Schema Propagation in Evolution Programs

Zur Erlangung des akademischen Grades eines

Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der Fakultät für Wirtschaftswissenschaften
der Universität Fridericiana zu Karlsruhe (TH)

genehmigte

Dissertation

von

Dipl.-Wirtschaftsing. und Dipl.-Ing. Andreas Frick

Tag der mündlichen Prüfung:	16.2.2005
Referent:	Prof. Dr. Diethard Pallaschke
Korreferent:	Prof. Dr. Detlef Günter Seese

Karlsruhe, 2005

Abstract

Evolution programs and their relatives have succeeded more and more in attacking hard optimization and learning problems during the last years. There are immense numbers of successful applications, which justify their importance in practice. One of the first attempts to explain the way they work was Holland's conjecture of schema propagation during evolution. Although his theory based on that is still contested, it has been developed further by many researchers and is still one of the pillars of the theory of evolution programs. But until now there is no empirical evidence of the relevance of this theory.

In this thesis Holland's conjecture of schema propagation during evolution is picked up and combined with other approaches that also explore the mechanics of artificial evolution. A comprehensive analysis of all these models and their deficiencies is given and their relations to the models of natural evolution are considered. As a result, a simple but tractable and rigorous theoretical model of schema propagation during evolution is formed. To show the relevance of this model empirically, suitable indices have been developed first to measure the schema reach in an artificial population and to generally describe its state. Using the new indices then the schema reach is monitored in practice during the run of an example evolution program on different instances of the traveling salesman problem.

Both the theoretical models and the performed experiments show the tight relation of schema propagation and successful evolution. Schema propagation is indeed a prerequisite for the improvement of the objective values during evolution but no guarantee. However, it can easily be obstructed by higher mutation and crossover rates. The driving force behind selection is the diversity of the population, which causes the selection pressure. The genetic drift induced by the randomness of the selection process cannot be neglected in the theoretical models. It drastically reduces the survival probability of new superior individuals and also promotes the schema propagation, despite of only small differences of the objective values in the population. Consequently, special cases can be constructed, for which a deterministic model of the evolution process leads to totally wrong results.

"The urge to understand is paired with the urge for truth. But the perception still need not be the truth at all."

Donald Brinkmann

Acknowledgments

This work has been sponsored partly by benefactions of Ursula Frick and scholarships granted by the state of Baden-Württemberg and the Institute of Applied Computer Science and Formal Description Methods of the Universität Karlsruhe (TH). The author especially thanks Diethard Pallaschke, Karl-Heinz Waldmann, Manfred Krtscha, Stefan Napel, Ulrike Stocker, Armin Haas, Kalyanmoy Deb, James David Schaffer and Alfred Müller for their helpful remarks and corrections, Nikolaus Geers for his help to master *Maple*, Annemarie Bauer and Wolfram Schüffel for their mental support, Pamela Hamblin for correcting the manuscript and the Institute of Applied Computer Science and Formal Description Methods of the Universität Karlsruhe (TH) for providing some computing facilities.

Contents

Chapter 1

Introduction

Today evolution programs and their relatives have set up themselves to be suitable to attack hard optimization and learning problems. There are immense numbers of successful applications, which justify their importance in practice. Related to the still increasing spread of evolution programs, the investigation of the way they work became an important topic of research. Due to their practical success, most of the research is focused on experimental work. Since evolution programs have many parameters, and data structures and evolutionary operators may be related to a specific problem class, this is a very wide research area. Nevertheless theoretical research has also been performed. Already Holland, who introduced genetic algorithms, which are a root of evolution programs, explained their success by the propagation of schemata in the artificial population during evolution [46] and also presented a theoretical model of the propagation process. Later on, other researchers tried to improve this model or to modify its interpretation. Although this theory is contested, it is still a pillar for explanations of the way evolution programs work. But there are no experimental observations of schema propagation in practical applications and, consequently, there is no evidence for its applicability. The theory of Markov chains has been independently applied to evolution programs [69]. Since nearly all approaches start from the most general case, only general statements about evolution programs can be derived from them and an application of this theory to a specific case is very difficult. Especially, there seems to be no attempt to relate Holland's schema propagation theory to the Markov models. Although evolution programs are inspired by the example of natural evolution, the results gained by it mathematically are used only rarely for research about evolution programs. This is very astonishing, because mathematical models had been developed in genetics a long time before the introduction of evolution programs. The foundations were laid by Wright [103] and Fisher [27], who both focused their work on the role of selection in the evolution process. Later their models were improved and extended by other researchers, which led to the theory of population genetics.

The aim of this work is to combine the different approaches mentioned above, to point out their relations and to fill their deficiencies. Holland's conjecture of schema propagation is taken up and shown to be a prerequisite of a successful evolution. To elucidate some points of the consideration also empirically, the traveling salesman problem is used as an example problem because of its clarity. The text also gives a concise but rigorous basic introduction to the working of evolution programs, also summing up many results scattered throughout the literature.

Chapter 2

Hard Problems

2.1 Overview

In this chapter the motivation is given to use evolutionary algorithms. Despite the advances in providing faster and faster computers many problems can not be solved as fast as one would wish. Since of its simplicity as an example of this problems the traveling salesman problem is proposed and its properties are considered. Extending the discussion the relations to other hard problems are explored and some approaches to compute at least admissible approximations or to improve them are proposed. The consideration follows and summarizes mostly the clear articles of Johnson and Papadimitriou [49, 50], which can be recommended a lot as a starting point for a deeper understanding of this subject. Alternatively one may read chapter 34 of Cormen's et aliorum nice textbook [17, p. 966*ff*].

2.2 The Traveling Salesman Problem

The traveling salesman has to visit every city in his territory exactly once and then to return to the starting point. Given the costs to travel between all cities he wants to minimize his total expenses. The problem has been named according to the title of an ancient German book [70] instructing real world traveling salesmen how to succeed in their business. Euler [24] already struggled with the same problem when he tried to solve the knight's tour problem: a knight should visit each of the 64 squares of a chessboard exactly once in a single tour.

The mathematical formulation of the problem is more complicated than the informal description given above [68, p. 442*ff*]. In the most general version there is a weighted, directed and complete graph $G = (\boldsymbol{V}, \boldsymbol{E}, C)$. The vertices $\boldsymbol{V} = \{1, \ldots, n\}$ represent the cities, the edges $\boldsymbol{E}$ the paths between the cities, and the distance matrix $C = \langle c_{ij} \rangle$ the nonnegative costs to travel from city i to j with $c_{ij} \in \mathbf{Z}$. If the costs of a path between two cities depend on the tour direction the problem is called the *asymmetric traveling salesman problem*. If the relation $c_{ij} = c_{ji}$ holds, which relaxes the general case to have distances independent from the tour direction, it is called the *symmetric traveling salesman problem*. Now, a binary integer programming problem has to be solved. The variables

$$x_{ij} := \begin{cases} 1, & \text{if in the tour city } j \text{ follows city } i \\ 0, & \text{else} \end{cases} \qquad \forall i, j \in \{1, \ldots, n\}$$

determine the course of the tour. Therefore, the function of the traveling expenses

$$f(x) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij}$$

has to be minimized while satisfying the conditions

$$\sum_{i=1}^{n} x_{ij} = 1 \qquad \forall j \in \{1, \ldots, n\}$$

and

$$\sum_{j=1}^{n} x_{ij} = 1 \qquad \forall i \in \{1, \ldots, n\}.$$

So far, the requirements only describe the *assignment problem* that would allow a solution consisting of several disjoint subtours. Thus a tour also has to satisfy the one-cycle conditions

$$\sum_{i \in \boldsymbol{U}} \sum_{j \in \boldsymbol{U}} x_{ij} \leq |\boldsymbol{U}| - 1, \qquad \forall \boldsymbol{U} \subset \boldsymbol{V}, \qquad \boldsymbol{U} \neq \emptyset.$$

They ensure that in a subset $\boldsymbol{U}$ of the vertices the number of edges of the related subtour is shorter than required for a cycle. There are $2^n - 2$ of these restrictions and there are furthermore other formulations of these one-cycle conditions possible [22, section 3.1].

The traveling salesman problem (TSP) is a member of the class of combinatorial optimization problems. I. e. there is a problem domain $\mathbf{Z}^n$ on which an objective function $f : \mathbf{Z}^n \mapsto \mathbf{Z}$ is defined, which has to be maximized or minimized. Furthermore, there may be a set of m constraints $r_1, \ldots, r_m$, which have to hold, defined also on the same domain as the objective function. Any $X \in \mathbf{Z}^n$ that satisfies all m constraints is called an *admissible* solution of the optimization problem. Finally, there are one or more admissible solutions, which are optimal. In case of a maximization problem, there is no higher objective value possible than the objective value of the optimal solution(s), or in case of a minimization problem, there is no lower objective value possible than the objective value of the optimal solution(s).

Corresponding to the original optimization problem there is also the traveling salesman decision problem. Given again a graph G like above and additionally a bound, then the question is, whether there is a tour visiting each vertex exactly once with a length lower than the given bound.

2.3 Growth of Functions

In many situations only the asymptotic behavior of a function is of interest. The function then can be approximated or bounded by a another function easier to deal with [17, p. 966*ff*].

A function $f(x)$ is bounded asymptotically from above by a second function $g(x)$ if there are positive constants c_1 and n_0 such that

$$0 \leq f(n) \leq c_1\, g(n)$$

holds for all $n \geq n_0$. This fact is abbreviated by the notation $f(x) = \mathsf{O}(g(x))$. Equivalently the ratio of the two functions

$$\frac{f(n)}{g(n)} = c_1$$

must be bounded from above for all $n \geq n_0$. Analogously a function $f(x)$ is bounded asymptotically from below by a second function $g(x)$ if there are positive constants c_2 and n_0 such that

$$0 \leq c_2\, g(n) \leq f(n)$$

holds for all $n \geq n_0$. This fact is abbreviated by the notation $f(x) = \Omega(g(x))$. Equivalently the inequality

$$0 < \frac{f(n)}{g(n)} = c_2$$

has to hold for all $n \geq n_0$. If for two functions $f(x)$ and $g(x)$ both conditions $f(x) = \mathsf{O}(g(x))$ and $f(x) = \Omega(g(x))$ hold, then this is noted by $f(x) = \Theta(g(x))$. Then the function $g(x)$ is an asymptotically tight bound of the function $f(x)$. If especially

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1 \tag{2.1}$$

holds, then $g(n)$ is called an *asymptotic approximation* of $f(n)$.

Generally the asymptotic upper bound given by the O-notation may or may not be asymptotically tight. E. g. the bound $2n = \mathsf{O}(n^2)$ is not tight. Consequently the function $g(x)$ is called to be of lower order than the function $f(x)$ if for any positive constant $c > 0$ there exists a constant $n_0 > 0$ such that

$$0 \leq f(n) < c\, g(n)$$

holds for all $n \geq n_0$. This fact is abbreviated by the notation $f(x) = \mathsf{o}(g(x))$. Consequently then for the ratio of the two functions

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$$

holds. Analogous definitions can be performed for the lower bound.

The above considerations also can be used to characterize the behavior of a function if its argument approaches a certain point. From equation 2.1 then

$$\lim_{n\downarrow n_0} \frac{f(n)}{g(n)} = 1 \tag{2.2}$$

follows for a function $g(n)$ to be an asymptotic approximation of the function $f(n)$ for $n \downarrow n_0$. The compliance is denoted by $f(n) \sim g(n)$ [47, p. 7]. Finally the approximation error

$$\mathsf{e}(n) = f(n) - g(n)$$

is of some interest. For it

$$\lim_{n\downarrow n_0} \frac{\mathsf{e}(n)}{g(n)} = \lim_{n\downarrow n_0} \frac{f(n)}{g(n)} - \frac{g(n)}{g(n)} = \lim_{n\downarrow n_0} \frac{f(n)}{g(n)} - 1$$

holds. For $f(n) \sim g(n)$ then with equation (2.2)

$$\lim_{n\downarrow n_0} \frac{\mathsf{e}(n)}{g(n)} = 0$$

follows and finally $\mathsf{e}(n) = \mathsf{o}(g(n))$. Then near n_0 the error induced by the approximation becomes insignificant in relation to the approximating function.

2.4 Computational Complexity

In theoretical computer science the worst case running time of a program solving a given problem depending on its size is an important property. Generally there are two classes of problems, the *tractable* and the *intractable* respectively *hard* ones, which can be further divided into subclasses.

Problems are called *tractable* if the worst case running time of the best known algorithm is asymptotically bounded from above by a polynomial function of the problem size n. Abbreviated these problems are solvable in polynomial time and consequently they are called to belong to the class $\mathcal{P}$. All other problems not solvable in polynomial time belong to the class of intractable respectively hard problems. Two prominent subclasses are problems solvable only in exponential time and problems, which are provable unsolvable. Unfortunately, for the traveling salesman problem no polynomial time algorithm is known until now. The chances to find one in the future are also extremely low, because the traveling salesman problem belongs to the class $\mathcal{NP}$ of closely related problems, which has been investigated very intensively during the last decades. If a polynomial time algorithm for one of the problems would be found, also all other problems of that class could be solved in polynomial time. On the other hand, these problems until now could not be proved to require exponential time in relation to the problem size. Thus there is the conjecture of $\mathcal{P} \subset \mathcal{NP}$ and consequently $\mathcal{P} \neq \mathcal{NP}$. In the following some of the properties of the problems belonging to the class $\mathcal{NP}$ are considered informally and shortly.

For the present the consideration is restricted to decision problems like the traveling salesman decision problem presented above. There the related question can be affirmed by giving an appropriate tour. Then one can check the claim by computing the tour length. The tour is called a *succinct certificate*, because in the case of the traveling salesman decision problem the length of the certificate is a linear function of the problem size and the claim can by proved in polynomial time. Correspondingly, a definition is possible by the introduction of nondeterministic algorithms, which are unrealistic but theoretically ingenious. There the computation can be divided into different branches which are executed in parallel. Repeating this, the computation can spread to a tree which at most can grow as much as an exponential function of the number of division levels. If the computation of any of the branches leads to a result that can be used as a certificate to a positive decision, the answer is affirmative, else negative. Now a problem belongs to the class $\mathcal{NP}$ of problems if it can be solved by a nondeterministic algorithm in polynomial time.

The considerations above can be further extended. If a problem is in the class $\mathcal{NP}$ and any other problem of that class can be reduced to it in polynomial time, then it is called to be in $\mathcal{NP}$-complete. Thus one can solve any problem of this class by transforming it to any other member, solve that and retransform the solution to that of the original problem. Since this is a class property it is sufficient for the membership of a new problem to show that it is in $\mathcal{NP}$ and one of the problems already known to be members of the class can be reduced in polynomial time to it. E. g. the traveling salesman decision problem can be proved to be in $\mathcal{NP}$-complete by the reduction of the Hamilton cycle problem to it. The original traveling salesman problem is not in $\mathcal{NP}$-complete because it is not a decision problem although it is of even algorithmic complexity as the members of that class. Furthermore there are problems not known to be members of $\mathcal{NP}$ but all problems belonging to $\mathcal{NP}$ can be reduced to them in polynomial time. These two sets are summarized in the class $\mathcal{NP}$-hard.

2.5 Approximation Algorithms

Since there is no hope to find a deterministic polynomial time algorithm for an optimization problem belonging to the class $\mathcal{NP}$, one at least will pursue to find an *approximation algorithm* that will calculate a near optimal admissible solution in polynomial time. Thus given a constant factor r with $1 \leq r < \infty$ for each problem *instance* I the approximative solution $A(I)$ should satisfy

$$A(I) \leq r\, O(I)$$

with $O(I)$ being the optimal solution. Unfortunately this claim can not be satisfied, because e. g. such an approximation for the traveling salesman problem could be used to solve the Hamiltonian cycle problem, which is proven to be in the class $\mathcal{NP}$-complete.

The situation relaxes a little, if the traveling salesman problem is restricted to conform to the *triangle inequality*. Thus for the distances c_{ij} between the cities i and j the inequality

$$c_{ij} \leq c_{ik} + c_{kj}$$

holds for a third city k and for all $i, j, k \in \{1, \ldots, n\}$. The inequality is satisfied e. g. for planar problems and euclidean distance and ensures a roundabout way not to be an abbreviation. Then polynomial approximation algorithms at least for the symmetric traveling salesman problem exist. They are based on the construction of a minimal spanning tree, which then is modified to become a valid tour. Both operations can be performed in a time polynomial in relation to the problem size. The best known algorithm has been proposed by Christofides [16] that calculates a tour with a length $C(I)$ holding

$$C(I) \leq \frac{3}{2}\, O(I).$$

There may exist better approximation algorithms, and also there may exist an *approximation scheme*. This is an algorithm which takes an instance I and an error bound $\varepsilon > 0$ as inputs an returns an a solution to the problem with value $A(I, \varepsilon)$ such that

$$\frac{|A(I, \varepsilon) - O(I)|}{O(I)} \leq \varepsilon.$$

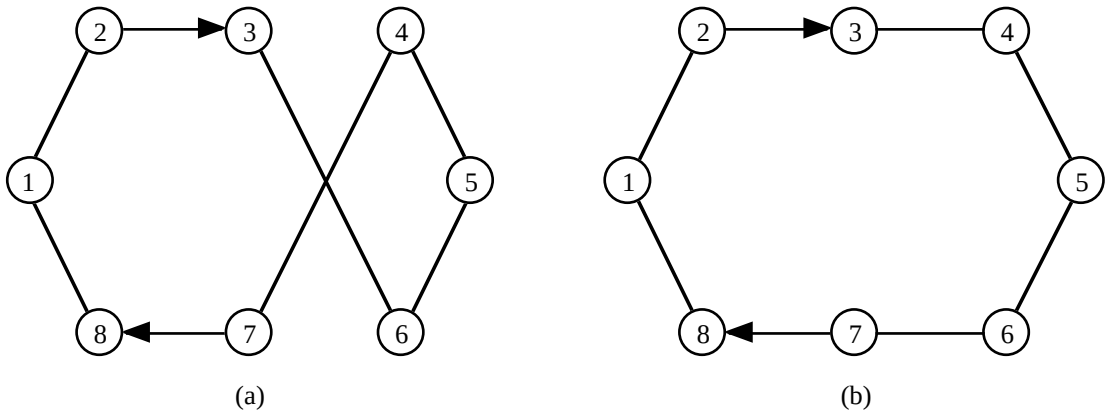


Figure 2.1: 2-change example: (a) current tour, (b) after exchange.

It would be useful to have a *fully polynomial time approximation schema*, i. e. a deterministic algorithm that calculates an approximative solution within a relative bound ε and that is polynomial in both the number of cities n and $1/\varepsilon$. Unfortunately it can be proved that no such scheme can exist for the traveling salesman problem assuming that $\mathcal{P} \neq \mathcal{NP}$.

2.6 Improvement Strategies

Improvement strategies start from an already existing admissible solution and try to improve it by local search. The local search requires a *neighborhood* in which the algorithm searches. If during the search a better admissible solution is found, it is taken as new starting point. The procedure is repeated until no further improvement is possible. The algorithm is more primitive than a hillclimbing algorithm, that searches for the maximal possible improvement before taking the related admissible solution as a new starting point.

In the case of the traveling salesman problem, a neighborhood can be defined to be the set of tours that can be created by exchanging a limited number of edges already in the tour with edges not in the tour. Then the distance between two tours can be defined to be the minimal number of exchange steps required to transform a tour into the other and vice versa. In the simplest case two edges are replaced, which is shown in figure 2.1 for a graph with vertices in a horizontal plane and consequently with an euclidean distance function. The replacement of the edges $e_{3,6}$ and $e_{4,7}$ by $e_{3,4}$ and $e_{6,7}$ shortens the distance of the new tour by eliminating the intersection. As a side effect of this 2-*change* the subtour between vertices 6 and 4 is reversed. This limits the usefulness of the approach to symmetric traveling salesman problems.

If the neighborhood is determined by the replacement of k edges the resulting final tour is called k-*optimal* [57]. Since higher values of k do not increase the quality of the resulting solutions as much as they raise the number of operations, values of $k = 2$ and $k = 3$ are most often used [51, p. 541]. An advanced algorithm by Lin and Kernighan [58] varies the number of exchanged edges k in an an ingenious manner. For many test problems it finds a tour, which can be proved to be optimal by other approaches. But generally a local search can not be guarantied to find a tour whose length is bounded by a constant multiple of

the tour length, provided that the neighborhood is bounded polynomially in relation to the problem size and that $\mathcal{P} \neq \mathcal{NP}$.

2.7 Classification of Approaches

Since there is no hope to find a deterministic polynomial time algorithm besides the proposed approximation and improvement algorithms a lot of other approaches have been tried besides the improvement strategies proposed above. Generally the algorithms to attack intractable problems can be divided into two classes, ones with provable success and ones without.

Branch and cut algorithms belong to the first group. They use polyhedral theory [37] and their running time also for large problems is acceptable although clearly not polynomial. Most of the known best tours of special test problems have been calculated using this approach and it can be applied to many other combinatorial optimization problems.

The approaches of the second group usually are called *heuristics*, although this label is controversial. In the strict sense a heuristic finds the optimal solution of non pathological problem instances, but special instances can be constructed, that deceive it. A typical example is the Lin-Kernighan algorithm mentioned above, which in many cases computes the optimal tour for a symmetric traveling salesman problem. In the wide sense a heuristic is just an algorithm calculating an admissible solution which hopefully may be close to the optimum. In his book Reinelt gives an overview about this class of algorithms for the traveling salesman problem and a comparison of their performance [84, Chapter 3, p. 31*ff*].

2.8 Problem Instances

Although a problem generally belongs to the class of $\mathcal{NP}$-hard problems, there are differences in the solubility of certain instances. E. g. for the traveling salesman problem one can construct an instance with all cities on a circle. Clearly each improvement strategy like the 2-change algorithm will quickly find the optimum tour. This can be rendered more difficult by constructing special instances that require hillclimbing respectively a greedy strategy. I. e. the optimum can only be found if in each improvement step a local search is performed and the highest possible improvement is chosen. Depending on the used algorithm, it is often also possible to construct special instances obstructing the algorithm totally which conforms with the properties considered above. Finally, in experiments proposed to illustrate the power of a new solution strategy often instances created at random are used. Unfortunately these are not very suitable for testing, because they usually are easy to solve for most heuristics in the strict sense.

2.9 Summary

The traveling salesman problem is a typical and clear example of a $\mathcal{NP}$-hard problem. In spite of the results found until now it is not definitively clear why it is so difficult to solve [38, p. 85]. Since it is very improbable to find an efficient deterministic algorithm, many other solution approaches besides the proposed ones have been tried. Furthermore,

there are a lot of general stochastic algorithms like simulated annealing [9, p. 305*ff*], which usually also are labeled heuristics and which can be applied also to the traveling salesman problem. One further approach to attack intractable optimization problems is to use *evolution programs*, which will be presented in the next chapter.

Chapter 3

Evolution Programs

3.1 Motivation

In this chapter a short and mostly informal introduction to evolution programs is given. Evolution programs are inspired by the process of natural evolution and mimic lots of the structure and working found there. Thus first some aspects of natural evolution are considered. Then genetic algorithms are proposed and generalized to evolution programs. Several aspects needed later are further explored. Clearly the consideration lacks the generality of a full size introduction to evolution programs because of this concentration. The reader may refer to the textbooks of Michalewicz [63], Bäck [4], Mitchell [64] and Goldberg [32] for a wider survey although the theoretical parts should be taken with care as will be considered later. Alternatively Beasly, Bull and Martin [5, 6] provide a reasonable short introduction into the subject.

3.2 Natural Evolution

The natural evolution has been the inspiration to develop evolution programs and their related variants. It is a very complicated process, which can be considered on several levels of abstraction. The first observations about changes of species over time in nature have been published by Jean-Baptiste Pierre Antoine de Monet, Chevalier de Lamarck [65]. But wrongly he conjectured that adaptations of individuals are transmitted to the next generation by heredity. Later on Charles Darwin laid a large part of the foundations of the evolutionary theory. He recognized the combination of reproduction, small changes and selection to be the source of the development of the species and the adaption to different environments [20]. A very similar model has been developed independently by Alfred Russel Wallace [98] and been proposed together with Darvin's one. Independently from this research the second branch of the theory of natural evolution has been founded by the monk Johann Gregor Mendel. From his experiments to breed flowers in the garden of his monastery at Brünn he conjectured that the combination of information stored and transmitted in the flowers determines certain of their properties. By deducing the laws of heredity he founded genetics [62]. His consideration has not been taken seriously for a quiet long time and temporarily also been lost. From around 1900 genetics succeeded and the process culminates in the discovery of the dual helix structure of the DNA by

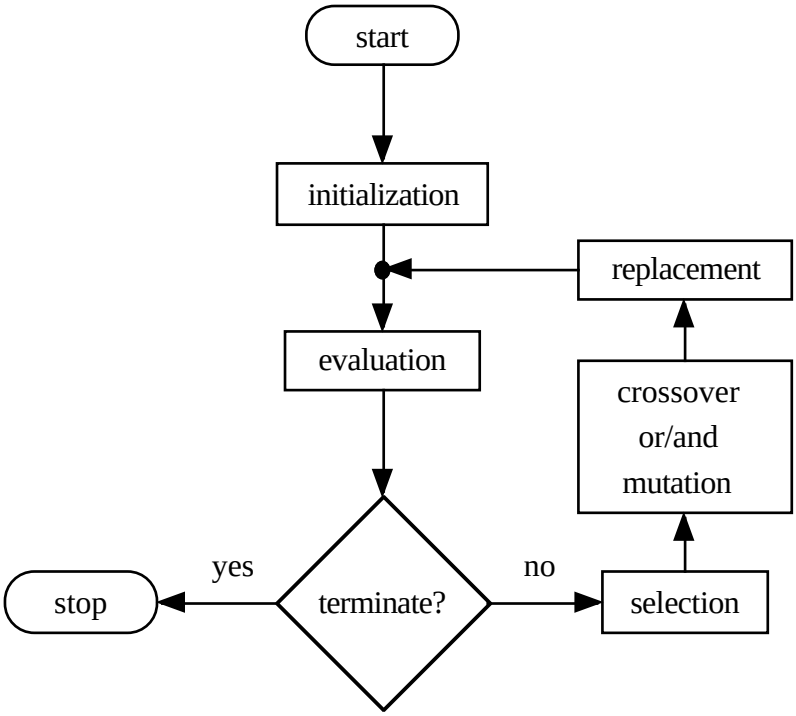


Figure 3.1: Flow chart of the classical genetic algorithm

James D. Watson and Francis H. C. Crick [100]. Both proposed branches of research have been joint to the synthetic evolutionary theory, which has been mainly influenced by Theodosius Dobzhansky [21] and Ernst Mayr [60]. A nice introduction into natural evolution with a detailed presentation of the history of the research is given by Storch, Welsch and Wink [93].

Neglected by most considerations about evolution from the biological point of view is the development of mathematical models for the propagation of genes. This field of *population genetics* has been mainly initiated by the work of Ronald H. Fisher [27] and Sewall Wright [103].

3.3 Genetic Algorithms

Genetic algorithms are one approach to use the elements of the natural evolution to attack hard problems. In their original version introduced by Holland [46], the domain space of the problem corresponds to a fixed length binary string. A candidate solution is encoded in it and the code is called an *individual*. The genetic algorithm then works on a set or *population* of such individuals. In figure 3.1 the process is visualized in a flow chart. First a start set is generated at random or by another heuristic. Then the process enters a loop. The individuals are evaluated to give some measure of their "fitness". Depending on it, individuals are selected randomly from the population for reproduction. Those with a high fitness get a better chance to be selected than those with a low fitness and an individual may be selected multiple times. The selected individuals are randomly modified by the genetic operations crossover or mutation or both. Crossover cuts two individuals at one or

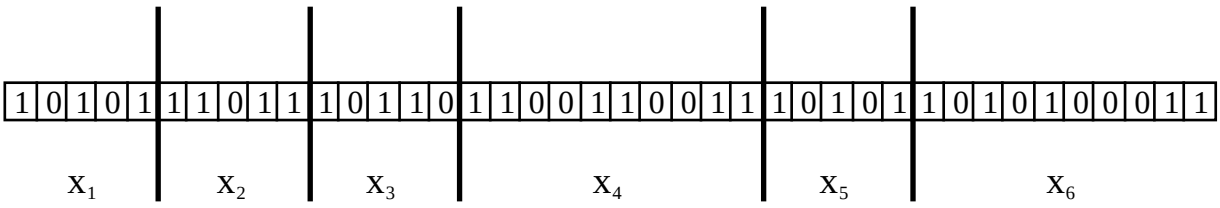


Figure 3.2: Binary encoded individual

more randomly chosen crossings and swaps the corresponding sections whereas mutation randomly changes single bits of an individual. Finally the set of newly created individuals replaces the current population either partly or completely forming a new population. This evolution process recurs until it reaches a stop criterion. During the run the fitness of the individuals should get improved so that both the fitness of the best individual and the average fitness increase. If there is no more improvement, the algorithm may be terminated. Thus the genetic algorithm tries to generate subsequent new and better solution sets by employing genetic operations on the individuals in each turn or generation.

The proposed genetic algorithm can be used to attack combinatorial optimization problems. Then usually each variable x_i corresponds to a substring of an individual. An example is depicted in figure 3.2 for the variables $x_1, \ldots, x_6$. The length of the substrings may vary depending on the domain sizes as shown in the example. The individual's fitness is calculated by a fitness function $f(x_1, x_2, \ldots, x_n)$, that takes the related objective value into account and also transforms a minimization to a maximization problem if required. If the possible solutions have to satisfy additional restrictions also, two approaches are possible. On one hand the fitness function may be modified to penalize violations of the restrictions by reducing the fitness of the individual. This may lead to additional local optima and the generated final solution candidates may not necessarily conform to all restrictions. On the other hand the new individuals may be repaired if necessary to conform to all restrictions before they replace the old individuals. This may lead to a lot of modifications of the solution candidates.

The original genetic algorithm has been extended in many different ways. There are several additional different types of crossover, e. g. two-point and uniform crossover and many selection methods like proportional and tournament selection. Finally the portion of the population being replaced by the newly created individuals can be varied.

Two important disadvantages of the proposed approach should be mentioned here. First the effect of a mutation depends on the position of the related bit in the substring and the used encoding. I. e. changing a single bit in a substring may have a higher effect then changing several bits. Second, a crossover point may also be located inside a substring, which would change the related variable instead of to exchange information. The recognition of these problems and the pursuit to avoid them has motivated some researchers to modify and to improve the original genetic algorithm.

3.4 Evolution Programs

It was Michalewicz [63] who first abstracted and generalized genetic algorithms to evolution programs. He modified the concept of genetic algorithms in several aspects. First the

internal encoding is extended to be any, but problem adequate data structure, consisting of several parts each representing a solution component. Second the genetic operations mutation and crossover are generalized to be problem specific. The unary mutation introduces new elements of the domain space and the binary crossover swaps corresponding solution elements between solution alternatives. Ideally, the encoding and the specific genetic operations of an evolution program respect the restrictions of the problem by generating only admissible solutions. Hence penalty functions and repair algorithms become superfluous.

In most cases, evolution programs outperform ordinary genetic algorithms [63, p. 9]. But each problem requires an especially adapted evolution program with an adequate solution encoding and mutation and crossover operators, which increases the implementation expense. One possible solution is, to use object-oriented programming languages and a design with extensibility and adaptability in mind [29].

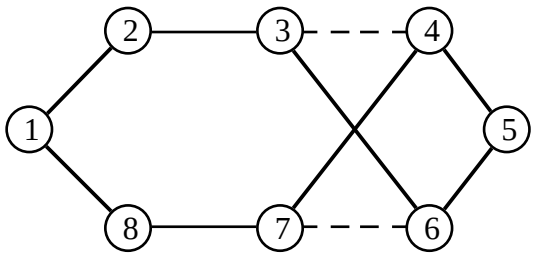
As an example in the next section an evolution program is proposed to attack the symmetric traveling salesman problem.

3.5 An Evolution Program for the TSP

A variety of approaches have been attempted so far to generate near optimal solutions with genetic algorithms or other evolution inspired techniques for the Traveling Salesman Problem also. They vary in the used representations, the genetic operators, whether they use penalty functions, repair algorithms or avoid both, and whether they use additional improvement strategies on newly generated individuals. Michalewicz [63, Chapter 10, p. 211*ff*] proposes the most important approaches and discusses their pros and cons. The program TSPGA uses a different approach aiming to tackle symmetric traveling salesman problems the simplest possible way [30].

The edges visited by a tour provide the important information for a tour description. Each used edge contributes its length to the total tour length. Thus TSPGA also emphasizes the edges used in a specific tour by representing a tour of n cities as an adjacency list $A = [a_1, \ldots, a_n]$ with $a_i = j$, if and only if the tour leads directly from city i to city j. I. e. the index of a list value represents the start city of an edge and the value at this index its destination. A drawback of this approach is, that the representation of a tour on the encoding level depends on its direction. Thus for each tour there are two totally different encoded representations, one for each direction. But both can quickly be transformed into one another by simply exchanging index and value. Nevertheless, for instances up to 1000 cities this is the best known tour representation, whereas for larger problems other representations are more efficient [28].

Unfortunately, a crossover produced by simply swapping the remainder following a randomly chosen crossing leads to illegal tours in this representation. Generally it is also not possible to incorporate a subtour of one admissible solution into another admissible solution without modifying it deeply to maintain the admissibility. Furthermore, for many other problems the simple one-point crossover performs not as well as the so called uniform crossover [94], which swaps corresponding parts of the candidate solutions separately. Thus, the aim is to find a crossover operator to exchange some edges between the individuals and a mutation operator to introduce new edges. Additionally, the resulting tours must



index	1	2	3	4	5	6	7	8
values before 2-change	2	3	6	7	4	5	8	1
values after 2-change	2	3	4	5	6	7	8	1

Figure 3.3: Effect of a 2-change operation on the tour representations

comply with the restrictions for admissible tours. Both operators require an operation that incorporates any given edge into a tour without destroying it.

TSPGA provides this facility by using Lin's 2-change algorithm [57] presented already in section 2.6. If the edge e_{ij} should be incorporated into the tour with the adjacency list $A = [a_1, \ldots, a_n]$, Then the edges e_{i,a_i} and e_{j,a_j} are replaced by the chosen edge e_{ij} and a new edge e_{a_i,a_j} in order to satisfy the tour restrictions. Furthermore, the edges between the vertices a_i to a_j have to be reversed in the adjacency list, because this part of the old tour is visited in reverse in the new tour. As an example the situation in figure 2.1 is reemphasized in figure 3.3. To introduce the new edge $e_{3,4}$ into the tour the edges $e_{3,6}$ and $e_{4,7}$ have to be removed and the second new edge $e_{6,7}$ closes the tour and ensures the admissibility. The enclosed subtour has to be reversed. On the phenotype and the encoding levels this changes the representation as shown in figure 3.3. Using the algorithm above, mutation and crossover are easily derived. For each position $h = 1, \ldots, n$ in the adjacency list the mutation operator checks whether it should change the accompanying edge or not. In this case, it randomly determines a new destination vertex k and incorporates the new edge $e_{h,k}$ into the tour. Similarly, for each position $i = 1, \ldots, n$ in the adjacency list the crossover operator checks whether it should take over a new destination vertex j from the individual to mate with. If so, it incorporates the new edge $e_{i,j}$ into the tour.

As a major advantage both randomized 2-change genetic operators process the essential information, i. e. which edges are used in the tour. Furthermore, in every stage the operators guarantee a tour meeting all restrictions. However, both operations introduce new edges into a tour to preserve its admissibility and the necessary reversion of the subtour between the new edges changes its representation completely. But similar problems are related to every known tour encoding and generally can not be circumvented.

3.6 Notation and Terminology

After proposing genetic algorithms and evolution programs rather fuzzily and by example now the situation and notation is refined and formalized.

In the following an evolution program is assumed to act on a population $\boldsymbol{P}$ of N individuals $I_1, \ldots, I_N$. An individual I_i representing a solution candidate of the related problem has an array of M different attributes $A_{i,1}, \ldots, A_{i,M}$ and each value $A_{i,j}$ is taken from its

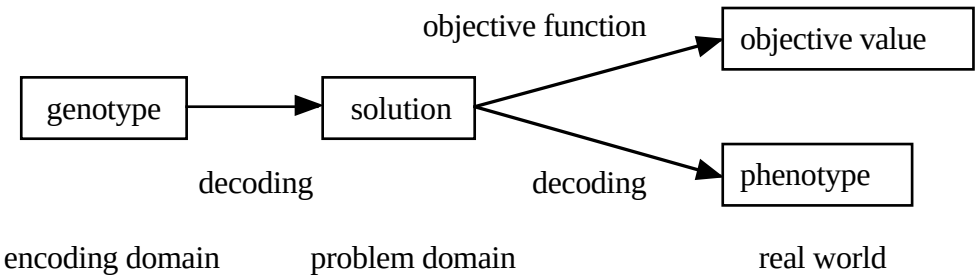


Figure 3.4: The relations between genotype, solution, phenotype and objective value and the different domain levels.

domain $\mathbf{A}_j = \{a_{j,1}, \ldots, a_{j,n_j}\}$. Usually this is called a string type representation. The combination of all attribute domains $\mathbf{A}_1 \times \mathbf{A}_2 \times \ldots \times \mathbf{A}_M = \mathbf{A}$ is called the encoding domain of the evolution program. If the attribute domains are restricted to the set $\{0, 1\}$, then the encoding is called binary, which is characteristic for genuine genetic algorithms. Finally, here each combination of attributes is called a *class*.

Using the decoding function $d : \mathbf{A} \mapsto \mathbf{Z}^n$, an individual is decoded into a solution in the problem domain. In evolution programs the solution usually is required to be admissible and there may be several different solutions, which in fact are equivalent. E. g. a tour belonging to a symmetric traveling salesman problem can start at each node and into both directions, thus it may have a lot of "different" but in fact equivalent representations on the problem and depending on the chosen representation also on the encoding domain. Finally from the solution in the problem domain the objective value belonging to the individual can be calculated and from that its fitness. By the way, choosing a suitable encoding heavily influences the performance of an evolution program. It may also be possible to omit the distinction between genotype level and problem domain. Then no decoding is necessary.

Analogously to the terms used in genetics a substring of the attribute array, which encodes a variable in the problem domain, is called a *gene*, its combination of values *allele* and its position *locus*. The whole attribute array corresponds to the *genotype* in genetics, whereas its partition into separate *chromosomes* in evolutionary computing is used only rarely. In evolution programs usually each attribute in the encoding domain maps to a variable in the problem domain, i. e. an attribute corresponds to a gene and its position j to the locus. Here this is assumed generally for convenience in the notation and terminology, but this does not restrict the generality of the consideration[1]. The real world representation of a solution or of a set of equivalent solutions is called *phenotype* of the individual. In figure 3.4 the relations between the different domain levels considered above and the terminology are presented.

3.7 Schema Propagation Conjecture

Already some approaches have been made to explain why genetic algorithms and evolution programs in many cases produce near optimal solutions. Already Holland [46, chapter 6,

[1] Sometimes also in genetic algorithms a single bit is called a gene and its position locus.

p. 89*ff*] tried to explain the improvement of the objective values respectively fitnesses during an evolution by introducing the notion of a *schema*, which otherwise would be called a *pattern*. In his considerations he took only a binary encoding into account but his argumentation can be transferred analogously to any string based representation like the one proposed above, which is performed in the following.

A schema consists of some fixed attribute values at certain positions of the attribute array whereas the other positions are arbitrary, which is denoted by the $*$ symbol. E. g. a population consisting of the four individuals having the chromosomes

I_1:	1	3	4	3	6	1
I_2:	2	3	8	2	6	4
I_3:	1	3	9	4	6	2
I_4:	5	3	8	1	1	3

may be assumed. Then all chromosomes match the schema $*3****$, the chromosomes of the individuals I_1, I_2 and I_3 the schema $*3**6*$ and the chromosomes of I_1 and I_3 the schema $13**6*$. Formalized a schema $\mathbf{S}$ describes a subspace of the encoding domain $\mathbf{A}$, i. e. $\mathbf{S} \subseteq \mathbf{A}$ holds. Then an individual matches a schema if it is in the related subspace, and the most specialized schema is a certain genotype. In an actual population the individuals belonging to genotypes matching a schema form a subset of the population.

Holland conjectured that the propagation of schemata in a population during the evolution is the main source of the improvement of the average and of the best individual's fitness during evolution. He also presents a rudimentary mathematical model and gives some proofs of an approximation for the schema propagation [46, chapter 6, p. 89*ff*]. Until now there is an ongoing discussion about the significance of his work. Conversely to the attention payed to Holland's conjecture there are only rare attempts to develop indices that describe the state and the reach of schemata in a population. Thus until now there is no empirical evidence for the propagation of schemata in a population during the run of an evolution program or genetic algorithm.

3.8 Consequences and Tasks

Picking up Holland's conjecture of schema propagation three tasks have to be performed to get deeper insight into the working of evolution programs and genetic algorithms.

1. A rigorous but clear theoretical model of the schema propagation process during evolution is needed. Although since Holland's initial work many articles have been published both supporting and doubting his conjecture, most of them are at least incomplete, only approximative, fuzzy or to complex. The model should incorporate also the approaches used already in population genetics. Then its properties have to be explored theoretically and compared with the existing models for evolution programs.

2. Until now there are no suitable indices to monitor the propagation of schemata in a population during a run of an evolution program. The schema reach in a population is one component of the state of the population. Consequently suitable indices have to be found that describe the state of a population and measure the schema reach in

it. In the existing work usually only the average fitness and the fitness of the best individual is recorded. Since the problem is similar to that of identifying patterns in multivariate sets by machine learning, the relations to indices used in that area of research also have to be explored.

3. Experiments have to be performed using the new indices to observe what really happens during the run of an evolution program. Although so many experiments have been performed, there are no observations supporting or refuting Holland's conjecture because of the missing suitable indices. The resulting observations have then to be compared with the properties of the theoretical model. Hopefully then there are no contradictions between the observations and the theoretical forecasts.

In the next chapter the first task is performed by considering a simplified model of evolution as a stochastic process.

Chapter 4

A Theoretical Model of Evolution

4.1 Introduction

As stated above, one component of evolution is selection. In natural evolution this is often referred to by the phrase "survival of the fittest", which has been introduced by Spencer [91] but often is attributed to Darwin. In nature an individual has no fitness property by himself, fitness turns out by survival and reproduction of an individual over the time. Nevertheless selection models have been proposed, assuming individuals having a fitness value. In evolutionary computation, fitness is related to the objective value of the generated solution candidates. To imitate the natural selection, often a stochastic selection process is used similar to that found in simple models of natural evolution. A further component of evolution is mutation. Usually it is applied stochastically to the selected individuals modifying some of their properties. A modified individual then belongs to a different species, which also may change its fitness. The selected and modified individuals then can be used to replace the old ones.

In population genetics mathematical models of the evolution process incorporating both components mentioned above have a long tradition. Fisher [27] introduced a model suitable for large populations, whereas Wright [103] investigated the influence of a small population size. As mentioned already above, in evolutionary computation Holland [46, p. 89*ff*] independently also tried to model the evolution in his algorithms mathematically to support his conjecture of propagation of better schemata as a main source of the success of evolution. Later on, other researchers extended his model and tried to modify its interpretation. But it is still one of the foundations of evolutionary computing and proposed in every introductory book of this subject.

The models used in both disciplines are very similar and can be used to form a unified approach, which then can be attached by standard methods developed for stochastic and deterministic dynamic systems. The results of both approaches have to be compared against each other and with that of existing other models. Finally, relations to the propagation of schemata must be proposed.

In the next section an evolution process is proposed using the variant *fitness proportional random selection* neglecting mutation and crossover or any other manipulation of the individuals, which is a simplification of the process taking place in many evolutionary algorithms.

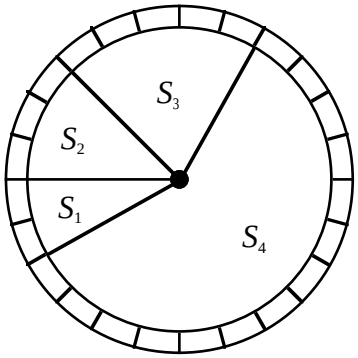


Figure 4.1: Roulette-wheel for selection out of four species. The slots are attached to a species S_j proportional to its relative fitness.

4.2 The General Model of Selection

A population $\boldsymbol{P}$ is a multiset of N individuals[1]. Each individual i belongs to a species S_j of the set $\boldsymbol{S}$ of M different species. Consequently, a species S_j is represented by n_j equivalent individuals in the population and the equation

$$\sum_{j=1}^{M} n_j = N \tag{4.1}$$

holds. Each unique combination of frequencies $n_1^{(l)}, \ldots, n_M^{(l)}$ is called a *state* l of the population. Furthermore each individual i has a fitness $F(i)$ determined by its affiliation to one of the species. If individual i belongs to species S_j, the equation $F(i) = F_j$ holds.

An individual is selected randomly for reproduction from the population in state l according to its fitness in relation to the sumfitness

$$\Phi = \sum_{i=1}^{N} F(i) = \sum_{k=1}^{M} n_k^{(l)} F_k.$$

Then the probability $p_s(S_j, l)$ to select an individual belonging to species S_j equals

$$p_s(S_j, l) = \frac{n_j^{(l)} F_j}{\Phi} = \frac{n_j^{(l)} F_j}{\sum_{k=1}^{M} n_k^{(l)} F_k}. \tag{4.2}$$

Goldberg [32, p. 10*ff*] has visualized this fitness proportional random selection by a roulette-wheel like the one depicted in figure 4.1. The wheel has a number of slots and to a species j belongs a portion corresponding to the sum of the fitness of its members $n_j^{(l)} F(j)$ in relation to the sum of the individuals' fitness. Throwing the ball, it will stop randomly in one of the slots with equal probability and an individual of the related species is selected.

[1] In population genetics usually a population size $2N$ is considered because individuals are assumed to be *diploid*.

In the model of evolution considered here, the selection process from a given population $\boldsymbol{P}(t)$ at time t is repeated N times and the chosen individuals are copied to form the a population $\boldsymbol{P}(t+1)$. In evolutionary computation this is known as generational approach in contrast to the non generational or steady-state approach, where in each generation one individual of the population is replaced by a new one, that is created by evolutionary operators from the current population [88]. In probability theory the generational selection process used here corresponds to sampling with replacement. Consequently, the frequencies $\boldsymbol{n}_1(t+1), \ldots, \boldsymbol{n}_M(t+1)$ at generation $t+1$ are stochastic variables just as the state $\boldsymbol{n}$ is one. The conditional probability to select as new population $\boldsymbol{P}(t+1)$ exactly $m_1^{(k)}, \ldots, m_M^{(k)}$ individuals of the species $S_1, \ldots, S_M$ forming the actual state k depending on the current population $\boldsymbol{P}(t)$ at generation t in actual state l with frequencies $n_1^{(l)}, \ldots, n_M^{(l)}$ follows a multinomial distribution and consequently is given by

$$\begin{aligned} (4.3)\quad & p(\boldsymbol{n}_1(t+1) = m_1^{(k)}, \ldots, \boldsymbol{n}_M(t+1) = m_M^{(k)} | \boldsymbol{n}_1(t) = n_1^{(l)}, \ldots, \boldsymbol{n}_M(t) = n_M^{(l)}) \\ & = \frac{N}{m_1^{(k)}! \cdots m_M^{(k)}!} \, p_s^{m_1^{(k)}}(S_1, l) \cdots p_s^{m_M^{(k)}}(S_M, l). \end{aligned}$$

This conditional probability depends not on the generation number respectively time t. Consequently the evolution is a homogeneous Markov chain and the sequence of populations forms a homogeneous Markov chain.

Following combinatorial theory there are

$$(4.4)\quad v = \binom{N+M-1}{N}$$

different states [69, p. 81], which are too many for a clear analysis. Consequently, in the following, the model is restricted to a set of two different species $\boldsymbol{S} = \{S_1, S_2\}$. There species S_2 corresponds to the species under consideration and species S_1 combines all other individuals. By doing so the evolution can be explored without complicating the situation unnecessarily by considering further species. In evolutionary computation this approach sometimes is called a "coarse grained" model of the evolution process [92]. Then from equation (4.1)

$$(4.5)\quad n_1 + n_2 = N$$

follows. Furthermore, the fitness F_2 of the individuals of species S_2 here is assumed to be related to the fitness F_1 of the individuals of the species S_1 by the equation $F_2 = \alpha F_1$ with the fitness factor $\alpha \geq 0$. Using that and equation (4.5) the equation (4.2) of the selection probability in state l is simplified to

$$(4.6)\quad p_s(S_j, l) = \frac{n_j^{(l)} F_j}{n_1^{(l)} F_1 + n_2^{(l)} F_2} = \frac{n_j^{(l)} F_j}{(n_1^{(l)} + \alpha(N - n_1^{(l)}))F_1}.$$

Since there are only two species, it is sufficient to consider only species S_2. Then the number of individuals $n_2^{(l)}$ belonging to it equals the actual state l of the population, which in the following will be noted by n. The probability $p_s(S_2, l)$ to select an individual belonging to species S_2 then is a function of the actual state $l = n$ only and from equation (4.6) for the selection probability

$$(4.7)\quad p_s(n) = \frac{n\alpha}{N + (\alpha - 1)n}$$

follows.

Analogous to the situation of M species, here the state $\boldsymbol{n}(t+1)$ of the new population $\boldsymbol{P}(t+1)$ is a binomially distributed stochastic variable depending on the actual state n of the old population $\boldsymbol{P}(t)$. With equation (4.7) the conditional transition probability

$$P\{\boldsymbol{n}(t+1) = j | \boldsymbol{n}(t) = i\} = \binom{N}{j} p_s^j(i)(1 - p_s(i))^{N-j} \tag{4.8}$$

$$= \binom{N}{j} \left(\frac{i\alpha}{N + (\alpha - 1)i}\right)^j \left(\frac{N - i}{N + (\alpha - 1)i}\right)^{N-j}. \tag{4.9}$$

follows. Equation (4.8) is the restriction of equation (4.3) to only two species. Consistently to the above formulas two special cases are defined:

$$P\{\boldsymbol{n}(t+1) = 0 | \boldsymbol{n}(t) = 0\} = 1 \quad \text{and} \quad P\{\boldsymbol{n}(t+1) = N | \boldsymbol{n}(t) = N\} = 1. \tag{4.10}$$

Thus, if a population consists of individuals belonging to only one species, then this state will be preserved because there is no mutation in the considered model.

The model proposed until now in fact is the Wright-Fisher model, which is widely used in population genetics [25, p. 16*ff*], restricted here to haploid individuals and with a special emphasis on its Markovian nature. It is also used sometimes in the theory of evolutionary computation, e. g. by Goldberg and Segrest [34] and Chakraborty, Deb and Chakraborty [15], but mostly without referring to its roots. One can also deduce it by restricting the model of Nix and Vose [69] and Vose [97] to consider only two species.

In the next section the probabilistic selection model is approximated to form a deterministic model, the properties of which then are discussed.

4.3 A Deterministic Selection Model

4.3.1 Approximation of the Selection Process

To characterize the evolution of the proposed model the development of the expected values respectively central moments is of special interest. To remember, depending on the current state $n(t)$, which is an abbreviation of $\boldsymbol{n}(t) = n$, the new state $\boldsymbol{n}(t+1)$ is a binomially distributed random variable with parameter N and selection probability $p_s(n(t))$. Then for the conditional expected value of the state

$$\mathrm{E}\{\boldsymbol{n}(t+1) | n(t)\} = \eta_{\boldsymbol{n}}(t+1)|_{n(t)} = p_s(n(t))\, N \tag{4.11}$$

and for the conditional variance of the state

$$\sigma^2_{\boldsymbol{n}}(t+1)|_{n(t)} = p_s(n(t))(1 - p_s(n(t)))\, N \tag{4.12}$$

hold [9, p. 20], where the last index measures the sampling error of the repeated stochastic selection. Using $p_s(n(t))$ from equation (4.7)

$$\mathrm{E}\{\boldsymbol{n}(t+1) | n(t)\} = \frac{n(t)\alpha N}{N + (\alpha - 1)n(t)} \tag{4.13}$$

and

$$\sigma^2_{\boldsymbol{n}}(t+1)|_{n(t)} = \frac{n(t)\alpha N(N-n(t))}{(N+(\alpha-1)n(t))^2} \tag{4.14}$$

follow. Introducing the portion of individuals of species S_2 in the population

$$r(t) = \frac{n(t)}{N} \tag{4.15}$$

from equation (4.11) for the conditional expected value of the portion

$$\mathrm{E}\{\boldsymbol{r}(t+1)|r(t)\} = \eta_{\boldsymbol{r}}(t+1)|_{r(t)} = p_s(r(t)) = \frac{\alpha r(t)}{1+(\alpha-1)r(t)} \tag{4.16}$$

follows and from equation (4.12) for the conditional variance of the portion

$$\sigma^2_{\boldsymbol{r}}(t+1)|_{r(t)} = \frac{1}{N}\, p_s(r(t))(1-p_s(r(t))). \tag{4.17}$$

Clearly $\lim_{N\to\infty} \sigma^2_{\boldsymbol{r}}(t+1)|_{r(t)} = 0$ holds independently from the selection function $p_s(r(t))$. Thus values of $r(t+1)$ different from $\mathrm{E}\{\boldsymbol{r}(t+1)|r(t)\}$ are very unlikely for sufficiently large populations and the portion $r(t+1)$ in the generation $t+1$ can be approximated by its conditional expected value, i. e.

$$r(t+1) \approx \mathrm{E}\{\boldsymbol{r}(t+1)|r(t)\}. \tag{4.18}$$

Then the errors caused by the probabilistic sampling are neglected and the portion in the new generation is determined by the selection probability, which picks up Fisher's [27] point of view. Finally, the domain of the state is transformed from a discrete to a continuous one. This introduces further errors, which only for large populations can be neglected. Often the use of this approximation is not stated explicitly but paraphrased by assuming an infinite population size. From the mathematical point of view this assumption is problematic, since then not only sampling errors are neglected but also other implicit assumptions about the population are no longer valid.

Using the above approximation from equation (4.16) the first order difference equation

$$r(t+1) = g(r(t)) = \frac{\alpha r(t)}{r(t)(\alpha-1)+1} \tag{4.19}$$

follows. It is well known as logistic Pielou difference equation [75, p. 22] and can be solved exactly, which is proposed in the next section. The deterministic selection easily can be extended for more than two species [13, p. 29*f*], which equals the approximation of the original multi species model of the previous section.

4.3.2 Properties of the Approximating Recursion Equation

By setting $r(t+1) = r(t) = z$ the fixed points of the recursive equation (4.19) can be found. Doing so

$$z(z-1)(1-\alpha) = 0 \tag{4.20}$$

follows. If $\alpha = 1$ holds, it is valid for all $0 \leq z \leq 1$, i. e. each z is a fixed point. If $\alpha \neq 1$ holds, then equation (4.20) is valid only for $z = 0$ and $z = 1$.

A fixed point z of a first order recursion equation

$$x(t+1) = g(x(t))$$

is asymptotically stable, if

$$\left|\frac{dg}{dx}(z)\right| = |g'(z)| < 1 \tag{4.21}$$

holds [45, p. 6*f*].

Transferred to equation (4.19) the inequality

$$\left|\frac{dg}{dr}(z)\right| = \left|\frac{d}{dr}\frac{\alpha r}{r(\alpha-1)+1}\right|_{r=z} = \left|\frac{\alpha}{(z(\alpha-1)+1)^2}\right| < 1$$

has to hold for a fixed point z to be asymptotically stable. For $0 \leq \alpha < 1$ this is true only for the fixed point $z = 0$ and for $\alpha > 1$ only for the fixed point $z = 1$. For $\alpha = 1$

$$\left|\frac{dg}{dr}(r)\right| = \left|\frac{\alpha}{(r(\alpha-1)+1)^2}\right| = 1$$

follows and all higher derivatives are zero. Consequently each z is stable, because a single small disturbance of the system remains and increases not over the time.

But for $\alpha = 1$ a small variation of the parameter α changes the behavior of the model totally depending on its direction. Thus this situation is called *structural unstable*.

The recursion equation (4.19) can be generalized to

$$r(t+1) = \frac{ar(t)}{cr(t)+d} \tag{4.22}$$

and easily solved by repeated insertion. For the transition from generation $t+1$ to generation $t+2$ the equation

$$r(t+2) = \frac{ar(t+1)}{cr(t+1)+d}$$

follows. Inserting there equation (4.22) gives

$$r(t+2) = \frac{a^2 r(t)}{c(a+d)r(t)+d^2}$$

and analogously

$$r(t+3) = \frac{a^3 r(t)}{c(a^2+ad+d^2)r(t)+d^3}.$$

Repeating this insertion, for $r(t+w)$ the equation

$$r(t+w) = \frac{a^w r(t)}{c(a^{w-1}+a^{w-2}d+\ldots+d^{w-1})r(t)+d^w} = \frac{a^w r(t)}{c\sum_{j=0}^{w-1} a^j d^{w-1-j} r(t)+d^w}$$

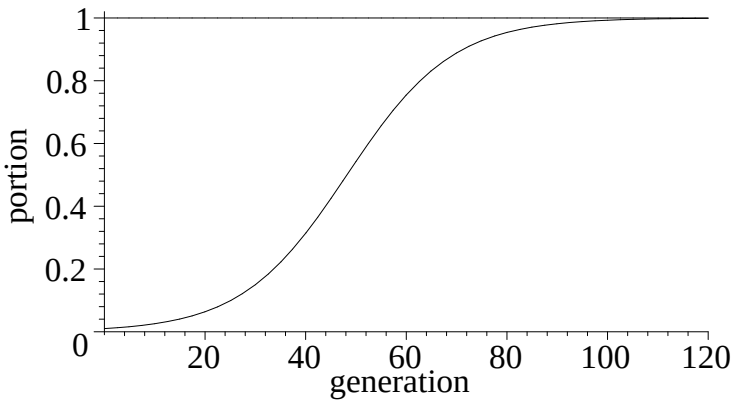


Figure 4.2: Development of the portion $r(t)$ with increasing generation number for $\alpha = 1.1$ and $r(0) = 0.01$ of equation (4.24).

follows. Setting $t = 0$ and using the sum formula (A.2) and some transformations the solution

$$r(t) = \frac{r(0)a^t(a-d)}{r(0)c(a^t - d^t) + d^t(a-d)} \tag{4.23}$$

is obtained and with resubstitution of the parameters the solution

$$r(t) = \frac{\alpha^t r(0)}{1 + r(0)(\alpha^t - 1)} \tag{4.24}$$

of equation (4.19). It includes the trivial solutions $r(t) = r(0) = 0$ and $r(t) = r(0) = 1$, which always are fixed points. For a start portion $0 < r(0) < 1$ the limit of function (4.24) is

$$\lim_{t\to\infty} \frac{\alpha^t r(0)}{1 + r(0)(\alpha^t - 1)} = \begin{cases} 1 & : \quad \alpha > 1 \\ r(0) & : \quad \alpha = 1 \\ 0 & : \quad 0 \leq \alpha < 1 \end{cases}$$

depending on the sign of the fitness factor α.

Thus corresponding with the fixed points and their stability the development of the portion depends on both the fitness factor α and the start portion $r(0)$. For $\alpha = 1$ the start portion persists resulting in $r(t) = r(0)$ and one orbit. If $\alpha \neq 1$ holds, then there are two orbits depending on the start portion. One orbit is yielded, if the start portion is the unstable fixed point, which persists, the second one, which tends to the asymptotic stable fixed point, from all other start portions.

In contrast to the general solution method proposed here Goldberg and Deb [33] solve equation (4.19) with a trick applicable only in this special case leading to the same result. Figure 4.2 shows its plot for a fitness factor $\alpha = 1.1$ and a start portion $r(0) = 0.01$.

The function (4.24) is continuous and differentiable. Differentiating it, the logistic Verhulst-Pearl differential equation [75, p. 20]

$$\dot{r} = r(1-r)\ln\alpha \tag{4.25}$$

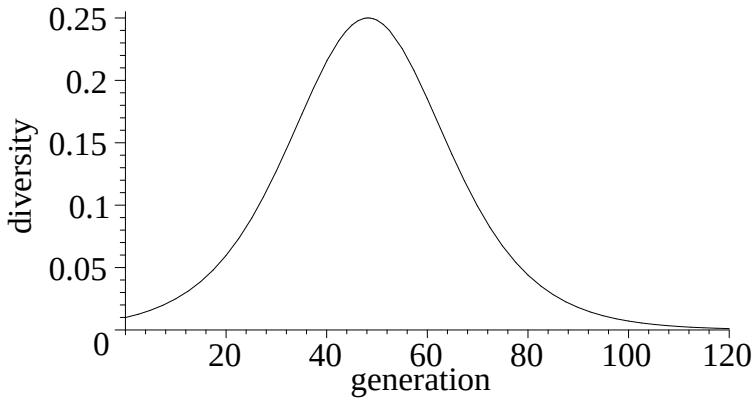


Figure 4.3: Development of the diversity D_s^H of equation (4.28) with increasing generation number for $\alpha = 1.1$ and $r(0) = 0.01$.

is obtained. Differentiating it again,

$$\ddot{r} = (1 - 2r) \ln \alpha$$

follows. For a turning point of equation (4.24) $\ddot{r} = 0$ has to hold, which is true only for $r_s = 1/2$. Setting $t = t_s$ and $r(t_s) = r_s = 1/2$ in equation (4.24) and applying some transformations leads to

$$t_s = \frac{ln\frac{1-r(0)}{r(0)}}{ln\alpha}.$$

Consequently $\dot{r}$ is maximum for $t = t_s$ and axial symmetrical to the ordinate at t_s. Then $r(t)$ must be point symmetrical to $(t_s, 0.5)$.

The derivative $\dot{r}$ on the left side of equation (4.25) is called the *selection pressure*. It is determined by the two components on the right side. One component is the logarithm of the fitness factor α. Thus increasing α increases the selection pressure only less than proportional. The second component $r(1 - r)$ plays an important role in population ecology, because

$$r(1-r) = \frac{1}{2}[1 - r^2 - (1-r)^2] = \frac{1}{2}\left(1 - \sum_{j=1}^{M} r_j\right) = \frac{1}{2} D_s^{GS}(r) \tag{4.26}$$

holds with $M = 2$ here. But $D_s^{GS}(r)$ is the Gini-Simpson-index of species diversity [81], which is widely used in population ecology to measure the heterogeneity of a population of an ecological system. It denotes the probability that two individuals drawn independently with replacement from the population belong to different species. Here the rate of change is maximum, if the heterogeneity of the population is also maximum, which is an analogy to the second part of Fisher's fundamental theorem [25, p. 14]. In population genetics this index is also called *heterozygosity*.

Depending on the portion r furthermore the current average fitness $\bar{f}(r)$ and the variance $s_f^2(r)$ of the fitness in the population can be calculated using statistical standard formulas. Between these three descriptive indices generally the relation

$$1 - D_s^{GS}(r) = \frac{1}{N}\left(\frac{s^2(r)}{\bar{f}^2(r)} + 1\right) \tag{4.27}$$

holds [12, p. 192]. In the next chapter indices based on respectively describing the state will be proposed in depth. For convenience in the following the species diversity index

$$D_s^H = r(1-r) = \frac{1}{2} D_s^{GS}$$

is used to indicate the heterogeneity of a population[2]. Then with equation 4.24 here for the diversity

$$D_s^H(t) = \frac{\alpha^t r(0)(1-r(0))}{(1+r(0)(\alpha^t-1))^2} \tag{4.28}$$

follows. In figure 4.3 its development over the generation is shown with identical initial values as for figure 4.2.

For small portions $r(t)$ equation (4.24) can be approximated asymptotically. To do so, the portion $r(0)$ is substituted by the parameter ε, from which the function

$$f(\varepsilon) = \frac{\alpha^t \varepsilon}{1+\varepsilon(\alpha^t-1)}$$

follows. Then for small t the function $\phi(\varepsilon) = \varepsilon\alpha^t$ is an asymptotic approximation of $f(\varepsilon)$ near $\varepsilon_0 = 0$. To show this, from condition (2.2) the equation

$$\lim_{\varepsilon \downarrow \varepsilon_0} \frac{1}{1+\varepsilon(\alpha^t-1)} = 1$$

has to hold, which is satisfied just for $\varepsilon_0 = 0$ and small t. Thus for small t and $r(0)$ equation (4.24) is asymptotically approximated by

$$\phi(t) = r(0)\alpha^t. \tag{4.29}$$

Although the portion $r(t)$ grows exponentially for small values, the real increase is only modest. Differentiating the function $\phi(t)$ in equation (4.29) according to t

$$\dot{r} = \frac{dr}{dt} \sim \alpha^t r(0) \ln \alpha \sim r(t) \ln \alpha$$

follows. For $0 < r(t) \ll 1$ and $1 < \alpha \ll e$ then the absolute increase per generation step of the portion $r(t)$ is very small. This can also be seen in the graph plotted in figure 4.2. If a new species is introduced in a population with a small start portion, it lasts many generations until it spreads. Consequently, in a real evolution a new individual has not only to survive this stage but also not to be mutated or crossed over. Thus exponential increase in this situation may be not sufficient.

In the next section the corresponding differential equation is proposed and its properties are considered.

[2]The status indices based on the current portion r introduced here should not be confused with the indices related to the state $\boldsymbol{n}$ introduced earlier! Here the indices are calculated from the deterministic portion and thus are descriptive whereas that of the state characterize some properties of the related stochastic variable.

4.3.3 Approximating Differential Equation and its Properties

Since differential equations are more wide-spread then difference equations and there are a lot more techniques to tackle them, many researchers usually try to transform discrete time into continuous time models, e. g. like in the article of Wright and Rowe [101]. Inverting the Euler approach [45, p. 25] from a first order difference a corresponding differential equation can be established. With the constant step size h, usually set to $h = 1$, between the two moments t and $t + 1$ then

$$r(t+1) = r(t) + h\,\dot{r}(t) \tag{4.30}$$

is obtained. Inserting equation (4.19) and setting $r(t) = r$ the first order differential equation

$$\dot{r} = \frac{1}{h}\frac{r(1-r)(\alpha-1)}{r(\alpha-1)+1} = f(r) \tag{4.31}$$

follows. The whole approach is only valid, if it can be shown, that the corresponding differential approximates the difference equation, which may be very difficult. The condition is true here, because the differential equation (4.31) approximates equation (4.25) asymptotically for $\alpha \approx 1$, which can be shown by using again the condition (2.2). Applying it results in

$$\lim_{\alpha\downarrow 1} \frac{r(1-r)\ln\alpha}{\frac{1}{h}\frac{r(1-r)(\alpha-1)}{r(\alpha-1)+1}} = \lim_{\alpha\downarrow 1} h\frac{\ln\alpha}{\alpha-1}.$$

Since counter and denominator both approach zero, L'Hospital's rule [47, p. 4] can be applied. Doing so one gets

$$\lim_{\alpha\downarrow 1} h\frac{1}{\alpha} = h.$$

Consequently setting $h = 1$ yields the desired result.

The fixed points of equation (4.31) can be found by setting $\dot{r} = 0$. Then analogously to the discrete model in the last section also equation (4.20) follows with the same fixed points. To consider their stability the differential equation (4.31) has to be linearized around them.

According to Hartman's and Grobman's theorem [40, p. 244] a differential equation

$$\dot{x} = f(x) \tag{4.32}$$

at a point $x = w$ behaves locally like its linearized form

$$\dot{y} = \frac{df}{dx}(w)\,y$$

at $y = 0$, if for the first derivative

$$f'(w) = \frac{df}{dx}(w) \neq 0$$

holds. Inferring from the properties of the linearized system the fixed point z is asymptotically stable if $f'(z) < 0$ holds.

Consequently for the differential equation (4.31), at an fixed point $r = z$ the inequality

$$\frac{df}{dr}(z) = f'(z) < 0$$

with

$$f'(r) = \frac{1}{h}\frac{d}{dr}\frac{r(t)(1-r(t))(\alpha-1)}{r(t)(\alpha-1)+1} \tag{4.33}$$

has to hold for asymptotic stability. With some transformations

$$f'(r) = \frac{1}{h}\frac{\alpha-1}{(r(\alpha-1)+1)^2}[1-2r-(\alpha-1)r^2] \tag{4.34}$$

follows. For $r = 0$ and $\alpha \neq 1$ then

$$f'(0) = \frac{\alpha-1}{h} < 0. \tag{4.35}$$

has to hold. This is true for $0 \leq \alpha < 1$ and $h > 0$ as well as for $\alpha > 1$ and $h < 0$. For $r = 1$ and $\alpha \neq 1$

$$f'(1) = -\frac{\alpha-1}{h\alpha} < 0. \tag{4.36}$$

has to hold. This is true for $a > 1$ and $h > 0$ as well as for $0 \leq a < 1$ and $h < 0$. The case $h = 0$ generally is of no interest.

For $\alpha = 1$ the equality $f'(r) = 0$ holds for all r. Consequently Hartman's and Grobman's theorem can not be applied. But then all higher derivatives also equal zero and $\dot{r} = f(r) = 0$ holds for all r. A small singular perturbation then remains and increases not over the time and therefore the system is stable. Conversely the system is structural unstable in this situation, because a small perturbation of the fitness factor α changes the stability of the fixed points totally.

The differential equation (4.31) can be solved by separation of the variables. The separation leads to the equation

$$\frac{\alpha-1}{h}\,dt = \frac{1+(\alpha-1)r}{r(1-r)}\,dr,$$

which can be integrated on the left side from t_1 to t_2 and on the right side from $r(t_1)$ to $r(t_2)$. Doing so

$$\frac{\alpha-1}{h}(t_2-t_1) = \int_{r(t_1)}^{r(t_2)} (\alpha-1)\frac{1}{1-r} + \frac{1}{(1-r)r}\,dr$$

follows and using Bronstein's formula 32 [11, p. 36]

$$\frac{\alpha-1}{h}(t_2-t_1) = -(\alpha-1)\ln(1-r)\Big|_{r(t_1)}^{r(t_2)} + \ln\frac{r}{1-r}\Big|_{r(t_1)}^{r(t_2)}.$$

Setting $t_1 = 0$, $t_2 = t$, $r_1 = r(0)$ and $r_2 = r(t)$ together with some simplifications

$$\frac{\alpha - 1}{h}\, t = \ln \frac{r(t)}{r(0)} - \alpha \ln \frac{r(t) - 1}{r(0) - 1}$$

results and finally with some further transformations

$$\gamma(r,t) = e^{-t\frac{\alpha-1}{h}}(1 - r(0))^{\alpha} r(t) - (1 - r(t))^{\alpha} r(0) = 0. \tag{4.37}$$

This implicit equation has the trivial solutions $r(t) = r(0)$ for $\alpha = 1$ and $r(t) = r(0) = 0$ and $r(t) = r(0) = 1$ for $\alpha \neq 1$, which are obviously the fixed points already calculated. A general solution could not be obtained explicitly, but there is exactly one, which is proved in the following.

The implicit equation (4.37) is continuous in the interval $0 < r < 1$. Differentiating with respect to r

$$\frac{d\gamma(r,t)}{dr} = \gamma'(r,t) = e^{-t\frac{\alpha-1}{h}}(1 - r(0))^{\alpha} + \alpha(1 - r(t))^{\alpha-1} r(0)$$

follows. This derivative is also a continuous function and for $0 < r(0), r(t) < 1$ the inequality $\gamma'(r,t) > 0$ holds. Furthermore for $t = 0$ equation (4.37) has the start solution $r(t) = r(0)$. Then according to the implicit function theorem [11, p. 282] there is exactly one function $r = \delta(t)$, that is defined in a neighborhood $U(0)$ around $t = 0$ with

$$\delta(0) = r(0) \quad , \qquad \forall t \in U(0): \quad \gamma(\delta(t), t) = 0.$$

Thus there is only one numerical solution $r(t)$. Since

$$\gamma(0,t) = -r(0) < 0 \quad \text{and} \quad \gamma(1,t) = e^{-t\frac{\alpha-1}{h}}\,(1 - r(0))^{\alpha} > 0$$

hold for finite t and $r(t)$ is continuous, the solution has to be in the interval $0 < r(t) < 1$, which follows from the intermediate value theorem for continuous functions [11, p. 258]. The numerical value can be found easily using Newton's approach (e. g. with *Maple*) because of the strict monotony of the function $\gamma(r,t)$.

The derivative in equation (4.31) is not axial symmetrical to a particular r because the denominator is a strictly increasing function of the portion r. Hence the solution of the implicit equation (4.37) can not be symmetrical also.

The limit values for $t \to \infty$ and $0 < r(0) < 1$ depend on

$$\lim_{t\to\infty} e^{-t\frac{\alpha-1}{h}} = \begin{cases} 0 & : \quad \alpha > 1, h > 0 \quad \text{or} \quad 0 \leq \alpha < 1, h < 0 \\ \infty & : \quad \alpha > 1, h < 0 \quad \text{or} \quad 0 \leq \alpha < 1, h > 0 \end{cases}.$$

In the first case from equation (4.37) follows immediately

$$\lim_{t\to\infty} r(t) = 1.$$

For the second case equation (4.37) has to be transformed to

$$\frac{(1 - r(0))^{\alpha}}{r(0)} \frac{r}{(1 - r)^{\alpha}} = e^{t\frac{\alpha-1}{h}}.$$

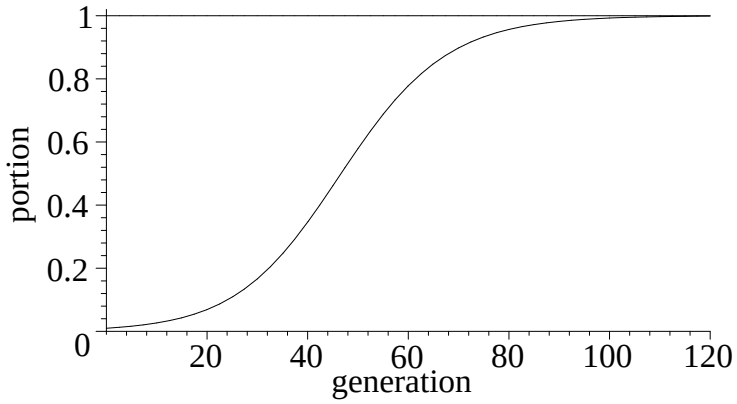


Figure 4.4: Development of the portion $r(t)$ with increasing generation number for $\alpha = 1.1$ and $r(0) = 0.01$ of the solution of equation (4.37).

For $t \to \infty$ the right side also approaches 0 and thus

$$\lim_{t\to\infty} r(t) = 0$$

holds. Finally, $\alpha = 1$ implies $r(t) = r(0)$. The case $h = 0$ is of no interest.

Consequently for $\alpha = 1$ there is one constant orbit and for $\alpha \neq 1$ there are two orbits. If the process starts in an unstable fixed point, this state persist. From all other states the portion converges to the asymptotic stable fixed point. Figure 4.4 shows the development of the portion $r(t)$ with increasing generation number for $\alpha = 1.1$ and $r(0) = 0.01$ of the solution of equation (4.37). The plot shows a high similarity to that of figure 4.2, which conforms to the approximation stated at the beginning of this section.

4.3.4 Discrete versus Continuous Approximation

In both models the individuals with inferior fitness extinct and only the individuals of the species with the highest fitness survive and spread. Randomly inserted superior individuals intersperse their species to take over the population, because of the instability of the fixed point referring to the existence of only inferior individuals in the population. If there are two species of equal fitness, then their portions remain constant but perturbations prevail because then each state is stable. Thus repeated perturbations in one direction may lead to the extinction of one species. This phenomenon is called *genetic drift* and will be discussed in more detail in section 4.4.3. Finally in both models the system is structurally unstable in this configuration. In appendix A.2 the relation of corresponding difference and differential equations in general is explored further.

Interestingly, in the considered model the difference equation (4.19) has an explicit solution but the corresponding differential equation (4.31) has not. Normally, the reverse is true and the differential equation is used, because there is a lot more knowledge how to solve a differential equation explicitly than a difference equation. Furthermore the proof of the applicability of the approximation uses the solution of the difference equation. Since that usually is not known, it is generally difficult to show the applicability of the reversed Euler approach. Consequently, for proportional generational selection the continuous ap-

proximation of the process should be abandoned altogether, which is a very important result.

4.4 Evolution as Markov-Chain

4.4.1 General Notes

Until now simplified models of the selection process have been discussed neglecting the variance of the state caused by sampling. In the following this simplification is dropped and the Markov chain proposed in section 4.2 is considered. A lot is already well known [25, p. 75*ff*], but needed later for a comparison with the deterministic models. Thus it is only stated without proof.

As considered above, the state $\boldsymbol{n}(t+1)$ in the next generation $t+1$ is a stochastic variable depending only on the current state $\boldsymbol{n}(t) = n$ and some fixed parameters, as has been proposed already above. But from the distribution of the states in the current generation t the distribution of the states in the next generation $t+1$ and from that additional indices like expected value and variance of the state can be calculated. With $p_{i,t} = p(\boldsymbol{n}(t) = i)$ and the transition probabilities $\pi_{i,j} = P\{\boldsymbol{n}(t+1) = j | \boldsymbol{n}(t) = i\}$ from equation (4.8) respectively from the definitions (4.10) the equation

$$p_{j,t+1} = \sum_{i=0}^{N} \pi_{i,j}\, p_{i,t} \tag{4.38}$$

follows using the total probability theorem [71, p. 30].

The notation can be simplified using matrices and vectors. The column vectors $P(t)$ consists of the state probabilities $p_{i,t}$ in generation t and the transition matrix Π of the state transition probabilities $\pi_{i,j}$. Then from equation (4.38)

$$P^T(t+1) = P^T(t)\,\Pi \tag{4.39}$$

follows and from the Markov property for the generation $t+k$

$$P^T(t+k) = P^T(t)\,\Pi^k. \tag{4.40}$$

To maintain consistency, $\Pi^0 = I$ is defined with the identity matrix I. The column vectors $V^T = (0, \ldots, k, \ldots, N)$, $W^T = (0, \ldots, k^2, \ldots, N^2)$ and $U^T = (0, \ldots, N(N-k)/N^2, \ldots, 0)$ are defined to calculate the expected value and the variance of the state and the expected value of the diversity. Then from the definitions the formulas

$$\eta_{\boldsymbol{n}}(t) = V^T P(t), \quad \sigma^2_{\boldsymbol{n}}(t) = W^T P(t) - \eta^2_{\boldsymbol{n}}(t) \quad \text{and} \quad D^H(t) = \frac{1}{2} D^{GS} = U^T P(t) \tag{4.41}$$

follow. Using equation (4.40) and the above formulas one can calculate the probability distribution, the expected value and the variance of the state and the expected value of the diversity at each generation k starting from a given state distribution $P(0)$ at generation $t = 0$.

If the population consists of only one species, which corresponds to the states $n = 0$ or $n = N$, then in the considered model this state will remain forever because of the definitions (4.10). Thus an extinction of a species is irreversible and consequently both states

are called absorbing. Since they are accessible from all the remaining non absorbing states, following the theory of Markov chains [9, p. 149*ff*] for a finite number of states $N+1$ the process will end up in one of them and stay there. To them the two terminal distributions $\tilde{P}_1^T = (1, 0, \ldots, 0)$ and $\tilde{P}_2^T = (0, \ldots, 0, 1)$ correspond, which are eigenvectors from the left of the eigenvalue $\lambda = 1$ because they satisfy the equation

$$P^T = P^T \,\Pi. \tag{4.42}$$

Clearly the above equation also holds for each linear combination $\tilde{P} = \beta\tilde{P}_1 + (1-\beta)\tilde{P}_2$ with $0 \leq \beta \leq 1$, which results in the terminal distribution $\tilde{P}^T = (\beta, 0, \ldots, 0, 1-\beta)$. Thus β denotes the probability to be absorbed in state $n = 0$ and is a function of the initial distribution $P(0)$. It can be calculated using the *first step analysis* [9, p. 65*ff*] just as the expected number of generations until absorption. For the terminal expected value of the state then follows

$$\tilde{\eta}_{\boldsymbol{n}} = N(1-\beta) \tag{4.43}$$

and for the terminal variance of the state

$$\tilde{\sigma}^2_{\boldsymbol{n}} = N^2(1-\beta) - N^2(1-\beta)^2 = N^2\beta(1-\beta) = \tilde{\eta}_{\boldsymbol{n}}(N - \tilde{\eta}_{\boldsymbol{n}}). \tag{4.44}$$

4.4.2 First Step Analysis

Defining $s_i = P\{\boldsymbol{n}(t) = 0 | \boldsymbol{n}(0) = i\}$, i. e. the probability to end up in state 0 after starting in state i, from the total probability theorem [71, p. 30] for each state i the equation

$$s_i = \sum_{j=0}^{N} \pi_{i,j} s_j \tag{4.45}$$

holds, since from state i the process switches to state j with probability $\pi_{i,j}$ and then ends up in state 0 with probability s_j. Finally for the absorbing states also

$$s_0 = 1 \quad \text{and} \quad s_N = 0,$$

hold. Defining the vector S consisting of the s_i the probability β to end up in state 0 starting from the initial distribution $P(0)$ is given by

$$\beta = S^T P(0). \tag{4.46}$$

Analogously the average time until absorption τ_i after starting in state i can be calculated. If the process currently is in state i, it transfers to state j in one time unit and then needs on an average additional τ_j time units until absorption. Using again the total probability theorem [71, p. 30] for the states $1 \leq i \leq N-1$ the formula

$$\tau_i = 1 + \sum_{j=1}^{N-1} \pi_{i,j} \tau_j \tag{4.47}$$

with the absorption conditions

$$\tau_0 = 0 \quad \text{and} \quad \tau_N = 0$$

follows. Defining the vector T consisting of the τ_i the equation

$$\tau = T^T P(0).$$

gives the average time τ until absorption in state 0 or N starting from the initial distribution $P(0)$[3].

Alternatively the probabilities of absorption and the average time to absorption can be calculated via the fundamental matrix [34]. Unfortunately in the general case only a numerical evaluation of the formulas proposed above is possible. For $\alpha = 1$ at least β can be calculated based on the development of the expected value, which is considered in the next section. Ewens [25, p. 75*f*] gives an approximation of equation (4.47) for $\alpha = 1$ and fixed start portion $r(0)$, which equals the mean absorption time got from the diffusion approximation of the process.

4.4.3 Development of the Moments

Until now only limits for certain indices of the considered Markov chain have been calculated. Now the aim is to find recursive equations for both the expected value and the variance of the state. The following consideration extends that of Hofbauer and Sigmund [44, p. 26*ff*] to cover also the case $\alpha \neq 1$.

The expected value $\eta_{\boldsymbol{n}}(t+1)$ of the state $\boldsymbol{n}(t+1)$ in the generation $t+1$ is defined to be

$$\eta_{\boldsymbol{n}}(t+1) = \mathrm{E}\{\boldsymbol{n}(t+1)\} = \sum_{i=0}^{N} i\, p_{i,t+1}.$$

Using equation (4.38)

$$\eta_{\boldsymbol{n}}(t+1) = \sum_{i=0}^{N} i \sum_{j=0}^{N} \pi_{j,i} p_{j,t}$$

follows and changing the sum order

$$\eta_{\boldsymbol{n}}(t+1) = \sum_{j=0}^{N} p_{j,t} \sum_{i=0}^{N} i\pi_{j,i} = \sum_{j=0}^{N} p_{j,t}\, \mathrm{E}\{\boldsymbol{n}(t+1)|\boldsymbol{n}(t) = j\}. \tag{4.48}$$

The conditional expected value $\mathrm{E}\{\boldsymbol{n}(t+1)|\boldsymbol{n}(t) = j\}$ already has been calculated in equation (4.13). Using it the equation

$$\eta_{\boldsymbol{n}}(t+1) = \sum_{j=0}^{N} j p_{j,t} \frac{N\alpha}{N + (\alpha - 1)j} \tag{4.49}$$

is obtained.

For the special case $\alpha = 1$ the above formula can be simplified to

$$\eta_{\boldsymbol{n}}(t+1) = \sum_{j=0}^{N} j p_{j,t} = \eta_{\boldsymbol{n}}(t), \tag{4.50}$$

[3] Ewens [25, p. 71] additionally gives a formula for the variance of the time to absorption.

i. e. the expected value of the state is constant, which also holds for the expected portion $\eta_{\boldsymbol{r}}(t)$. In the theory of stochastic processes such a process is called a *martingale* [9, p. 179]. Finally from the equations (4.50) and (4.43) for the probability of extinction

$$\beta = 1 - \frac{\eta_{\boldsymbol{n}}(0)}{N} = 1 - \eta_r(0)$$

follows.

In the general case $\alpha \neq 1$ the recursion of equation (4.49) can not be solved explicitly, since then the expected value of the state in the next generation depends on the distribution of states in the current generation. Thus the equation

$$\eta_{\boldsymbol{n}}(t+1) = f(\eta_{\boldsymbol{n}}(t)) \tag{4.51}$$

is impossible. But from the *first step analysis* in section 4.4.2 it is known, that the process converges to a terminal distribution $\tilde{P}$ with the expected value $\tilde{\eta}_{\boldsymbol{n}}$. Furthermore for $\alpha > 1$

$$\alpha \geq \frac{N\alpha}{N + (\alpha - 1)j} \geq 1$$

for all j holds, with equality on the left side only for $j = 0$ and on the right side only for $j = N$. Using that, from equation (4.49)

$$\alpha\eta_{\boldsymbol{n}}(t) > \eta_{\boldsymbol{n}}(t+1) > \eta_{\boldsymbol{n}}(t) \tag{4.52}$$

follows, if the process at time t is not in a terminal distribution $\tilde{P}$. Consequently, the expected state $\eta_{\boldsymbol{n}}(t)$ is a strictly monotonous increasing function of the time t, limited from above by an geometric increase, which clearly is never reached. Since it approaches the limit value $\tilde{\eta}_{\boldsymbol{n}}$ from below, for sufficiently large t it has to be quasi concave.

The development of the variance of the state is more difficult to consider, since it also depends on the expected value of the state. Therefore first the development of $\mathrm{E}\{\boldsymbol{n}^2(t)\}$ is explored. From its definition for the generation $t+1$

$$\mathrm{E}\{\boldsymbol{n}^2(t+1)\} = \sum_{i=0}^{N} i^2 \, p_{i,t+1}.$$

with equation (4.38)

$$\mathrm{E}\{\boldsymbol{n}^2(t+1)\} = \sum_{i=0}^{N} i^2 \sum_{j=0}^{N} \pi_{j,i} p_{j,t}$$

follows and by exchanging the sum order

$$\begin{aligned} \mathrm{E}\{\boldsymbol{n}^2(t+1)\} &= \sum_{j=0}^{N} p_{j,t} \sum_{i=0}^{N} i^2 \pi_{j,i} = \sum_{j=0}^{N} p_{j,t} \, \mathrm{E}\{\boldsymbol{n}^2(t+1) | \boldsymbol{n}(t) = j\} \\ &= \sum_{j=0}^{N} p_{j,t} [\eta_{\boldsymbol{n}}^2(t+1)|_{\boldsymbol{n}(t)=j} + \sigma_{\boldsymbol{n}}^2(t+1)|_{n(t)=j}]. \end{aligned} \tag{4.53}$$

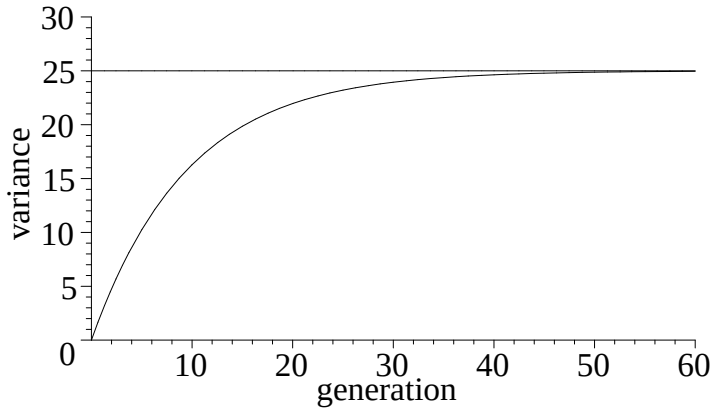


Figure 4.5: Development of the variance of the state for equal fitness against its limit in a population of size $N = 10$ starting with equal portions.

The conditional expected value $\eta_{\boldsymbol{n}}(t+1)|_{n(t)=j}$ has been calculated already above in equation (4.13) and the conditional variance of the state $\sigma^2_{\boldsymbol{n}}(t+1)|_{\boldsymbol{n}(t)=j}$ in equation (4.14). Using both formulas with some transformations, it follows that

$$\text{(4.54)}\quad \mathrm{E}\{\boldsymbol{n}^2(t+1)\} = \sum_{j=0}^{N} p_{j,t} j \frac{N^2\alpha}{(N+(\alpha-1)j)^2} + \sum_{j=0}^{N} p_{j,t} j^2 \frac{N\alpha(\alpha N-1)}{(N+(\alpha-1)j)^2}.$$

For the special case $\alpha = 1$ the above equation is simplified to

$$\text{(4.55)}\quad \mathrm{E}\{\boldsymbol{n}^2(t+1)\} = \eta_{\boldsymbol{n}}(t) + \left(1-\frac{1}{N}\right)\mathrm{E}\{\boldsymbol{n}^2(t)\} = \eta_{\boldsymbol{n}}(0) + \left(1-\frac{1}{N}\right)\mathrm{E}\{\boldsymbol{n}^2(t)\},$$

since then from equation (4.50) $\eta_{\boldsymbol{n}}(t) = \eta_{\boldsymbol{n}}(0)$ follows. The recursion can be solved analogously to equation (4.19). Equation (4.55) is of the type

$$\text{(4.56)}\quad x(t+1) = a + b\,x(t).$$

Continuing this way

$$x(t) = a\left(1 + b + b^2 + \cdots + b^{t-1}\right) + b^t x(0)$$

follows. Using the sum formula (A.1) with some transformations

$$\text{(4.57)}\quad x(t) = b^t\left(x(0) - \frac{a}{1-b}\right) + \frac{a}{1-b}$$

is obtained and from that with resubstitution

$$\mathrm{E}\{\boldsymbol{n}^2(t)\} = \left(1-\frac{1}{N}\right)^t\left(\mathrm{E}\{\boldsymbol{n}^2(0)\} - \eta_{\boldsymbol{n}}(0)N\right) + \eta_{\boldsymbol{n}}(0)N.$$

With the equations $\mathrm{E}\{\boldsymbol{n}^2(t)\} = \sigma^2_{\boldsymbol{n}}(t) + \eta^2_{\boldsymbol{n}}(t)$ and $\eta_{\boldsymbol{n}}(t) = \eta_{\boldsymbol{n}}(0)$ then

$$\text{(4.58)}\quad \sigma^2_{\boldsymbol{n}}(t) = (N-\eta_{\boldsymbol{n}}(0))\eta_{\boldsymbol{n}}(0) + \left(1-\frac{1}{N}\right)^t\left(\sigma^2_{\boldsymbol{n}}(0) - \eta_{\boldsymbol{n}}(0)(N-\eta_{\boldsymbol{n}}(0))\right)$$

follows leading to the limit

$$\lim_{t\to\infty} \sigma_{\boldsymbol{n}}^2(t) = \tilde{\sigma}_{\boldsymbol{n}}^2 = (N - \eta_{\boldsymbol{n}}(0))\eta_{\boldsymbol{n}}(0)$$

conforming with equation (4.44). For the portion $\boldsymbol{r}$ analogously

$$\sigma_{\boldsymbol{r}}^2(t) = (1 - \eta_{\boldsymbol{r}}(0))\eta_{\boldsymbol{r}}(0) + \left(1 - \frac{1}{N}\right)^t \left(\sigma_{\boldsymbol{r}}^2(0) - \eta_{\boldsymbol{r}}(0)(1 - \eta_{\boldsymbol{r}}(0))\right)$$

holds. Thus the variance of the state $\boldsymbol{n}$ is a strictly monotonously increasing and quasi concave function of the generation t, converging against $\tilde{\sigma}_{\boldsymbol{n}}^2$. Its increase during the evolution usually is called *genetic drift* and finally causes the extinction of one of the species. In figure 4.5 an iteration for a population of ten individuals is plotted, starting from the state $n = 5$. The average time to absorption calculated using the first step analysis in this case is $\tau_5 = 12.6$. Although the population is very small the convergence to the limit is very slow and the time until approaching it is approximately five times longer than the average time to absorption.

For $\alpha \neq 1$ the recursive equation (4.54) could not be solved explicitly. Nevertheless some statements about the development of the variance of the state are possible based on properties for $\alpha = 1$ and the limits. The evolution ends up in a terminal distribution $\tilde{P}$, having the expected value $\tilde{\eta}_{\boldsymbol{n}}$ and according to equation (4.44) the variance of the state

$$\tilde{\sigma}_{\boldsymbol{n}}^2 = \tilde{\eta}_{\boldsymbol{n}}(N - \tilde{\eta}_{\boldsymbol{n}}). \tag{4.59}$$

Clearly the limit variance of the state is concave in $\tilde{\eta}_{\boldsymbol{n}}$ and maximum for $\tilde{\eta}_{\boldsymbol{n}} = N/2$. Generally here the variance $\sigma_{\boldsymbol{n}}^2$ is limited from above by the inequality

$$\sigma_{\boldsymbol{n}}^2(t) \leq \eta_{\boldsymbol{n}}(t)(N - \eta_{\boldsymbol{n}}(t))$$

with equality only, if the distribution is one of the two terminal ones. Finally $\eta_{\boldsymbol{n}}(t)$ for $\alpha > 1$ has been shown above to be a strictly monotonous increasing function of the time t. In the following $\alpha > 1$ is assumed and a start distribution being not a terminal distribution.

If $\eta_{\boldsymbol{n}}(0) > N/2$ holds, the selection reduces the final variance of the state compared to the case of equal fitness, since $\tilde{\eta}_{\boldsymbol{n}} > \eta_{\boldsymbol{n}}(0)$ holds and according to equation (4.59) the final variance of the state is a strictly monotonous decreasing function of $\tilde{\eta}_{\boldsymbol{n}}$ for $\tilde{\eta}_{\boldsymbol{n}} > N/2$.

If $\eta_{\boldsymbol{n}}(0) < \tilde{\eta}_{\boldsymbol{n}} < N/2$ holds, then the final variance of the state is higher compared to the case of equal fitness, because according to equation (4.59) the final variance of the state is a strictly monotonous increasing function of $\tilde{\eta}_{\boldsymbol{n}}$ for $\tilde{\eta}_{\boldsymbol{n}} < N/2$. Summed up, $N^2/4 > \tilde{\sigma}_{\boldsymbol{n}}^2 > (N - \eta_{\boldsymbol{n}}(0))\eta_{\boldsymbol{n}}(0)$ follows. There the last term equals the limit variance of the state for $\alpha = 1$.

4.4.4 Development of the Diversity

Additionally to the development of the variance the development of the expected value of the diversity is of interest. In section 4.3.2 the Gini-Simpson index has been noted to be an appropriate measure, whose half $D^H(t)$ here is used for convenience. First the conditional

expected value of the diversity assuming a given portion $r(t)$ is considered. Again the notion $r(t)$ is an abbreviation of $\boldsymbol{r}(t) = r$. Using equation (4.26) and some transformations

$$\begin{aligned} \mathrm{E}\left\{D^H(t+1)\middle|r(t)\right\} &= \mathrm{E}\left\{\boldsymbol{r}(t+1)(1-\boldsymbol{r}(t+1))|r(t)\right\} \\ &= \mathrm{E}\{\boldsymbol{r}(t+1)\} - \mathrm{E}^2\{\boldsymbol{r}(t+1)\} - \sigma^2_{\boldsymbol{r}}(t+1)|_{r(t)} \end{aligned}$$

follows and with equations (4.16) and (4.17) finally

$$\mathrm{E}\left\{D^H(t+1)|r(t)\right\} = p(r(t))[1-p(r(t))]\left(1-\frac{1}{N}\right) \tag{4.60}$$

with $p(r(t))$ given also by equation (4.16). Ewens [25, p. 17f] already has proposed a specialization of this formula for $\alpha = 1$ and infers from it an extreme slow decrease of the expected value of the diversity. In fact, this statement assumes

$$\mathrm{E}\{D^H(t+1)|r(t)\} \approx r(t+1)(1-r(t+1)) \tag{4.61}$$

analogously to the approximation (4.18) to get a recursion equation. But with the results of the last section at least for the special case $\alpha = 1$ also an exact formula for the development of the expected expected value of the diversity is possible. Clearly

$$\mathrm{E}\{D^H(t+1)\} = \mathrm{E}\{\boldsymbol{r}(t+1)(1-\boldsymbol{r}(t+1))\} = \frac{1}{N}\mathrm{E}\{\boldsymbol{n}(t+1)\} - \frac{1}{N^2}\mathrm{E}\{\boldsymbol{n}^2(t+1)\} \tag{4.62}$$

holds. With equations (4.50) and (4.55) from that

$$\mathrm{E}\{D^H(t+1)\} = \frac{1}{N}\mathrm{E}\{\boldsymbol{n}(t)\} - \frac{1}{N^2}\left(\mathrm{E}\{\boldsymbol{n}(t)\} + \left(1-\frac{1}{N}\right)\mathrm{E}\{\boldsymbol{n}^2(t)\}\right)$$

follows and with some transformations

$$\mathrm{E}\{D^H(t+1)\} = \left(1-\frac{1}{N}\right)\mathrm{E}\{\boldsymbol{r}(t)(1-\boldsymbol{r}(t))\} = \left(1-\frac{1}{N}\right)\mathrm{E}\{D^H(t)\}. \tag{4.63}$$

This again is a recursion equation which also proves the approximation (4.61) to be in fact exact. Thus for $\alpha = 1.0$ the evolution reduces the expected value of the diversity of the population very slowly, which conforms to the slow convergence of the variance against its limit. The development from the initial state $n = 5$ for a population size $N = 10$ is shown in figure 4.6. Again there is a notable delay between the average time to absorption $\tau_5 = 12.6$ and the approach of the function to zero. Unfortunately for $\alpha \neq 1.0$ no explicit solution could be found. But because always one species extincts, the limit of the expected value of the diversity for large t always is zero.

From equation (4.60) for the leading non unit eigenvalue $\lambda_3 = 1 - 1/N$ follows [25, p. 305]. Interestingly this eigenvalue depends not on the selection probability $p_s(r(t))$ as long as the two absorbing states prevail, but it is caused by the the sampling with replacement, which leads to a binomial distribution. Both in equation (4.58) and (4.63) the leading non unit eigenvalue determines the speed of the geometric convergence of the variance of the state respectively of the expected value of the diversity.

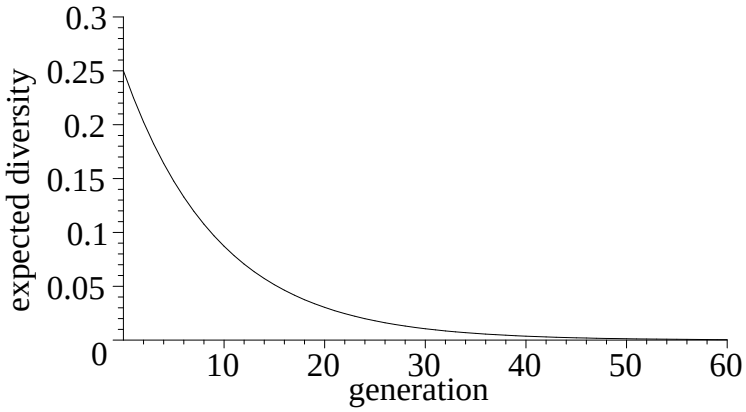


Figure 4.6: Development of the expected value of the diversity for equal fitness in a population of size $N = 10$ starting with equal portions.

4.4.5 Survival Probability of a Single New Individual

During evolution sometimes an individual of a new species is created. Then the question arises, whether the new species will prevail and replace the other one or whether it will extinct. In evolutionary computation the probability, that a single individual of a new species survives and its descendants finally replace the whole population, is called *takeover probability* [88]. It is a very important special case, because a new species is introduced usually only with a single individual. The takeover probability is also used as an index of the selection pressure of an evolutionary algorithm. If a superior new species can not extinct, then the same purpose can be achieved with the *takeover time*, which measures the time until the extinction of all other species.

First the probability of extinction of a new individual in the next generation after its introduction is considered. From equation (4.8) for $n = 1$ and $k = 0$

$$P\{\boldsymbol{n}(t+1) = 0 | \boldsymbol{n}(t) = 1\} = \binom{N}{0} p_s^0 (1-p_s)^N = (1-p_s)^N$$

follows with

$$p_s = \frac{\alpha}{N+\alpha-1} \quad \text{and} \quad q_s = 1 - p_s = \frac{N-1}{N+\alpha-1}$$

and with that finally

$$P\{\boldsymbol{n}(t+1) = 0 | \boldsymbol{n}(t) = 1\} = \left(\frac{N-1}{N-1+\alpha}\right)^N.$$

This term can be approximated further. For the selection probability p_s the limit

$$\lim_{N\to\infty} p_s = \lim_{N\to\infty} \frac{\alpha}{N-1+\alpha} = 0$$

holds and furthermore for the expected number of selected individuals

$$\lim_{N\to\infty} N p_s = \lim_{N\to\infty} \frac{N\alpha}{N-1+\alpha} = \alpha.$$

Both expressions are the conditions to be satisfied for the application of the Poisson-Theorem [71, p. 56]. Using it

$$\lim_{N\to\infty} \binom{N}{k} p_s^k (1-p_s)^{N-k} = e^{-\alpha}\frac{\alpha^k}{k!}. \tag{4.64}$$

follows and finally for $k = 0$

$$P\{\boldsymbol{n}(t+1) = 0 | \boldsymbol{n}(t) = 1\} \approx e^{-\alpha}. \tag{4.65}$$

If $\alpha = 1.1$ holds, then $P\{\boldsymbol{n}(t+1) = 0|\boldsymbol{n}(t) = 1\} \approx 0.33$ is obtained. I. e. in a sufficiently large population a new individual having a fitness $\alpha = 1.1$ times higher than the remaining other individual will be removed with probability 0.33 in the next selection step.

To calculate the general extinction probability again two cases have to be distinguished. If $\alpha = 1$ holds, from equation (4.50) $1 = \eta_{\boldsymbol{n}}(0) = \tilde{\eta}_{\boldsymbol{n}}$ follows. Then from equation (4.43) the probability of extinction $\beta = 1 - 1/N$ follows. Conversely the probability to be absorbed in state N and to take over the population is $1/N$. Finally the limit of the variance of the state is $\tilde{\sigma}^2_{\boldsymbol{n}} = N - 1$ according to equation (4.44). If $\alpha \neq 1$ holds, only a numerical solution can be calculated using the first step analysis proposed in section 4.4.2. Like above a fitness factor $\alpha = 1.1$ is assumed. This is realistic, because in the final stage of the run of an evolution program the individuals forming the population will differ only slightly in their fitness. Furthermore again a population size $N = 10$ is used. Then a limit of the expected value $\tilde{\eta}_{\boldsymbol{n}} = 2.05$ and of the variance $\tilde{\sigma}^2_{\boldsymbol{n}} = 16.27$ is obtained and an average time to absorption $\tau_1 = 6.70$. The probability of extinction is $\beta = 0.795$. For $\alpha = 1.0$ one gets $\beta = 0.9$, $\tilde{\eta}_{\boldsymbol{n}} = 1$, $\tilde{\sigma}^2_{\boldsymbol{n}} = 9$ and $\tau_1 = 5.75$. Thus although the population size is low and consequently the fitness of a new superior individual covers a high portion of the total fitness of the population, the probability to take over the population is relatively low. Furthermore the final variance of the state has increased compared to the case of equal fitness. In figures 4.7 and 4.8 the development of the expected value and the variance of the state during numerical iteration and their limits are plotted. The graphs are quasi concave and converge very slowly compared to the average time to absorption. The development of the expected value of the diversity in this case is shown in figure 4.9. A higher fitness coefficient retards the decrease to zero slightly. Also in this situation there is a notable delay between the average time to absorption and the approach of the indices to their limits.

Furthermore the dependence of survival probability and average absorption time from the population size are considered. Again both quantities are calculated using the first step analysis proposed in section 4.4.2.

In figure 4.10 the probability of a single new individual to take over the population is plotted over the logarithm of the population size for a fitness factor $\alpha = 1.1$. Increasing the population size lowers the survival probability slightly to a limit, a result which is different from that obtained by Rudolph [88] for tournament selection and non-generational population replacement. Additionally the survival probability for $\alpha = 1.0$, $1/N$, is also plotted.

Finally the average time to absorption (not to confuse with the takeover time) in relation to the logarithm of the population size is shown in figure 4.11. For $\alpha = 1.0$ the slope is lower than for $\alpha = 1.1$ and both graphs are nearly straight lines. Thus if a single

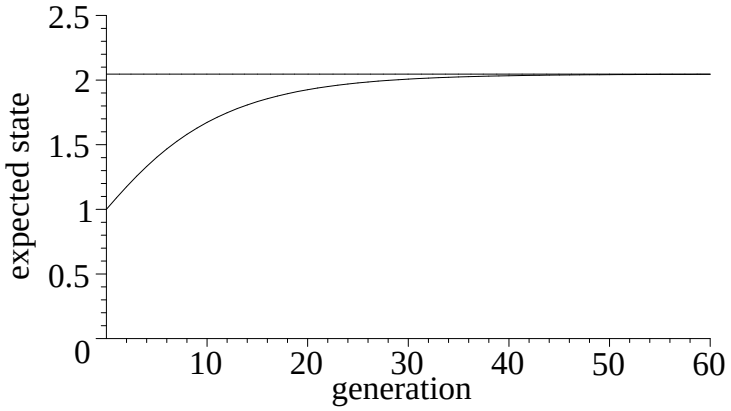


Figure 4.7: Development of the expected value of the state with its limit for $\alpha = 1.1$ starting at $\boldsymbol{n}(0) = 1$ in a population of size $N = 10$.

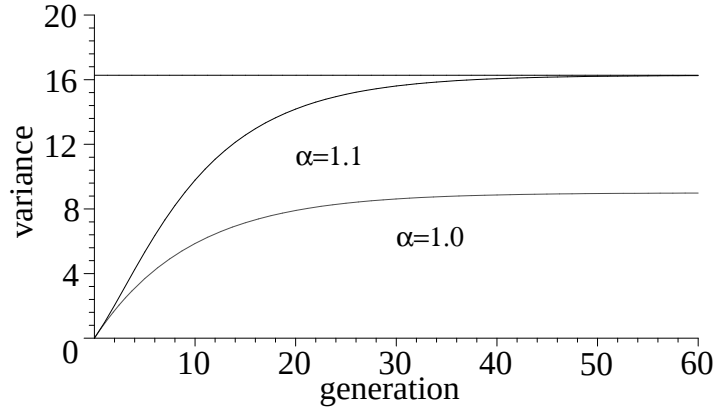


Figure 4.8: Development of the variance of the state for $\alpha = 1.0$ and $\alpha = 1.1$ with their limits starting at $\boldsymbol{n}(0) = 1$ in a population of size $N = 10$.

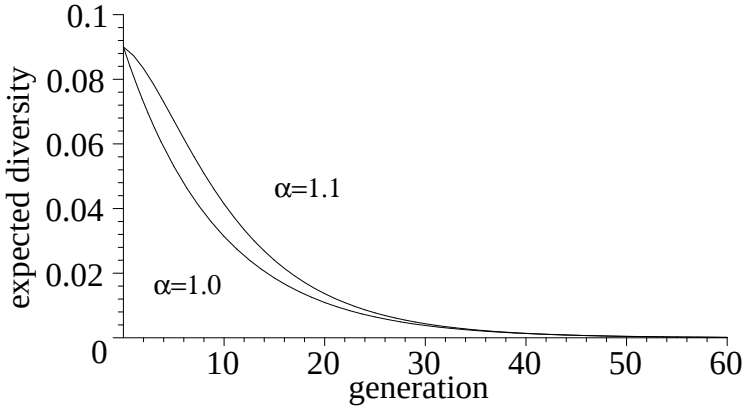


Figure 4.9: Development of the expected value of the diversity for $\alpha = 1.0$ and $\alpha = 1.1$ starting at $\boldsymbol{n}(0) = 1$ in a population of size $N = 10$.

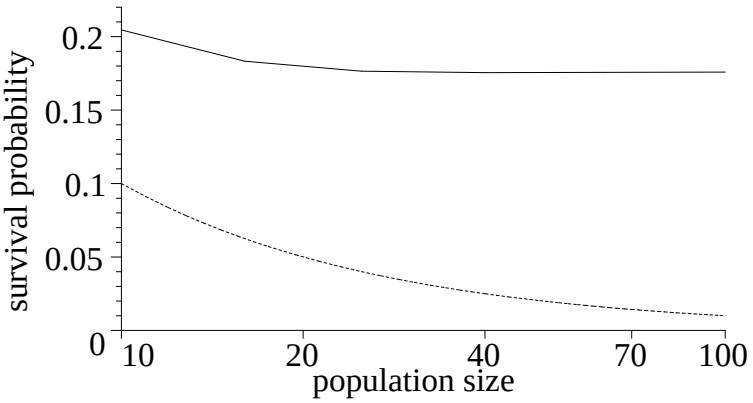


Figure 4.10: Dependence of the survival probability of a single new individual for $\alpha = 1.0$ (dotted line) and $\alpha = 1.1$ (solid line) from the population size.

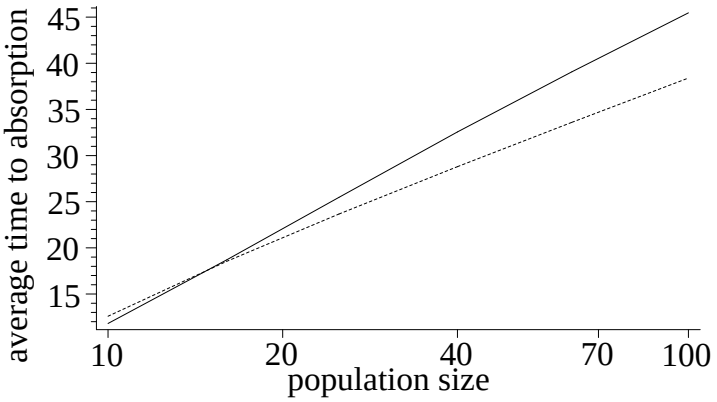


Figure 4.11: Dependence of the average time to absorption starting with a single new individual for $\alpha = 1.0$ (dotted line) and $\alpha = 1.1$ (solid line) from the population size.

new individual is slightly superior than the others, the increased drift to the state N also increases the time to absorption in sufficiently large populations[4].

Summed up, increasing the population size only increases the average time to absorption, but does not improve the survival probability of a single new superior individual as one would expect. In evolution programs large populations should be avoided, because it takes to long until a slightly superior new individual takes over a sufficiently high portion of the population, if at all it survives the selection process. During this long time its portion is likely to be reduced by mutation and crossover which is not taken into account here.

4.4.6 Selection in Large Populations

The main result of the last subsection has been, that if a single superior new individual enters a population, then its survival probability is very low. The situation changes if a sufficiently high number of new superior individuals is introduced in the population, respectively the start state $n(0)$ is high enough. Then the probability of extinction of the superior species is negligible and the development of the expected value of the state resembles that of the state in the deterministic models. Thus there is a threshold of a critical number of individuals of the superior species above which the superior species takes over the population with high probability. Furthermore, if the start fraction is small enough, with increasing population size the plot approaches the logistic shape of the deterministic approximation. E. g. for a population size $N = 200$ starting in state $n = 20$ the development of the expected value of the state is shown in figure 4.12. The variance of the state and the expected value of the diversity in this case show an interesting development. The development of the variance of the state is shown in figure 4.13 and that of the expected value of the diversity in figure 4.14. After an immediate very high peak caused by the strengthening interaction of drift and selection in the initial stage of the process both indices rapidly approach their final values, which is typical for such stochastic processes. Furthermore there is nearly no delay between the average time to absorption $\tau_{20} = 68$ and the approach of the indices to their limits. In both plots for comparison also the development for equal fitness is shown, in which case $\tau_{20} = 275$ holds.

4.5 Evolution with Mutation

4.5.1 General Model

Until now in the proposed model only selection has been considered. But in natural and artificial evolution there is usually also mutation and crossover. Both may lead to superior individuals, but also from superior individuals inferior ones may be derived. Generally it can be assumed, that it is much more probable to get inferior descendants then superior ones. This is especially true for the evolution used in genetic algorithms, if the tackled problem is hard. Since it is much to complicated to incorporate mutation and crossover

[4]For their diploid model Wright and Fisher propose an approximation of the time to absorption derived by another approach for equal fitness of the individuals [25, p. 18f]. For different fitness they propose a complicated approximation which can only be evaluated numerically [25, p. 19f]. They also give a formula for $1 - \beta$ derived from that.

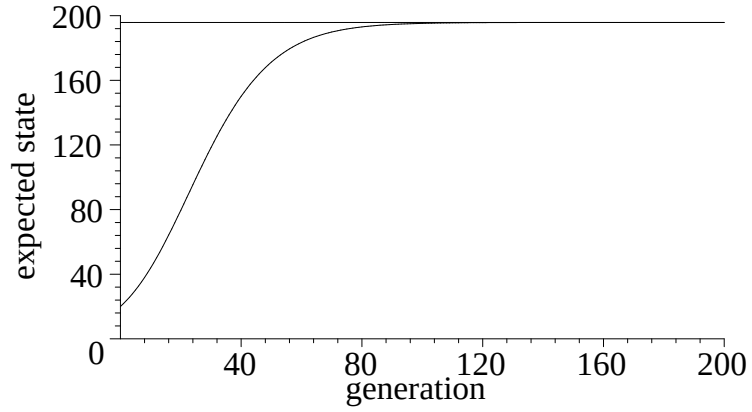


Figure 4.12: Development of the expected value of the state in a population of size $N = 200$ with start portion $\boldsymbol{r}(0) = 0.1$.

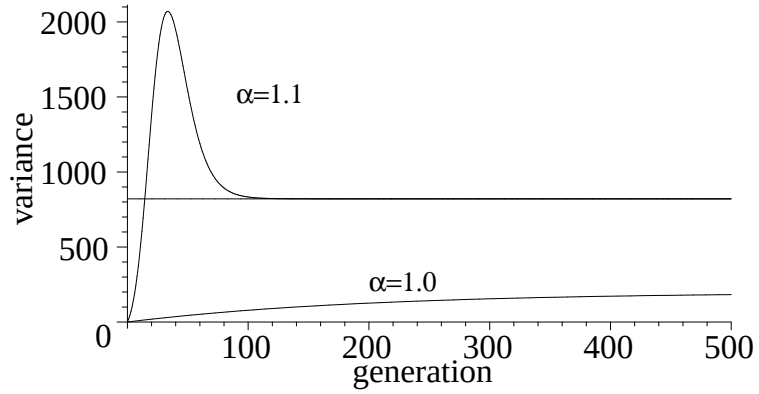


Figure 4.13: Development of the variance of the state in a population of size $N = 200$ with start portion $\boldsymbol{r}(0) = 0.1$.

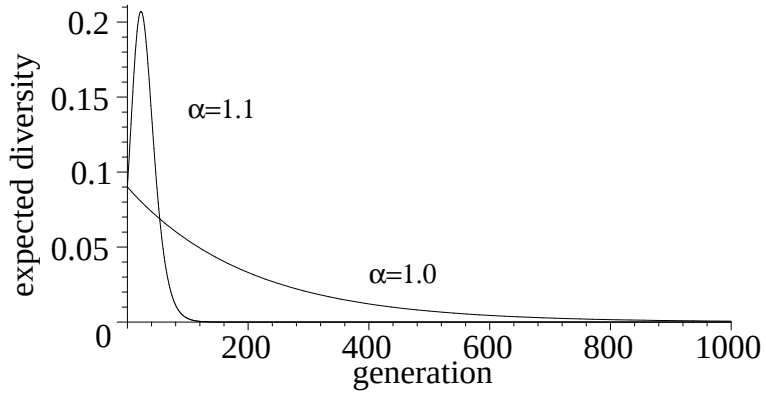


Figure 4.14: Development of the expected value of the diversity in a population of size $N = 200$ with start portion $\boldsymbol{r}(0) = 0.1$.

in the model considered until now, only mutation is added in a manner, that reflects the different probabilities discussed above and thus subsumes any modification of the individuals. Consequently, after selection of an individual from the population there may happen a mutation. The probability depends on the species it belongs to. If an individual of the superior species is selected, then with probability u a mutation takes place resulting in an individual of the worse species. Analogously with probability v an individual of the worse species is transformed to one of the superior one. As stated above generally here $v \ll u$ is assumed to hold and clearly $0 < u, v < 1$. Finally for convenience $0 < u + v < 1$, which for practical cases usually is no relevant constraint, and $\alpha \geq 1$ is asserted, i. e. the second species is as least as fit as the first one.

Following the conventions used until now the probability is required, to get an individual of the second and superior species after selection and mutation. The desired result can be performed by first selecting a superior individual with the probability $p_s(n)$, which then is not mutated with probability $1 - u$, or by selecting a worse individual with probability $1 - p_s(n)$, which then is mutated with probability v. Since selection and mutation are stochastically independent, the effective selection probability

$$\hat{p}_s(n) = p_s(n)(1-u) + (1-p_s(n))v = p_s(n)(1-u-v) + v \tag{4.66}$$

follows to get an individual of the second and superior species. This selection probability with mutation is bounded from above by the selection probability without mutation, which is proved by the following short calculation:

$$\begin{aligned} \hat{p}_s(n) &= p_s(n)(1-u-v) + v \\ &\leq p_s(n)(1-u-v) + v = p_s(n) - up_s(n) + v \\ &\leq p_s(n) - u + v \\ &\leq p_s(n). \end{aligned} \tag{4.67}$$

Using the selection probability of equation (4.7) from equation (4.66)

$$\hat{p}_s(n) = \frac{n(\alpha(1-u) - v) + vN}{N + (\alpha - 1)n} \tag{4.68}$$

follows. Analogously to the original model the selection process is repeated N times, which again implies a binomial distribution of the state in the next generation. Clearly, the process is also a homogeneous Markov chain and corresponding to equation (4.8) the transition probability is given by

$$\begin{aligned} P\{\boldsymbol{n}(t+1) = j | \boldsymbol{n}(t) = i\} &= \binom{N}{j} \hat{p}_s^j(i)(1-\hat{p}_s(i))^{N-j} \\ &= \binom{N}{j} (p_s(i)(1-u-v)+v)^j \left(1 - p_s(i)(1-u-v) - v\right)^{N-j}. \end{aligned}$$

For the equations (4.10) there are no correspondences due to the missing absorbing states. Goldberg and Segrest [34, p. 4*ff*] propose a different model to incorporate mutation by multiplying the original selection matrix by a special mutation matrix.

4.5.2 Deterministic Approximation

Establishing the Recursion

Since the introduction of mutation has only changed the selection probability from $p_s(n)$ to $\hat{p}_s(n)$, the analysis of the process can be performed similarly to the case without mutation. Modifying equation (4.11) for the conditional expected value of the state

$$\text{(4.69)}\quad \mathrm{E}\{\boldsymbol{n}(t+1)|n(t)\} = \eta_{\boldsymbol{n}}(t+1)|_{n(t)} = \hat{p}_s(n(t))\,N = \frac{n(t)N(\alpha(1-u)-v)+vN^2}{N+(\alpha-1)n(t)}$$

follows and using equation (4.15)

$$\text{(4.70)}\quad \mathrm{E}\{\boldsymbol{r}(t+1)|r(t)\} = \eta_{\boldsymbol{r}}(t+1)|_{r(t)} = \hat{p}_s(r(t)) = \frac{r(t)(\alpha(1-u)-v)+v}{1+(\alpha-1)r(t)}.$$

With the approximation (4.18) the first order difference equation

$$\text{(4.71)}\quad r(t+1) = g(r(t)) = \frac{r(t)(\alpha(1-u)-v)+v}{r(t)(\alpha-1)+1}$$

follows, which is of the type

$$\text{(4.72)}\quad r(t+1) = g(r(t)) = \frac{a\,r(t)+b}{c\,r(t)+d}\,.$$

Properties of the Recursion Equation

By setting $r(t+1) = r(t) = z$ the fixed points of the recursive equation (4.72) can be found. Doing so

$$\text{(4.73)}\quad cz^2 + z(d-a) - b = 0$$

follows. Now two different cases have to be considered.

If $c = 0$ holds, which implies $\alpha = 1$, then equation (4.73) has only one solution

$$\text{(4.74)}\quad z = \frac{b}{d-a} = \frac{v}{u+v}.$$

Its asymptotic stability can be proved analogous to section 4.3.2. From equation (4.72)

$$\text{(4.75)}\quad r(t+1) = g(r(t)) = \frac{a}{d}r(t) + \frac{b}{d} = (1-u-v)r(t) + v$$

follows. Applying the inequality (4.21) leads to

$$\left|\frac{d}{dr}((1-u-v)r+v)\right|_{r=z} = |1-u-v| < 1,$$

because $0 < u, v < 1$ has been asserted, which proves the asymptotic stability of the fixed point.

The recursion equation (4.75) is of the same type as equation (4.56). Consequently its solution is

$$r(t) = (1-u-v)^t \left(r(0) - \frac{v}{u+v} \right) + \frac{v}{u+v} \tag{4.76}$$

with the limit

$$\lim_{t\to\infty} r(t) = \frac{v}{u+v} = z.$$

The term in the parenthesis in equation (4.76) denotes the distance between start portion and fixed point. The geometric convergence is very slow because u and v usually are very small and consequently the factor $1-u-v$, which in this case is the leading non unit eigenvalue of the transition matrix Π [25, p. 82], is close to one.

If $c \neq 0$ holds, which implies $\alpha \neq 1$, then equation (4.73) generally has exactly two solutions

$$z_1 = \frac{a-d}{2c} + \sqrt{\frac{(a-d)^2}{4c^2} + \frac{b}{c}} \quad \text{and} \quad z_2 = \frac{a-d}{2c} - \sqrt{\frac{(a-d)^2}{4c^2} + \frac{b}{c}} \tag{4.77}$$

The constraint $\alpha \geq 1$ given initially implies $c > 1$, and additionally $v = b > 0$ has been assumed. Consequently, the inequality

$$\left| \frac{a-d}{2c} \right| < \sqrt{\frac{(a-d)^2}{4c^2} + \frac{b}{c}} \tag{4.78}$$

holds. Since also the constraint $0 \leq z \leq 1$ has to be hold, z_1 is the only valid solution.

If the constraints are relaxed to include the case $v = b = 0$, then for this special case $n = 0$ becomes an absorbing state in the Markov model and from equation (4.77) it follows that

$$z_1 = \frac{a-d}{c} = 1 - \frac{\alpha u}{\alpha - 1} \quad \text{and} \quad z_2 = 0. \tag{4.79}$$

Consequently, for the mutation rate

$$u \leq 1 - \frac{1}{\alpha} \tag{4.80}$$

has to hold to ensure $z_1 \geq 0$ and thus to be a valid solution. Setting $\alpha \leq 1$ contradicts this condition and consequently only z_2 remains.

Finally, if $u = 0$ is set, then $z_1 = 1$ follows and $n = N$ becomes an absorbing state in the Markov model.

To prove the asymptotic stability of the solutions, the inequality (4.21) has to be hold for the general recursion equation (4.72). Consequently

$$\left| \frac{d}{dr} g(r(t)) \right|_{r(t)=z} = \left| \frac{d}{dr} \frac{a\,r(t)+b}{c\,r(t)+d} \right|_{r(t)=z} = \left| \frac{ad-bc}{(c\,z+d)^2} \right| = \frac{\alpha(1-u-v)}{((\alpha-1)z+1)^2} < 1$$

must be satisfied, from which the condition

$$\alpha(1-u-v) < ((\alpha-1)z+1)^2 \tag{4.81}$$

follows. The right side is a strong monotonous increasing function of z. Thus, to prove stability it is sufficient to show, that the last inequality holds for an estimator of z from below. From inequality (4.78) for z_1

$$\frac{a-d}{c} = \frac{\alpha - 1 - \alpha u - v}{\alpha - 1} \leq z_1 \tag{4.82}$$

follows. Inserting the left side into inequality (4.81)

$$\alpha(1-u-v) < ((\alpha(1-u-v) + (\alpha-1)v)^2$$

follows. Clearly the inequality holds for $\alpha(1-u-v) \geq 1$. If this condition is not satisfied, then asymptotic stability follows from inequality (4.81) for any $z \geq 0$. This is especially the case if $\alpha \leq 1$. Consequently, under the given constraints the system is structurally stable in relation to α, i. e. a change over the point $\alpha = 1$, which is critical for the model without mutation, generally only changes the value of the fixed point but not the overall behavior. Setting $u = 0$ also does not change this situation.

If $b = v = 0$ is set, then z_1 is asymptotically stable for $\alpha > 1$ and from the condition (4.81) for $z_2 = 0$ follows the condition

$$u > 1 - \frac{1}{\alpha},$$

which is the opposite of condition (4.80). Thus $z_2 = 0$ is solely stable if it is the only fixed point.

Summed up, if the selection is strong enough then there is a stable fixed point with a balance of selection and mutation. If selection is not strong enough, the system can not escape from the otherwise unstable state $r(t) = 0$, which thus is then a stable fixed point. The border is given by

$$u = 1 - \frac{1}{\alpha},$$

where both fixed points are equal. This situation is very similar to a situation in the model of Eigen, McCaskill and Schuster [23, p. 174*ff*]. It also models an evolution process with selection and mutation deterministically. But their approach is continuous in time and there are no explicit solutions of the related differential equations. In their terminology the coexistence of both species is called *organization*.

In the case considered here the general difference equation (4.72) is reduced to

$$r(t+1) = \frac{a\, r(t)}{c\, r(t) + d},$$

which already has been solved in section 4.3.2. Resubstituting the parameters into the solution (4.23) leads to

$$r(t) = \frac{r(0)\alpha^t(1-u)^t(\alpha(1-u)-1)}{r(0)(\alpha-1)(\alpha^t(1-u)^t-1) + \alpha(1-u) - 1} \tag{4.83}$$

with the special solution $r(t) = r(0) = 0$.

Unfortunately an explicit solution of the general recurrence equation (4.71) can not be given, only a numerical iteration is possible. In figure 4.15 the development of the portion

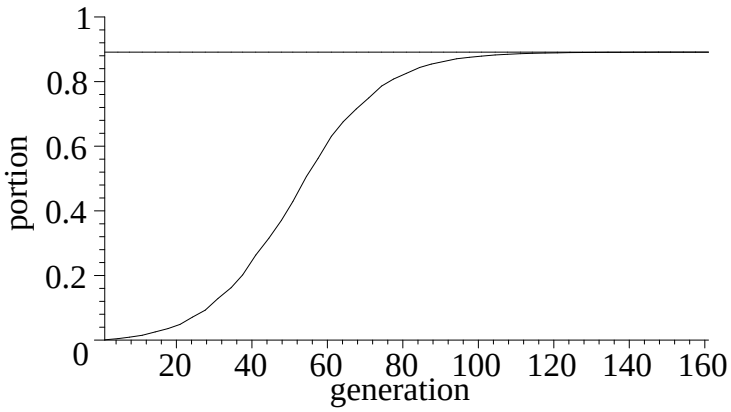


Figure 4.15: Development of the portion $r(t)$ with increasing generation number for deterministic selection combined with mutation.

is plotted using the start value $r(0) = 0.001$, the mutation rates $u = 0.01$ and $v = 0.001$ and the fitness factor $\alpha = 1.1$. The logistic development resembles that of the deterministic models without mutation except for the limit portion due the fixed point z, which is smaller than one. Setting $v = 0$ does not change the plot significantly, only the limit respectively fixed point is lowered. Of course, then also the special solution $r(t) = r(0) = 0$ is possible. The development of the diversity also is very similar to the case without mutation already shown in figure 4.3. Again the most prominent difference is the limit value depending on the fixed point.

Since for the model without mutation the investigation of the corresponding differential equation model leads to unsatisfying results, the related considerations have not been performed. Furthermore, the difference equation can not be solved explicitly and consequently no corresponding exact differential equation can be deduced from its solution.

4.5.3 Markov Model with Mutation

General Notes

Opposed to the model without mutation, in the general case with mutation in both directions there are no absorbing states, and the Markov matrix is positive. Thus, one is a unique eigenvalue, to which a stationary respectively balance distribution $\tilde{P}$ corresponds. Consequently, $\tilde{P}$ also satisfies equation (4.42), which can be used to calculate it together with the constraint

$$\sum_{i=0}^{N} \tilde{p}_i = 1.$$

But trying to solve the resulting homogeneous system of linear equations

$$P^T(\Pi - I) = P^T \Pi^{\heartsuit} = \mathsf{0} \tag{4.84}$$

with $\mathsf{0}^T = (0, \ldots, 0)$ usually only leads to the trivial solution $P = \mathsf{0}$ due to numerical inaccuracies. But the homogeneous system of linear equations can be transformed into an

inhomogeneous one, which can be solved, and from which the solution of the original homogeneous system can be obtained. From the theory of linear equations the rank of the matrix $\Pi^{\heartsuit}$ with dimension N in equation (4.84) has to be $N-1$ to have a solution different from the trivial one, which always is true for a positive Markov matrix. Consequently, the column vectors $\Pi_j^{\heartsuit}$ are linearly dependent and with $\lambda_j \neq 0$ the equation

$$\sum_{j=0}^{N} \lambda_j \Pi_j^{\heartsuit} = \mathsf{0} \tag{4.85}$$

holds. There one of the λ_j can be chosen arbitrarily, on which then the others depend. Clearly, $P^T = \lambda^T = (\lambda_0, \ldots, \lambda_N)$ is a solution of equation (4.84). Now the corresponding inhomogeneous equation

$$\sum_{j=0}^{N-1} \xi_j \Pi_j^{\heartsuit} + \xi_N \Upsilon = -\Pi_N^{\heartsuit} \tag{4.86}$$

with $\Upsilon^T = (1, \ldots, 1)$ is considered. It has exactly one solution $\Xi = (\xi_0, \ldots, \xi_N)$, because the rank of the new matrix

$$\Pi^{\star} = (\Pi_0^{\heartsuit} \Pi_1^{\heartsuit} \ldots \Pi_{N-1}^{\heartsuit} \Upsilon)$$

is N. To insure the correspondence of equations (4.85) and (4.86) $\lambda_N = 1$ and $\xi_N = 0$ have to hold, from which $\xi_j = \lambda_j$ for $j = 0, \ldots, N-1$ follows. Thus by solving the inhomogeneous equation (4.86) one can obtain a solution Λ of the homogeneous equation (4.85), which normed by the sum of its components is a valid stationary distribution $\tilde{P}$. Usually in practical calculations due to numerical inaccuracies $\xi_N = 0$ holds only approximately but sufficiently accurate.

According to the Perron Frobenius Theorem [9, p. 197*ff*], no matter in which distribution of states the process starts, it will converge to the stationary distribution. The convergence is approximately geometric, and the relative speed is determined by the second highest eigenvalue. Unfortunately, in most cases it is difficult to calculate the eigenvalues of a transition matrix. Then as without mutation alternatively one can iterate the Markov chain and calculate the expected value of the state, its variance and the expected value of the diversity. The development of these indices is sufficient to rate the behavior of the process and sometimes explicit solutions can be found for them.

Development of the Indices

Proceeding analogous to section 4.4.3 first the development of the expected value of the state is considered. From equation (4.48) together with equation (4.69) for the expected value of the state

$$\eta_{\boldsymbol{n}}(t+1) = \sum_{j=0}^{N} p_{j,t} \hat{p}(j) N = \sum_{j=0}^{N} p_{j,t} \frac{jN(\alpha(1-u)-v) + vN^2}{N + (\alpha - 1)j} \tag{4.87}$$

follows. Since already for the corresponding deterministic recursion equation (4.71) no explicit solution could be found, this is true also here. But for the special case $\alpha = 1$ the recursion can be solved. Then from equation (4.87)

$$\eta_{\boldsymbol{n}}(t+1) = vN + (1-u-v)\sum_{j=0}^{N} jp_{j,t} = vN + (1-u-v)\eta_n(t) \tag{4.88}$$

follows, which corresponds to the deterministic recursion equation (4.75). Analogous the solution is

$$\eta_{\boldsymbol{n}}(t) = (1-u-v)^t\left(\eta_n(0) - \frac{v}{u+v}\right) + \frac{v}{u+v}. \tag{4.89}$$

Thus for $\alpha = 1$ generally the development of the portion in the deterministic model equals that of the expected value in the probabilistic model. The geometric convergence determined by the highest non unit eigenvalue also is a correspondence to the Perron Frobenius theorem. Ewens [25, p. 70] also in fact got the recursion equation (4.88), but only used it to calculate the limit. Finally it should be emphasized that the convergence speed does not depend on the population size N.

Since the following calculations are lengthy, equation (4.89) is simplified to

$$\eta_{\boldsymbol{n}}(t) = d^t(\eta_{\boldsymbol{n}}(0) - f) + f = d^t e + f. \tag{4.90}$$

The development of the variance of the state and of the expected value of the diversity again only can be calculated indirectly by first considering $\mathrm{E}\{\boldsymbol{n}^2(t)\}$. From equation (4.53) together with equations (4.11) and (4.12)

$$\mathrm{E}\{\boldsymbol{n}^2(t+1)\} = \sum_{j=0}^{N} p_{j,t} N p_s(n(t))[p_s(n(t))(N-1)+1]$$

follows. Taking mutation into account, $\hat{p}_s(n)$ from equation (4.68) is used instead of $p_s(n)$ and with some transformations

$$\mathrm{E}\{\boldsymbol{n}^2(t+1)\} = \sum_{j=0}^{N} p_{j,t} \frac{jN(\alpha(1-u)-v) + vN^2}{(j(\alpha-1)+N)^2}\left(j[N\alpha - (N-1)(\alpha u - v) - 1] + N(v(N-1)+1)\right)$$

follows. This stochastic recursion equation is quadratic in both counter and denominator. Clearly, no general explicit solution could be found. But again for equal fitness the equation is simply enough to be solved explicitly. Thus setting $\alpha = 1$ and performing some transformations

$$\mathrm{E}\{\boldsymbol{n}^2(t+1)\} = \sum_{j=0}^{N} p_{j,t} j^2 (1-u-v)^2 \frac{N-1}{N} + j(1-u-v)[2(N-1)v+1] + vN^2(v(N-1)+1)$$

follows and finally

$$\begin{aligned}\mathrm{E}\{\boldsymbol{n}^2(t+1)\} &= \mathrm{E}\{\boldsymbol{n}^2(t)\}(1-u-v)^2\frac{N-1}{N}\\ &\quad +\mathrm{E}\{\boldsymbol{n}(t)\}(1-u-v)[2(N-1)v+1]+vN^2(v(N-1)+1).\end{aligned}$$

To simplify the further calculations, in the above equation the factors are combined to symbolic constants, which leads to

$$\mathrm{E}\{\boldsymbol{n}^2(t+1)\} = a\,\mathrm{E}\{\boldsymbol{n}^2(t)\} + b\,\mathrm{E}\{\boldsymbol{n}(t)\} + c$$

and using equation (4.90) finally to

$$\mathrm{E}\{\boldsymbol{n}^2(t+1)\} = a\,\mathrm{E}\{\boldsymbol{n}^2(t)\} + bed^t + fb + c,$$

which can be further combined to

$$\mathrm{E}\{\boldsymbol{n}^2(t+1)\} = a\,\mathrm{E}\{\boldsymbol{n}^2(t)\} + gd^t + h. \tag{4.91}$$

This recursive equation again can be solved by repeated insertion. For $t+2$

$$\mathrm{E}\{\boldsymbol{n}^2(t+2)\} = a\,\mathrm{E}\{\boldsymbol{n}^2(t+1)\} + gd^{t+1} + h$$

and using equation (4.91)

$$\mathrm{E}\{\boldsymbol{n}^2(t+2)\} = a^2\,\mathrm{E}\{\boldsymbol{n}^2(t)\} + d^tg(a+d) + h(a+1).$$

Repeating this procedure leads to

$$\mathrm{E}\{\boldsymbol{n}^2(t)\} = a^t\,\mathrm{E}\{\boldsymbol{n}^2(0)\}+g^t(a^{t-1}+a^{t-2}g+a^{t-3}g^2+\ldots+g^{t-1})+h(a^{t-1}+a^{t-2}+\ldots+1). \tag{4.92}$$

Now the sum formulas (A.1) and (A.2) can by applied. Doing so and performing some further simplifications

$$\mathrm{E}\{\boldsymbol{n}^2(t)\} = a^t\left(\mathrm{E}\{\boldsymbol{n}^2(0)\} - \frac{g}{d-a} + \frac{h}{a-1}\right) + d^t\frac{g}{d-a} - \frac{h}{a-1}$$

follows and finally resubstituting g and h

$$\mathrm{E}\{\boldsymbol{n}^2(t)\} = a^t\left(\mathrm{E}\{\boldsymbol{n}^2(0)\} - \frac{be}{d-a} + \frac{fb+c}{a-1}\right) + d^t\frac{be}{d-a} - \frac{fb+c}{a-1}. \tag{4.93}$$

Since

$$a = (1-u-v)\frac{N-1}{N} < 1 \quad \text{and} \quad d = (1-u-v) < 1 \tag{4.94}$$

hold, the limit

$$\lim_{t\to\infty}\mathrm{E}\{\boldsymbol{n}^2(t)\} = \frac{fb+c}{1-a}$$

is obtained. Using the intermediate results now the remaining indices can be calculated. With the general formula

$$\sigma_n^2(t) = \mathrm{E}\{\boldsymbol{n}^2(t)\} - \mathrm{E}^2\{\boldsymbol{n}(t)\}$$

and equations (4.93) and (4.90) with some transformations

$$\sigma_n^2(t) = a^t \left(\mathrm{E}^2\{\boldsymbol{n}(0)\} + \sigma_n^2(0) - \frac{b\,\mathrm{E}\{\boldsymbol{n}(0)\} - bf}{d-a} + \frac{fb+c}{a-1} \right)$$
$$+d^t \left(\mathrm{E}\{\boldsymbol{n}(0)\} - f\right) \left(\frac{b}{d-a} - 2f \right) - d^{2t} \left(\mathrm{E}\{\boldsymbol{n}(0)\} - f\right)^2 - \frac{fb+c}{a-1} + f^2$$

follows. With the same argumentation as above the limit is

$$\lim_{t\to\infty} \sigma_n^2(t) = f^2 - \frac{fb+c}{a-1}.$$

Finally from equation (4.62) together with (4.93) and (4.90)

$$\mathrm{E}\{D^H(t)\} = d^t \left(\mathrm{E}\{\boldsymbol{n}(0)\} - f\right) \frac{1}{N} \left(1 - \frac{1}{N} \frac{b}{d-a} \right)$$
$$-a^t \frac{1}{N^2} \left(\mathrm{E}\{\boldsymbol{n}^2(0)\} - (\mathrm{E}\{\boldsymbol{n}(0)\} - f) \frac{b}{d-a} + \frac{fb+c}{a-1} \right) + \frac{f}{N} - \frac{1}{N^2} \frac{fb+c}{a-1}$$

is obtained for the expected value of the diversity. Analogous the limit is

$$\lim_{t\to\infty} \mathrm{E}\{D^H(t)\} = \frac{f}{N} - \frac{1}{N^2} \frac{fb+c}{a-1}.$$

Although the resulting above formulas are complicated, it is easy to see using the equations (4.94), that the leading non unit eigenvalue $1-u-v$ again plays an important role. Thus not only the development of the expected value of the state but also of the variance of the state and the expected value of the diversity is controlled by it.

As an example a population of size $N = 10$ is considered. The mutation rates are $v = 0.001$ and $u = 0.01$ and initially both species have equal portions. In figure 4.16 the development of the expected value of the state is presented according to equation (4.89). The limit is given by the mutation rates only. The development of the variance of the state is shown in figure 4.17. After a rapid increase to the maximum the variance of the state decreases geometrically to its limit. Corresponding the development of the expected value of the diversity is presented in figure 4.18. It shows an initial rapid decrease followed by a slow final convergence. Increasing the population size slows down the development both of the variance of the state and the expected value of the diversity and raises their limit values, but the general shape of the plots remains. Clearly, the variance of the portion decreases according to equation (4.17), but preserves its general shape.

As noted above for $\alpha \neq 1$ closed explicit formulas for the indices can not be found. Consequently an evaluation of the stationary distribution and an iteration of the process is possible only numerically. Again here a single new superior individual in an otherwise uniform population with a fitness factor $\alpha = 1.1$ is considered. The mutation rates still are $v = 0.001$ and $u = 0.01$.

For large populations of size $N \geq 200$ the development of the expected value resembles that of the portion in the deterministic model already shown in figure 4.15. In contrast, for a small population of size $N = 10$ the development of the expected value of the state is shown in figure 4.19, which does not show a logistic shape as in figure 4.15. Generally the final expected portion of the superior species for small populations is a lot lower than the

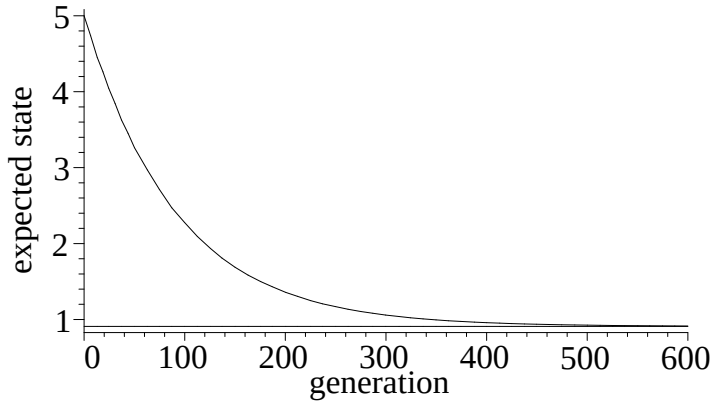


Figure 4.16: Development of the expected value of the state for equal fitness for a population of size $N = 10$ starting with equal portions.

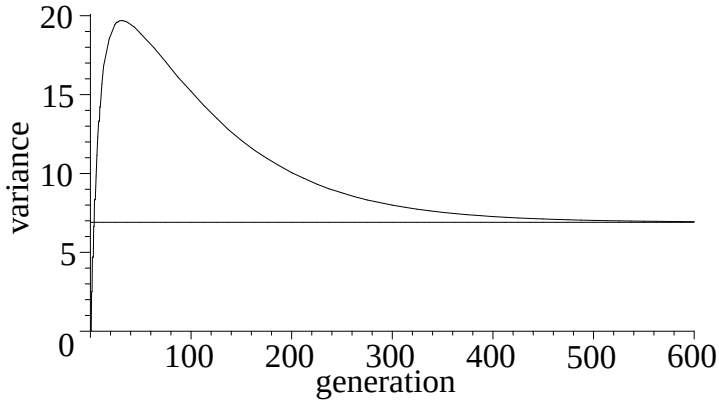


Figure 4.17: Development of the variance of the state for equal fitness for a population of size $N = 10$ starting with equal portions.

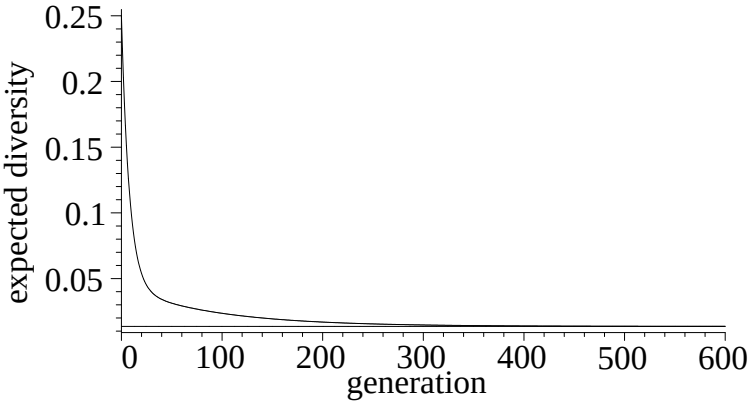


Figure 4.18: Development of the expected value of the diversity for equal fitness for a population of size $N = 10$ starting with equal portions.

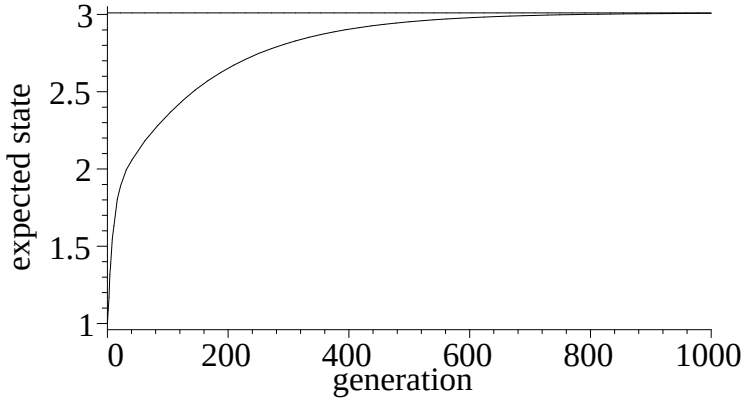


Figure 4.19: Development of the expected value of the state for different fitness for a population of size $N = 10$ starting with a single superior individual.

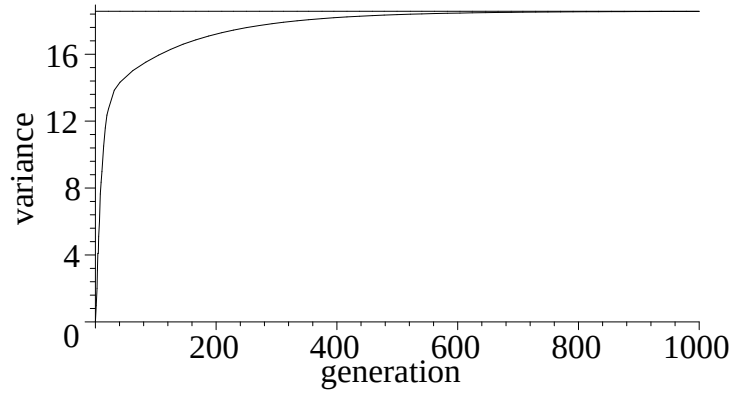


Figure 4.20: Development of the variance of the state for a population of size $N = 10$ starting with a single superior individual.

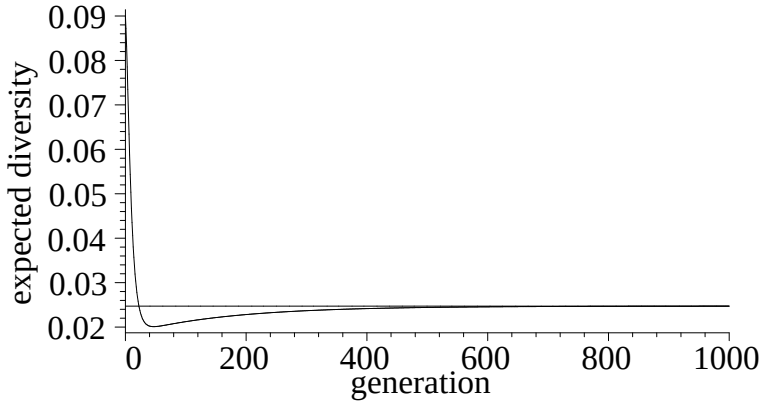


Figure 4.21: Development of the expected value of the diversity for a population of size $N = 10$ starting with a single superior individual.

fixed point of the deterministic model, to which it converges with increasing population size. This is due to the genetic drift caused by sampling errors, which for small populations has an important impact. If the number of individuals of the superior species is small, the genetic drift may lead to a temporary extinction which is improbable otherwise. Starting with the same portions, the initial number of superior individuals is a linear function of the population size which explains the differences.

The development of the variance of the state and of the expected value of the diversity in this setup also changes significantly with the population size. For a small population of size $N = 10$ the development of the variance of the state is depicted in figure 4.20 and of the expected value of the diversity in figure 4.21. Increasing the population size to $N = 200$ leads to a development of the variance of the state shown in figure 4.22 and of the expected value of the diversity shown in figure 4.23. Both changes are also due to the different influence of the genetic drift. Generally the final variance of the portion decreases with the population size whereas the final expected value of the diversity increases as expected.

Finally the important special case $v = 0$ is considered, where mutation and selection work against each other. Then the resulting Markov chain has a single absorbing state at $n = 0$, in which it finally will end always, and to which the asymptotically unstable fixed point $z_2 = 0$ corresponds. But there is no absorbing state related to the asymptotically stable fixed point z_1 of equation (4.79). The average time to absorption again can be calculated using the first step analysis. The behavior of the process also shows up in the plot of an numerical iteration depicted in figure 4.24. The population size is $N = 20$, the fitness factor $\alpha = 1.1$, the mutation rate $u = 0.01$ and the process starts with the portion $\boldsymbol{r}(0) = 0.1$. After a short initial increase the expected value of the portion converges to zero as expected because of the related absorbing state. But in contrast to the extremely slow convergence of the plot the average time to absorption calculated with the first step analysis is approximately 214 generations.

4.6 Discussion

4.6.1 Comparison of the Two Species Models

In the two species model considered here the genetic drift generally plays an important role. It is caused by the sampling errors of the random selection of individuals transforming one generation to the next and can not be neglected especially for small populations. Its impact is maximum if there is no mutation and the fitness of the individuals of both species is equal. Then it leads to the extinction of one of the species because of the related two absorbing states. The absorption or takeover probabilities depend on the start distribution of the state. In contrast, the expected value remains constant pretending the preservation of the start configuration. The first step analysis of the Markov model indicates a short time to absorption, which in evolution programs is called *premature convergence*. It contradicts the very low convergence speed of the variance of the state and the expected diversity, a discrepancy which already has been recognized by Ewens [25, p. 79]. In the corresponding deterministically approximated system each state is a stable fixed point, which preserves the current state in the absence of disturbances and prevents the further increase of single

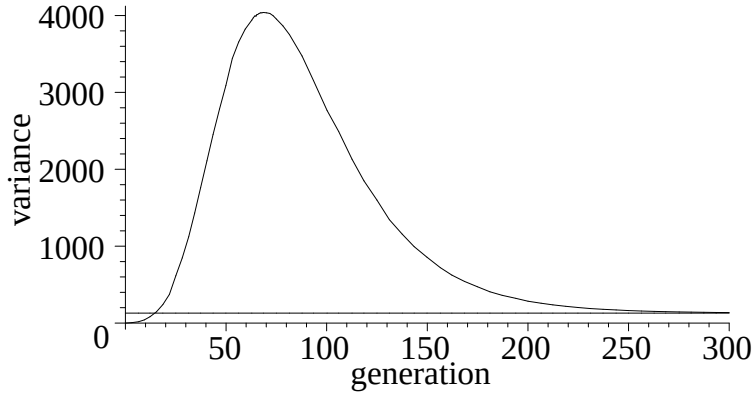


Figure 4.22: Development of the variance of the state for a population of size $N = 200$ starting with a single superior individual.

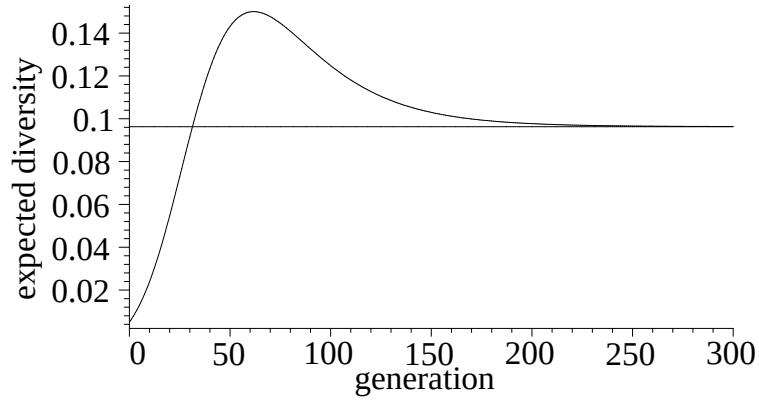


Figure 4.23: Development of the expected value of the diversity for a population of size $N = 200$ starting with a single superior individual.

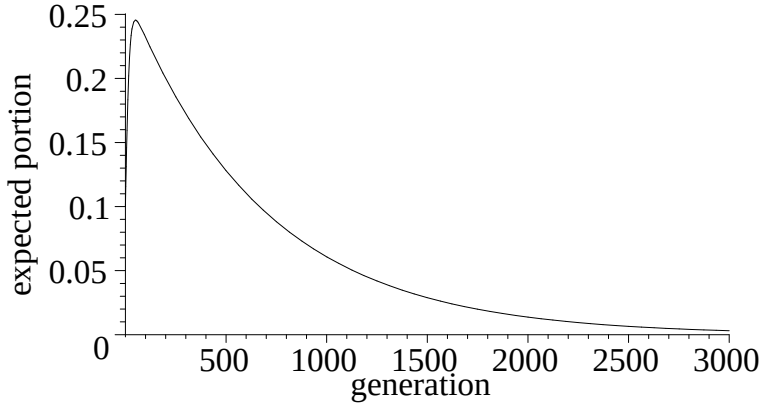


Figure 4.24: Development of the expected value of the portion for stochastic selection against mutation in a population of size $N = 20$.

disturbances. But the stability is not asymptotic, thus the system is not led back to the state prior to the disturbance, which thus prevails. Consequently sampling errors are not corrected but accumulate. Finally the whole system is structural unstable in respect to the fitness factor α, since a small change away from $\alpha = 1$ respectively equal fitness changes the fixed points and their stability totally. Summed up, although the constant development of the expected value of the portion in the Markov model and of the portion in the deterministic models coincide, alone they are not sufficient to describe the system. But the structural instability of the system and the missing asymptoticly stability of the fixed points correspond very well with the genetic drift.

If the species specific fitness differ, then in the deterministic system there are only two fixed points corresponding to the absorbing states of the Markov model. The fixed point related to an uniform population of superior individuals is asymptotically stable, whereas the other one is unstable. Due to these properties the deterministic portion will always converge to the stable fixed point unless the system starts in the unstable fixed point. If the fitness factor crosses $\alpha = 1$, the stability of the fixed points changes as considered already above. Finally the plot of the portion shows a typical logistic shape induced by the underlying Verhulst-Pearl differential equation. The selection pressure is determined both by the logarithm of the fitness factor and the diversity in the population. Thus the diversity is a moving force of the selection process. In the Markov model the development depends on the population size and the start state respectively number of individuals of the superior species. If the population is large and the start number of individuals of the superior species is sufficiently high, then the probability of the extinction of the superior species is negligible and it will prevail one self with high probability. This property corresponds to many real world processes, where also a critical threshold has to be passed to ensure success with high probability. Furthermore, then there is also nearly no delay between the average time to absorption and the approach of the indices to their limit values. Unfortunately this situation is of little practical relevance. Finally in this situation the development of the expected portion equals that of the portion of the deterministic model. Thus the genetic drift can be neglected, which conforms also to the vanishing conditional variance of the portion for large populations.

If an individual of a new species with slightly superior fitness is introduced in an otherwise uniform population, then its probability to survive and to prevail is only slightly increased compared to the case of equal fitness. The cause is the genetic drift, which in one direction supports the prevalence of the new species and in the other one its extinction. If the amount of the considered species is low, then the probability that a sampling error will lead to extinction is not negligible, which is true especially for a single individual. This is a severe problem in the final stage of an evolutionary algorithm, where fitness differences are very small and an innovation usually is introduced by a single individual. Informally the considered selection process is called to have a low selection pressure [3]. Again the average time to absorption is much shorter than the time the indices take to approach their limit values, leading rapidly to uniform populations. In evolution programs this premature loss of diversity frequently is tried to be prevented by fitness scaling. The fitness of superior individuals is reduced compared to that of inferior ones (here the fitness coefficient α would be changed to be closer to one). Clearly this does not remedy the cause but reduces the already low chances of single superior individuals to survive. Finally, in the deterministic model the selection pressure is only proportional to the logarithm of the fitness coefficient α.

Thus, to influence the selection process significantly, fitness scaling must be strong enough. Unfortunately the survival probability of a single superior individual can not be improved by an increasing population size. Thus to insure the survival, the selection mechanism has to be modified, e. g. by using *elitist selection*, which introduces some determinism. Summed up, in this situation the deterministic approximation of the process is not adequate. Yet the discrete and the continuous deterministic models resemble each other very closely. But the transformation of the difference to a differential equation does not simplify the solubility, since the differential equation has no explicit solution, which is the opposite one would expect. Consequently, the transformation is inadequate especially also for more complicated variants of this model.

The situation gets worst if mutation from individuals of the superior species to individuals of the worse one with probability u is added. Then the absorbing state corresponding to a uniform population of individuals of the superior species vanishes. If the fitness factor is high enough in relation to the mutation rate, then in the deterministic model there are two fixed points, one asymptotic stable caused by a balance of selection and mutation and one unstable corresponding to the remaining absorbing state, which is equivalent to a uniform population of individuals of the worse species. If the fitness factor is to low in relation to the mutation rate, the only stable fixed point is a uniform population of the worse species. But due to the properties of the Markov process, the expected value of the portion of the superior species converges very slowly to its extinction, whereas the first step analysis indicates a rapid absorption. Thus in this special case there is a contradiction not only between the deterministic and the Markov approach, but also within the latter one. At least, the deterministic model here is totally unsuitable.

Adding mutation also from individuals of the worse species to individuals of the superior one with probability v solves the unpleasant situation. Then in the Markov model there is a stationary distribution, to which the process converges, but no more absorbing states. Thus, the behavior of the process changes totally. Correspondingly, the deterministic model generally has a single asymptotic stable fixed point. If the individuals of both species exhibit equal fitness then the development of the expected portion of the Markov model and of the portion in the deterministic model are equal. The drift induced by the sampling errors only causes the final variance and an expected diversity different from the final diversity of the corresponding deterministic model. If the fitness differs, for small populations the stationary portions of the Markov model are lower than the deterministic fixed point and converge to it with increasing population size. Thus the deterministic model overestimates the stationary expected value of the portion for small populations and the drift obstructs the prevalence of the superior species in small populations in the presence of mutation.

Compared to the case with no mutation in the Markov model with different fitnesses the final expected portion in the case with mutation may also be higher although following equation 4.67 the effective selection probability is lower. This is again due to the missing extinction probability of the superior species. An example is the development of the expected value of the portion in figures 4.7 and 4.19. Both cases mentioned already earlier differ only by the mutation present in the second one.

Summed up, the mutation stabilizes the system by partly compensating the genetic drift and guaranties the prevalence of the superior species. If the population size is high and the mutation rates are very low, then the development of the expected value of the Markov

model with mutation also is approximated by the discrete deterministic model neglecting mutation, which could be solved explicitly. Since the selection probability with mutation $\tilde{p}_s$ is lower than without it as shown in equation (4.67), this approximation overestimates the expected value of the portion.

4.6.2 Multi Species Extension

In the preceding sections a very simplified model consisting only of two species has been explored. The main reason of the simplification has been to get a model simply enough to be tractable as a stochastic process. Fortunately most results can be transferred directly to the case of more than two species, which has been the starting point in section 4.2.

Considering only selection without mutation the extension back to more than two species results in a transition of the conditional distribution of the states from binomial to multinomial[5]. Then the conditional expected value of the frequency of species S_j is

$$\text{(4.95)}\quad \mathrm{E}\{\boldsymbol{n}_j(t+1)|n_1^{(l)}(t),\ldots,n_M^{(l)}(t)\} = p_s(S_j,l)\,N$$

corresponding to equation (4.11) and the conditional variance of the frequency of species S_j is

$$\text{(4.96)}\quad \sigma^2_{\boldsymbol{n}_j}(t+1)|_{n_1^{(l)}(t),\ldots,n_M^{(l)}(t)} = p_s(S_j,l)(1-p_s(S_j,l))\,N$$

corresponding to equation (4.12) for the two species case. In both equations the selection probability $p_s(S_j,l)$ in generation t is given by equation (4.2) depending on the current state l at generation t given by the frequencies $n_1^{(l)}(t),\ldots,n_M^{(l)}(t)$. The further proceeding is the same as for two species. Especially from equation (4.95)

$$\text{(4.97)}\quad \mathrm{E}\{\boldsymbol{n}_j(t+1)|n_1^{(l)}(t),\ldots,n_M^{(l)}(t)\} = \frac{n_j^{(l)}(t)F_j}{\sum_{k=1}^{M} n_k^{(l)}(t)F_k}\,N$$

follows. Consequently also here for large populations the actual portion $r_j(t+1)$ of a species S_j in generation $t+1$ can be approximated by the related conditional expected value, i. e.

$$\text{(4.98)}\quad r_j(t+1) \approx \mathrm{E}\{\boldsymbol{r}_j(t+1)|r_1(t),\ldots,r_M(t)\}.$$

holds corresponding to equation (4.18). Using this approximation one can construct a system of first order multidimensional difference equations, which behaves like its special case of two dimensions already discussed above [1].

If there is no mutation, the fittest species contained in the start population takes over the population, since one can think of the less fit species to be combined in one inferior species. If there are two or more equal fit species having the highest fitness value, the evolution will end up in a combination of them with portions depending on their start portions. Analogous to the two species case then this balance point will be structurally unstable. Finally, if the start portions are small enough, the development of the best

[5] In fact the binomial distribution is just a special case of the multinomial one.

species also shows a logistic shape. These statements can be proved easily by using multi-dimensional extensions of the techniques used in section 4.3 for the two species case. Also if mutation is taken into account, there is no principle difference to the two species case. If there is mutation from each species to each other, then there is exactly one fixed point, which is asymptotically stable. If there is a species, whose individuals can not be mutated, this results in a fixed point corresponding to a uniform population consisting only of that species, which may be asymptotically stable, if the selection pressure away from it is not strong enough.

The Markov model also does not change its behavior by extending the number of possible species beyond two. If there are absorbing states, the process finally will be absorbed in one of them with probabilities depending on the start state. If there is no absorbing state or class of states, then there also will be one stationary distribution of the states and the related expected value corresponds to the asymptotically stable fixed point of the approximating differential equation system. A general analysis of the properties of the Markov matrix including further special cases has been already proposed by Rudolph [87].

4.6.3 Schema Propagation

The models considered until now easily can be extended to calculate the propagation of a schema during the evolution. In the population a schema $\boldsymbol{W}$ is just the set of different species matching the related pattern. Clearly the probability to select an individual belonging to a schema then is the sum of the selection probabilities of the related species depending on the actual state l. Neglecting mutation from equation (4.2) the selection probability

$$p_s(\boldsymbol{W}, l) = \sum_{j \in \boldsymbol{W}} p_s(S_j, l) = \frac{\sum_{j \in \boldsymbol{W}} n_j^{(l)} F_j}{\sum_{k=1}^{M} n_k^{(l)} F_k} \tag{4.99}$$

follows, Combining this with equation (4.95) leads to the conditional expected value of the frequency

$$\mathrm{E}\{\boldsymbol{n}_{\boldsymbol{W}}(t+1) | n_1^{(l)}, \ldots, n_M^{(l)}\} = \frac{\sum_{j \in \boldsymbol{W}} n_j^{(l)} F_j}{\sum_{k=1}^{M} n_k^{(l)} F_k} N \tag{4.100}$$

of a schema $\boldsymbol{W}$ in the next generation $t+1$. Then proceeding analogously to equation (4.48) for the expected value

$$\begin{aligned} \mathrm{E}\{\boldsymbol{n}_{\boldsymbol{W}}(t+1)\} &= \sum_{l=1}^{v} \mathrm{E}\{\boldsymbol{n}_{\boldsymbol{W}}(t+1) | n_1^{(l)}(t), \ldots, n_M^{(l)}(t)\}\, p(n_1^{(l)}(t), \ldots, n_M^{(l)}(t)) \\ &= N \sum_{l=1}^{v} \frac{\sum_{j \in \boldsymbol{W}} n_j^{(l)} F_j}{\sum_{k=1}^{M} n_k^{(l)} F_k}\, p(n_1^{(l)}(t), \ldots, n_M^{(l)}(t)) \end{aligned} \tag{4.101}$$

follows, whereby the index l again denotes the actual state given by the related frequencies $n_1^{(l)}, \ldots, n_M^{(l)}$. The equation exactly denotes the expected value of the frequency of a

schema $\boldsymbol{W}$ in the next generation and corresponds to equation (4.48) for the two species case. Analogously it is not sufficient to know the expected value of the frequency of a schema in the current population to calculate the expected number in the next generation if all fitnesses are not equal. Thus here also a formula analogous to equation (4.51) generally is not possible. Instead the distribution of states in the current generation t is needed, which usually is not known. Adding mutation further complicates the whole approach, since it must be taken into account whether the source and the target species belong to the schema or not. But since the selection of a schema works analogously to that of a single species, its propagation can be treated like that and consequently be abstracted like in the two species model. Then also mutation can be considered, since there is only mutation away from and to the schema.

4.6.4 Comparison with Existing Models

Already Holland tried to support his conjecture of schema propagation by a theoretical model [46, p. 89*ff*]. It seems, that he has realized the stochastic nature of the sampling with replacement during the random selection, since he claims to estimate from below the "expected proportion" $P(\xi, t+1)$ of a schema ξ in generation $t+1$ depending on $P(\xi, t)$, which consequently is the "expected proportion" of schema ξ in generation t. This clearly contradicts the considerations of the last section and especially equation (4.101). In his formulas Holland did not take the sampling into account, which then he would have had to specify. In fact, although he claims to consider expected values Holland's formulas generally imply that he uses a deterministic approximation of the selection process and under this modification his formulas are correct. Finally, since Holland seems not to have been aware of these problems, he also did not show the applicability of his implicit approximation. As has been shown earlier, deterministic approximations may have solutions totally different from the related stochastic ones. Especially, all deterministic approximation he uses to estimate the portion of a superior schema in the next generation from below correspond to Markov models with an absorbing state related to extinction of the schema under consideration like in the case of selection against mutation presented above.

Similar problems are evident in the presentations of Holland's model in Bäck's [4, p. 123*ff*], Mitchell's [64, p. 27*ff*] and Michalewicz's [63, p. 43*ff*] textbooks, who altogether seem not have detected the problems of Holland's approach. They claim to consider expected values but do not even mention the sampling problem and in fact present deterministic approximations although they claim to consider expected values.

Goldberg [32, p. 30] picks up Holland's considerations and proposes them in a slightly different manner. He also claims to consider expected values but in fact presents a deterministic approximation of the selection process. Using the notation of section 4.2 he introduces the assumption

$$F_j = (1+\beta)\frac{\Phi}{N}$$

with $\beta > 0$ for a superior species S_j (or schema). From the above equation he infers an exponential propagation of the considered species or schema. Clearly, Goldberg's assumption implies an increase of the fitness of the species or schema S_j under consideration per generation due to the feedback of its increasing portion to the total fitness Φ, which also

is a function of the generation t. This complicates the model a lot, which to discuss is beyond the scope of the current consideration. In a later article together with Deb [33] he recognizes at least the logistic nature of the recursion for constant fitnesses but states the difference equation to hold for the expected value of the portion, which above already has been shown to be incorrect.

Also in later publications authors claim to consider the development of expected values of the portion of a superior species or schema while in fact they consider a deterministic approximation of the selection process, e. g. Wright [102] and Stephens and Waelbroeck [92].

Generally the models mentioned above and those based on them neglect the effects of sampling and the special properties of Markov chains. It is very important to consider the state as a stochastic variable depending on the generation and distinguish it from its actual value. Furthermore, its expected value has to be discriminated from the conditional one depending on some prerequisites. E. g. Reeves and Rove [83, p. 68] in fact calculate the *conditional* expected value of the number of descendants depending on the current state, although they call it the expected value. In the general case for a species or schema under consideration it is only possible to calculate that conditional expected value of its portion or frequency in the next generation assuming a given state in the current generation. To calculate the expected value the distribution of the state in the current generation has to be known, the expected value alone is not sufficient. For large populations, a deterministic approximation is possible, but its applicability has to be shown, which may be very difficult. If the number of states is small enough like in the cases proposed here, then a numerical calculation of the Markov model is possible, which can show the applicability of the approximation. At least Poli [76] recognized some of the problems mentioned above, but also did not take the properties of the Markov chains into account.

The above argumentation can be transfered analogously to the results of Rogers and Prügel-Bennett [86], who in fact also calculate the *conditional* expected sample variance of the fitness in the next generation using neutral selection depending on the current state. Thus their result corresponds to equation (4.60) due to equation (4.27), which denotes the relation between diversity and sample variance.

4.6.5 Consequences for Evolution Programs

In a real world evolution program during its run new species are created from the existing ones by applying crossover and mutation. In the models considered above this process is subsumed by mutation only. If the problem is complex enough it can be assumed that after its extinction a species is newly created only after a relative long time due to low mutation rates. This behaviors can be characterized by calling the related class of states “temporary absorbing”. Consequently, if the mutation rates are low enough for the introduction of a new species the proposed stochastic two species model neglecting mutation can be used, which overestimates the survival respectively takeover probability somewhat. Generally, the random fitness proportional selection process leads to an almost uniform population, which in the model shows up by a low final expected species diversity. The survival of single individuals of superior species can not be assured and especially after reaching an almost uniform population single individuals of new but only slightly superior species have only low chances to survive and to spread. Then drift is the dominating process. The same considerations can be transferred to the propagation of a schema. Consequently, it can be

expected that in a real world evolution program after a sufficiently number of generations one schema is dominating and the individuals are very similar. Thus Holland's schema propagation conjecture is correct. Without mutation the takeover probability of a new superior individual is very low. Consistently with mutation in reality it may take a lot of time until the best species dominates the population, because the may be very long periods of its extinction.

The question is, whether a deterministic fitness proportional selection in evolutionary algorithms would be better suited than a random one. Especially in small populations the deterministic selection combined with mutation causes a higher final portion of the superior species. Furthermore the running time of the algorithms may be shortened, since the time consuming stochastic sampling is omitted. To ensure the survival of a superior species over the time at least some elitism has to be introduced, as already has been shown by Rudolph [87].

4.7 Conclusions

Deterministic models in many cases are not suitable to model the random selection, because they neglect the genetic drift and the probability of extinction. If mutation and selection work against each other without mutation the deterministic model pretends a development of the portion, that is totally different from that of the expected value of the the portion in the Markov model. Additionally, then the first step analysis result contradicts the slow convergence of the indices. Generally here mutation stabilizes the selection process and compensates the genetic drift at least partly. Then a deterministic approximation for large populations is possible. Summed up, one first has to compare the different models and only then can decide, whether the deterministic model is sufficiently accurate or not.

The considered two species model can be extended by additional species. Then the conditional distribution of states changes from binomial to multinomial, the number of states increases extremely and the situation becomes more complex than necessary but the approach and the results do not change. But especially the logistic growth of the portion of the best species in the deterministic model caused by the negative feedback of the propagation in this case is difficult to discover [13, p. 29f]. These problems have been the main reasons to limit the consideration to a simple but tractable model.

The propagation of a schema in a population behaves much like that of a single species and thus deserves no special treatment. It is not possible to calculate the expected value of the portion of a a species or schema in the next generation depending solely on that in the current generation as many authors claim. In fact their models are deterministic approximations although not stated, whereas they left open the applicability of the approximation.

Finally an approximation by a diffusion model is possible. This approach is already used in population genetics and leads to explicit solutions [25].

Chapter 5

Population State Indices

5.1 Introduction

In the last chapter theoretical models of schema propagation have been considered very deeply. Although they are very simple compared to what is happening in a real evolution program, it can be conjectured, that a similar propagation process happens there. As already stated in the beginning, to observe schema propagation in experiments, the reach of the schemata has to be measured. But until now this subject has not been payed a lot of attention to and there are no suitable indices. Thus although a large number of experiments has been performed already, there are hardly observations available that support Holland's conjecture because of the missing suitable indices.

Consequently indices have to be found to monitor a possible schema propagation in the population during the run of an evolution program. This task can be extended to generally find suitable indices to describe the state of a population. Until now usually with some exceptions only the fitness of the best individual, the average fitness and sometimes the variance of the fitness of the current population is recorded, which all provide not enough information. Since in ecology there are similar problems, the indices used there have to be considered for their applicability and suitability for evolution programs and the differences between both situations have to be explored.

5.2 Convergence versus Improvement and Assimilation

5.2.1 Convergence

During the execution of an evolution program the objective values of the individuals should approach the optimum. To measure the distance to the optimal objective value, the difference between the best objective value or the average objective value of the population and the optimal objective value can be used. This works well, if the optimal solution is known, i. e. the evolution program runs on a test problem instance, which already has been solved by an algorithm that calculates a provable optimal solution. Furthermore, during the run of the evolution program there may be stages of deterioration after which the improvement of the objective values respectively fitnesses is resurrected. This causes a jagged output

when plotting the graph. One can interpret the jags to be noise which is superposed on the "true" improvement process. If the distance to the optimal objective value approaches zero, then this is normally called convergence. Here this will be referred to as *convergence in objective*. In contrast to the strict definition of convergence in analysis or in stochastic processes, the definition given here is very sloppy.

To transfer the considerations above related to the objective value also to the encoding and problem domain, first a distance function has to be defined on both. The notion of distance is an important topic and can be defined in different ways. In the following the approach of Bergmann and Richter [85, Chapter 6] is proposed with some supplements.

Considering a set $\mathbf{M}$ of arbitrary elements a distance function $\Delta : \mathbf{M} \times \mathbf{M} \mapsto \mathbf{R}^+$ is defined to hold the following properties:

Minimality: $\forall x, y \in \mathbf{M} \quad \Delta(x, y) \geq \Delta(y, y)$

Symmetry: $\forall x, y \in \mathbf{M} \quad \Delta(x, y) = \Delta(y, x)$

Reflexivity: $\forall x, y \in \mathbf{M} \quad \Delta(x, x) = \Delta(y, y)$

It is called a *metric*[1], if the additional properties

Identity: $\forall x, y \in \mathbf{M} \quad \Delta(x, y) = 0 \quad \text{iff} \quad x = y$

Triangle Inequality: $\forall x, y, z \in \mathbf{M} \quad \Delta(x, y) + \Delta(y, z) \geq \Delta(x, z)$

hold. Alternatively, there are other requirements possible based on psychological models of similarity, which subsume the properties claimed above [96].

The above model can be transferred to the genotype level of an artificial population $\boldsymbol{P}$ and equivalently on the problem domain. Then the distance function can be separated into two components. First, there are local distance functions $\delta_j : \mathbf{A}_j^2 \mapsto \mathbf{R}$ for each attribute index or locus j of two individuals I_k, I_l, which depend on the attribute type. Second, an amalgamation (or constructor) function $F : \mathbf{R}^M \mapsto \mathbf{R}$ calculates the global distance $\Delta(x, y) = F(\delta_1(I_k, I_l), \ldots, \delta_M(I_k, I_l))$ depending on the local distances $\delta_j(I_k, I_l)$. To conform to the above requirements, F has to be monotonous in each argument, i. e. an increase of a single argument does not decrease the function value, and the constraint $F(0, \ldots, 0) = 0$ has to be hold.

Both in artificial populations and in genetics it is only appropriate to decide whether the values of an attribute respectively the alleles of a gene of two individuals are equal or not [80, p. 27], because it is not possible to establish an order or scale. Thus the local distance of a nominal scaled attribute is defined to be

$$\delta_j(I_k, I_l) = \begin{cases} 0 : A_{k,j} = A_{l,j} \\ 1 : A_{k,j} \neq A_{l,j}. \end{cases} \tag{5.1}$$

Since each attribute has the same importance, F can be defined to be the sum of the local distances, resulting in the global distance function

$$\Delta(I_k, I_l) = F(\delta_1(I_k, I_l), \ldots, \delta_M(I_k, I_l)) = \sum_{j=1}^{M} \delta_j(I_k, I_l). \tag{5.2}$$

[1] The Greek word $\mu\acute{\varepsilon}\tau\varrho o\nu$ in fact means measure. Thus a distance function holding not all of the requirements is better called an index.

This is in fact the *Hamming distance* counting simply the number of different attribute values and establishing a metric on the attribute space, which easily can be verified using the properties presented above. Standardized on the number of attributes M it is called *generalized M-coefficient* in multivariate statistics [26, p. 446]. If the generalized M-coefficient of two individuals is close to zero, both are called to be *similar*[2].

Now using the Hamming distance the distance between the genotype of an individual to a genotype, that represents an optimal solution, can be calculated or its population average. Again this is only possible, if an optimal solution is known and also can be done on the problem domain level. The approach of the average distance between the genotypes and an optimal solution genotype to zero is referred here as *convergence in domain*. Further problems arise, if there is more than one optimal solution, and as for the convergence in objective, the observed convergence process may have stages of deterioration and resurgence.

Convergence in objective and convergence in domain generally may not imply each other. Especially for hard problems there may be admissible solutions with objective values close to the optimum value but with a high distance on the genotype level to the representation of an optimum. Conversely, there may be admissible solutions very close to an optimal solution on the genotype level but with a high distance between their objective values.

5.2.2 Improvement and Assimilation

If an evolution program approaches a problem which has no known optimal solutions, the convergence criteria proposed above can not be applied. Then other criteria have to be taken into account to characterize the state of the evolution, which are of general interest also.

Usually the smoothed proceeding of the objective value of the best individual and the average objective value of the population shows an improvement with increasing generation number during the evolution. Thus the approach to zero of the gradient of the best or average objective value with respect to the generation number, i. e. the end of the improvement process, can be taken as an indicator of "convergence". To not confuse this with the definition of convergence in objective, it will be named *saturation in objective*. Furthermore the variance of the objective values of the individuals in the population can be monitored during the run.

Unfortunately the criterion introduced above can not be transferred directly to the genotype or solution level. However the average or mean distance

$$D^G(\boldsymbol{P}) = \frac{1}{N(N-1)} \sum_{k=1}^{N} \sum_{l=1}^{N} \Delta(I_k, I_l), \tag{5.3}$$

[2]There may be situations, in which the approach presented above is not appropriate and a more detailed rating depending on both values is necessary. Clearly, to set the distance of an attribute value to all other values with the value one is a severe restriction.

between the individuals of the population can be calculated. Equivalently, the different classes of equivalent genotypes can be considered. Having n different classes and a class i has N_i specimens this results in

$$D^G(\boldsymbol{P}) = \frac{1}{N(N-1)} \sum_{k=1}^{n} \sum_{l=1}^{n} N_k N_l \Delta(S_k, S_l) \tag{5.4}$$

Incorporating the additive distance function defined in equation (5.2) gets

$$D^G(\boldsymbol{P}) = \frac{1}{N(N-1)} \sum_{j=1}^{M} \sum_{k=1}^{n} \sum_{l=1}^{n} N_k N_l \delta_j(S_k, S_l). \tag{5.5}$$

The value of an attribute may be the same for more than one class of genotypes, thus each attribute can be considered separately. Using the local distance function of equation (5.1) for nominal attributes, each of the $N_{j,k}$ individuals having attribute value $a_{j,k}$ at position or locus j has the local distance of one to the remaining $N - N_{j,k}$ individuals having a different value. From this relation the equation

$$\sum_{l=1}^{n} N_{j,k} N_{j,l} \delta_j(S_k, S_l) = N_{j,k}(N - N_{j,k})$$

follows and using that from equation (5.5) the *average Hamming distance*

$$D^H(\boldsymbol{P}) = \frac{1}{N(N-1)} \sum_{j=1}^{M} \sum_{k=1}^{n_j} N_{j,k}(N - N_{j,k}) \tag{5.6}$$

of the population $\boldsymbol{P}$. Here n_j denotes the number of different possible allele at locus j. With the definition of the relative frequency

$$r_{j,k} = \frac{N_{j,k}}{N}$$

and some simple transformations

$$D^H(\boldsymbol{P}) = \frac{N}{N-1} \sum_{j=1}^{M} \left(1 - \sum_{k=1}^{n_j} r_{j,k}^2 \right) \tag{5.7}$$

is obtained finally. Divided by the number of attributes respectively loci the *average generalized M-coefficient*

$$d^M(\boldsymbol{P}) = \frac{N}{M(N-1)} \sum_{j=1}^{M} \left(1 - \sum_{k=1}^{n_j} r_{j,k}^2 \right)$$

between zero and one follows.

An important property of the average Hamming distance $D^H(\boldsymbol{P})$ is concavity. Thus if there are two populations $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ of size N_1 and N_2, then with

$$\lambda = \frac{N_1}{N_1 + N_2} \tag{5.8}$$

the inequality

$$D^H(\boldsymbol{P}_1 \cup \boldsymbol{P}_2) \geq \lambda D^H(\boldsymbol{P}_1) + (1-\lambda) D^H(\boldsymbol{P}_2) \tag{5.9}$$

holds. Since of the additivity and nonnegativity with respect to the loci it is sufficient to show concavity for only a single attribute. Then from equation (5.6) for the population $\boldsymbol{P}_i$

$$D^H(\boldsymbol{P}_i) = \frac{1}{N_i(N_i-1)} \sum_{k=1}^{n} N_{i,k}(N - N_{i,k})$$

follows and with

$$N_i = \sum_{k=1}^{n} N_{i,k}$$

for their union $\boldsymbol{P}$

$$D^H(\boldsymbol{P}) = \frac{1}{(N_1+N_2)(N_1+N_2-1)} \sum_{k=1}^{n} (N_{1,k}+N_{2,k})(N - N_{1,k} - N_{2,k}).$$

By insertion of these identities and some transformations equation (5.9) finally is proved.

Knowing the average Hamming distance may not be sufficient, because it would be interesting to calculate a virtual individual, having the same minimal distance to all individuals of the population. Unfortunately, this problem is NP-hard [43]. Thus it is intractable to determine the current center of representation during the run of an evolution program, even if it is not claimed to be an admissible solution.

Summarized, the mean distance and the average generalized M-coefficient are indices of the *dissimilarity* of a population, and if they decreases during the run of an evolution program, the individuals become more similar on the average and the whole process can be called *assimilation*. The reduction of the gradient of the average Hamming distance with respect to the generation number is called here *saturation in similarity*.

Besides the indices proposed until now, other indices may be used to monitor the state of a population. One important index is *diversity*, which already plays an important role in ecology and population genetics but also can be used for evolution programs. Thus in the next section it is discussed in detail.

5.3 Diversity

5.3.1 Definition

In the context of evolutionary computation diversity is rarely defined even verbally, although many authors use the term extensively, e. g. Burke, Gustafson and Kendall [14]. In ecology it is related primarily to the species found in an ecological community. To elucidate the term more deeply first some desirable properties of a diversity index are considered.

According to Hurlbert [48], species diversity should be both a function of the number of species present (*species richness* or *species abundance*) and the evenness of the distribution of the individuals into these species (*species evenness* or *species equitability*). Furthermore a diversity index should not presume a specific distribution and be without any external

parameters. Thus a species diversity $D(\boldsymbol{P})$ generally is a function only of the frequencies of the species or the number of individuals and the relative frequencies of the species

$$D(\boldsymbol{P}) = D(N_1, \ldots, N_n) = D(N, r_1, \ldots, r_{n-1}). \tag{5.10}$$

These sloppy claims can be refined to the following ones:

Minimality: The diversity has to be zero iff all individuals belong to the same species.

Maximality: For a given number of individuals N and a given number of species s, diversity reaches its maximum, if the distribution of the individuals into the species is as even as possible. That is, if N is a multiple of s, when $N_i = N/s = m$ for all i, or, if $N = s\,m' + r$ (with m' and r whole numbers and $r < s$), when $s - r$ of the values are represented by m' individuals each and the remaining r values by $m' + 1$ individuals each [73, p. 372]. This in fact is equivalent to the Pigou-Dalton transfer principle [2, p. 5].

Concavity: If there are two populations $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ of size N_1 and N_2, then the inequality

$$D(\boldsymbol{P}_1 \cup \boldsymbol{P}_2) \geq \lambda D(\boldsymbol{P}_1) + (1 - \lambda) D(\boldsymbol{P}_2)$$

holds with

$$\lambda = \frac{N_1}{N_1 + N_2}.$$

Monotony: For a given number of individuals N and a given number of species n the diversity increases, if a single species S_i is replaced by m different new species S'_k, so that $\sum_{k=1}^{m} N'_k = N_i$, $N'_k > 0$ and $n' = n - 1 + m$ is the new number of possible species. Thus, the diversity increases as the classification is refined.

Insensitivity to non present species: The diversity is unchanged if new species are introduced but none of the individuals in the population actually belongs to one of them.

Invariance under cloning: For given relative frequencies $r_j = N_j/N$ the diversity is not a function of the population size N. Usually this property is called replication axiom [53, p. 110]. It is not intuitively reasonable for many people, i. e. Pielou regards a small population less diverse than a large one [73, p. 373].

Symmetry: For each permutation of the attachment of the given frequencies $N_1, \ldots, N_n$ to the species $S_1, \ldots, S_n$ of the population $\boldsymbol{P}$ the diversity $D(\boldsymbol{P})$ remains unchanged.

There are further ingenious requirements and there is an ongoing discussion which of them have to hold [48, 72, 80][3].

[3] If symmetry and concavity hold for a diversity function, then it is also *Schur-concave* [53, p. 107], i. e. from $D(\boldsymbol{P}_1) \leq D(\boldsymbol{P}_2)$ follows that $\boldsymbol{P}_1$ majors $\boldsymbol{P}_2$ [2, p. 16].

5.3.2 Indices

Based on the above requirements now some diversity indices already proposed by various authors are considered.

Often the diversity of a population $\boldsymbol{P}$ is synonymous with the number s of different species found in it. Forming a diversity index leads to

$$D^{NC}(\boldsymbol{P}) = s - 1 = \frac{1}{N}\sum_{j=1}^{n} N - N_j$$

with N_j being the frequency of species j and n the number of possible species. The index can be transformed easily to the coefficient

$$d^{NC}(\boldsymbol{P}) = \frac{D^{NC}(\boldsymbol{P})}{N-1} = \frac{1}{N(N-1)}\sum_{j=1}^{n}(N - N_j)$$

between zero and one. The diversity index D^{NC} satisfies all claims except maximality because it does not depend on the evenness of the distribution.

Simpson [90] proposed a parameterless index of *concentration*, the opposite of diversity,

$$C(\boldsymbol{P}) = \frac{1}{N^2}\sum_{j=1}^{n} N_j^2 = \sum_{j=1}^{n} r_j^2,$$

which has been proposed independently also by Herfindahl [42]. Using it, an index of diversity

$$D^{GS}(\boldsymbol{P}) = 1 - C(\boldsymbol{P}) = 1 - \frac{1}{N^2}\sum_{j=1}^{n} N_j^2 = 1 - \sum_{j=1}^{n} r_j^2 = \frac{1}{N^2}\sum_{j=1}^{n} N_j(N - N_j) \tag{5.11}$$

can be defined [81]. Since it already has been proposed earlier by Gini [31], it is usually called Gini-Simpson index. From that a coefficient

$$d^{GS}(\boldsymbol{P}) = \frac{N}{N-1}D^{GS}(\boldsymbol{P}) = \frac{1}{N(N-1)}\sum_{j=1}^{n} N_j(N - N_j)$$

between zero and one can be derived, which is exactly the average distance of a single locus j in equation (5.6). By using the Gini-Simpson-index the species diversity can be interpreted to be the average distance between two individuals drawn independently at random with replacement, at which the distance between two species is set always to one analogously to the definition in equation (5.1) for each attribute. As already stated there, this is a very severe restriction. But the Gini-Simpson-index D^{GS} satisfies all of the above claims which is easy to prove.

Rao [80, p. 25*ff*] generalized the concept of diversity to be an average distance to the definition

$$D^{D}(\boldsymbol{P}) = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\Delta(I_k, I_l). \tag{5.12}$$

Then using the Hamming distance of equation (5.2) analogously to equation (5.7) the formula

$$D^{RH}(\boldsymbol{P}) = \frac{1}{N^2}\sum_{j=1}^{M}\sum_{k=1}^{n_j} N_{j,k}(N - N_{j,k}) = \sum_{j=1}^{M}\left(1 - \sum_{k=1}^{n_j} r_{j,k}^2\right) \tag{5.13}$$

follows. Clearly, this is the sum of the Gini-Simpson-indices of the M attributes.

Generally it is an important question, which properties have to hold for the distance function $\Delta(I_k, I_l)$ to insure at least concavity. It is not sufficient to claim the distance function to be a metric if there are more then four distinctive species [79, p. 72]. The requirements can be further extended to form Rao's *quadratic entropy*, from which finally a metric on a set of populations can be established [82].

Finally Shannon's entropy H [89] can be used as a diversity index by substituting the probabilities p_j of the different alternative outcomes by their relative frequencies r_j [59, 74]. Then the definition

$$D^S(\boldsymbol{P}) = -\sum_{j=1}^{n} r_j \ln r_j = -\sum_{j=1}^{n} \frac{N_j}{N} \ln \frac{N_j}{N} \tag{5.14}$$

follows[4]. The Shannon diversity D^S satisfies all properties claimed above [67] and easily the corresponding Shannon diversity coefficient

$$d^S(\boldsymbol{P}) = \frac{D^S}{\ln N} = -\frac{1}{\ln N}\sum_{j=1}^{n} r_j \ln r_j$$

between zero and one can be deduced, because D^S gets maximum for $r_j = 1/N$ for all j, from which

$$\max D^S(\boldsymbol{P}) = -\sum_{j=1}^{n} \frac{1}{N} \ln \frac{1}{N} = \ln N \tag{5.15}$$

follows.

Additionally an estimation from below [53, p. 100*ff*] can be performed. Using the inequality

$$-\ln x \geq 1 - x$$

from equation (5.14) the inequality

$$D^S(\boldsymbol{P}) = -\sum_{j=1}^{n} r_j \ln r_j \geq \sum_{j=1}^{n} r_j(1 - r_j) = 1 - \sum_{j=1}^{n} r_j^2 \tag{5.16}$$

follows, where the right side equals the Gini-Simpson-index D^{GS} of equation (5.11).

In ecology the properties of a population usually have to be estimated from a small sample. I. e. the whole population can not be censused, but only a random subset can be observed and thus estimators for the proposed indices and coefficients have to be used, which often are hard to find. Fortunately, in the field of evolutionary computation this is not a problem because here a population is always censused.

[4] In this work the natural logarithm ln is used. In information theory the base usually is two, because there digital signals are considered. Since a transformation between different bases is equivalent to a multiplication by a constant factor, the base only influences the numerical values.

5.3.3 Diversity Indices in Evolution Programs

The above indices related to species in ecology also can be applied to evolution programs and genetic algorithms. To do so, each different class of artificial genotypes is treated analogously to a natural species and vice versa. Thus a species is equivalent to a class in the following. Consequently, the species diversity $D_S(\boldsymbol{P})$ of an artificial population $\boldsymbol{P}$ is defined to be

$$D_S(\boldsymbol{P}) = D(N_1, N_2, \cdots, N_n) \tag{5.17}$$

with N_i being the number of individuals belonging to class or species i and n the number of possible different classes or species, which are given by the combinations of the M attributes or loci. Furthermore both a distance based diversity index like the Gini-Simpson-index D^{GS} or the entropy based D^S can be used as specialization of the general diversity index D.

The species diversity is required for a successful evolution and determines several properties of the selection process.

In the Verhulst-Pearl differential equation (4.25), which approximates the selection process between two species deterministically and continuously, the selection pressure is proportional to the species diversity given by the Gini-Simpson-index, which has been shown in equation (4.26). The rate of change is maximum, if the species diversity of the population is also maximum, which is an analogy to the second part of Fisher's fundamental theorem [25, p. 14]. Thus beneath the logarithm of the fitness factor α, which denotes the relation of the fitnesses, the diversity is the moving force that causes selection pressure.

Until now the species diversity usually is not evaluated. The only exception known by the author is Hatjimihail's work [41].

But a diversity index can be applied not only with respect to the species respectively class of genotype affiliation but also to each separate attribute respectively locus. Consequently the population diversity $D_P(\boldsymbol{P})$ of a population $\boldsymbol{P}$ consisting of individuals having genotypes with M loci is defined to be

$$D_P(\boldsymbol{P}) = \sum_{j=1}^{M} D(N_{j,1}, \ldots, N_{j,i}, \ldots, N_{j,n_j}), \tag{5.18}$$

where $N_{j,i}$ is the number of individuals having attribute value respectively allele a_i at locus j. I. e. it is the sum of the attribute diversities of each locus j. This corresponds to the average distance of equation (5.7) and the distance based diversity of a population using the Gini-Simpson-index in equation (5.13). Clearly, all these quantities consider the attributes independently from each other. Thus the population diversity also can be called the *genetic potential* of the population.

The population diversity also is an important factor for a successful evolution, because an operative crossover requires differing individuals. Without mutation the offspring individuals can differ from the parent individuals only as much as these differ themselves. If the parent individuals respectively their genotypes are very similar, the resulting offspring individuals also are very similar to their parents and mutation remains to be the only force to explore new solution alternatives. The population diversity calculated using the Gini-Simpson-index is a suitable index to measure the dissimilarity, because it equals Rao's generalized average distance of equation (5.13). But that in fact is the average distance of

two individuals drawn randomly with replacement from the population, which is just the method used by many selection algorithms. Since following inequality (5.16) the Shannon diversity D^S bounds the Gini-Simpson-index D^{GS} from above, also using Shannon's index the population diversity indicates the average dissimilarity analogously to Rao's generalized average distance.

Thus generally a decrease of the population diversity during the run of an evolution program indicates assimilation and the reduction of the gradient of the population diversity with respect to the generation number can also be called *saturation in similarity* analogously to the terminology for the objective value. If an assimilation happens before near optimum solutions have been found respectively the improvement of the objective values is nearly finished, this usually is called *premature convergence* in genetic algorithms and evolution programs [32, p. 73*ff*]. To avoid confusion with the convergence in objective discussed earlier, it is suggested to label this phenomenon *premature assimilation*, which is more appropriate. It should be avoided, because during the search the whole domain of each locus should be "covered" by the alleles to explore the whole search space not only by mutation. Consequently the population diversity should be as high as possible at least in the initial stage.

The average dissimilarity indicated by the population diversity or the average distance is related also to the appearance of schemata. To remember from above, if some individuals belong to a schema, then they share some alleles at certain positions and form a subset. Thus within a schema the distance between the related individuals is lower than the maximum and consequently the average generalized M-coefficient is lower than one. If a schema spreads in the population, then the average distance and the average generalized M-coefficient decrease indicating an assimilation. On the other hand assimilation clearly reduces the number of schemata present in the population. If the N individuals differ pairwise in each of the M loci, then there are $(N+1)^M - M + 1$ different schemata, whereas there are only 2^M schemata, if the population is totally uniform, i. e. there is only one species.

The population diversity or deduced indices already have been used sometimes in evolutionary computation to keep track of what is happening during the run of an evolution program. Grefenstette [36] has evaluated an index similar to the population diversity based on Shannon's formula normalized by the number of loci M, i. e. in fact an index equivalent to the average diversity per locus respectively attribute. Mori et al. [66] have introduced a similar formula, although their approach was inspired by the role of entropy in thermodynamics. They also try to preserve the population diversity by using special operations. Generally both approaches use Shannon's formula as diversity indices without any justification, explanation or references to the vast literature in mathematical biology.

5.3.4 Limitations

As stated above in the population diversity indices proposed until now each attribute is evaluated independently. But there may be *dependences* between attribute values. Then considering a randomly selected individual from the values of some attributes the values of some of the remaining attributes can be determined independently from its species affiliation respectively the uncertainty about their values is reduced. In genetics such a relation is called a *linkage*. This property also is related closely to the properties of

a schema. If a schema covers more than one species, which generally is not required by the definition of a schema, then from the fixed attribute values at certain positions neither the values of the remaining attributes not constituting the schema nor the species affiliation can be determined. To the extreme, if in a population schemata have spread maximally among the species, then all attributes have to be evaluated to determine the species respectively class affiliation of an individual. On the other hand, if the evaluation of any single attribute is sufficient to determine also the values of the remaining attributes and the species affiliation, then no schema has spread across different species. Using only the population diversity, that treats the attributes respectively loci independently, a dependence and thus schema propagation among different species or classes of genotypes can not be detected. Consequently, an index is needed, which is capable of that.

Since different species respectively classes of genotypes are distinguished by their different attribute values respectively alleles, there must be a relation between the species diversity and the diversities of the attributes shaping the species respectively classes of genotypes in artificial populations, whose sum has been defined to be the population diversity. Until now this relation has not been explored, that therefor has to be also performed. This is not an issue in ecology, because there a species is determined by more than a limited set of attributes.

5.4 Relation between Diversities versus Dependence

5.4.1 The Relation of the Diversities for Two Attributes

First the relation between species and population diversity is explored. Recapitulating the considerations of section 3.6 in artificial populations each combination of attributes forms a class or species. Thus an individual or object is classified into a species by evaluation of its attribute values. For the sake of simplicity first the consideration is limited to a population $\boldsymbol{P}$ of N objects having two attributes A_1 and A_2 of the domains $\mathbf{A}_1 = \{a_{1,1}, \ldots, a_{1,n_1}\}$ and $\mathbf{A}_2 = \{a_{2,1}, \ldots, a_{2,n_2}\}$. Then $N_{i,j}$ is the number of individuals having value $a_{1,i}$ of attribute A_1 and value $a_{2,j}$ of attribute A_2 together, which forms the class or species $S_{i,j}$. Conforming to the definition of equation (5.10) the species diversity $D_S(A_1A_2)$ of the population is defined to be

$$D_S(\boldsymbol{P}) = D(A_1A_2) = D(N_{1,1}, N_{1,2}, \ldots, N_{n_1,n_2-1}, N_{n_1,n_2}). \tag{5.19}$$

The number of individuals having value $a_{1,i}$ of attribute A_1 is

$$N_{i,*} = \sum_{j=1}^{n_2} N_{i,j}$$

and the number of individuals having value $a_{2,j}$ of attribute A_2 is

$$N_{*,j} = \sum_{i=1}^{n_1} N_{i,j}.$$

Using these definitions the attribute diversities

$$D(A_1) = D(N_{1,*}, \ldots, N_{n_1,*}) \quad \text{and} \quad D(A_2) = D(N_{*,1}, \ldots, N_{*,n_2}) \tag{5.20}$$

are defined analogously. Furthermore the diversity of attribute A_1 under the condition $A_2 = a_{2,j}$ is

$$D(A_1|A_2 = a_{2,j}) = D(N_{1,j}, \ldots, N_{n_1,j}). \tag{5.21}$$

A corresponding formula can be obtained for $D(A_2|A_1 = a_{1,i})$. Using equation (5.21) the conditional diversity $D(A_2|A_1)$ is defined to be the average of $D(A_1|A_2 = a_{2,j})$ over all $a_{2,j}$, i. e.

$$D(A_1|A_2) = \sum_{j=1}^{n_2} \frac{N_{*,j}}{N} D(A_1|A_2 = a_{2,j}).$$

Again a corresponding formula for $D(A_2|A_1)$ can be obtained. Now using the above definitions the relation

$$D(A_1A_2) = D(A_1) + D(A_2|A_1) = D(A_1|A_2) + D(A_2) \tag{5.22}$$

can be proved for using the Shannon diversity D^S as diversity function D [71, p. 551]. Unfortunately this relation is generally not true for the Gini-Simpson-index D^{GS}. From its definition in equation (5.11) follows

$$\begin{aligned}
D^{GS}(A_1A_2) &= 1 - \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \left(\frac{N_{i,j}}{N}\right)^2 \\
&= 1 - \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \left(\frac{N_{i,j}}{N_{*,j}}\right)^2 \left(\frac{N_{*,j}}{N}\right)^2 \\
&= 1 - \sum_{j=1}^{n_2} \left(\frac{N_{*,j}}{N}\right)^2 \sum_{i=1}^{n_1} \left(\frac{N_{i,j}}{N_{*,j}}\right)^2 + \sum_{j=1}^{n_2} \left(\frac{N_{*,j}}{N}\right)^2 - \sum_{j=1}^{n_2} \left(\frac{N_{*,j}}{N}\right)^2 \\
&= 1 - \sum_{j=1}^{n_2} \left(\frac{N_{*,j}}{N}\right)^2 + \sum_{j=1}^{n_2} \left(\frac{N_{*,j}}{N}\right)^2 \left(1 - \sum_{i=1}^{n_1} \left(\frac{N_{i,j}}{N_{*,j}}\right)^2\right) \qquad (5.23) \\
&= D^{GS}(A_2) + \sum_{j=1}^{n_2} \left(\frac{N_{*,j}}{N}\right)^2 D^{GS}(A_1|A_2 = a_{2,j}) \qquad (5.24) \\
&\neq D^{GS}(A_2) + D^{GS}(A_1|A_2).
\end{aligned}$$

unless

$$\sum_{j=1}^{n_2} \left(\frac{N_{*,j}}{N}\right)^2 = 1 \tag{5.25}$$

holds. This condition can only be satisfied by claiming $D^{GS}(A_2) = 0$. Consequently using the Gini-Simpson-index it is not possible to specify the relation between the species diversity and the diversities of the attributes generally. This a very disenchant result, which is reemphasized later in section 5.4.2.

For the Shannon diversity from equation (5.22) the inequalities

$$D^S(A_1) \leq D^S(A_1A_2) \quad \text{and} \quad D^S(A_2) \leq D^S(A_1A_2) \tag{5.26}$$

follow, because diversities are nonnegative. Furthermore, $D^S(A_1|A_2) \leq D^S(A_1)$ respectively $D^S(A_2|A_1) \leq D^S(A_2)$ hold, because conditioning can not increase diversity. Thus, the inequalities

$$0 \leq D^S(A_1|A_2) \leq D^S(A_1) \leq D^S(A_1A_2) \tag{5.27}$$

and

$$0 \leq D^S(A_2|A_1) \leq D^S(A_2) \leq D^S(A_1A_2) \tag{5.28}$$

are obtained. Combined with equation (5.22) the inequality

$$\max(D^S(A_1), D^S(A_2)) \leq D^S(A_1A_2) \leq D^S(A_1) + D^S(A_2) \leq 2\, D^S(A_1A_2) \tag{5.29}$$

follows. Thus the sum of the attribute diversities is the maximum obtainable species diversity. But itself it is limited by the twofold of the species diversity.

5.4.2 Statistical Independence of Attributes

After considering the relation between species and population diversity now the dependence of the attribute values is explored. To do so a special situation needs some additional consideration.

If the relative frequencies of the attribute values satisfy the condition

$$r_{i,j} = r_{i,*}\, r_{*,j} \tag{5.30}$$

or equivalently the absolute frequencies the condition

$$N_{i,j} = \frac{N_{i,*}\, N_{*,j}}{N}, \tag{5.31}$$

then for the Shannon diversity D^S the equation

$$D^S(A_1A_2) = D^S(A_1) + D^S(A_2) \tag{5.32}$$

holds analogously to the relation for the entropies of two independent stochastic variables [71, p. 551]. Therefore property (5.31) is called *statistical independence* here. Equivalently

$$D^S(A_2|A_1) = D^S(A_2) = D^S(A_2|A_1 = a_i) \tag{5.33}$$

holds for all i, because from condition (5.31) for the conditional relative frequencies

$$\frac{N_{i,j}}{N_{i,*}} = \frac{N_{*,j}}{N} = r_{*,j} = \text{const} \tag{5.34}$$

follows for each i with $N_{i,*} > 0$. Corresponding relations can be obtained for $D^S(A_1|A_2)$. Then knowing one attribute does not reduce the diversity of the second one and thus gives no advantage for estimating the second attribute. A special case of the above situation arises if one attribute is the same for all individuals, e. g. $A_1 = a_{1,i}$. Then $N_{i,*} = N$ follows for that i and from that $D^S(A_1) = 0$ and $D^S(A_1A_2) = D^S(A_2)$. Consequently an attribute

having no diversity is always statistically independent to the remaining attribute and vice versa.

Although equation (5.33) also holds for the Gini-Simpson-index D^{GS} in case of statistical independence of the attributes, equation (5.32) generally does not, because equation (5.23) and (5.24) imply

$$D^{GS}(A_1A_2) = D^{GS}(A_2) + \sum_{j=1}^{n_2}\left(\frac{N_{*,j}}{N}\right)^2\left(1 - \sum_{i=1}^{n_1}\left(\frac{N_{i,j}}{N_{*,j}}\right)^2\right)$$

and with equation (5.31)

$$\begin{aligned} D^{GS}(A_1A_2) &= D^{GS}(A_2) + \sum_{j=1}^{n_2}\left(\frac{N_{*,j}}{N}\right)^2\left(1 - \sum_{i=1}^{n_1}\left(\frac{N_{*,j}}{N}\right)^2\right) \\ &= D^{GS}(A_2) + \sum_{j=1}^{n_2}\left(\frac{N_{*,j}}{N}\right)^2 D^{GS}(A_1) \\ &\neq D^{GS}(A_2) + D^{GS}(A_1) \end{aligned}$$

unless again the condition (5.25) holds. Thus the diversity D^S based on Shannon's entropy remains the only known index to be generally additive for two statistical independent attributes. For measuring economic inequality at least Cowell has recognized the importance of this fact [19, p. 48], whereas Rao seems to ignore this subject altogether.

If nothing else is stated explicitly in the following the diversity D is synonymous with the Shannon diversity D^S.

5.4.3 Mutual and Redundant Information

Continuing from inequality (5.29) the function

(5.35) $I(A_1, A_2) = D(A_1) + D(A_2) - D(A_1A_2)$

is defined to be the *mutual information* of the attributes A_1 and A_2 [18, p. 18*ff*]. From equation (5.22)

(5.36) $I(A_1, A_2) = D(A_1) - D(A_1|A_2) = D(A_2) - D(A_2|A_1)$

and furthermore

(5.37) $D(A_1A_2) = D(A_1|A_2) + I(A_1, A_2) + D(A_2|A_1)$

follow. The equations (5.35) and (5.36) combined with the inequalities (5.27) and (5.28) directly lead to

(5.38) $0 \leq I(A_1, A_2) \leq \min(D(A_1), D(A_2)) \leq D(A_1A_2).$

Clearly the mutual information $I(A_1, A_2)$ also equals the *redundant information* $\Theta(A_1, A_2)$, which is gained twice if both attributes A_1 and A_2 are evaluated separately. Thus

(5.39) $I(A_1, A_2) = \Theta(A_1, A_2)$

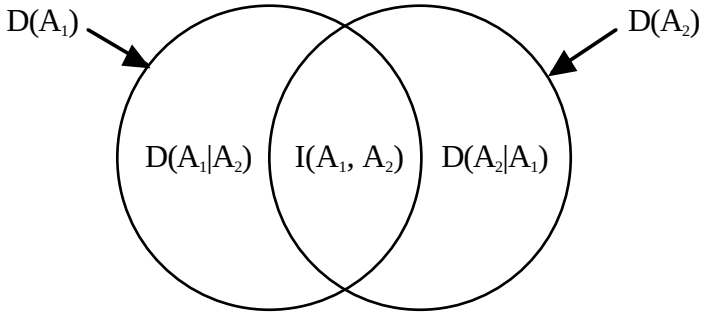


Figure 5.1: Venn diagram of the relationship between the diversities respectively entropies and the mutual information [18, p. 20].

holds and both terms can be used synonymously here. From equation (5.38) the coefficient

$$\vartheta = \frac{\Theta(A_1, A_2)}{D(A_1 A_2)}$$

between zero and one can be defined. The term $\gamma = 1 - \vartheta$ equals Rajski's metric coefficient of independence [78].

Cover and Thomas [18, p. 20] have tried to elucidate the above equations using the Venn diagram in figure 5.1. Clearly, from equation (5.36)

$$D(A_1|A_2) + I(A_1, A_2) = D(A_1)$$

follows, which corresponds to the set operations

$$(D(A_1) \setminus D(A_2)) \cup (D(A_1) \cap D(A_2)) = D(A_1)$$

in the Venn diagram. Consequently, the mutual information of the attributes A_1 and A_2 corresponds to the intersection of the sets representing them. The whole area in the diagram can be partitioned corresponding to equation (5.22) and (5.37) into non overlapping parts. But there is no related set operation for equation (5.35) possible, because the set theoretic union operator $\cup$ does not take into account whether an element is included in both sets or only in one. Thus it corresponds to the $+$ operator only, if both related sets are disjoined. But if the diversities are treated to be areas in the Venn diagram, then in the addition in equation (5.35) the area corresponding to $I(A_1, A_2)$ respectively $\Theta(A_1, A_2)$ is contained both in the areas representing $D(A_1)$ and $D(A_2)$ and therefore has to be subtracted to get the area corresponding to the species diversity $D(A_1 A_2)$. Thus the species diversity is free of redundancy. This is an analogy to the probability of the occurrence of two events. The probability, that at least one of them happens, is the sum of their single probabilities of occurrence reduced by the probability of their combined occurrence. Unfortunately until now there is no algebraic interpretation of this situation[5].

Mutual information can also be introduced from another point of view. Analogously to the *relative entropy* or *Kullback Leibler distance* between two probability mass functions

[5]Maybe the situation can be elucidated better by the use of *multisets* [7], which are also called *bags*. Their elements also have a cardinality, which can help to manage the problems pointed out above.

in information theory [18, p. 18] the *relative diversity* between two arbitrary distributions of the relative frequencies r_i and s_i of an attribute i is defined to be

$$D(r||s) = \sum_{i=1}^{n} r_i \ln \frac{r_i}{s_i}. \tag{5.40}$$

It is always nonnegative and zero iff $r_i = s_i$ for all i. Thus it is an index of the similarity of the distributions of the two variables r and s but not a true distance function because it is not symmetric as claimed in section 5.2.1.

Starting from equation (5.36) applying some transformations the equation

$$I(A_1, A_2) = D(A_1) - D(A_1|A_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} r_{i,j} \ln \frac{r_{i,j}}{r_{i,*} r_{*,j}} = D(r_{i,j}||r_{i,*} r_{*,j}) \tag{5.41}$$

can be deduced [18, p. 19]. Finally, from the above equation

$$I(A_1, A_2) = I(A_2, A_1) \quad \text{and} \quad I(A_1, A_1) = D(A_1)$$

follows.

5.4.4 Statistical Dependence of Attributes

Transferred to the initial problem the mutual information here measures the dependence between the values of the two considered attributes, which will be eventually elucidated by the consideration of the two limit situations.

If the value of the attribute A_1 of an individual implies the value of attribute A_2, then $D(A_2|A_1) = 0$ holds and from equation (5.22) $D(A_1 A_2) = D(A_1)$ follows. Combining this again with equation (5.22) the inequality

$$D(A_1) = D(A_2) + D(A_1|A_2) \geq D(A_2)$$

and thus $\min(D(A_1), D(A_2)) = D(A_2)$ results. In fact, this is the maximum of $I(A_1, A_2)$ set by inequality (5.38). Furthermore $D(A_2|A_1) = 0$ substituted in equation (5.36) yields $I(A_1, A_2) = D(A_2)$, thus it reaches its maximum just calculated. If also attribute A_2 implies the value of attribute A_1, then

$$I(A_1, A_2) = D(A_1) = D(A_2) = D(A_1 A_2) \tag{5.42}$$

and thus $\vartheta = 1$ hold, i. e. two individuals of different species differ in both attributes and the relative frequencies of the attributes equal the relative frequencies of the species. Thus from the observation of one attribute value the value of the second attribute and also the species affiliation are clearly determined.

Conversely, if the attributes are statistically independent, then equation (5.32) holds and from equation (5.35) $I(A_1, A_2) = 0$ and thus $\vartheta = 0$ follows. The sum of the attribute diversities has reached its minimum, which at least is needed to provide the species diversity $D(A_1 A_2)$ and there is no redundant information. In this situation also equation (5.33) is valid. Consequently, observing one attribute does not reduce the diversity respectively uncertainty of the second one and the species affiliation. As already stated in section 5.4.2 this situation also arises if one attribute has only one value for all objects.

Taking this into account also equation (5.41) can be interpreted. The mutual information compares the observed distribution of the combinations of both attributes with their fictive distribution under the hypothesis of their statistical independence using the relative diversity.

5.4.5 Multiple Attribute Extension

The considerations proposed until now can be extended to more than two attributes. For the sake of simplicity first only three attributes A_1, A_2, A_3 are considered. Starting from $D(A_1A_2A_3)$ using the chain rules [18, p. 21*ff*] the transformation

$$\begin{aligned} D(A_1A_2A_3) &= D(A_1) + D(A_2A_3|A_1) \\ &= D(A_1) + D(A_2|A_1) + D(A_3|A_2A_1) \\ &= D(A_1) + D(A_2|A_1) + D(A_2) - D(A_2) + \\ &\quad D(A_3|A_1A_2) + D(A_3) - D(A_3) \\ &= D(A_1) + D(A_2) + D(A_3) - I(A_1, A_2) - I(A_3, A_1A_2) \qquad (5.44) \\ &= D(A_1) + D(A_2) + D(A_3) - \Theta(A_1, A_2, A_3). \qquad (5.45) \end{aligned}$$

can be carried out. Thus the redundant information $\Theta(A_1, A_2, A_3)$ combines the mutual information between the attributes. Using equation (5.41) the transformation

$$\begin{aligned} \Theta(A_1, A_2, A_3) &= I(A_1, A_2) + I(A_3, A_1A_2) \\ &= \sum_{i=1,j=1}^{n_1,n_2} r_{i,j} \ln \frac{r_{i,j}}{r_{i,*}r_{*,j}} + \sum_{i=1,j=1,k=1}^{n_1,n_2,n_3} r_{i,j,k} \ln \frac{r_{i,j,k}}{r_{*,*,k}r_{i,j,*}} \\ &= \sum_{i=1,j=1,k=1}^{n_1,n_2,n_3} r_{i,j,k} \ln \frac{r_{*,*,k}r_{i,j,*}}{r_{*,*,k}r_{i,*,*}r_{*,k,*}} + \sum_{i=1,j=1,k=1}^{n_1,n_2,n_3} r_{i,j,k} \ln \frac{r_{i,j,k}}{r_{*,*,k}r_{i,j,*}} \\ &= \sum_{i=1,j=1,k=1}^{n_1,n_2,n_3} r_{i,j,k} \ln \frac{r_{i,j,k}}{r_{i,*,*}r_{*,j,*}r_{*,*,k}} \end{aligned} \qquad (5.46)$$

follows. Thus again the real joined distribution of the attributes is compared against the fictive joined distribution assuming statistical independence using the relative diversity. Since for two attributes redundant information and mutual information are equal, then the comparison also gives the mutual information as in equation (5.41). This is not true for more than two attributes. The situation is depicted in the Venn diagram of figure 5.2. The mutual information $I(A_1, A_2, A_3)$ is common to all three attributes, thus is gained by evaluating any one of them. Consequently it is defined to be

$$I(A_1, A_2, A_3) = I(A_1, A_2) - I(A_1|A_3, A_2|A_3) = I(A_1, A_2) - I(A_1, A_2|A_3) \qquad (5.47)$$

From that with some transformations

$$I(A_1, A_2, A_3) = I(A_1, A_2) + I(A_1, A_3) - I(A_1, A_2A_3)$$

follows. There are corresponding formulas for the other permutations of the attributes. In the Venn diagram of figure 5.2 this corresponds to the set operation

$$I(A_1, A_2, A_3) = I(A_1, A_2) \cap I(A_1, A_3).$$

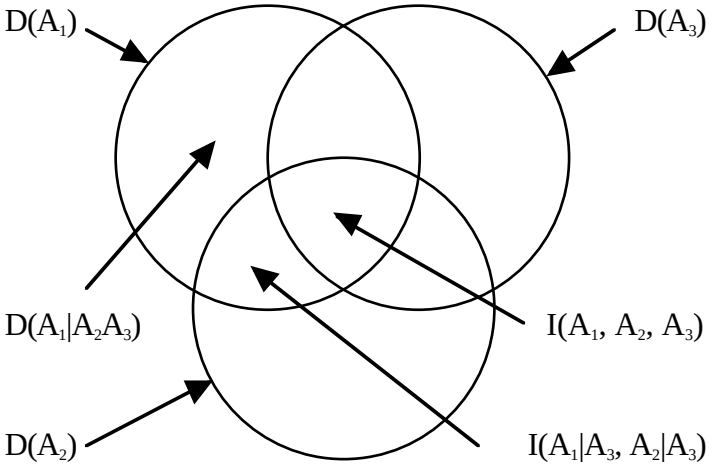


Figure 5.2: Venn diagram of the relationship between three diversities.

From that correspondence one normally would infer the inequality

(5.48) $0 \leq I(A_1, A_2, A_3)$,

which unfortunately does not hold [61], because contradictory examples can be found [71, problem 8-2, p. 237]. Thus for more than two attributes a graphical representation of the dependence structure is only possible for special cases. But without further proof from figure 5.2 at least

(5.49) $I(A_1, A_2, A_3) \leq \min(I(A_1, A_2), I(A_2, A_3), I(A_1, A_3))$.

follows. Consequently any independence between the attributes implies the non positivity of the triple mutual information. The redundant information $\Theta(A_1, A_2, A_3)$ can not even be shown in the situation depicted in figure 5.2, because expanding the mutual informations in equation (5.44) shows, that the triple mutual information $I(A_1, A_2, A_3)$ is contained twice in the redundant information.

The equations (5.43), (5.45) and (5.47) can be further extended from three to M attributes. Consequently

(5.50) $D(A_1 \cdots A_M) = D(A_M) + D(A_{M-1}|A_M) + \cdots + D(A_1|A_2 \cdots A_M)$

and

(5.51) $D(A_1 \cdots A_M) = \sum_{l=1}^{M} D(A_l) - \Theta(A_1, \ldots, A_M)$

follow with

$$\Theta(A_1, \ldots, A_M) = I(A_1, A_2) + I(A_3, A_1A_2) + \ldots + I(A_M, A_1 \cdots A_{M-1}).$$

Finally for the mutual information

$$\begin{aligned} I(A_1, \ldots, A_M) &= I(A_1, \ldots, A_{M-1}) - I(A_1, \ldots, A_{M-1}|A_M) \\ &= I(A_1, A_2) - I(A_1, A_2|A_3) - \ldots - I(A_1, \ldots, A_{M-1}|A_M) \end{aligned}$$

holds.

Analogously the inequalities deduced until now for two attributes also can be extended for M attributes. Using corresponding argumentation as for two attributes from equation (5.50) and (5.51) the inequality

$$\max(D(A_1),\ldots,D(A_M)) \leq D(A_1\cdots A_M) \leq \sum_{l=1}^{M} D(A_l) \leq M\, D(A_1\cdots A_M) \tag{5.52}$$

follows analogously to inequality (5.29) and the inequality

$$0 \leq \Theta(A_1,\ldots,A_M) \leq (M-1)D(A_1\cdots A_M). \tag{5.53}$$

analogously to inequality (5.38). From that the coefficient

$$\vartheta = \frac{\Theta(A_1,\ldots,A_M)}{(M-1)D(A_1\cdots A_M)} = \frac{\sum_{l=1}^{M} D(A_l) - D(A_1\cdots A_M)}{(M-1)D(A_1\cdots A_M)} \tag{5.54}$$

with $0 \leq \vartheta \leq 1$ can be derived. For the mutual information

$$I(A_1,\ldots,A_M) \leq \min(I(A_1,A_2),\ldots,I(A_1,A_M),I(A_2,A_3),\ldots,I(A_{M-1},A_M))$$

follows. Its further properties are not considered here, because they are of no interest for the main subject of this work. Nevertheless their exploration will be worthwhile.

5.4.6 Dependence versus Schema Reach

Like for two attributes in section 5.4.3, where it equals the mutual information, the redundant information $\Theta(A_1,\ldots,A_M)$ and the corresponding coefficient ϑ measure not only the redundancy but also the *dependence* respectively *determination* between the values of the M attributes. If $\vartheta = 1$ respectively $\Theta(A_1,\ldots,A_M) = (M-1)D(A_1\cdots A_M)$ hold, then individuals from different species respectively classes differ in each attribute. Furthermore the distributions of the attribute values are the same for all attributes. Conversely, if $\Theta(A_1,\ldots,A_M) = \vartheta = 0$ holds, then the attributes are statistical independent. Thus, like in section 5.4.2 the sum of the attribute diversities has reached its minimum which at least is needed to provide the species diversity $D(A_1\cdots A_M)$.

The above considerations can be combined with the notion of a schema proposed in section 3.7. Clearly, if $\vartheta = 1$ holds, then in the population there is no schema reaching across species borders. Conversely, if $\vartheta = \Theta(A_1,\ldots,A_M) = 0$ holds, then the attributes are statistical independent and the schemata have spread among the different species in the population as much as possible with respect to the species diversity $D(A_1\cdots A_M)$ and the number of attributes M. Particularly, there may be attributes having the same value for all classes respectively species. Clearly the propagation of schemata across species borders in a population reduces the diversity of the related attributes respectively genes. If the species diversity in spite of this remains constant, then the redundant information $\Theta(A_1,\ldots,A_M)$ and the dependence coefficient ϑ decrease. Consequently both are indices of the *heterogeneity* of the different species with respect to the species diversity $D(A_1\cdots A_M)$. If further attributes are introduced, which have the same value for all species, then $\Theta(A_1,\ldots,A_M)$ remains constant, but ϑ decreases, as it is expected, because M increases. Summed up,

schema propagation always leads to a reduction of the sum of the attribute diversities and thus the population diversity. But such a reduction does not guarantee a schema propagation between species. It may be induced also by the increase of the portions of some dominating species, which manifests itself by a reduction of the species diversity.

Finally, the dependence respectively determination coefficient ϑ can replace the normed contingency coefficient in statistics[6]. It also compares the real joint distribution against the fictive distribution assuming statistical independence, but uses relative diversity instead of χ^2 statistics [54, p. 155*ff*]. Furthermore it can be used easily for more than two dimensions, which is a great advantage.

Having explored the relation between species and population diversity to dependence and schema propagation, finally the application to evolution programs has to be considered, which is performed in the next section.

5.5 Status Indices for Evolution Programs

To monitor the reach of schemata during evolution the use of the population diversity is needed. Since schemata usually are not required to reach across genotype class borders, to use the Gini-Simpson-index is sufficient. But then it is not possible to decide whether only dominant classes of genotypes spread or whether there is an assimilation between the classes. Additional information can be obtained by monitoring also the species diversity. Using it, uniformity in the population can be detected. If the species and the population diversity are evaluated using Shannon's index, then using the number of loci M the dependence coefficient ϑ can be calculated, which is based on the relation of both diversity indices. A value near one indicates that the schemata in the population have hardly spread across different classes, whereas a value near zero indicates a near maximum possible assimilation of the different classes. Generally using Shannon's index the species diversity limits the population diversity from below, which follows from inequality (5.52) and corresponds to a dependence coefficient $\vartheta = 0$. Conversely the population diversity is limited from above by the M fold of the species diversity, which follows from the same equation and corresponds to a dependence coefficient $\vartheta = 1$.

Usually an evolution program or genetic algorithm starts from a randomly generated start population of size N. Then a species diversity close to the maximum $D_S^{\max}(\boldsymbol{P}) = \ln N$ given by equation (5.15) can be expected, because it is improbable that there are identical individuals in the population and assumed that the number of possible classes of genotypes is larger than the population size N. Furthermore each allele of a locus should occur with the same relative frequency. If the population size N is smaller than the number of possible alleles n_j at a specific locus j then not all possible alleles can occur in the population simultaneously. Thus on the encoding level not the whole domain can be covered. Considering this restriction, from equations (5.14) and (5.18) the maximum population diversity

$$D_P^{\max}(\boldsymbol{P}) = \sum_{j=1}^{M} \ln \min(n_j, N) \tag{5.55}$$

[6]The term contingency is misleading in this context because for independence the related coefficient in statistics has to be zero. In fact then there is no dependence, thus it also should be named dependence coefficient.

follows and the start population should have a population diversity close to it. If $n_j > N$ holds, then the dependence coefficient ϑ should be close to one, because then the diversities of the loci equal approximately the species diversity. Conversely, if the opposite holds, then the initial diversities of the attributes are smaller than the species diversity and consequently the dependence coefficient ϑ will be significantly lower than one. Consequently, in such a random population already some schemata reach across different classes of genotypes and their number increases with the population size N.

During the run of an evolution program a propagation of schemata will reduce the population diversity, whereas the species diversity should not decrease significantly, because that would indicate the propagation of equivalent individuals respectively an increase of uniformity, which is not intended. Consequently a significant reduction of ϑ can be expected during a run.

Using a binary encoding, which is a characteristic of the genuine genetic algorithm, a complete coverage of the encoding domain always can be achieved despite of a relatively small population size, because already a population consisting of two individuals differing at each bit achieves the maximum possible population diversity. But trying to preserve the coverage needs appropriate operators and is expensive not only with respect to the computing time [66]. From the considerations proposed above also follows that in a randomly created binary encoded population always a lot of schemata should cover different classes of genotypes.

It is also advantageous to partition a population into several subpopulations and to process them separately, because the diversity of the whole population always is at least as high as the average of the subpopulations, which follows from the concavity of the diversity. This advantage is used implicitly already e. g. in the program "Asparagos" [35].

5.6 Summary

In this chapter the species and the population diversity have been introduced to be suitable indices to monitor the state of a population. A reduction of the population diversity during the run of an evolution program indicates the propagation of schemata and an assimilation, whereas a reduction of the species diversity indicates an increase of uniformity. If both indices are calculated using Shannon's entropy as diversity index, then from them the new dependence coefficient ϑ also can be calculated. It indicates the presence of dependences in the population, which is related to the reach of schemata across different classes of genotypes. If the dependence coefficient decreases during the run of an evolution program or genetic algorithm, schemata spread in the population between different classes of genotypes and thus cause an assimilation without an increase of uniformity. Shannon's entropy function has to be used here because it is the only diversity index providing additivity in the case of statistical independence, which is needed to identify dependences. The Gini-Simpson-index favored by Rao does not have this property and thus is not capable to detect dependences.

Using the proposed indices having a strong theoretical background the relation between a successful improvement of the objective values close to the optimum during evolution and a propagation of schemata in general and especially across genotype classes can be explored.

The properties of the dependence coefficient ϑ have not been explored to the end until now. Furthermore its statistic has to be considered to use it also for statistical tests on independence. Finally the properties of the mutual information for more than two variables have to be explored more deeply.

Chapter 6

Sample Experimental Observations

6.1 Introduction

Due to missing suitable indices exact observations of schema propagation in experiments could not be performed until now. Thus there is hardly an evidence of the relevance of the theoretical models supporting that conjecture. Using the new indices developed in the last chapter this gap can be filled. In the following observations are proposed gained with an improved version of TSPGA [30] on some sample instances of the traveling salesman problem. TSPGA has been extended to calculate both the species and the population diversity using Shannon's index and from that the dependence coefficient. In the experiments the objective has been to explore schema propagation and other phenomena expected from the theoretical selection model, but not to show the superiority of TSPGA over other evolution programs for the symmetric traveling salesman problem. All experiments proposed in the following have been performed with fitness proportional selection, although TSPGA due to its use of PGAPack [56] provides also other selection methods.

As stated already above, three classes of instances are considered. First, there are instances especially constructed to mislead heuristics. Second, there are problem instances created at random. Normally in this case most heuristics find at least near optimal solutions. Third, there are trivial instances, where the optimum can be found by any deterministic improvement strategy or at least by a greedy algorithm.

Since the problem size respectively number of cities M in the chosen sample instances varies from 100 to 159 two population sizes of 50 and 300 individuals have been selected for presentation. This results in maximum species diversities of $D_S^{\max}(\boldsymbol{P}) = \ln 50 \approx 3.91$ and $D_S^{\max}(\boldsymbol{P}) = \ln 300 \approx 5.70$ and for a random initial population the observed values of this indices can be expected to be close to that values, because each individual should belong to a separate species.

The population size of 50 individuals is lower than number of cities and consequently the attribute values at each locus can not cover the whole domain, which is a usual situation for very large instances. Then according to equation 5.55 the maximum population diversity is $D_P^{\max}(\boldsymbol{P}) = M \ln 50$ and again for a random initial population the observed value can be expected to be close to that value. Following the argumentation of section 5.5 in this situation the dependence coefficients of a random initial population $\vartheta(\boldsymbol{P}(0))$ should be close to one and two randomly drawn individuals should differ in almost every locus.

The population size of 300 individuals is higher than the number of cities. Thus the whole domain can be covered at each locus. According to equation 5.55 the maximum population diversity is $D_P^{\max}(\boldsymbol{P}) = M \ln M$, which implies a uniform distribution of the possible attribute values at each locus. But the population size of 300 individuals is to small in relation to the number of different possible attribute values respectively cities to approach a uniform distribution closely. Consequently here in a random initial population the observed value can be expected to be significantly lower than that value. Following the argumentation of section 5.5 in this situation the dependence coefficients of a random initial population $\vartheta(\boldsymbol{P}(0))$ should significantly lower than one, because the species diversity is not limited by of the problem size, as considered already above. Thus although a population is generated at random, there are already schemata reaching across different species. Clearly, the dependence coefficient for random populations decreases for populations sizes greater than the tourlength, because then the population diversity retains its near maximum value, but the species diversity increases under the condition, that each individual belongs to a different species, which generally should hold for a random population.

In the following observations with three sample problem instances are proposed. All experiments have been finished after 5000 generations, which seems to be long enough to observe all important effects.

6.2 Constructed Instance

6.2.1 General Observations

Now the observations with instance u159, a 159 cities planar problem instance from the traveling salesman problem collection TSPlib95 [84] are proposed. The length of the optimal tour of this instance is 42081. As expected in all experiments the initial population $\boldsymbol{P}(0)$ exhibits a species diversity $D_S(\boldsymbol{P}(0))$ close to the maximum possible value stated above. Thus normally each individual belongs to a separate species and two individuals differ at least in one locus. From the considerations above for the population size of 50 individuals a maximum population diversity $D_P^{\max}(\boldsymbol{P}) = 159 \ln 50 \approx 622$ follows and for the population size of 300 individuals a maximum population diversity $D_P^{\max}(\boldsymbol{P}) = 159 \ln 159 \approx 806$. The initial population diversities always are slightly lower than these maximum possible values. For the populations of 50 individuals $D_P(\boldsymbol{P}(0)) \approx 590$ and for the populations of 300 individuals $D_P(\boldsymbol{P}(0)) \approx 757$ have been observed, which meets the conjecture of imperfect uniform distribution. The initial diversity values result in dependence coefficients of $\vartheta(\boldsymbol{P}(0)) \approx 0.9$ for the 50 individuals populations and of $\vartheta(\boldsymbol{P}(0)) \approx 0.8$ for the 300 individuals populations, which also meet the expectations. Finally, the populations exhibit a tourlength of the best individual of approximately 400,000 and an average of approximately 450,000.

6.2.2 Example Experiments

In the first considered series an elitist evolution program is realized by taking over the 10 % best individuals from the old population to the new one. As a general example experiment 53 with crossover probability 0.8, uniform crossover probability of 0.1, mutation

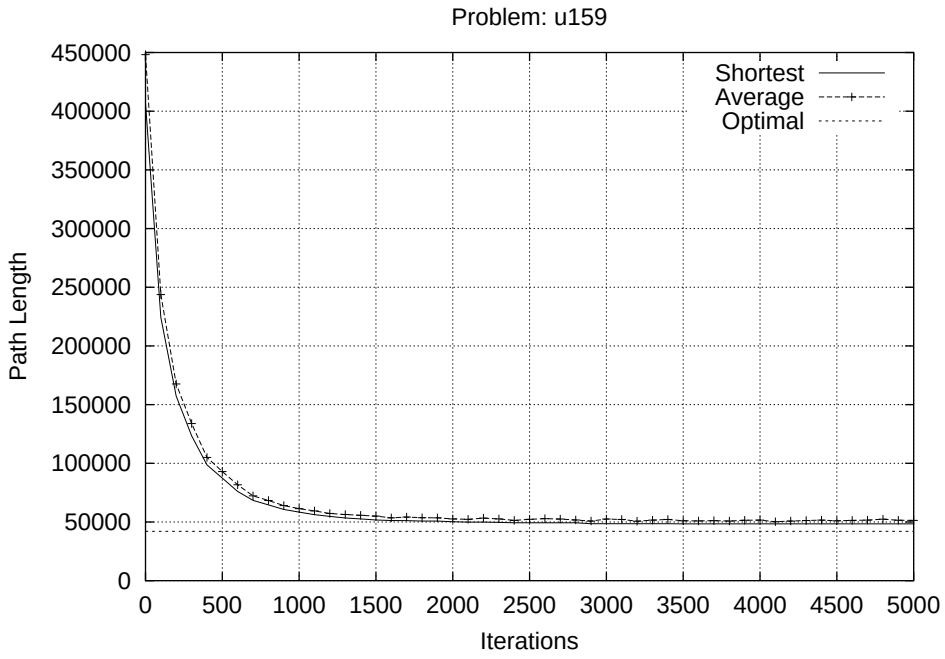


Figure 6.1: The development of best and average tourlengths in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

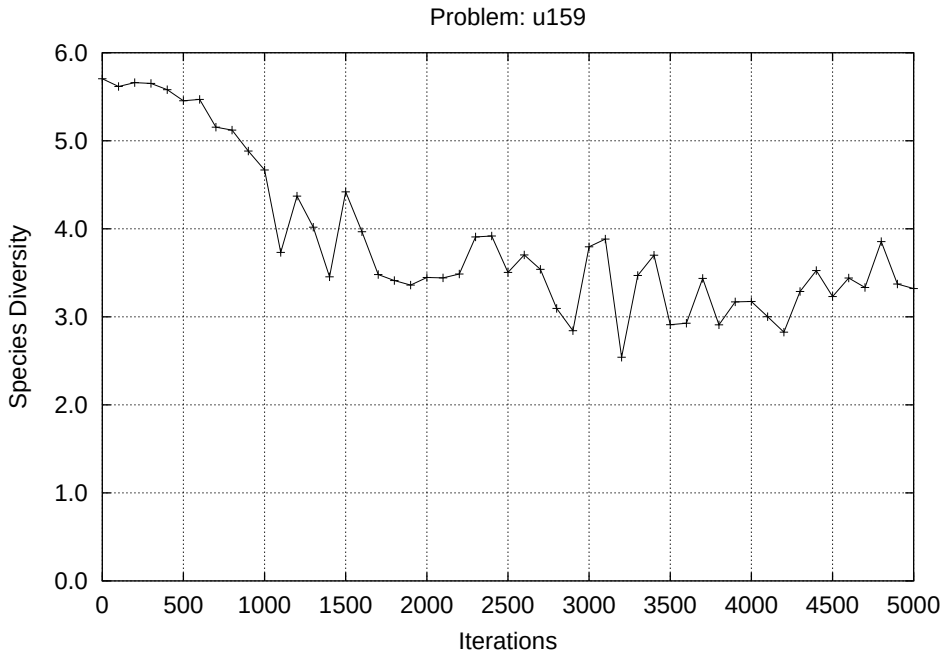


Figure 6.2: The development of the species diversity in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

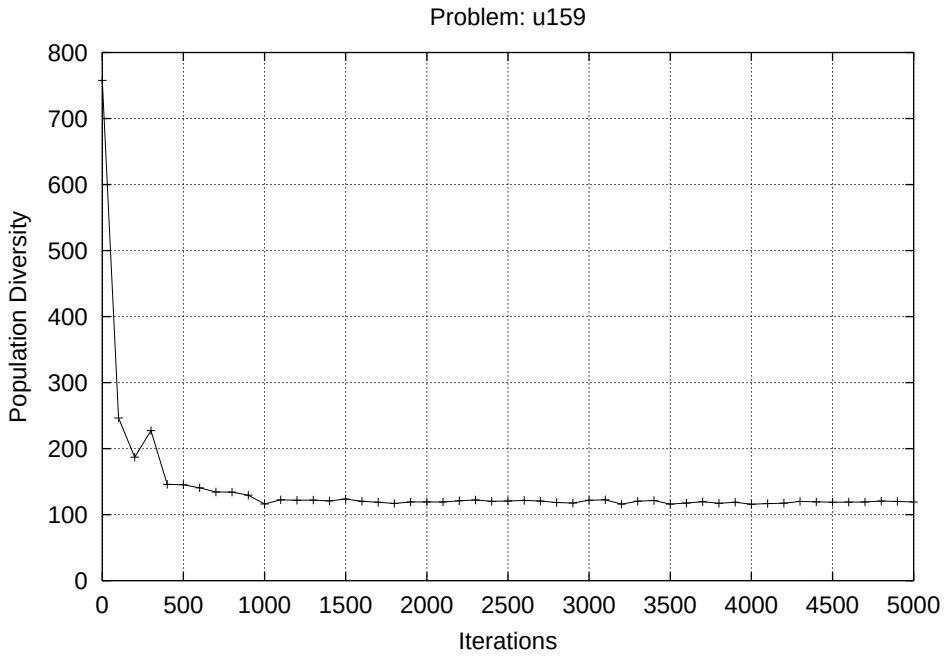


Figure 6.3: The development of the population diversity in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

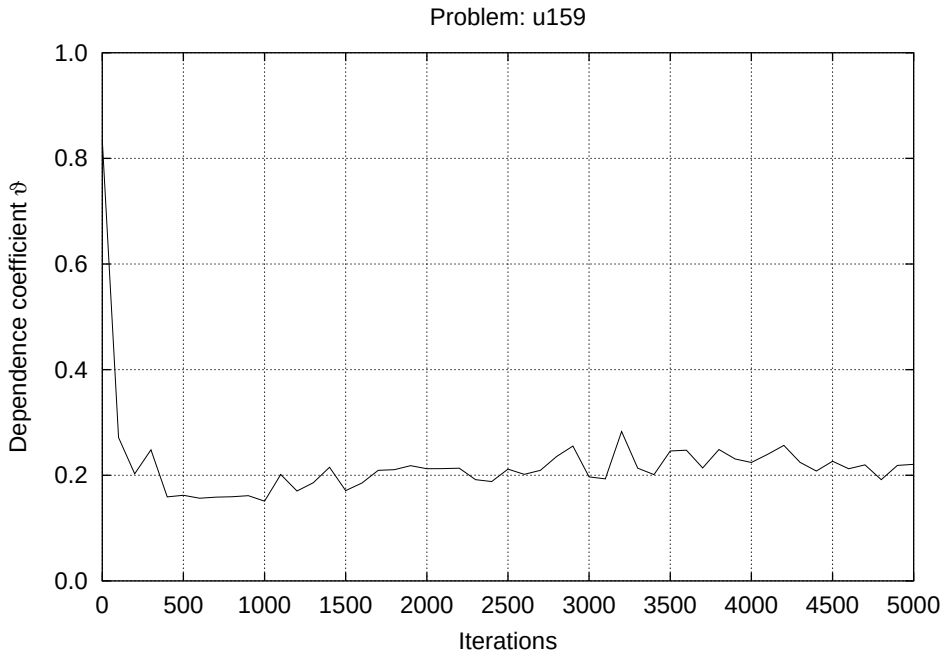


Figure 6.4: The development of the dependence coefficient in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

probability 0.025 and population of 300 individuals is considered. Figure 6.1 shows the development of the tourlength of the best individual and the population average during the run. Both values decrease strongly during the first stage and approach a value relatively close to the optimum although they do not reach it and remarkably they are always relatively close together. This improvement close to the optimum in the context of evolutionary computing is called convergence, but which has been defined above to be called saturation in objective. The development of the species diversity is shown in figure 6.2. First it stays close to the maximum possible value and then decreases with some jags to approximately half of the initial value. This indicates an increase of uniformity during the evolution. Conversely, figure 6.3 shows an initial drastic reduction of the population diversity to a limit value followed by a nearly constant continuation. This indicates an assimilation of the individuals to a specific level, which has been defined above to be saturation in domain. The assimilation is finished significantly before the tourlength values have approached their limit. The finish of the assimilation process coincides with the start of the decrease of the species diversity. From the species and the population diversities, the dependence coefficient ϑ can be calculated, whose development is shown in figure 6.4. The strong reduction caused by the assimilation indicates a propagation of schemata in the population between different species respectively classes of genotypes, which reaches its limit already before the improvement process has been finished. This phenomenon is the premature assimilation discussed already above. Later on there is a slight increase, which is caused by the decrease of the species diversity.

Summed up, these observations are more or less what is expected from a standard evolution program. Changing the parameters results in slightly different plots of the species diversity and the dependence coefficient.

Reducing the population size in experiment 52 to 50 individuals retards the improvement process significantly, which is shown in figure 6.5. The species diversity is reduced similarly as for the larger population size but with more jags, which is shown in figure 6.6. The development of the population diversity is shown in figure 6.7. It is reduced quickly to nearly the same level as for a population of 300 individuals. Due to the decreasing species diversity and its jags the dependence coefficient shown in figure 6.8 after its strong initial reduction increases slightly and also the plot becomes a lot more jaggy than for 300 individuals.

Generally changes of the parameters do not affect the saturation both in fitness and domain strongly, i. e. the fitness and population diversity plots do not change principally. Most importantly the time gap between the end of the assimilation and the improvement changes. Summed up fitness proportional selection with elitist population replacement is very robust against parameter and population size changes. Schemata propagate rapidly and using appropriate parameter values the species diversity can be preserved at least partly preventing uniformity from domination.

But increasing the population replacement fraction to one changes the behavior totally. Then there is no more elitism, but the algorithm becomes totally generational. With otherwise identical parameter setting as in experiment 53 with a population of 300 individuals proposed above, in experiment 65 nearly no improvement takes place, which is shown in figure 6.9. The species diversity shown in figure 6.10 remains nearly constant and the population diversity is only hardly reduced, which is shown in figure 6.11. Consequently also the dependence coefficient presented in figure 6.12 remains on a very high level. Thus there

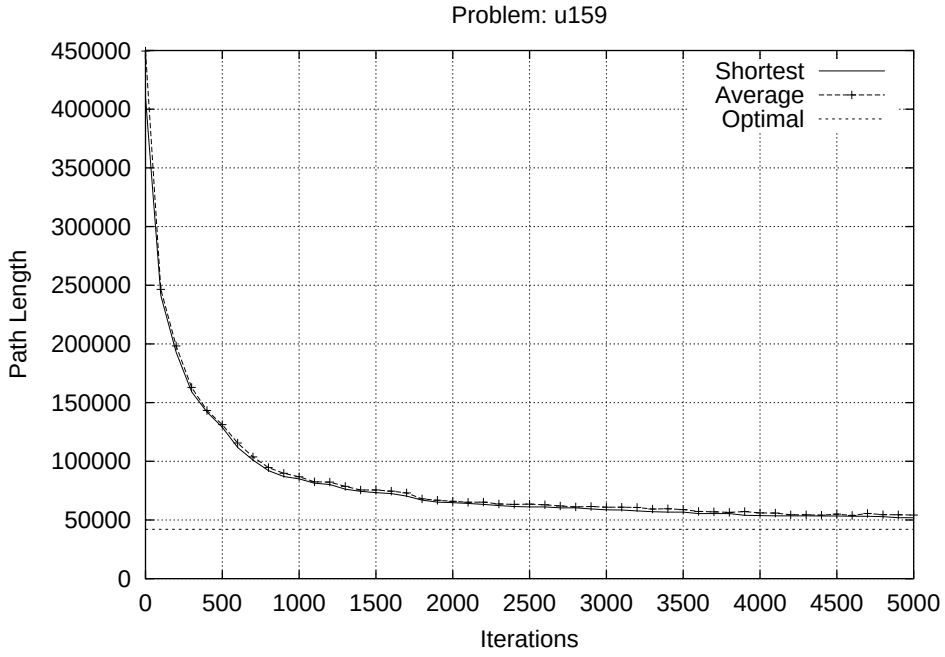


Figure 6.5: The development of best and average tourlengths in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

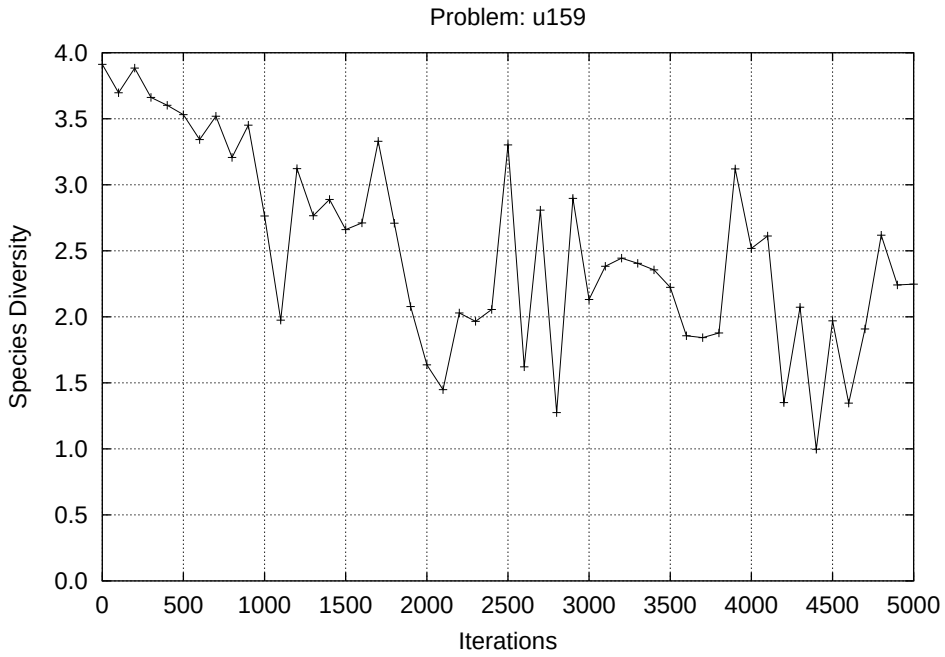


Figure 6.6: The development of the species diversity in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

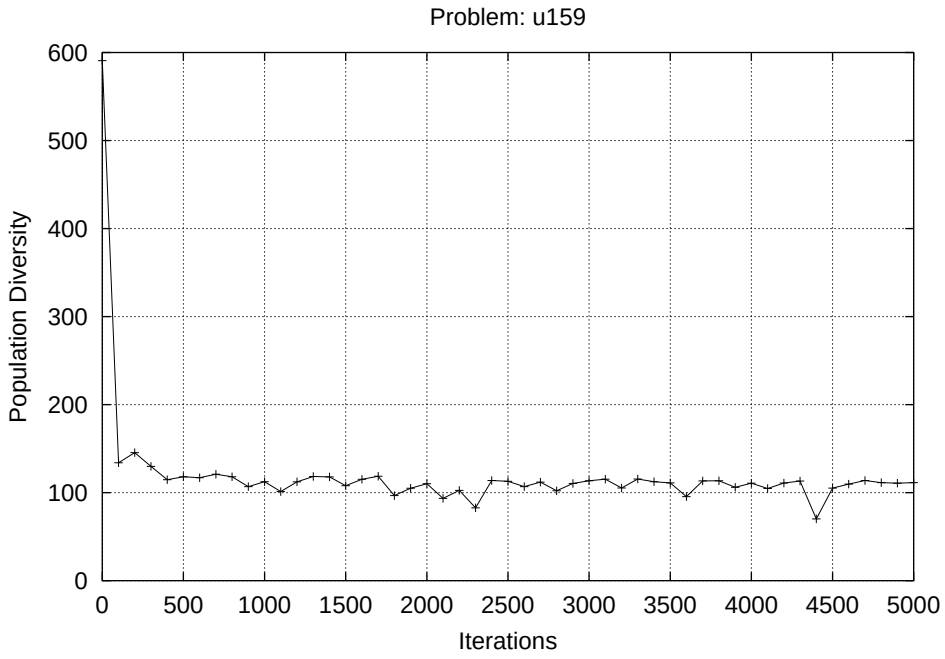


Figure 6.7: The development of the population diversity in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

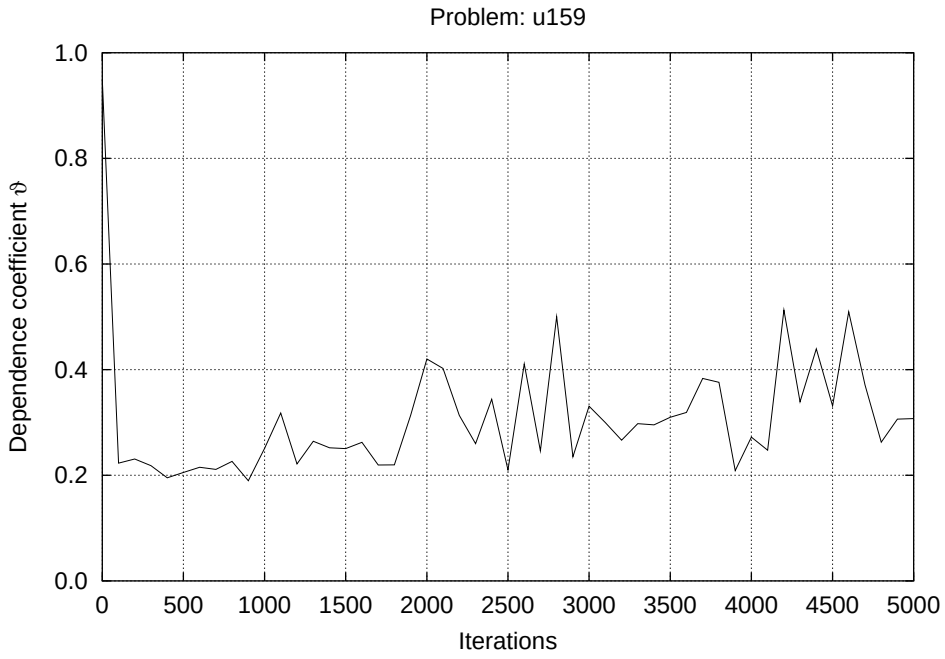


Figure 6.8: The development of the dependence coefficient in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

is nearly no improvement and no schema propagation during the evolution. A reduction of the population size to 50 individuals improves the behavior somewhat. Then there is a noticeable improvement and the population diversity is lowered slightly more than for 300 individuals.

If the mutation and the uniform crossover rates are lowered enough like in experiment 70, then the population diversity is reduced to an intermediate level, as shown in figure 6.15. Since the species diversity remains nearly constant, which is shown in figure 6.14, also the dependence coefficient is reduced to an intermediate level as shown in figure 6.16. Thus now there is schema propagation and it takes place between different species respectively classes of genotypes. Related to that there is a reduction of the tourlengths, but unfortunately not to a near optimum level. This incomplete improvement is shown in figure 6.13.

If both mutation and and crossover rates are further reduced and tuned, then the improvement process can be enforced. In experiment 73 the population diversity is reduced almost instantly to the same level as with elitist selection, as shown in figure 6.19. Since the species diversity remains nearly constant, which is shown in figure 6.18, also the dependence coefficient is reduced to a low level as shown in figure 6.20. Thus schemata spread already in the initial stage and reach also across different species respectively classes of genotypes. Related to that there is a strong reduction of the tourlengths, but unfortunately not to a near optimum level within the monitored number of generations, as shown in figure 6.17. The improvement respectively saturation in objective is very slow. Consequently there is a large generation gap between the instant assimilation and the improvement. Generally the behavior changes very smoothly, i. e. small changes of the parameters do not induce a drastic change of the behavior.

6.2.3 Summary

Summed up, without the help of elitist population replacement the propagation of schemata using fitness proportional selection is easily obstructed and the improvement is insufficient. On the other hand elitism slightly promotes the propagation of uniformity in the population after assimilation has finished, which is indicated by a decreasing species diversity. Generally, all experiments showing final tourlengths close to the optimum value also exhibit an initial reduction of the population diversity to a low value, which indicates a strong assimilation of the individuals on the average. On the other hand a missing assimilation always indicates a missing improvement process.

6.3 Random Instance

6.3.1 General Observations

In the following the observations with instance rd100, a random 100 cities planar problem instance taken also from the traveling salesman problem collection TSPlib95 [84] are proposed. The length of the optimal tour of this instance is 7910. For a population size of 50 individuals a maximum population diversity $D_P^{\max}(\boldsymbol{P}) = 100 \ln 50 \approx 391$ and for 300 individuals a maximum population diversity of $D_P^{\max}(\boldsymbol{P}) = 100 \ln 100 \approx 461$ follows.

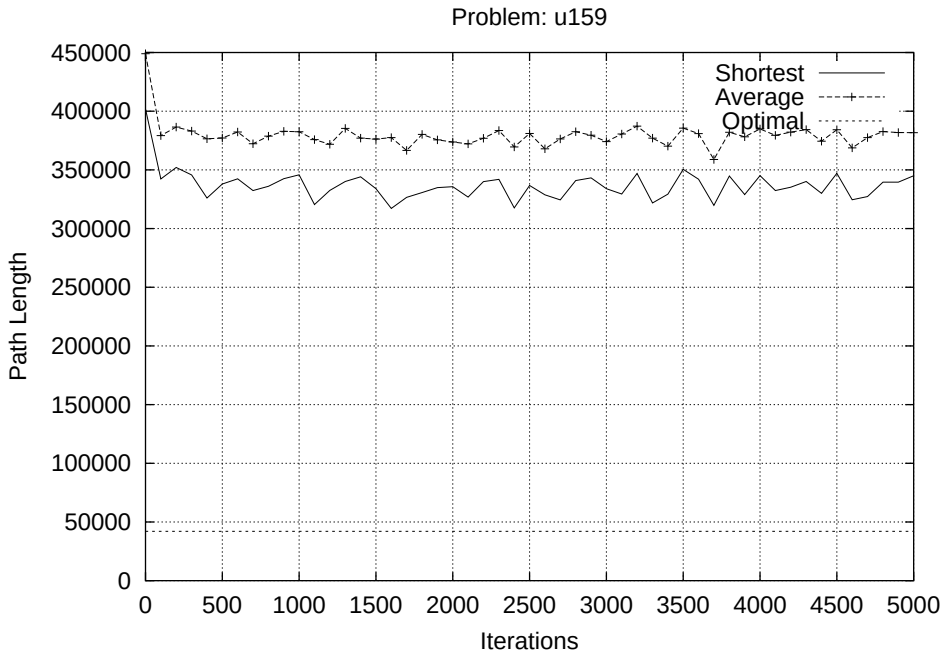


Figure 6.9: The development of best and average tourlengths in experiment 65 with a population of 300 individuals using a non elitist replacement strategy.

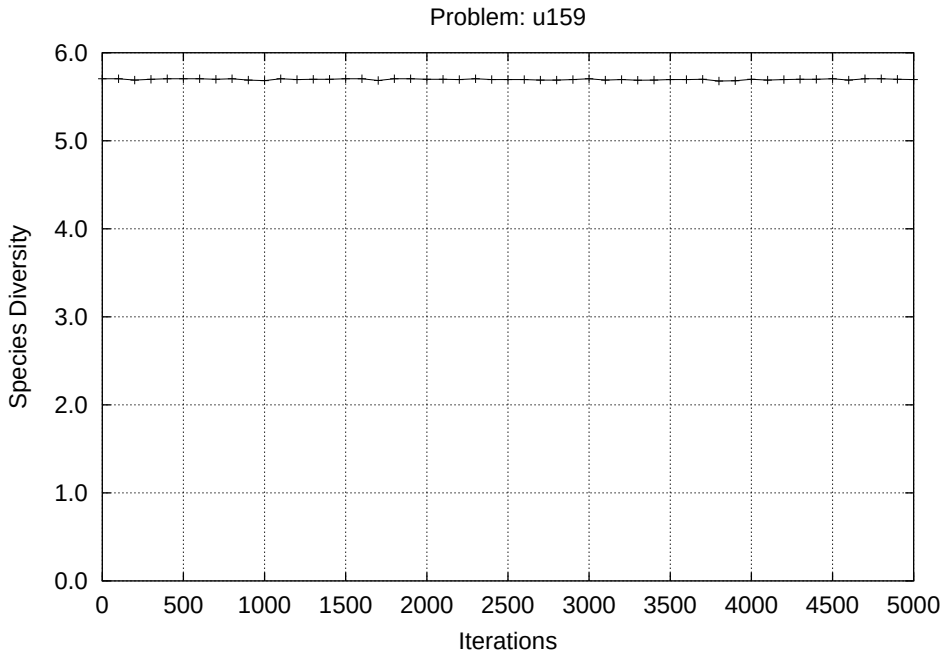


Figure 6.10: The development of the species diversity in experiment 65 with a population of 300 individuals using a non elitist replacement strategy.

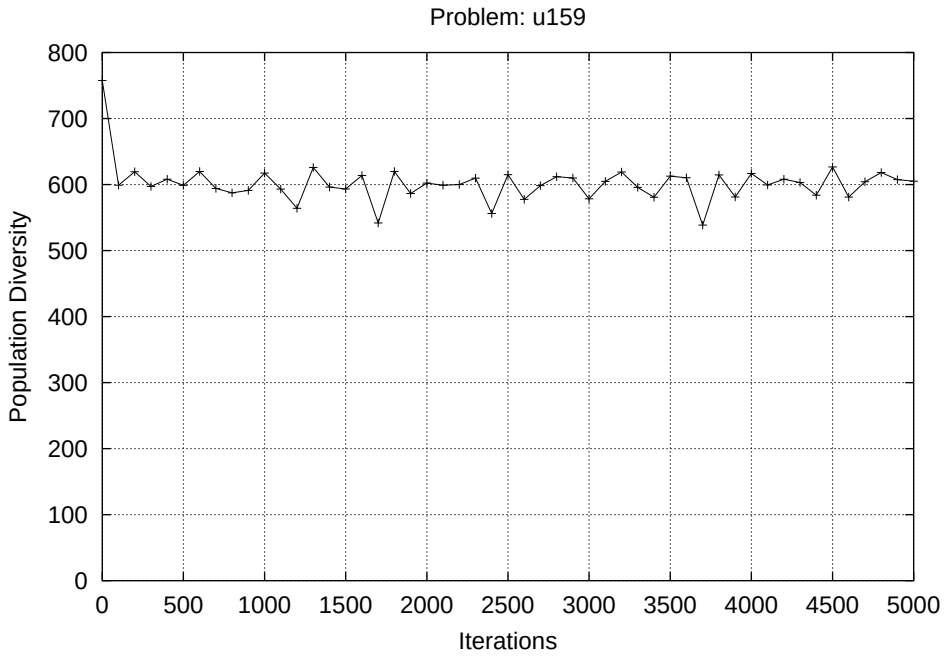


Figure 6.11: The development of the population diversity in experiment 65 with a population of 300 individuals using a non elitist replacement strategy.

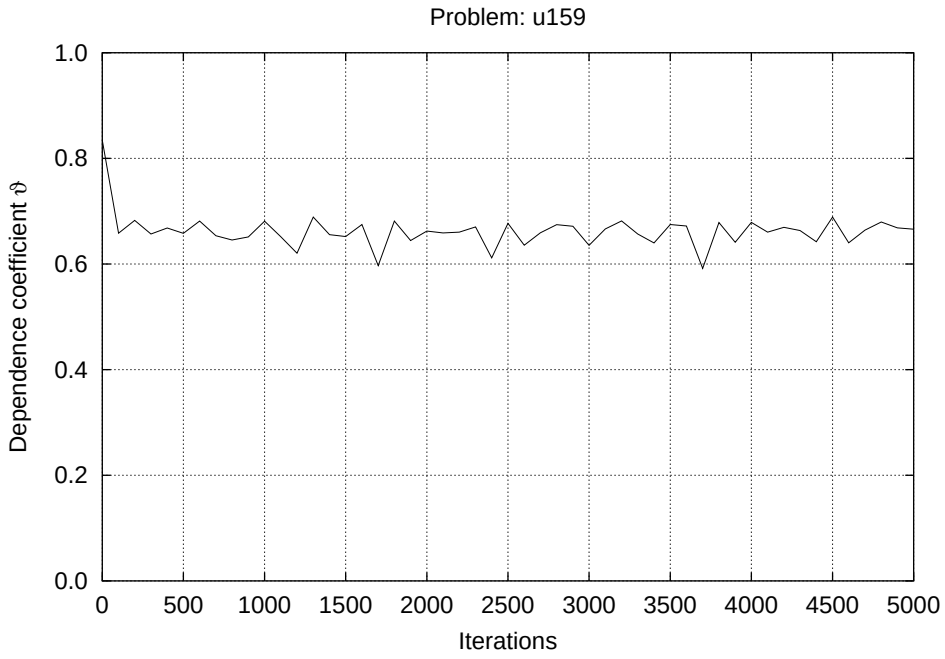


Figure 6.12: The development of the dependence coefficient in experiment 65 with a population of 300 individuals using a non elitist replacement strategy.

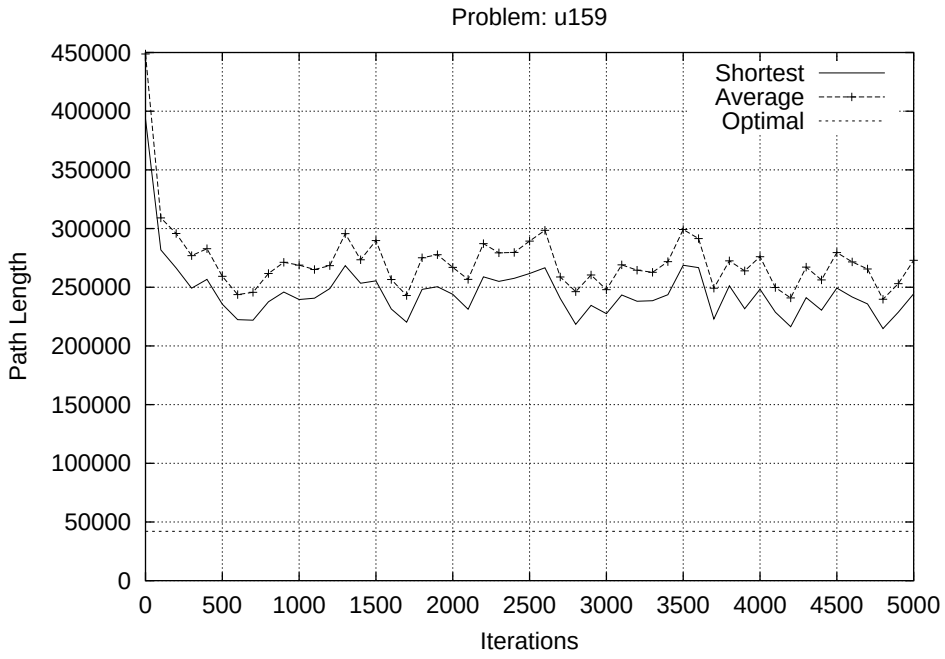


Figure 6.13: The development of best and average tourlengths in experiment 70 with a population of 300 individuals using a non elitist replacement strategy.

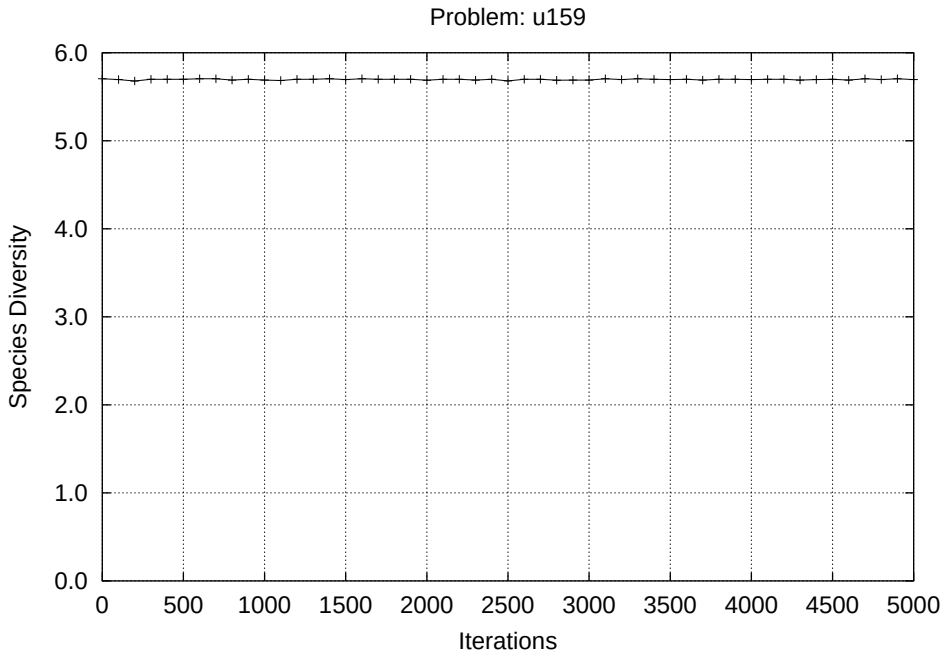


Figure 6.14: The development of the species diversity in experiment 70 with a population of 300 individuals using a non elitist replacement strategy.

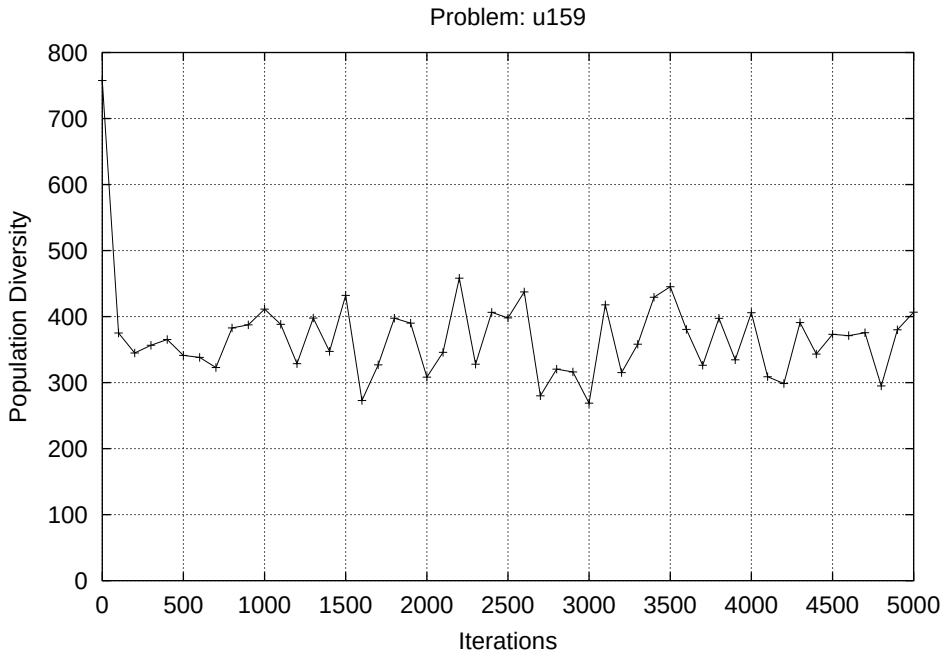


Figure 6.15: The development of the population diversity in experiment 70 with a population of 300 individuals using a non elitist replacement strategy.

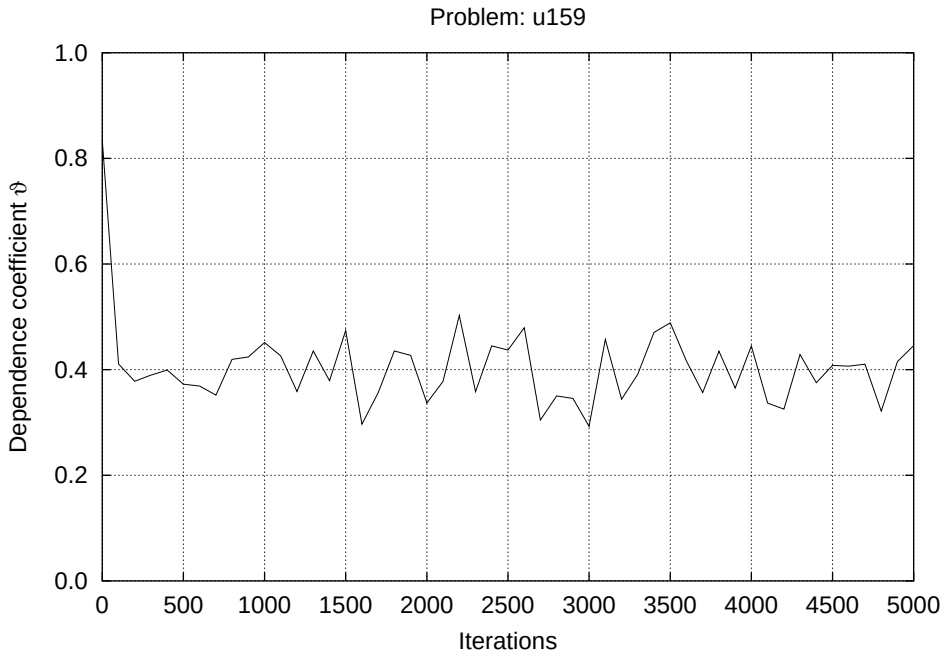


Figure 6.16: The development of the dependence coefficient in experiment 70 with a population of 300 individuals using a non elitist replacement strategy.

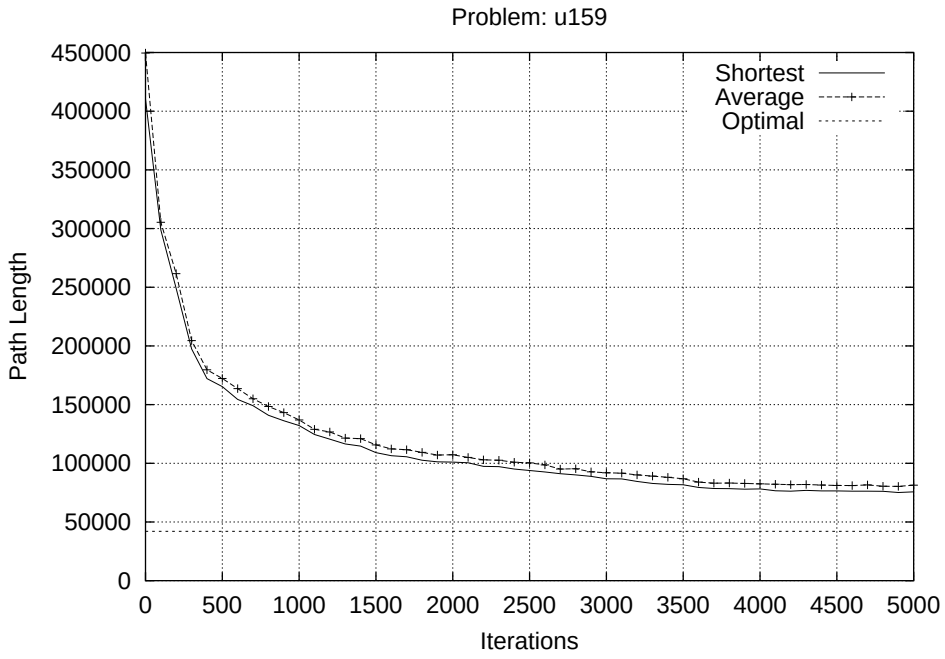


Figure 6.17: The development of best and average tourlengths in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

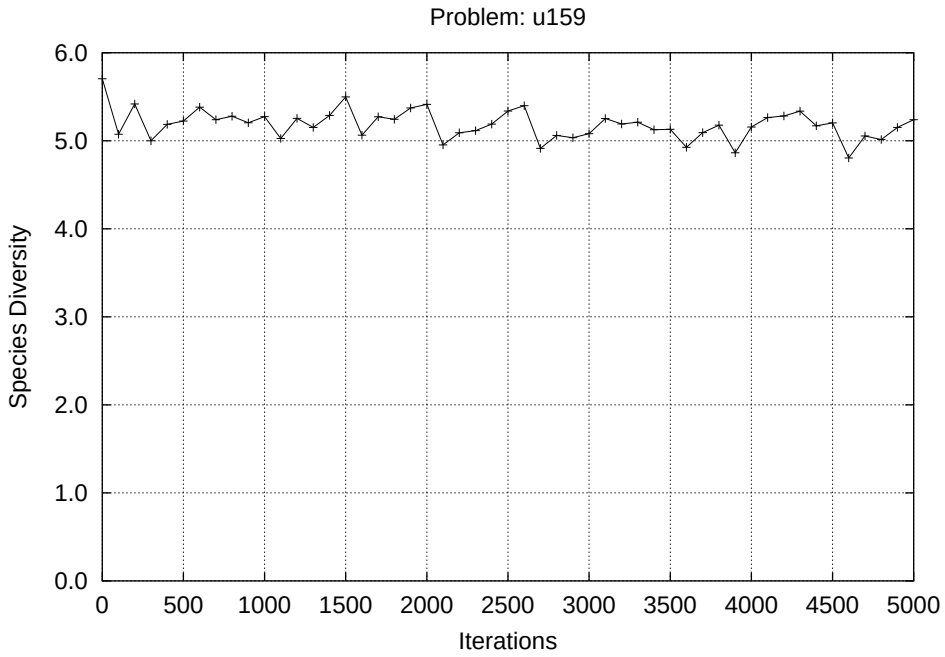


Figure 6.18: The development of the species diversity in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

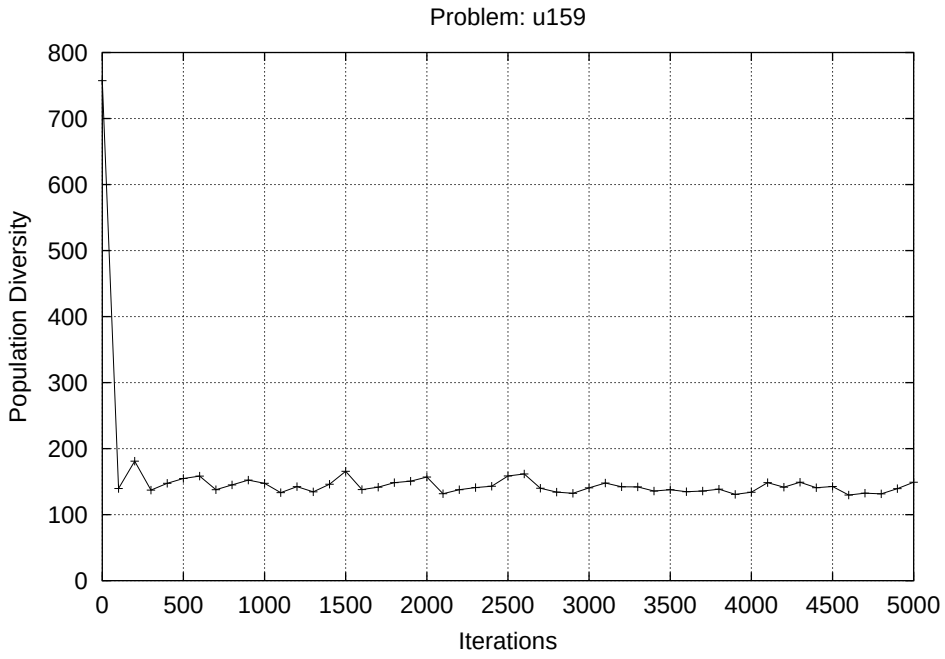


Figure 6.19: The development of the population diversity in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

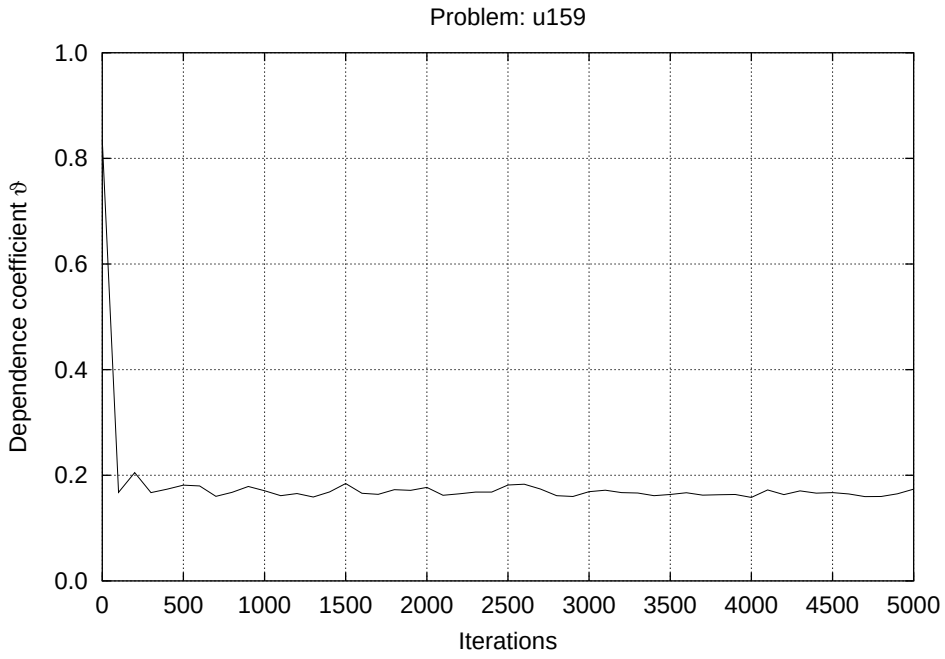


Figure 6.20: The development of the dependence coefficient in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

Analogous to the observations for problem instance u159 presented earlier, the initial species diversities are close to the maxima, whereas the population diversities are slightly lower. Furthermore the dependence coefficients are on the same levels and the populations exhibit a tourlength of the best individual of approximately 49,160 and an average of approximately 55,370.

6.3.2 Example Experiments

Again first an elitist evolution program is realized by taking over the 10 % best individuals from the old population to the new one. As an example, experiment 53 with crossover probability 0.8, uniform crossover probability of 0.1, mutation probability 0.025 and a population of 300 individuals is considered. The plots are very similar to that of instance u159 with the same parameter setting presented above. Figure 6.21 shows the development of the tourlength of the best individual and the population average during the run, figure 6.22 the development of the species diversity, figure 6.23 the development of the population diversity and figure 6.24 the development of the dependence coefficient ϑ. The most notable difference to the behavior with instance u159 is the better and faster saturation in fitness. I. e. the shortest tourlength in the population here is closer to the optimum than for instance u159 and it takes less generations to finish the improvement process. Again uniformity increases after the finish of the assimilation.

Reducing the population size in experiment 52 to 50 individuals again retards the improvement process significantly, which is shown in figure 6.5. The plot of the development of the species diversity in figure 6.6 shows more jags than for instance u159. This also leads to a plot of the dependence coefficient with more jags shown in figure 6.8, because the development of the population diversity shown in figure 6.7 remains relatively noise free. Again the gap between the finish of assimilation and that of the improvement increases with reduced population size.

Switching to a non elitist population replacement scheme again leads to results very similar to that proposed already for instance u159. As an example again the plots of experiment 73 are proposed. The development of the tourlengths is shown in figure 6.29, that of the species diversity in figure 6.30, that of the population diversity in figure 6.31 and that of the dependence coefficient in figure 6.32. The improvement is not as successful as with elitist selection and a lot slower, and there is a large generation gap between the instant assimilation and the improvement to a near optimal level. But again the species diversity remains on the initial near maximum level.

6.3.3 Summary

Summed up, the results for instance rd100 are very similar to that for instance u159. The improvement process is slightly stronger and faster, which probably is caused not only by the smaller problem size but also by the random distribution of the cities, which usually facilitates the working of heuristics.

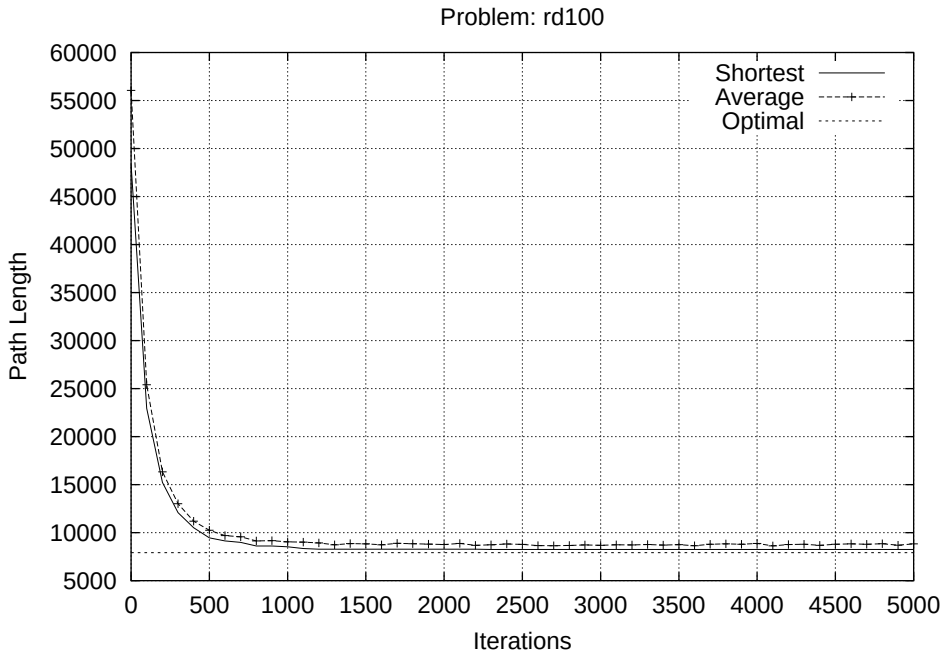


Figure 6.21: The development of best and average tourlengths in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

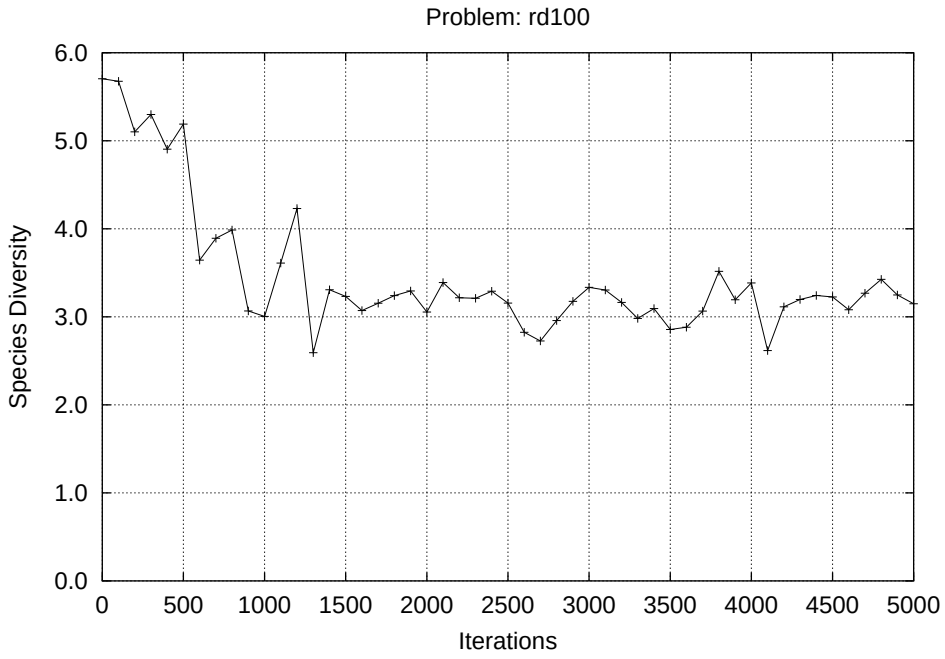


Figure 6.22: The development of the species diversity in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

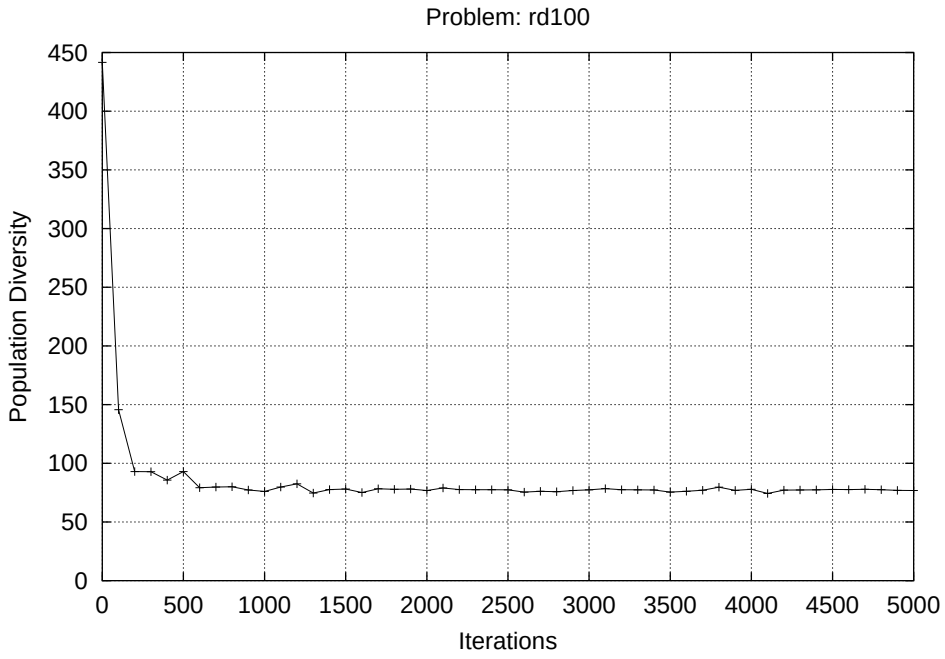


Figure 6.23: The development of the population diversity in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

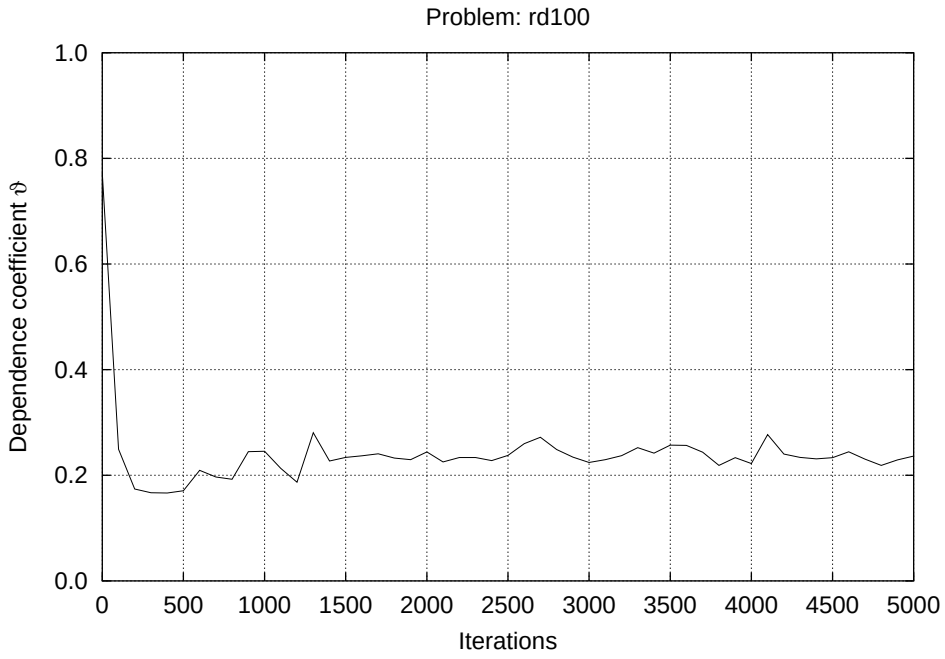


Figure 6.24: The development of the dependence coefficient in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

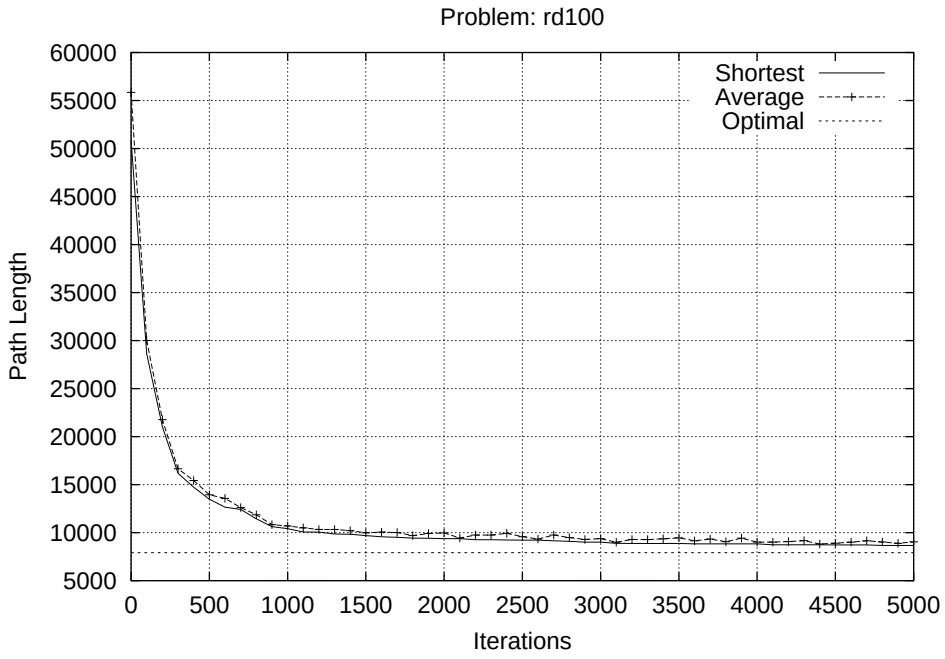


Figure 6.25: The development of best and average tourlengths in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

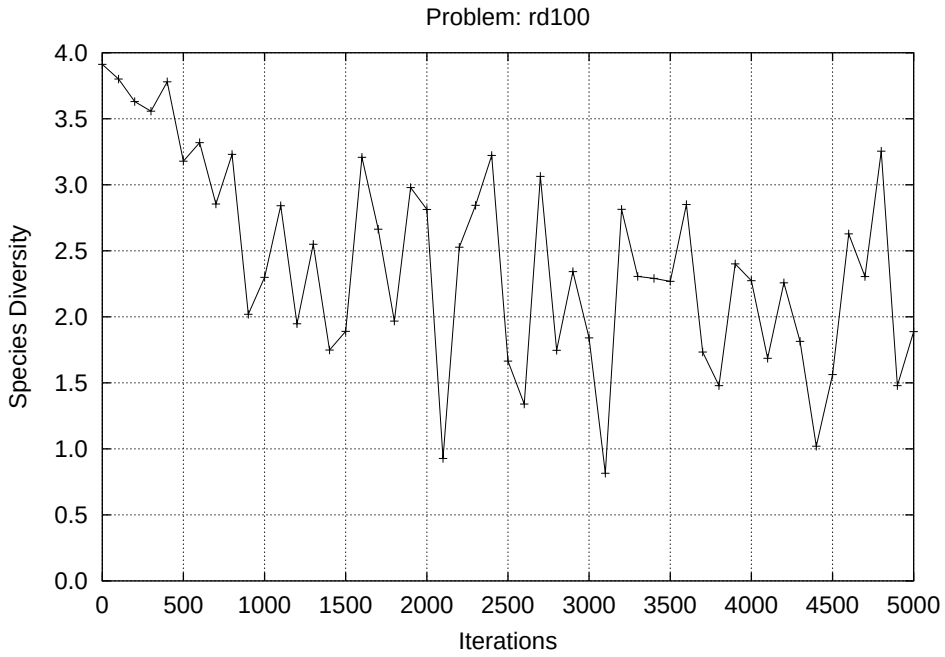


Figure 6.26: The development of the species diversity in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

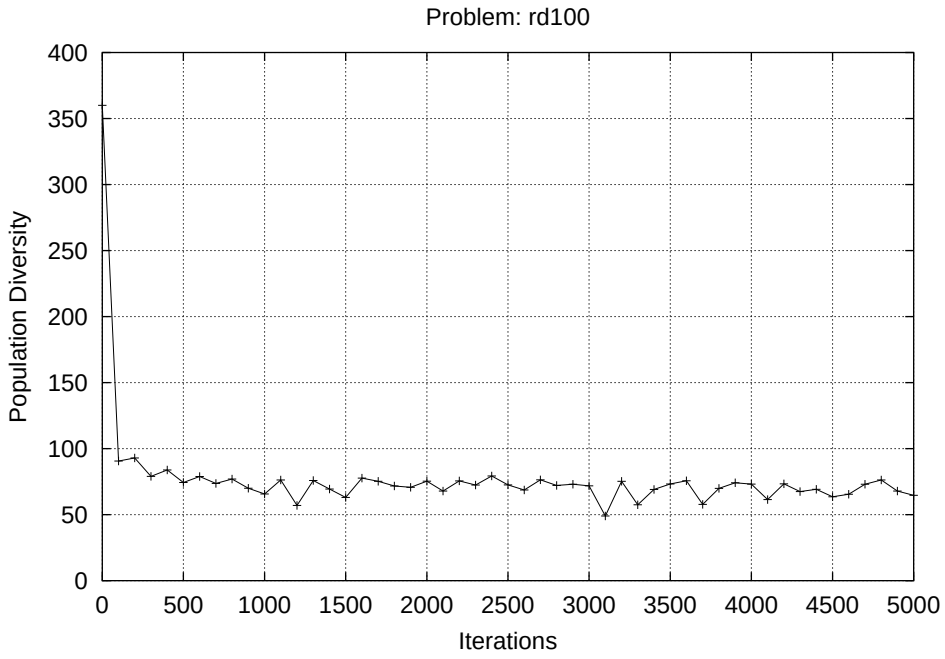


Figure 6.27: The development of the population diversity in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

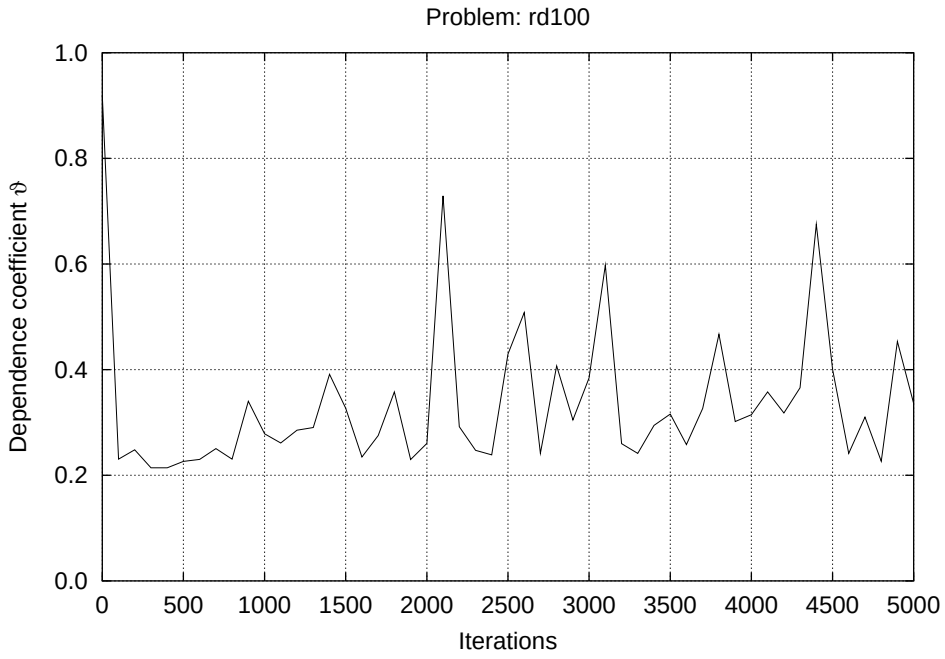


Figure 6.28: The development of the dependence coefficient in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

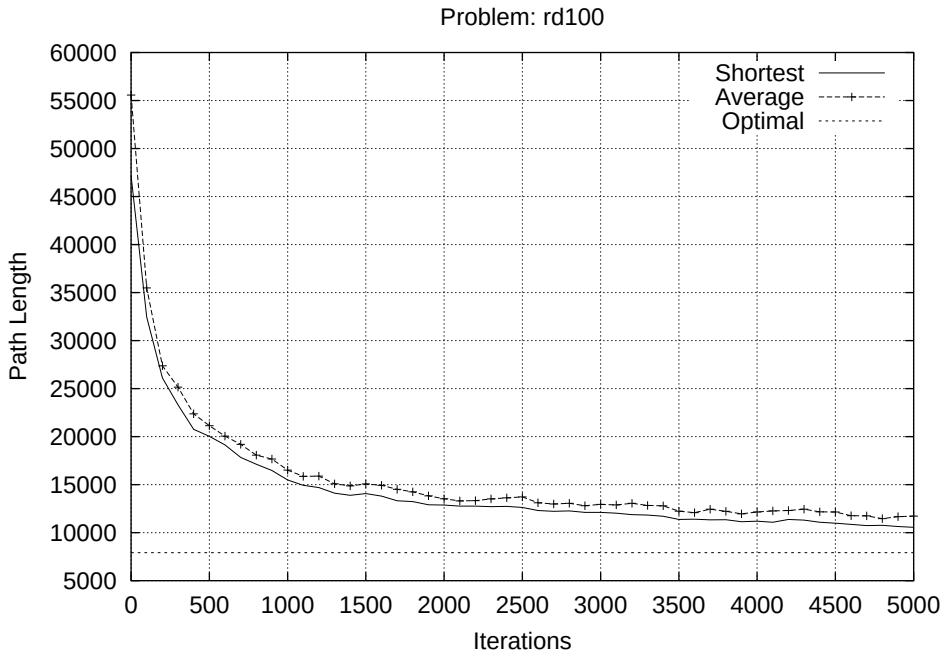


Figure 6.29: The development of best and average tourlengths in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

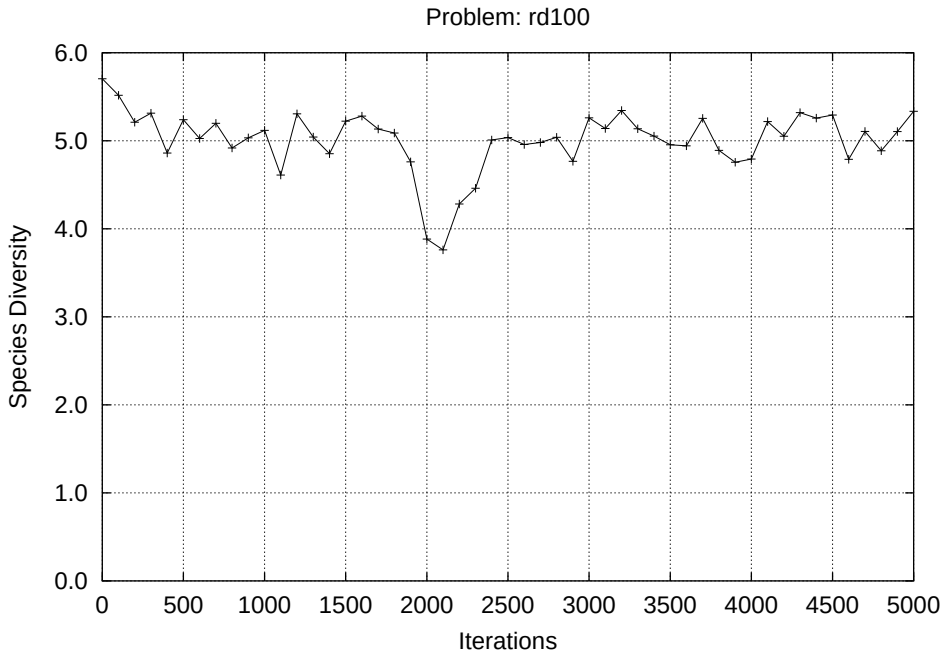


Figure 6.30: The development of the species diversity in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

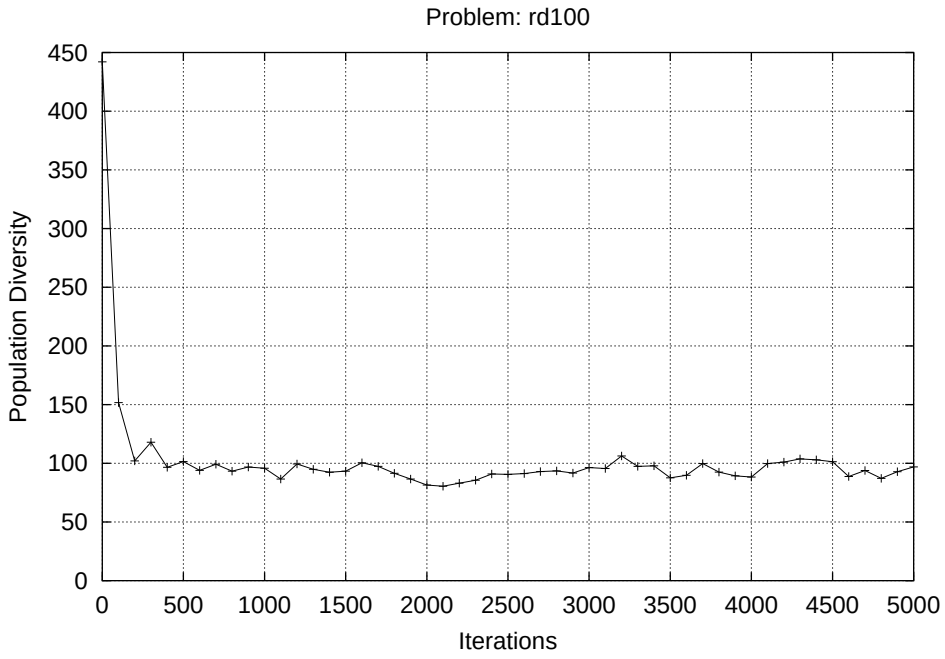


Figure 6.31: The development of the population diversity in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

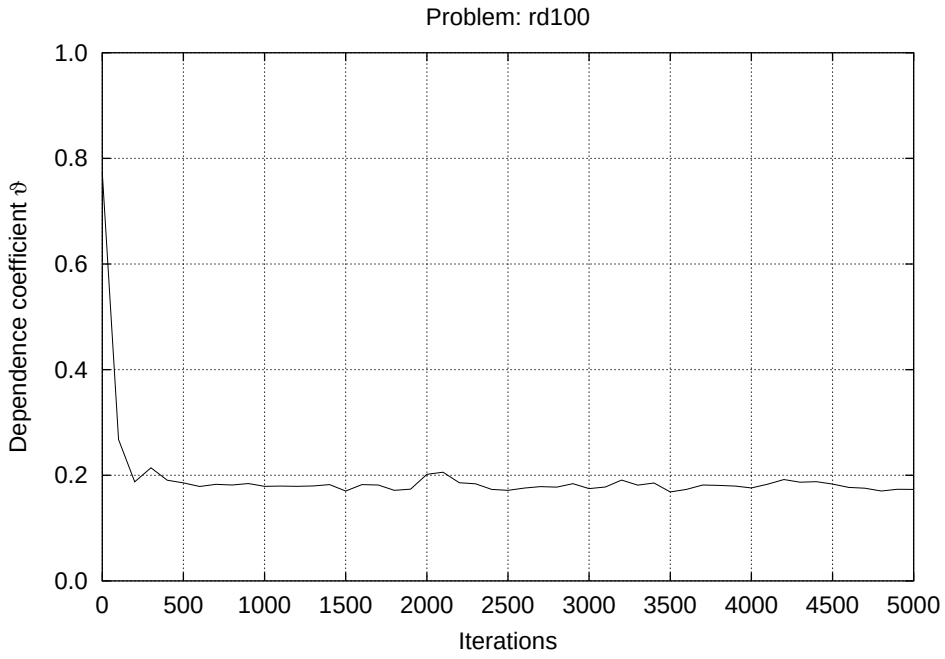


Figure 6.32: The development of the dependence coefficient in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

6.4 Trivial Instance

6.4.1 General Observations

Finally the observations with instance cc100, a 100 cities planar problem instance are proposed. In this instance all cities are placed on a circle and the length of the optimal tour of this instance is 6300. The initial diversities exhibit values analogous to that for the last instance. The tourlength of the best random individual is approximately 104,000 and the average random tourlength is approximately 129,000.

6.4.2 Example Experiments

Again first an elitist evolution program is realized by taking over the 10 % best individuals from the old population to the new one. As an example experiment 53 with crossover probability 0.8, uniform crossover probability of 0.1, mutation probability 0.025 and a population of 300 individuals is considered. The plots are very similar to that of the instances u159 and rd100 with the same parameter setting presented above. Figure 6.33 shows the development of the tourlength of the best individual and the population average during the run. Fortunately TSPGA is able to find the optimal tour with this parameter setting. Figure 6.34 shows the development of the species diversity, figure 6.35 the development of the population diversity and figure 6.36 the development of the dependence coefficient ϑ.

Reducing the population size in experiment 52 to 50 individuals again retards the improvement process significantly, which is shown in figure 6.37. But still the optimum tour is found, although it takes a lot of generations. The species diversity is reduced similarly as for instance rd100 with many jags, which is shown in figure 6.38. The development of the population diversity is shown in figure 6.39. It is reduced quickly to nearly the same level as for a population of 300 individuals. Due to the decreasing species diversity and its jags the dependence coefficient shown in figure 6.40 after its strong initial reduction increases slightly and also the plot becomes very jaggy.

Switching to a non elitist population replacement scheme again leads to results very similar to that proposed already for the instances u159 and rd100. As an example again the plots of experiment 73 are proposed. The tourlengths are improved only very slowly and the algorithm fails to find the optimal tour within the limit number of generations, which is shown in figure 6.41. The development of the species diversity is shown in figure 6.42, that of the population diversity in figure 6.43 and that of the dependence coefficient in figure 6.44.

6.4.3 Summary

Summed up, using an elitist population replacement scheme the optimal solution of this trivial problem instance can be found. If the population size is small, this may take many generations. Without the help of elitism within the limit number of generations the optimum tour could not be found.

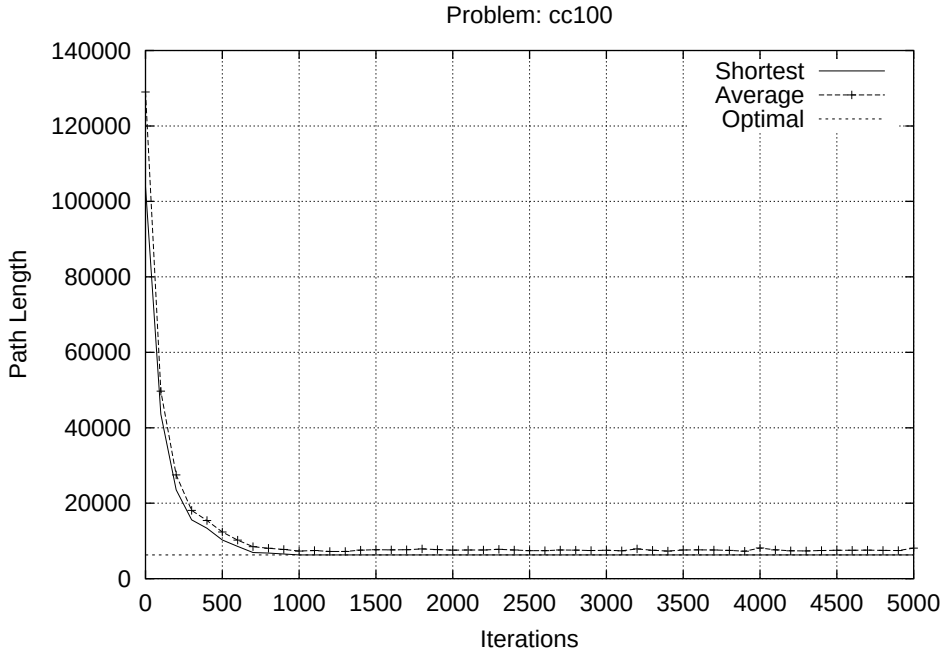


Figure 6.33: The development of best and average tourlengths in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

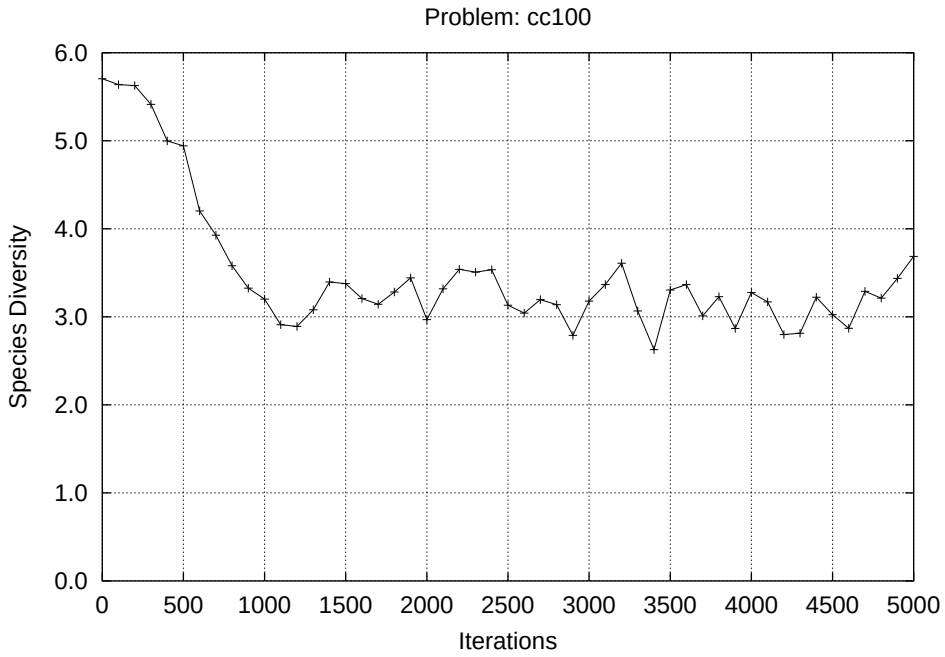


Figure 6.34: The development of the species diversity in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

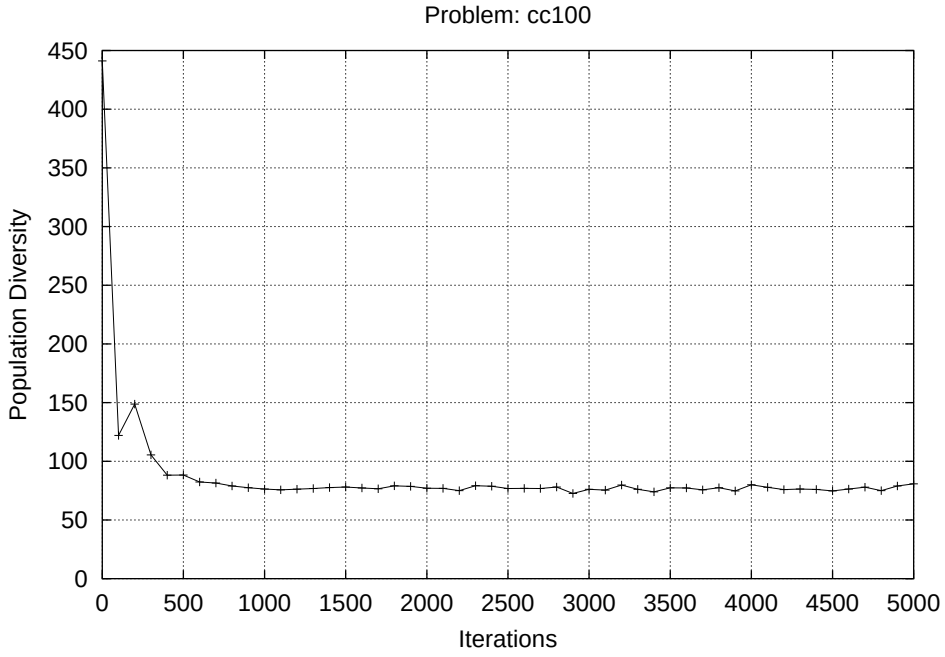


Figure 6.35: The development of the population diversity in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

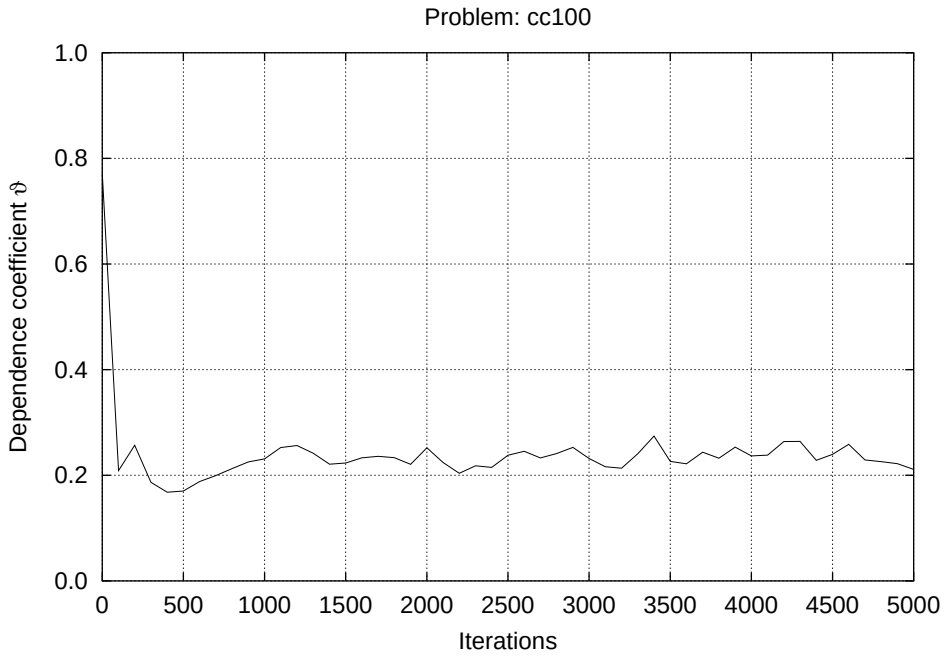


Figure 6.36: The development of the dependence coefficient in experiment 53 with a population of 300 individuals using an elitist replacement strategy.

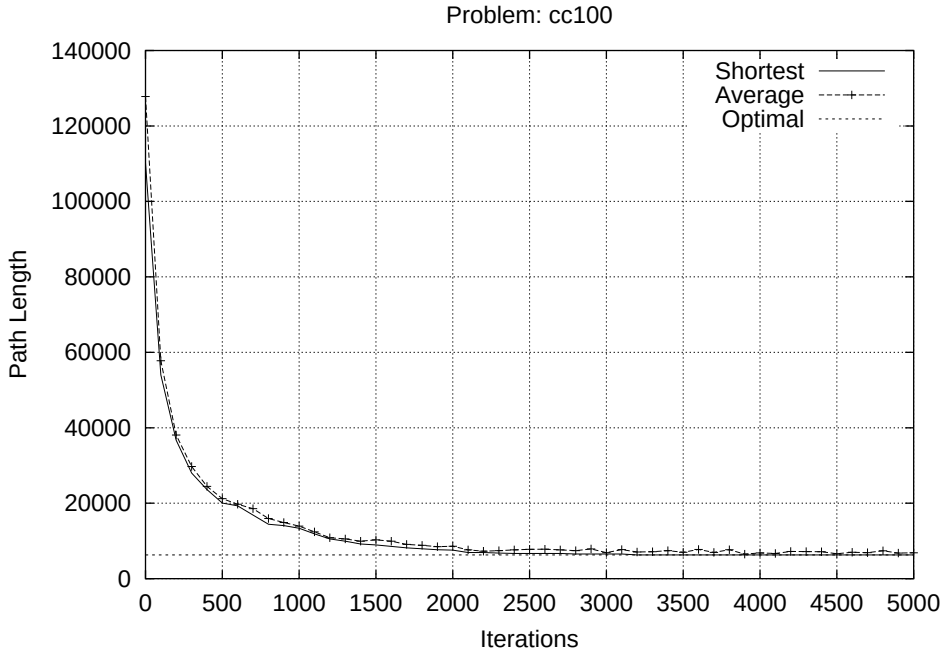


Figure 6.37: The development of best and average tourlengths in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

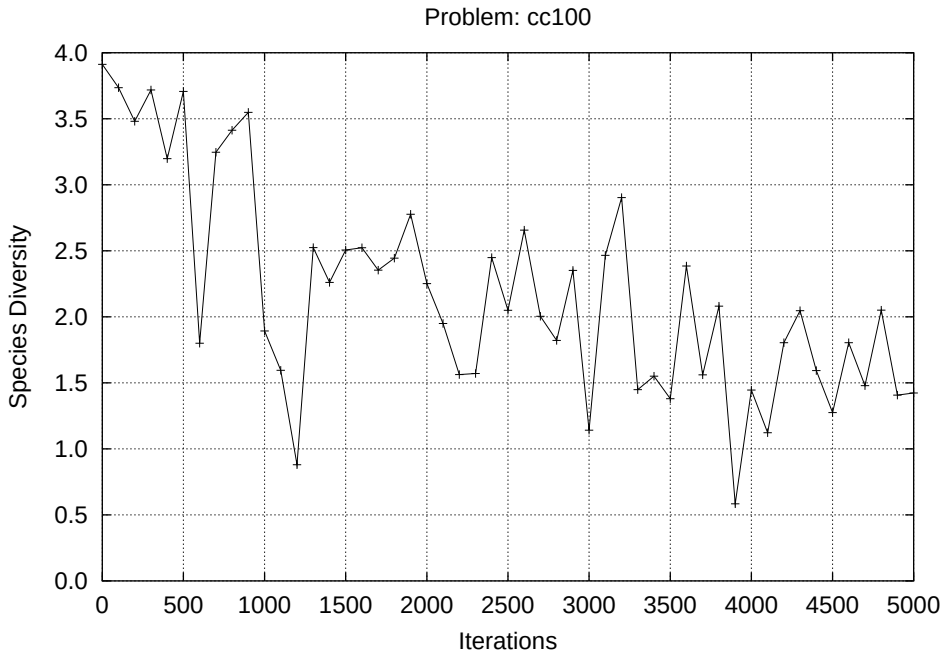


Figure 6.38: The development of the species diversity in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

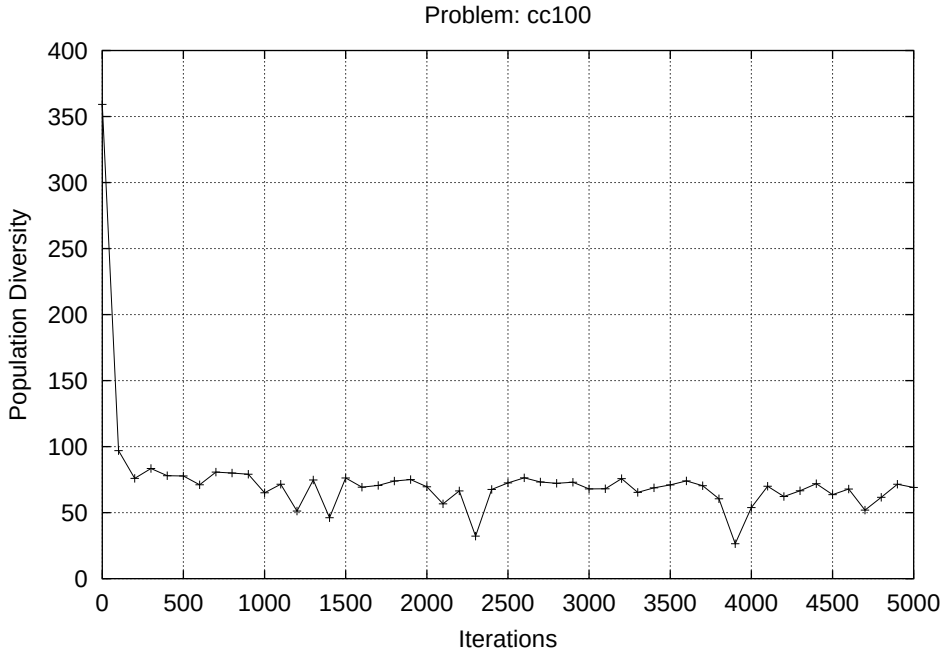


Figure 6.39: The development of the population diversity in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

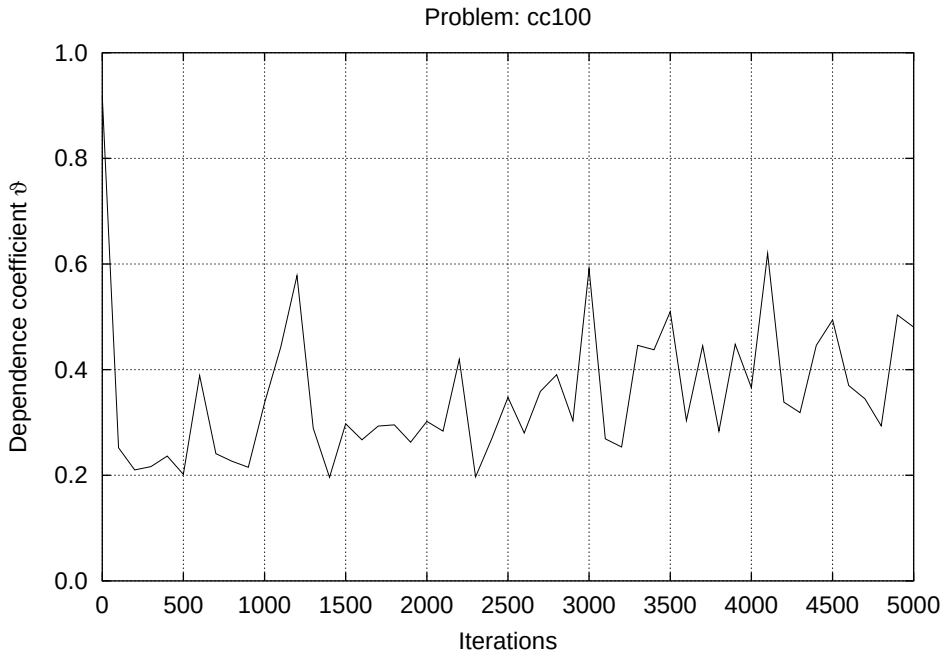


Figure 6.40: The development of the dependence coefficient in experiment 52 with a population of 50 individuals using an elitist replacement strategy.

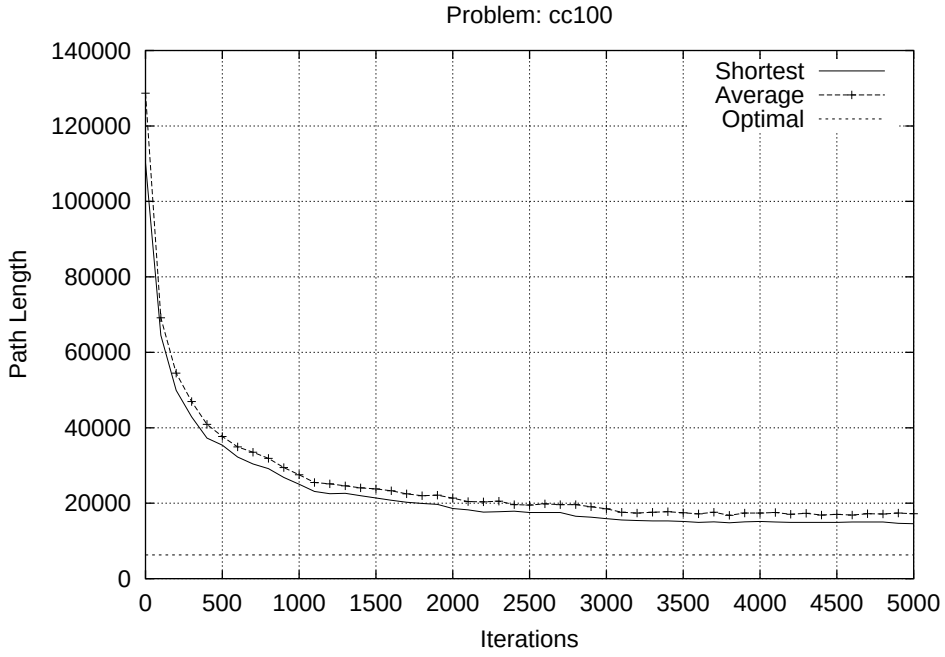


Figure 6.41: The development of best and average tourlengths in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

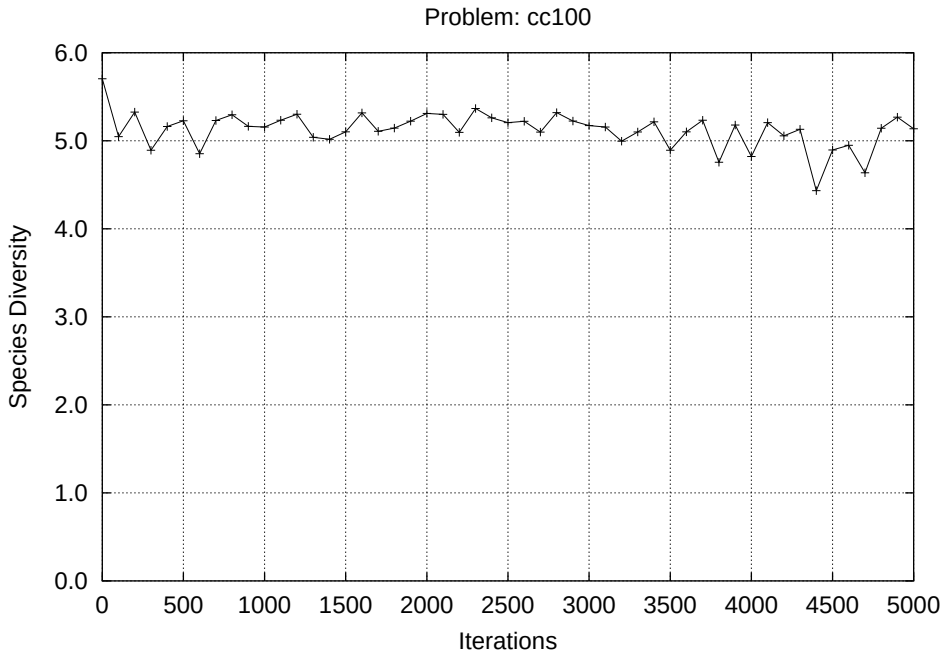


Figure 6.42: The development of the species diversity in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

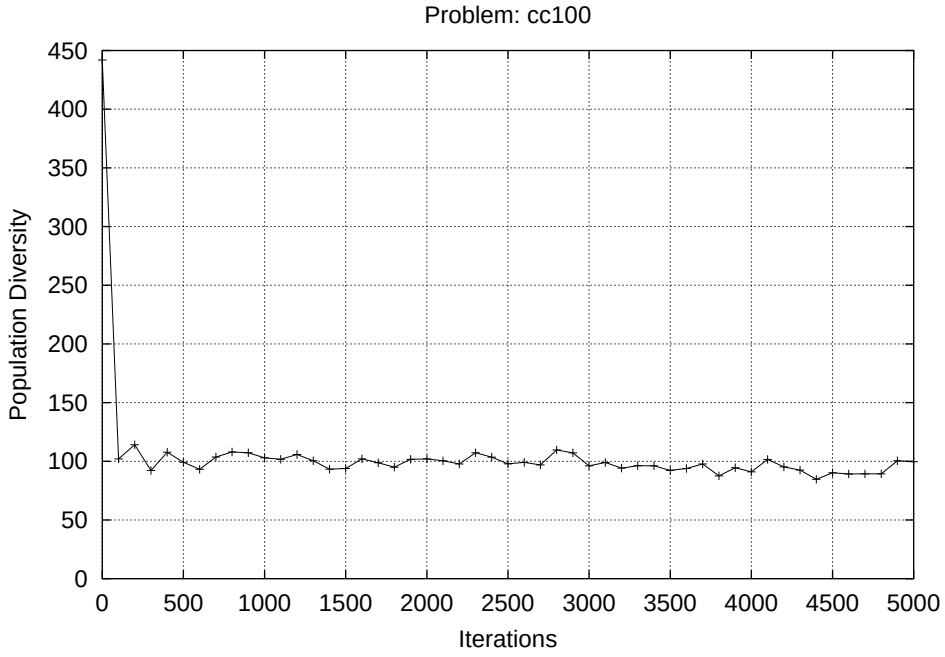


Figure 6.43: The development of the population diversity in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

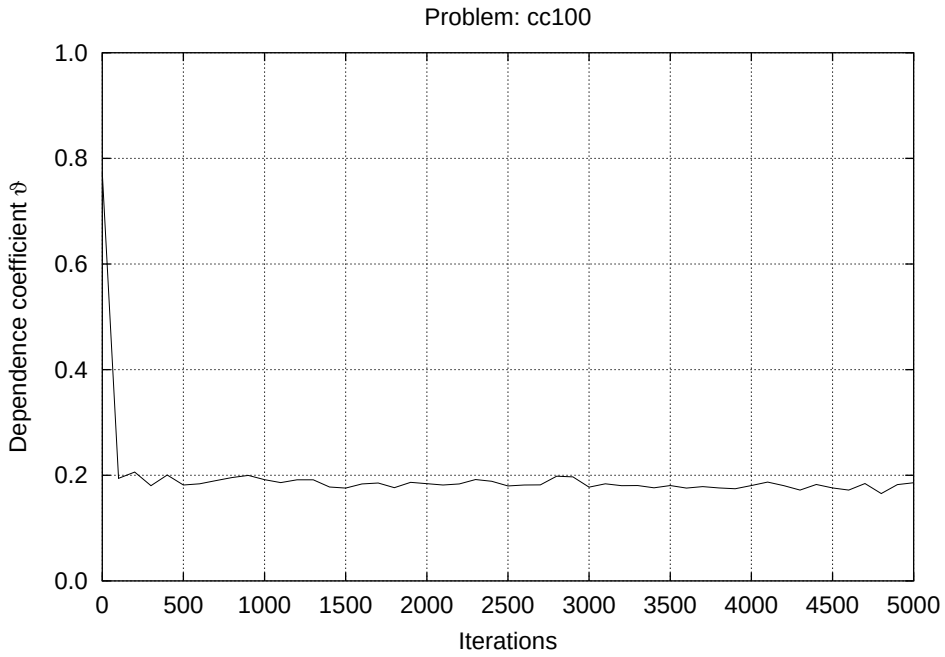


Figure 6.44: The development of the dependence coefficient in experiment 73 with a population of 300 individuals using a non elitist replacement strategy.

6.5 Discussion

The experiments show some interesting behavior and relations between the development of the different indices. In fact three effects have been observed:

improvement (emendation): the development of the best and average fitness in direction to the optimum.

assimilation: the reduction of the population diversity to a low value.

schema propagation between different species: the reduction of the dependence coefficient to a small value.

The last two effects are related by the development of the species diversity. If it remains constantly close to its maximum possible value, both effects coincide.

Elitist population replacement is very robust against a bad choice of parameter values. Also provoked by high mutation and uniform crossover probabilities the experiments show a good improvement behavior and finally tours with lengths close to the optimum are found. For the trivial instance even the optimum is found. Related to the improvement is a strong assimilation which is completed already before the improvement process has finished. Thus there may be a quite long time lag between the completion of both processes. Also during the initial stage schemata are propagated and the reached high amount persists until the program is aborted. But after assimilation has finished the species diversity decreases significantly indicating an increase of uniformity in the population.

Non elitist population is not that successful. The performance of the improvement depends on the parameter values and the population size. If the mutation and uniform crossover probabilities are too high, then there is nearly no improvement and no assimilation. Reducing the probabilities there is a smooth transition from this behavior to successful improvement, assimilation and schema propagation across species borders. But within the monitored number of generations the found solutions are a lot worse than that found with elitist population replacement. Also for the trivial problem the optimum has not been found. Furthermore the species diversity remains nearly constantly on the initial near maximum level. Consequently there is no increase of uniformity during evolution.

The observations of the development of the dependence coefficient ϑ clearly indicate the propagation of schemata across species borders during a successful evolution. It is caused by the assimilation because in the stage of schema propagation the species diversity remains nearly constantly close to the maximum possible value. But for non elitist population replacement the propagation of schemata easily can be obstructed by a bad parameter choice. Also the schema propagation is no guarantee of a good improvement performance, because although there is a strong assimilation there may be an unsatisfactory convergence in fitness due to a non elitist population replacement scheme. But in all experiments performed for this work, assimilation has been a prerequisite of improvement. Thus if there is no assimilation, there will also be no improvement.

The working of the simple evolution program is to focus the search on a specific part of the domain by schema propagation. This is indicated by the straight reduction of the population diversity and the dependence coefficient in the first stage of a successful evolution. The area of constant population diversity then is shifted to further improve the

objective value. Especially for relatively small populations and low mutation and uniform crossover probabilities there is a time lag between the concentration of the search and the approach of the objective value to the optimum, which is the already considered *premature assimilation*. Thus the propagation of schemata goes ahead of the improvement process. After the assimilation has finished, one can assume that a balance of population diversity increasing and reducing processes has established itself. Mutation increases the population diversity if a change of a gene leads to a more uniform distribution at the respective locus. The probability to do so increases as the diversity of the corresponding locus decreases. Crossover normally does not influence the population diversity because it changes only the affiliation of attribute values to individuals of the population. In TSPGA however the diversity is influenced because the reversion of a subtour in a 2-change step changes the representation at loci not involved directly in the edge exchange. But several sequential steps may be without any effect to the diversity. Both processes also increase the species diversity, because the resulting new individuals may belong to species not present in the population. Finally selection removes individuals and therefore also their genes from the population whereas already existing are copied, which reduces the diversities and promotes uniformity, because copied individuals may not be altered by mutation or crossover.

Chapter 7

Conclusions

In this work, two approaches to evolution programs have been presented until now to explore schema propagation during evolution. First, a theoretical model of the random fitness proportional selection process has been considered deeply and in several directions. Second, experiments have been performed using the indices developed earlier to monitor the schema reach in a population. Now the relationship between theory and experimental observations must be confirmed.

Generally in all experiments the genetic drift seems to play an important role, because the fitness of the best individual is close to the average fitness of the population. Thus transferred to the theoretical model of two species selection the fitness coefficient is close to one. In the experiments a uniform population is prevented by the genetic operators mutation and crossover, which constantly introduce small changes that retain species diversity. But the genetic drift promotes the dominance of a few schemata in the population reaching across different species if mutation and crossover rates are low enough, which is indicated by the final low dependence coefficient of approximately 0.2 after successful evolution. This also gives rise to the assumption that the schemata are very similar. Consequently, the genetic drift mainly causes one main direction of search to be there also if the fitness differences in a population are low. In fact, the genetic drift is the main cause of premature assimilation, which is therefore an unavoidable accompaniment of the random fitness proportional selection process.

The genetic drift also causes the low survival chance of a new but superior individual in the Markov model without mutation, which corresponds to the relatively bad performance of the evolution program TSPGA in the experiments without elitism. There the probability to recreate an individual of an extincted class from a persisting one by mutation and crossover decreases with an increasing dissimilarity to the remaining population. Thus the real world selection process without elitism can be suitably modeled by the Markov selection model for two species without mutation. At least some superior classes will randomly survive and spread, which then improves the fitness of the population. But this process can easily be obstructed by setting the crossover and mutation rates to high. Then individuals of a superior class are destroyed before they can spread, which inhibits evolution and also assimilation. Consequently also from a theoretical point of view, a missing assimilation is an indicator of an unsuccessful evolution. If an assimilation takes place, then that happens immediately after the start and takes only a relative small number

of generations, which corresponds to the low average time to absorption in the Markov model without mutation.

Using elitist population replacement the effects considered above are inhibited at least partly. Then the fraction of the best individuals is always copied unchanged to the new population. Consequently, a new superior individual can no longer be eliminated by genetic drift or changed by the genetic operators. Also in the following generations it repeatedly has the chance to spread, which explains the better performance of TSPGA when using an elitist population replacement scheme. After improvement has finished, the persistence of the best fraction leads to an increase of uniformity, because if a copy of the best individual is randomly not changed by the genetic operators, it will also persist. This effect has been observed in the experiments with elitism by the decreasing species diversity. In the fraction not covered by the elitist population replacement, all effects mentioned earlier are also working. Thus, there is also a rapid assimilation caused by the genetic drift, which eliminates classes randomly from the population and promotes the dominance of a few schemata, which are closely related.

The final balance distribution of the two species Markov model with mutation corresponds to the steady respectively saturated state, to which the evolution process converges. The mutation and crossover destroys existing individuals as long as they are not persistent due to elitism. From them, individuals belonging to other possibly new classes are created, which counteracts the genetic drift trying to enforce uniformity. Consequently, there is a balance of creation and elimination, which has been modeled in the two species model by the different mutation rates.

Summed up, in evolution schema propagation is a prerequisite of improvement. Without schema propagation the evolution cannot focus on a subdomain of the problem domain, from which then a further search starts. Schema propagation can be observed during the run of an evolution program or genetic algorithm by monitoring the suitable indices proposed in this work. The different variants of the theoretical model proposed here can be used to explain the experimental results. But the simplification to two species, that enables the tractability of the proposed model, always requires a careful decision, which variant must be used to explain a certain effect. Since there are not even explicit solutions for the the simple Markov models with different fitnesses, only a numerical analysis will be possible for more complicated situations.

Appendix A

Mathematical Prerequisites

A.1 Sum Formulas

The sum

$$\sum_{j=0}^{t-1} a^j b^{t-1-j} = a^{t-1} + a^{t-2}b + a^{t-3}b^2 + \ldots + b^{t-1}$$

often needs to be calculated. It can be transformed easily to

$$\sum_{j=0}^{t-1} a^j b^{t-1-j} = a^{t-1} \sum_{j=0}^{t-1} \left(\frac{d}{a}\right)^j.$$

Now the well known sum formula [45, p. 27]

$$\sum_{j=0}^{w-1} \alpha^j = \frac{\alpha^w - 1}{\alpha - 1} \tag{A.1}$$

can be applied, which leads to

$$\sum_{j=0}^{t-1} a^j b^{t-1-j} = \frac{a^t - b^t}{a - b}. \tag{A.2}$$

A.2 Difference versus Differential Equations

Generally corresponding differential and difference equations have the same fixed points, but their asymptotic stability properties may differ [45, p. 25]. As stated above, to a first order differential equation (4.32) with equation (4.30) and constant step size h the differential equation

$$x(t+1) = x(t) + h\, f(x(t)) = g(x(t)) \tag{A.3}$$

corresponds. As already noted, this is a reversal of Euler's approach to solve a differential equation numerically, which can not be solved explicitly.

Since of the stability criterion (4.21) an asymptotically stable fixed point z of the difference equation has to satisfy the inequality

$$\left|\frac{dg}{dx}(z)\right| = \left|1 + h\,\frac{df}{dx}(z)\right| < 1.$$

From that with a case separation for $h > 0$

$$-\frac{2}{h} < \frac{df}{dx}(z) < 0$$

follows for a fixed point z of the differential equation to be stable also for the corresponding difference equation. Consequently the step size h has to be chosen small enough to ensure asymptotic stability in this case.

In the selection model considered in section 4.3 the approach is reversed. Asymptotic stability of a fixed point of the difference equation generally ensures that also for the corresponding differential equation independently of the step size $h > 0$.

But generally there must be no similarity of the solutions of a difference equation and its corresponding differential equation constructed using the approach of equation (A.3). E. g. to the difference equation

$$y(t+1) = 1 + y(t) + y^2(t),$$

which has no singularities, the differential equation

$$\dot{y} = \frac{1}{h}\left(1 + y^2\right)$$

corresponds. Clearly its solution is $y(t) = \frac{1}{h}\tan t$, which is even, i. e. $y(t) = -y(-t)$ holds, periodic, i. e. $y(t) = y(t + n\pi)$ holds with $n \in \mathbf{Z}$, and has singularities for $t = \frac{\pi}{2} + n\pi$. Consequently, the solutions of the corresponding difference and differential equations are totally different. Thus the similarity of the difference equation (4.19) and the corresponding differential equation given by equation (4.31) at least for $\alpha \approx 1$ is remarkable. The theory behind this subject in depth is considered by Kato [52, p. 509–515], which is mainly based on Trotter's article [95].

Appendix B

Extensions to the Population State Indices

B.1 Information Theoretic Additions

B.1.1 From Information to Entropy

The consideration of information is closely related to the notion of a *message*. A message $\mathcal{M}$ is an ordered sequence of symbols $s_1\, s_2\, \ldots\, s_N$. Each symbol s_i of the message is taken from a domain or alphabet $\mathbf{A} = \{a_1, \ldots, a_n\}$. Alternatively this is often called a *string*. Now following Brillouin [10, Chapter 1] the information of a message is defined to be

$$I = K \ln P \tag{B.1}$$

with P being the number of different messages that can be composed from the symbols contained in the message and K an arbitrary constant. But the number of different messages P is the number of permutations of M with repetitions, thus by following combinatorial theory [11, p. 110]

$$P = \frac{N!}{N_1!N_2!\cdots N_n!} \tag{B.2}$$

with N_j being the frequency of symbol a_j in the message. Combining equations (B.1) and (B.2) leads to the information

$$I = K \ln \frac{N!}{\prod_{j=1}^{n} N_j!}. \tag{B.3}$$

If two separate messages are concatenated, then each permutation of the first message can be combined with each of the second. Doing so, the number of possible combined messages is

$$P = P_1 \cdot P_2$$

giving

$$I = K \ln(P_1 \cdot P_2) = I_1 + I_2 \tag{B.4}$$

with

$$I_1 = K \ln P_1 \quad \text{and} \quad I_2 = K \ln P_2$$

for their individual informations. Thus an important property of the information is its additivity under the condition of the independence of the related messages. In section B.1.2 this property will be reexamined and some consequences be considered.

Setting $K = 1$ from equation (B.3) the average information per symbol

$$H^B = \frac{1}{N} \ln \frac{N!}{\prod_{j=1}^{n} N_j!}. \tag{B.5}$$

follows, which will be referred as Brillouin entropy in the following. If all the N_j are very large, then the approximation

$$\ln n! \approx n(\ln n - 1) \tag{B.6}$$

can be applied. Doing so with some transformations leads to Shannon's entropy H [89] of equation (5.14) with the relative frequencies r_j in the message instead of the probabilities p_j.

Brillouin's average information per symbol H^B also is used as species diversity index [59, 74], but it does not satisfy invariance under cloning [53, p. 113]. Instead for fixed relative frequencies r_j it is an increasing function of the number of symbols N of the message [73, p. 372].

Finally the Brillouin diversity coefficient

$$d^B = \frac{N\, D^B}{\ln N!} = \frac{1}{\ln N!} \ln \frac{N!}{\prod_{j=1}^{n} N_j!}$$

between zero and one can be deduced. Its properties here are not explored any further.

B.1.2 Diversity of Combined Sets

The combination of two submessages to a message in equation (B.4) shall be reconsidered here to be the combination of subpopulations to a population using two attributes. Then the diversity of the first attribute A_1 is examined and a second attribute A_2 specifies the affiliation of an individual to a subpopulation. Thus $D(A_1|A_2 = a_j)$ is the diversity of attribute A_1 in subpopulation a_j, which equals the information per character in the corresponding submessage. Combining the subpopulations the average subpopulation diversity $D(A_1|A_2)$ is obtained, which is equivalent to the information per character of a whole message combined from the submessages under the condition of their independence. Contrary, the diversity of attribute A_1 respectively information per character in the whole population is $D(A_1)$. From equation (5.36) using the Shannon diversity

$$D^S(A_1) - D^S(A_1|A_2) = I(A_1, A_2) \geq 0 \tag{B.7}$$

follows, which also proves the concavity of the Shannon diversity used here.

Analogously to the considerations in the last section these relations can be interpreted. If the subpopulations are statistical independent, then $D^S(A_1|A_2) = D^S(A_1)$ and

$I(A_1, A_2) = 0$ hold. Consequently, with equation (5.34) the relative frequencies of the considered attribute A_1 in each subpopulation must be constant. Thus the subpopulations share the same structure.

Conversely, if in each subpopulation there is only one unique value of the considered attribute A_1 and therefore $D^S(A_1|A_2 = a_j) = 0$ for all subpopulations $1 \leq j \leq m$, then there is a strict one to one dependence between attribute A_1 and attribute A_2, which specifies the affiliation of an individual to a subpopulation. Consequently, $D^S(A_1|A_2) = D^S(A_2|A_1) = 0$ follows. Furthermore because of equations (5.22) and (5.36) the equation

$$D^S(A_1) = D^S(A_2) = D^S(A_1A_2) = I(A_1, A_2)$$

holds. Thus, the increase of diversity is limited by the diversity of attribute A_2, which specifies the decomposition of the set into subsets. Consequently

$$D^S(A_1) \leq D^S(A_1|A_2) + D^S(A_2) \quad \text{and} \quad I(A_1, A_2) \leq D^S(A_2)$$

hold. Unfortunately, knowing only $D^S(A_1|A_2)$ and $D^S(A_2)$ it is impossible to calculate $D^S(A_1)$ directly from both quantities.

If the diversities D are calculated using an index based on a generalized average distance, e. g. the Gini-Simpson-index, then the average distance of the whole set can also be separated into distance portions within the subsets and among them. Rao called that *apportionment of diversity* [79, p. 72*ff*]. Doing so, the distance portion among the sets corresponds to the mutual information considered above.

Summed up, the increase of diversity resulting from a combination of sets depends on their *heterogeneity*, i. e. here the dissimilarity of their attribute value distributions. Using the Shannon diversity it is measured by the mutual information, which is limited by the diversity of the subset assignment. If the considered attribute has the same value within a subset, then the diversity of the union equals the heterogeneity of the subsets. Since each set can be partitioned into subsets of uniform individuals, the heterogeneity and the diversity of a population are synonymous.

B.2 Application to Classification Theory and Cluster Analysis

B.2.1 Heterogeneity

In classification theory and cluster analysis the heterogeneity of a set of multivariate objects often is defined to be the average distance of equation (5.3) [8, p. 91*f*], which can be generalized to Rao's diversity of equation (5.12) alternatively. Thus also there heterogeneity in fact is equivalent to diversity. For sets of objects with nominal scaled attributes then the Hamming distance of equation (5.2) is the natural distance function, which results in the heterogeneity indices of equation (5.7) respectively (5.13). This fact generally has not been recognized [53, p. 96*ff*]. E. g. Bock [8, p. 99*f*] has proposed the Gini-Simpson index of equation (5.11), but has not considered its relation to the average distance. Instead he has recommended the *information content* or *total information*

$$I(\boldsymbol{P}) = N \sum_{j=1}^{M} D^S(N_{j,1}, \ldots, N_{j,i}, \ldots, N_{j,n_j}) \tag{B.8}$$

to measure the heterogeneity of the set or population $\boldsymbol{P}$ of size N [8, p. 95]. Clearly, this quantity does not satisfy the properties of invariance under cloning [53, p. 103] and of concavity, which have been claimed for diversities in section 5.3. Instead from the concavity of the Shannon diversity

$$I(\boldsymbol{P}) \geq \sum_{i=1}^{n} I(\boldsymbol{P}_i)$$

follows, which is called superadditivity. Consequently the total information is not a suitable heterogeneity index.

B.2.2 Patterns versus Schemata

In machine learning one important task is to identify patterns in a dataset of objects having multiple nominal attributes and then to infer classification rules or cluster the dataset. In fact a pattern is equivalent to a schema in evolution programs. But the heterogeneity indices used in machine learning usually are distance based or similar to the population diversity like the total information proposed above. Thus they consider each attribute independent from the remaining. This situation is very schizophrenic because the task is to identify dependences. But until now no index has been proposed in machine learning to measure the dependences and pattern reach in a data set.

The considerations made for populations processed by evolution programs also can be applied here. Consequently, it is impossible to extract any rules from a set of objects having $\vartheta = 1$ and easiest if $\vartheta = 0$ holds. Thus it is now possible to rate the performance of rule extraction of a machine learning algorithm in relation to the dependencies inherent in the set of objects on which it acts. The heterogeneity also can be interpreted to be the redundancy during a classification process. If $\vartheta = 1$ holds then each attribute is sufficient for the classification and the remaining can be used to correct errors. Conversely, if $\vartheta = 0$ holds then all attributes have to be evaluated and an error always leads to a false classification.

Finally an application to cluster analysis and classification theory is possible. An approach based on ideas similarly to the one introduced in this article already has been proposed by Watanabe [99, p. 409*ff*] and later been adopted by Guiaşu [39, p. 72*ff*]. But their realization is very strange, because it results in the subadditivity of the entropy respectively diversity in case of the union of the subsets. Also Bock was not very convinced of this approach [8, p. 99], although he esteemed it worth to discuss.

B.2.3 Combining Sets

Now the situation of combining subsets of objects in section B.1.2 is reconsidered and extended to three attributes A_1, A_2, A_3, where A_3 specifies the affiliation to a subset. Then for each of the subpopulations $\boldsymbol{P}_k$ from equation (5.35)

$$D(A_1A_2|A_3 = a_{3,k}) = D(A_1|A_3 = a_{3,k}) + D(A_2|A_3 = a_{3,k}) - I(A_1, A_2|A_3 = a_{3,k})$$

follows. Forming the average leads to

$$D(A_1A_2|A_3) = D(A_1|A_3) + D(A_2|A_3) - I(A_1, A_2|A_3)$$

and combining the subpopulations to equation (5.35)

$$I(A_1, A_2) = D(A_1) + D(A_2) - D(A_1 A_2).$$

For each diversity the heterogeneity between the subsets is defined to be the increase induced by unconditioning respectively combining because of the concavity of the diversity. For the redundant information

$$\Theta(A_1, A_2) - \Theta(A_1, A_2|A_3) = I(A_1, A_2, A_3)$$

follows from equation (5.47) because of the equivalence of mutual and redundant information for two arguments. For a successful classification the conditional dependence, which equals the average dependence of the subsets, should be lower than the dependence in the whole set, because a low dependence indicates a high reach of patterns. Thus $I(A_1, A_2, A_3)$ has to be positive. From this consideration new classification and clustering algorithms can be inferred, which is beyond the scope of this work. Preuss and Vorkauf [77] already proposed an approach based on concepts similar to those presented above.

Bibliography

[1] Lee Altenberg. Evolutionary computation models from population genetics: Part 2: An historical toolbox. slides presented at the Congress on Evolutionary Computation, La Jolla Marriott, San Diego, USA, July 16–19, 2000.

[2] Barry C. Arnold. *Majorization and the Lorenz Order: A Brief Introduction*. Number 43 in Lecture Notes in Statistics. Springer-Verlag, Berlin et al., 1987.

[3] Thomas Bäck. Selective pressure in evolutionary algorithms: A characterization of selection mechanims. In *Proceedings of the first IEEE International Conference on Evolutionary Computation, 1994 (ICEC'94)*, pages 57–62, Orlando, FL, June 27–29, 1994. IEEE, IEEE Service Center, Piscataway, NJ, 1994.

[4] Thomas Bäck. *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, New York, NY, 1996.

[5] David Beasley, David R. Bull, and Ralph R. Martin. An overview of genetic algorithms: Part 1, fundamentials. *University Computing*, 15(2): 58–69, 1993.

[6] David Beasley, David R. Bull, and Ralph R. Martin. An overview of genetic algorithms: Part 2, research topics. *University Computing*, 15(4): 170–181, 1993.

[7] Wayne D. Blizard. Multiset theory. *Notre Dame Journal of Formal Logic*, 30(1): 36–66, 1989.

[8] Hans Hermann Bock. *Automatische Klassifikation: Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten (Cluster-Analyse)*, volume 24 of *Studia Mathematica/Mathematische Lehrbücher*. Vandenhoeck & Ruprecht, Göttingen, 1974.

[9] Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Number 31 in Texts in Applied Mathematics. Springer-Verlag, New York, NY, et al., 1999.

[10] Léon Brillouin. *Science and Information Theory*. Academic Press, New York, NY, et al., second edition, 1962.

[11] I. N. Bronstein and K. A. Semendjajew. *Taschenbuch der Mathematik*. Verlag Harri Deutsch, Thun, Frankfurt am Main, 22th edition, 1985.

[12] Gerhart Bruckmann. *Konzentrationsmessung*, chapter 26, pages 191–196. WiSt-Studienkurs. Verlag Franz Vahlen, München, 11th, revised edition, 1998.

[13] Reinhard Bürger. *The Mathematical Theory of Selection, Recombination, and Mutation*. Wiley Series in Mathematical and Computational Biology. John Wiley & Sons, Chichester et al., 2000.

[14] E. K. Burke, S. Gustafson, and G. Kendall. Diversity in genetic programming: An analysis of measures and correlation with fitness. *IEEE Transactions on Evolutionary Computation*, 8(1): 47–62, February 2004.

[15] Uday Chakraborty, Kalyanmoy Deb, and Mandira Chakraborty. Analysis of selection algorithms: A markov chain approach. *Evolutionary Computation*, 4(2): 133–167, 1996.

[16] Nicos Christofides. Worst case analysis of a new heuristic for the travelling salesman problem. report 388, Graduate School of Administration, Carnegie-Mellon University, Pittsburgh, PA, 1976.

[17] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, Cambridge, MA, second edition, 2001.

[18] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, 1991.

[19] Frank A. Cowell. *Measuring Inequality*. LSE Handbooks in Economics. Prentice Hall/Harvester Wheatsheaf, London, second edition, 1995. data disk enclosed.

[20] Charles Darwin. *On the Origin of Species*. John Murray, London, 1859.

[21] Theodosius Dobzhansky. *Genetics and the Origin of Species*. 1937.

[22] Wolfgang Domschke. *Logistik: Rundreisen und Touren*. Oldenbourgs Handbücher der Wirtschaftswissenschaften und Sozialwissenschaften. R. Oldenbourg Verlag, München, 1985.

[23] M. Eigen, J. McCaskill, and P. Schuster. The molecular quasispecies. *Advances in Chemical Physics*, 75: 149–263, 1989.

[24] Leonhard Euler. Solution d'une question curieuse qui ne paroit soumise à aucune analyse. *Mémoires de l'Académie Royale des Sciences et Belles-Lettres*, (15): 310–337, 1759.

[25] Warren J[ohn] Ewens. *Mathematical Population Genetics*, volume 9 of *Biomathematics*. Springer-Verlag, Berlin et al., 1979.

[26] Ludwig Fahrmeir, Alfred Hamerle, and Gerhard Tutz, editors. *Multivariate statistische Verfahren*. Walter de Gruyter, Berlin, New York, NY, second, revised edition, 1996.

[27] Ronald H. Fisher. *The Genetical Theory of Natural Selection*. Dover Publications, New York, NY, second, revised edition, 1958. Revised and enlarged version of the work originaly published 1930.

[28] M. L. Fredman, D. S. Johnson, L. A. McGeoch, and G. Ostheimer. Data structures for the traveling salesman. *Journal of Algorithms*, 18: 432–479, 1995.

[29] Andreas Frick. A universal object-oriented framework for evolution programs. In Achim Sydow, editor, *IMACS World Congress (15th) on Scientific Computation, Modelling and Applied Mathematics: Numerical Mathematics*, volume 2, pages 639–644, Berlin, August 24–29, 1997. International Association for Mathematics and Computers in Simulation, Wissenschaft und Technik Verlag, Berlin, 1997.

[30] Andreas Frick. TSPGA – an evolution program for the symmetric traveling salesman problem. In Hans-Jürgen Zimmermann, editor, *EUFIT'98 – 6th European Congress on Intelligent Techniques and Soft Computing*, volume 1, pages 513–517, Aachen, September 7–10, 1998. ELITE – European Laboratory for Intelligent Techniques Engineering, Mainz Verlag, Aachen, 1998.

[31] Corrado W. Gini. Variabilità e Mutabilità. In *Studi Economico – Giuridici*, volume 3, pages 3–159. G. Dessi, Cagliari, 1912. Pubblicati per cura della Facoltà di Giurisprudenza, R. Università di Cagliari, Instituto Economico-Giuridico.

[32] David E[dward] Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison-Wesley Publishing Company, Reading, MA, 1989.

[33] David E[dward] Goldberg and Kalyanmoy Deb. A comparative analysis of selection schemes used in genetic algorithms. In Gregory J. E. Rawlins, editor, *The First Workshop on the Foundations of Genetic Algorithms and Classifier Systems*, pages 69–93, Bloomington, IN, 1990. Indiana University, Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[34] David E[dward] Goldberg and Philip Segrest. Finite markov chain analysis of genetic algorithms. In John J. Grefenstette, editor, *Genetic Algorithms and their Applications: Proceedings of the Second International Conference on Genetic Algorithms*, pages 1–8, Cambridge, MA, July 28–31, 1987. Massachusetts Institute of Technology, Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

[35] Martina Gorges-Schleuter. Asparagos96 and the traveling salesman problem. In *Proceedings of the IEEE International Conference on Evolutionary Computation, 1997 (ICEC'97)*, Indianapolis, IN, April 13–16, 1997. IEEE, IEEE Service Center, Piscataway, NJ, 1997.

[36] John J. Grefenstette. Incorporating problem specific knowledge into gentic algorithms. In Lawrence Davis, editor, *Genetic Algorithms and Simulated Annealing*, Research Notes in Artificial Intelligence, pages 42–60. Morgan Kaufmann Publishers, Los Altos, CA, 1987.

[37] M[artin] Grötschel and M[anfred] W. Padberg. Polyhedral theory. In Lawler et al. [55], chapter 8, pages 251–305.

[38] Martin Grötschel and Manfred [W.] Padberg. Die optimierte Odyssee. *Spektrum der Wissenschaft*, pages 76–85, April 1999.

[39] Silviu Guiaşu. *Information Theory with Applications*. Advanced Book Program. McGraw-Hill International Book Company, New York, NY, et al., 1977.

[40] Philip Hartman. *Ordinary Differential Equations*. John Wiley & Sons, New York, NY, 1964.

[41] Aristides T. Hatjimihail. Entropy and genetic algorithms: Definition, and some graphs. technical report I, Hellenic Complex Systems Laboratory, Drama, Greece, October 1993.

[42] Orris C[lemens] Herfindahl. *Concentration in the Steel Industry*. PhD thesis, Columbia University, New York, NY, 1950.

[43] C. de la Higuera and F. Casacuberta. Topology of strings: Median string is NP-complete. *Theoretical Computer Science*, 230(1–2): 39–48, January 2000.

[44] Josef Hofbauer and Karl Sigmund. *Evolutionstheorie und dynamische Systeme: Mathematische Aspekte der Selektion*. Verlag Paul Parey, Berlin, Hamburg, 1984.

[45] Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, UK, et al., 1998. partly replaces [44].

[46] John H. Holland. *Adaption in Natural and Artificial Systems*. The MIT Press, Cambridge, MA, 1975.

[47] Mark H. Holmes. *Introduction to Perturbation Methods*. Number 20 in Texts in Applied Mathematics. Springer-Verlag, Berlin et al., 1995.

[48] Stuart H. Hurlbert. The nonconcept of species diversity: A critique and alternative parameters. *Ecology*, 52(4): 577–586, summer 1971.

[49] D. S. Johnson and C[hristos] H. Papadimitriou. Computational complexity. In Lawler et al. [55], chapter 3, pages 37–85.

[50] D. S. Johnson and C[hristos] H. Papadimitriou. Performance guarantees for heuristics. In Lawler et al. [55], chapter 5, pages 145–180.

[51] Dieter Jungnickel. *Graphen, Netzwerke und Algorithmen*. BI Wissenschaftsverlag, Mannheim et al., third edition, 1994.

[52] Tosio Kato. *Perturbation Theory for Linear Operators*. Classics in Mathematics. Springer-Verlag, Berlin et al., 1995.

[53] Manfred Krtscha. *Zur Axiomatik der Ungleichheitsmessung in der Wirtschaftswissenschaft*. Habilitationsschrift, Universität Karlsruhe (TH), Fakultät für Wirtschaftswissenschaften, April 1996.

[54] Solomon Kullback. *Information Theory and Statistics*. Dover Publications, New York, NY, 1959.

[55] E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, and D. B. Shmoys, editors. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Chichester et al., 1985.

[56] David Levine. *Users Guide to the PGAPack Parallel Genetic Algorithm Library*. Mathematics and Computer Science Division, Argonne National Laboratory, 1996.

[57] S[hen] Lin. Computer solutions of the traveling salesman problem. *Bell System Technical Journal*, 44: 2245–2269, 1965.

[58] S[hen] Lin and B[rian] W. Kernighan. An effective heuristic algorithm for the traveling salesman problem. *Operation Research*, 21: 498–516, 1973.

[59] D. R. Magalef. Information theory in ecology. *General Systems*, 3: 36–71, 1958.

[60] Ernst Mayr. *Systematics and the Origin of Species*. Columbia University Press, New York, NY, 1942.

[61] William J. McGill. Multivariante information transmission. *Psychometrika. A Journal of Quantitative Psychology*, 19(2): 97–116, June 1954.

[62] Johann Gregor Mendel. Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn, Abhandlungen*, 4: 3–47, 1866.

[63] Zbigniew Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, New York, NY, et al., second, extended edition, 1992.

[64] Melanie Mitchell. *An Introduction to Genetic Algorithms*. The MIT Press, Cambridge, MA, 1996.

[65] Jean-Baptiste Pierre Antoine de Monet, Chevalier de Lamarck. *Philosophie zoologique ou exposition des considérations relatives à l'hist. naturelle des animaux*. Paris, 1909.

[66] Naoki Mori, Junji Yoshida, Hisashi Tamori, Hajime Kita, and Yoshikatzu Nishikawa. A thermodynamical selection rule for the genetic algorithm. In *Proceedings of the 2nd IEEE International Conference on Evolutionary Computation, 1995 (ICEC'95)*, pages 188–192, Perth, Western Australia, November 29–December 1, 1995. IEEE, IEEE Service Center, Piscataway, NJ, 1995.

[67] Tapan Kumar Nayak. On diversity measures based on entropy functions. *Communications in Statistics: Theory and Methods*, 14(1): 203–215, 1985.

[68] Klaus Neumann and Martin Morlock. *Operations Research*. Carl Hanser Verlag, München, Wien, 1993.

[69] Allen E. Nix and Michael D. Vose. Modeling genetic algorithms with markov chains. *Annals of Mathematics and Artificial Intelligence*, 5(1): 79–88, 1992.

[70] N. N. *Der Handlungsreisende, wie er sein soll und was er zu thun hat, um Aufträge zu erhalten und eines glücklichen Erfolgs in seinen Geschäften gewiss zu sein. Von einem alten Commis-Voyageur*. Voigt, Ilmenau, 1832.

[71] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill Series in Electrical Engineering. McGraw-Hill Book Company, New York, NY, et al., third edition, 1991.

[72] G. P. Patil and C. Taillie. Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77(379): 548–561, September 1982.

[73] E[velyn] C. Pielou. Species-diversity and pattern-diversity in the study of ecological sucsession. *Journal of Theoretical Biology*, 10(2): 370–383, February 1966.

[74] E[velyn] C. Pielou. *Ecological Diversity*. A Wiley-Interscience Publication. John Wiley & Sons, New York, NY, et al., 1975.

[75] E[velyn] C. Pielou. *Mathematical Ecology*. A Wiley-Interscience Publication. John Wiley & Sons, New York, NY, et al., second, revised edition, 1977. First edition published as "An Introduction to Mathematical Ecology".

[76] Riccardo Poli. Why the schema theorem is correct also in the presence of stochastic effects. In *Proceedings of the 2000 Congress on Evolutionary Computation CEC00*, pages 487–492, La Jolla, CA, July 16–19, 2000. IEEE Neural Networks Council, IEEE Press.

[77] Lucien Preuss and Helmut Vorkauf. The knowledge content of statistical data. *Psychometrika. A Journal of Quantitative Psychology*, 62(1): 133–161, 1997.

[78] C. Rajski. A metric space of discrete probability distributions. *Information and Control*, 4(4): 371–377, December 1961.

[79] C. Radhakrishna Rao. Convexity properties of entropy functions and analysis of diversity. In Y. L. Tong, editor, *Inequalities in Statistics and Probability*, volume 5 of *Lecture Notes – Monograph Series*, pages 68–77, Lincoln, NE, October 27–30, 1982. Institute of Mathematical Statistics, Hayward, CA, 1984.

[80] C. Radhakrishna Rao. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21: 24–43, 1982.

[81] C. Radhakrishna Rao. Gini-simpson index of diversity: A characterisation, generalization and application. *Utilitas Mathematica*, 21B: 273–282, 1982.

[82] C. Radhakrishna Rao and Tapan N. Nayak. Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Transactions on Information Theory*, IT-31(5): 589–593, 1985.

[83] Colin R. Reeves and Jonathan E. Rowe. *Genetic Algorithms – Priciples and Perspectives*. Number 20 in Operations research/computer science interfaces series. Kluwer Academic Publishers, Boston, MA, 2003.

[84] Gerhard Reinelt. *The Traveling Salesman: Computational Solutions for TSP Applications*. Number 840 in Lecture Notes in Computer Science. Springer-Verlag, Berlin et al., 1994.

[85] Michael M. Richter and Ralph Bergmann. Knowledge management for e-commerce. slides of the lecture at the Universität Kaiserslautern, 2001.

[86] Alex Rogers and Adam Prügel-Bennett. Genetic drift in genetic algorithm selection schemes. *IEEE Transactions on Evolutionary Computation*, 3(4): 298–303, November 1999.

[87] Günter Rudolph. Convergence analysis of canonical genetic algorithms. *IEEE Transactions on Neural Networks*, 5(1): 96–101, January 1994.

[88] Günther Rudolph. Takeover times and probabilities of non-generational selection rules. In Darrell Whitley, David Goldberg, Erick Cantu-Paz, Lee Spector, Ian Parmee, and Hans-Georg Beyer, editors, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000)*, pages 903–910, Las Vegas, NV, July 10–12, 2000. Morgan Kaufmann Publishers, San Mateo, CA, 2000.

[89] Claude E[lwood] Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27: 379–423 and 623–656, July and October 1948.

[90] E. H. Simpson. Measurement of diversity. *Nature*, 163(4148): 688, April 30, 1949.

[91] Herbert Spencer. *SOCIAL STATICS: or, the Conditions Essential to Human Happiness Specified, and the First of them developed*. Chapman, London, UK, 1851.

[92] Christopher R. Stephens and Henri Waelbroeck. Schemata evolution and building blocks. *Evolutionary Computation*, 7(2): 109–124, 1999.

[93] V[olker] Storch, U[lrich] Welsch, and M[ichael] Wink. *Evolutionsbiologie*. Springer-Verlag, Berlin et al., 2001.

[94] Gilbert Syswerda. Uniform crossover in genetic algorithms. In J[ames] David Schaffer, editor, *Proceedings of the Third Internatinal Conference on Genetic Algorithms (ICGA 3)*, pages 2–9, Fairfax, VA, June 4–7, 1989. George Mason University, Morgan Kaufmann Publishers, San Mateo, CA, 1989.

[95] H. F. Trotter. Approximation of semi-groups of operators. *Pacific Journal of Mathematics*, 8: 887–919, 1958.

[96] Amos Tverski and Itamar Gati. Similarity, separability and the triangle inequality. *Psychological Review*, 89(2): 123–154, 1982.

[97] Michael D. Vose. *Simple Genetic Algorithm: Foundations and Theory*. The MIT Press, Cambridge, MA, 1998.

[98] Alfred Russel Wallace. On the tendency of varieties to depart indefinitely from the original type. *Journal of the Proceedings of the Linnean Society: Zoology*, 3(9): 53–62, August 20 1858.

[99] Satosi Watanabe. *Knowing and Guessing: A Quantitative Study of Inference and Information*. John Wiley & Sons, New York, NY, et al., 1969.

[100] J[ames] D. Watson and F[rances] H. C. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171: 737–738, 1953.

[101] Alden H. Wright and Jonathan E. Rowe. Continuous dynamical system models of steady-state genetic algorithms. In Worthy N. Martin and William M. Spears, editors, *Foundations of Genetic Algorithms 6 (FOGA-6)*, Charlottesville, VA, July 21–23, 2000. Morgan Kaufmann Publishers, San Francisco, CA, 2000.

[102] Alden T. Wright. The exact schema theorem. discussion paper, January 28, 2000.

[103] Sewall Wright. Evolution in mendelian populations. *Genetics*, 16: 97–159, March 1931.