Genetic Algorithms in Engineering and Computer Science

GC 26/7/1995 12:30—PAGE PROOFS for John Wiley & Sons Ltd (jwbook.sty v3.0, 12-1-1995)

Genetic Algorithms in Engineering and Computer Science

Edited by J. Périaux and G. Winter

JOHN WILEY & SONS
Chichester · New York · Brisbane · Toronto · Singapore

7

Evolution Strategies II: Theoretical Aspects

Hans-Paul Schwefel[1] and Thomas Bäck[2]

7.1 SOME DEFINITIONS FROM PROBABILITY THEORY AND STATISTICS

This section shortly presents some notions and corollaries from the theory of stochastic events that will be helpful later on in discussing the convergence reliability and convergence velocity of evolution strategies (ESs). More detailed foundations may be found in textbooks on statistics and probability theory.

7.1.1 Random variables, distribution and density functions

A continuous random variable $X : \Omega \rightarrow I\!R$ defined over a probability space $(\Omega, \mathcal{A}, p)$ with Ω as set of elementary events, $\mathcal{A}$ as algebra of the events, and p as probability measure is characterized by the distribution function

$$\begin{aligned} F_X : I\!R &\rightarrow [0,1] \\ x &\mapsto F_X(x) = p(X \leq x). \end{aligned}$$

The density function $f_X : I\!R \rightarrow I\!R$ of the random variable X is implicitly defined by the equation

$$F_X(x) = \int_{-\infty}^{x} f_X(z)dz \qquad \forall x \in I\!R. \tag{7.1}$$

[1] University of Dortmund, Department of Computer Science, 44221 Dortmund, Germany. E-mail: schwefel@ls11.informatik.uni-dortmund.de
[2] Informatik Centrum Dortmund, Joseph-von-Fraunhofer-Str. 20, 44227 Dortmund, Germany. E-mail: baeck@ls11.informatik.uni-dortmund.de
Genetic Algorithms in Engineering and Computer Science
Editor J. Périaux and G. Winter

As *support*(X) one denotes the set $\{x \in I\!R \mid F_X(x) > 0\}$ of all real numbers with strictly positive density.

7.1.2 Characteristic values of probability distributions

One dimensional distributions

The expectation of a one dimensional continuous random variable X with density f_X is defined by

$$\xi \;=\; \mathrm{E}[X] \;:=\; \int_{-\infty}^{\infty} x f_X(x) dx. \tag{7.2}$$

As measures for the dispersion the variance

$$\sigma^2 \;=\; D^2[X] \;:=\; \int_{-\infty}^{\infty} (x - \mathrm{E}[X])^2 f_X(x) dx$$

and the standard deviation $\sigma = D[X]$ are used. Let $h : I\!R \to I\!R$ be a monotonous function of a continuous random variable with density f_X, then the expectation of h can be calculated as

$$\mathrm{E}[h(X)] \;=\; \int_{-\infty}^{\infty} h(x) f_X(x) dx.$$

Multidimensional distributions

The vector of expectations of an n-dimensional random set $X = (X_1, \ldots, X_n)^T : \Omega \to I\!R^n$ is defined as

$$\xi \;=\; \mathrm{E}[X] \;:=\; (\mathrm{E}[X_1], \ldots, \mathrm{E}[X_n])^T .$$

The dispersion of two random variables X_i and X_j may be presented in the form of a covariance matrix $\mathrm{Cov}[X_i, X_j]$ with elements

$$\sigma_{ij} = \sigma_{ji} \;=\; \mathrm{E}\left[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])\right].$$

Diagonal elements $\sigma_{ii} = \sigma_i^2$ correspond to the individual one dimensional variances.

The covariance matrix

$$\Sigma_X \;:=\; \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{pmatrix}$$

contains the complete information on the dispersion of an n-dimensional random set. If $\sigma_{ij} = 0$ for any two random variables X_i and X_j, then they are termed uncorrelated. Stochastically independent random variables are always uncorrelated; the inverse, however, does not hold, generally.

7.1.3 Special distributions

The Gaussian or normal distribution

A distribution for a random variable like

$$F_X(x) \quad = \quad \Phi(x) \quad = \quad \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-\xi)^2}{2\sigma^2}\right) dz$$

is called a normal or Gaussian distribution if ξ is identical to its expectation and σ^2 identical to its variance. The corresponding density of the $N(\xi, \sigma^2)$ distribution is

$$f_X(x) \quad = \quad \phi(x) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right).$$

In the following Φ and ϕ shall always denote the distribution and density functions of a standard normal distribution $N(0,1)$. Furthermore, N, N', N'', and $N_i (i = 1, \ldots, n)$ always denote standard normally distributed random variables.

A normally distributed random variable $X \sim N(\xi, \sigma^2)$ can always be transformed into a standard normally distributed random variable $Y \sim N(0,1)$ by means of the operation $Y := (X - \xi)/\sigma$.

An important and later on often used quality of normal distributions is given by their addition theorem: Let $X_1, \ldots, X_n$ be stochastically independent $N_i(\xi_i, \sigma_i^2)$ distributed random variables, then

$$\sum_{i=1}^{n} X_i \quad \sim \quad N\left(\sum_{i=1}^{n} \xi_i, \sum_{i=1}^{n} \sigma_i^2\right) \tag{7.3}$$

is valid. The sum over normally distributed random variables is always also normally distributed, and the resulting expectation as well as the resulting variance may be calculated easily as sums of the individual items according to (7.3).

The n-dimensional normal distribution

The multidimensional normal distribution is a simple and natural extension of the one dimensional normal distribution introduced above to a vector $X = (X_1, \ldots, X_n)^T$ of random variables. All marginal distributions are standard normal distributions. The density function of the vector X has the form

$$f_X(x) \quad = \quad \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|\Sigma_X|}} \exp\left(-\frac{1}{2}(x-\xi)^T \Sigma_X^{-1}(x-\xi)\right)$$

where ξ is identical to the vector of the individual expectations and Σ_X is identical to the covariance matrix, $|\Sigma_X|$ being its determinant.

An important characteristic of the multidimensional normal distribution is the equivalence of the two terms 'stochastic independence' and 'uncorrelatedness' of the single components of the random vector. If the correlation matrix is a diagonal matrix, then the single components of the random vector are stochastically independent from each other, and the elements of the main diagonal are equal to the variances of the single random variables, the marginal distributions of which are all normally distributed. Helpful for all following conclusions is the next definition:

Definition 7.1.1 *An n-dimensional random vector Z is called elliptically distributed if it has a representation $Z = RX$ where the random vector X is uniformly distributed over an n-dimensional ellipsoid and R is a non-negative random variable that is stochastically independent of X. Z is called spherically distributed if X is uniformly distributed over an n-dimensional sphere in addition to the demands above.*

The multidimensional normal distribution belongs to the class of elliptical distributions. In case of stochastic independence of the components of the random vector and identical variances it is spherically distributed. The distribution of the parameter R (sometimes called total mutation step size in ESs) may be reduced to a χ^2 distribution in that case.

The χ^2 distribution

Let $X_1, \ldots, X_n$ be stochastically independent and identically standard normal distributed components, then the distribution of $Y := \sum_{i=1}^{n} (X_i)^2$ is a χ^2 distribution with n degrees of freedom. Its expectation is n, its variance $2n$. For large n the χ^2_n distribution can be approximated by an $N(n, 2n)$ distribution, a fact that can be expressed by

$$\chi^2_n \rightarrow n + \sqrt{2n} N. \tag{7.4}$$

$R := +\sqrt{Y}$ then may be handled as if $R \sim N(\sqrt{n}, \sqrt{\frac{1}{2}})$.

7.1.4 Order statistics

Let $X_1, \ldots, X_n$ be stochastically independent continuous random variables with distribution function F_X and density function f_X, then the distribution function F_{Y_1} of $Y_1 := \max\{X_1, \ldots, X_n\}$ can be calculated by

$$F_{Y_1}(x) \quad = \quad \prod_{i=1}^{n} F_{X_i}(x). \tag{7.5}$$

If in addition the X_i are identically distributed, then

$$F_{Y_1}(x) \quad = \quad [F_X(x)]^n \tag{7.6}$$

is valid for the distribution function of the ensemble, and after (7.1) the density of the maximum of the X_i is

$$f_{Y_1}(x) \quad = \quad \frac{\partial}{\partial x} F_{Y_1}(x) \quad = \quad n f_X(x) \left[F_X(x)\right]^{n-1}.$$

This is a special case only of the more general result from order statistics, which states the distribution function of the mth largest X_i to be

$$f_{Y_m}(x) \quad = \quad n \binom{n-1}{m-1} f_X(x) \left[F_X(x)\right]^{n-m} \left[1 - F_X(x)\right]^{m-1}.$$

7.2 ON THE CONVERGENCE BEHAVIOR OF EVOLUTION STRATEGIES

This section summarizes so far known results on the convergence properties of evolution strategies. At first, the simple (1+1) ES will be considered. The extension to $(\mu + \lambda)$ ESs is trivial since they exert results that are at least as good as those of the (1+1) ES. More difficult are assertions about (μ, λ) ESs. A (1,1) ES, for example, represents a simple random walk strategy without selection. Moreover, figure 7.1 shows that it always diverges in nonlinear cases except for $\sigma = 0$ (lowermost curve).

Criteria for the merits of an optimization technique comprise the reliability of finally reaching an ϵ-vicinity of the optimum that is sought and, even more important, the time or number of iterations necessary to reach that goal. More specifically, the convergence rate should say something about the reduction of the distance to the goal over time — may it be with respect to the objective function values or the Euclidean distance in the space of the decision variables. So far, results of this kind have been achieved only for a very limited class of situations. Nevertheless, such theoretical results are very helpful in tuning external strategy parameters to actual situations at hand.

7.2.1 Convergence reliability

In this subsection we consider a (1+1) ES starting from $x^{(0)}$ with a mutation operator $x^{(t+1)} := x^{(t)} + Z$ where $Z \sim N(0, \sigma^2 I_n)$ is a normally distributed vector with stochastically independent components of same variance σ^2. I_n denotes the n-dimensional unit matrix.

In order to guarantee the convergence of an ES, one has to restrict oneself to the class of regular optimization problems. This restriction is not stark, however, since the features to be demanded to the objective function are rather weak and moreover essential for most other optimization procedures, as well.

Definition 7.2.1 *An optimization problem*

$$f^* \quad = \quad f(x^*) \quad = \quad \min\{f(x) \mid x \in M \subseteq I\!R^n\}$$

is called regular if and only if

- $f^* > -\infty$,
- $x^* \in \mathrm{int}(M)$, and
- $\mu(\{x \in M \mid f(x) \in U_\epsilon(f^*)\}) > 0 \qquad \forall \epsilon > 0$,

where μ is the Lebesgue measure, $\mathrm{int}(M)$ *the set of internal points of M, and U_ϵ an ϵ-environment of its argument.*

One calls f the objective function, f^* the global minimum, and x^* the solution to the optimization problem.

The necessity of the first requirement is immediately clear, whereas the second one only simplifies the analysis and is used in the proof of the following convergence theorem. The third requirement excludes optimization tasks with isolated global optima, which cannot be reached with a probability greater than zero.

Theorem 7.2.1 *Let $\epsilon > 0$ and $p_t := p(x^{(t)} \in \{x \in M \mid f(x) \in U_\epsilon(f^*)\})$ be the probability that a population of the (1+1) ES has reached the point $x^{(t)}$ at iteration t, the objective function value belonging to which is closer to the goal f^* than ϵ. Then, assuming*

$$\sum_{t=0}^{\infty} p_t \quad = \quad \infty \tag{7.7}$$

implies that

$$p(\lim_{t \to \infty} (f(x^{(t)}) - f^*) = 0) \quad = \quad 1$$

for any starting point $x^{(0)} \in M$.

The condition (7.7) in Theorem 7.2.1 looks rather abstract, but one can show that it is the result of a more obvious condition that is fulfilled in many practical situations:

Lemma 7.2.1 *If $M \subseteq support(f_Z)$, where f_Z denotes the probability density of the random vector Z of the mutation operator, and M is bounded, then (7.7) is valid.*

Of course, this result is of academic interest only since nobody likes to wait for a result until the end of time. More interesting are results about the expected convergence velocity.

7.2.2 Convergence velocity

The following definition is useful in order to give some quantitative contents to the term convergence velocity:

Definition 7.2.2 *The value $\delta_t := \mathrm{E}[f(x^{(t)}) - f^*]$ is called expected error at step t. An algorithm has polynomial convergence order if $\delta_t = O(t^{-\alpha})$ with $\alpha > 0$; its convergence rate is exponential if $\delta_t = O(\beta^t)$ with $\beta \in (0, 1)$.*

In order to gain statements about the convergence rate, one must restrict oneself to a special class of problems, here to the class of strictly convex problems. Strict convexity implies among other features continuous differentiability and unimodality, and thus a stark restriction. Especially, strictly convex problems may easily be solved by means of gradient type strategies and are thus not the domain of evolutionary algorithms. EAs, however, should guarantee something at least in such simple situations as well, and there is some hope that it will be possible to weaken these conditions in the future in order the arrive at results for a significantly broader and more interesting class of optimization problems.

Theorem 7.2.2 *Let $f : M \to \mathbb{R}$ be strictly convex and the mutation step size of a* (1+1) ES *be spherically distributed with $Z = RU$ where $support(R) = (0, a) \neq \emptyset$. Then the expected error δ_t at step t for any start position $x^{(0)} \in M$ is*

$$\delta_t \quad = \quad \begin{cases} O(t^{-2/n}) & \text{for a constant mutation step size} \\ O(\beta^t) & \text{for an adaptive mutation step size} \end{cases}$$

with $\beta \in (0, 1)$ and step size adaptation according to $R^{(t+1)} = \|\nabla f(x^{(t)})\| R^{(t)}$.

For a static mutation probability distribution one can thus state only polynomial convergence rates, whereas exponential convergence rates may be achieved in case of an appropriate step size control. Rappl [Rap84] proved, in addition, that a rule similar to the so-called success rule introduced by Rechenberg [Rec94] to adapt the variance of the mutation operator, can yield an exponential convergence rate.

Whereas the definition of the term convergence rate above only covers the order of magnitude of the convergence velocity, the factor β is of decisive interest in practical applications. That is why it is interesting to calculate optimal parameters for the step size control. For a simple strictly convex objective function such calculations have been performed [BRS93, Rec94, Schw95]. The results will be summarized in the following.

7.3 THE SPHERE MODEL

The following section contains results gained for an objective function of type $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $f(x) := \|x - x^*\|^2 = r^2$. For different ES versions that all use normally distributed mutation vectors with mean $\xi = 0$ and covariance matrix $\sigma^2 I_n$ (i.e., with just one strategy parameter σ) optimal values σ^* for the mutation step sizes shall be calculated. On that basis one can derive the corresponding success probability , i.e., the mean number of cases in which the successor is better than its predecessor. This figure is independent of the current distance r to the optimum and thus may be used to control the step size adaptation.

There is some hope that the same control mechanism may yield success probabilities of the same order of magnitude also for situations that differ largely from that of the sphere model. At least one completely different situation has been identified for which this is valid, i.e., the so-called corridor model, a simple linear function with constraints [Rec94, Schw95].

7.3.1 The (1+1) ES or two membered evolution strategy

Here we want to find out the optimal value of the standard deviation σ in case of a (1+1) ES with mutation operator $x^{(t+1)} := x^{(t)} + Z$ where $Z \sim N(0, \sigma^2 I_n)$ is n-dimensionally normal distributed. For that purpose we consider the relative progress during step t

$$P := \frac{r^{(t)} - r^{(t+1)}}{r^{(t)}} \tag{7.8}$$

that denotes the relative approximation to the goal in the space of the decision variables. Using the addition theorem of the normal distribution and relation (7.4), one gets (writing simply r instead of $r^{(t)}$):

$$\begin{aligned} r^{(t+1)} &= \|x^{(t)} - x^* + Z\| \\ &= \sqrt{\sum_{i=1}^{n} \left((x_i^{(t)} - x_i^*) + \sigma N_i \right)^2} \end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\sum_{i=1}^{n}(x_i^{(t)}-x_i^*)^2 + 2\sigma\sum_{i=1}^{n}(x_i^{(t)}-x_i^*)N_i + \sigma^2\sum_{i=1}^{n}N_i^2} \\
&\sim \sqrt{r^2 + 2\sigma r N + \sigma^2(n+\sqrt{2n}N')} \\
&\sim \sqrt{r^2 + \sigma^2 n + \sqrt{2n\sigma^4 + 4r^2\sigma^2}N''}.
\end{aligned}
$$

Inserting this result into (7.8) yields the relative progress in terms of the dimensionless item $\check{\sigma} := \sigma n/r$ (using a Taylor series for $\sqrt{1+x}$ and restricting to the linear term):

$$
\begin{aligned}
P &= 1 - \sqrt{\frac{r^2 + \sigma^2 n + \sqrt{2n\sigma^4 + 4r^2\sigma^2}N''}{r^2}} \\
&= 1 - \sqrt{1 + \frac{\check{\sigma}^2}{n} + \frac{\check{\sigma}^2}{n}\sqrt{\frac{4}{\check{\sigma}^2} + \frac{2}{n}}N''} \\
&\approx 1 - \sqrt{1 + \frac{\check{\sigma}^2}{n} + \frac{2\check{\sigma}}{n}N''} \quad \approx \quad 1 - \left(1 + \frac{\check{\sigma}^2}{2n} + \frac{\check{\sigma}}{n}N''\right) \\
&= -\frac{\check{\sigma}^2}{2n} - \frac{\check{\sigma}}{n}N'' \quad \sim \quad N\left(-\frac{\check{\sigma}^2}{2n}, \frac{\check{\sigma}^2}{n^2}\right). \qquad (7.9)
\end{aligned}
$$

Since the (1+1) ES only accepts improvements that in case of the sphere model occur if the distance to the optimum is diminished, the expectation of the random variable $P_{1+1} := \max\{0, P\}$ has to be considered only. This can be done in the following way

$$
\begin{aligned}
\mathrm{E}[P_{1+1}] &= \int_0^\infty x f_P(x)dx \quad =: \quad \frac{\varphi_{1+1}}{r} = \frac{\check{\varphi}_{1+1}}{n} \\
&= \int_0^\infty \frac{nx}{\sqrt{2\pi}\check{\sigma}} \exp\left[-\frac{1}{2}\left(\frac{nx + \check{\sigma}^2/2}{\check{\sigma}}\right)^2\right] dx \\
&= \frac{1}{n}\left\{\frac{\check{\sigma}}{\sqrt{2\pi}}\exp\left(-\frac{\check{\sigma}^2}{8}\right) - \frac{\check{\sigma}^2}{2}\left[1 - \Phi\left(\frac{\check{\sigma}}{2}\right)\right]\right\}
\end{aligned}
$$

where Φ is the distribution function of a standard normal random variable and φ Rechenberg's progress velocity, $\check{\varphi}_{1+1} = \varphi_{1+1}n/r$ denoting the dimensionless correspondent to $\check{\sigma}$. The solid line in figure 7.1 shows that this mean value is maximal for $\check{\sigma}^* = 1.224$ at $\check{\varphi}^*_{1+1} = 0.2025$ and thus for

$$
\sigma^* = \frac{1.224}{n}r = \frac{0.612}{n}\|\nabla f(x)\| \quad \text{at} \quad \varphi^*_{1+1} = \frac{0.2025}{n}r.
$$

The corresponding success probability p_s to the optimally adjusted step size σ^* becomes

$$
\begin{aligned}
p_s^* &= p(P>0)\,|_{\sigma=\sigma^*} = p\left(-\frac{{\sigma^*}^2}{2n} - \frac{\sigma^*}{n}N'' > 0\right) \\
&= \Phi\left(-\frac{\sigma^*}{2}\right) \approx 0.270.
\end{aligned}
$$

This result has been the basis for designing an algorithm that adapts its current step size automatically.

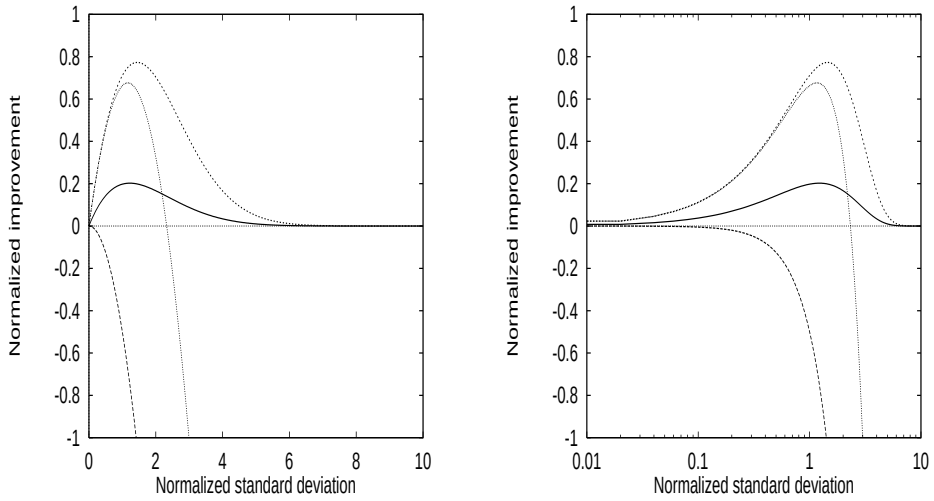

Figure 7.1: The normalized expected improvement $\check{\varphi}$ over the normalized standard deviation $\check{\sigma}$ (from top to bottom the curves correspond to $(1+5)$-, $(1,5)$-, $(1+1)$-, and $(1,1)$ evolution strategies; normal scale on the left, logarithmic scale for $\check{\sigma}$ on the right hand side)

7.3.2 The $(\mu+1)$ evolution strategy

Rather early, the two membered evolution strategy has been the basis for expanding the population size. Three different ways of realizing this idea can be identified:

- Increasing the number of predecessors,
- Increasing the number of successors,
- Increasing both numbers.

The corresponding ES versions have been termed $(\mu+1)$-, $(1+\lambda)$-, and $(\mu+\lambda)$ ESs, a variant of the latter being called (μ,λ) evolution strategy.

Within the $(\mu+1)$ ES only the number of predecessors is increased from 1 to μ. It still works in a strictly sequential way by creating just one successor at a time. Only if the successor outperforms at least one of the μ predecessors, it replaces one of the latter - generally the worst one. Therefore, one might speak of an 'extinction-of-the-worst' principle that can be found in a couple of other optimization techniques, as well, e.g., in the EVOP- or evolutionary operation method of G. E. P. Box[Box57], a deterministic factorial design technique for experimental optimization, and in all polyhedron strategies like Nelder and Mead's Simplex[NM65]- or M. J. Box's Complex[Box65] methods.

The $(\mu+1)$ evolution strategy invites straightforwardly the incorporation of another evolution principle: recombination. Any mix of two (or even more, especially when used as a computer program) parental genomes may be transferred to the successor before the mutation operator does its work. Rechenberg[Rec94] showed already in the early 1970ies that recombination may enhance the efficiency of an evolution strategy

considerably. On the other hand it consumes additional storage capacity, which was relatively expensive in those days. The main reason why the $(\mu + 1)$ ES is no longer used today lies in its inability to incorporate self-adaptation of the internal strategy parameters. We therefore skip all theoretical analyses of that ES variant here.

7.3.3 The $(1 + \lambda)$ evolution strategy

Increasing the size of the predecessor's progeny per generation from just only 1 to λ improves the convergence velocity per iteration cycle considerably. The same problem as above, the sphere model, is considered again. Only the best successor becomes predecessor of the next cycle here, and if it turns out to be worse than its predecessor, then the predecessor 'survives'.

According to (7.9), the relative change in the distance to the goal for each successor is already known to be

$$P \approx -\frac{\check{\sigma}^2}{2n} - \frac{\check{\sigma}}{n} N''.$$

The distribution function thus is $F_P(x) = \Phi((x-\theta)/\eta)$ with expectation $\theta = -\check{\sigma}^2/2n$ and standard deviation $\eta = \check{\sigma}/n$. According to subsection 7.1.4 on order statistics the maximal change P_λ, i.e. that of the best successor, is, according to the stochastic independence of the successors, distributed according to

$$F_{P_\lambda}(x) = \Phi^\lambda\left(\frac{x-\theta}{\eta}\right).$$

The expectation of P_λ can be calculated by means of (7.2) as

$$\mathrm{E}[P_\lambda] = \frac{1}{\eta}\int_{-\infty}^{\infty} x f_{P_\lambda}\left(\frac{x-\theta}{\eta}\right) dx, \tag{7.10}$$

where the density of P_λ results from (7.1) as

$$f_{P_\lambda}(x) = \frac{d}{dx}\Phi^\lambda(x) = \frac{\lambda}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \Phi^{\lambda-1}(x).$$

The expectation of the relative improvement with respect to the random variable $P_{1+\lambda} := \max\{0, P_\lambda\}$ can be calculated in analogy to the (1+1) ES case according to

$$\mathrm{E}[P_{1+\lambda}] = \frac{1}{\eta}\int_{0}^{\infty} x f_{P_\lambda}\left(\frac{x-\theta}{\eta}\right) dx.$$

Substituting $z := (x-\theta)/\eta$ one arrives at

$$\begin{aligned}\mathrm{E}[P_{1+\lambda}] &= \int_{-\theta/\eta}^{\infty} (z\eta + \theta) f_{P_\lambda}(z) dz \\ &= \frac{\check{\sigma}}{n}\int_{\check{\sigma}/2}^{\infty} z f_{P_\lambda}(z) dz - \frac{\check{\sigma}^2}{2n}\left[1 - \Phi^\lambda\left(\frac{\check{\sigma}}{2}\right)\right].\end{aligned}$$

A numerical evaluation is presented in figures 7.2 and 7.3 (solid lines on the left hand sides for the $(1+\lambda)$ ES): Increasing the number λ of successors should be accompanied by enlarging the mutation step size. Since only the best successor is of importance here, the increasing number of failures is completely ignored.

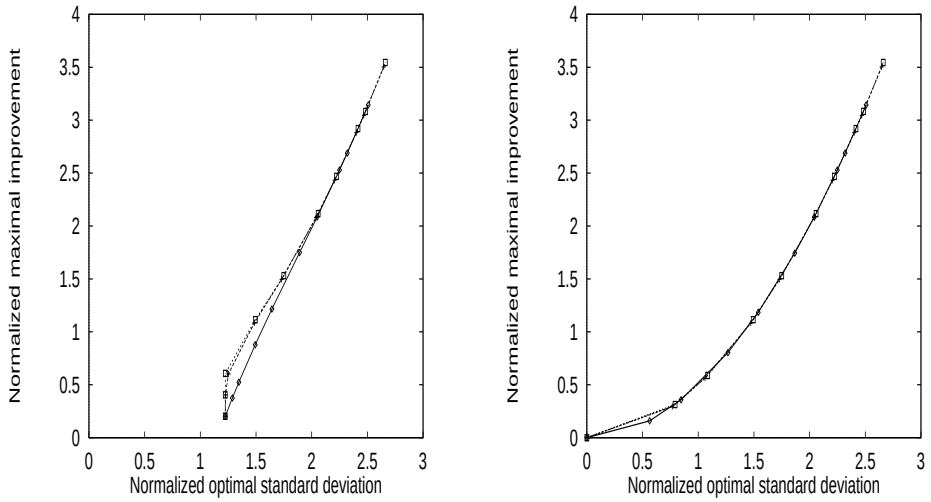

Figure 7.2: The normalized maximal improvement $\breve{\varphi}^*$ over the normalized optimal standard deviation $\breve{\sigma}^*$ for plus-ESs (left) and comma-ESs (right) (from bottom to top in each figure the curves correspond to the cases $\mu = 1$, $\mu = 10$, and $\mu = 50$ with varying ratios $\frac{\lambda}{\mu}$)

7.3.4 The (1,λ) evolution strategy

Each non-elitist (or comma-) version of the evolution strategies that ignores the so far achieved best intermediate result during the selection process, is prone to diverge — at least temporarily. A convergence proof thus is a difficult task. Nevertheless, this task has been solved meanwhile [Rud94]. We shall not go into the details here, however.

For certain optimization problems it is even possible to calculate mean convergence rates. We shall demonstrate that here for the above introduced objective function $f(x) := \|x - x^*\|^2 = r^2$.

In contrary to all plus-versions of the ES above, the $(1, \lambda)$ ES may lead to intermediary deterioration. The expected relative improvement is given by P_λ directly and thus may be negative, especially. The substitution $z = (x - \theta)/\eta$ in (7.10) yields

$$\begin{aligned} \mathrm{E}[P_{1,\lambda}] &= \int_{-\infty}^{\infty} (z\eta + \theta) f_{P_\lambda}(z) dz \\ &= \frac{\breve{\sigma}}{n} \int_{-\infty}^{\infty} z f_{P_\lambda}(z) dz - \frac{\breve{\sigma}^2}{2n} \\ &= (2\breve{\sigma} c_{1,\lambda} - \breve{\sigma}^2)/2n, \end{aligned} \tag{7.11}$$

where

$$c_{1,\lambda} := \int_{-\infty}^{\infty} z f_{P_\lambda}(z) dz$$

denotes the expectation of the maximum of λ stochastically independent standard normally distributed random variables, a constant that depends on λ only.

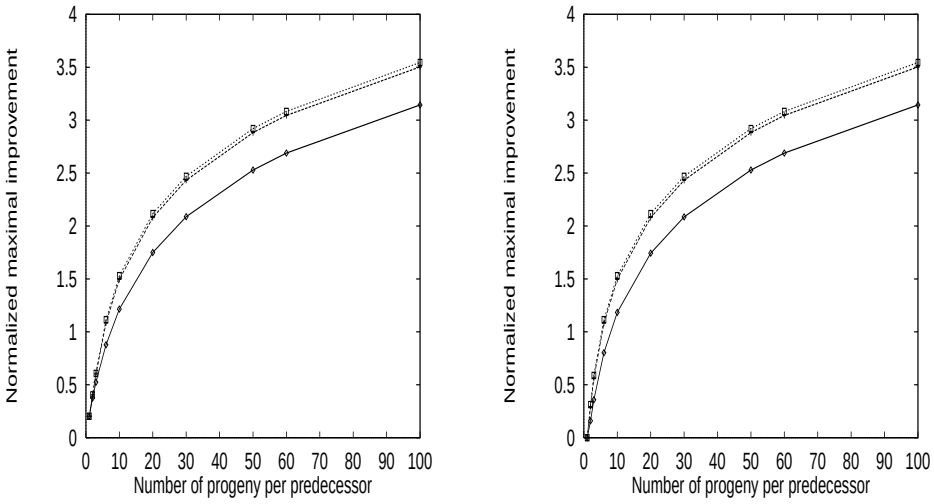

Figure 7.3: The normalized maximal improvement $\breve{\varphi}^*$ over the ratio $\frac{\lambda}{\mu}$ for plus-ESs (left) and comma-ESs (right) (from bottom to top in each figure the curves correspond to the cases $\mu = 1$, $\mu = 10$, and $\mu = 50$ with varying ratios $\frac{\lambda}{\mu}$)

Differentiating equation (7.11) with respect to $\breve{\sigma}$, one obtains due to

$$\partial^2 \mathrm{E}[P_{1,\lambda}]/\partial \breve{\sigma}^2 \equiv -1 < 0$$

as optimal normalized standard deviation $\breve{\sigma}^* = c_{1,\lambda}$. The optimal mean mutation step size thus becomes

$$\sigma^* \quad = \quad \frac{c_{1,\lambda}}{n} r \quad = \quad \frac{c_{1,\lambda}}{2n} \|\nabla f(x)\|.$$

The expectation of an improvement corresponds to $\mathrm{E}^*[P_{1,\lambda}] = c_{1,\lambda}^2/2n$, or in normalized terms, to $\breve{\varphi}^*_{1,\lambda} = \frac{1}{2} c_{1,\lambda}^2 = \frac{1}{2} \breve{\sigma}^{*2}$, respectively. Values for $c_{1,\lambda}$ are tabulated and may be found in [Rec94]. For large n and large λ one may make use of an approximation reported in [BRS93] saying that $c_{1,\lambda}$ is approximately proportional to $\sqrt{2 \ln \lambda}$. With this result one may conclude for the expectation of the optimal relative improvement per generation

$$\mathrm{E}^*[P_{1,\lambda}] \quad \sim \quad \frac{\ln \lambda}{n} \quad \text{or} \quad \breve{\varphi}^*_{1,\lambda} \quad \sim \quad \ln \lambda \tag{7.12}$$

in case of large n and large λ. This result is approximately valid also for the $(1 + \lambda)$ ES under the same assumptions since for large λ the probability that all successors are worse than their predecessor diminishes towards zero. A numerical evaluation is presented in figures 7.2 and 7.3 (solid lines on the right hand sides for the $(1, \lambda)$ ES): Differences to the $(1+\lambda)$ ES obviously appear only if λ is chosen to be small, especially for $\lambda \leq 5$.

If the progeny of one generation can be generated and evaluated in parallel, then the approximation above helps to calculate the order of the speedup for a $(1,\lambda)$ ES over the merely sequential $(1+1)$ ES. Using

$$\beta := \mathrm{E}\left[\frac{r^{(t)} - r^{(t+1)}}{r^{(t)}}\right] = \begin{cases} \mathrm{E}[P_{1+1}] &=& 0.2025/n & \text{for the } (1+1) \text{ ES} \\ \mathrm{E}[P_{1,\lambda}] &=& \ln\lambda/n & \text{for the } (1,\lambda) \text{ ES}. \end{cases} \tag{7.13}$$

we find by induction from (7.13) the expected error in generation t

$$\mathrm{E}[r^{(t)}] \quad = \quad (1-\beta)r^{(t-1)} \quad = \quad (1-\beta)^t r^{(0)}.$$

A simple transformation yields

$$\mathrm{E}[r^{(t)}] \quad = \quad (1-\beta)^t r^{(0)} < \epsilon \Leftrightarrow t > \log_{1-\beta}(\epsilon/r^{(0)}),$$

and as expected number of steps that a (1+1)- or $(1,\lambda)$ ES needs for reaching an ϵ-vicinity of the optimum we get

$$\mathrm{E}[t_\epsilon] \quad = \quad \log_{1-\beta}(\epsilon/r^{(0)}).$$

The speedup can be calculated by means of a Taylor series for $\log(1-x)$ cut off after the linear term:

$$\begin{aligned} S_\lambda \quad := \quad \frac{\mathrm{E}[t_{\epsilon,1}]}{\mathrm{E}[t_{\epsilon,\lambda}]} \quad &= \quad \frac{\log_{(1-\beta_1)}(\epsilon/r^{(0)})}{\log_{(1-\beta_\lambda)}(\epsilon/r^{(0)})} \\ &= \quad \frac{\ln(\epsilon/r^{(0)})}{\ln(1-0.2025/n)} \frac{\ln(1-\ln\lambda/n)}{\ln(\epsilon/r^{(0)})} \\ &\approx \quad \frac{(\ln\lambda)/n}{0.2025/n} \quad = \quad O(\ln\lambda). \end{aligned}$$

Thus, a $(1,\lambda)$ ES provides an expected speedup that is only logarithmic in λ. The same is true for a $(1+\lambda)$ ES since (7.12) is valid for large λ in both cases.

7.3.5 The (μ,λ) evolution strategy

In order to achieve a higher speedup one has to increase the number of survivors during the selection process. Only at the first sight this reduces the convergence velocity. Recently it has been shown, at least approximately [Bey94], that equation (7.12) has to be rewritten for a (μ,λ) ES as

$$\check{\varphi}^*_{\mu,\lambda} \quad \sim \quad \ln\frac{\lambda}{\mu}.$$

A numerical evaluation is presented within figures 7.2 and 7.3: Especially figure 7.3 demonstrates that the ratio $\frac{\lambda}{\mu}$ dominates the speed of convergence. Increasing the population size beyond $\mu = 10$ does not help in case of no recombination.

Nature's trick for a speedup lies in the recombination itself. We shall not dive into the rather lengthly calculation, but just provide the result from [Bey94] here. There is no difference in the expected maximum of the convergence velocity between two kinds

of recombination, i.e., intermediary recombination of all μ predecessors and global discrete recombination:

$$\breve{\varphi}^*_{\mu,\kappa=1,\lambda,\rho=\mu} \quad \sim \quad \mu \ln \frac{\lambda}{\mu}$$

The mechanisms of the two types of recombination being quite different from each other, it is not astonishing that the corresponding optimal mutation step sizes are different as well:

$$\sigma^* = \begin{cases} \sqrt{2\varphi^*} & \text{for uniform crossover} \\ \sqrt{2\mu\varphi^*} & \text{for global intermediary multirecombination.} \end{cases} \tag{7.14}$$

With respect to a theory of the autoadaptation to those optimal step sizes we refer to recent results of Beyer [Bey95].

References

[BRS93] T. Bäck, G. Rudolph and H.–P. Schwefel, Evolutionary Programming and Evolution strategies: Similarities and Differences, in: D. B. Fogel and W. Atmar, editors, *Proceedings of the Second Annual Conference on Evolutionary Programming*, San Diego, Feb. 25-26, 1993 Evolutionary Programming Society, La Jolla CA, pp. 11–22.

[Bey94] H.–G. Beyer, *Towards a theory of 'evolution strategies'—results from the N-dependent* (μ, λ) *and the multi-recombinant* $(\mu/\mu, \lambda)$ *theory*, technical report SYS-5/94, Systems Analysis Research Group, University of Dortmund, Department of Computer Science, Oct. 1994.

[Bey95] H.–G. Beyer, *Towards a theory of 'evolution strategies': The* $(1, \lambda)$*-self-adaptation*, technical report SYS-1/95, Systems Analysis Research Group, University of Dortmund, Department of Computer Science, May 1995.

[Box57] G. E. P. Box, Evolutionary operation—a method for increasing industrial productivity, Appl. Stat. **6**(1957), 81–101.

[Box65] M. J. Box, A new method of constrained optimization and a comparison with other methods, Comp. J. **8**(1965), 42–52.

[NM65] J. A. Nelder and R. Mead, A simplex method for function minimization, Comp. J. **7**(1965), 308–313.

[Rap84] G. Rappl, *Konvergenzraten von Random Search Verfahren zur globalen Optimierung*, PhD Thesis, Hochschule der Bundeswehr, Munich, 1984.

[Rec94] I. Rechenberg, *Evolutionsstrategie '94*, Frommann-Holzboog, Stuttgart, 1994.

[Rud94] G. Rudolph, Convergence of non-elitist strategies, in: Z. Michalewicz, J. D. Schaffer, H.–P. Schwefel, and D. B. Fogel, editors, *Proceedings of the 1st IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence*, Orlando FL, June 27-29, 1994, vol. 1, pp. 63–66.

[Schw95] H.–P. Schwefel, *Evolution and Optimum Seeking*, Wiley, New York, 1995.

Genetic Algorithms in Engineering and Computer Science
Editor J. Périaux and G. Winter