nm)$
5. Apply T^{-1}: $\mathcal{O}(mk)$

It is possible to see that most of the FCA steps have a computational complexity which is linear or constant with respect to the number of variables n. Unfortunately we have that the choice of L by means of the greedy KLD minimization strategy (6.3) has a complexity that goes with the square of the number of variables while the complexity of other univariate EDAs is generally linear in the number of variables and in the population size. This is because the cardinality of $\mathcal{L}^1$ is $n(n-1) \approx n^2$.

Another aspect which has to be considered is that the the parameters k and m can be related to n, e.g. it is possible to empirically determine the minimum population size and $\mathcal{L}^1$ concatenation length for the FCA search strategy to exhibit good performances as a function of the number of variables n. Thus the complexity of the second step is in general higher than n^2.

As a further remark, note that the number of fitness function evaluations is critical since in most applications this task is non trivial and could hide the core algorithm complexity. Note that FCA evaluates the fitness of at most m individual per iteration, like PBIL and most of the other EDAs. This means that in most of the real world applications, in which the fitness evaluation is computationally expensive, FCA could have a running time comparable to the PBIL one.

6.3 The two Variable Case

In this section we analyse the behaviour of the FCA search strategy dealing with a two variables fitness function. We first try to derive analytically the behaviour of the algorithm when the infinite population hypothesis holds, then we compare the results with the observed behaviour of FCA.

6.3.1 Theoretical Analysis

In this section we assume that the expected value of the fitness function calculated on the independence model has a saddle point and we try to characterize the starting distributions for which the KLD minimization strategy

proposed in the previous section leads to the choice of the correct model $\mathcal{M}_T$.

In order to do this, we use a similar technique as in Section 4.3.4: first we make hypothesis on which individuals are contained in the population after selection, then we compute which distributions can lead to these populations after sampling and selection with rate α. Knowing the composition of the selected population allows us to write the expressions for $D[P : P^{(1)}]$, $D[P_T : P_T^{(1)}]$ and thus decide if a variable map is applied by FCA or not for the starting distribution considered.

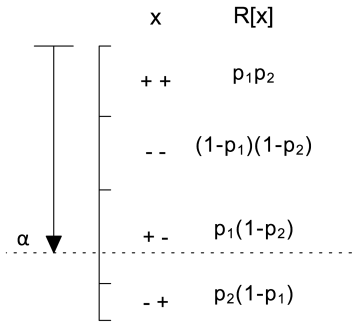

Figure 6.1: Representation of the population after sampling and before selection

Consider a 2 variables pseudo-boolean function whose fitness ordering over the domain is, without loss of generality, $f(1,1) > f(-1,-1) > f(1,-1) > f(-1,1)$. Recall the structure of the population after selection, sketched in Figure 6.1. Let s be a starting independent distribution for the FCA search strategy such as its marginal probabilities $p_{01} = p(x_{01} = 1)$ and $p_{10} = p(x_{10} = 1)$ satisfy

$$p_{01}p_{10} < \alpha < p_{01}p_{10} + (1 - p_{01})(1 - p_{10}) \tag{6.5}$$

where α is the truncation selection rate. The distributions that satisfy this condition are the one in the highlighted region of Figure 6.2. Condition (6.5) specifies that, if the infinite population hypothesis holds, after truncation selection the population is composed only by individuals $(1,1)$ and $(-1,-1)$. The distribution representing the population is shown in the following table.

	x_{10}	
x_{01}	-1	1
-1	$1 - \frac{p_{01}p_{10}}{\alpha}$	0
1	0	$\frac{p_{01}p_{10}}{\alpha}$

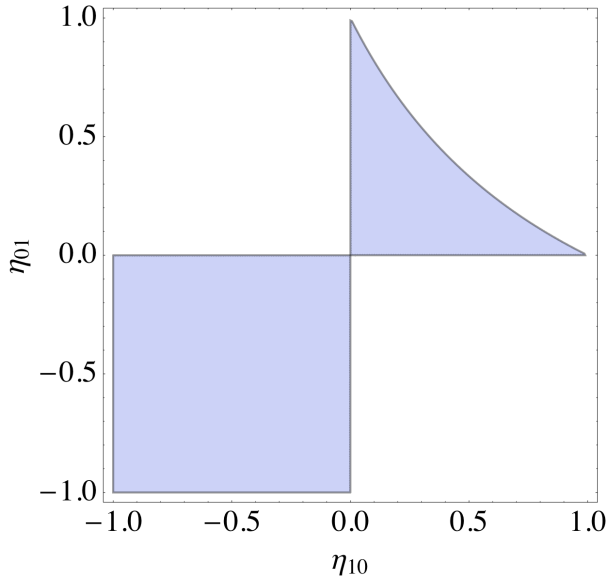

Figure 6.2: Independent distributions which satisfy (6.5) in the η parametrization, with $\alpha = \frac{1}{2}$

Let us now compute the Kullback-Leibler Divergence between this distribution and its projection on the independence model $s^{(1)}$. This is easily found with the Max-Likelihood estimators of the marginal probabilities, which both reads as

$$\hat{p}_{10} = \hat{p}_{01} = \frac{p_{01}p_{10}}{\alpha}$$

Thus the $s^{(1)}$ distribution is

	x_{10}	
x_{01}	-1	1
-1	$(1 - \frac{p_{01}p_{10}}{\alpha})^2$	$\frac{p_{01}p_{10}}{\alpha}(1 - \frac{p_{01}p_{10}}{\alpha})$
1	$\frac{p_{01}p_{10}}{\alpha}(1 - \frac{p_{01}p_{10}}{\alpha})$	$(\frac{p_{01}p_{10}}{\alpha})^2$

The Kullback-Leibler Divergence reads as

$$D[s : s^{(1)}] = \frac{p_{01}p_{10}}{\alpha} \log\left[\frac{\alpha}{p_{01}p_{10}}\right] + \left(1 - \frac{p_{01}p_{10}}{\alpha}\right) \log\left[\frac{\alpha}{\alpha - p_{01}p_{10}}\right] > 0$$

Remember that $D[p : q] \geq 0$ for every p, q and $D[p : q] = 0$ if and only if $p = q$. Since clearly $s^{(1)} \neq s$ for every value of $\alpha \in (0, 1)$ we have that the KLD between s and $s^{(1)}$ is never null.

Consider now to apply the $\mathcal{L}^1$ map $T = L(x_{10}, x_{01})$. The other map in the class $\mathcal{L}^1$, $L(x_{01}, x_{10})$ leads to identical results and it is never considered in this section. Since $T(1, 1) = (1, 1)$ and $T(-1, -1) = (1, -1)$, after all the

individuals have been mapped with T the distribution s_T representing the transformed population is:

	y_{10}	
y_{01}	-1	1
-1	0	$1 - \frac{p_{01}p_{10}}{\alpha}$
1	0	$\frac{p_{01}p_{10}}{\alpha}$

To compute $D[s_T : s_T^{(1)}]$ one first estimates the marginal probabilities $p_\beta = p(y_\beta = 1)$.

$$\begin{cases} \hat{p}_{10} = 1 \\ \hat{p}_{01} = \frac{p_{10}p_{01}}{\alpha} \end{cases}$$

Looking at the marginal probabilities it is easy to see that $s_T = s_T^{(1)}$ thus we have that $D[s_T : s_T^{(1)}] = 0$ for all α, s such as condition (6.5) is satisfied. This means that when FCA searches for the map $T \in \mathcal{L}^1$ which minimizes the $D[s_T : s_T^{(1)}]$ it chooses $T = L(x_{10},\, x_{01})$ for every starting distribution that satisfy (6.5) and for every function with the same values ordering as f. Consider what follows.

- We know that in the two variable case if the expected value of f calculated over the independence model has a saddle point inside the parameters domain then this never holds for $\mathbb{E}[f \circ L(x_{10},\, x_{01})]$.
- If the function f has no saddle point it is possible to see that opposite results holds and FCA never chooses to apply the map $L(x_{10}, x_{01})$ for distribution which satisfy (6.5).

This allows us to conclude that, at least for the two variable case and the starting distributions considered, the KLD minimization strategy employed by FCA leads to a favourable model choice. Note that the PBIL search strategy, if the infinite population holds, never converge to the global optimum for the starting distributions considered.

Consider now another set of starting distributions such that they all satisfy

$$p_{01}p_{10}+(1-p_{01})(1-p_{10}) < \alpha < p_{01}p_{10}+(1-p_{01})(1-p_{10})+p_{10}(1-p_{01}) \quad (6.6)$$

Look at the highlighted region in Figure 6.3. If a distribution s satisfies this condition it means that sampling from s and then performing truncation selection with rate α the resulting population is composed only by the individuals $(1, 1)$, $(-1, -1)$ and $(1, -1)$. The distribution s representing these populations is shown in the following table.

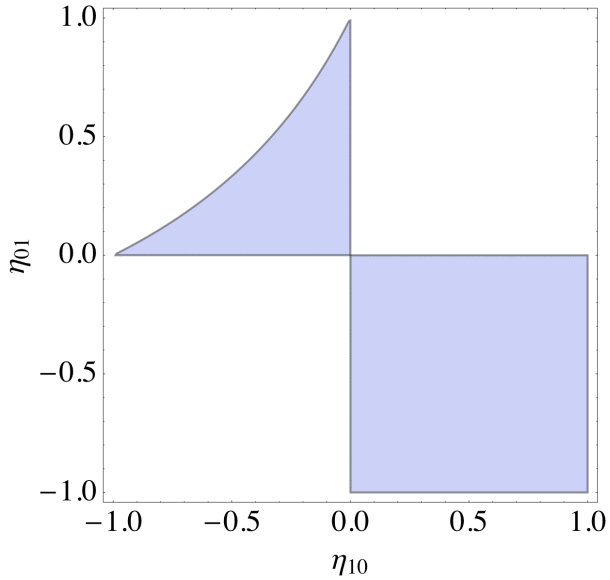

Figure 6.3: Independent distributions which satisfy (6.6) in the η parametrization, with $\alpha = \frac{1}{2}$

	x_{10}	
x_{01}	-1	1
-1	$\frac{(1-p_{01})(1-p_{10})}{\alpha}$	k
1	0	$\frac{p_{01}p_{10}}{\alpha}$

$$k = 1 - \frac{p_{01}p_{10} + (1-p_{01})(1-p_{10})}{\alpha}$$

The expressions of $D[s : s^{(1)}]$ and $D[s_T : s_T^{(1)}]$, where $T = L(x_{01}, x_{10})$, are quite verbose and difficult to compare analytically. We know that the expected fitness of f calculated on the independence model has a saddle point so we are interested in which are the distributions satisfying (6.6) such as FCA find convenient to apply T and thus switch to an independence model over the $y = T(x)$ variables.

We have determined numerically the subset of s satisfying condition (6.6) and such as

$$D[s : s^{(1)}] > D[s_T : s_T^{(1)}] \tag{6.7}$$

The set of distributions for which (6.7) holds is highlighted in red in Figure 6.4

It is possible to see in Figure 6.4 that there exist a set of distributions near the vertex $(1, -1)$ for which the independence model on the x variables is nearer in KLD to the distribution representing the selected population

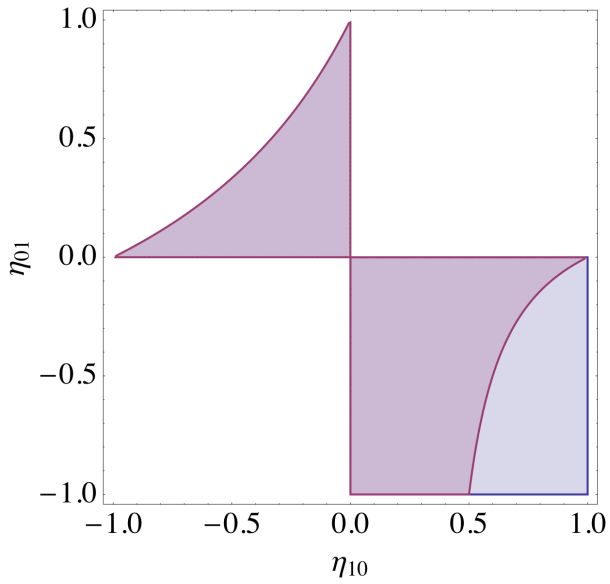

Figure 6.4: The set of independent distributions which satisfy conditions (6.6), in red, with $\alpha = \frac{1}{2}$

and thus no map T is applied. Remember that FCA behaves like PBIL if no map is applied, thus the trajectories starting from the neighbourhood of $\eta = (1, -1)$ move towards the center of the parameter space and enter the region in which the independence model over $L(x_{10}, x_{01})$ becomes nearer.

The third set of starting distributions considered is the one for which holds that

$$p_{01}p_{10} + (1 - p_{01})(1 - p_{10}) + p_{10}(1 - p_{01}) < \alpha < 1 \tag{6.8}$$

The distributions which satisfy this condition are shown in Figure 6.5. After sampling from s satisfying (6.8) and performing truncation selection with rate α each variable assignment still exists in the selected population. This case is similar to the previous one. We have determined numerically the set of distribution for which $D[s : s^{(1)}] > D[s_T : s_T^{(1)}]$, highlighted in red in Figure 6.6.

Like in the previous case, there exist a set of distributions near the vertex $(-1, 1)$ for which the independence model on the x variables is nearer in KLD to the distribution representing the selected population and thus no map T is applied. Note the symmetry of such distributions with respect to the ones in Figure 6.4. Again, like in the previous case the same reasoning lead us to conclude that trajectories starting in the neighbourhood of $(-1, 1)$ move towards the center and then to the global optimum, first on the independence

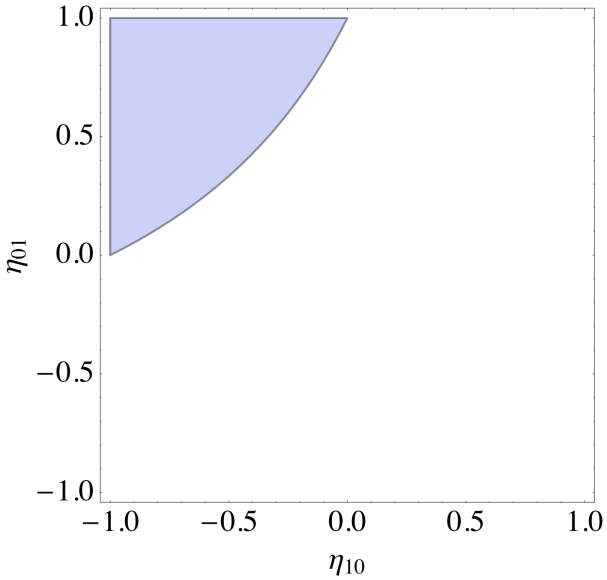

Figure 6.5: The set of independent distributions which satisfy conditions (6.8) with $\alpha = \frac{1}{2}$

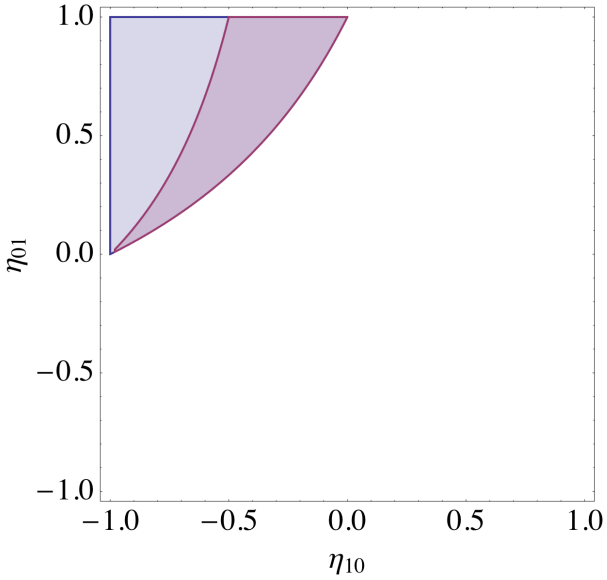

Figure 6.6: The set of independent distributions which satisfy conditions (6.6), in red, with $\alpha = \frac{1}{2}$

model over the x variables and then on the one defined over the transformed variables $y = T(x)$.

The last subset of starting independent distributions, i.e., the ones for

which holds

$$0 < \alpha < p_{01}p_{10} \tag{6.9}$$

is trivial since the population after selection is composed only by individuals $(1, 1)$, which are global optima for f. Since the resulting distribution s has reduced support on one vertex of the probability simplex it belongs to both the independence models on the x and the $y = T(x)$ variables. Thus both $D[s : s^{(1)}]$ and $D[s_T : s_T^{(1)}]$ are zero. Convergence to the global optimum is thus achieved in one step for every starting distribution that satisfies (6.9).

We have seen in this section how selection reveals correlations between variable values of the best solutions and how the KLD minimization strategy proposed is able to suggest the correct model in the two dimensional case.

6.3.2 Observed Algorithm Behaviour

In this section we discuss the behaviour of our implementation of FCA and we show that with large enough population it resembles the one derived in the previous section.

Here we deal with the sample function f whose coefficient vector is

$$c = \left(\frac{5}{2},\ \frac{1}{2},\ 0,\ 1\right)$$

This function has the same fitness ordering assumed in the previous section analysis: $f(1, 1) > f(-1, -1) > f(1, -1) > f(-1, 1)$. We already know that this ordering implies the presence of a saddle point in $\mathbb{E}[f]$ calculated on the independence model. From the f coefficients we have that its coordinates are $\eta = (0, -\frac{1}{2})$. In Figure 6.9(b) we have plotted the dynamics of the PBIL search strategy dealing with the function f.

To represent the dynamics of FCA three dimensions are needed since in general the distributions the populations are sampled from do not belong to the independence model. We have already seen in the previous chapter that there exists three two-dimensional models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ which can be defined over the x variables in terms of variables maps in $\mathcal{L}^1$. See Section 5.2.3.

All the trajectories have been plotted running the FCA algorithm on function f with population size 10000, selection rate $\alpha = 0.5$ and learning rate 0.7 to slow down the convergence. In this section we always refer to the models $\mathcal{M}_1$, the independence model over the x variables, purple in the figures, and $\mathcal{M}_2$, the independence model associated with the map $L_1 = L(x_{10}, x_{01})$, blue in the Figures.

In firgure 6.7 we have plotted the trajectory of FCA starting from the initial independent distribution $\eta = (-\frac{1}{2}, -\frac{1}{2})$. As foreseen, the model $\mathcal{M}_2$ is

nearer then $\mathcal{M}_1$ to the distribution of the selected individuals, thus the map $L(x_{10}, x_{01})$ is applied. It is possible to see that once $L(x_{10}, x_{01})$ is chosen $\mathcal{M}_2$ is never left and trajectories progressively tend to the model till convergence to the global optimum in $(1,\ 1,\ +\infty)$. Note that if the learning rate is set to 1, i.e., every individual of the population at iteration $k-1$ is discarded, for this starting condition the distribution representing the selected part of the population already has $\theta_{11} = +\infty$ at iteration 2. Thus the algorithm immediately jumps on the edge $(-1,-1)$, $(1,1)$ of the probability simplex and then follows it till convergence in $(1,1)$.

Running FCA starting from distributions satisfying condition (6.5) it is possible to see that there is a small set of distributions in the neighbourhood of the independent distribution $\eta = (-1,-1)$ starting from which FCA converges to $(-1,-1)$. This region gets smaller and smaller as the population increases, better approximating the infinite population behaviour.

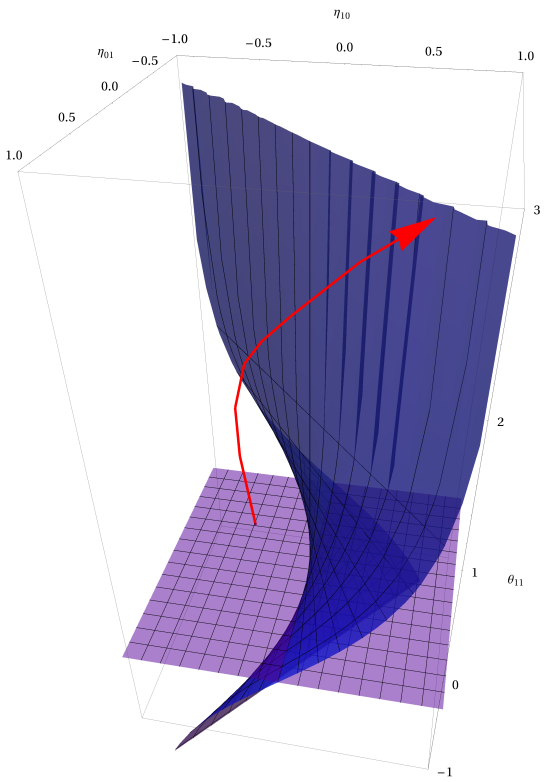

Figure 6.7: A FCA trajectory starting from distribution the independent distribution $\eta = (-\frac{1}{2}, -\frac{1}{2})$

In the previous section we have seen that there exist two sets of distribution in the neighbourhood of $\eta = (-1,\ 1)$ and $\eta = (1,\ -1)$ for which the independence model is preferred to both $\mathcal{M}_2$ and $\mathcal{M}_3$. In Figure 6.8 we have plotted the trajectory starting from the distribution $\eta = (-0.95,\ 0.95)$, which satisfies condition (6.7). It is possible to see that FCA moves towards the inner part of the independence model for some iterations, then applies the map $L(x_{10}, x_{01})$, switches to $\mathcal{M}_2$ and keeps following it till convergence in $(1, 1, +\infty)$.

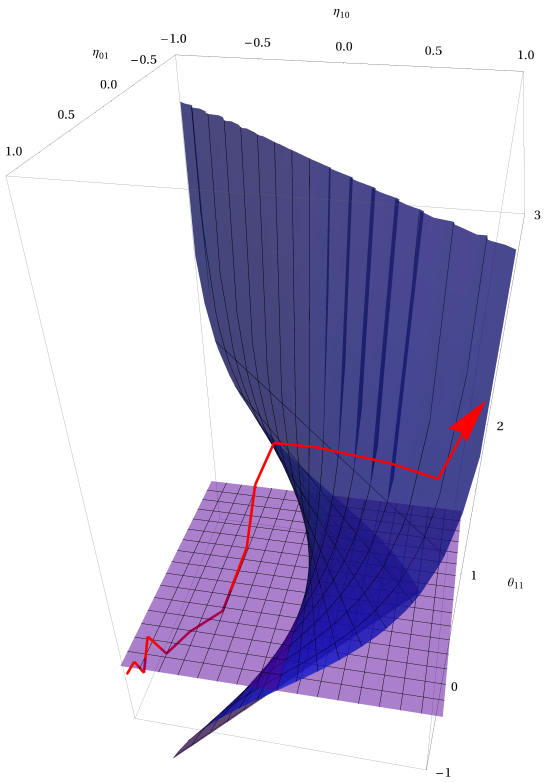

Figure 6.8: A FCA trajectory starting from distribution the independent distribution $\eta = (-0.95, 0.95)$

It is interesting to look at the projection of trajectories on the independence model, some of which have been plotted in Figure 6.9(a). It is important to note the trajectories in general do not lay on the independence model and this is the reason why they are so different from the PBIL ones. From Figure 6.9(a) is clear that almost all the starting distribution are in-

cluded in the attraction basin of the global optimum $(1, 1)$. This confirms the analysis of the previous section.

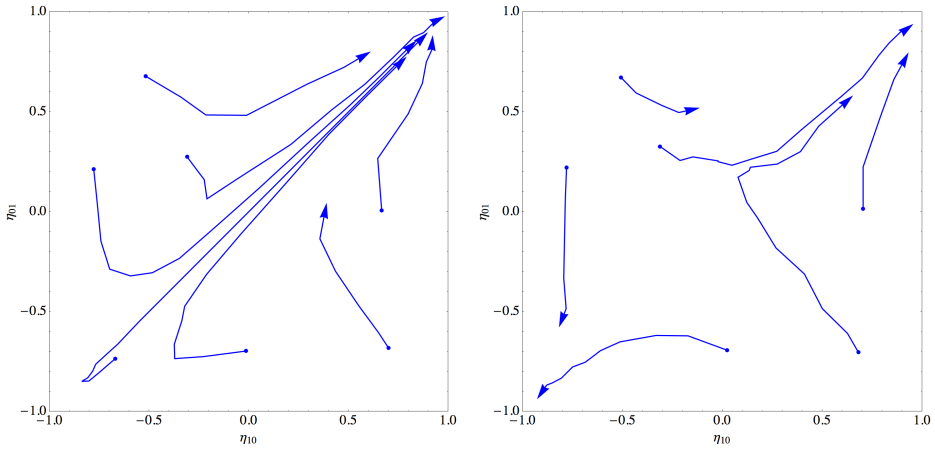

Figure 6.9: The projection on the independence model of some FCA trajectories on the left and some of PBIL on the right both maximizing function f

Consider now the fitness function g whose coefficient vector is $c = (0,\ 1,\ 3,\ 0.5)$. The expected value of g calculated over the independence model has no saddle point inside the η domain, thus we expect that no map T is applied by FCA and that its behaviour resembles the PBIL one. In Figure 6.10(a) and 6.10(b) some of the trajectories of FCA and PBIL have been plotted for different starting distributions. These figure confirms the analytical results for functions with no saddle point in their expected fitness over the independence model.

In this section we have analysed the behaviour of the proposed the proposed map selection strategy based on the greedy minimization of the Kullback-Leibler Divergence. We have seen that this strategy leads to the choice of a good map for the two dimensional case by means of analytical analysis of the KLD expressions and of simple experiments with our implementation of FCA

6.4 Conclusions

In this chapter we have proposed a novel search strategy based on the choice of univariate models associated with variables map belonging to the class $\mathcal{L}^k$. We consider this strategy as the first and most intuitive way to try to exploit the previous chapter results and intuitions.

FCA employs an univariate probability model defined over the variables

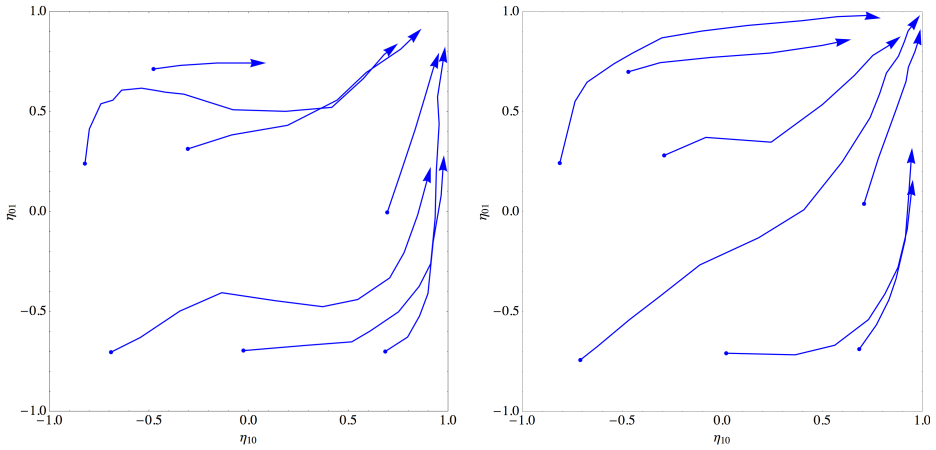

Figure 6.10: The projection on the independence model of some FCA trajectories on the left and some of PBIL on the right both maximizing function g

$y = T(x)$ where T is an opportune map in $\mathcal{L}^k$. The choice of the map T is done by means of a greedy Kullback-Leibler Divergence minimization technique which tries to minimize the amount of mutual information lost during the traditional EDAs estimation and sampling steps.

We have analysed the behaviour of this strategy in the two variable case by means of analytical comparison of the KLD expressions and then we observed the real trajectories followed by our FCA implementation. The results of the last section suggest that the KLD minimization procedure proposed could actually lead to the choice of a model on which the expected value of the fitness function is easy to traverse for a gradient based search strategy.

In the following chapter we analyse the behaviour of the FCA search strategy on bigger and more complex fitness functions.

Chapter 7

Experimental Results

In Chapter 4 and 5 we have presented the results of an analysis of the relations between the expected value of a pseudo-boolean function f, calculated over a certain probability model $\mathcal{M}$, and the performance of search strategies for the (R) problem. In Chapter 6 we have proposed the Function Composition Algorithm, a novel search strategy belonging to the Estimation of Distribution Algorithms family. FCA exploits a wide family of univariate models associated with variable maps of the class $\mathcal{L}^k$ which are chosen by means of a Kullback-Leibler Divergence minimization technique. The analysis conducted on the two variable case gave encouraging results. In this chapter we present the preliminary experimental results which came from the application of the FCA search strategy on two well known test functions, Alternate Bits and F3-Deceptive. After an introduction to these fitness functions and the software tools employed, we discuss a preliminary tuning of the critical parameters of FCA: selection and learning rates, KLD minimization threshold and population size. Then we compare the performances of FCA with the already mentioned Population Based Incremental Learning and Stochastic Gradient Descent strategy in terms of best and average solution found after a fixed number of iterations.

7.1 Experimental Framework

In this section we describe the experimental environment in which FCA has been tested and compared with other reference search strategies. This

include the two pseudo-boolean fitness functions Alternate Bits and F3-Deceptive, the two reference search strategies PBIL and SGD and the Evoptool software suite which was used to implement FCA and run the experiments.

7.1.1 Test Fitness Functions

In this sections we discuss the two fitness function we have used in the FCA experiments: Alternate Bits and F3-Deceptive.

Alternate Bits

Alternate Bits [10], also called 1D Checkerboard, is a pseudo-boolean function which introduce dependencies between couples of adjacent variables defining a chain-like structure. The value of a variable relative to its neighbours in the chain is taken into account and higher fitness are achieved when adjacent variables take different values.

More precisely, starting from the second variable one fitness point is gained if $x_h = -x_{h-1}$. The first variable value has no relevance. It is easy to see that there exists two global optima for Alternate Bits, which are obtained with $x_1 = \pm 1$ and $x_h = -x_{h-1}$, and that the maximum for $f(x)$ is $n - 1$. A three variables example is the following:

x_{100}	x_{010}	x_{001}	$f(x)$
-1	-1	-1	0
-1	-1	1	1
-1	1	-1	2
-1	1	1	1
1	-1	-1	1
1	-1	1	2
1	1	-1	1
1	1	1	0

The c_α coefficients are the followings. The pairwise dependencies between neighbours variables are visible in the coefficients c_{110} and c_{011}.

$$c_{000} = 1 \quad c_{001},\, c_{010},\, c_{100} = 0 \quad c_{110},\, c_{011} = -\tfrac{1}{2} \quad c_{101},\, c_{111} = 0$$

This fitness function is interesting since it is one of the simplest which include structured pairwise dependencies. The lack of the coefficients c_α with $|\alpha| = 1$ makes essential to employ probability models more complex than the independence one to enclose the features of good fitness individuals.

F3-Deceptive

The F4-Deceptive fitness function belongs to the wider class of Deceptive Functions which appeared in the literature to show the limit of Genetic Algorithms. As presented in Chapter 2 One of the key assumptions of genetic algorithms is that the best fitness individual can be obtained composing small subsets of variable assignments, often called Building Blocks, which by themselves produce good fitness. It can be shown that this is equivalent to assume that the maximum order of dependencies among variables is at most $k \ll n$, where n is the number of variables.

In deceptive functions there are building blocks which can be composed to produce high fitness but the optimum obtained in this way is not the global one. In Figure 7.1 it is possible to see the fitness landscape of a five variables deceptive function. Note that there are position independent building blocks, i.e., $x_h = 1$, which composed lead to fitness 4. Instead the highest fitness is achieved with $x = \{0,\ 0,\ 0,\ 0,\ 0\}$.

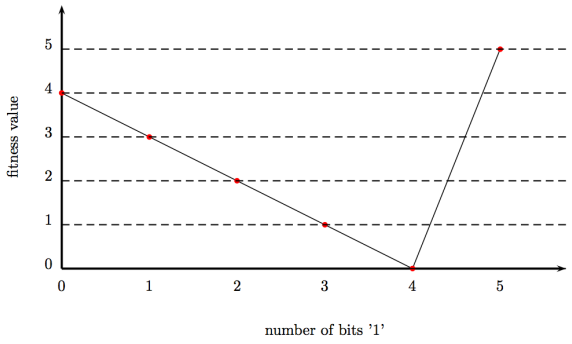

Figure 7.1: The fitness landscape of a 5 variables trap function

For the FCA experiments we have employed a order three deceptive function whose fitness values over the domain Ω are shown in the following table.

x_{100}	x_{010}	x_{100}	$f(x_{100},\ x_{010}\, x_{001})$
-1	-1	-1	1
-1	-1	1	0
-1	1	-1	0
-1	1	1	0.8
1	-1	-1	0
1	-1	1	0.8
1	1	-1	0.8
1	1	1	0.9

The coefficients of the polynomial expansion of f are

$$c = \Big(0.5375,\ 0.0875,\ 0.0875,\ 0.0875,\ 0.1375,\ ,0.1375,\ 0.1375,\ -0.3125\Big)$$

The expected value of f calculated over the independence model has two saddle points in

$$\overline{\eta} = \Big\{(-0.248186, -0.248186, -0.248186), (1.12819, 1.12819, 1.12819)\Big\}$$

In our experiments we have juxtaposed k F3-Deceptive [32] to form an $n = 3k$ variables fitness function. These test functions are interesting since they allow us to test the performance of search strategies dealing with several high order structured dependencies between variables.

7.1.2 Reference Algorithms

For this preliminary experiments on the performances of FCA, two search strategies have been chosen as reference: Population Based Incremental Learning and Stochastic Gradient Descent.

The first has been already described in Chapter 2 and it was analysed as a reference all throughout this work. At every iteration PBIL estimates the independent distribution which better fits the selected population sample by means of max-likelihood estimators of the marginal probabilities. A new population is then sampled from this distribution. PBIL explicitly moves along the independence model defined over the x variables.

Like PBIL, FCA employs an independence model, but defines it over the variables y obtained mapping x with an opportune variable map T in the class $\mathcal{L}^k$. It has been shown that this actually enrich the class of models available while keeping fixed and equal to n the number of free parameters required to fully characterize a probability distribution. PBIL was thus chosen since it allows to evaluate precisely the effect of the function compositions while keeping fixed the rest of the search strategy parameters.

The other reference algorithm, Stochastic Gradient Descent, tries to approximate the theoretical Exact Gradient Descent strategy on the independence model kept as reference all throughout this work. At every iteration an approximation of $\nabla\mathbb{E}[f]$ over the independence model is calculated using a subset of selected individuals from the population by means of the covariance besed estimator proposed in Theorem 5. Details about SGD can be found in [21]. This strategy, when the infinite population assumption holds, follows the projection of the gradient $\nabla\mathbb{E}[f]$, calculated for distributions in $\mathcal{P}$, on the model $\mathcal{M}$ employed, which in our case is the independence model. Thus if $\mathbb{E}_{\mathcal{M}}[f]$ has a saddle point SGD should converge to a local minima of

$\mathbb{E}[f]$ for certain starting conditions. We employ SGD with the independence model and we keep its performances as a reference when the expected fitness calculated over the model employed is known to have saddle points.

7.1.3 Figures of Merit

For this preliminary analysis of FCA we have chosen two figures of merit:

- Best fitness individual found
- Mean population fitness

We have chosen to sample this values after 25 iterations. This value has been chosen to be high enough to allow all algorithms to converge. To obtain meaningful values and noise independent values we have considered the average of these figures of merit over various runs of the compared algorithms.

7.1.4 Software Tools: Evoptool

Evoptool is an optimization toolkit that implements a set of algorithms based on the Evolutionary Computation paradigm. Evoptool provides a common platform for the development and test of new search strategies, in order to facilitate the performance comparison activity. The toolkit offers a wide set of benchmark problems, from classical toy examples to complex tasks, and a collection of implementations of algorithms from the Genetic Algorithms and Estimation of Distribution Algorithms paradigms. We have extended the Evoptool implementation of the PBIL algorithm to include the proposed KLD minimization strategy for the choice of $T \in \mathcal{L}^k$.

It is important to note that the graphs produced by Evoptool are scaled in a way that every fitness function f implemented returns values in the range [0, 100]. In Evoptool all the algorithms are meant to maximize the value of the fitness function.

7.2 Tuning Algorithm Parameters

In this section we discuss a preliminary analysis on the effect of the parameters which characterize the FCA search strategy: the selection and learning rates, and threshold on the minimum acceptable KLD reduction during the iterative building of the map $T \in \mathcal{L}^k$.

7.2.1 The KLD Threshold

As it was already anticipated in the previous chapter, to avoid model overfitting on a noisy selected population, especially with low selection rates or small populations, it seems reasonable to stop the application of further maps in $\mathcal{L}^1$ when these produce a KLD reduction lower than a certain threshold.

Since the Kullback-Leibler Divergence depends on a number of factors such as the number of variables n, the size of the selected population, the fitness function f and the parametrization of the distributions considered, it seems unreasonable that a fixed value for the threshold leads to good performances. An example of this fact is discussed in later sections.

In our experiments we have chosen to limit the overall maximum KLD reduction which can be achieved composing f with a map in $\mathcal{L}^k$. More precisely, we allow the concatenation of the current map T with another element $L \in \mathcal{L}^1$ if the KLD associated with $T \circ L$ is greater than 0.2 times the one associated with T_{id}, i.e., the identity map. All the experiment discussed in this chapter are run with this policy.

However, it could also be reasonable to consider no threshold at all and limit the overall number of concatenations allowed with the parameter k. The performances obtained with this setting are discussed in one case in later sections.

7.2.2 Learning and Selection Rates

Here we discuss the effects of the learning and selection rates on the performances of the FCA search strategy. A number of experiments have been performed on the maximization of the 20 variables Alternate Bits function with different selection rate α and learning rate γ. The k parameter characterizing the class $\mathcal{L}^k$ of maps considered was set to 10000 to approximate the behaviour with $\mathcal{L}^\infty$. We observed different behaviours with different population sizes.

In Figure 7.2 the best fitness and the average population fitness (averages of 100 runs of FCA) at iteration 25 are plotted as a function of the selection rate α. γ has been kept fixed and equal to 1 for all the runs. It is possible to see that when the population is very small with respect to the size of Ω the best results are obtained with average selection rates.

In Figure 7.3 again the best fitness found at iteration 25 has been plotted as a function of the learning rate γ for the best and the worst selection rates, 0.4 and 0.8 to check the effects of the learning rate when the selection rate is kept fixed. It is possible to see that when the learning rate is introduced,

i.e., $\gamma < 1$, and part of the population of the previous iteration is reinserted in the current one, the performances deteriorate.

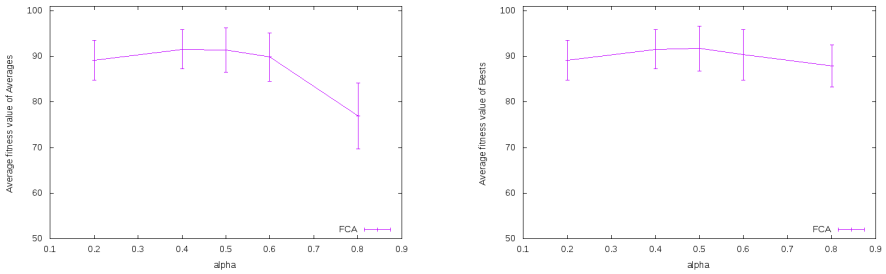

Figure 7.2: $\overline{f}$ *and best fitness found at iteration* 25 *with* $n = 20$, $m = 100$, $k = 10000$, $\gamma = 1$. *Average of* 100 *runs.*

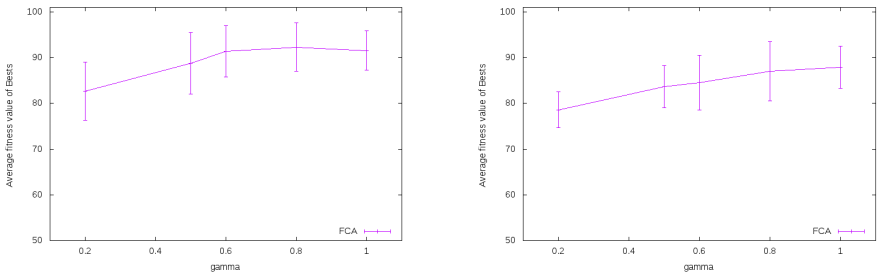

Figure 7.3: Best fitness found at iteration 25 *with* $n = 20$, $m = 100$, $k = 10000$, $alpha = 0.4$ *on the left,* $\alpha = 0.8$ *on the right. Averages of* 100 *runs.*

A different behaviour is observed when the population gets bigger, for example $m = 2000$. In Figure 7.4 the best fitness and the average population fitness (averages of 100 runs of FCA) at iteration 25 are plotted as a function of the selection rate α. γ has been kept fixed and equal to 1 for all the runs. It is possible to see that the performances deteriorate for low selection rates when the figures of merit are sampled at the end of iteration 25. This happens mainly because the velocity of the parameters change is slowed down with low selection rates and at iteration 25 convergence is not reached. Another effect is that low fitness individuals are preserved in the selected population, are translated into the model and reintroduced in the next iteration.

In Figure 7.5 we have again plotted the best fitness found at iteration 25

for the best and the worst selection rates, 0.2 and 0.8 to check the effects of the learning rate when the selection rate is kept fixed. It is possible to see that when the selection rate is 0.2 the learning rate γ does not change the performances, i.e., the global optimum $f(x) = 100$ is found for every rate γ. When α is 0.8 the introduction of a learning rate generally worsen the FCA performances.

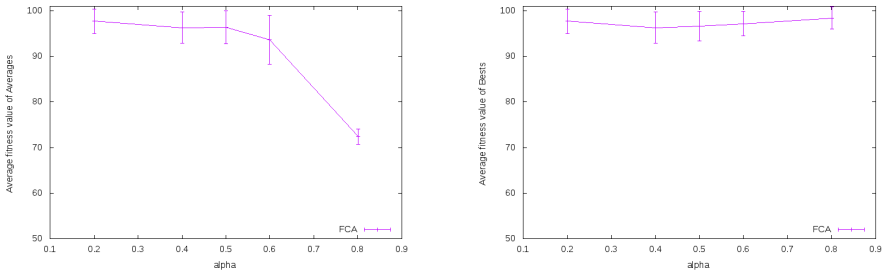

Figure 7.4: $\overline{f}$ and best fitness found at iteration 25 *with* $n = 20$, $m = 2000$, $k = 10000$, $\gamma = 1$. *Average of* 100 *runs.*

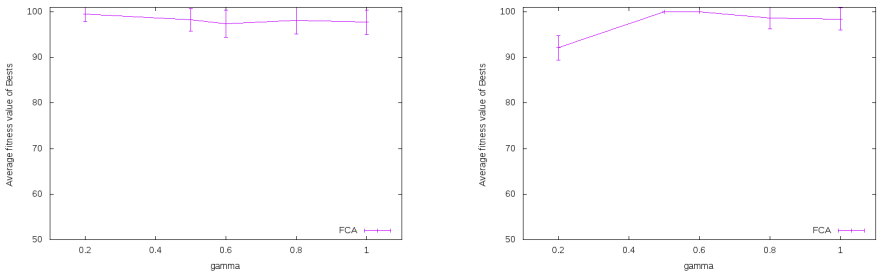

Figure 7.5: Best fitness found at iteration 25 *with* $n = 20$, $m = 2000$, $k = 10000$, $alpha = 0.2$ *on the left,* $\alpha = 0.8$ *on the right. Averages of* 100 *runs.*

These preliminary results lead us to conclude that learning rate 1 and high selection rates are good parameters choice. However, we use $\alpha = 0.5$ and for the rest of experiments, in order to compare FCA with other reference algorithms for which this parameter is fixed to 0.5 for implementation reasons. We employed $\gamma = 1$ both for FCA and PBIL.

Note that, because of the way the learning rate has been defined in FCA, at every iteration $1 - \gamma$ new individuals are sampled to keep the population size fixed to n. This means that if $\gamma < 1$ fewer fitness evaluations are

performed and thus at iteration t a smaller fraction of the search space Ω has been explored w.r.t. the case $\gamma = 1$. In some real-world applications the number of fitness evaluation is critical since this task is complex and time consuming. A more accurate analysis of the effects of the learning rate should take into account the best fitness individual found as a function of the number of fitness evaluations performed.

7.3 Test Case: Alternate Bits

In this section we analyse some experimental results of FCA on Alternate Bits function. In particular, we discuss the case with $n = 20$, $k = 1 \div 10000$, $\alpha = 0.5$, $\gamma = 1$ and $m = 20 \div 1000$, in relation with the reference algorithms, after 25 iterations and 50 runs. We consider selection rate 0.5 in order to maintain an uniform comparison with the Evoptool implementation of PBIL. Note that from now on all reference algorithms figures of merit are calculated on the average of 200 runs.

7.3.1 The Effects of the Choice of $\mathcal{L}^k$

The Figures 7.6 and 7.7 show the average and the best fitness found as a function of the maximum number of $\mathcal{L}^1$ maps k that the algorithm can concatenate. Note that FCA gives good results starting from $k \approx n$. We verified, with further tests, that this observation is, in general, always true, for different n and f. We consider $k = 2n$ a general rule in order to ensure best results in terms of best fitness found and computational time. Since FCA behaves like PBIL when $k = 0$, note that the use of maps always increases the performances.

7.3.2 The Population Size

The Figures 7.8 and 7.9 show the average and the best fitness found as a function of the population size m. Note that the performances of FCA are worse than PBIL for small population sizes. We conjecture that this is related to the fact that with too small population sizes the stochastic noise in the selected population becomes dominant and the choice of the $\mathcal{L}^k$ maps is inaccurate.

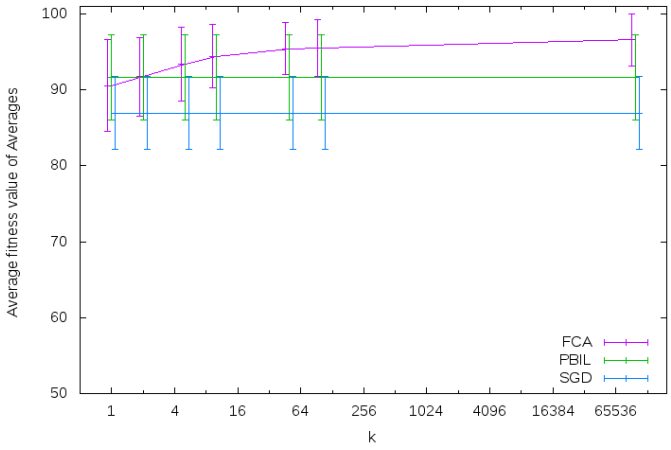

Figure 7.6: Alternate Bits: $\overline{f}$ at iteration 25, as function of k, with $n = 20$, $m = 1000$, $alpha = 0.5$. Averages of 100 runs.

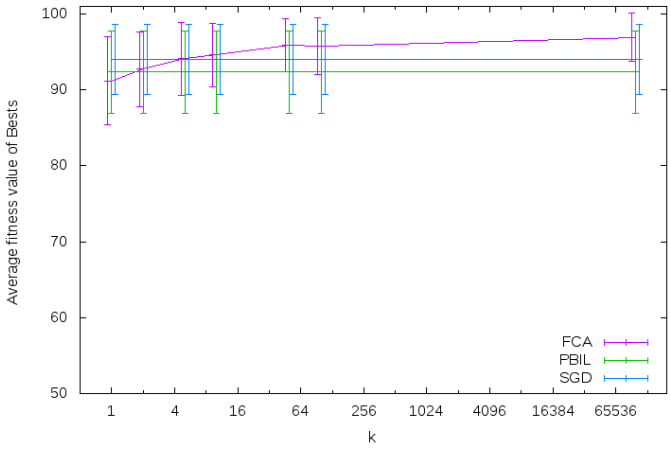

Figure 7.7: Alternate Bits: Best fitness at iteration 25, as function of k, with $n = 20$, $m = 1000$, $alpha = 0.5$. Averages of 100 runs.

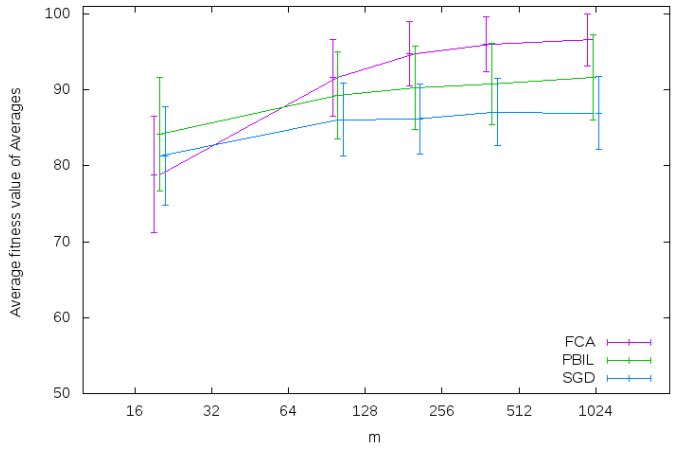

Figure 7.8: Alternate Bits: $\overline{f}$ at iteration 25, as function of m, with $n = 20$, $m = 100000$, $alpha = 0.5$. Averages of 100 runs.

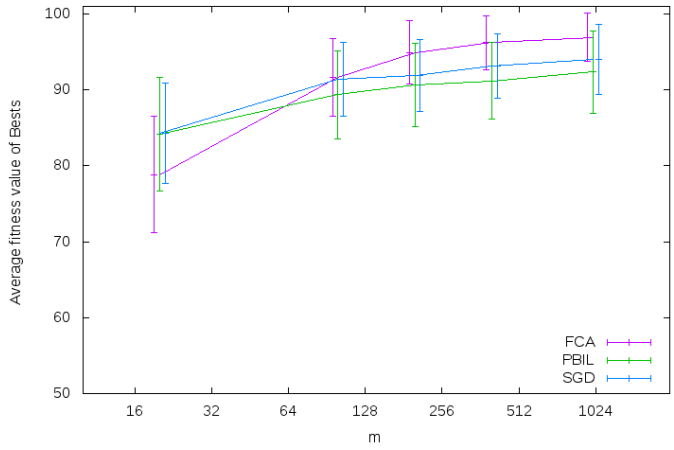

Figure 7.9: Alternate Bits: Best fitness at iteration 25, as function of m, with $n = 20$, $k = 100000$, $alpha = 0.5$. Averages of 100 runs.

7.4 Test Case: F3-Deceptive

In this section we analyse the behaviour of FCA on the F3-Deceptive fitness function in a 3-variables and n-variables cases.

7.4.1 3-variables Case

In this section we analyse one execution of FCA on 3-variable F3-Deceptive function with $m = 10000$ (infinite population approximation) and the $k = 2$ (with no threshold level). In the first iteration, the KLD values associated to any $\mathcal{L}^1$ map are shown in Table 7.1, where $T_1 \circ T_2 = T_{F3}$ is the selected map.

T_1	$KLD(f \circ T_1)$	$KLD(f)$
$L(x_{100}, x_{010})$	0.40695	0.45402
$L(x_{100}, x_{001})$	0.40695	0.45402
$L(x_{010}, x_{100})$	0.405697	0.45402
$L(x_{010}, x_{001})$	0.408407	0.45402
$L(x_{001}, x_{100})$	0.405697	0.45402
$L(x_{001}, x_{100})$	0.408407	0.45402
$T_1 = L(x_{010}, x_{100})$		
T_2	$KLD(f \circ T_1 \circ T_2)$	$KLD(f \circ T_1)$
$L(x_{100}, x_{010})$	0.40695	0.405697
$L(x_{100}, x_{001})$	0.358627	0.405697
$L(x_{010}, x_{100})$	0.45402	0.405697
$L(x_{010}, x_{001})$	0.339054	0.405697
$L(x_{001}, x_{100})$	0.357374	0.405697
$L(x_{001}, x_{100})$	0.290731	0.405697
$T_2 = L(x_{001}, x_{100})$		

Table 7.1: KLD values associated to any $\mathcal{L}^1$ map for the first iteration of the FCA algorithm on F3-Deceptive

From the composition of the function f with the map T_{F3}, we obtain the function g characterized by the coefficients:

$$c_{F3} = \Big(0.5375,\ 0.0875,\ 0.0875,\ 0.0875,\ 0.1375,\ 0.1375,\ 0.1375,\ -0.3125\Big)$$

and the fitness values:

x_{100}	x_{010}	x_{001}	$g(x) = f \circ T_{F3}$
-1	-1	-1	0
-1	-1	1	0.8
-1	1	-1	0
-1	1	1	1
1	-1	-1	0.8
1	-1	1	0
1	1	-1	0.8
1	1	1	0.9

The expected value of the fitness function g has no saddle points. Since the n-variables F3-Deceptive function is decomposable in equal 3-variables functions (building blocks), the absence of saddle points from the expected fitness function implies that we expect better performances for a Exact Gradient Descent search strategy, when the T_{F3} map is applied on every 3-variables function, e.g., 6-variables function, $T_{F3,6} = \Big(L(x_{010000}, x_{100000}) \circ L(x_{001000}, x_{010000}) \circ L(x_{000010}, x_{000100}) \circ L(x_{000001}, x_{000010})\Big)$. Note that the KLD minimization strategy proposed returned a map T such that a $\mathbb{E}[f \circ T]$, calculated on the independence model, has no saddle points.

At the second iteration, the algorithm reselect the same map. In the third iteration the map changes in $T_5 \circ T_6$, where $T_5 = L(x_{100}, x_{010})$, $T_5 = L(x_{010}, x_{001})$. The analytical analysis of $g(f \circ T_5 \circ T_6)$, omitted here, shows that the maps $T_5 \circ T_6$ removes critical point from the expected fitness function, as we have already seen for the map T_{F3}, thus we can consider the map $T_5 \circ T_6$ and T_{F3} equivalent. In later iterations, the distribution representing the selected population has reduced support and move towards a vertex of the probability simplex. In these situations no map is applied since every vertex belongs to all the models associated with maps in $\mathcal{L}^1$.

This example confirms that the KLD minimization could represent a good map selection strategy for the F3-Deceptive function.

It is possible to see that a fixed constant threshold on the minimum accepted KLD reduction for a map in $\mathcal{L}^1$ to be considered, for example 0.1 for this case, could lead FCA to apply no map and thus employ the independence model and achieving no saddle point removal. This is a further confirm that this parameter is critical for the FCA performances.

7.4.2 n-variables Case

Now we analyse some experimental results of FCA on F3-Deceptive function, when $n = 21$, $k = 1 \div 10000$, $\alpha = 0.5$, $\gamma = 1$ and $m = 21 \div 2100$, in relation with the reference algorithms, after 25 iterations and 50 runs.

As it can be seen in Figures 7.10 and 7.11, FCA reaches its best performance starting from $k \approx 21 = n$. This confirms the results previously seen in the Alternate Bits case, concerning to the choice of k. Note, in Figures 7.12 and 7.13, that, starting from small population size ($m = 42$), the average and best population fitness of FCA are always greater than those of reference algorithm PBIL. In this case, the modular structure of the F3 function permits to obtain a good KLD estimation with a small population. Note that the population size seems to weakly influence the performances of PBIL and SGD. This is a reasonable behaviour for those algorithms when the function is deceptive.

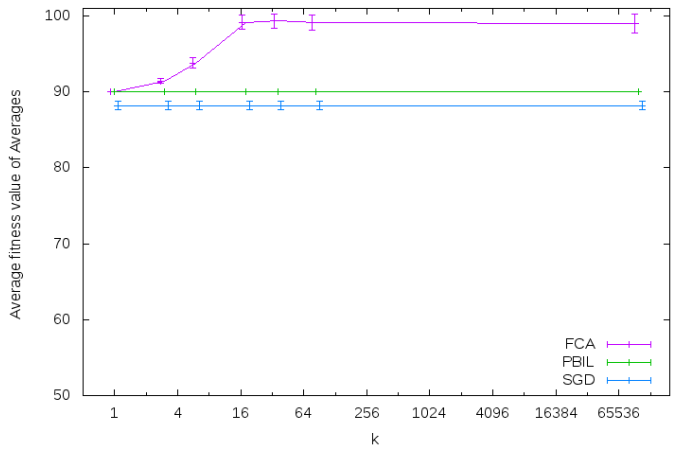

Figure 7.10: F3-Deceptive: $\overline{f}$ at iteration 25*, as function of* k*, with* $n = 21$*,* $m = 1050$*,* $alpha = 0.5$*. Averages of* 100 *runs.*

7.5 FCA without the KLD Threshold

The experimental results that we have presented in the previous section are generated with FCA with the KLD threshold, in order to limit a number of $\mathcal{L}^1$ maps applied at each iteration and reduce a computational time. However, the best performance of FCA are obtained when the threshold is not applied, because the algorithm can reach a better KLD reduction. Let consider an Alternate bits function, with $n = 20$, $k = 1 \div 10000$, $\alpha = 0.5$, $\gamma = 1$ and $m = 20 \div 1000$. The Figures 7.14 and 7.15 show the average and the best fitness of each iteration. As it can be seen, the optimum is always

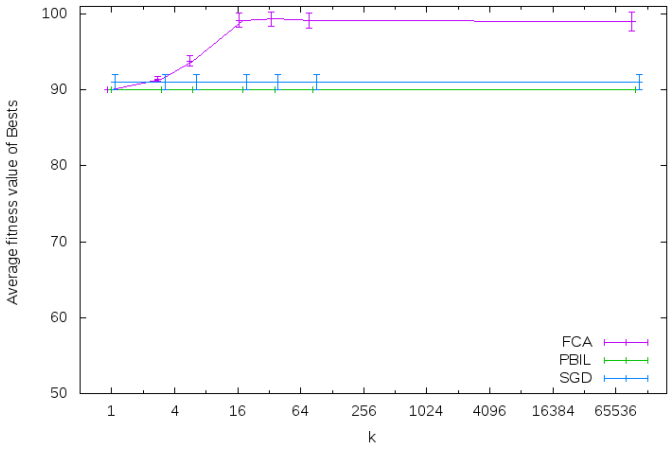

Figure 7.11: F3-Deceptive: Best fitness found at iteration 25*, as function of* k*, with* $n = 21$*,* $m = 1050$*,* $alpha = 0.5$*. Averages of* 100 *runs.*

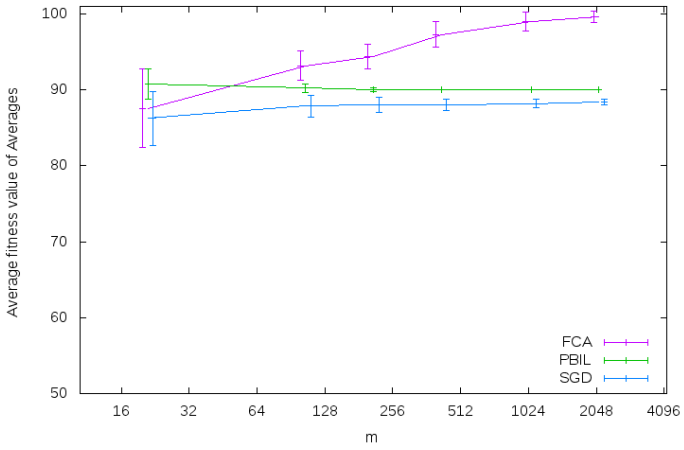

Figure 7.12: F3-Deceptive: $\overline{f}$ *at iteration* 25*, as function of* m*, with* $n = 21$*,* $k = 100000$*,* $alpha = 0.5$*. Averages of* 100 *runs.*

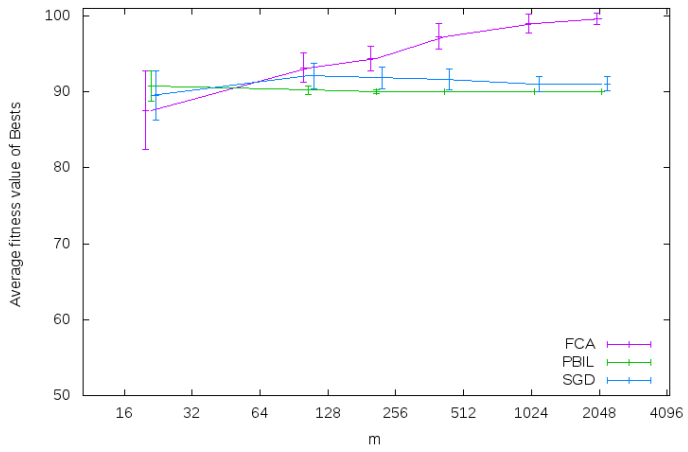

Figure 7.13: F3-Deceptive: Best fitness found at iteration 25, *as function of* m, *with* $n = 21$, $k = 100000$, $alpha = 0.5$. *Averages of* 100 *runs.*

reached from FCA, starting from the first iterations. This results confirm that the absence of the KLD threshold increments always the performances.

7.6 Conclusions

In this chapter we presented experimental results of FCA maximizing two well known fitness functions: Alternate Bits and F3-Deceptive. This functions are known to be difficult to optimize for search strategies which does not employ a complex probability model expressive enough to encode order two and three interactions between variables.

The preliminary results presented in this chapter seem to confirm the proposed KLD minimization strategy as a good model selection technique. We conjecture that the expected value of the fitness functions evaluated on the n parameters models found by FCA has less saddle points and attraction basins with respect to the independence model. This is the main reason why FCA performs better than SGD and PBIL. Note that these two algorithms employ the independence model while FCA uses a family of n free parameters models associated with maps in the class $\mathcal{L}^k$.

However, a number of aspects of FCA should be deeper investigated. These include:

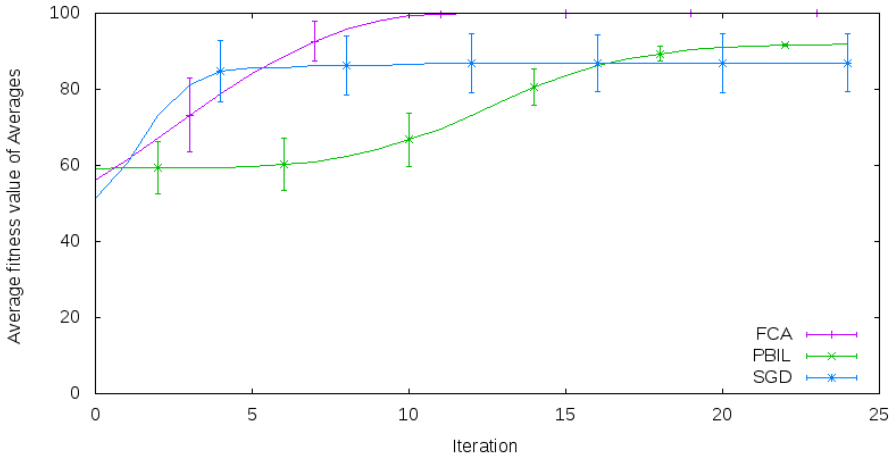

Figure 7.14: Alternate bits: $\overline{f}$ over iterations, with $n = 21$, $k = 100000$, $alpha = 0.5$, $m = 1000$. Averages of 100 runs.

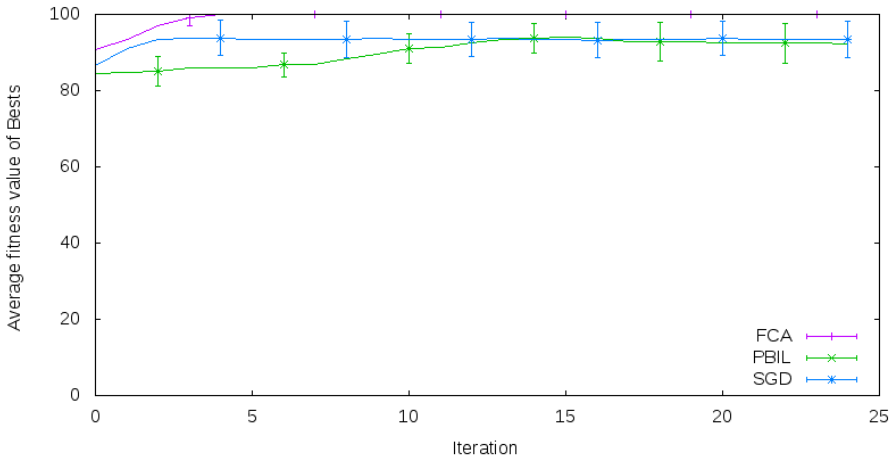

Figure 7.15: Alternate bits: Best fitness over iterations, with $n = 21$, $k = 100000$, $alpha = 0.5$, $m = 1000$. Averages of 100 runs.

- Best and average fitness found as a function of the number of fitness evaluations.
- FCA comparison with reference search strategy with similar complexity.
- The effective number of attraction basins for the Exact Gradient Descent strategy on the models found by FCA.
- Compare FCA with SGD when the model employed by SGD is more complex than the independence model.
- Timing analysis.

These analysis could result in FCA optimizations and further performances improvements.

Chapter 8

Conclusions and Further Work

In this work we have addressed some of the issues which arise when evolutionary search strategies are employed in the problem of optimizing a pseudo-boolean function f. In particular we focused on the Estimation of Distribution Algorithms. In these search strategies a population of candidate solutions is employed and a the fraction of "good" ones is used to estimate a probability distribution belonging to a model $\mathcal{M}$, which is later sampled to form a new generation of individuals. The main issue in the design of EDAs is the fact that the trajectory drawn by these search strategies in the $\mathcal{M}$ sub-manifold often converges to a distribution with reduced support in one local optimum for f.

In this work we related this search strategies to the stochastic relaxation of the function f which is the problem of optimizing its expected value calculated over a distribution belonging to the model $\mathcal{M}$. We have shown various examples of how the convergence properties of EDAs could be related to the presence of various attractors for the Exact Gradient Descent strategy on $\mathbb{E}_{\mathcal{M}}[f]$. In EDAs, a common approach to the problem of convergence to sub-optimal solutions for f is usually addressed increasing the complexity of the model $\mathcal{M}$ employed. A well known example of this solution is the already cited Bayesian Optimization Algorithm, which employees a very complex and expressive family of probability distributions based on Bayesian Networks to perform the estimation and sampling phases. This allows to

reduce the mutual information loss during the estimation and sampling steps and thus progressively enclose the features of good solution in the estimated distribution. In this way the sampled population resembles the selected fractions of the individuals and thus the exploration of the search space is guided by the fitness relevant features which appears in the population. Note that in most of the cases the family of the probability distributions employed by the algorithms is chosen at design time and thus is fixed.

In this work we have proposed an alternative approach: instead of choosing a fixed and expressive model at the beginning of the search process, we propose to switch dynamically among a bundle of simple models that we appositely defined. In particular we introduced the concept of one-to-one maps which associate each variable value assignment in the search space Ω with another one, eventually different. These maps between individuals can be employed to define an equivalence relation between probability models. A bundle of univariate models are obtained defining an independence model over a vector of binary variables and then applying a map to the support of the distributions belonging to this model.

The model choice is done during the whole search process by means of an Information Geometry criterion based on the minimization of the Kullback-Leibler Divergence, a notion of distance for the manifold of the probability distributions. In particular, we propose to chose every iteration the model which contains the nearest distribution to the selected population in terms of KLD. This procedure aims to minimize the mutual information loss during the estimation and sampling steps while keeping fixed and low the model complexity. This is considered positive since the estimation and sampling phases are easier in terms of computational complexity when simple models, e.g. univariate, are employed.

We designed and implemented an novel EDA, Function Composition Algorithm which at every iteration searches for a map T and then maps the selected population with T and estimates an independent distribution from these individuals. The model selection is achieved indirectly thanks to the composition of f with the map T. The addition of the map building phase based on the KLD minimization increases the complexity which is quadratic in the number of variables.

Experiments have been done to test the performances of FCA and compare it with Population Based Incremental Learning, another strategy which employs an univariate probability model, and Stochastic Gradient Descent, which instead tries to approximate the behaviour of the theoretical Exact Gradient Descent strategy. The experiments regarded the maximization of the pseudo-boolean functions Alternate Bits and F3-Deceptive, which are

known for containing structured order 2 or 3 interactions between variables. The preliminary results are encouraging, showing that the addition of the map selection and application step improves the performances in every case, leading FCA to reach the global optimum for f in almost all runs, given big enough populations. However, this is achieved with an increased complexity. Further investigations should be done comparing FCA with similar complexity algorithms, or, differently, when the same maximum running time or the number of fitness evaluation are fixed.

Further Work

In this work we have presented the idea of composing the function f to be optimized with a one-to-one map T. We have shown that if T is properly chosen, search strategies could find better solutions when optimizing $f \circ T$. The proposed strategy FCA is an extension of the PBIL search strategy to include a step in which the map T is built with an iterative greedy KLD minimization strategy. This approach can be employed to extend other search strategy, for example Bayesian Optimization Algorithm. The composition $f \circ T$ can be made transparent for the overlying search strategy. This is done simply applying T^{-1} every time the fitness of an individual has to be evaluated. This would lead to a further enrichment of the class of the models employed by search strategies at the price of an extra computationally intensive step in which the map T to compose with f is chosen properly.

A different idea came from the analysis of the FCA behaviour on the test fitness function employed. We observed that the choice of the map T at the first iteration is often decisive. More precisely, the maps chosen in following iterations are often very similar to the one at the first iteration. This suggest that the KLD minimization technique could be also employed as the model building step for strategies which are known to perform very good once a proper model has been selected. For example, a Stochastic Gradient Descent could be performed on the model employed in the first iteration of FCA. This could also be an experiment to solve the main open question about the KLD minimization heuristic which is, essentially, if it is able to find the model in which $\mathbb{E}_{\mathcal{M}}[f]$ has no saddle points or more than one attractor still exists for an Exact Gradient Descent search strategy.

A completely different direction is given when the composition of maps T with the function f to optimize is interpreted as a change of the encoding or, equivalently, as a non trivial *genotype-phenotype* mapping. The main observation is the fact that certain representations of the same information are more favourable for the evolutionary process. It has been shown in [34]

that a selection pressure on the representation (the *genotype*) is indirectly induced by the fitness based selection on the individuals (the *phenotype*) every time the encoding is not fixed. This means that there are certain encodings which more likely produce good fitness individuals. This suggest that the active searching of the map T could be replaced by an evolutionary process in the space of the representations which proceeds along with the one on the individuals. This is called σ-*evolution*. One comprehensive theory of the evolution of the representation analysed with information geometry can be found in [35].

Bibliography

[1] S. Aaronson. Guest column: Np-complete problems and physical reality. *SIGACT News*, 36:30–52, March 2005.

[2] S.I. Amari. Differential geometry of a parametric family of invertible linear systems—Riemannian metric, dual affine connections, and divergence. *Theory of Computing Systems*, 20(1):53–82, 1987.

[3] S.I. Amari. Information geometry on hierarchical decomposition of stochastic interactions. Citeseer preprint, 1999.

[4] S.I. Amari. Information geometry on hierarchy of probability distributions. *Information Theory, IEEE Transactions on*, 47(5):1701–1711, 2001.

[5] S.I. Amari and H. Nagaoka. *Methods of information geometry*. Amer Mathematical Society, 2007.

[6] S. Baluja. Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning. Technical report, Carnegie Mellon University, Pittsburgh, PA, USA, 1994.

[7] O. E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, New York, 1978.

[8] E. Boros and P.L. Hammer. Pseudo-boolean optimization. *Discrete Applied Mathematics*, 123(1-3):155–225, 2002.

[9] L. D. Brown. *Fundamentals of statistical exponential families: with applications in statistical decision theory*. Institute of Mathematical Statistics, Hayworth, CA, USA, 1986.

[10] A.E.I. Brownlee. *Multivariate Markov networks for fitness modelling in an estimation of distribution algorithm*. PhD thesis, The Robert Gordon University, 2009.

[11] I. Csiszár. On topological properties of F-Divergences. *Studia Sci. Math. Hungar.*, 2:329–339, 1967.

[12] I. Csiszár, T. Cover, and B.S. Choi. Conditional limit theorems under Markov conditioning. *Information Theory, IEEE Transactions on*, 33(6):788–801, 2002.

[13] I. Csiszár and F. Matúš. Closures of exponential families. *The Annals of Probability*, 33(2):pp. 582–600, 2005.

[14] C. Darwin. *On the origin of species, 1859*. New York University Press, 1988.

[15] S. Geman and D. Geman. *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, pages 452–472. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[16] D.E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989.

[17] G.R. Harik, F.G. Lobo, and D.E. Goldberg. The compact genetic algorithm. *Evolutionary Computation, IEEE Transactions on*, 3(4):287–297, 2002.

[18] M. Hohfeld and G. Rudolph. Towards a theory of population-based incremental learning. In *Proceedings of the 4th IEEE conference on evolutionary computation*, pages 1–5. IEEE Press, 1997.

[19] J.H. Holland. *Adaptation in natural and artificial systems*. MIT Press, Cambridge, MA, USA, 1992.

[20] S. Kullback. The Kullback-Leibler distance. *The American Statistician*, 41(4), 1987.

[21] L. Malagò, M. Matteucci, and B. Dal Seno. An information geometry perspective on estimation of distribution algorithms: boundary analysis. In *Proceedings of the 2008 GECCO conference companion on Genetic and evolutionary computation*, pages 2081–2088. ACM, 2008.

[22] L. Malagò, M. Matteucci, and G. Pistone. Towards the geometry of estimation of distribution algorithms based on the exponential family. In *In Proceedings of XI Foundation of Genetic Algorithms*, January To appear.

[23] M. Mitchell. *An introduction to genetic algorithms*. The MIT press, 1998.

[24] H. Mühlenbein. The equation for response to selection and its use for prediction. *Evol. Comput.*, 5:303–346, September 1997.

[25] H. Mühlenbein and T. Mahnig. Mathematical analysis of evolutionary algorithms. *Essays and Surveys in Metaheuristics, Operations Research/Computer Science Interface Series*, pages 525–556, 2002.

[26] H. Mühlenbein and G. Paaß. From recombination of genes to the estimation of distributions i. binary parameters. In *PPSN*, pages 178–187, 1996.

[27] H. Mühlenbein and D. Schlierkamp-Voosen. Analysis of selection, mutation and recombination in genetic algorithms. *Evolution and Biocomputation*, pages 142–168, 1995.

[28] H. Nagaoka and S.I. Amari. Differential geometry of smooth families of probability distributions. *Univ. Tokyo, Tokyo, Japan, METR*, pages 82–7, 1982.

[29] M. Pelikan, D.E. Goldberg, and E.E. Cantù-paz. Linkage problem, distribution estimation, and bayesian networks. *Evol. Comput.*, 8:311–340, September 2000.

[30] C.R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Breakthroughs in Statistics: Foundations and basic theory*, page 235, 1992.

[31] J. Rauh, T. Kahle, and N. Ay. Support sets in exponential families and oriented matroid theory. 2009.

[32] C. Sangkavichitr and P. Chongstitvattana. Dictionary based estimation of distribution algorithms. In *Communications and Information Technologies, 2007. ISCIT'07. International Symposium on*, pages 364–369. IEEE, 2007.

[33] C.E. Shannon and W. Weaver. The mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[34] M. Toussaint. Demonstrating the evolution of complex genetic representations: An evolution of artificial plants. In *Genetic and Evolutionary Computation—GECCO 2003*, pages 200–200. Springer, 2003.

[35] M. Toussaint. On the evolution of phenotypic exploration distributions. *Foundations of Genetic Algorithms*, 7:169–182, 2003.

[36] M. Toussaint. The structure of evolutionary exploration: On crossover, buildings blocks, and Estimation-Of-Distribution algorithms. In *Genetic and Evolutionary Computation—GECCO 2003*, pages 207–207. Springer, 2003.

[37] M. Toussaint. Compact genetic codes as a search strategy of evolutionary processes. In *Foundations of Genetic Algorithms*, volume 3469 of *Lecture Notes in Computer Science*, pages 364–366. Springer Berlin / Heidelberg, 2005.

[38] M.D. Vose. *The simple genetic algorithm: foundations and theory.* The MIT Press, 1999.