Chapter 4

Random Strategies

One group of optimization methods has been completely ignored in Chapter 3: methods in which the parameters are varied according to probabilistic instead of deterministic rules; even the methods of stochastic approximation are deterministic. As indicated by the title there is not one random strategy but many, some of which differ considerably from each other.

It is common to resort to random decisions in optimization whenever deterministic rules do not have the desired success, or lead to a dead end; on the other hand random strategies are often supposed to be essentially more costly. The opinion is widely held that with careful thought leading to cleverly constructed deterministic rules, better results can always be achieved than with decisions that are in some way made randomly. The strategies that follow should show that randomness is not, however, the same as arbitrariness, but can also be made to obey very refined rules. Sometimes only this kind of method solves a problem effectively.

Profound considerations do not underlie all the procedures used in hill climbing strategies. The cyclic choice of coordinate directions in the Gauss-Seidel strategy could just as well be replaced by a random sequence. One can also consider increasing the number of directions used. Since there is no good reason for preferring to search for the optimum along directions parallel to the axes, one could also use, instead of only n different unit vectors, any number of randomly chosen direction vectors. In fact, suggestions along these lines have been made (Brooks, 1958) in order to avoid a premature termination of the minimum search in narrow oblique valleys (compare Chap. 3, Sect. 3.2.1.1). Similar concepts have been developed for example by O'Hagan and Moler (after Wilde and Beightler, 1967), Emery and O'Hagan (1966), Lawrence and Steiglitz (1972), and Beltrami and Indusi (1972), to improve the pattern search of Hooke and Jeeves (1961, see Chap. 3, Sect. 3.2.1.2). The limitation to a finite number of search directions is not only a disadvantage in narrow oblique valleys but also at the border of the feasible region as determined by inequality constraints. All the deterministic remedies against prematurely ending the iteration sequence assume that more information can be gathered, for example in the form of partial derivatives of the constraint functions (see Klingman and Himmelblau, 1964; Glass and Cooper, 1965; Paviani and Himmelblau, 1969). Providing this information usually means a high extra cost and is sometimes not possible at all.

Random directions that are not oriented with respect to the structure of the objective function and the allowed region also imply a higher cost because they do not take optimal single steps. They can, however, be applied in every case.

Many deterministic optimization methods, especially those which are guided by the gradient of the objective function, have convergence difficulties at points where the partial derivatives are discontinuous. On the contour diagram of a two parameter objective function, of which the maximum is sought, such positions correspond to sharp ridges leading to the summit (e.g., Zwart, 1970). A narrow valley–the geometric picture in the case of minimization–leads to the same problem if the finite step lengths are greater than its width. Then all attempts fail to make improvements in the coordinate directions or, from trial steps in these directions, fail to predict a locally best direction in which to continue (gradient direction). The same phenomenon can also occur when the partial derivatives are specified analytically, because of the rounding errors involved in computing with a finite number of significant figures. To avoid premature termination of a search in such cases, Norkin (1961) has suggested the following procedure. When the optimization according to the conventional scheme has ended, a step is taken away from the supposed optimum in an arbitrary coordinate direction. The extremum is sought again, excluding this one variable, and the search is only finally ended when deviations in all directions have led back to the same point. This rule should also prevent stagnation at saddle points.

Even the simplex method of linear programming makes random decisions if the search for the extremum threatens to be endless because the problem is *degenerate*. Then following Dantzig's suggestion (1966) the iteration scheme should be interrupted in favor of a random exchange step. A problem is only degenerate, however, because the general rules do not cover the special case (see also Chap. 6, Sect. 6.2). A further example of resorting to chance when a dead end has been reached is Brent's modification of the strategy with conjugate directions (Brent, 1973). Powell's algorithm (Powell, 1964) when applied to problems in many dimensions tends to generate linearly dependent directions and then to proceed within a subspace of $\mathbb{R}^n$. For this reason Brent now and then interrupts the line searches with steps in randomly chosen directions (see also Chap. 3, Sect. 3.2.2.1).

One very frequently comes across proposals to let chance take control when the problem is to find global minima of multimodal objective functions. Such problems frequently crop up in process design (Motskus, 1967; Mockus, 1971) but can also be the result of recasting discrete problems into continuous form (Katkovnik and Shimelevich, 1972). Practically all sequential search procedures can only lead to a local optimum–as a rule, the one nearest to the starting point. There are a few proposals for ensuring global convergence of sequential optimization methods (e.g., Motskus and Feldbaum, 1963; Chichinadze, 1967, 1969; Goldstein and Price, 1971; Ueing, 1971, 1972; Branin and Hoo, 1972; McCormick, 1972; Sutti, Trabattoni, and Brughiera, 1972; Treccani, Trabattoni, and Szegö, 1972; Brent, 1973; Hesse, 1973; Opačić, 1973; Ritter and Tui as mentioned by Zwart, 1973). They are often in the form of additional, heuristic rules. Gran (1973), for example, considers gradient methods that are supposed to achieve global convergence by the addition of a random process to the deterministic changes. Hill (1964; see also Hill and Gibson, 1965) suggests subdividing the interval to be explored and gathering sufficient information in each section to carry out a cubic interpolation. The best of the results for the parts is taken as an approximation to the global optimum. However, for n-dimensional interpolations the cost increases rapidly with n; this scheme thus looks impractical for more than two variables. To work with several, randomly chosen starting points and to compare each of the local minima (or maxima) obtained is usually regarded as the only course of action for determining the global optimum with at least a certain probability (so-called *multistart* techniques). Proposals along these lines have been made by, among others, Gelfand and Tsetlin (1961), Bromberg (1962), Bocharov and Feldbaum (1962), Zellnik, Sondak, and Davis (1962), Krasovskii (1962), Gurin and Lobac (1963), Flood and Leon (1964, 1966), Kwakernaak (1965), Casey and Rustay (1966), Weisman and Wood (1966), Pugh (1966), McGhee (1967), Crippen and Scheraga (1971), and Brent (1973).

A further problem faces deterministic strategies if the calculated or measured values of the objective function are subject to stochastic perturbations. In the experimental field, for example in the on-line optimum search, or for control of the optimal conditions in processes, perturbations must be taken into account from the start (e.g., Tovstucha, 1960; Feldbaum, 1960, 1962; Krasovskii, 1963; Medvedev, 1963, 1968; Kwakernaak, 1966; Zypkin, 1967). However, in computational optimization too, where the objective function is analytically specified, a similar effect arises because of rounding errors (Brent, 1973), especially if one uses hybrid analogue computers for solving functional optimization problems (e.g., Gilbert, 1967; Korn and Korn, 1964; Bekey and Karplus, 1971). A simple, if expensive (in the sense of cost in computations or trials) method of dealing with this is the repetition of measurements until a definite conclusion is possible. This is the procedure adopted by Box and Wilson (1951) in the experimental gradient method, and by Box (1957) in his EVOP strategy. Instead of a fixed number of repetitions, which while on the safe side may be unnecessarily high, one can follow the concept of sequential analysis of statistical data (Wald, 1966; see also Zigangirov, 1965; Schumer, 1969; Kivelidi and Khurgin, 1970; Langguth, 1972), which is to make only as many trials as the trial results seem to make absolutely necessary. More detailed investigations on this subject have been made, for example, by Mlynski (1964a,b, 1966a,b).

As opposed to attempting to improve the decisive data, Brooks and Mickey (1961) have found that one should work with the minimum number of $n + 1$ comparison points in order to determine a gradient direction, even if this is a perturbed one. One must however depart from the requirement that each step should yield a success, or even the locally greatest success. The motto that following locally the best possible route seldom leads to the best overall result is true not only for first order gradient strategies but also for Newton and quasi-Newton methods . Harkins (1964), for example, maintains that inexact line searches not only do not worsen the convergence of a minimization procedure but in some cases actually improve it. Similar experiences led Davies, Swann, and Campey in their strategy (see Chap. 3, Sect. 3.2.1.4) to make only one quadratic interpolation in each direction. Also Spendley, Hext, and Himsworth (1962), in the formulation of their simplex method, which generates only near-optimal directions, work on the assumption that random decisions are not necessarily a total disadvantage (see also Himsworth, 1962). Based on similar arguments, the modification of this strategy by M. J. Box (1965) sets up the initial simplex or complex by means of random numbers. Imamura et al. (1970) even go so far as to superimpose artificial stochastic variations on an objective function in order to prevent convergence to inferior local optima.

The rigidity of an algorithm based on a fixed internal model of the objective function, with which the information gathered during the iterations is interpreted, is advantageous if the objective function corresponds closely enough to the model. If this is not the case, the advantage disappears and may even turn into a disadvantage. Second order methods with quadratic models seem more sensitive in this respect than first order methods with only linear models. Even more robust are the direct search strategies that work without an explicit model, such as the strategy of Hooke and Jeeves (1961). It makes no use of the sizes of the changes in the objective function values, but only of their signs.

A method that uses a kind of minimal model of the objective function is the stochastic approximation (Schmetterer, 1961; see also Chap. 2, Sect. 2.3). This purely deterministic method assumes that the measured or calculated function values are samples of a normally distributed random quantity, of which the expectation value is to be minimized or maximized. The method feels its way to the optimum with alternating exploratory and work steps, whose lengths form convergent series with prescribed bounds and sums. In the multidimensional case this standard concept can be the basis of various strategies for choosing the directions of the work steps (Fabian, 1968). Usually gradient methods show themselves to best advantage here. The stochastic approximation itself is very versatile. Constraints can be taken into account (Kaplinskii and Propoi, 1970), and problems of functional optimization can be treated (Gersht and Kaplinskii, 1971) as well as dynamic problems of maintaining or seeking optima (Chang, 1968). Tsypkin (1968a,b,c, 1970a,b; see also Zypkin, 1966, 1967, 1970) discusses these topics very thoroughly. There are also, however, arguments against the reliability of convergence for certain types of objective function (Aizerman, Braverman and Rozonoer, 1965). The usefulness of the strategy in the multidimensional case is limited by its high cost. Hence there has been no shortage of attempts to accelerate the convergence (Fabian, 1967; Berlin, 1969; Saridis, 1968, 1970; Saridis and Gilbert, 1970; Janáč, 1971; Kwatny, 1972; see also Chap. 2, Sect. 2.3). Ideas for using random directions look especially promising; some of the many investigations of this topic which have been published are Loginov (1966), Stratonovich (1968, 1970), Schmitt (1969), Ermoliev (1970), Svechinskii (1971), Tsypkin (1971), Antonov and Katkovnik (1972), Berlin (1972), Katkovnik and Kulchitskii (1972), Kulchitskii (1972), Poznyak (1972), and Tsypkin and Poznyak (1972).

The original method is not able to determine global extrema reliably. Extensions of the strategy in this direction are due to Kushner (1963, 1972) and Vaysbord and Yudin (1968). The sequence of work steps is so designed that the probability of the following state being the global optimum is maximized. In contrast to the gradient concept, the information gathered is not interpreted in terms of local but of global properties of the objective function. In the case of two local minima, the effort of the search is gradually concentrated in their neighborhood and only when one of them is significantly better is the other abandoned in favor of the one that is also a global minimum. In terms of the cost of the strategy, the acceleration of the local search and the reliability of the global search are diametrically opposed. Hill and Gibson (1965) show that their global strategy is superior to Kushner's, as well as to one of Bocharov and Feldbaum. However, they only treat cases with $n \leq 2$ parameters. More recent research results have been presented by

Pardalos and Rosen (1987), Törn and Žilinskas (1989), Floudas and Pardalos (1990), Zhigljavsky (1991), and Rudolph (1991, 1992b). Now there are even specialized journals established in the field, see Horst (1991).

All the strategies mentioned so far are fundamentally deterministic. They only resort to chance in dead-end situations, or they operate on the assumption that the objective function is stochastically perturbed. Jarvis (1968), who compares deterministic and probabilistic optimization methods, finds that random methods that do not stick to any particular model are most suitable when an optimum must be located under particularly difficult conditions, such as a perturbed objective function or a "pathological" problem structure with several extrema, discontinuities, plateaus, forbidden regions, etc. The homeostat of Ashby (1960) is probably the oldest example of the application of a random strategy. Its objective is to maintain a condition of equilibrium against stochastic disturbances. It may happen that no optimum is sought, but only a point in an allowed region (today one calls such task a *constraints satisfaction problem* or CSP). Nevertheless, corresponding solution methods are closely tied to optimization, and there are a series of various *heuristic planning methods* available (e.g., Weinberg and Zehnder, 1969). Ashby's strategy, which he calls a blind homeostatic process, becomes active whenever the apparatus strays from equilibrium. Then the controllable parameters are randomly varied until the desired condition is restored. The finite number (in this case) of discrete settings of the variables all enter the search process with equal probability. Chichinadze (1960) later constructed an electronic model on the same principle and used it for synthesizing simple optimal control systems.

Brooks (1958), probably stimulated by R. L. Anderson (1953), is generally regarded as the initiator of the use of random strategies for optimization problems. He describes the simple, later also called *blind* or *pure random search* for finding a minimum or maximum in the experimental field. In a closed interval $a \leq x \leq b$ several points are chosen at random. The probability density $w(x)$ is constant everywhere within the region and zero outside.

$$w(x) = \begin{cases} 1/V\,, & \text{for all } a \leq x \leq b \\ 0\,, & \text{otherwise} \end{cases}$$

V, the volume of the cube with corners a_i and b_i for $i = 1(1)n$, is given by

$$V = \prod_{i=1}^{n} (b_i - a_i)$$

The value of the objective function must be determined at all selected points. The point that has the lowest or highest function value is taken as optimum. How well the true extremum is approximated depends on the number of trials as well as on the actual random results. Thus one can only give a probability p that the optimum will be found within a given number N of trials with a prescribed accuracy.

$$p = 1 - (1 - v/V)^N \qquad (4.1)$$

The volume $v < V < \infty$ contains all points that satisfy the accuracy requirement. By rearranging Equation (4.1), the number of trials is obtained

$$N = \frac{\ln(1-p)}{\ln(1-\frac{v}{V})} \tag{4.2}$$

that is required in order to place with probability p at least one trial in the volume v. Brooks concludes from this that the cost is independent of the number of variables. In their criticism Hooke and Jeeves (1958) point out that it is not feasible to consider the accuracy in terms of the volume ratio for problems with many variables. For $n = 100$ parameters, a volume ratio of $\frac{v}{V} = 0.1$ corresponds to a length ratio of the side length D of V and d of v of

$$\frac{d}{D} = \sqrt[n]{\left(\frac{v}{V}\right)} \simeq 0.98$$

This means that the uncertainty in the variables x_i is 98% of the original interval $[a_i, b_i]$, although the volume containing the optimum has been reduced to one tenth of the original. Shimizu (1969) makes the same mistake as Brooks and attempts to implement the strategy for problems with more general constraints.

A comparison of the pure random search and deterministic search methods known at the time for experimental optimization problems (Brooks, 1959) also shows no advantage of the stochastic strategy. The test only covers four different objective functions, each with two variables. Brooks then recommends applying his random method if the number of parameters is large or if the determination of objective function values is subject to large perturbations. McArthur (1961) concludes on the basis of numerical experiments that *the random strategy* is also preferable for complicated problem structures. Just this circumstance has led to the use, even today, of the pure random search, often called the *Monte-Carlo method*, for example in computer optimization of building construction (Goliński and Leśniak, 1966; Leśniak, 1970; Hupfer, 1970).

In principle, all the trials of the simple random strategy can be made simultaneously. It is thus numbered among the simultaneous optimization methods. The decision to choose a particular state vector of variables does not depend on the results of preceding trials, since the probability of scoring according to the uniform distribution is the same at all times. However, in applications on the traditional, serially operating computers, the trials must be made sequentially. This can be used to advantage by storing the current best value of the objective function and its associated variable value. In Chapter 3, Section 3.1.1 and 3.2 the *grid or tabulation method* was referred to as optimal in the minimax sense. The blind random strategy should thus not be any better. Defining the interval length $D_i = b_i - a_i$ for the variable x_i, with required accuracy d_i, and assuming that all the $D_i = D$ and $d_i = d$ for $i = 1(1)n$, then for the volume ratio in Equations (4.1) and (4.2)

$$\frac{v}{V} = \left(\frac{d}{D}\right)^n$$

If $\frac{v}{V}$ is small, which when there are many variables must be the case, one can use the approximation

$$\ln(1+y) \simeq y \quad \text{for } y \ll 1$$

to write the number of required trials as

$$N \simeq -\ln(1-p)\left(\frac{D}{d}\right)^n$$

Assuming that $\dfrac{D}{d}$ is an integer, the grid method requires

$$N = \left(\frac{D}{d}\right)^n$$

trials (compare Chap. 3, Sect. 3.2, Equation (3.19)). The value is the same for both procedures if $p \simeq 0.63$. Supposing that the probability of at least one score of the required accuracy is $p = 0.90$, then the random strategy results in

$$N \simeq 2.3\left(\frac{D}{d}\right)^n$$

which is clearly worse than the grid strategy (Spang, 1962). The reason for the extra cost, however, should not be attributed to the randomness of decisions itself, but to the fact that for an equiprobable, continuous selection of variables, the trials can be very close together or, in the discrete case, they can repeat themselves. If one can avoid that, the disadvantage would no longer exist. A randomized sequence of trials even might hit upon the optimal result earlier than an ordered one. Nevertheless Spang's proof has for some time brought all random methods, not only the simple Monte-Carlo strategy, into disrepute.

Nowadays the term *Monte-Carlo methods* is understood to cover, in general, simulation methods that have to do with stochastic events. They are applied effectively to solving difficult differential equations (Little, 1966) or for evaluating integrals (Cowdrey and Reeves, 1963; McGhee and Walford, 1968). Besides the simple *hit-or-miss scheme*, however, greatly improved variants have been developed (e.g., W. F. Bauer, 1958; Hammersley and Handscomb, 1964; Korn, 1966, 1968; Hull, 1967; Brandl, 1969). Amann (1968a,b) reports a Monte-Carlo method with information storage and a sequential extension for the solution of a linear boundary value problem, and Curtiss (1956) describes a Monte-Carlo procedure for solving systems of linear equations. Both are supposed to be less costly than comparable deterministic strategies. Pinkham (1964) and Pincus (1970) describe modifications for the problems of finding zeros of a non-linear function and of constrained optimization. Since only relatively few publications treat random optimization methods in any depth (Karnopp, 1961, 1963; Idelsohn, 1964; Dickinson, 1964; Rastrigin, 1963, 1965a,b, 1966, 1967, 1968, 1969, 1972; Lavi and Vogl, 1966; Schumer, 1967; Jarvis, 1968; Heydt, 1970; Cockrell, 1970; White, 1970, 1971; Aoki, 1971; Kregting and White, 1971), the improved strategies will be briefly presented here. They all operate with sequential and sometimes both simultaneous and sequential random trials and in one way or another exploit the information from preceding trials to accelerate the convergence.

Brooks himself already suggests several improvements. Thus to exclude repetitions or closely situated trials, the volume to be investigated can be subdivided into, for example, cubic subspaces, into each of which only one random trial is placed. According to one's knowledge of the approximate position of the optimum, the subspaces will be assigned different sizes (Idelsohn, 1964). The original uniform distribution is thereby replaced by one with a greater density in the neighborhood of the expected optimum. Karnopp (1961, 1963, 1966) has treated this problem in detail without, however, giving any practical procedure. Mathematically based investigations of the same topic are due to Motskus (1965), Hupfer (1970), Pluznikov, Andreyev, and Klimenko (1971), Yudin (1965, 1966, 1972), Vaysbord (1967, 1968, 1969), Taran (1968a,b), Karumidze (1969), and Meerkov (1972). If after several (simultaneous) samples the search is continued in an especially promising looking subregion, the procedure becomes sequential in character. Suggestions of this kind have been made for example by McArthur (1961), Motskus (1965), and Hupfer (1970) (*shrinkage random search*). Zakharov (1969, 1970) applies the stochastic approximation for the successive shrinkage of the region in which Monte-Carlo samples are placed. The most thoroughly worked out strategy is that of McMurtry and Fu (1966, *probabilistic automaton*; see also McMurtry, 1965). The problem considered is to adjust the variable parameters of a control system for a dynamic process in such a way that the optimum of the system is found and maintained despite perturbations and (slow) drift (Hill, McMurtry, and Fu, 1964; Hill and Fu, 1965). Initially the probabilities are equal for all subregions, at the center of which the function values are measured (assumed to be stochastically perturbed). In the course of the iterations the probability matrix is altered so that regions with better objective function values are tested more often than others. The search ends when only one subregion remains: the one with the highest probability of containing the global optimum. McMurtry and Fu use a so-called linear intensification to adjust the probability matrix. Suggestions for further improving the convergence rate have been made by Nikolić and Fu (1966), Fu and Nikolić (1966), Shapiro and Narendra (1969), Asai and Kitajima (1972), Viswanathan and Narendra (1972), and Witten (1972). Strongin (1970, 1971) treats the same problem from the point of view of decision theory.

All these methods lay great emphasis on the reliability of global convergence. The quality of the approximation depends to a large extent on the number of subdivisions of the n-dimensional region under investigation. High accuracy requirements cannot be met for many variables since, at least initially, the number of subregions to investigate rises exponentially with the number of parameters. To improve the local convergence properties, there are suggestions for replacing the midpoint tests in a subvolume by the result of an extreme value search. This could be done with one of the familiar search strategies such as a gradient method (Hill, 1969) or any other purely sequential random search method (Jarvis 1968, 1970) with a high convergence rate, even if it were only guaranteed to converge locally. Application, however, is limited to problems with at most seven or eight variables, as reported.

Another possibility for giving a sequential character to random methods consists of gradually shifting the expectation value of a random variable with a restricted probability density distribution. Brooks (1958) calls his proposal of this type the *creeping random search.* Suitable random numbers are provided for example by a Gaussian distribution with expectation value ξ and standard deviation σ. Starting from a chosen initial condition $x^{(0)}$, several simultaneous trials are made, which most likely fall in the neighborhood of the starting point ($\xi = x^{(0)}$). The coordinates of the point with the best function value form the expectation value for the next set of random trials. In contrast to other procedures, the data from the other trials are not exploited to construct a linear or even quadratic model from which to calculate a best possible step (e.g., Brooks and Mickey, 1961; Aleksandrov, Sysoyev, and Shemeneva, 1968; Pugachev, 1970). For small σ and a large number of samples, the best value will in any case fall in the locally most favorable direction. In order to approach a solution with high accuracy, the variance σ^2 must be successively reduced. Brooks, however, gives no practical rule for this adjustment. Many algorithms have since been published that are extensions of Brooks' basic concept of the creeping random search. Most of them no longer choose the best of several trials; they accept each improvement and reject each worsening (Favreau and Franks, 1958; Munson and Rubin, 1959; Wheeling, 1960).

The iteration rule of a creeping random search is, for the minimum search:

$$x^{(k+1)} = \begin{cases} x^{(k)} + z^{(k)}, & \text{if } F(x^{(k)} + z^{(k)}) \leq F(x^{(k)}) \quad \text{(success)} \\ x^{(k)}, & \text{otherwise} \quad \text{(failure)} \end{cases}$$

The random vector $z^{(k)}$, which in this notation effects the change in the state vector x, belongs to an n-dimensional $(0, \sigma^2)$ normal distribution with the expectation value $\xi = 0$ and the variance σ^2, which in the simplest case is the same for all components. One can thus regard σ, or better $\sigma\sqrt{n}$, as a kind of average step length. The direction of $z^{(k)}$ is uniformly distributed in $\mathbb{R}^n$, i.e., purely random. Gaussian distributions for the increments are also used by Bekey et al. (1966), Stewart, Kavanaugh, and Brocker (1967), and De Graag (1970). Gonzalez (1970) and White (1970) use instead of a normal distribution a uniform distribution that covers a small region in the form of an n-dimensional cube centered on the starting point. This clearly favors the diagonal directions, in which the total step lengths are on average a factor $\sqrt{n}$ greater than in the coordinate directions. Pierre (1969) therefore restricts the uniformly distributed random probe to an n-dimensional hypersphere of fixed radius. Rastrigin (1960–1972) gives the total step length

$$s = \sqrt{\sum_{i=1}^{n} z_i^2}$$

a fixed value. Instead of the normal distribution he thus obtains a circumferential or hypersphere-surface distribution. In addition, he repeats the evaluation of the objective function when there is a failure in order to reduce the effect of stochastic perturbations. Taking two model functions

$$\begin{aligned} F_1(x) &= F_1(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i && \text{(inclined plane)} \\ F_2(x) &= F_2(x_1, \ldots, x_n) = \sqrt{\sum_{i=1}^{n} x_i^2} && \text{(hypercone)} \end{aligned}$$

he investigates the average convergence rate of his strategy and compares it with that of an experimental gradient method, in which the partial derivatives are approximated by quotients of differences obtained from exploratory steps . He shows that for a linear problem structure like F_1 the random strategy needs only $O(\sqrt{n})$ trials, whereas the gradient strategy needs $O(n)$ trials to cover a prescribed distance. For $n > 3$, the random strategy is always superior to the deterministic method. Whereas Rastrigin shows that the random search always does better than the gradient search in the spherically symmetric field F_2, Movshovich (1966) maintains the opposite. The discrepancy can be traced to differing assumptions about the choice of step length (see also Yvon, 1972; Gaviano and Fagiuoli, 1972).

To choose suitable step lengths or variances poses the same problems for sequential random searches as are familiar from deterministic strategies. Here too, a closely related problem is to achieve global convergence with reference to a suitable termination rule, the convergence criterion, and with a degree of reliability. Khovanov (1967) has conceived an individual manner of controlling the random step lengths. He accepts every random change, irrespective of success or failure, increases the variance at each failure and reduces it otherwise. The objective is to increase the probability of lingering in the more promising regions and to abandon states that are irrelevant to the optimum search. No applications of the strategy are known to the author. Favreau and Franks (1958), Bekey et al. (1966), and Adams and Lew (1966) use a constant ratio between σ_i and x_i for $i = 1(1)n$. This measure does have the effect of continuously altering the "step lengths," but its merit is not obvious. Just because a variable value x_i is small in no way indicates that it is near to the extreme position being sought. Karnopp (1961) was the first to propose a step length rule based on the number of successes or failures, according to which the σ_i or s are all uniformly reduced or enlarged such that a success always occurs after two or three trials. Schumer (1967), and Schumer and Steiglitz (1968), submit Rastrigin's circumferential random direction method to a thorough examination by probability theory. For the model

$$F_3(x) = \sum_{i=1}^{n} x_i^2 = r^2$$

with the condition $n \gg 1$ and the continuously optimal step length

$$s \simeq 1.225 \, \frac{r}{\sqrt{n}}$$

they obtain a rate of progress φ, which is the average distance covered in the direction of the objective (minimum) per random step:

$$\varphi \simeq 0.203 \, \frac{r}{n}$$

and a success rate w_s which is the average number of successes per trial:

$$w_s \simeq 0.270$$

They are only able to treat the general quadratic case theoretically for $n = 2$. Their result can be interpreted in the sense that φ is dependent on the smallest radius of curvature ρ of the elliptic contour passing through r. Since neither r nor s can be assumed to be known in advance, it is not clear how to keep to the optimal step length. Schumer and Steiglitz (1968) give an adaptive method with which the correct size of s can be maintained at least approximately during the course of the iterations. At the starting point $x^{(0)}$ two random changes are made with step lengths $s^{(0)}$ and $s^{(0)}(1 + a)$, where $0 < a \leq 1$. If both samples are successful, for the next iteration $s^{(1)} = s^{(0)}(1 + a)$ is taken, i.e., the greater value. If only one sample yields an improvement in the objective function, its step length is taken; finally if no success is scored, $s^{(1)}$ remains equal to $s^{(0)}$.

A reduction in s is only made if several consecutive trials are unsuccessful. This is also the procedure of Maybach (1966). This adjustment to the local conditions assists the strategy in achieving high convergence rates but reduces the chances of locating global optima among several local ones. For this reason a sample with a significantly larger step length ($a > 1$) should be included from time to time. Numerical tests show that the computation cost, or number of trials, actually only increases linearly with the number of variables. Schumer and Steiglitz have tested this using the model functions F_3 and

$$F_4(x) = \sum_{i=1}^{n} x_i^4$$

A comparison with a Newton-Raphson strategy, in which the partial first and second derivatives are determined numerically and the cost increases as $O(n^2)$, favors the random method when $n > 78$ for F_3 and when $n > 2$ for F_4. For the second, biquadratic model function, Nelder and Mead (1965) state that the number of trials or function evaluations in their simplex strategy grows as $O(n^{2.11})$, so that the sequential random method is superior from $n > 10$. White and Day (1971) report numerical tests in which the cost in iterations with Schumer's strategy increases more sharply than linearly with n, whereas a modification by White (1970) shows exact linear dependence. A comparison with the strategy of Fletcher and Powell (1963) favors the latter, especially for truly quadratic functions.

Rechenberg (1973), with an n-dimensional normal distribution (see Chap. 5, Sect. 5.1), reaches almost the same theoretical results as Schumer for the circumferential distribution, if one notes that the overall step length

$$\sigma_{tot} = \sqrt{\sum_{i=1}^{n} \sigma_i^2} = \sigma\sqrt{n}$$

for equal variances $\sigma_i^2 = \sigma^2$ in each random component z_i is proportional to the square root of the number of variables. The reason for this lies in the property of Euclidean space that, as the number of dimensions increases, the volume of a hypersphere becomes concentrated more and more in the boundary region near the surface. Rechenberg's adaptation rule is founded on the relation between optimal variance and probability of success derived from two essentially different models of the objective function. The adaptation rule which is thereby formulated makes the frequency and size of the σ increments respectively dependent on the number of variables and independent of the structure of the objective function. This will be discussed in more detail in Chapter 5, Section 5.1.

Convergence proofs for the sequential random strategy have been given by Matyas (1965, 1967) and Rechenberg (1973) only for the case of constant variance σ^2. Gurin (1966) has proved convergence also for stochastically perturbed objective functions. The convergence rate is still reduced by perturbations (Gurin and Rastrigin, 1965), but not as much as in gradient methods. Global convergence can be achieved if the reference value of the objective function is measured more than once at the comparison point (Saridis and Gilbert, 1970). As soon as any attempt is made to achieve higher rates of convergence by adjusting the variances or step lengths, the chance of finding a global optimum diminishes. Then the random strategy itself becomes a *path-oriented* instead of a *volume-oriented* strategy. The probability of global convergence still always remains finite; it may simply become very small, especially in the case of many dimensions.

Apart from adjusting the step lengths, one can consider modifying the directions. Several proposals of this kind have been published: Satterthwaite (1959a; following McArthur, 1961), Wheeling (1960), Smith and Rudd (1964; following Dickinson, 1964), Matyas (1965, 1967), Bekey et al. (1966), Stewart, Kavanaugh, and Brocker (1967), De Graag (1970), and Lawrence and Emad (1973). They are all heuristic in nature. In the simplest case of a *directed random search*, a successful random direction is maintained until a failure occurs (Satterthwaite). Bekey, Lawrence, and Rastrigin actually make use of each random direction. If the first step leads to a failure, they use the opposite direction (positive and negative absolute biasing). Smith and Rudd store the two currently best points from a larger series of samples and obtain from their separation a step length for continuing the optimization. Wheeling's *history vector method* adds to each random increment a deterministic portion, derived from experience. This additional vector is initially zero. It is increased at each success by a fraction of the increment vector, and correspondingly decreased at each failure. Such a *learning and forgetting* process also forms the basis of the algorithms of De Graag and Matyas. The latter has received the most attention, in spite of the fact that it gives no precise guidance on how to choose the variances. Schrack and Borowski (1972), who apply their own step length rule in Matyas' strategy, were able to show by numerical tests that the simple algorithm of Schumer and Steiglitz, without direction orientation, is at least as good as Matyas' for unperturbed as well as perturbed measurements of the objective function. A quite different kind of method, due to Kjellström (1965), in which the random search takes place in varying three dimensional subspaces of the space $\mathbb{R}^n$, shows itself here to be very much worse.

Another method that sets out to accept only especially favorable directions is the *threshold strategy* of Stewart, Kavanaugh and Brocker (1967), in which only those random changes are accepted that result in a specified minimum improvement in the objective function value. A more recent version of the same idea has been given by Dueck and Scheuer (1990). The simultaneous adjustment of step lengths and directions has seldom been attempted. The suggestions of Favreau and Franks (1958) and Matyas (1965, 1967) remain too imprecise to be practicable. Gaidukov (1966; see also Hupfer, 1970) and Fürst, Müller, and Nollau (1968) provide more exact information for this purpose, based on either the concepts of Rastrigin or Matyas. Modification of the expectation values and variances of the random vectors is made according to the success or failure of iterations. No applications of the strategy are known, however, so that for the time being the observation of Schrack and Borowski (1972) still stands, namely that a careful choice of the step lengths is the most important prerequisite for the rapid convergence of a random method.

A method devised by Rastrigin (1965a,b, 1968) and developed further by Heydt (1970) works entirely with a restricted choice of directions. With a fixed step length, a direction can be randomly selected only from within an n-dimensional hypercone. The angle subtended by the cone and its height (and thus the overall step length) are controlled in an adaptive way. For a spherical objective function, e.g., the model functions F_2 (hypercone), F_3 (hypersphere), or F_4 (something intermediate between hypersphere and hypercube), there is no improvement in the convergence behavior. Advantages can only be gained if the search has to follow a particular direction for a long time along a narrow valley. Sudden changes in direction present a problem, however, which leads Heydt to consider substituting for the cone configuration a hyper-parabolic or hyper-hyperbolic distribution, with which at least small step lengths would retain sufficient freedom of direction.

In every case the striving for rapid convergence is directly opposed to the reliability of global convergence. This has led Jarvis (1968, 1970) to investigate a combination of the method of Matyas (1965, 1967) with that of McMurtry and Fu (1966). Numerical tests by Cockrell (1969, 1970; see also Fu and Cockrell, 1970) show that even here the basic strategy of Matyas (1965) or Schumer and Steiglitz (1967) is clearly the better alternative. It offers high convergence rates besides a fair chance of locating global optima, at least for a small number of variables. In the case of many dimensions, every attempt to reach global reliability is thwarted by the excessive cost. This leaves the globally convergent stochastic approximation method of Vaysbord and Yudin (1968) far behind the rest of the field. Furthermore, the sequential or creeping random search is the least susceptible if perturbations act on the objective function.

Users of random strategies always draw attention to their simplicity, flexibility and resistance to perturbations. These properties are especially important if one wishes to construct automatic optimalizers (e.g., Feldbaum, 1958; Herschel, 1961; Medvedev and Ruban, 1967; Krasnushkin, 1970). Rastrigin actually built the first optimalizer with a random search strategy, which was designed for automatic frequency control of an electric motor. Mitchell (1964) describes an extreme value controller that consists of an analogue computer with a permanently wired-in digital part. The digital part serves for storage and flow control, while the analogue part evaluates the objective function. The development of hybrid analogue computers, in which the computational inaccuracy is determined by the system, has helped to bring random methods, especially of the sequential type, into more general use. For examples of applications besides those of the authors mentioned above, the following publications can be referred to: Meissinger (1964), Meissinger and Bekey (1966), Kavanaugh, Stewart, and Brocker (1968), Korn and Kosako (1970), Johannsen (1970, 1973), and Chatterji and Chatterjee (1971). Hybrid computers can be applied to best advantage for problems of optimal control and parameter identification, because they are able to carry out integrations and differentiations more rapidly than digital computers. Mutseniyeks and Rastrigin (1964) have devised a special algorithm for the dynamic control problem of keeping an optimum. Instead of the variable position vector x, a velocity vector with components $\partial x_i / \partial t$ is varied. A randomly chosen combination is retained as long as the objective function is decreasing in value (for minimization $\partial F / \partial t < 0$). As soon as it begins to increase again, a new velocity vector is chosen at random.

It is always striking, if one observes living beings, how well adapted they are in shape, function, and lifestyle . In many cases, biological structures, processes, and systems even surpass the capabilities of highly developed technical systems. Recognition of this has for years led many authors to suspect that nature is in possession of optimal solutions to her problems. In some cases the optimality of biological subsystems can even be demonstrated mathematically, for example for the ratios of diameters in branching arteries (Cohn, 1954), for the hematocrit value (the volume fraction of solid particles in the blood; Lew, 1972), and the position of branch points in a level system of blood vessels (Kamiya and Togawa, 1972; see also Grassmann, 1967, 1968; Rosen, 1967; Rein and Schneider, 1971).

According to the theory of the descent of the species, all organisms that exist today are the (intermediate) result of a long process of development: evolution. Based on the multitude of finds of transitional species that have since become extinct, paleontology is providing a gradually more complete picture of this development. Leaving aside supernatural explanations, one must assume that the development of optimal or at least very good structures is a property of evolution, i.e., evolution is, or possesses, an optimization (or better, meliorization) strategy.

In evolution, the mechanism of variation is the occurrence of random exchanges, even "errors," in the transfer of genetic information from one generation to the next. The selection criterion favors the better suited individuals in the so-called *struggle for existence.* The similarity of variation and selection to the iteration rules of direct optimization methods is, in fact, striking. This analogy is most often drawn for random strategies, since mutations can best be interpreted as random changes. Thus Ashby (1960) regards as mutations the stochastic parameter variations in his blind homeostatic process. For many variables, however, the pure or blind random search requires so many trials that it offers no acceptable explanation of the capabilities of natural structures, processes, and systems. With the highest possible physical rate of transfer of information, as given by Bremermann (1962; see also Ashby, 1965, 1968) of 10^{47} bits per second and gram of computer mass, the mass of the earth and the extent of its lifetime up to now would not be sufficient to solve even simple combinatorial problems by complete enumeration or a blind random search, never mind to determine the optimal configuration of the 10^4 to 10^5 genes with their information content of around 10^{10} bits (Bremermann, 1963). Evolution must rather be considered as a sequential process that exploits the information from preceding successes and failures in order to follow a trajectory, although not a completely deterministic one, in the n-dimensional parameter space. Brooks (1958) and Favreau and Franks (1958) are therefore right to compare their creeping random search with biological evolution. Yet it is also certainly a very much simplified imitation of the natural process of development. In the 1960s, two proposals that consciously think of higher evolution principles as optimization rules to be simulated are due to Rechenberg (1964, 1973) and Bremermann (1962, 1963, 1967, 1968a,b,c, 1970, 1971, 1973a,b; see also Bremermann, Rogson, and Salaff, 1965, 1966; Bremermann and Lam, 1970). Bremermann reasons from the (nowadays!) low mutation rates observed in nature that only one component of the variable vector should be varied at a time. He then encounters with this scheme the same difficulties as arise in the coordinate method. On the basis of his failure with the mutation-selection scheme, for example on linear programming problems, he comes to the conclusion that ecological niches are actually only stagnation points in development, and they do not represent optimal states of adaptation. None of his many attempts to invoke the principles of population, sexual inheritance, recombination, dominance, and recessiveness to improve the convergence behavior yield the hoped for breakthrough. He thus eventually resigns himself to a largely deterministic strategy. In the linear programming problem, he chooses from the starting point several random directions and follows these in turn up to the boundary of the feasible region. The best states on the individual bounding hyperplanes are used to determine a new starting point by taking the arithmetic mean of the component parameters. Because of the convexity of the allowed region, the new starting point is always within it. The simultaneous choice of several search directions is supposed to be the analogue of the population principle and the construction of the average the analogue of recombination in sexual propagation. To tackle the problem of finding the minimum or maximum of an unconstrained, non-linear function, Bremermann even applies a five point Lagrangian interpolation to determine relative extrema in the random directions.

Rechenberg's *evolution strategy* changes all the components of the variable vector at each mutation. In his case, the low *mutation rate* for many dimensions is expressed by choosing small values for the step lengths, or the spread in the random changes. On the basis of theoretical work with two model functions he finds that the standard deviations of the random components are set optimally when they are inversely proportional to the number of parameters. His two membered evolution strategy resembles the scheme of Schumer and Steiglitz (1968), which is acknowledged to be particularly good, except that a $(0, \sigma^2)$ normally distributed random quantity replaces the fixed step length s. He has also added to it a step length modification rule, again derived from theory, which makes this look a very promising search method. It is refined in Chapter 5, Section 5.1 to meet the requirements of numerical optimization with digital computers. A multimembered strategy is treated in Section 5.2, which follows the same basic concept; however, by imitating the principles of population and recombination, it can operate without external control of the step lengths. Incorporating more than one descendant at a time and forgetting "parental wisdom" at the end of each iteration loop has provoked fierce objections against a more natural evolution strategy.

Box (1957) also considers that his EVOP (evolutionary operation) strategy resembles the biological mutation-selection process. He regards the vertices of his pattern of trial points, of which the best becomes the center of the next pattern, as individuals of a population, of which only the best "survives." The "offspring" are, however, generated by purely deterministic rules. Random decisions, as used by Satterthwaite (1959a; after Lowe, 1964) in his REVOP (random evolutionary operation) variant, are actually explicitly rejected by Box (see Youden et al., 1959; Satterthwaite, 1959b; Budne, 1959; Anscombe, 1959).

From a biological or cybernetic point of view, Pask (1962, 1971), Schmalhausen (1964), Berg and Timofejew-Ressowski (1964), Dobzhansky (1965), Moran (1967), and Kussul and Luk (1971) among others have examined the analogy between optimization and evolution. The fact that no practical algorithms have come out of this is no doubt because the evolutionary processes are described only verbally. Although they sometimes even include their more subtle effects, they have so far not produced a really quantitative, predictive theory. Exceptions, such as the work of Eigen (1971; see also Schuster, 1972), Merzenich

(1972), and Papentin (1972) are so different in emphasis that they are not applicable to the kind of problems considered here. The ways in which a process of *mathematization* can be implemented in theoretical biology are documented in for example the books by Waddington (1968) and Locker (1973), which contain a number of contributions of interest from the optimization point of view, as well as many articles in the journal *Mathematical Biosciences*, which has been published by R. W. Bellman since 1967, and some papers from two Berkeley symposia (LeCam and Neyman, 1967; LeCam, Neyman, and Scott, 1972). Whereas many modern books on biology, such as Riedl (1976) and Roughgarden (1979), still give mainly verbal explanations of organic evolution, in general, this is no longer the case. Physicists like Ebeling and Feistel (see Feistel and Ebeling, 1989) and biologists like Maynard Smith (1982, 1989) meanwhile have contributed mathematical models. The following two paragraphs thus no longer represent the actual situation, but before we add some new aspects they will be presented, nevertheless, to characterize the situation as perceived by the author in the early 1970s (Schwefel, 1975a):

> Relationships have been seen between random strategies and biological evolution on the one hand and the psychology of recognition processes on the other, for example, by Campbell (1960) and Khovanov (1967). The imitation of such processes–the catch phrase is *artificial intelligence*–always leads to the problem of choosing or designing a suitable search algorithm, which should rather be heuristic than strictly deterministic. Their simplicity, reliability (even in extreme, unfamiliar situations), and flexibility give the random strategies a special rôle in this field. The topic will not be discussed more fully here, except to mention some publications that explicitly deal with the relationship to optimization strategies: Friedberg (1958), Friedberg, Dunham, and North (1959), Minsky (1961), Samuel (1963), J. L. Barnes (1965), Vagin and Rudelson (1968), Thom (1969), Minot (1969), Ivakhnenko (1970), Michie (1971), and Slagle (1972). A particularly impressive example is given by the work of Fogel, Owens, and Walsh (1965, 1966a,b), in which imitation of the biological evolutionary principles of mutation and selection gives a (mathematical) automaton the ability to recognize prescribed sequences of numbers.
>
> It may be that in order to match the capabilities of the human brain–and to understand them–there must be a move away from the digital methods of present serial computers to quite different kinds of switching elements and coupling principles. Such concepts, as pursued in neurocybernetics and neurobionics, are described, for example, by Brajnes and Svečinskij (1971). The development of the *perceptron* by Rosenblatt (1958) can be seen as a first step in this direction.

Two research teams that have emphasized the adaptive capacity of evolutionary procedures and who have shown interesting computer simulations are Allen and McGlade (1986), and Galar, Kwasnicka, and Kwasnicki (see Galar, Kwasnicka, and Kwasnicki, 1980; Galar, 1994). In terms of the optimization tasks looked at throughout this book, one might call their point of view dynamic or on-line optimization, including optimum holding against environmental changes. As Schwefel and Kursawe (1992) have pointed out, a limited life span of all individuals is an important ingredient in such cases (*principle of forgetting*).

Two others who have tried to explain brain processes on an evolutionary, at least selectionist, basis are Edelman (1987) and Conrad (1988). Though their approach has not yet been embraced by the main stream of neural network research, this might happen in the near future (e.g., Banzhaf and Schmutz, 1992).

An even more general paradigm shift in the field of artificial intelligence (AI) has emerged under the label *artificial life* (AL; see Langton, 1989, 1994a,b; Langton et al., 1992; Varela and Bourgine, 1992). Whereas Lindenmayer (see Prusinkiewicz and Lindenmayer, 1990) demonstrates the possibility of (re-)creating plant forms by means of rather simple computer algorithms, the AL community tries to imitate animal behavior computationally. In most cases the goal is to design "intelligent" *robots*, sometimes called *knowbots* or *animats* (Meyer and Wilson, 1991; Meyer, 1992; Meyer, Roitblat, and Wilson, 1993).

The attraction of even simple evolutionary models (re-)producing fairly complex behavior of multi-individual systems simulated on computers is already spreading across the narrow bounds of computer science as such. New ideas are emerging from evolutionary computation, not only towards the organization of software development (Huberman, 1988), but also into the field of economics (e.g., Witt, 1992; Nissen, 1993, 1994) and even beyond (Schwefel, 1988; Haefner, 1992). It may be questionable whether worthwhile conclusions from the new findings can reach as far as that, but ecology at least should be a field that could benefit from a consequent use of evolutionary thinking (see Wolff, Soeder, and Drepper, 1988).

Computers have opened a third way of systems analysis aside from the classical mathematical/analytical and experimental/empirical main roads: i.e., numerical and/or symbolical simulation experiments. There is some hope that we may learn this way quickly enough so that we can maintain life on earth before we more or less unconsciously destroy it. Real evolution always had to deal with unpredictable environmental changes, not only those resulting from exogenous influences, but also self-induced endogenous ones. The landscape is some kind of n-dimensional trampoline, and every good imitation of organic evolution, whether it be called adaptive or meliorizing, must be able to work properly under such hard conditions. The multimembered evolution strategy (see Chap. 5, Sect. 5.2) with limited life span of the individuals fulfills that requirement to some extent.

Chapter 5

Evolution Strategies for Numerical Optimization

The task of mimicking biological structures and processes with the object of solving technical problems is as old as engineering itself. Mimicry itself, as a natural "strategy", is even older than mankind. The legend of Daedalus and Icarus bears early witness to such human endeavor. A sign of its scientific coming of age is the formation of the distinct branch of science known as bionics (e.g., Hertel, 1963; Gérardin, 1968; Beier and Glaß, 1968; Nachtigall, 1971; Heynert, 1972; Zerbst, 1987), which is concerned with the recognition of existing biological solutions to problems that also happen to arise in engineering, and with the adequate emulation of these examples. It is always thereby supposed that evolution has found particularly good, perhaps even optimal solutions. This assumption has often proved to be correct, or at any rate useful. Only a few attempts to imitate the actual methods of natural development are known (Ashby, 1960; Bremermann, 1962–1973; Rechenberg, 1964, 1973; Fogel, Owens, and Walsh, 1965, 1966a,b; Holland, 1975; see also Chap. 4) since they are curiously regarded a priori as being especially bad, meaning costly.

Rechenberg proposed the hypothesis "that the method of organic evolution represents an optimal strategy for the adaptation of living things to their environment," and he concludes "it should therefore be worthwhile to take over the principles of biological evolution for the optimization of technical systems."

5.1 The Two Membered Evolution Strategy

Rechenberg's two membered evolution scheme, suggested in similar form by other authors as a random strategy (see Chap. 4) will be expressed in this chapter as an algorithm for solving non-discrete, non-stochastic, parameter optimization problems. As in Chapter 3, the problem is

$$F(x) \to \min$$

where $x \in \mathbb{R}^n$. In the constrained case x has to be in an allowed region $G \subseteq \mathbb{R}^n$, where

$$G = \{x \in \mathbb{R}^n \,|\, G_j(x) \geq 0\,, j \;=\; 1(1)n\,, G_j \text{ restriction functions}\}$$

In this, as in all direct search methods, it is not possible to deal with constraints in the form of equalities.

5.1.1 The Basic Algorithm

The two membered scheme is the minimal concept for an imitation of organic evolution. The two principles of mutation and selection, which Darwin (1859) recognized to be most important, are taken as rules for variation of the parameters and for filtering during the iteration sequence respectively.

In the language of biology, the rules are as follows:

Step 0: (Initialization)
A given population consists of two individuals, one parent and one descendant. They are each identified by their genotype according to a set of n genes. Only the parental genotype has to be specified as starting point.

Step 1: (Mutation)
The parent $E^{(g)}$ of the generation g produces a descendant $N^{(g)}$, whose genotype is slightly different from that of the parent. The deviations refer to the individual genes and are random and independent of each other.

Step 2: (Selection)
Because of their different genotypes, the two individuals have a different capacity for survival (in the same environment). Only one of them can produce further descendants in the next generation, namely the one which represents the higher survival value. It becomes the parent $E^{(g+1)}$ of the generation $g+1$.

Thus the simplest possible assumptions are made:

- The population size remains constant
- An individual has in principle an infinitely long life span and capacity for producing descendants (asexually)
- No difference exists between genotype (encoding) and phenotype (appearance), or that one is unambiguously and reproducibly associated with the other
- Only point mutations occur, independently of each other at all single parameter locations
- The environment and thus the criterion of survival is constant over time

This minimal concept takes no account of the evolutionary factors familiar to the modern synthetic evolution theory (e.g., Stebbins, 1968; Čížek and Hodáňová, 1971; Osche, 1972), such as chromosome mutations, bisexuality, recombination, diploidy, dominance and recessiveness, population size, niching, isolation, migration, etc. Even the concepts of mutation and selection are not applied here with their full biological meaning. Natural selection does not simply mean the struggle between just two individuals in which the better survives, but far more accurately that an individual with more favorable properties produces on average more descendants than one less well adapted to the environment. Neither does the present work go more deeply into the connections between cause and effect in the transmission of inherited information, of which so much has been revealed by molecular biology. Mutation is used in the widest biological sense as a synonym for all types of alteration of the substance of inheritance. In his book *Evolutionsstrategie*, Rechenberg (1973) examines in more detail the analogy between natural evolution and technical optimization. He compares in particular the biological with the technical parameter space, and interprets mutations as steps in the nucleotide space.

Expressed in mathematical language, the rules are as follows:

Step 0: (Initialization)
There should be storage allocated in a (digital) computer for two points of an n-dimensional Euclidean space. Each point is characterized by a position vector consisting of a set of n components.

Step 1: (Variation)
Starting from point $E^{(g)}$, with position vector $x_E^{(g)}$, in iteration g, a second point $N^{(g)}$, with position vector $x_N^{(g)}$, is generated, the components $x_{N,i}^{(g)}$ of which differ only slightly from the $x_{E,i}^{(g)}$. The differences come about by the addition of (pseudo) random numbers $z_i^{(g)}$, which are mutually independent.

Step 2: (Filtering)
The two points or vectors are associated with different values of an objective function $F(x)$. Only one of them serves as a starting point for the new variation in the next iteration $g + 1$: namely the one with the better (for minimization, smaller) value of the objective function.

Taking account of constraints in the form of a barrier penalty function, this algorithm can be formalized as follows:

Step 0: (Initialization)
Define $x_E^{(0)} = \{x_{E,i}^{(0)}, i = 1(1)n\}^T$, such that $G_j(x_E^{(0)}) \geq 0$ for all $j = 1(1)m$.
Set $g = 0$.

Step 1: (Mutation)
Construct $x_N^{(g)} = x_E^{(g)} + z^{(g)}$ with components
$x_{N,i}^{(g)} = x_{E,i}^{(g)} + z_i^{(g)}$ for all $i = 1(1)n$.

Step 2: (Selection)
Decide

$$x_E^{(g+1)} = \begin{cases} x_N^{(g)}, & \text{if } F(x_N^{(g)}) \leq F(x_E^{(g)}) \text{ and } G_j(x_N^{(g)}) \geq 0 \text{ for all } j = 1(1)m \\ x_E^{(g)}, & \text{otherwise.} \end{cases}$$

Increase $g \leftarrow g + 1$ and go to step 1 as long as the termination criterion does not hold.

The question remains of how to choose the random vectors $z^{(g)}$. This choice has the rôle of mutation. Mutations are understood nowadays to be random, purposeless events, which furthermore only occur very rarely. If one interprets them, as is done here, as a sum of many individual events, it is natural to choose a probability distribution according to which small changes occur frequently, but large ones only rarely (the central limit theorem of statistics). For discrete variations one can use a *binomial distribution*, for example, for continuous variations a *Gaussian* or *normal distribution*.

Two requirements then arise together by analogy with natural evolution:

- That the expectation value ξ_i for a component z_i has the value zero
- That the variance σ_i^2, the average squared deviation from the mean, is small

The probability density function for normally distributed random events is (e.g., Heinhold and Gaede, 1972):

$$w(z_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\frac{(z_i - \xi_i)^2}{2\,\sigma_i^2}\right) \tag{5.1}$$

If $\xi_i = 0$, one obtains a so-called $(0, \sigma_i^2)$ normal distribution. There are still however a total of n free parameters $\{\sigma_i,\ i = 1(1)n\}$ with which to specify the standard deviations of the individual random components. By analogy with other, deterministic search strategies, the σ_i can be called step lengths, in the sense that they represent average values of the lengths of the random steps.

For the occurrence of a particular random vector $z = \{z_i,\ i = 1(1)n\}$, with the independent $(0, \sigma_i^2)$ distributed components z_i, the probability density function is

$$w(z_1, z_2, \ldots, z_n) = \prod_{i=1}^{n} w(z_i) = \frac{1}{(2\pi)^{\frac{n}{2}} \prod\limits_{i=1}^{n} \sigma_i} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{z_i}{\sigma_i}\right)^2\right) \tag{5.2}$$

or more compactly, if $\sigma_i = \sigma$ for all $i = 1(1)n$,

$$w(z) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(\frac{-z\,z^T}{2\,\sigma^2}\right) \tag{5.3}$$

For the length of the overall random vector $S = \sqrt{\sum_{i=1}^{n} z_i^2}$ a $\sigma\sqrt{\chi^2}$ distribution is obtained. The χ^2 distribution with n degrees of freedom approximates, for large n, to a $(\sigma\sqrt{n - \frac{1}{2}}, \frac{\sigma^2}{2})$ normal distribution. Thus the expectation value for the total length of the random vector for many variables is $E(S) = \sigma\sqrt{n}$, the variance is $D^2(S) = E((S - E(S))^2) = \frac{\sigma^2}{2}$, and the coefficient of variation is

$$\frac{D(S)}{E(S)} = \frac{1}{\sqrt{2\,n}}$$

This means that the most probable value for the length of the random vector at constant σ increases as the square root of the number of variables and the relative width of variation decreases with the reciprocal square root of parameters.

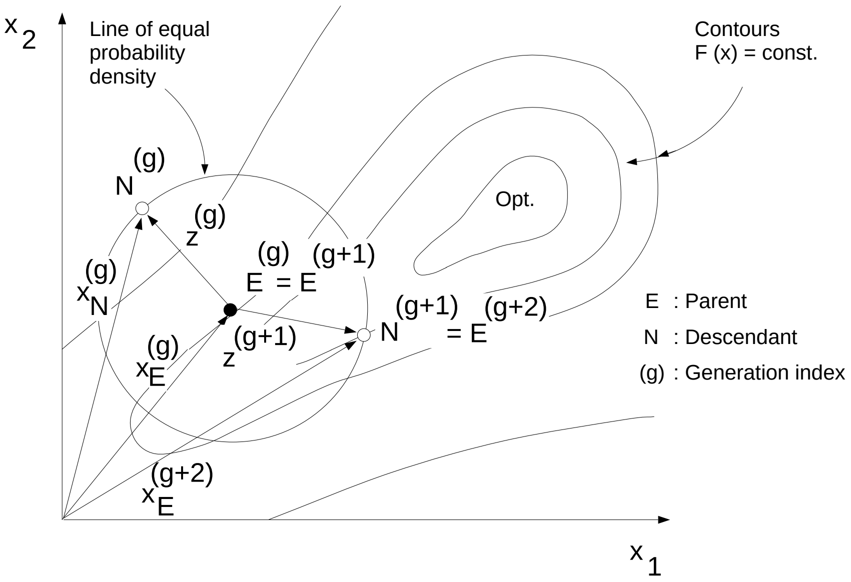

Figure 5.1: Two membered evolution strategy

The geometric locus of equally likely changes in variation of the variables can be derived immediately from the probability density function, Equation (5.2). It is an n-dimensional hyperellipsoid (n-fold variance ellipse) with the equation

$$\sum_{i=1}^{n} \left(\frac{z_i}{\sigma_i} \right)^2 = const.$$

referred to its center, which is the starting point $x_E^{(g)}$. In the multidimensional case, the random changes can be regarded as a vector ending on the surface of a hyperellipsoid with the semi-axes σ_i; or if $\sigma_i = \sigma$ for all $i = 1(1)n$, in the language of two dimensions they are distributed circumferentially. Figure 5.1 serves to illustrate two iteration steps of the evolution strategy on a two dimensional contour diagram. Whereas in other, fully deterministic search strategies both the direction and length of the search step are determined in the procedure in a fixed manner, or on the basis of previously gathered information and plausible assumptions about the topology of the objective function, in the evolution strategy the direction is purely random and the step length–except for a small number of variables–is practically fixed. This should be emphasized again to distinguish this random method from Monte-Carlo procedures, in which the selected trial point is always fully independent of the previous choice and its outcome. Darwin (1874) himself emphasized that the evolution of living things is not a purely random process. Yet against his theory of descendancy, a polemic is still waged in which the impossibility is demonstrated that life could arise by a purely random process (e.g., Jordan, 1970). Even at the level of the simplest imitation of organic evolution, a suitable choice of the step lengths or variances turns out to be of fundamental significance.

5.1.2 The Step Length Control

In experimental optimization, the appropriate step lengths can frequently be predicted. The values of the variables usually have to be determined exactly at only a few points. Thus constant values of the variances are often all that is required to complete an extreme value search. It is a matter of fact that in most experimental applications of the simple evolution strategy fixed (and discrete) distributions of mutations have been used.

By contrast, in mathematically formulated problems that are to be solved on a digital computer, the variables often run over much of the number range of the computer, which corresponds to many powers of 10. In a numerical optimum search the step lengths must be continuously modified if the algorithm is to be efficient–a problem reminiscent of steering safely between Scylla and Charybdis; for if the step length is too small the search takes an unnecessarily large number of iterations; if it is too large, on the other hand, the optimum can only be crudely approached and the search can even get stuck far from the optimum, for example, if the route to the minimum passes along a narrow valley. Thus in all optimization strategies the step length control is the most important part of the algorithm after the recursion formula, and it is furthermore closely linked to the convergence behavior.

The corresponding remarks hold for the evolution strategy, with the following difference: In place of a predetermined step length for a parameter of the objective function there is the variance of the random changes in this parameter, and instead of the statement that an improvement will or will not be made in a given direction with a specified step length, there can only be a statement of probability of the success or failure for a chosen variance.

In his theoretical investigations of the two membered evolution strategy, Rechenberg discovered using two basically different model objective functions (sphere model = Problem 1.1, corridor model = Problem 3.8 of the problem catalogue; see Appendix A) that the maximal rate of convergence corresponds to a particular value for the probability of a success, i.e., an improvement in the objective function value. He was thus led to formulate the following rule for controlling the size of the random changes:

> *The* 1/5 *success rule:*
>
> From time to time during the optimum search obtain the frequency of successes, i.e., the ratio of the number of successes to the total number of trials (mutations). If the ratio is greater than 1/5, increase the variance, if it is less than 1/5, decrease the variance.

In many problems this rule proves to be extremely effective in maintaining approximately the highest possible rate of progress towards the optimum. While in the right-angled corridor model the variances are adjusted once and for all in accordance with this rule and subsequently remain constant, in the sphere model they must steadily become smaller. The question then arises as to how often the success criterion should be tested and by what factor the variances are most effectively reduced or increased.

This question will be answered with reference to the sphere model introduced by Rechenberg, since this is the simplest non-linear model objective function and requires

the greatest and most frequent changes in the step lengths. The following results of Rechenberg's theory can be used here:
For the maximum rate of progress

$$\varphi_{\max} = k_1 \frac{r}{n}, \qquad k_1 \simeq 0.2025 \tag{5.4}$$

with a common variance σ^2, which is always optimal given by

$$\sigma_{opt} = k_2 \frac{r}{n}, \qquad k_2 \simeq 1.224 \tag{5.5}$$

for all components z_i of the random vector z. In these expressions r is the current distance from the goal (optimum) and n is the number of variables. The rate of progress is defined as the expectation value of the radial difference covered per trial (mutation), as illustrated in Figure 5.2.

$$\varphi^{(g)} = r^{(g)} - r^{(g+1)} \tag{5.6}$$

From Equations (5.4) to (5.6) one obtains a relation for the changes in the variance after a generation (iteration, or mutation) under the condition of maximum convergence rate

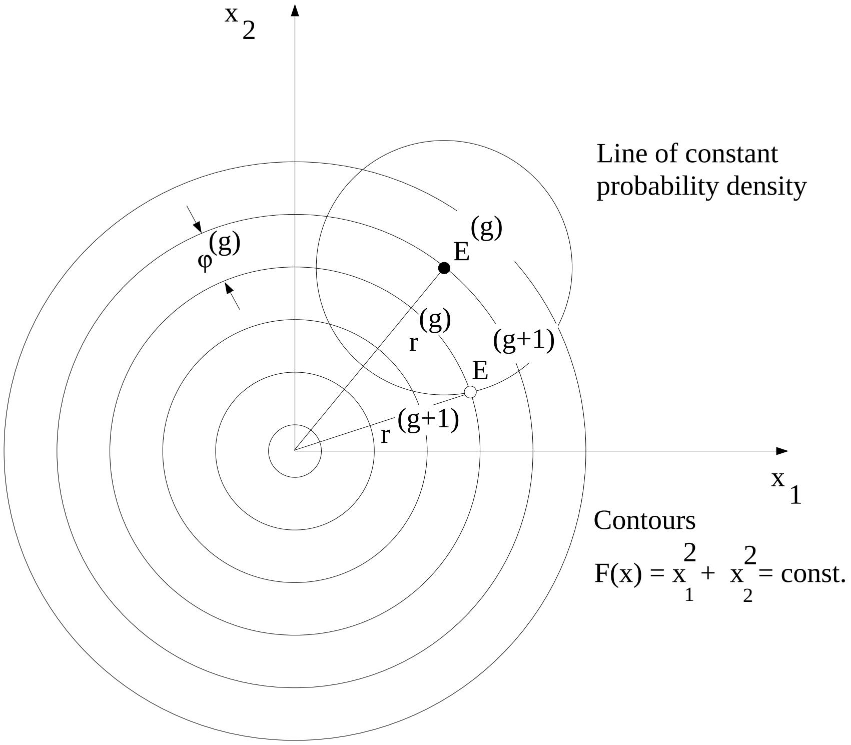

Figure 5.2: The rate of progress for the sphere model

$$\frac{\sigma_{opt}^{(g+1)}}{\sigma_{opt}^{(g)}} = \frac{r^{(g+1)}}{r^{(g)}} = 1 - \frac{k_1}{n}$$

or after n generations

$$\frac{\sigma_{opt}^{(g+n)}}{\sigma_{opt}^{(g)}} = \left(1 - \frac{k_1}{n}\right)^n$$

If n is large compared to one, and the formulae derived by Rechenberg are only valid under this assumption, the step length factor tends to a constant:

$$\lim_{n \to \infty} \left(1 - \frac{k_1}{n}\right)^n = e^{-k_1} \simeq 0.817 \simeq \frac{1}{1.224}$$

The same result is obtained by considering the rate of progress as a differential quotient $\varphi = dr/dg$, in which g represents the iteration number.

This matches the limiting case of very many variables because, according to Equation (5.4) the rate of progress is inversely proportional to the number of variables. The fact that the rate of progress φ near its maximum is quite insensitive to small changes in the variances, together with the fact that the probability of success can only be determined from an average over several mutations, leads to the following more precise formulation of the 1/5 success rule for numerical optimization:

> After every n mutations, check how many successes have occurred over the preceding $10\,n$ mutations. If this number is less than $2\,n$, multiply the step lengths by the factor 0.85; divide them by 0.85 if more than $2\,n$ successes occurred.

The 1/5 success rule enables the step lengths or variances of the random variations to be controlled. One might do even better by looking for a control mechanism with additional differential and integral coefficients to avoid the oscillatory behavior of a mere proportional feedback. However, the probability of success unfortunately gives no indication of how appropriate are the ratios of the variances σ_i^2 to each other. The step lengths can only be all reduced together, or all increased. One would sometimes rather like to build in a scaling of the variables, i.e., to determine ratios of the step lengths to each other. This can be achieved by a suitable formulation of the objective function, in which new parameters are introduced in place of the original variables. The functional dependence can be freely chosen and in the simplest case it is given by multiplicative factors. In the formulation of the numerical procedure for the two membered evolution strategy (Appendix B, Sect. B.1) the possibility is therefore included of specifying an initial step length for each individual variable. The ratios of the variances to each other remain constant during the optimum search, unless specified lower bounds to the step lengths are not operating at the same time.

All digital computers handle data only in the form of a finite number of units of information (bits). The number of significant figures and the range of numbers is thereby limited. If a quantity is repeatedly divided by a factor greater than one, the stored value of

the quantity eventually becomes zero after a finite number of divisions. Every subsequent multiplication leaves the value as zero. If it happens to one of the standard deviations σ_i, the affected variable x_i remains constant thereafter. The optimization continues only in a subspace of $\mathbb{R}^n$. To guard against this it must be required that $\sigma_i > 0$ for all $i = 1(1)n$. The random changes should furthermore be sufficiently large that at least the last stored place of a variable is altered. There are therefore two requirements:
Lower limits for the "step lengths":

$$\sigma_i^{(g)} \geq \varepsilon_a, \qquad \text{for all } i = 1(1)n$$

and

$$\sigma_i^{(g)} \geq \varepsilon_b \left| x_i^{(g)} \right|, \quad \text{for all } i = 1(1)n$$

where

$$\left.\begin{array}{r} \varepsilon_a > 0 \\ 1 + \varepsilon_b > 1 \end{array}\right\} \text{ according to the computational accuracy}$$

It is thereby ensured that the random variations are always active and the region of the search stays spanned in all dimensions.

5.1.3 The Convergence Criterion

In experimental optimization it is usually decided heuristically when to terminate the series of trials: for example, when the trial results indicate that no further significant improvement can be gained. One always has an overall view of how the experiment is running. In numerical optimization, if the calculations are made by computer, one must build into the program a rule saying when the iteration sequence is to be terminated. For this purpose objective, quantitative criteria are needed that refer to the data available at any time. Sometimes, although not always, one will be concerned to obtain a solution as exactly as possible, i.e., accurate to the last stored digit. This requirement can relate to the variables or to the objective function. Remember that the optimum may be a weak one.

Towards the minimum, the step lengths and distances covered normally become smaller and smaller. A frequently used convergence criterion consists of ending the search when the changes in the variables become zero (in which case no further improvement in the objective function is made), or when the step lengths have become zero. As a rule one sets the lower bound not to zero but to a sufficiently small, finite value. This procedure has however one disadvantage that can be serious. Small step lengths occur not only if the minimum is nearby, but also if the search is moving through a narrow valley. The optimization may then be practically halted long before the extreme value being sought is reached. In Equations (5.4) and (5.5), r can equally well be thought of as the local radius of curvature. Neither φ, the distance covered, nor σ, the step length, are a measure of the closeness to the optimum. Rather they convey information about the complexity of the minimum problem: the number of variables and the narrowness of valleys encountered. The requirement $\sigma > \varepsilon$ or $\|x^{(g)} - x^{(g-1)}\| > \varepsilon$ for the continuation of the search is thus no guarantee of sufficient convergence.

Gradient methods, which seek a point with vanishing first derivatives, frequently also apply this necessary condition for the existence of an extremum as a termination criterion. Alternatively the search can be continued until $\triangle F = F(x^{(k-1)}) - F(x^{(k)})$, the change in the objective function value in one iteration, goes to zero or to below a prescribed limit. But this requirement can also be fulfilled far from the minimum if the valley in which the deepest point is sought happens to be very flat in shape. In this case the step length control of the two membered evolution strategy ensures that the variances become larger, and thus the function value differences between two successful trials also on average become larger. This is guaranteed even if the function values are equal (within computational accuracy), since a change in the variables is always then registered as a success. One thus has only to take care that $\triangle F$ is summed over a number of results in order to derive a termination criterion. Just as lower bounds are defined for the step lengths, an absolute and a relative bound can be specified here:

Termination rule:

End the search if

$$F(x_E^{(g-\triangle g)}) - F(x_E^{(g)}) \le \varepsilon_c$$

or

$$\frac{1}{\varepsilon_d}\left[F(x_E^{(g-\triangle g)}) - F(x_E^{(g)})\right] \le |F(x_E^{(g)})|$$

where

$$\triangle g \ge 20\,n$$

and

$$\left.\begin{array}{l} \varepsilon_c > 0 \\ 1 + \varepsilon_d > 1 \end{array}\right\} \text{according to the computational accuracy}$$

The condition $\triangle g \ge 20\,n$ is designed to ensure that in the extreme case the standard deviations are reduced or increased within the test period by at least the factor $(0.85)^{\pm 20} \simeq (25)^{\pm 1}$, in accordance with the 1/5 success rule. This will prevent the search being terminated only because the variances are forced to change suddenly. It is clear from Equation (5.4) that the more variables are involved in the problem, the slower is the rate of progress. Hence it does not make sense to test the convergence criterion very frequently. A recommended procedure is to make a test every $20\,n$ mutations. Only one additional function value then needs to be stored.

Another reason can be adduced for linking the termination of the search to the function value changes. While every success in an optimum search means, in the end, an economic profit, every iteration costs computer time and thus money. If the costs exceed the profit, the optimization may well provide useful information, but it is certainly not on the whole of any economic value. Thus someone who only wishes to optimize from an economic point of view can, by a suitable choice of values for the accuracy parameters, restrain the search process as soon as it starts running into a loss.

5.1.4 The Treatment of Constraints

Inequality constraints $G_j(x) \geq 0$ for all $j = 1(1)m$ are quite acceptable. Sign conditions may be formulated in the same manner and do not receive any special treatment. In contrast to linear and non-linear programming, no sign conditions need to be set in order to keep within a bounded region. If a mutation falls in the forbidden region it is assessed as a worsening (in the sense of a lethal mutation) and the variation of the variables is not accepted.

No particular penalty function, such as Rosenbrock chooses for his method of rotating coordinates, has been developed for the evolution strategy. The user is free to use the techniques for example of Carroll (1961), Fiacco and McCormick (1968), or Bandler and Charalambous (1974), to construct a suitable sequence of substitute objective functions and to solve the original constrained problem as a sequence of unconstrained problems. This, however, can be done outside the procedure.

It is sometimes difficult to specify an allowed initial vector of the variables. If one were to wait until by chance a mutation satisfied all the constraints, it could take a very long time. Besides, during this search period the success checks could not be carried out as described above. It would nevertheless be desirable to apply the normal search algorithm effectively to find an allowed state. Box (1965) has given in the description of his complex method a simple way of proceeding from a forbidden starting point. He constructs an auxiliary objective function from the sum of the constraint function values of the violated constraints:

$$\tilde{F}(x) = \sum_{j=1}^{m} G_j(x)\, \delta_j(x)$$

where

$$\delta_j(x) = \begin{cases} -1\,, & \text{if } G_j(x) < 0 \\ 0\,, & \text{otherwise} \end{cases} \tag{5.7}$$

Each decrease in the value of $\tilde{F}(x)$ represents an approach to the feasible region. When eventually $\tilde{F}(x) = 0$, then x satisfies all the constraints and can serve as a starting vector for the optimization proper. This procedure can be taken over without modification for the evolution strategy.

5.1.5 Further Details of the Subroutine EVOL

In Appendix B, Section B.1 a complete FORTRAN listing is given of a subroutine corresponding to the two membered evolution scheme that has been described. Thus no detailed algorithm will be formulated here, but a few further programming details will be mentioned.

In nearly all digital computers there are library subroutines for generating uniformly distributed pseudorandom numbers. They work, as a rule, according to the multiplicative or additive congruence method (see Jöhnk, 1969; Niederreiter, 1992; Press et al., 1992). From any two numbers taken at random from a uniform distribution in the range $[0, 1]$, by using the transformation rules of Box and Muller (1958) one can generate two independent, normally distributed random numbers with the expectation values zero and the variances unity. The formulae are

$$\begin{aligned} Z'_1 &= \sqrt{-2\,\ln Y_1}\,\sin(2\pi\,Y_2) \\ \text{and} \qquad Z'_2 &= \sqrt{-2\,\ln Y_1}\,\cos(2\pi\,Y_2) \end{aligned} \tag{5.8}$$

where the Y_i are the uniformly distributed and the Z'_i $(0,1)$-normally distributed random numbers respectively. To obtain a distribution with a variance different from unity, the Z'_i must simply be multiplied by the desired standard deviation σ_i (the "step length"):

$$Z_i = \sigma_i\,Z'_i$$

The transformation rules are contained in a function procedure separate from the actual subroutine. To make use of both Equations (5.8) a switch with two settings is defined, the condition of which must be preset in the subroutine once and for all. In spite of Neave's (1973) objection to the use of these rules with uniformly distributed random numbers that have been generated by a multiplicative congruence method, no significant differences could be observed in the behavior of the evolution strategy when other random generators were used. On the other hand the trapezium method of Ahrens and Dieter (1972) is considerably faster.

Most algorithms for parameter optimization include a second termination rule, independent of the actual convergence criterion. They end the search after no more than a specified number of iterations, in order to avoid an infinite series of iterations in case the convergence criterion should fail. Such a rule is effectively a bound on the computation time. The program libraries of computers usually contain a procedure with which the CPU time used by the program can be determined. Thus instead of giving a maximum number of iterations one could specify a maximum computation time as a termination criterion. In the present program the latter option is adopted. After every n iterations the elapsed CPU time is checked. As soon as the limit is reached the search ends and output of the results can be initiated from the main program.

The 1/5 success rule assumes that there is always some combination of variances $\sigma_i > 0$ with which, on average, at least one improvement can be expected within five mutations. In Figure 5.3 two contour diagrams are shown for which the above condition cannot always be met. At some points the probability of a success cannot exceed 1/5 : for example, at points where the objective function has discontinuous first partial derivatives or at the edge of the allowed region. Especially in the latter case, the selection principle progressively forces the sequence of iteration points closer up to the boundary and the step lengths are continuously reduced in size, without the optimum being approached with comparable accuracy.

Even in the corridor model (Problem 3.8 of Appendix A, Sect. A.3) difficulties can arise. In this case the rate of progress and probability of success depend on the current position relative to the edges of the corridor. Whereas the maximum probability of success in the middle of the corridor is 1/2, at the corners it is only 2^{-n}. If one happens to be in the neighborhood of the edge of the corridor for several mutations, the probability of success

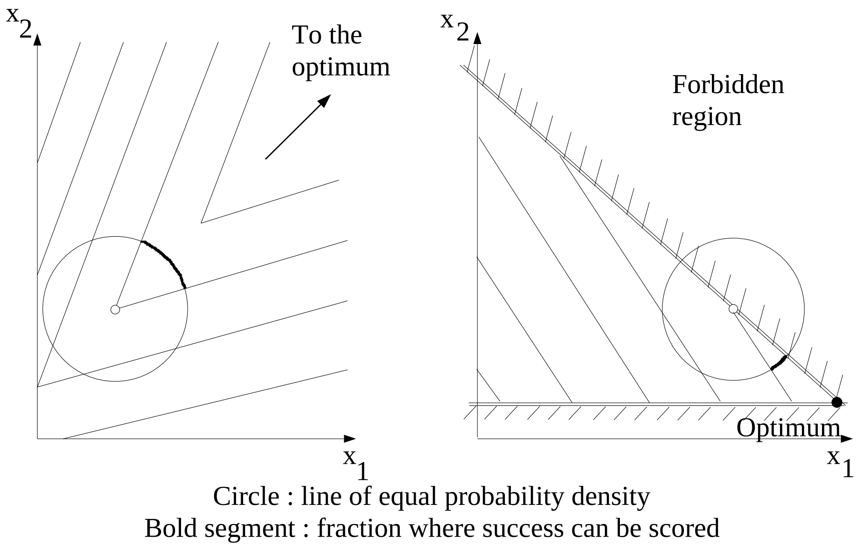
x2
To the optimum
x1
x2
Forbidden region
Optimum
x1
Circle : line of equal probability density
Bold segment : fraction where success can be scored

Figure 5.3: Failure of the 1/5 success rule

calculated by the above rule will be very different from that associated with the same step length if an average over the corridor cross section were taken. If now, on the basis of this low estimate of the success probability, the step length is further reduced, there is a corresponding decrease in the probability of escaping from the edge of the corridor. It would therefore be desirable in this special case to average the probability of success over a longer time period. Opposed to this, however, is the requirement from the sphere model that the step lengths should be adjusted to the topology as directly as possible. The present subroutine offers several means of dealing with the problem. For example, the lower bounds on the variances (variables EA, EB in the subprogram EVOL) can be chosen to be relatively large, or the number of mutations (the variable LS) after which the convergence criterion is tested can be altered by the user. The user has besides a free choice with regard to the required probability of success (variable LR) and the multiplier of the variance (variable SN). The rate of change of the step lengths, given by the factor 0.85 per n mutations, was fixed on the basis of the sphere model. It is not ideal for all types of problems but rather in the nature of a lower bound. If it seems reasonable to operate with constant variances, the parameter in question should be set equal to one.

An indication of a suitable choice for the initial step lengths (variable array SM) can be obtained from Equation (5.4). Since r increases as the root of the number of parameters, one is led to set

$$\sigma_i^{(0)} = \frac{\triangle x_i}{\sqrt{n}}$$

in which $\triangle x_i$ is a rough measure of the expected distance from the optimum. This does not actually give the optimal step length because r is a kind of local scale of curvature of the contours of the objective function. However, it does no harm to start with variances that are too large; they will quickly be reduced to a suitable size by the 1/5 success rule. During this transition phase there is still a chance of escaping from the neighborhood of a merely local optimum but very little chance afterwards. The global convergence property (see Rechenberg, 1973) of the evolution strategy can only be proved under the condition of constant step lengths. With the introduction of the success rule, it is lost, or to be more precise: the probability of finding the global minimum among several local minima decreases continuously as a local minimum is approached with continuous reduction in the step lengths. Rapid convergence and reliable global convergence behavior are two contradictory requirements. They cannot be reconciled if one has absolutely no knowledge of the topology of the objective function. The 1/5 success rule is aimed at high convergence rates. If several local optima are expected, it is thus advisable to keep the variances large and constant, or at least to start with large $\sigma_i^{(0)}$ and perhaps to require a lower success probability than 1/5. This measure naturally costs extra computation time. Once one is sure of having located a point near the global extremum, the accuracy can be improved subsequently in a follow-up computation. For more sophisticated investigations of the global convergence see Born (1978), Rappl (1984), Scheel (1985), Bäck, Rudolph, and Schwefel (1993), and Beyer (1993).

5.2 A Multimembered Evolution Strategy

While the simple, two membered evolution strategy is successful in application to many optimization problems, it is not a satisfactory method of solving certain types of problem. As we have seen, by following the 1/5 success rule, the step lengths can be permanently reduced in size without thereby improving the rate of progress. This phenomenon occurs frequently if constraints become active during the search, and greatly reduce the size of the success scoring region. A possible remedy would be to alter the probability distribution of the random steps in such a way as to keep the success probability sufficiently large. To do so the standard deviations σ_i would have to be individually adjustable. The contour surfaces of equal probability could then be stretched or contracted along the coordinate axes into ellipsoids. Further possibilities for adjustment would arise if the random components were allowed to depend on each other. For an arbitrary quadratic problem the rate of convergence of the sphere model could even be achieved if the random changes of the individual variables were correlated so as to make the regression line of the random vector run parallel to the concentric ellipsoids $F(x) = const.$, which now lie at some angle in the space. To put this into practice, information about the topology of the objective function would have to be gathered and analyzed during the optimum search. This would start to turn the evolution strategy into something resembling one of the familiar deterministic optimization methods, as Marti (1980) and recently again Ostermeier (1992) have done; this is contrary to the line pursued here, which is to apply biological evolution principles to the numerical solution of optimization problems. Following Rechenberg's hypothesis, construction of an improved strategy should therefore be attempted by taking into account further evolution principles.

5.2.1 The Basic Algorithm

When the ground rules of the two membered evolution strategy were formulated in the language of biology, reference was to one parent and one offspring; the basic population thus consisted of two individuals. In order to reach a higher level of imitation of the evolutionary process, the number of individuals must be increased. This is precisely the concept behind the evolution strategy referred to in the following as *multimembered.* In his basic work (Rechenberg, 1973), Rechenberg already presented a scheme for a multimembered evolution. The one considered here is somewhat different. It turns out to be particularly useful with respect to the individual control of several step lengths to be described later. As yet, however, no detailed comparison of the two variants has been undertaken.

It is useful to introduce at this point a *nomenclature* for the different evolution strategies. We shall call the number of parents of a generation μ, and the number of descendants λ, so that the selection takes place between $\mu + \lambda = 1 + 1 = 2$ individuals in the two membered strategy. We thus characterize the simplest imitation of evolution in abbreviated notation as the (1+1) strategy. Since the multimembered evolution scheme described by Rechenberg allows a selection between $\mu > 1$ parents and $\lambda = 1$ offspring it should be called the (μ+1) strategy. Accordingly a more general form, a (μ+λ) evolution strategy, should be formulated in such a way that a basic population of μ parents of generation g produces λ offspring. The process of selection only allows the μ best of all $\mu + \lambda$ individuals to proceed as parents of the following generation, be they offspring of generation g or their parents. In this model it could happen that a parent, because of its vitality, is far superior to the other parents in the same generation, "lives" for a very long time, and continues to produce further offspring. This is at variance to the biological fact of a *limited life span*, or more precisely a limited capacity for reproduction. Aging phenomena do not, as far as is known, affect biological selection (see Savage, 1966; Osche, 1972). As a further conceptual model, therefore, let us introduce a population in which μ parents produce $\lambda > \mu$ offspring but the μ parents are not included in the selection. Rather the parents of the following generation should be selected only from the λ offspring. To preserve a constant population size, we require that each time the μ best of the λ offspring become parents of the following generation. We will refer to this scheme in what follows as the (μ , λ) strategy. As for the (1+1) strategy in Section 5.1.1, the algorithm of the multimembered (μ , λ) strategy will first be formulated in the language of biology.

Step 0: (Initialization)
A given population consists of μ individuals. Each is characterized by its genotype consisting of n genes, which unambiguously determine the vitality, or fitness for survival.

Step 1: (Variation)
Each individual parent produces λ/μ offspring on average, so that a total of λ new individuals are available. The genotype of a descendant differs only slightly from that of its parents. The number of genes, however, remains to be n in the following, i.e., neither gene duplication nor gene deletion occurs.

Step 2: (Filtering)
Only the μ best of the λ offspring become parents of the following generation.

In mathematical notation, taking constraints into account, the rules are as follows:

Step 0: (Initialization)
Define $x_k^{(0)} = x_{E_k}^{(0)} = (x_{k,1}^{(0)}, \ldots, x_{k,n}^{(0)})^T$ for all $k = 1(1)\mu$.
$x_k^{(0)} = x_{E_k}^{(0)}$ is the vector of the kth parent E_k, such that
$G_j(x_k^{(0)}) \geq 0$ for all $k = 1(1)\mu$ and all $j = 1(1)m$.
Set the generation counter $g = 0$.

Step 1: (Mutation)
Generate $x_\ell^{(g+1)} = x_k^{(g+1)} + z^{(g\lambda+\ell)}$,
such that $G_j(x_\ell^{(g+1)}) \geq 0,\ j = 1(1)m, \ell = 1(1)\lambda$,
where $k \in [1, \mu]$

$$\text{e.g.,} \quad k = \begin{cases} \mu, & \text{if } \ell = p\,\mu\,, p\,\text{integer} \\ \ell(\text{mod}\,\mu), & \text{otherwise.} \end{cases}$$

$x_\ell^{(g+1)} = x_{N_\ell}^{(g+1)} = (x_{\ell,1}^{(g+1)}, \ldots, x_{\ell,n}^{(g+1)})^T$ is the vector of the ℓth offspring N_ℓ,
and $z^{(g\lambda+\ell)}$ is a normally distributed random vector with n components.

Step 2: (Selection)
Sort the $x_\ell^{(g+1)}$ for all $\ell = 1(1)\lambda$ so that
$F(x_{\ell_1}^{(g+1)}) \leq F(x_{\ell_2}^{(g+1)})$, for all $\ell_1 = 1(1)\mu$, $\ell_2 = \mu + 1(1)\lambda$
Assign $x_k^{(g+2)} = x_{\ell_1}^{(g+1)}$, for all $k, \ell_1 = 1(1)\mu$.
Increase the generation counter $g \leftarrow g + 1$.
Go to step 1, unless some termination criterion is fulfilled.

What happens in one generation for a $(2\,,4)$ evolution strategy is shown schematically on the two dimensional contour diagram of a non-linear optimization problem in Figure 5.4.

5.2.2 The Rate of Progress of the $(1\,,\lambda)$ Evolution Strategy

In this section we attempt to obtain approximately the rate of progress of the multimembered, or $(\mu\,,\lambda)$ strategy–at least for $\mu = 1$. For this purpose the n-dimensional sphere and corridor models, as used by Rechenberg (1973), are employed for calculating the progress for the (1+1) strategy.

In the two membered evolution strategy φ was the expectation value of the useful distance covered in each mutation. It is convenient here to define the rate in terms of the number of generations.

$$\varphi = \text{expectation value}\left(\|\hat{x} - \bar{x}^{(g)}\| - \|\hat{x} - \bar{x}^{(g-1)}\|\right)$$

where $\hat{x}$ is the vector of the optimum and $\bar{x}^{(g)}$ is the average vector of the parents of generation g.

From the chosen n-dimensional normal distribution of the random vector, which has expectation value zero and variance σ^2 for all independent vector components, the probability density for going from a point E with vector $x_E = (x_{E,1}, \ldots, x_{E,n})^T$ to another

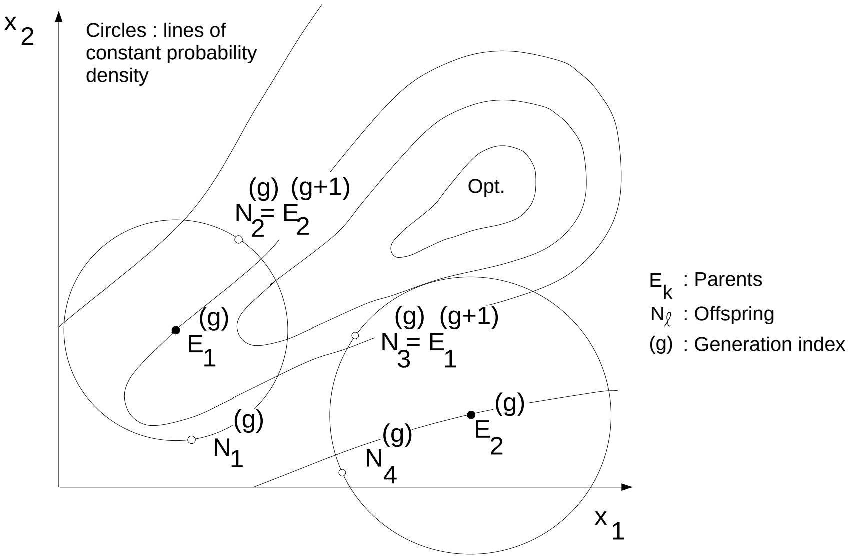

Figure 5.4: Multimembered (2 , 4) evolution strategy

point N with vector $x_N = (x_{N,1}, \ldots, x_{N,n})^T$ is

$$w(E \rightarrow N) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(x_{E,i} - x_{N,i})^2\right) \tag{5.9}$$

The distance $\|x_E - x_N\|$ between x_E and x_N is

$$\|x_E - x_N\| = \sqrt{\sum_{i=1}^{n}(x_{E,i} - x_{N,i})^2}$$

But of this, only a part, $s = f(x_E, x_N)$, is useful in the sense of approaching the objective. To discover the total probability density for covering a useful distance s, an integration must be performed over the locus of points for which the useful distance is s, measured from the starting point x_E. This locus is the surface of a finite region in n-dimensional space:

$$p(s) = \underset{f(x_E, x_N) = s}{\int \cdots \int} w(E \rightarrow N)\, dx_{N,1}\, dx_{N,2} \ldots dx_{N,n} \tag{5.10}$$

The result of the integration depends on the weighting function $f(x_E, x_N)$ and thus on the topology of the objective function $F(x)$.

So far only one random change has been considered. In the multimembered evolution strategy, however, the average over the μ best of the λ offspring must be taken, in which each of the offspring is to be associated with its own distance s_ℓ. We first have to find the probability density $w_\nu(s')$ for the νth best descendant of a generation to cover the useful distance s'. It is a combinatorial product of

- The probability density $w(s_{\ell_1} = s')$ that a particular descendant N_{ℓ_1} gets exactly s' closer to the objective
- The probability $p(s_{\ell_2} > s')$ that a descendant N_{ℓ_2} advances further than s'
- The probability $p(s_{\ell_3} < s')$ that a descendant N_{ℓ_3} advances less than s'

Better results must be given by $\nu - 1$ descendants and worse by $\lambda - \nu$. This results in a large number of combinations, since it is of no significance which descendant is in which place.

$$\begin{aligned} w_\nu(s') = \sum_{\ell_1=1}^{\lambda} \Big\{ w(s_{\ell_1} = s') \cdot \sum_{\substack{\ell_2=1\\ \ell_2 \neq \ell_1}}^{\lambda-\nu+2} \Big\{ p(s_{\ell_2} > s') \cdot \sum_{\substack{\ell_3=\ell_2+1\\ \ell_3 \neq \ell_1}}^{\lambda-\nu+3} \Big\{ p(s_{\ell_3} > s') \cdot \\ \cdot \sum_{\substack{\ell_4=\ell_3+1\\ \ell_4 \notin \{\ell_1,\ell_2\}}}^{\lambda-\nu+4} \Big\{ p(s_{\ell_4} > s') \quad \cdots \quad \sum_{\substack{\ell_\nu=\ell_{\nu-1}+1\\ \ell_\nu \notin \{\ell_1,\ell_2,\ldots,\ell_{\nu-2}\}}}^{\lambda} \Big\{ p(s_{\ell_\nu} > s') \cdot \\ \cdot \prod_{\substack{\ell_{\nu+1}=1\\ \ell_{\nu+1} \notin \{\ell_1,\ell_2,\ldots,\ell_\nu\}}}^{\lambda} p(s_{\ell_{\nu+1}} < s') \Big\} \cdots \Big\}\Big\}\Big\}\Big\} \end{aligned} \qquad (5.11)$$

As an average of the μ best descendants one obtains

$$w(s') = \frac{1}{\mu} \sum_{\nu=1}^{\mu} w_\nu(s') \qquad (5.12)$$

and hence the rate of progress

$$\varphi = \int_{s'=s_u}^{\infty} s' \, w(s') \, ds' \qquad (5.13)$$

The meaning of s_u will be described later.

To evaluate φ, besides μ and λ, all components of the position vectors of all parents of the generation must be given, together with the values of σ for producing each descendant. If φ is to become independent of a particular initial configuration, it is necessary to define representative or average values of the relative positions of the parents, which are established during the optimization as a function of the topology. To do so would require setting up and solving an integral equation. This has not yet been achieved.

To be able to say something nevertheless about the rate of convergence some simplifying assumptions will be made. All parents will be represented by a single position vector x_k, and the standard deviations $\sigma_{\ell,i}$ will be assumed equal for all components $i = 1(1)n$ and for the descendants $\ell = 1(1)\lambda$. Equation (5.11) thereby simplifies to

$$w_\nu(s') = \lambda \left(\begin{array}{c} \lambda - 1 \\ \nu - 1 \end{array} \right) w(s_\ell = s') \, [p(s_\ell < s')]^{\lambda-\nu} \, [p(s_\ell > s')]^{\nu-1}$$

Since

$$p(s_\ell > s') + p(s_\ell < s') = 1$$

and

$$\left(\begin{array}{c} \lambda - 1 \\ \nu - 1 \end{array} \right) = \frac{(\lambda - 1)\,!}{(\nu - 1)\,!\,(\lambda - \nu)\,!}$$

we have

$$w_\nu(s') = \frac{\lambda\,!}{(\nu - 1)\,!\,(\lambda - \nu)\,!}\, w(s_\ell = s')\, [p(s_\ell < s')]^{\lambda - \nu}\, [1 - p(s_\ell < s')]^{\nu - 1} \qquad (5.14)$$

Henceforth the number of parents is reduced to $\mu = 1$. One parent produces all λ descendants. Of these, because of the assumption of constant population size, only the best survives. All the others are rejected. Accordingly we are now dealing with a $(1\,,\lambda)$ strategy, for which Equation (5.12) reduces to

$$w(s') = w_1(s') = \lambda\, w(s_\ell = s')\, [p(s_\ell < s')]^{\lambda - 1} \qquad (5.15)$$

where

$$w(s_\ell = s') = \underset{f(x_E, x_N) = s}{\int \cdots \int} \left(\frac{1}{\sqrt{2\pi}\,\sigma} \right)^n \exp\left(-\frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (x_{E,i} - x_{N,i})^2 \right) dx_{N,1} \ldots dx_{N,n}$$

and

$$p(s_\ell < s') = \int_{s_\ell = -\infty}^{s'} w(s_\ell = s')\, ds_\ell$$

If we now make use of the corridor and sphere model objective functions, as chosen by Rechenberg in his work, we can directly take over some of his results; in particular the integrations for the calculation of $w(s_\ell = s')$ and $p(s_\ell < s')$. The final integration (Equation (5.13)) for determining φ turns out to be impossible to evaluate in closed form. To find a suitable way around this let us take a closer look at Equation (5.13). It has the form of an equation for the mean value (expectation value) of the probability density $w(s')$ in the interval $s_u \leq s' \leq \infty$. The lower limit s_u of the range depends on whether, in cases when none of the offspring represent an improvement over the parent, the selection allows either the parent to survive (so-called "plus" version), or only the best of all offspring (so-called "comma" version), in which case the chance of deterioration is greater than zero. It will turn out later that the optimization can actually benefit if setbacks are permitted.

We therefore distinguish the two cases:

- The parent is included in the selection process and can in theory survive an infinite number of generations: $s_u = 0, (1{+}\lambda)$ strategy

- The parent is no longer considered in the selection: $s_u = -\infty$, $(1\,,\lambda)$ strategy

In the second case the integral for p extends over the total interval in which the variable of integration s' can lie. Now if the function $w(s')$ happens to be symmetrical and unimodal, the expectation value can be found in a different way. Namely, it would be equal to the value of s' at which the probability density $w(s')$ reaches its maximum. For a skew distribution this is not the case. Perhaps, however, the skewness is only slight, so that one can determine at least approximately the expectation value from the position of the maximum.

Before treating the sphere and corridor models in this way, we will check the usefulness of the scheme with an even simpler objective function.

5.2.2.1 The Linear Model (Inclined Plane)

The simplest way the objective function can depend on the variables is linearly. Imagining the function to be a terrain in the $(n+1)$-dimensional space, it appears as an inclined plane. In the two dimensional projection the contours are straight, parallel lines in this case. Without loss of generality one can orient the coordinate system so that the plane only slopes in the direction of one axis x_1 and the starting point or parent under consideration lies at the origin (Fig. 5.5).

The useful distance s_ℓ towards the objective that is covered by descendant N_ℓ of the parent E is just the part of the random vector z lying along the x_1 axis. Since the components z_i of z are independent, we have

$$w(s_\ell = s') = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{s'^2}{2\,\sigma^2}\right)$$

and

$$p(s_\ell < s') = \int\limits_{s_\ell=-\infty}^{s'} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{s_\ell^2}{2\,\sigma^2}\right) ds_\ell = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{s'}{\sqrt{2}\,\sigma}\right)\right]$$

Substituting these two results in Equation (5.15) we obtain the probability density for the best of λ offspring of a parent covering the useful distance s':

$$w(s') = \lambda\,\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(\frac{-s'^2}{2\,\sigma^2}\right)\left(\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{s'}{\sqrt{2}\,\sigma}\right)\right]\right)^{\lambda-1} \tag{5.16}$$

To obtain the position of the maximum we differentiate with respect to s' and set the result equal to zero. The associated value of s' is then the sought for approximation $\tilde{\varphi}$ to the rate of progress φ.

From

$$\left.\frac{\partial w(s')}{\partial s'}\right|_{s'=\tilde{\varphi}} \overset{!}{=} 0$$

it follows that

$$\lambda = 1 + \frac{\sqrt{\pi}\,\tilde{\varphi}}{\sqrt{2}\,\sigma} \exp\left(\frac{\tilde{\varphi}^2}{2\,\sigma^2}\right)\left[1 + \operatorname{erf}\left(\frac{\tilde{\varphi}}{\sqrt{2}\,\sigma}\right)\right] \tag{5.17}$$

Figure 5.6 shows how the function $\tilde{\varphi}/\sigma$, which is just $\tilde{\varphi}$ for $\sigma = 1$, depends on λ. For $\lambda = 1$ the rate of progress is equal to zero, independent of the step length. This must be so because for only one descendant the probability of improvement is the same as that of worsening. As the number of descendants increases so does the rate of progress,

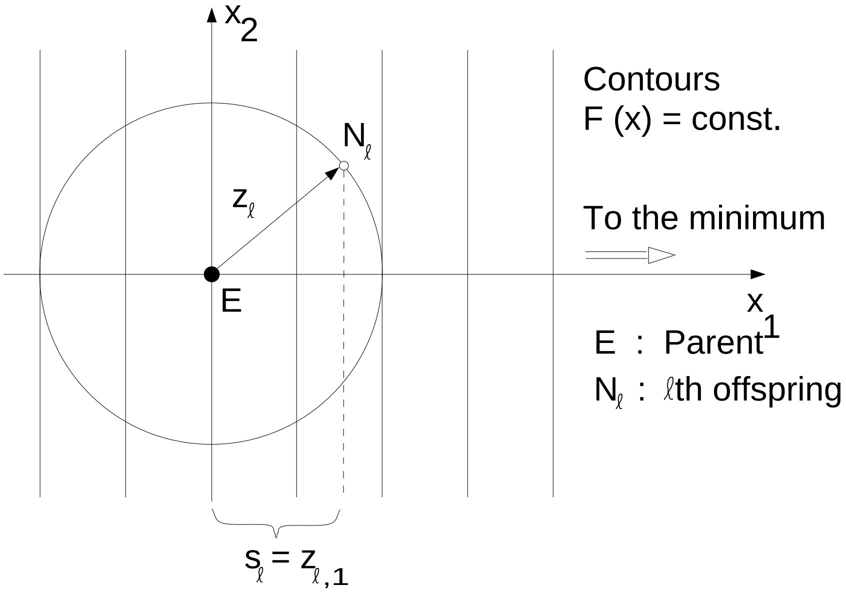

Figure 5.5: The inclined plane model function

sublinearly however, probably proportional to the logarithm of λ. To compare the above approximation $\tilde{\varphi}$ with the exact value φ the following integral must be evaluated:

$$\varphi = \int\limits_{s'=s_u}^{\infty} \lambda \frac{s'}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{s'^2}{2\sigma^2}\right) \left(\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{s'}{\sqrt{2}\,\sigma}\right)\right]\right)^{\lambda-1} ds'$$

For small values of λ the integration can be performed by elementary methods, but not for general values of λ. The value of φ was therefore obtained by simulation on the computer; first for the case in which the parent survives if the best of the descendants is worse than the parent (φ_{sur} with $s_u = 0$) and secondly for the case in which the parent is no longer considered in the selection (φ_{ext} with $s_u = -\infty$). The two results are shown in Figure 5.6 for comparison with the approximate solution $\tilde{\varphi}$. It is immediately striking that for only five offspring the extinction of the parent has hardly any effect on the rate of progress, i.e., for $\lambda \geq 5$ it is as good as certain that at least one of the descendants will be better than the parent. The greatest differences between φ_{sur} and φ_{ext} naturally appear when $\lambda = 1$. Whereas φ_{ext} goes to zero, φ_{sur} keeps a finite value. This can be determined exactly. Omitting here the details of the derivation, which is straightforward, the result is simply

$$\varphi_{sur}(\lambda = 1) = \frac{\sigma}{\sqrt{2\pi}}$$

The relationship to the (1+1) evolution scheme is thereby established. The differences between the approximate theory ($\tilde{\varphi}$) and the simulation (φ_{ext}) indicate that the assumption of the symmetry of $w(s')$ is not correct. The discrepancy with regard to φ/σ seems to tend to a constant value as λ increases. While the approximate theory is shown by this

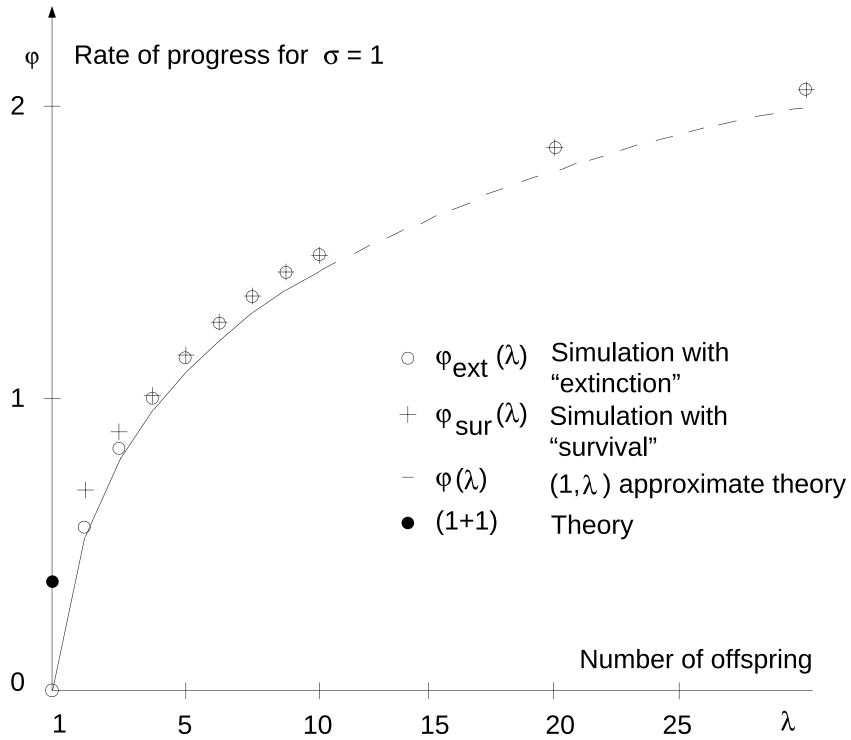

Figure 5.6: Rate of progress for the inclined plane model

comparison to be poor for making exact quantitative predictions, it nevertheless correctly reproduces the qualitative relation between the rate of progress and the number of descendants in a generation. The probability distributions $w(s')$ are illustrated in Figure 5.7 for five different values of $\lambda \in \{1, 3, 10, 30, 100\}$, according to Equation (5.16).

For the inclined plane model the question of an optimal step length does not arise. The rate of progress increases linearly with the step length. Another question that does arise, however, is how to choose the optimal number of offspring per parent in a generation. The immediate answer is: the bigger λ is, the faster the evolution advances. But in nature, since resources are limited (territory, food, etc.) it is not possible to increase the number of descendants arbitrarily. Likewise in applications of the strategy to solving problems on the digital computer, the requirements for computation time impose limits. The computers in common use today can only work in a serial rather than parallel way. Thus all the mutations must be produced one after the other, and the more descendants the longer the computation time. We should therefore turn our attention instead to finding the optimum value of φ/λ. In the case where the parent survives if it is not bettered by any descendant, we have the trivial solution

$$\lambda_{opt} = 1$$

The corresponding value for the $(1\,,\lambda)$ strategy is, however, larger. With Equation (5.17)

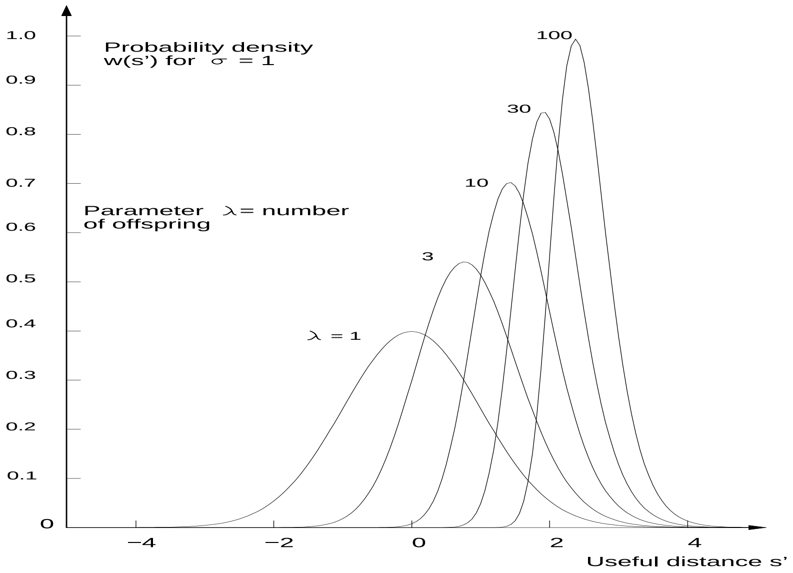

Figure 5.7: Probability distribution $w(s')$

one obtains from the requirement

$$\frac{\partial}{\partial\lambda}\left(\frac{\tilde{\varphi}}{\lambda}\right)\Bigg|_{\lambda=\lambda_{opt}} \stackrel{!}{=} 0$$

the relation

$$\lambda_{opt} = \tilde{\varphi}\;\frac{\partial\lambda}{\partial\tilde{\varphi}}\Bigg|_{\lambda=\lambda_{opt}} = \frac{\sigma^2}{\tilde{\varphi}^2}$$

and, by substituting it back in Equation (5.17), the result

$$\lambda_{opt} = 1 + \sqrt{\frac{\pi}{2\,\lambda_{opt}}}\;\exp\left(\frac{1}{2\,\lambda_{opt}}\right)\left[1 + \;\mathrm{erf}\left(\frac{1}{\sqrt{2\,\lambda_{opt}}}\right)\right]$$

The value obtained iteratively is

$$\lambda_{opt} \;\simeq\; 2.5 \quad (\text{as an integer: } \lambda_{opt} = 2 \text{ or } 3)$$

5.2.2.2 The Sphere Model

We will now try to calculate the rate of progress for the simple spherically symmetrical model, which is of importance for considering the convergence rate properties of the strategy. The contours of the objective function $F(x)$ are concentric hypersphere surfaces, given for example by

$$F(x) = \sum_{i=1}^{n} x_i^2 = const.$$

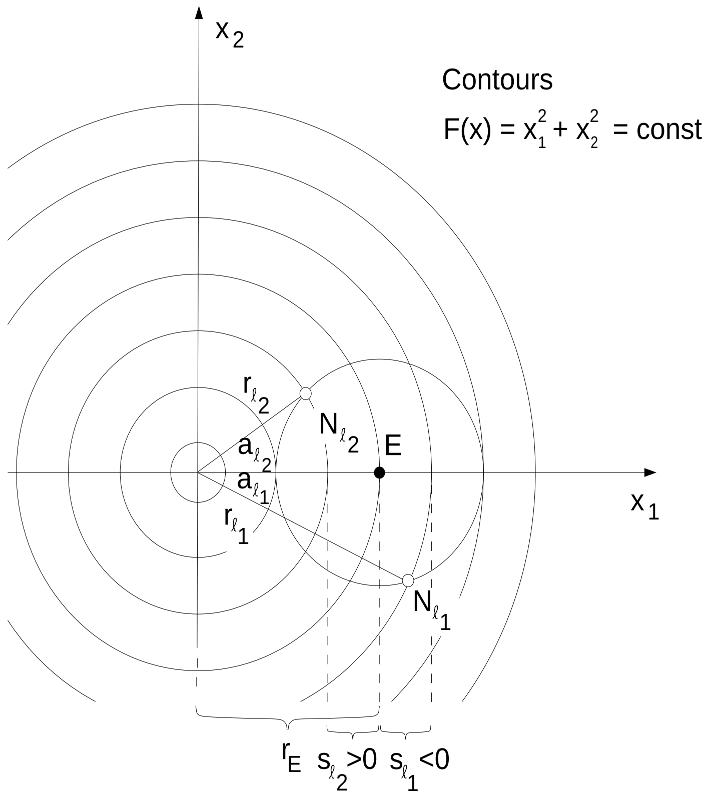

Figure 5.8: Hypersphere model function

Figure 5.8 illustrates the case of two variables. The solution is obtained in much the same way as for the inclined plane. We shall take over some of the steps and subsidiary results from the derivation of Rechenberg (1973) in the two membered evolution strategy.

The normalized probability density for production of a descendant N_ℓ with position vector $x_\ell = (x_{\ell,1}, \ldots, x_{\ell,n})^T$ from parent E with position vector $x_E = (x_{E,1}, \ldots, x_{E,n})^T$ again corresponds to an n-dimensional normal distribution with expectation value $\xi = 0$ and variance σ^2 (the same for all vector components). Without affecting the generality of the result, the special position $x_E = (r_E, 0, \ldots, 0)^T$ can be selected for the starting point E in relation to the coordinate system.

With the notation

$$r_\ell = \sqrt{\sum_{i=1}^{n} x_{\ell,i}^2}$$

Equation (5.9) yields

$$w(E \rightarrow N_\ell) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(-\frac{1}{2\,\sigma^2}(r_\ell^2 + r_E^2 - 2\,r_E\,x_{\ell,1})\right)$$

For the distance covered towards the objective, s_ℓ, the portion is now calculated that contributes to an improvement of the objective function, i.e., in this case the radial difference $s_\ell = r_E - r_\ell$ (see Fig. 5.8). The locus of all points N_ℓ for which s_ℓ is the same is the surface of the n-dimensional hypersphere about the origin with radius $r_\ell = r_E - s_\ell$. Accordingly the total probability density that a mutation (index ℓ) starting from point E will cover the distance s_ℓ is the n-fold line integral:

$$w(s_\ell) = \underset{r_E - r_\ell = s_\ell}{\int \cdots \int} \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(-\frac{1}{2\,\sigma^2}\left(r_\ell^2 + r_E^2 - 2\,r_E\,x_{\ell,1}\right)\right) dx_{\ell,1} \ldots dx_{\ell,n}$$

By transforming to spherical coordinates one obtains a simple integral

$$w(s_\ell) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \frac{\pi^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} \exp\left(-\frac{r_E^2 + r_\ell^2}{2\,\sigma^2}\right) r_\ell^{n-1} \int\limits_{\alpha=0}^{2\pi} \exp\left(\frac{r_E\,r_\ell\,\cos\alpha}{\sigma^2}\right) \sin^{n-2}\alpha\,d\alpha$$

The remaining integral can be expressed as a modified Bessel function:

$$w(s_\ell) = \frac{r_\ell^{\frac{n}{2}}\,r_E^{1-\frac{n}{2}}}{\sigma^2} \exp\left(-\frac{r_E^2 + r_\ell^2}{2\,\sigma^2}\right) I_{\frac{n}{2}-1}\left(\frac{r_E\,r_\ell}{\sigma^2}\right)$$

To simplify the notation we now introduce the following definitions:

$$\nu = \frac{n}{2},\ a = \frac{r_E^2}{\sigma^2},\ v = \frac{r_\ell}{r_E}$$

We thereby obtain

$$w(s_\ell) = \frac{a}{r_E}\ e^{-\frac{a}{2}}\ v^\nu\ e^{-\frac{a\,v^2}{2}}\ I_{\nu-1}(a\,v) \quad \text{with } s_\ell = r_E\,(1-v)$$

In order to use Equation (5.15) to calculate the total probability that the best of λ descendants will cover the distance

$$s' = \max_\ell\{s_\ell \mid \ell = 1(1)\lambda\} = r_E - r'$$

the following quantities are still required:

$$w(s_\ell = s') = \frac{a}{r_E}\ e^{-\frac{a}{2}}\ u^\nu\ e^{-\frac{a\,u^2}{2}}\ I_{\nu-1}(a\,u)$$

with

$$\frac{r'}{r_E} = u \quad \text{and} \quad s' = r_E\,(1-u)$$

and

$$\begin{aligned} p(s_\ell < s') &= 1 - p(s_\ell > s') \\ &= 1 - \int\limits_{s_\ell = r_E}^{s'} w(s_\ell)\,ds_\ell \\ &= 1 - \int\limits_{v=0}^{u} a\,e^{-\frac{a}{2}}\,v^\nu\,e^{-\frac{a\,v^2}{2}}\,I_{\nu-1}(a\,v)\,dv \end{aligned}$$

This finally gives the probability function for the useful distance s' covered in one generation, expressed in units of u:

$$w(s') = \frac{a}{r_E}\, e^{-\frac{a}{2}}\, u^\nu\, e^{-\frac{a\,u^2}{2}}\, I_{\nu-1}(a\,u) \left(1 - a\, e^{-\frac{a}{2}} \int\limits_{v=0}^{u} v^\nu\, e^{-\frac{a\,v^2}{2}}\, I_{\nu-1}(a\,v)\, dv\right)^{\lambda-1}$$

Since the expectation value of this distribution is not readily obtainable, we shall determine its maximum to give an approximation $\tilde{\varphi}$. From the necessary condition

$$\left.\frac{\partial w(s')}{\partial s'}\right|_{s'=\tilde{\varphi}} \overset{!}{=} 0$$

with the more concise notation

$$D(y) = a\;\; e^{-\frac{a}{2}}\;\; y^\nu\;\; e^{-\frac{a\,y^2}{2}}\;\; I_{\nu-1}(a\,y)$$

we obtain the relation

$$\lambda = 1 + \left.\frac{\partial D(u)}{\partial u}\right|_{u=1-\tilde{\varphi}/r_E} [D(1-\tilde{\varphi}/r_E)]^{-2} \left(1 - \int\limits_{v=0}^{1-\tilde{\varphi}/r_E} D(v)\, dv\right) \qquad (5.18)$$

Except for the upper limit of integration, this is the same integral that made it so difficult to obtain the exact solution for the rate of progress in the (1+1) evolution strategy (see Rechenberg, 1973). Under the condition $\nu \gg 1$ and $\nu/a \ll 1$, which means for many variables and at a large enough distance from the optimum, Rechenberg obtained an estimate by expanding Debye's asymptotic series representation of the Bessel function (e.g., Jahnke-Emde-Lösch, 1966) in powers of ν/a. Without giving here the individual steps in the derivation, the result is

$$\int\limits_{v=0}^{1} D(v)\, dv \simeq \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\nu}{\sqrt{2\,a}}\right)\right] + \frac{\sqrt{a}}{2\,\nu\,\sqrt{2\,\pi}}\left[\exp\left(-\frac{(\nu-1)^2}{2\,a}\right) - \exp\left(-\frac{\nu^2}{2\,a}\right)\right] \quad (5.19)$$

It is clear from Equation (5.4) that the rate of progress of the (1+1) strategy for the two membered evolution varies inversely as the number of variables. Even if a higher convergence rate is expected from the multimembered scheme, with many descendants per parent, there will be no change in the relation to n, the number of parameters. In addition to the assumptions already made regarding ν and ν/a, without further risk to the validity of the approximate theory we can assume that $1 - \tilde{\varphi}/r_E \simeq 1$. Equation (5.19) can now also be applied here.

For the partial differential

$$\left.\frac{\partial D(u)}{\partial u}\right|_{u=1-\tilde{\varphi}/r_E}$$

we obtain with the use of the Debye series again:

$$\left.\frac{\partial D(u)}{\partial u}\right|_{u=1-\tilde{\varphi}/r_E} = D(1-\tilde{\varphi}/r_E)\left[a\,\exp\left(\frac{\nu}{a\,(1-\tilde{\varphi}/r_E)}\right) + \frac{1}{1-\tilde{\varphi}/r_E} - a\,(1-\tilde{\varphi}/r_E)\right]$$

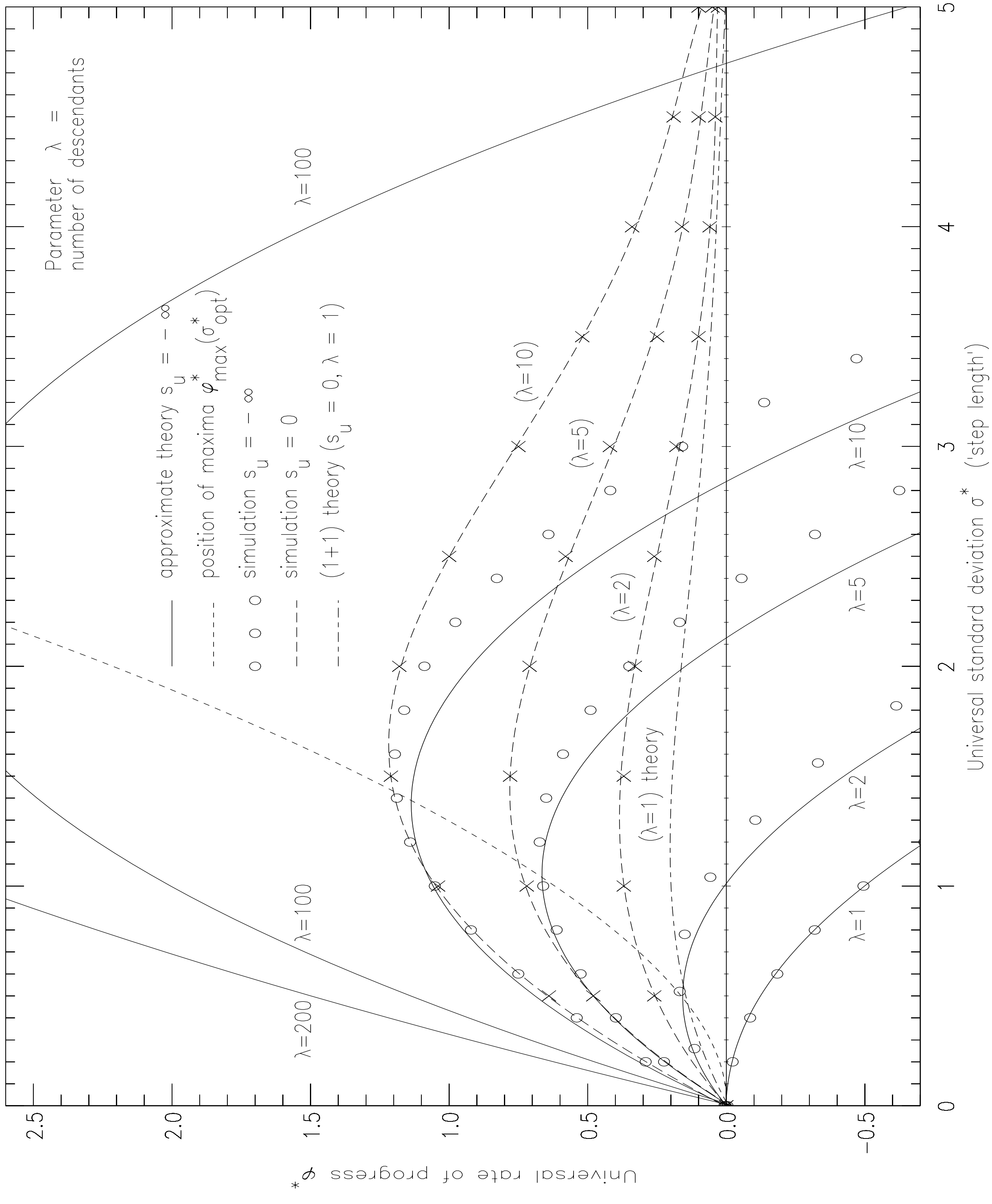

Figure 5.9: Rate of progress for the sphere model

If the result is substituted into Equation (5.18) a longer expression is obtained, of the form:

$$\lambda = \lambda(\tilde{\varphi}, \sigma, r_E, n)$$

In the expectation of an end result similar to Equation (5.4) and since a particular starting point r_E is of no interest, we will introduce new variables:

$$\varphi^* = \frac{\tilde{\varphi}\, n}{r_E} \quad \text{and} \quad \sigma^* = \frac{\sigma\, n}{r_E}$$

If $\tilde{\varphi}$ and σ are now replaced by $\varphi *$ and σ^*, taking the limit

$$\lim_{n \to \infty} \lambda(\varphi^*, \sigma^*, r_E, n)$$

we find that the quantities n and r_E disappear from the parameter list of λ. φ^* and σ^* can therefore be regarded as "universal" variables. We obtain

$$\lambda = \lambda(\varphi^*, \sigma^*) = 1 + \sqrt{\pi}\left(\frac{\varphi *}{\sqrt{2}\,\sigma^*} + \frac{\sigma^*}{\sqrt{8}}\right) \exp\left[\left(\frac{\varphi *}{\sqrt{2}\,\sigma^*} + \frac{\sigma^*}{\sqrt{8}}\right)^2\right]\left[1 + \operatorname{erf}\left(\frac{\sigma^*}{\sqrt{8}}\right)\right] \qquad (5.20)$$

As in the case of the inclined plane considered previously, this equation cannot be simply solved for φ^*. Figure 5.9 shows the family of curves $\varphi^* = \varphi^*(\sigma^*, \lambda)$.

For $\sigma^* \to 0$, as expected, $\varphi^* \to 0$. For $\lambda = 1$, the rate of progress is always negative. Since the parent in the $(1\,,\lambda)$ strategy is not included in the selection after it has served to produce a descendant, $\lambda = 1$ means that every mutation is adopted, whether better or worse. For the sphere model, except for $\sigma^* = 0$, the region of success is always smaller than half of the variable space. With increasing σ^*, the ratio becomes even worse; φ^* is thus always ≤ 0, and more strongly negative the greater is σ^*.

For $\lambda \geq 2$ the rate of progress increases at first as a function of the variance, reaches a maximum, and then decreases continuously until it becomes negative. From this behavior one can see even more clearly than in the (1+1) strategy how important the correct choice of variance is for the optimization.

In the $(1\,,\lambda)$ strategy, the progress can turn retrograde if all the offspring are worse than the parent that produced them. Only with an immortal parent having an infinite capacity for reproduction would progress be guaranteed or, at least, would retrogression be ruled out. We shall see later why the model with "extinction" is nevertheless advantageous. Except for small values of λ, the maximum rate of progress is almost the same in the "survival" and "extinction" cases. So if the optimal variance can be maintained throughout, leaving the parents out of the selection is not a disadvantage.

The position of the maxima of φ^* with respect to σ^* at a constant λ is obtained by simple differentiation and equating the partial derivative to zero. Defining

$$\frac{\sigma^*_{opt}}{\sqrt{8}} = \sigma^+ \text{ and } \frac{\varphi^*_{max}}{\sqrt{2}\,\sigma^*_{opt}} = \varphi^+$$

the equation is

$$\sigma^+(\varphi^+ + \sigma^+)\exp(-\sigma^{+2}) + \sqrt{\pi}(\sigma^+ - \varphi^+)\left(\frac{1}{2} + (\varphi^+ + \sigma^+)^2\right)\left(1 + \operatorname{erf}(\sigma^+)\right) \overset{!}{=} 0 \qquad (5.21)$$

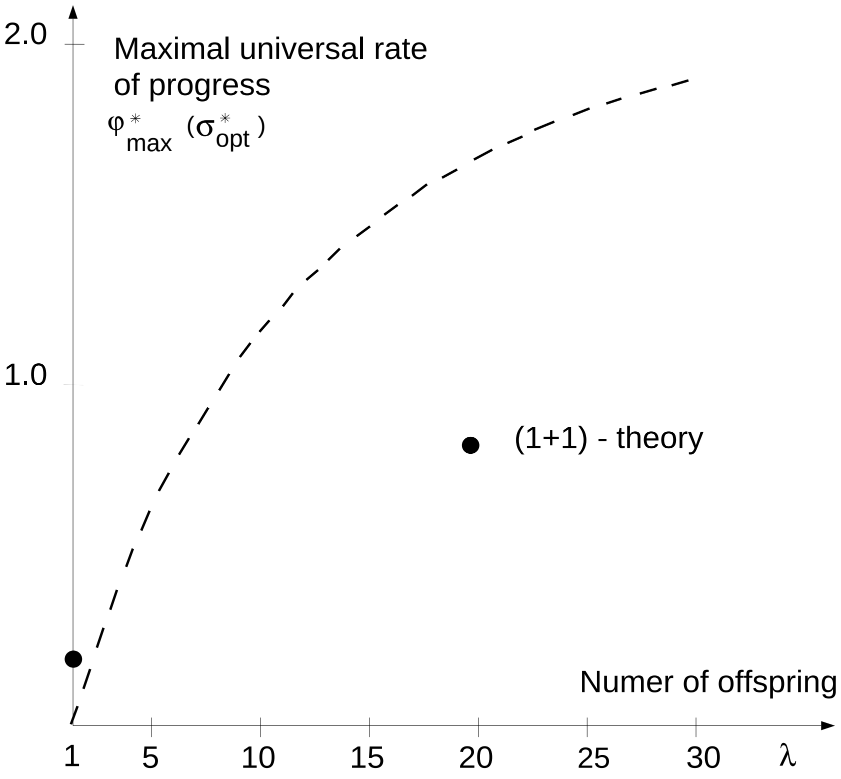

Figure 5.10: Maximal rate of progress for the sphere model

Points on the curve $\varphi^*_{max} = \varphi^*(\sigma^* = \sigma^*_{opt})$ can only be obtained iteratively. To express $\lambda = \lambda(\varphi^*_{max})$, the non-linear system of equations consisting of Equations (5.20) and (5.21) must be solved. The results as obtained with the multimembered evolution strategy are shown in Figure 5.10. A convenient formula can only be obtained by assuming

$$\varphi^+ \simeq \sigma^+, \text{ i.e., } 2\,\varphi^*_{max} \simeq \sigma^{*2}_{opt}$$

This estimate is actually not far wrong, since the second term in Equation (5.21) goes to zero. We thus find

$$\lambda \simeq 1 + \sqrt{\pi\,\varphi^*_{max}}\,\exp(\varphi^*_{max})\left[1 + \mathrm{erf}\left(\frac{1}{2}\sqrt{\varphi^*_{max}}\right)\right] \tag{5.22}$$

a relation with comparable structure to the result for the inclined plane.

Finally we ask whether φ^*_{max}/λ has a maximum, as in the inclined plane case. If the parent can survive the offspring, $\lambda_{opt} = 1$ here too; if not the condition

$$\lambda_{opt} = 2\sqrt{\pi}\left[\frac{1}{2} + (\varphi^+ + \sigma^+)^2\right]\exp[(\varphi^+ + \sigma^+)^2]\,[1 + \mathrm{erf}(\sigma^+)]\,\varphi^+ \tag{5.23}$$

must be added to Equations (5.20) and (5.21). The solution, obtained iteratively, is:

$$\lambda_{opt} \simeq 4.7 \quad (\text{as an integer: } \lambda_{opt} = 5)$$

Both the $(1\,,\lambda)$ and $(1{+}\lambda)$ schemes were run on the computer for the sphere model, with $n = 100, r_E = 100$, and variable σ. In each case φ was evaluated over $10,000$ generations. The resulting data are shown in terms of φ^* and σ^* in Figure 5.9. In comparison with the approximate theory, deviations are apparent mainly for $\sigma^* > \sigma^*_{opt}$. The skewness of the probability distribution $w(s')$ and the error in the estimate of the integral $\int D(y)\,dy$ have only a weak effect in the region of greatest interest, where the rate of progress is maximum. Furthermore, the results of the simulation fall closer to the approximate theory if n is taken to be greater than 100; however, the computation time then becomes excessive. For large values of λ the possible survival of the parent only becomes noticeable when the variance is too large to allow rapid convergence. The greatest differences, as expected, appear for $\lambda = 1$.

On the whole we see that the theory worked out here gives at least a qualitative account of the behavior of the $(1\,,\lambda)$ strategy. A much more elegant method yielding an even better approximation may be found in Bäck, Rudolph, and Schwefel (1993), or Beyer (1993, 1994a,b).

5.2.2.3 The Corridor Model

As a third and last model objective function, we will now consider the right-angled corridor. The contours of $F(x)$ in the two dimensional picture (Fig. 5.11) are straight and parallel, but not necessarily equidistant.

$$F(x) = c_0 + \sum_{i=1}^{n} c_i\, x_i$$

For the sake of simplifying the calculation we will again give the coordinate system a particular position and orientation with $c_1 = -1$, $c_i = 0$ for all $i = 2, 3, \ldots, n$. The right-angled corridor (Problem 2.37, see Appendix A, Sect. A.2)–we are using here three dimensional concepts for the essentially n-dimensional case–is defined by constraints of the form

$$G_j(x) = |x_j| \leq b, \qquad \text{for } j = 2(1)n$$

It has the width $2\,b$ for all coordinate directions x_i, $i = 2(1)n$; hence the cross section $(2\,b)^{n-1}$. As a starting point, the position x_E of the parent E, we choose the origin with respect to $x_1 = 0$. The useful part of a random step is just its component z_1 in the x_1 direction, which is the negative gradient direction. The formulae for $w(s_\ell = s')$ and $p(s_\ell < s')$ derived previously for the inclined plane also apply here. We cannot, however, insert them immediately into Equation (5.15); first we must pay attention to the rule that mutants that violate one or more of the constraints are not accepted.

For a given mutation, the probability of not jumping through the corridor wall associated with the variable x_i, $i = 2(1)n$, is

$$p(|x_{\ell,i}| \leq b) = \int\limits_{x_{\ell,i}=-b}^{b} \frac{1}{\sqrt{2\pi}\,\sigma}\, \exp\left[-\frac{(x_{E,i} - x_{\ell,i})^2}{2\,\sigma^2}\right] dx_{\ell,i}$$

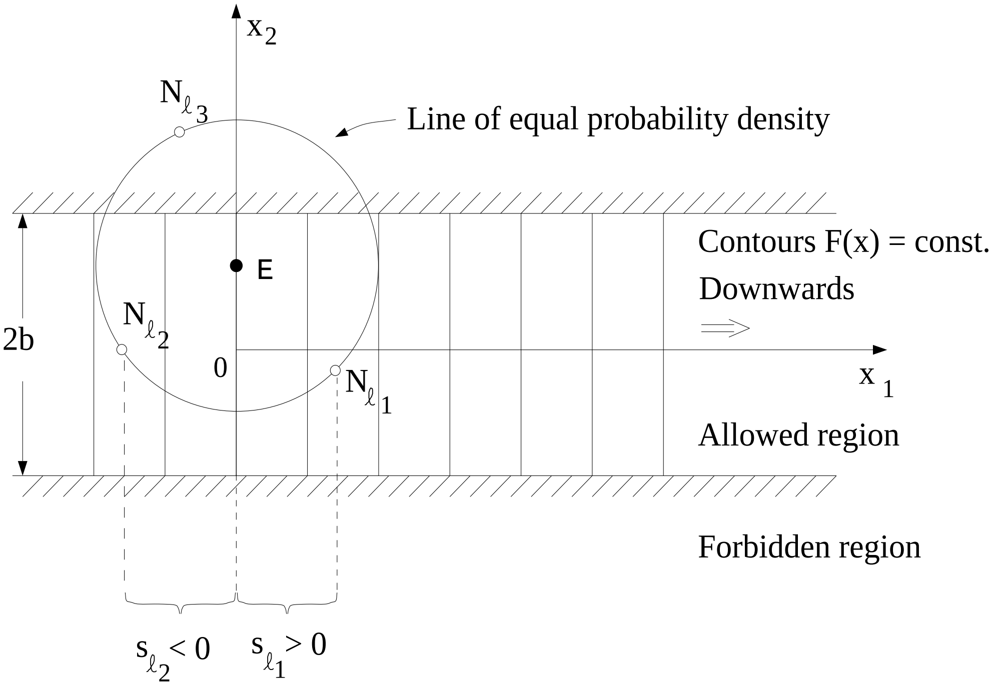

Figure 5.11: Corridor model function

$$= \frac{1}{2}\left[\,\mathrm{erf}\left(\frac{b - x_{E,i}}{\sqrt{2}\,\sigma}\right) + \ \mathrm{erf}\left(\frac{b + x_{E,i}}{\sqrt{2}\,\sigma}\right)\right]$$

That is, the probability depends on the current position $x_{E,i}$ of the starting point E. We can only construct an average value for all possible situations if we know the probability p_a of certain situations occurring. It could well be that, during the minimum search, positions near the border are occupied less often than others. The same problem of finding the occupation probability p_a has arisen already in the theoretical treatment of the (1+1) strategy. Rechenberg (1973) discovered that

$$p_a = \frac{1}{2\,b} \text{ (with respect to one of the variables } x_i, \ i \ = \ 2(1)n)$$

which is a constant independent of the current values of the variables. We will assume that this also holds here. Thus the average probability that one of the $n-1$ constrained variables will remain within the corridor can be given as:

$$\tilde{p}(|x_{\ell,i}| \le b) = \int\limits_{x_{E,i}=-b}^{b} p_a\, p(|x_{\ell,i}| \le b)\, dx_{E,i}$$

$$= \frac{1}{4\,b} \int\limits_{x_{E,i}=-b}^{b} \left[\,\mathrm{erf}\left(\frac{b - x_{E,i}}{\sqrt{2}\,\sigma}\right) + \ \mathrm{erf}\left(\frac{b + x_{E,i}}{\sqrt{2}\,\sigma}\right)\right] dx_{E,i}$$

Making use of the relation (see Ryshik and Gradstein, 1963)

$$\int\limits_{y=0}^{p} \operatorname{erf}(\alpha\, y)\, dy = p\ \operatorname{erf}(\alpha\, p) + \frac{\exp(-\alpha^2\, p^2) - 1}{\sqrt{\pi}\, \alpha}$$

one finally obtains

$$\tilde{p}(|x_{\ell,i}| \leq b) = \ \operatorname{erf}\left(\frac{\sqrt{2}\, b}{\sigma}\right) + \frac{1}{\pi}\frac{\sigma}{\sqrt{2}\, b}\left[\exp\left(-\frac{2\, b^2}{\sigma^2}\right) - 1\right] \tag{5.24}$$

In the following we refer to this expression as item v.

$$v = \tilde{p}(|x_{\ell,i}| \leq b)$$

With the above definition of v, the total probability that a descendant N_ℓ is feasible, i.e., that it satisfies all the constraints, is

$$\begin{aligned} p_{feas} &= \prod_{i=2}^{n} \tilde{p}(|x_{\ell,i}| \leq b) \\ &= v^{n-1} \end{aligned}$$

and the probability that N_ℓ is lethal is

$$p_{leth} = 1 - p_{feas} = 1 - v^{n-1}$$

Only non-lethal mutants come into consideration as parents of the next generation. Hence, instead of $w(s_\ell\ =\ s')$ we must insert into Equation (5.15) the expression

$$w(s_\ell = s')\, p_{feas} = \frac{1}{\sqrt{2\pi}\, \sigma}\ \exp\left(-\frac{s'^2}{2\, \sigma^2}\right)\ v^{n-1}$$

and instead of $p(s_\ell\ <\ s')$ we should take

$$p(s_\ell < s')\, p_{feas} + p_{leth} = \frac{1}{2}\left[1 +\ \operatorname{erf}\left(\frac{s'}{\sqrt{2}\, \sigma}\right)\right]\ v^{n-1} + 1 - v^{n-1}$$

The first term expresses the probability that the descendant N_ℓ both falls within the allowed region and progresses by s'; the second term represents the probability that a descendant N_ℓ is either non-lethal and advanced by $s_\ell\ <\ s'$, or lethal. If we now insert both these quantities into Equation (5.15) we obtain

$$\tilde{w}(s') = \frac{\lambda\, v^{n-1}}{\sqrt{2\pi}\, \sigma\, 2^{\lambda-1}}\ \exp\left(-\frac{s'^2}{2\, \sigma^2}\right)\left(\left[1 +\ \operatorname{erf}\left(\frac{s'}{\sqrt{2}\, \sigma}\right)\right]\ v^{n-1} + 2\,(1 - v^{n-1})\right)^{\lambda-1} \tag{5.25}$$

where v is given by Equation (5.24).

So far we have not considered the special case of all the descendants being lethal mutants. If we were to abide by the rules of the $(1\,,\lambda)$ strategy as followed up to now, the outcome would be extinction of the population and the rate of progress would no longer be defined. The probability of extinction of the population is given by the product of the lethal probabilities:

$$p_{stop} = (1 - v^{n-1})^{\lambda}$$

To be able to optimize further in such situations let us adopt the following procedure: If all the mutations lead to forbidden points, the parent will survive and produce another generation of descendants. Thus for this generation the rate of progress takes the value zero. Equation (5.25) then only holds for $s' \neq 0$ and we must reformulate the probability of advancing by s' in one generation as follows:

$$w(s') = \tilde{w}(s') + \delta\, p_{stop}$$

where

$$\delta = \begin{cases} 0\,, & \text{if } s' \neq 0 \\ 1\,, & \text{if } s' = 0 \end{cases}$$

The distribution $w(s')$ is no longer continuous, and even if $w'(s')$ is symmetric we cannot assume that the maximum of the distribution is a useful approximation to the average rate of progress (Fig. 5.12). The following condition must be satisfied:

$$\int\limits_{s'=-\infty}^{\infty} w(s')\, ds' = \int\limits_{s'=-\infty}^{\infty} \tilde{w}(s')\, ds' + w_{stop} = 1 \qquad (5.26)$$

We can think of $w(s')$ as a superposition of two density distributions, with conditional mathematical expectation values

$$\varphi_1 = \int\limits_{s'=-\infty}^{\infty} s'\, \tilde{w}(s')\, ds'$$

and

$$\varphi_2 = 0$$

and with associated frequencies

$$p_1 = \int\limits_{s'=-\infty}^{\infty} \tilde{w}(s')\, ds' = 1 - p_{stop}$$

and

$$p_2 = p_{stop}$$

The events belonging to the two density distributions are mutually exclusive and by virtue of Equation (5.26) they make up together a complete set of events. The expectation value is then given by (e.g., Gnedenko, 1970; Sweschnikow , 1970).

$$\varphi = \int\limits_{s'=-\infty}^{\infty} s'\, w(s')\, ds' = \varphi_1\, p_1 + \varphi_2\, p_2 = \varphi_1\,(1 - p_{stop})$$

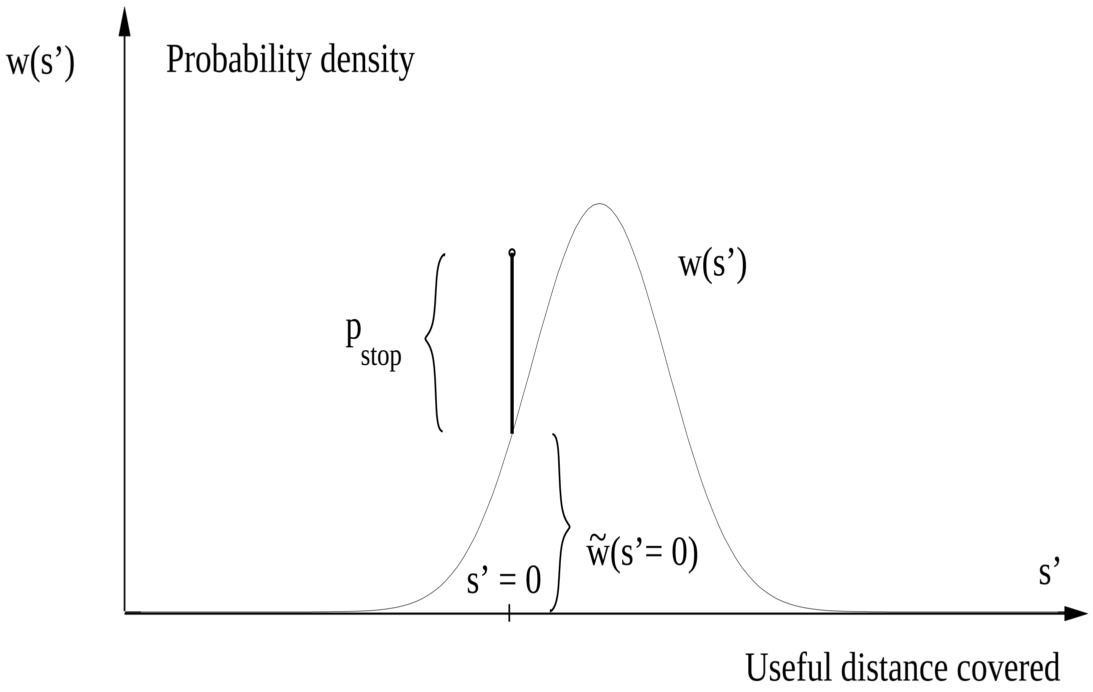

Figure 5.12: Estimation of the rate of progress from the probability density for the corridor model

Since we are unable to calculate φ_1 directly, we make an approximation:

$$\tilde{\varphi} = \hat{\varphi}\,(1 - p_{stop}) = \hat{\varphi}[1 - (1 - v^{n-1})^{\lambda}] \qquad (5.27)$$

taking for $\hat{\varphi}$ the position of the maximum of $\tilde{w}(s')$.
We require

$$\left.\frac{\partial \tilde{w}(s')}{\partial s'}\right|_{s'=\hat{\varphi}} \overset{!}{=} 0$$

By differentiating Equation (5.25) and setting the first derivative to zero:

$$\lambda = 1 + \sqrt{\pi}\frac{\hat{\varphi}}{\sqrt{2}\,\sigma}\,\exp\left(\frac{\hat{\varphi}^2}{2\,\sigma^2}\right)\left[1 + \ \mathrm{erf}\left(\frac{\hat{\varphi}}{\sqrt{2}\,\sigma}\right) + 2\,(v^{1-n} - 1)\right] \qquad (5.28)$$

Apart from an extra term, this formula is similar to the relation $\lambda = \lambda(\tilde{\varphi}, \sigma)$ found for the inclined plane (Equation (5.17)). The main difference here, however, is that in place of $\tilde{\varphi}, \hat{\varphi}$ appears, as defined by Equation (5.27).

As in the case of the sphere model, we will introduce here "universal parameters"

$$\varphi^* = \frac{\tilde{\varphi}\,n}{b} \quad \text{and} \quad \sigma^* = \frac{\sigma\,n}{b}$$

and take the limit $n \rightarrow \infty$ in order to arrive at a practically useful relation $\lambda = \lambda(\varphi^*, \sigma^*)$.

With the new quantities φ^* and σ^*, Equation (5.24) for v becomes

$$v = \operatorname{erf}\left(\frac{\sqrt{2}\,n}{\sigma^*}\right) - \frac{\sigma^*}{\sqrt{2\pi}\,n}\left[1 - \exp\left(-\frac{2\,n^2}{\sigma^{*2}}\right)\right]$$

Since the argument of the error function increases as n, the number of variables, the approximation

$$\operatorname{erf}(y) \simeq 1 - \frac{1}{\sqrt{\pi}\,y}\,\exp\left(-y^2\right)$$

can be used to give

$$v = 1 - \frac{\sigma^*}{n\,\sqrt{2\pi}} \qquad \text{for } n \gg 1$$

and with

$$\lim_{n\to\infty}\left(1 + \frac{1}{n}\right)^n = e$$

finally

$$v^{1-n} = \exp\left(\frac{\sigma^*}{\sqrt{2\pi}}\right)$$

The desired relation $\lambda = \lambda(\varphi^*, \sigma^*)$ is thus

$$\lambda = 1 + \frac{\sqrt{\pi}\,\tilde{\varphi}^*}{\sqrt{2}\,\sigma^*}\,\exp\left[\left(\frac{\tilde{\varphi}^*}{\sqrt{2}\,\sigma^*}\right)^2\right]\left[\operatorname{erf}\left(\frac{\tilde{\varphi}^*}{\sqrt{2}\,\sigma^*}\right) + 2\,\exp\left(\frac{\sigma^*}{\sqrt{2\pi}}\right) - 1\right] \tag{5.29}$$

in which, from Equation (5.27),

$$\tilde{\varphi}^* = \frac{\varphi^*}{1 - \left[1 - \exp\left(-\frac{\sigma^*}{\sqrt{2\pi}}\right)\right]^\lambda}$$

Pairs of values obtained iteratively are shown in Figure 5.13 together with simulation results for the cases of "survival" and "extinction" of the parent ($n = 100$, $b = 100$, average over $10{,}000$ successful generations).

As in the case of the sphere model, the deviations can be attributed to the simplifying assumptions made in deriving the approximate theory. For $\lambda = 1$, φ^* is always zero if the parent is not included in the selection. The transition to the inclined plane model is correctly reproduced in this respect. Negative rates of progress cannot occur.

The position of the maxima $\varphi^*_{max} = \varphi^*(\sigma^* = \sigma^*_{opt})$ at constant λ are obtained in the same way as for the sphere model. The condition to be added to Equation (5.29) is

$$\left(\frac{\lambda}{c}\,\exp\left(-\sigma^+\right)\left[1 - \exp\left(-\sigma^+\right)\right]^{\lambda-1} - \frac{1}{\sigma^+}\right) \times$$

$$\times\left(\left[\,\operatorname{erf}(\varphi^+) + 2\,\exp\left(\sigma^+\right) - 1\right]\left[1 + 2\,\varphi^{+2}\right] + \frac{2}{\sqrt{\pi}}\,\varphi^+\,\exp\left(-\varphi^{+2}\right)\right) + 2\,\exp\left(\sigma^+\right) \overset{!}{=} 0 \tag{5.30}$$

in which the following new quantities are introduced again for compactness:

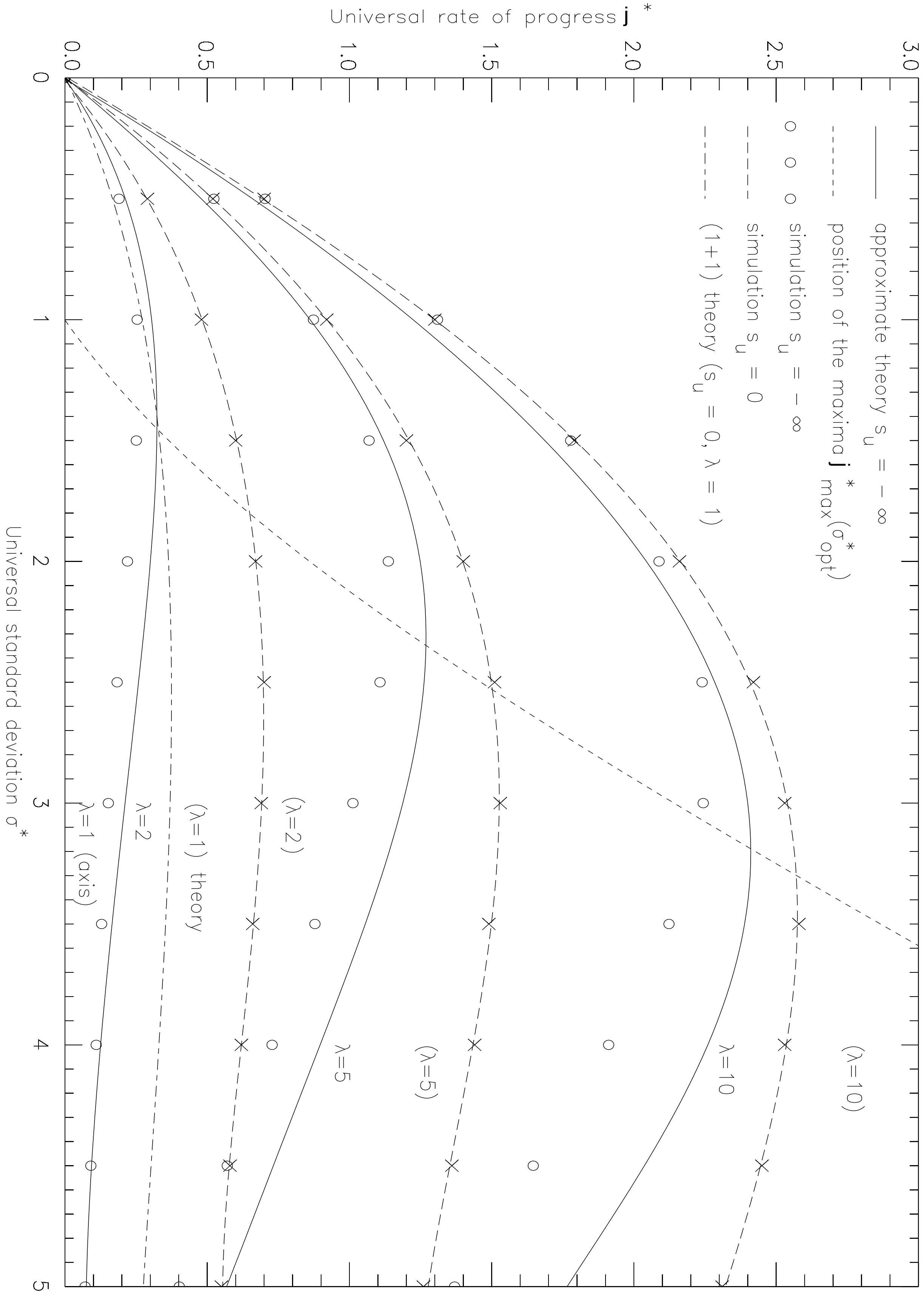

Figure 5.13: Rate of progress for the corridor model

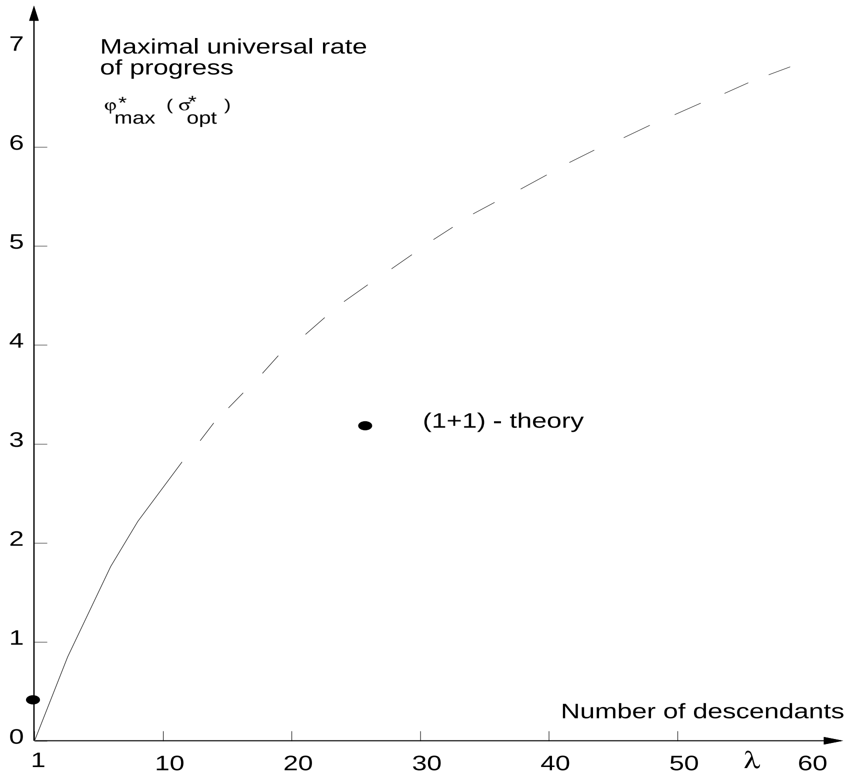

Figure 5.14: Maximal rate of progress for the corridor model

$$\begin{aligned}
\sigma^+ &= \frac{\sigma^*_{opt}}{\sqrt{2\pi}} \\
\varphi^+ &= \frac{\varphi^*_{max}}{\sqrt{2}\,\sigma^*_{opt}\,c} \\
c &= 1 - \left[1 - \exp\left(-\frac{\sigma^*_{opt}}{\sqrt{2\pi}}\right)\right]^\lambda
\end{aligned}$$

Pairs of values found by iteration are shown in Figure 5.13. Figure 5.14 shows φ^*_{max} versus λ. To determine λ_{opt} for the $(1\,,\lambda)$ strategy, i.e., the value of λ for which φ^*_{max}/λ is a maximum, it is necessary to solve the system of three non-linear equations, comprising Equation (5.29), Equation (5.30), and

$$\begin{aligned}
\lambda_{opt} &= \varphi^+ \left\{\sqrt{\pi}\,\exp\,(\varphi^{+2})[\,\mathrm{erf}\,(\varphi^+) + 2\,\exp\,(\sigma^+) - 1]\,[1 + 2\varphi^{+2}] + 2\varphi^+\right\} \times \\
&\times \left\{\frac{\lambda_{opt}}{c}\,[1 - \exp\,(-\sigma^+)]^\lambda \ln[1 - \exp(-\sigma^+)] + 1\right\} \qquad (5.31)
\end{aligned}$$

The result is

$$\lambda_{opt} \simeq 6.0 \quad (\text{as an integer: } \lambda_{opt} = 6)$$

5.2.3 The Step Length Control

How should one proceed in order to still achieve the maximum rate of progress, i.e., to maintain the optimum variances σ_i^2, $i = 1(1)n$, for the case of the multimembered evolution scheme? For the (1+1) strategy this aim was met by the 1/5 success rule, which was based on the probability of success at maximum convergence rate of the *sphere* and *corridor model* functions. Such control from outside the actual mutation-selection game does not correspond to the biological paradigm. It should rather be assumed that the step lengths, or more precisely the variances, have adapted and are still adapting to circumstances arising in the course of natural evolution. Although the environmentally induced rate of mutation cannot be interfered with directly, the existence of mutator genes and repair enzymes strongly suggests that the consequences of such environmental influences are always reduced to the appropriate level. In the multimembered evolution strategy the fact that the observed rates of mutation are also small, indeed that they must be small to be optimal, comes out of the universal rate of progress and standard deviation introduced above, which require σ to be inversely proportional to the number of variables, as in the (1+1) strategy.

If we wish to imitate organic evolution, we can proceed as follows. Besides the variables $x_{E,i}$, $i = 1(1)n$, a set of parameters $\sigma_{E,i}$, $i = 1(1)n$, is assigned to a parent E. These describe the variances of the random changes. Each descendant N_ℓ of the parent E should differ from it both in $x_{\ell,i}$ and $\sigma_{\ell,i}$. The changes in the variances should also be random and small, and the most probable case should be that there is no change at all. Whether a descendant can become a parent of the next generation depends on its vitality, thus only on its $x_{\ell,i}$. Which values of the variables it represents depends, however, not only on the $x_{E,i}$ of the parent, but also on the standard deviations $\sigma_{\ell,i}$, which affect the size of the changes $z_i = x_{\ell,i} - x_{E,i}$. In this way the "step lengths" also play an indirect rôle in the selection mechanism.

The highest possible probability that a descendant is better than the parent is normally

$$w_{e_{max}} = 0.5$$

It is attained in the inclined plane case, for example, and for other model functions in the limit of infinitely small step lengths. In order to prevent that a reduction of the σ_i always gives rise to a selection advantage, λ must be at least ≥ 2. But the optimal step lengths can only take effect if

$$\lambda > \frac{1}{w_{e_{opt}}}$$

This means that on average at least one descendant represents an improvement of the value of the objective function. The number of descendants per parent thus plays a decisive rôle in the multimembered scheme, just as does the check on the success ratio in the two membered evolution scheme. For comparison let us tabulate here the λ_{opt} of the $(1\,,\lambda)$ strategy and $w_{e_{opt}}$ of the (1+1) strategy for the three model functions considered. The values of $w_{e_{opt}}$ are taken from the work of Rechenberg (1973).

Model function	$w_{e_{opt}}$	$\frac{1}{w_{e_{opt}}}$	λ_{opt}
Inclined plane	$\frac{1}{2}$	2	2.5
Sphere	0.27	3.7	4.7
Corridor	$\frac{1}{2e}$	5.4	6.0

How should the step lengths now be altered? We shall first consider only a single variance σ^2 for changes in all the variables. In the production of the random changes, the standard deviation σ is always a positive factor. It is therefore reasonable to generate new step lengths from the old by a multiplicative rather than additive process, according to the scheme

$$\sigma_N^{(g)} = \sigma_E^{(g)} \, \bar{Z}^{(g)} \tag{5.32}$$

The median $\bar{\xi}$ of the random distribution for the quantity $\bar{Z}$ must equal one to satisfy the condition that there is no deterministic drift without selection. Furthermore an increase of the step length should occur with the same frequency as a decrease; more precisely, the probability of occurrence of a particular random value must be the same as that of its reciprocal. The third requirement is that small changes should occur more often than large ones. All three requirements are satisfied by the log-normal distribution. Random quantities obeying this distribution are obtained from $(0,\ \tau^2)$ normally distributed numbers Y by the process

$$\bar{Z} = e^Y \tag{5.33}$$

The probability distribution for $\bar{Z}$ is then

$$w(\bar{z}) = \frac{1}{\sqrt{2\pi}\,\tau} \, \frac{1}{\bar{z}} \, \exp\left(-\frac{(\ln \bar{z})^2}{2\,\tau^2}\right)$$

The next question concerns the choice of τ, and we shall answer it, in the same way as for the (1+1) strategy, with reference to the rate of change of step lengths that maintains the maximum rate of progress in the sphere model. Regarding φ as a differential quotient $-dr/dg$ leads to the relation (see Sect. 5.1.2)

$$\frac{\sigma_{opt}^{(g+1)}}{\sigma_{opt}^{(g)}} = \exp\left(-\frac{\varphi_{max}^*}{n}\right) \tag{5.34}$$

for the optimal step lengths of two consecutive generations, where φ_{max}^* now has a different, larger value that depends on λ and μ. The actual size of the average changes in the variances, using the proposed mutation scheme based on Equations (5.32) and (5.33), depends on the topology of the objective function and the number of parents and descendants. If n, the number of variables, is large, the optimal variance will only change slightly from generation to generation. We will therefore assume that the selection in any generation is more or less indifferent to reductions and increases in the step length. We thereby obtain the multiplicative change in the random quantity X, averaged over n generations:

$$X = \left(\prod_{g=1}^{n} \bar{Z}^{(g)}\right)^{\frac{1}{n}} = \exp\left(\frac{1}{n}\sum_{g=1}^{n} Y^{(g)}\right)$$

Since the $Y^{(g)}$ are all $(0,\, \tau^2)$ normally distributed, it follows from the addition theorem of the normal distribution (Heinhold and Gaede, 1972) that

$$\frac{1}{n} \sum_{g=1}^{n} Y^{(g)}$$

is a $(0,\, \tau^2/n)$ normally distributed random quantity. Accordingly, the two quantities $\exp(\pm\tau/\sqrt{n})$ are characteristic of the average changes (minus sign for reduction) in the step lengths per generation. The median of $w(\bar{z})$ is of course just $e^0 = 1$. Together with Equation (5.34), our observation leads us to the requirement

$$\exp\left(\frac{\varphi^*_{max}}{n}\right) \simeq \exp\left(\frac{\tau}{\sqrt{n}}\right)$$

or

$$\tau \simeq \frac{\varphi^*_{max}}{\sqrt{n}} \tag{5.35}$$

The variance τ^2 of the normally distributed random numbers Y, from which the log-normally distributed random multipliers for the standard deviations ("step sizes") of the changes in the object variables are produced, thus must vary inversely as the number of variables. Its actual value should depend on the expected rate of convergence φ^* and hence on the choice of the number of descendants λ.

Instead of only one common strategy parameter σ, each individual can now have a complete set of n different σ_i, $i = 1(1)n$, for every alteration in the corresponding n object variables x_i, $i = 1(1)n$. The two following schemes can be envisioned:

$$\sigma^{(g)}_{N,i} = \sigma^{(g)}_{E,i}\, \bar{Z}^{(g)}_i \tag{5.36}$$

or

$$\sigma^{(g)}_{N,i} = \sigma^{(g)}_{E,i}\, \bar{Z}^{(g)}_i\, \bar{Z}^{(g)}_0 \tag{5.37}$$

But only the second one should be taken into further consideration, because otherwise in the case of $n \gg 1$ the average overall step size of the offspring

$$s_N = \sqrt{\sum_{i=1}^{n} \sigma^2_{N,i}}$$

could not be substantially different from that of its parent

$$s_E = \sqrt{\sum_{i=1}^{n} \sigma^2_{E,i}}$$

due to the levelling effect of the many random multiplication events (law of large number of events). In order to split the mutation effects to the overall step size and the individual step sizes one could choose

$$\tau_0 \simeq \frac{\varphi^*}{\sqrt{2\,n}}, \qquad \text{for } \bar{Z}_0 \tag{5.38}$$

$$\tau \simeq \frac{\varphi^*}{\sqrt{2\sqrt{n}}}, \qquad \text{for all } \bar{Z}_i,\ i = 1(1)n \tag{5.39}$$

We shall not go into further details since another kind of individual step length control will offer itself later, i.e., recombination.

At this point a further word should be said about the alternative $(1{+}\lambda)$ or $(1\,,\lambda)$ strategies. Let us assume that by virtue of a jump landing far from the expectation value, a descendant has made a very large and useful step towards the optimum, thus becoming a parent of the next generation. While the variance allocated to it was eminently suitable for the preceding situation, it is not suited to the new one, being in general much too big. The probability that one of the new descendants will be successful is thereby low. Because the $(1{+}\lambda)$ strategy permits no worsening of the objective function value, the parent survives–and may do so for many generations. This increases the probability of a successful mutation still having a poorly adapted step length. In the $(1\,,\lambda)$ strategy such a stray member will indeed also turn up in a generation, but it will be in effect revoked in the following generation. The descendant that regresses the least survives and is therefore probably the one that most reduces the variance. The scheme thus has better adaptation properties with regard to the step length. In fact this phenomenon can be observed in the simulation. Since we have seen that for $\lambda \geq 5$ the maximum rate of progress is practically independent of whether or not the parent survives, we should favor a $(\mu\,,\lambda)$ strategy, at least when λ/μ is not chosen to be very small, e.g., less than 5 or 6.

5.2.4 The Convergence Criterion for $\mu > 1$ Parents

In Section 5.2.2 we were really looking for the rate of progress of a $(\mu\,,\lambda)$ evolution method. Because of the analytical difficulties, however, we had to fall back on the $\mu = 1$ case, with only one parent. We shall now proceed again on the assumption that $\mu > 1$. In each generation μ state vectors x_E and associated step lengths are stored, which should always be the μ best of the λ mutants of the previous generation. We naturally require more storage space for doing this on the computer, but on the other hand we have more suitable values at our disposal for each variable. Supposing that the topology of the objective function is complicated or even "pathological," and an individual reaches a point that is unfavorable to further progress, we still have sufficient alternative starting points, which may even be much more favorable. According to the usefulness of their parameter sets, some parents place more mutants in the prime group of descendants than others. In general the best individuals of a generation will differ with respect to their variable vectors and objective function values as long as the optimum has not been reached. This provides us with a simple convergence criterion.

From the population of μ parents E_k, $k = 1(1)\mu$, we let F_b be the best objective function value:

$$F_b = \min_k \{F(x_k^{(g)}),\ k = 1(1)\mu\}$$

and F_w the worst

$$F_w = \max_k \{F(x_k^{(g)}),\ k = 1(1)\mu\}$$

Then for ending the search we require that either

$$F_w - F_b \leq \varepsilon_c$$

or

$$\frac{\mu}{\varepsilon_d}\,(F_w - F_b) \leq \left|\sum_{k=1}^{\mu} F(x_k^{(g)})\right|$$

where ε_c and ε_d are to be defined such that

$$\left.\begin{array}{l} \varepsilon_c > 0 \\ 1 + \varepsilon_d > 1 \end{array}\right\} \text{ according to the computational accuracy}$$

Either absolutely or relatively, the objective function values of the parents in a generation must fall closely together before convergence is accepted. The reason for basing the criterion on function values, rather than variable values or step lengths, has already been discussed in connection with the (1+1) strategy (see Sect. 5.1.3).

5.2.5 Scaling of the Variables by Recombination

The $(\mu\,,\lambda)$ method opens up the possibility of imitating a further principle of organic evolution, which is of particular interest from the point of view of numerical optimization problems, namely sexual propagation. By combining the genes of two parents a new source of variation is added to point mutation. The fact that only a few primitive organisms do without this mechanism of recombination leads us to expect that it is very favorable for evolution. Instead of one vector $x_E^{(g)}$ now there are μ distinct vectors $x_k^{(g)}$ for $k = 1(1)\mu$ in a population. In biology, the totality of all genes in a generation is known as a *gene pool*. Among the concerns of population genetics (e.g., Wilson and Bossert, 1973) is the frequency distribution of certain alleles in a population, the so-called gene frequencies. Until now, we did not argue on that level of detail, nor did we go down to the floor of only four nucleic acids in order to model, for example, the mutation process within evolution strategies. This might be worthwhile for quaternary optimization, but not in our case of continuous parameters. It would be a tedious task to model all the intermediate processes from nucleic acids to proteins, cell, organs, etc., taking into account the genetic code and the whole epigenetic apparatus. We shall now apply the *principle of recombination* to numerical optimization with continuous parameters, once again in a simplified fashion.

In our population of μ parents we have stored μ different values of each component x_i, $i = 1(1)n$. From this gene pool we now draw one of the μ values of x_i for each $i = 1(1)n$. The draw should be random so that the probability that an x_i comes from any particular parent (k) of the μ is just $1/\mu$ for all $k = 1(1)\mu$. The variable vector constructed in this way forms the starting point for the subsequent variation of the components. The Figure 5.15 should help to clarify that kind of global recombination.

By imitating recombination in this way we have, so as to speak, replaced bisexuality by multisexuality. This was less for reasons of principle than as a result of practical considerations of programming. A crude test yielded only a slight further increase in the rate of progress in changing from the bisexual to the multisexual scheme, whereas appreciable acceleration was achieved by introducing the bisexual in place of the asexual scheme, which allowed no recombination. A more detailed and exact comparison has yet to be carried out. Without some guidance from theory it is hard to choose the correct initial step lengths and rates of change of step lengths for each of the different algorithms.

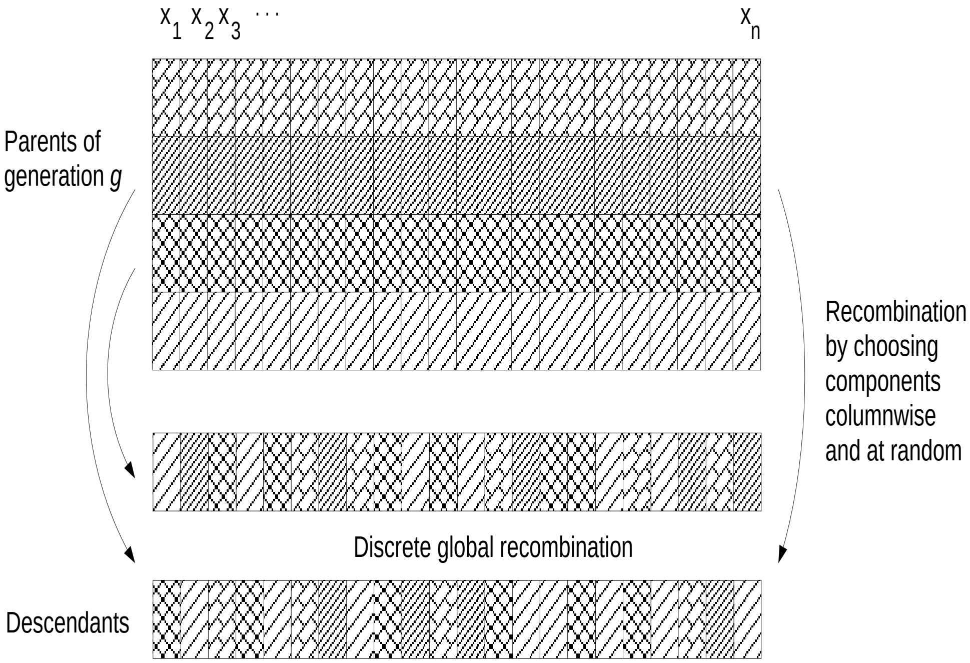
x_1 x_2 x_3 ...
x_n
Parents of generation *g*
Recombination by choosing components columnwise and at random
Discrete global recombination
Descendants

Figure 5.15: Scheme of global uniform recombination

This is, however, the only way to arrive at quantitative statements, free from confusing side effects.

It is thus hard to explain the origin of the accelerating effect of recombination. It may, for example, lie in the fact that instead of μ different starting points, the bisexual scheme offers

$$\mu^2 + \mu\,(\mu - 1) \sum_{i=1}^{n-2} 2^i$$

possible combinations in the case of n variables. With multirecombination, as chosen here, there are as many as μ^n, which is far more than could be put into effect. A more detailed investigation may be found in Bäck (1994a).

So far we have only considered recombination of the object variables, but the strategy variables, the step lengths, can be recombined in just the same way. Even if all the parents start with equal $\sigma_i = \sigma$ for all $i = 1(1)n$, and if all the step length components are varied by a common random factor in the production of descendants, the variances σ_i of all the individuals for each $i = 1(1)n$ differ from each other in the subsequent generations.

Thus by recombination is it possible for the step lengths to adapt individually in this way to circumstances. A better combination affords a higher chance of survival to its bearer. It can therefore be expected that in the course of the optimum search, the currently best combination of the $\{\sigma_i,\ i = 1(1)n\}$ prevails–the one that is associated with the fastest rate of progress. In attempting to verify this in a practical test, an unpleasant phenomenon occurs. It can happen that one of the standard deviations σ_i is suddenly (e.g., by a random value very far from the expectation value) so much reduced in size that the associated variable x_i can now hardly be changed. The total change in the vector x is then roughly speaking within an $(n - 1)$-dimensional subspace of $\mathbb{R}^n$. Contrary to what one might hope, that such a descendant would have less chance of surviving than others, it turns out that the survival of such a descendant is actually favored. The reason is that the rate of progress with an optimal step length is proportional to $1/n$. If the number of variables n decreases, the rate of convergence, together with the optimal step length, increases. The optimum search therefore only proceeds in a subspace of $\mathbb{R}^n$. Not until the only improvement in the objective function entails changing the variable that has hitherto been omitted from the variation will the mutation-selection mechanism operate to increase its associated variance and so restore it to the range for which noticeable changes are possible.

The minimum search proceeds by jumps in the value of the objective function and with rates of progress that vary alternately above and below what would otherwise be smooth convergence. Such unstable behavior is most pronounced when μ, the number of parents, is small. With sufficiently large μ the reserve of step length combinations in the gene pool is always big enough to avoid *overadaptation*, or to compensate for it quickly. From an experimental study (Schwefel, 1987) the conclusion could be drawn that *punctuated equilibrium* evolution (Gould and Eldredge, 1977, 1993) can be avoided by using a sufficiently large population ($\mu > 1$) and a sufficiently low selection pressure ($\lambda/\mu \simeq 7$). A further improvement can be made by using as the starting point in the variation of the step lengths the current average of two parents' variances, rather than the value from only one or the other parent. This measure too has its biological justification; it represents an imitation of what is called *intermediary recombination* (instead of *discrete recombination*).

In this context chromosome mutations should be very effective, those in which for example, the positions of two individual step lengths are exchanged. As well as the haploid scheme of inheritance on which the present work is based, some forms of life also exhibit the diploid scheme. In this case each individual stores two sets of variable values. Whilst the formation of the phenotype only makes use of one allele, the production of offspring brings both alleles into the gene pool. If both alleles are the same one speaks of homozygosity, otherwise of heterozygosity. Heterozygote alleles enlarge the set of variants in the gene pool and thus the range of possible combinations. With regard to the stability of the evolutionary process this also appears to be advantageous. The true gain made by diploidy only becomes apparent, however, when the additional evolutionary factors of recessiveness and dominance are included. For multiple criteria optimization, the usefulness of this concept has been demonstrated by Kursawe (1991, 1992). Many possible extensions of the multimembered scheme have yet to be put into practice. To find their theoretical effect on the rate of progress, one would first have to construct a theory of the (μ , λ) strategy for $\mu > 1$. If one goes beyond the $\mu = 1$ scheme followed here, significant differences between approximate theory and simulation results arise for $\mu > 1$ because of the greater asymmetry of the probability distribution $w(s')$.

5.2.6 Global Convergence

In our discussion of deterministic optimization methods (Chap. 3) we have established that only simultaneous strategies are capable of locating with certainty global minima of arbitrary objective functions. The computational cost of their application increases with the volume of the space under consideration and thus with the power of n. The dynamic programming technique of Bellman allows the reliability of global convergence to be maintained at less cost, but only if the objective function has a rather special structure, such that only a part of the space $\mathbb{R}^n$ needs to be investigated. Of the stochastic search procedures, the Monte-Carlo method has the best chance of global convergence; it offers a high probability rather than certainty of finding the global optimum. If one requires a 90% probability, its cost is greater than that of the equidistant grid search. However, the (1+1) evolution strategy can also be credited with a finite probability of global convergence if the step lengths (variances) of the random changes are held constant (see Rechenberg, 1973; Born, 1978; Beyer, 1989, 1990). How great the chance is of finding an absolute minimum among several local minima depends on the topology, in particular on the disposition and "width" of the minima.

If the user wishes to realize the possibility of a jump from a local to a global extremum, it requires a trial of patience. The requirement of approaching an optimum as quickly and as accurately as possible is always diametrically opposed to maintaining the reliability of global convergence. In the formulation of the algorithms of the evolution strategies we have mainly strived to satisfy the first requirement of rapid convergence, by adaptation of the step lengths. Thus for both strategies no claims can be made for good global convergence properties.

With $\mu > 1$ in the multimembered evolution scheme, several state vectors $x_k^{(g)} \in \mathbb{R}^n$, $k = 1(1)\mu$, are stored in each generation g. If the $x_k^{(g)}$ are very different, the probability is greater that at least one point is situated near the global optimum and that the others will approach it in the process of generation. The likelihood of this is less if the $x_k^{(g)}$ fall close together, with the associated reduction in the step lengths. It always remains finite, however, and increases with μ, the number of parents. This advantage over the (1+1) strategy is best exploited if one starts the search with μ initial vectors $x_k^{(0)}$ roughly evenly distributed over the whole region of interest, and chooses fairly large initial values of the standard deviations $\sigma_k^{(0)} \in \mathbb{R}^n$, $k = 1(1)\mu$. Here too the (μ, λ) scheme is preferable to the $(\mu + \lambda)$ because concentration at a locally very favorable position is at least delayed.

5.2.7 Program Details of the $(\mu \overset{+}{,} \lambda)$ ES Subroutines

Appendix A, Section A.2 contains FORTRAN listings of the multimembered $(\mu \overset{+}{,} \lambda)$ evolution strategy developed here, with the alternatives

GRUP without recombination
REKO with recombination (intermediary recombination for the step lengths)
KORR the so far most general form with correlated mutations as well as five different recombination types (see Chap. 7)

In the choice of μ (number of parents) and λ (number of descendants) there is no need to ensure that λ is exactly divisible by μ. The association of descendants to parents is made by a random selection of uniformly distributed random integers from the range $[1, \mu]$. It is only necessary that λ exceeds μ by a sufficient margin that on average at least one descendant can be better than its parent. From the results of Section 5.2.3 a suitable choice would be for example $\lambda \geq 6\,\mu$.

The transformation from $[0,1]$ evenly distributed random numbers to $(0, \sigma^2)$ normally distributed pseudorandom numbers is carried out in the same way as in subroutine EVOL of the (1+1) strategy (see Sect. 5.1.5). The log-normally distributed variance multipliers are produced by the exponential function. The step lengths (standard deviations of the individual random components) can initially be specified individually. During the subsequent process of generation they satisfy the constraints

$$\left.\begin{array}{rl} & \sigma_i^{(g)} \geq \varepsilon_a \\ \text{and} & \sigma_i^{(g)} \geq \varepsilon_b \, |x_i^{(g)}| \end{array}\right\} \qquad \text{for all } i = 1(1)n$$

where

$$\left.\begin{array}{rl} & \varepsilon_a > 0 \\ \text{and} & 1 + \varepsilon_b > 1 \end{array}\right\} \text{ according to the computational accuracy}$$

can be specified in advance.

The parameter τ which influences the average rate of change of the step lengths should be given a value roughly proportional to $1/\sqrt{n}$; in case of two factors (the case to be preferred), a global and an individual one, the values given in Section 5.2.3 are recommended. The constant of proportionality depends mainly on another adjustable feature, λ/μ, which may be called the selection pressure. For a $(10\,,100)$ strategy it should be set at about unity to allow the fastest convergence of simple optimization problems like the hypersphere. With increasing λ this value φ^* can be changed sublinearly according to

$$\lambda \sim \sqrt{\varphi^*}\, e^{\varphi^*}$$

(compare Equation (5.22)).

If the initial step lengths $\sigma_i^{(0)}$ are chosen to be too large, what may have been an especially well situated starting point $x^{(0)}$ can be thrown away. Nevertheless, this step backwards in the first generation works in favor of reaching a global minimum among several local minima. In principle, for $\mu > 1$ each of the μ different starting vectors $x_k^{(0)} \in \mathbb{R}^n$ and $\sigma_k^{(0)} \in \mathbb{R}^n$, $k = 1(1)\mu$ can be specified. In the present program this differentiation of the parent generation is carried out automatically; the $x_k^{(0)}$ are produced from $x^{(0)}$ by addition of $(0,\, (\sigma^{(0)})^2)$ normally distributed random vectors. The $\sigma_k^{(0)} = \sigma^{(0)}$ are initially equal for all parents.

The convergence criterion is described in Section 5.2.4. It is based on the difference in objective function values between the current best and worst parents of a generation. As accuracy parameters, an absolute and a relative quantity (ε_c and ε_d) must be specified (compare Sect. 5.1.3). Furthermore, an upper bound on the computation time for the search can be given so that whatever the outcome results can be output from the main program (see also Sect. 5.1.5).

Inequality constraints are treated as described for subroutine EVOL (Sect. 5.1.4); so too is the case of the starting point $x^{(0)}$ lying outside the feasible region.

Whereas the subroutine GRUP with option REKO has been taken into account in the test series of Chapter 6, this is not so for the third version KORR, which was created later (Schwefel, 1974). Still, more often than any multimembered version, the (1+1) strategy has been used in practice. Nonetheless it has proved its usefulness in several applications: for example, in conjunction with a linearization method for minimizing quadratic functions in surface fitting problems (Plaschko and Wagner, 1973). In this case the evolution process provides useful approximate values that enable the deterministic method to converge. It should also serve to locate the global minimum of the multimodal objective function. Another practically oriented multiparameter case was to find the optimum weight disposition of lightweight rigidly jointed frameworks (Höfler, Leyßner, and Wiedemann, 1973; Leyßner, 1974). Here again the evolution strategy is combined with another method, this time the simplex method of linear programming. Each strategy is applied in turn until the possible improvements remaining at a step are very small. The usefulness of this procedure is demonstrated by checking against known solutions. A third example is provided by Hartmann (1974), who seeks the optimal geometry of a statically loaded shell support. He parameterizes the functional optimization problem by assuming that the shape of the cross section of the cylindrical shell is described by a suitable polynomial. Its coefficients are to be determined such that the largest absolute value of the transverse moment is as small as possible. For various cases of loading, Hartmann finds optimal shell geometries differing considerably from the shape of circular cylinders, with sometimes almost vanishingly small transverse moments. More examples are mentioned in Chapter 7.

5.3 Genetic Algorithms

At almost the same time that evolution strategies (ESs) were developed and used at the Technical University of Berlin, two other lines of *evolutionary algorithms* (EAs) emerged in the U.S.A., all independently of each other. One of them, *evolutionary programming* (EP), was mentioned at the end of Chapter 4 and goes back to the work of L. J. Fogel (1962; see also Fogel, Owens, and Walsh, 1965, 1966a,b). For a long time, activity on this front seemed to have become quiet. However, in 1992 a series of yearly conferences was started by D. B. Fogel and others (Fogel and Atmar, 1992, 1993; Sebald and Fogel, 1994) to disseminate recent results on the theory and applications of EP. Since EP uses concepts that are rather similar to either ESs or genetic algorithms (GAs) (Fogel, 1991, 1992), it will not be described in detail here, nor will it be compared to ESs on the basis of test results. This was done in a paper presented at the second EP conference (Bäck, Rudolph, and Schwefel, 1993). Similarly, contributions to comparing ESs and GAs in detail may be found in Hoffmeister and Bäck (1990, 1991, 1992; see also Bäck, Hoffmeister, and Schwefel, 1991; Bäck and Schwefel, 1993).

The third line of EAs mentioned above, *genetic algorithms*, has become rather popular today and differs from the others in several aspects. This approach will be explained in the following according to its classical (also called *canonical*) form.

Even to attentive scientists, GAs did not become apparent before 1975 when the first book of Holland (1975) and the dissertation of De Jong (1975) were published. Thus this work was unknown in Europe at the time when Rechenberg's and the author's dissertations were completed and, later on, published as books. Only 10 years later, however, in 1985, a series of biennial conferences (ICGA, International Conferences on Genetic Algorithms) has been started (Grefenstette, 1985, 1987; Schaffer, 1989; Belew and Booker, 1991; Forrest, 1993) to bring together those who are interested in the theory or application of GAs. On the Eastern side of the Atlantic, a similar revival of the field began in 1990 with the first conference on parallel problem solving from nature (PPSN) (Schwefel and Männer, 1991; Männer and Manderick, 1992; Davidor, Schwefel, and Männer, 1994). During the PPSN 90 and the ICGA 91 events, proponents of GAs and ESs agreed upon the common denominators *evolutionary algorithms* (EAs) for both approaches as well as *evolutionary computation* (EC) for a new international journal (see De Jong, 1993). The latter term has been adopted among others by the Institute of Electrical and Electronics Engineers (IEEE) for an international conference during the 1994 World congress on *computational intelligence* (WCCI). Surveys of the history have been attempted by De Jong and Spears (1993) and Spears et al. (1993). As forerunners of the genetic simulation, Fraser (1957), Friedberg (1958), and Hollstien (1971) should at least be mentioned here.

5.3.1 The Canonical Genetic Algorithm for Parameter Optimization

Even if the originators of the GA approach emphasized that GAs were designed for general adaptation processes, most applications reported up to now concern numerical optimization by means of digital computers, including discrete as well as *combinatorial optimization.* Books by Ackley (1987), Goldberg (1989), Davis (1987, 1991), Davidor (1990), Rawlins (1991), Michalewicz (1992, 1994), Stender (1993), and Whitley (1993) may serve as sources for more details in this field. As for so-called classifier systems (CS; see Holland et al., 1986) and genetic programming (GP; see Koza, 1992), two very interesting special areas of evolutionary computation–in which GAs play an important rôle in searching for production rules in so-called knowledge-based systems and for correct expressions in computer programs, respectively–the reader must be referred to the relevant and vast literature (Alander, 1994; he compiled more than 3,000 references).

The GA for parameter optimization usually has been presented in the following general form:

Step 0: (Initialization)
A given population consists of λ individuals. Each is characterized by its genotype consisting of n genes, which determine the vitality, or fitness for survival. Each individual's genotype is represented by a (binary) bit string, representing the object parameter values either directly or by means of an encoding scheme.

Step 1: (Selection)
Two parents are chosen with probabilities proportional to their relative position in the current population, either measured by their contribution to the

mean objective function value of the generation (*proportional selection*) or by their rank (e.g., linear ranking selection).

Step 2: (Recombination)
Two different preliminary offspring are produced by recombination of two parental genotypes by means of crossover at a given recombination probability p_c; only one of those offspring (at random) is actually taken into further consideration.
Steps 1 and 2 are repeated until λ individuals represent the (next) generation.

Step 3: (Mutation)
The offspring eventually (with a given fixed and small probability p_m) underly further modification by means of point mutations working on individual bits, either by reversing a one to a zero, or vice versa; or by throwing a dice for choosing a zero or a one, independent of the original value.

At first glance, this scheme looks very similar to that of a multimembered ES with discrete recombination. To reveal the differences one has to take a closer look at the so-called operators, "selection (S)", "mutation (M)", and "recombination (R)." The GA sequence of events, i.e., S – R – M, as opposed to M – R – S within ESs, should not matter significantly since the whole process is a circular one, and whether one likes to reverse the order of mutation and recombination is a matter of avoiding unnecessary operations or not. In applications, the evaluation of the individuals with respect to their corresponding objective function values normally dominates all other operations. Canonical values for the recombination probability are $p_c = 0.6$, for the number of crossover points $n_c = 2$, and for the mutation probability $p_m = 0.001$.

5.3.2 Representation of Individuals

One of the most apparent differences between GAs and ESs is the fact, that completely different representations of the object variables are used. Organic evolution uses four different nucleotides to encode the genotype in pairs of triplets. By means of the genetic code these are translated to 20 different amino acids. Since there are $4^3 = 64$ different triplets, the genetic code is largely redundant. A closer look reveals its property of maintaining similarity on the amino acid level despite most of the small variations on the level of single nucleotides. Similar transmission laws between chains of amino acids and proteins, proteins and higher aggregates like cells and organs, up to the overall *phenotype* are called *the epigenetic apparatus* (Riedl, 1976). As a matter of fact, biologists as well as behaviorists report that differences among several children of the same parents as well as differences between two consecutive generations can well be described by normal distributions with zero mean and characteristic, probably genetically coded, variances. That is why ESs, when used for seeking optimal values for continuous variables use the more aggregate model of normal distributions for mutations and discrete or intermediary recombination as described in Sections 5.1 and 5.2.

GAs, however, rely on *binary representations* of the object variables. One might call this genotypic modelling of the variation process, instead of phenotypic modelling as is practiced in ESs and EP. An important link between both levels, i.e., the genetic code as well as the so-called epigenetic apparatus, is neglected at least in the canonical GA. For dealing with integer or real values on the level of the object variables GAs make use of a normal Boolean representation or they use the so-called Gray code. Both, however, present the difficulty of so-called *Hamming cliffs*. Depending on its position, a single bit reversal thus can lead to small or very large changes on the phenotypic level. This important fact has advantages and disadvantages. The advantage lies in the broad range of different phenotypes available in a GA population at the same time, a matter affecting its global convergence reliability (for a thorough convergence analysis of the canonical GA see Rudolph, 1994a). The corresponding disadvantage stems from the other side of the same coin, i.e., the inability to focus the search effort in a close enough vicinity of the current positions of individuals in one generation.

There is a second reason to cling to binary representations of object variables within GAs, i.e., Holland's *schema theorem* (Holland, 1975, 1992). This theorem tries to assure exponential penetration of the population by individuals with above average fitness under proportional selection, with sufficiently higher reproduction rates for better individuals, one point crossover with fixed crossover probability, and small, fixed mutation rates.

If, at some time, especially when starting the search, the population contains the globally optimal solution, this will persist in the case where there are zero probabilities for mutation and recombination. Mutation, according to the theorem, is an always destructive force and thus called a subordinate operator. It only serves to introduce missing or reintroduce lost correct bits into finite populations. Recombination (here, *one point crossover*) may or may not be destructive, depending on whether the crossover point happens to lie within a so-called *building block*, i.e., a short substring of the bit string that contributes to above-average fitness of one of the mating individuals, or not. Building blocks are especially important in case of decomposable objective functions (for a more detailed description see Goldberg, 1989).

GAs in their original form do not permit the handling of implicit inequality or equality constraints. On the other hand, explicit upper and lower bounds have to be provided for the range of the object variables:

$$u_i \leq x_i \leq v_i, \qquad \text{for all } i = 1(1)n$$

in order to have a basis for the binary decoding and encoding process, e.g.,

$$x_i = u_i + \frac{v_i - u_i}{2^l - 1} \sum_{j=1}^{l} a_{i,j} \, 2^{j-1}$$

where $a_{i,j}$ for $j = 1(1)l$ represents the bit string segment of length l for encoding the ith element of the object variable vector x.

Instead of this Boolean mapping one also may choose the Gray code, which has the property that neighboring values for the x_i differ in one bit position only. Looking for the probability distribution $p(\Delta x_i)$ of phenotypic changes Δx_i from one generation to the next at a given position $x_i^{(0)}$ and a given mutation probability p_m shows that changing the code from Boolean to Gray only shifts, but never avoids, the so-called Hamming

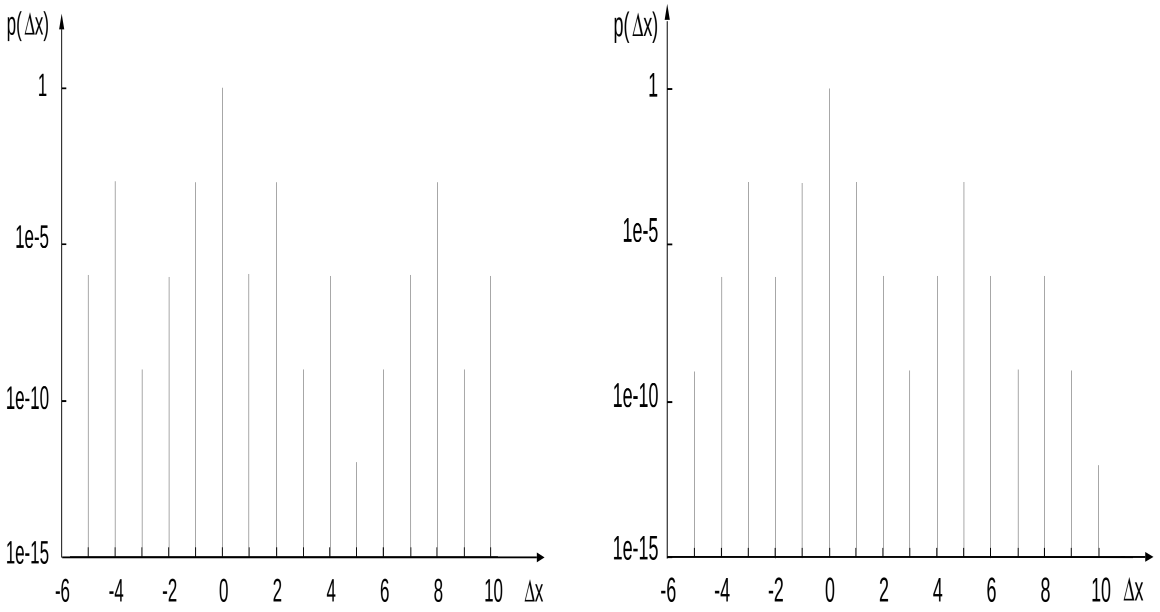

Figure 5.16: Probability distributions for GA mutations / left: normal binary code; right: Gray code

cliffs. As Figure 5.16 clearly shows for a one dimensional case with $x^{(0)} = 5$, $l = 4$, and $p_m = 0.001$, the expectation values for changes Δx are different from zero in both cases, and the distribution is in no case unimodal.

5.3.3 Recombination and Mutation

Innovation during evolutionary processes occurs in two different ways, for so-called higher organisms at least. Only the most early and primitive species operate asexually. People have often said that GAs can do their work without mutations, which, according to the schema theorem, always hamper the adaptation or optimization process, and that, on the other hand, ESs can do their work without recombination. The latter is not true if self-adaptation of the individual mutation variances and covariances is to work properly (see Schwefel, 1987), whereas the former conjecture has been disproved by Bäck (1993, 1994a,b). For a GA the probability of containing the correct bits for the global solution, dispersed over its random start population, is $1 - L\,2^{-\lambda}$, which may be close enough to 1 for $\lambda = 50$ as population size and $L = 1000$ as length of the bit string (actually it is 0.999999999999); however, it cannot be guaranteed that those bits will not get lost in the course of generations. Whether this happens or not, largely depends on the problem structure, the phenomenon being called deception (e.g., Whitley, 1991; Page and Richardson, 1992).

If one looks for recombination effects within GAs on the level of phenotypes, one stumbles over the fact that a recombined offspring of two parents that are close together in the phenotype space may largely deviate from both parental positions there. This

Table 5.1: Two point crossover within a GA and its effect on the phenotypes

	Bit strings	Phenotype
Parent 1	0111 1100	7 12
Parent 2	**1000 1011**	8 11
	Two point crossover	
Offspring 1	0**000 10**00	0 8
Offspring 2	**1**111 11**11**	15 15

completely contradicts the proverbial saying that the apple never falls far from the tree. Table 5.1 shows a simple situation with two parents producing two offspring by means of two point crossover, on a bit string of length 8, and encoding two phenotypic variables in the range $[0, 15]$ in the standard Boolean form. Neither discrete nor intermediary recombination within ESs can be that disruptive; intermediary recombination always delivers phenotypic values for the offspring *between* those of their parents. The assumption that mutations are not necessary for the GA process may even stem from that disruptive character of recombination that permits crossover points not only at the boundaries of meaningful parental information but also within the genes themselves.

ESs obey the general rule, that mutations are undirected, by means of using normally distributed changes with zero mean–even in the case of correlated mutations. That this is not so for GAs can easily be seen from Figure 5.16. Without selection, the GA process thus provides biased genetic drift, depending on the actual situation.

Table 5.2 presents the probability transition matrix for one phenotypic integer variable x_i in the range $[0, 3]$ encoded by means of two bits only. Let

$$p = p_m \ll \frac{1}{2} \text{ single bit inversion probability and}$$

$$q = 1 - p_m \text{ probability of not inverting the bit}$$

From Table 5.2 it is obvious that among all possible transitions (except for those with-

Table 5.2: Transition probabilities for mutations within a GA

			x_i new			
	Genotype		00	01	10	11
		Phenotype	0	1	2	3
x_i old	00	0	q^2	$p\ q$	$p\ q$	p^2
	01	1	$p\ q$	q^2	p^2	$p\ q$
	10	2	$p\ q$	p^2	q^2	$p\ q$
	11	3	p^2	$p\ q$	$p\ q$	q^2

out any change) between the four different genetic states $00, 01, 10, 11$ (e.g., phenotypes $0, 1, 2, 3$), those from 01 to 10 and from 10 to 01 are the most improbable ones despite their phenotypic vicinity. Let $p_m = 10^{-3}$, then $q^2 = 0.998001, p\,q = 0.000999$, and $p^2 = 0.000001$.

5.3.4 Reproduction and Selection

Whether selection is the first or last operator in the generation loop of EAs should not matter except for the first iteration. The difference in this respect between ESs and GAs, however, is that both mingle several aspects of the generation transition. Let us look first, therefore, at the biological facts to be modelled by a selection operator.

An offspring may or may not be able to survive the time span between birth and reproduction. If it is vital up to its reproductive age it may have varying numbers of offspring with one or more partners of its own generation. Thus, the term “selection” in EAs comprises at least three different aspects:

- Survival to adult state (ontogeny)
- Mating behavior (perhaps including promiscuity)
- Reproductive activity

Both ESs and GAs select parents for each offspring anew, thus modelling maximal promiscuity. GAs assign higher mating and reproductive activities to individuals with better objective function values (both for proportional as well as linear or other ranking selection). But even the worst offspring of generation g may become parents for generation $g + 1$. The probability, however, may be very low. If this is the case, most offspring are descendants of a few best parents only. The corresponding loss of diversity in the population may lead to premature stagnation (not convergence!) of the evolutionary seeking process. Reducing the proportionality factor in the selection function, on the other hand, ultimately leads to random walk behavior. This enhances the reliability in multimodal situations, but reduces the convergence velocity and the precision of locating the optimum.

For proportional selection, after Holland derived from an analogy to the game-theoretic multiarmed bandit problem, the average number of offspring for an individual with genotype a_k, phenotype x_k, and vitality $f(x_k)$ is

$$\eta(a_k) = \lambda\, p_s(a_k) = \frac{\Phi(f(x_k))}{\dfrac{1}{\lambda}\sum\limits_{i=1}^{\lambda}\Phi(f(x_i))} = \frac{\Phi_k}{\bar{\Phi}}$$

The transformation $\Phi(f)$ is necessary for introducing the proportionality factor mentioned above as well as for dealing with negative values of the objective function. p_s often is called the survival probability, which is misleading. No parent really survives its generation except in an *elitist* GA version. Then the best parent is put into the next generation without applying the selection operator. Otherwise it may happen simply by chance that one or the other descendant is not different from one of its parents.

In contrast to ESs, the number of offspring always is equal to the number of parents ($\mu = \lambda$). There is no surplus of descendants to cope with lethal mutations and recombinations. ESs need that kind of surplus for handling constraints, at least. In the non-preserving case of its comma-version, a multimembered ES also needs a surplus ($\lambda > \mu$) for the selection process. The $\lambda - \mu$ worst offspring are handled as if they do not survive to the adult reproductive state; the μ best, however, have the same reproduction probability $p_s = 1/\mu$, which does not depend on their individual phenotypes or corresponding objective function values. Thus, on average, every parent has λ/μ descendants. This is depicted on the left-hand side of Figure 5.17, where the average number of descendants of the two best of $\lambda = 10$ descendants (evenly distributed on the fitness scale just for simplification purposes) is just $\lambda/\mu = 5$ for a (2,10) ES, and zero for all others.

Within a GA it largely depends on the scaling function $\Phi(f)$, how many offspring are produced on average by their ancestors. The right-hand part of Figure 5.17 presents two possible situations. Crosses (+) belong to a steep, triangles ($\triangle$) to a flat reproduction probability curve (average number of offspring) over the fitness of the individuals. In the former case it typically happens that, just like in ESs, only the best individuals produce offspring (here the best parent has 6, the second best 3, the third best only 1, and all others zero offspring). One would call this strong selection. Weak selection, on the contrary, characterizes the other case (only the worst parent has no offspring, the best one just 2, and all others 1). It will strongly depend on the actual topology how one should choose the proportionality factor and it may even be necessary to change it during one optimum seeking process.

Self-adaptation of internal strategy parameters is possible within the framework of GAs, too. Bäck (1992a,b, 1993, 1994a,b) has demonstrated this with respect to the mutation rate. For that purpose he adopts the selection mechanism of the multimembered ES.

Last but not least, the question remains whether a stochastic or a deterministic approach to modelling selection is more appropriate. The argument that a stochastic model is closer to reality, is not sufficient for the purpose at hand: optimization and adaptation.

5.3.5 Further Remarks

Of course, one would like to incorporate at least one close-to-canonical GA version into the comparative test series with all the other optimization procedures. But there are problems with that kind of endeavor. First, GAs do not permit general inequality constraints. This does not matter too much, since there are other algorithms that are not applicable directly in such cases, too. Next, GAs must be provided with lower and upper bounds for all parameters, which of course have to be chosen to contain the solution, probably in or near the middle of the hypercube defined by the explicit bounds. The GA thus would be provided with information that is not available for the other algorithms.

For all other methods the starting point is of great importance, not only because it

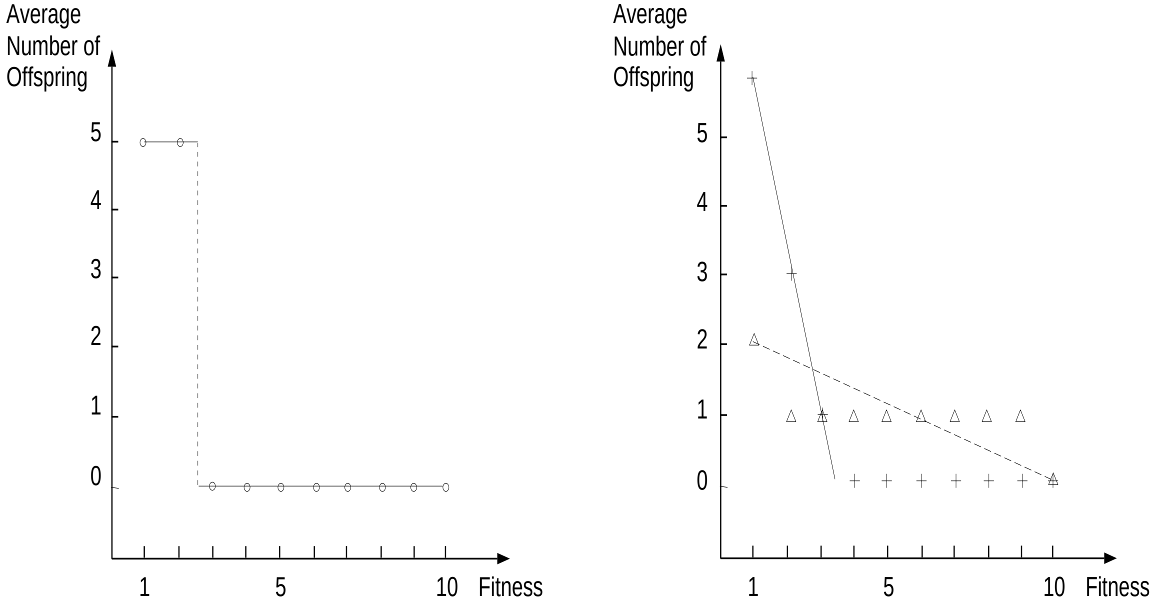

Figure 5.17: Comparison of selection consequences in EAs
left: ES; right: GA

defines the initial distance from the optimum and thus determines largely the number of iterations needed to approximate the solution at the predefined accuracy, but also because it may provide more or less topological difficulties in its vicinity. GAs, however, should be started at random in the whole hypercube defined by the lower and upper bounds of the variables, in order to give them a chance of approaching the global or, at least, a very good local optimum. Reliability tests (see Appendix A, Sect. A.2), especially in cases of multimodal functions would thus be biased against all other methods, if one allows the GA to start from many points at the same time and if one gives the GA the needed extra information about the relevant search region that is not available for the other methods. One might provide special test conditions to compare different EAs with each other without giving one of them an advantage from the very beginning, but no large effort of this kind has been made so far.

Even in cases of special constraints or side conditions one may formulate appropriate instantiations of suitable GA versions. This has been done, for example, for the combinatorial optimization task of solving the travelling salesperson problem (TSP) by Gorges-Schleuter (1991a,b); repair mechanisms were used in cases where unfeasible tours were caused by recombination. Beyer (1992) has investigated ESs for solving TSP-like optimization problems. It is much better to look for data structures fitted to the special task and to redefine the genetic operators to keep to the feasible solution set (see Michalewicz, 1992, 1994). The time for developing such special EAs must be added to the run time on the computer, and one argument in favor of EAs is lost, i.e., their simplicity of use or generality of application.

As the short analysis of GA mutation and recombination operators above has clearly shown, GAs other than ESs favor in-breadth search and thus are especially prepared to solve global and discrete optimization problems, where a volume-oriented approach is more appropriate than a path-oriented one. They have so far done their best in all kinds of combinatorial optimization (e.g., Lawler et al., 1985), a field that has not been pursued in depth throughout this book. One example in the domain of computational intelligence has been the combined topology and parameter optimization of artificial neural networks (e.g., Mandischer, 1993); another is the optimization of membership function parameters within fuzzy controllers (e.g., Meredith, Karr, and Kumar, 1992).

5.4 Simulated Annealing

The simulated annealing approach to solve optimization problems does not really belong to the biologically motivated evolutionary algorithms. However, it belongs to the realm of problem solving methods that make use of other natural paradigms. This is the reason why this section has not been placed elsewhere among the traditional hill climbing strategies.

In order to harden steel one first heats it up to a high temperature not far away from the transition to its liquid phase. Subsequently one cools down the steel more or less rapidly. This process is known as annealing. According to the cooling schedule the atoms or molecules have more or less time to find positions in an ordered pattern (e.g., a crystal structure). The highest order, which corresponds to a global minimum of the free energy, can be achieved only when the cooling proceeds slowly enough. Otherwise the frozen status will be characterized by one or the other local energy minimum only. Similar phenomena arise in all kinds of phase transitions from gaseous to liquid and from liquid to solid states.

A descriptive mathematical model abstracts from local particle-to-particle interactions. It describes statistically the correspondences between macro variables like density, temperature, and entropy. It was Boltzmann who first formulated a probability law to link the temperature with the relative frequencies of the very many possible micro states. Metropolis et al. (1953) simulated on that basis the evolution of a solid in a heat bath towards thermal equilibrium. By means of a Monte-Carlo method new particle configurations were generated. Their free energy E_{new} was compared with that of the former state (E_{old}). If $E_{new} \leq E_{old}$ then the new configuration "survives" and forms the basis for the next perturbation. The new state may survive also if $E_{new} > E_{old}$, but only with a certain probability w

$$w = \frac{1}{c} \exp\left(\frac{E_{old} - E_{new}}{K\,T}\right)$$

where K denotes the famous Boltzmann constant and T the current temperature. The constant c serves to normalize the probability distribution. This Metropolis algorithm thus is in line with the probability law of Boltzmann.

Kirkpatrick, Gelatt, and Vecchi (1983) and Černy (1985) published optimization methods based on Metropolis' simulation algorithm. These methods are used quite frequently nowadays as *simulated annealing* (SA) procedures. Due to the fact that good intermediate positions may be "forgotten" during the search for a minimum or maximum, the algorithm is able to escape from local extrema and finally might reach the global optimum.

There are two loops within the SA process:

- Lowering the temperature (outer loop)
  $T_{new} = f(T_{old}) < T_{old}$, e.g., $T_{new} = \alpha\, T_{old}$, $0 < \alpha < 1$
  until the ground state $T = 0$ is reached

- Waiting until the equilibrium state is found (inner loop)
  Metropolis simulations are performed at $T = const.$ until no further improvements occur

Two questions arise immediately. First, how long should the equilibration phase last, or which constructive criterion should be used for stopping the search for an optimum at a given temperature? Secondly, how large should the cooling steps be? Another question concerns the step size for the perturbations of the variables during the equilibration stage.

There are many empirical suggestions for partial answers to the questions; a lot of successful applications of the method, e.g., for the combinatorial optimization of the travelling salesperson problem (TSP); as well as some rigorous theoretical results concerning the global convergence; but very few investigations about the convergence rates that can be obtained. A good summary may be found in the books of van Laarhoven and Aarts (1987), Aarts and Korst (1989), and Azencott (1992). The relation between SA and evolutionary algorithms (EAs) has been stressed by Rudolph (1993), especially under the parallel computing point of view.

In the following a more detailed pseudocode is given:

```
Step 0:   (Initialization)
          Choose a start position $x^{(0,0)}$,
                 a start temperature $T^{(0)}$,
                 a start width $d^{(0)}$ for the variations of $x$.
          Set $x^* = \tilde{x} = x^{(0,0)}$, $k = 0$, and $\ell = 1$.

Step 1:   (Metropolis simulation)
          Construct $x^{(k,l)} = \tilde{x} + d^{(k)}\, z$,
          where $z$ is uniformly distributed for all components
          $z_i$,  for all $i = 1(1)n$ in the range $z_i \in [-\frac{1}{2}, +\frac{1}{2}]$
          or normally distributed according to $w(z_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\, z_i^2\right)$.
          If $F(x^{(k,l)}) < F(x^*)$ , set $x^* = x^{(k,l)}$.
          If $F(x^{(k,l)}) < F(\tilde{x})$ , go to step 3;
          otherwise draw a uniform random number, $\chi$, from the interval $[0,1]$.
          If $\chi \le \exp\left(\frac{F(x^{(k,l)})-F(\tilde{x})}{T^{(k)}}\right)$ , go to step 3.

Step 2:   (Check for equilibrium)
          If $F(x^*)$ has not been improved within the last $N$ trials,
          go to step 4.

Step 3:   (Inner loop)
          Set $\tilde{x} = x^{(k,l)}$, increase $\ell \leftarrow \ell + 1$, and go to step 1.
```

```
Step 4:  (Termination criterion)
         If $T^{(k)} \le \varepsilon$, end the search with result $x^*$.

Step 5:  (Cooling, outer loop)
         Set $x^{(k+1,0)} = x^*,\ \tilde{x} = x^*$,
         and $T^{(k+1)} = \alpha\, T^{(k)},\ 0 < \alpha < 1$.
         Eventually, decrease $d^{(k+1)} = \beta\, d^{(k)},\ 0 < \beta < 1$.
         Set $\ell = 1$, increase $k \leftarrow k + 1$, and go to step 1.
```

The most important feature of the SA algorithm is its ability to escape from inferior local optima by allowing deteriorations with a certain probability. This kind of *forgetting principle* cannot be found in most numerical optimization routines. In EAs, however, it is more or less built-in as well.

Though the overall structure of the algorithm is rather simple, it turns out to be quite difficult to decide upon the free parameters

$T^{(0)}$	the temperature to start with
$d^{(0)}$	the start width for the step sizes
α	the cooling factor
β	the step size reduction factor
N	the criterion upon which to state "equilibrium"
ε	the lower bound on the temperature

All rules that have been devised rely upon assumptions concerning the special type of objective function. The reader is referred to the literature in this special field, which is closely related to the field of global and stochastic optimization. Laußermair (1992a,b) recently devised a special set of rules called *hyperplane annealing*, and Rudolph (1993) points to similarities with ESs in case of parallel function evaluations.

5.5 Tabu Search and Other Hybrid Concepts

Many heuristic optimum seeking methods, especially those that are called more or less greedy, are in danger of getting trapped in inferior local optima in case of multimodal objective functions. This is especially enhanced by measures to achieve ultimate efficiency, e.g., by controlling the step size or search domain. Tabu search (TS) is a metastrategy aimed at avoiding the local optimality trap and can be superimposed onto many traditional direct optimization methods.

Glover (1986, 1989; see also Glover and Greenberg, 1989) tries to overcome the problem by setting up short-, medium-, and long-term memories of successful as well as unsuccessful trials. According to that history of events, some rules are set up to alternate between three modes of operation:

- Aggressive exploration
- Intensification
- Diversification

Aggressive exploration using a short-term memory forms the core of the TS. From a candidate list of (non-exhaustive) moves the best admissible one is chosen. The decision is based on tabu restrictions on the one hand and on aspiration criteria on the other. Whereas aspiration criteria aim at perpetuating former successful operations, tabu restrictions help to avoid stepping back to inferior solutions and repeating already investigated trial moves. Although the best admissible step does not necessarily lead to an improvement, only better solutions are stored as real moves. Successes and failures are used to update the tabu list and the aspiration memory. If no further improvements can be found, or after a specified number of iterations, one transfers the results to the longer-term memories and switches to either an intensification or a diversification mode.

Intensification combined with the medium-term memory refers to procedures for reinforcing move combinations historically found good, whereas diversification combined with the long-term memory refers to exploring new regions of the search space. The first articles of Glover (1986, 1989) present many ideas to decide upon switching back and forth between the three modes. Many more have been conceived and published together with application results. In some cases complete procedures from other optimization paradigms have been used within the different phases of the TS, e.g., line search or gradient-like techniques during intensification, and GAs during diversification.

Instead of going into further details here, it seems appropriate to give some hints that point to rather similar hybrid methods, more or less centered around either GAs, ESs, or SA as the main strategy.

One could start again with Powell's rule to look for further restart points in the vicinity of the final solutions of his conjugate direction method (Chap. 3, Sect. 3.2.2.1) or with the restart rule of the simplex method according to Nelder and Mead (Chap. 3, Sect. 3.2.1.5), in order to interpret them in terms of some kind of diversification phase. But in general, both approaches cannot be classified as better ideas than starting a specific optimum seeking method from different initial solutions and simply comparing all the (maybe different) outcomes, and choosing the best one as the final solution. It might even be more promising to use different strategies from the same starting point and to select the overall best outcome again as a new start condition. On MIMD (multiple instructions, multiple data) parallel computers or nets of workstations the competition of different search methods could even be used to set up a knowledge base that adapts to a specific situation (e.g., Peters, 1989, 1991). Only individual conclusions for one or the other special application can be drawn from this kind of metastrategic approach, however.

At the close of this general survey, only a few further hints will be given regarding the vast number of recent proposals.

Ablay (1987), for example, uses a basic search routine similar to Rechenberg's (1+1) ES and interrupts it more or less frequently by a pure random search in order to avoid premature stagnation as well as convergence to a non-global local optimum.

The *replicator algorithm* of Voigt (1989) also refers to organic evolution as a metaphor (see also Voigt, Mühlenbein, and Schwefel, 1990). Its modelling technique may be called descriptive, according to earlier work of Feistel and Ebeling (1989). Ebeling (1992) even proposes to incorporate *ontogenetic learning* features (so-called *Haeckel strategy*).

Mühlenbein and Schlierkamp-Voosen (1993a,b) proposed a so-called breeder GA, which combines a greedy algorithm to locate nearest local optima very quickly, with a genetic algorithm to allocate recombined start positions for further local optimum seeking cycles. This has proven to be very successful in special situations where the local optima are situated in a regular pattern in the search space.

Dueck and Scheuer (1990) have devised a so-called *threshold accepting strategy*, which is rather similar to the simulated annealing approach but pretends to deliver superior results. Later on Dueck (1993) elaborated his *great deluge algorithm*, which adds to the threshold accepting method some kind of diversification mode like the tabu search in order to avoid premature stagnation at a non-global local optimum.

Lohmann (1992) and Herdy (1992) propose a hierarchical ES according to Rechenberg's extended notation (Rechenberg, 1978, 1989, 1994) of the multimembered scheme to solve so-called structural optimization problems. Whereas this term normally points to situations in which a solid structure subject to stresses and deformations has to be designed in order to have least weight or production cost, Lohmann and Herdy do not mean anything else than a mixed-integer optimization problem. The solution is sought for in an outer ES-loop that varies the integer object variables only and an inner ES-loop that varies the real-valued variables. Thus the outer loop compares relative optima found in the inner loops. This kind of cyclical subspace search, somehow similar to the Gauss-Seidel approach, must not represent the ultimate solution to mixed-integer problems, however. It is more or less prone to finding non-global local optima only. A more general evolutionary algorithm should be able to change–at the same time, by appropriate mutation and recombination operators–both the discrete and the real-valued object variables. But this speculation must be proved in forthcoming further steps towards a more general evolutionary algorithm, perhaps a hybrid of ES and GA ingredients.