Toward a Theory of Evolution Strategies: The (μ, λ)-Theory

Hans-Georg Beyer,
University of Dortmund,
Department of Computer Science,
Systems Analysis Research Group,
D-44221 Dortmund,
Germany

beyer@LS11.informatik.uni-dortmund.de

Abstract
The multi-membered Evolution Strategy (ES) acting on μ parents and λ offspring is analyzed for real-valued, N-dimensional parameter spaces ($N \gtrapprox 30$). N-dependent progress rate formulas are derived for $(1, \lambda)$ and (μ, λ) strategies on spherical models. The analytical results obtained are compared with simulation experiments for the (hyper)sphere and the inclined (hyper)plane.

Keywords
Evolution Strategy(ES), progress rate theory, multi-membered ES, (μ, λ)-selection, truncation selection

1 Introduction

State-of-the-art Evolution Strategies (ES) for parameter optimization are multi-membered strategies acting on one or more populations of individuals. This paper deals with strategies denoted by (μ, λ), that is, the best μ offspring of a population (consisting of λ individuals) are selected to be the parents for the next generation. Such selection schemes are sometimes called 'truncation selection' (Mühlenbein & Schlierkamp-Voosen, 1993). Recombination is not included in this paper. It will be treated in (Beyer, 1995).

The main goal of a theoretical analysis is the determination of the progress rate φ depending on the mutation strength σ.[1] Such an analysis is difficult and depends on the functional structure of the fitness function $Q(\mathbf{y})$ to be optimized.

In (Beyer, 1993) the author developed the asymptotic $(1, \lambda)$ theory for the N-dimensional sphere. It is an intention of this paper to continue that work in several directions.

First, in section 2 the normalized progress formula of the $(1, \lambda)$-ES will be re-visited. Numerical experiments have indicated that (even) the normalized progress rate $\varphi^\star$ depends on the

[1] Often referred to as *mutation width* or *mutation step size*, however, the standard deviation of a single component of the mutation vector is actually meant.

dimension N of the parameter space. I. e., the quadratic progress formula (see, for example (Rechenberg, 1989)) holds for small mutations ($\sigma \to 0$) only. A theory will be developed that takes the N-dependence of the progress rate $\varphi^\star$ into account.

Section 3 is devoted to the (μ, λ)-ES. In contrast to $(1, \lambda)$ strategies, $\mu > 1$ indicates that there is a population of μ parents evolving in the parameter space. A progress rate theory has to consider the *self-consistent distribution* of the parents - a rather difficult task.

The central problem to be solved is the determination of the distribution of a single trial (in this paper often referred to as *offspring distribution*). Such a single trial (offspring) is produced by randomly sampling the μ parents and applying the mutation operator. The single trial distribution is connected to the parental distribution by a nonlinear functional equation not explicitly derived in this paper - an approximation is used instead. Even this sophisticated approach using Hermite polynomials for the series expansion of the distributions is rather long. In this article only a sketch of the derivation can be given covering the key ideas and techniques used. More details can be found in (Beyer, 1994b).

Notations used in this paper are following the conventions used in (Beyer, 1993). Throughout this paper φ denotes the progress rate (in (Beyer, 1993) E has been used). The fitness function F depends on N parameters y_i to be optimized

$$F(\mathbf{y}) = F(y_1, y_2, \dots y_N). \tag{1}$$

It is assumed that the optimum point is at

$$\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots \hat{y}_N)^T.$$

By changing the coordinate system and introduction of the distance-to-optimum vector $\mathbf{R}$

$$\mathbf{y} := \hat{\mathbf{y}} + \mathbf{R}$$

Equation (1) can be written as $F(\mathbf{y}) = F(\hat{\mathbf{y}} + \mathbf{R}) =: Q(\mathbf{R})$. Due to the spherical model $Q(\mathbf{R})$ is a function depending on the length $R = \|\mathbf{R}\|$ only, that is, $Q = Q(R)$.
Mutations $\mathbf{Z} = (z_1, z_2, \dots z_N)^T$ are produced by independent random numbers z_i

$$\tilde{\mathbf{y}} = \mathbf{y} + \mathbf{Z}. \tag{2}$$

All components z_i of $\mathbf{Z}$ are independently, identically normally distributed (iid normal) with the 'mutation strength' σ, i. e. the density for a component $z := z_i$ reads

$$p(z) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{z}{\sigma}\right)^2}. \tag{3}$$

Further definitions will be given in the text.

2 The $(1, \lambda)$ Theory Re-Visited

In this section the N-dependent theory for the $(1, \lambda)$ spherical model will be developed. It is presupposed that the reader is familiar with the algorithm, the theoretical concepts of the progress rate, and the spherical model (for details see (Beyer, 1993)).

It is assumed that the parent $\mathbf{y}$ has a distance R to the optimum $\hat{\mathbf{y}}$. By mutations λ offspring $\tilde{\mathbf{y}}_l$ are generated (according to (2)) with distance $\tilde{R}_l$ $(l = 1, \dots \lambda)$ to $\hat{\mathbf{y}}$. The $\tilde{R}_l$ are samplings from a r-distribution $p(r)$ to be determined in 2. 1. Knowing the $p(r)$ distribution one can

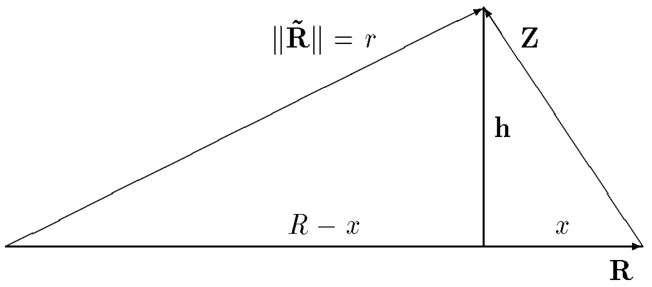

Figure 1: Decomposition of $\mathbf{Z}$ in x and $\mathbf{h}$.

calculate the distribution $p_{1,\lambda}(r)$ of the best offspring, i. e. that $\tilde{\mathbf{y}}_l$ with the smallest $\tilde{R}$. The density $p_{1,\lambda}(r)$ - a special case of order statistics - reads

$$p_{1,\lambda}(r) = p_{1:\lambda}(r) = \lambda\, p(r) \left[P(\tilde{R} > r)\right]^{\lambda-1}. \tag{4}$$

The order statistics notation $m : \lambda$ (Arnold, Balakrishnan, & Nagaraja, 1992) is used indicating the ordering of the λ observations $\tilde{R}_l$ in nondecreasing order

$$\tilde{R}_{1:\lambda} \le \tilde{R}_{2:\lambda} \le \ldots \le \tilde{R}_{m:\lambda} \le \ldots \le \tilde{R}_{\lambda:\lambda}. \tag{5}$$

The progress rate φ can be expressed as the expected r-change from R to $\tilde{R}_{1:\lambda}$

$$\varphi_{1,\lambda} = E\{R - \tilde{R}_{1:\lambda}\} = \int_0^\infty (R - r)\, p_{1,\lambda}(r)\, dr. \tag{6}$$

2.1 The $p(r)$-Density

Due to the spherical symmetry of the joint density of the mutation vectors $\mathbf{Z}$ (cp. (Beyer, 1993), p. 168) a mutation can be decomposed into a part x parallel to $-\mathbf{R}$ and a perpendicular part $\mathbf{h}$. Fig. 2.1 shows this decomposition

$$\mathbf{Z} = -x\mathbf{e}_R + \mathbf{h}, \qquad \mathbf{e}_R := \frac{\mathbf{R}}{||\mathbf{R}||}.$$

For the r^2-variate elementary geometry delivers

$$r^2 = \mathbf{h}^2 + (R - x)^2 = \mathbf{h}^2 + R^2 - 2Rx + x^2 \approx \mathbf{h}^2 + R^2 - 2Rx, \tag{7}$$

i. e. r^2 is constituted from the independent variates $\mathbf{h}^2$ and $2Rx$ (neglecting x^2, because $x^2 \ll \mathbf{h}^2$ holds). Since x is $N(0, \sigma)$ normally distributed (eq. (3) and the rotational symmetry of $p(\mathbf{Z})$) the variate $2Rx$ is $N(0, 2R\sigma)$ distributed.[2] $\mathbf{h}^2$ as a sum of $(N-1)$ iid z_i^2 (each z_i has distribution (3)) can be well approximated (if $N \gtrapprox 30$) by a normal distribution with mean

$$\overline{\mathbf{h}^2} = \sigma^2(N-1) \approx N\sigma^2 \tag{8}$$

and standard deviation

$$D\{\mathbf{h}^2\} = \sqrt{\overline{(\mathbf{h}^2)^2} - \left(\overline{\mathbf{h}^2}\right)^2} = \sqrt{2(N-1)\,\sigma^2} \approx \sqrt{2N}\,\sigma^2. \tag{9}$$

[2] R is a constant.

This is due to the central limit theorem of statistics (cp. e. g. (Fisz, 1971)). Thus, we have for the independent variates $v_1 := 2Rx$ and $v_2 := \mathbf{h}^2$

$$p_{2Rx}(v_1) = \frac{1}{\sqrt{2\pi}\, 2R\sigma} e^{-\frac{1}{2}\left(\frac{v_1}{2R\sigma}\right)^2}, \quad p_{h^2}(v_2) = \frac{1}{\sqrt{2\pi}\,\sqrt{2N}\,\sigma^2} e^{-\frac{1}{2}\left(\frac{v_2 - N\sigma^2}{\sqrt{2N}\,\sigma^2}\right)^2}. \tag{10}$$

As a sum over the normal variates v_1, v_2 and a constant R^2 (cp. eq. 7) r^2 itself is normally distributed with mean

$$\overline{r^2} = R^2 + N\sigma^2$$

and standard deviation $D = \sqrt{D^2\{\mathbf{h}^2\} + D^2\{2Rx\}}$

$$D = \sqrt{2N\sigma^4 + 4R^2\sigma^2} = 2\sigma R\sqrt{1 + \frac{\sigma^2 N}{2R^2}}.$$

Now, we can write down the cdf (cumulative distribution function)

$$P(\tilde{R}^2 \le r^2) = \frac{1}{2} + \Phi_0\left(\frac{r^2 - R^2 - \sigma^2 N}{2\sigma R\sqrt{1 + \frac{\sigma^2 N}{2R^2}}}\right), \tag{11}$$

with the Gauß-integral

$$\Phi_0(x) := \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}t^2}\, dt = \frac{1}{2}\, erf\left(\frac{x}{\sqrt{2}}\right). \tag{12}$$

Because

$$P(\tilde{R}^2 \le r^2) = P(\tilde{R} \le r) \quad \text{and} \quad p(r) = \frac{d}{dr}\, P(\tilde{R} \le r)$$

we finally obtain the pdf (probability density function) of the r-variate

$$p(r) = \frac{r}{\sqrt{2\pi}\,\sigma R\sqrt{1 + \frac{\sigma^2 N}{2R^2}}}\ \exp -\frac{1}{2}\left(\frac{r^2 - R^2 - \sigma^2 N}{2\sigma R\sqrt{1 + \frac{\sigma^2 N}{2R^2}}}\right)^2. \tag{13}$$

2.2 Derivation of the Progress Rate φ

The progress rate $\varphi_{1,\lambda}$ is given by eq. (6). Combination of (4) and (6) gives

$$\varphi_{1,\lambda} \quad = \quad R \; - \; \lambda \int_0^\infty r\, p(r) \left[1 - P(\tilde{R} \le r)\right]^{\lambda - 1} dr. \tag{14}$$

The usage of (11, 13, 14) and the substitution $t = t(r)$

$$t = \left(\frac{r^2 - R^2 - \sigma^2 N}{2\sigma\sqrt{R^2 + \frac{\sigma^2 N}{2}}}\right) \qquad \Longrightarrow \qquad r = \sqrt{R^2 + \sigma^2 N}\ \sqrt{1 + 2\sigma\frac{\sqrt{R^2 + \sigma^2 N/2}}{R^2 + \sigma^2 N}t}$$

yields

$$\varphi_{1,\lambda} = R \; - \; \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{R^2 + \sigma^2 N}\ \sqrt{1 + 2\sigma\frac{\sqrt{R^2 + \sigma^2 N/2}}{R^2 + \sigma^2 N}t}\ \ e^{-\frac{1}{2}t^2} \left[\frac{1}{2} - \Phi_0(t)\right]^{\lambda - 1} dt. \tag{15}$$

It should be noted that the lower limit $t = t(0)$ has been extended to $-\infty$, because the integrand is concentrated about $t = 0$. For the same reason the second term in the square root of the integrand can be assumed to be small compared to 1. Therefore, a good approximation is obtained by Taylor expansion breaking off after the linear term

$$\varphi_{1,\lambda} = R - \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left\{ \sqrt{R^2 + \sigma^2 N} + \sigma \sqrt{\frac{R^2 + \sigma^2 N/2}{R^2 + \sigma^2 N}}\, t \right\} e^{-\frac{1}{2}t^2} \left[\frac{1}{2} - \Phi_0(t)\right]^{\lambda-1} dt. \qquad (16)$$

This gives for (15)

$$\varphi_{1,\lambda} = R - R\sqrt{1 + \frac{\sigma^2 N}{R^2}} + \sigma \sqrt{\frac{R^2 + \sigma^2 N/2}{R^2 + \sigma^2 N}}\, \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t\, e^{-\frac{1}{2}t^2} \left[\frac{1}{2} + \Phi_0(t)\right]^{\lambda-1} dt. \qquad (17)$$

The remaining integral in (17) defines the well-known *progress coefficient* $c_{1,\lambda}$ (Beyer, 1993)

$$c_{1,\lambda} = \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t\, e^{-\frac{1}{2}t^2} \left[\frac{1}{2} + \Phi_0(t)\right]^{\lambda-1} dt. \qquad (18)$$

If changed to normalized values $\varphi^\star$ and $\sigma^\star$

$$\varphi^\star := \varphi \frac{N}{R} \qquad \text{and} \qquad \sigma^\star := \sigma \frac{N}{R} \qquad (19)$$

one finally obtains the *N-dependent progress rate*

$$\boxed{\varphi^\star_{1,\lambda}(\sigma^\star) = N\left(1 - \sqrt{1 + \frac{(\sigma^\star)^2}{N}}\right) + c_{1,\lambda}\,\sigma^\star \sqrt{\frac{1 + \frac{(\sigma^\star)^2}{2N}}{1 + \frac{(\sigma^\star)^2}{N}}}. \qquad (20)}$$

2.3 Comparison with Experiments

The predictions of (20) can be easily tested by experiments. From a fixed point $\mathbf{y}$ with distance R to the optimum $\hat{\mathbf{y}}$, G 'one generation' $(1,\lambda)$ experiments (each of them consisting of λ trials) are to perform with $\sigma = const.$ The actually measured progress values $(R - \tilde{R}_{1:\lambda})$ are averaged and normalized. As an example the $(1,8)$-ES, $N = 100$, has been tested with $G = 20,000$ 'one-generation' experiments per $\sigma^\star$-value. The results are depicted as dots in Fig. 2.

One observes a good agreement between the N-dependent theory (20) and the experiment. The dashed curves come from Rechenberg's formula (Rechenberg, 1989; Beyer, 1993)

$$\varphi^\star_{1,\lambda}(\sigma^\star) = -\frac{(\sigma^\star)^2}{2} + c_{1,\lambda}\sigma^\star. \qquad (21)$$

The progress coefficient (18) needed for (20) and (21) has been computed numerically: $c_{1,8} = 1.4236$. As can be seen (21) fails for normalized mutations greater than $\sigma^\star \approx 10$. Eq. (21) exhibits an asymptotically quadratic behavior whereas eq. (20) a linear one. For small $\sigma^\star$ values the results are comparable, and indeed, expanding the roots in (20) such that $(\sigma^\star)^2/N \ll 1$ holds[3] will recover (21). Therefore, the inequality

$$\frac{(\sigma^\star)^2}{N} < 1$$

[3]In his new book Rechenberg (1994, p. 23) explicitly emphasizes this condition for the applicability of the quadratic progress law (21).

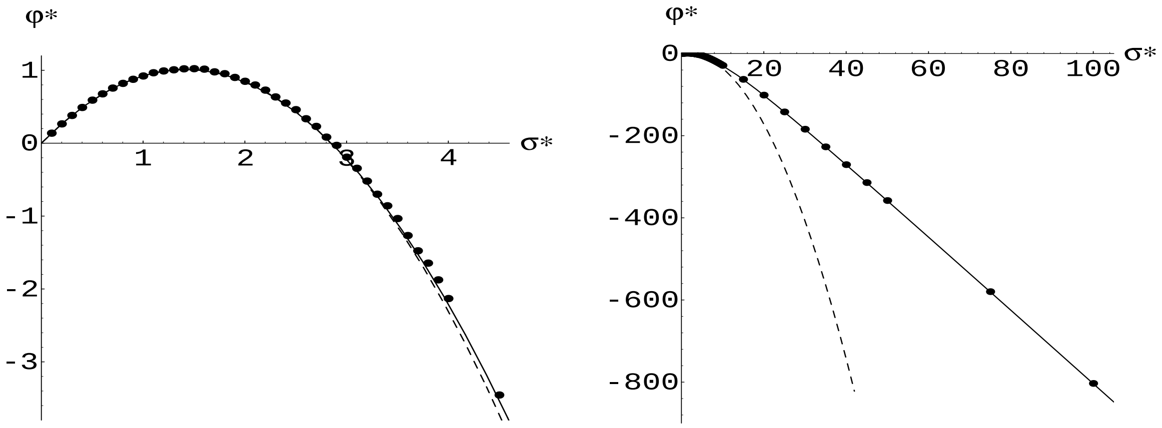

Figure 2: Comparison: experiment and theory for the $(1, 8)$-ES, $N = 100$. The dashed curves are due to the Rechenberg formula. Note, the standard deviations of the averages displayed are smaller than the diameter of the dots.

can serve as a condition for the validity of (21).

One might argue that (21) suffices for most of the applications, since (positive) progress is guaranteed for $\sigma^\star < 2\,c_{1,\lambda}$ only and $c_{1,\lambda}$ is smaller than 3.242 if $\lambda < 1{,}000$ ($c_{1,\lambda} \sim \sqrt{2\ln(\lambda)}$, cp. (Beyer, 1993)). However, this does not hold for *multi-recombinant* Evolution Strategies which is investigated in (Beyer, 1995) and for the *σ-self-adaptation* theory to be published in a forthcoming article.

2.4 The $p(r)$-Density Revised

In order to get manageable expressions for the (μ, λ) theory it will be worth using the simplest approximation of $p(r)$ which allows the derivation of (16). Instead of (13) we are wishing a normal approximation to $p(r)$. This can be obtained from (16) by back-substitution of the curly brace standing for r

$$t = \frac{r - \sqrt{R^2 + \sigma^2 N}}{\tilde{\sigma}} \qquad \text{with} \qquad \tilde{\sigma} := \sigma \sqrt{\frac{R^2 + \frac{\sigma^2 N}{2}}{R^2 + \sigma^2 N}}.$$

The progress rate integral reads

$$\varphi_{1,\lambda} = R - \lambda \int_0^\infty r \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \exp -\frac{1}{2}\left(\frac{r - \sqrt{R^2 + \sigma^2 N}}{\tilde{\sigma}}\right)^2 \left[\frac{1}{2} - \Phi_0\left(\frac{r - \sqrt{R^2 + \sigma^2 N}}{\tilde{\sigma}}\right)\right]^{\lambda - 1} dr$$

and the normal approximation for $p(r)$ becomes

$$p(r) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \exp -\frac{1}{2}\left(\frac{r - \sqrt{R^2 + \sigma^2 N}}{\tilde{\sigma}}\right)^2. \qquad (22)$$

3 The (μ, λ) Theory

3.1 Basic Aspects of the (μ, λ) Theory

3.1.1 The (μ, λ) Algorithm and the $\varphi_{\mu,\lambda}$ Definition

In (μ, λ) strategies (Schwefel, 1981) μ parents at different locations $\mathbf{y}_m$ $(m = 1, \ldots \mu)$ in the parameter space are to produce λ offspring $\tilde{\mathbf{y}}_l$ $(l = 1, \ldots \lambda)$. Each of the λ offspring is generated by random selection of one parent and mutation in accordance with equations 2 and 3 [4]

$$\tilde{\mathbf{y}}_l = \mathbf{y}_m + \mathbf{Z}, \qquad m = \mathtt{Random}\{1 \ldots \mu\}. \tag{23}$$

The distance of the parent $\mathbf{y}_m$ to the optimum may be given by R_m and by $\tilde{R}_l$ for the offspring l. The μ fittest offspring are selected obeying the ordering

$$\tilde{R}_{1:\lambda} \le \tilde{R}_{2:\lambda} \le \ldots \le \tilde{R}_{m:\lambda} \le \ldots \le \tilde{R}_{\mu:\lambda} \tag{24}$$

(cp. (5)). Now it is possible to define the average progress $\langle \Delta R \rangle_{\mu,\lambda}$

$$\langle \Delta R \rangle_{\mu,\lambda} := \langle R \rangle - \langle \tilde{R} \rangle_{\mu,\lambda} \tag{25}$$

with

$$\langle R \rangle := \frac{1}{\mu} \sum_{m=1}^{\mu} R_m \qquad \text{and} \qquad \langle \tilde{R} \rangle_{\mu,\lambda} := \frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{R}_{m:\lambda}. \tag{26}$$

Since the $\tilde{R}_{m:\lambda}$ are random variates the average is a random variate, too. Let $P_{\mu,\lambda}(r) = P(\langle \tilde{R} \rangle_{\mu,\lambda} \le r)$ be the cdf of this average and $p_{\mu,\lambda}(r) = \frac{d}{dr} P(\langle \tilde{R} \rangle_{\mu,\lambda} \le r)$ the pdf. Then, it is possible to define the progress rate φ as the expectation of (25)

$$\varphi_{\mu,\lambda} = E\{\langle \Delta R \rangle_{\mu,\lambda}\} = \langle R \rangle - \int_0^\infty r \, p_{\mu,\lambda}(r) \, dr. \tag{27}$$

Due to the spherical symmetry the problem reduces to a one-dimensional one. It is interesting to note that there is no selection pressure for components independent from r. Therefore, the population can perform a random walk on the surface of the hypersphere. At least for the sphere, *the (μ, λ)-ES is a really explorative strategy.*

3.1.2 The Progress Rate Integral

Eq. (27) can be re-formulated using the second expression from (26)

$$\begin{aligned} \varphi_{\mu,\lambda} &= \langle R \rangle - \frac{1}{\mu} \sum_{m=1}^{\mu} E\{\tilde{R}_{m:\lambda}\} \\ \varphi_{\mu,\lambda} &= \langle R \rangle - \frac{1}{\mu} \sum_{m=1}^{\mu} \int_0^\infty r \, p_{m:\lambda}(r) \, dr. \end{aligned} \tag{28}$$

$p_{m:\lambda}(r)$ is the pdf of the mth order statistics. Let $p(r)$ be the pdf and $P(r)$ the cdf of a single trial (i. e. the density of an offspring produced by (23)), then $p_{m:\lambda}(r)$ reads

$$p_{m:\lambda}(r) = \frac{\lambda!}{(m-1)!\,(\lambda - m)!} \, p(r) [P(r)]^{m-1} [1 - P(r)]^{\lambda - m}. \tag{29}$$

[4] The $\mathtt{Random}$ notation means that m is a random sample from the integer set $1 \le m \le \mu$. Each element of this set is picked up with the same probability $1/\mu$.

Eq. (29) can be easily understood by taking into account that the m-best trial is characterized by $p(r)$ AND $(m-1)$ trials smaller than r (i. e. $P(\tilde{R} \le r) =: P(r)$) AND $(\lambda - m)$ trials larger than r (i. e. $P(\tilde{R} > r) = 1 - P(r)$). The factor $\lambda!/((m-1)!(\lambda-m)!)$ counts all possibilities of arranging λ trials in such a way. With (29) the progress rate integral (28) reads

$$\varphi_{\mu,\lambda} = \langle R \rangle - \frac{\lambda}{\mu} \int_0^\infty r p(r) \sum_{m=1}^{\mu} \frac{(\lambda-1)!}{(m-1)!\,(\lambda-m)!} \, [P(r)]^{m-1} [1-P(r)]^{\lambda-m} dr. \tag{30}$$

For most derivations to come it will be convenient to express the sum in (30) by an integral

$$\begin{aligned}
&\sum_{m=1}^{\mu} \frac{(\lambda-1)!}{(m-1)!\,(\lambda-m)!} \, P^{m-1} [1-P]^{\lambda-m} \\
&\qquad = \sum_{k=0}^{\mu-1} \frac{(\lambda-1)!}{k!\,(\lambda-k-1)!} \, P^k [1-P]^{\lambda-k-1} \\
&\qquad = \sum_{k=0}^{\mu-1} \binom{\lambda-1}{k} P^k [1-P]^{\lambda-k-1} = 1 - \sum_{k=\mu}^{\lambda-1} \binom{\lambda-1}{k} P^k [1-P]^{\lambda-k-1} \\
&\qquad = 1 - I_P(\mu, \lambda-\mu) = I_{1-P}(\lambda-\mu, \mu).
\end{aligned}$$

Here we have used the binomial theorem and the relation to the regularized incomplete beta function $I_y(a,b)$ (Abramowitz & Stegun, 1984) which can be expressed by the integral

$$I_{1-P(r)}(\lambda-\mu, \mu) = \frac{(\lambda-1)!}{(\lambda-\mu-1)!(\mu-1)!} \int_0^{1-P(r)} x^{\lambda-\mu-1} (1-x)^{\mu-1} dx.$$

Thus,

$$\sum_{m=1}^{\mu} \frac{P^{m-1}[1-P]^{\lambda-m}}{(m-1)!\,(\lambda-m)!} = \frac{1}{(\lambda-\mu-1)!(\mu-1)!} \int_0^{1-P(r)} x^{\lambda-\mu-1} (1-x)^{\mu-1} dx \tag{31}$$

holds and (30) becomes

$$\varphi_{\mu,\lambda} = \langle R \rangle - \frac{\lambda!}{(\lambda-\mu-1)!\,\mu!} \int_{r=0}^{r=\infty} \int_{x=0}^{x=1-P(r)} r p(r) x^{\lambda-\mu-1} (1-x)^{\mu-1} dx\, dr. \tag{32}$$

By changing the order of integration one obtains

$$\varphi_{\mu,\lambda} = \langle R \rangle - (\lambda-\mu) \binom{\lambda}{\mu} \int_{x=0}^{x=1} \int_{r=0}^{r=P^{-1}(1-x)} r p(r) x^{\lambda-\mu-1} (1-x)^{\mu-1} dr\, dx,$$

with the inverse cdf P^{-1}. The substitution $x = 1 - P(t)$, $dx = -p(t)\,dt$, finally yields

$$\varphi_{\mu,\lambda} = \langle R \rangle - (\lambda-\mu) \binom{\lambda}{\mu} \int_{t=0}^{t=\infty} p(t) [1-P(t)]^{\lambda-\mu-1} [P(t)]^{\mu-1} \int_{r=0}^{r=t} r p(r)\, dr\, dt. \tag{33}$$

3.1.3 Formal Approximation of the Offspring Distribution $p(r)$

All problems concerning the (μ, λ) theory are concentrated in determining the $p(r)$ distribution needed for integral (33). $p(r)$ is the pdf of a single trial consisting of a randomly chosen parent and a mutation. It is obvious that $p(r)$ changes from generation to generation. Starting from a common 'forebear' $p(r)$ evolves successively to a steady state distribution. The question arises how to describe and to determine this distribution in a self-consistent manner. It is quite clear that there is no analytical solution to $p(r)$. Therefore one has to rely on approximations. A very natural technique is the expansion of $p(r)$ in a series of orthogonal Hermite polynomials He_k (Beyer, 1994a)

$$He_k(x) := (-1)^k e^{\frac{1}{2}x^2} \frac{d^k}{dx^k} \left(e^{-\frac{1}{2}x^2} \right) \tag{34}$$

for the standardized variate x

$$x := \frac{r - M}{D}, \tag{35}$$

where M is the mean of r

$$M := E\{r\} = \int_0^\infty r\, p(r)\, dr \tag{36}$$

and D the standard deviation

$$D^2 := E\{(r-M)^2\} = \int_0^\infty r^2 p(r)\, dr \; - \; M^2. \tag{37}$$

The expansion reads

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \sum_{k=0}^{\infty} \frac{a_k}{k!} He_k(x). \tag{38}$$

Due to the orthonormality relation the Fourier coefficients are determined by

$$a_k = \int_{-\infty}^{\infty} p(x) He_k(x)\, dx. \tag{39}$$

For the first four terms in (38) one easily finds

$$\begin{array}{ll} He_0(x) = 1, & a_0 = 1, \\ He_1(x) = x, & a_1 = 0, \\ He_2(x) = x^2 - 1, & a_2 = 0, \\ He_3(x) = x^3 - 3x, & a_3 = M_3/D^3, \end{array} \tag{40}$$

where M_3 is the third central moment of r

$$M_3 := \int_0^\infty (r - M)^3 p(r)\, dr. \tag{41}$$

Now it is possible to express $p(x)$ by means of (38) and (40). Because we are searching for an approximation of $p(r)$ we transform back to r via (35). The expansion up to the fourth term reads

$$p(r) = \frac{1}{\sqrt{2\pi}\, D} e^{-\frac{1}{2}\left(\frac{r-M}{D}\right)^2} \left[1 + \gamma\, He_3\left(\frac{r-M}{D} \right) \right], \qquad \gamma := \frac{1}{6} \frac{M_3}{D^3}. \tag{42}$$

For the cdf of the same order one obtains

$$P(r) = \frac{1}{2} + \Phi_0 \left(\frac{r-M}{D} \right) - \frac{\gamma}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{r-M}{D}\right)^2} He_2 \left(\frac{r-M}{D} \right) \tag{43}$$

as can be proved by differentiation. Higher order approximations can be obtained straightforwardly from (34 – 39) but will not be used in this paper because of the enormous computational effort in the remaining derivations.

The basic idea behind approximation (42) is that the steady state distribution will be almost Gaußian shaped with a slight skewness measured by γ. Due to the expansion (42) the problem of finding the steady state distribution is reduced to that of M, D, and γ.

3.1.4 Estimates for M, D, and γ

In order to bring life to the formal $p(r)$-approximation, the parameters M, D, and γ have to be derived from 'first principles'. That is, M, D, and γ are interpreted as statistical estimates from the *actual* offspring distribution $p(r)$. This distribution is constituted by the μ parents $\mathbf{y}_m$ with their actual distances R_m to the optimum $\hat{\mathbf{y}}$ and the mutations $\mathbf{Z}$. If a certain parent m is randomly chosen the conditional r-distribution $p(r|R_m)$ is given by (13) (spherical model!). Since this formula is rather difficult to handle we will use the simpler normal approximation (22) from section 2.4

$$p(r|R_m) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}_m} \exp -\frac{1}{2}\left(\frac{r - \sqrt{R_m^2 + \sigma^2 N}}{\tilde{\sigma}_m}\right)^2, \qquad \tilde{\sigma}_m := \sigma\sqrt{\frac{R_m^2 + \frac{\sigma^2 N}{2}}{R_m^2 + \sigma^2 N}}. \tag{44}$$

All parents are chosen with equal probability $1/\mu$, thus, the actual offspring pdf $p(r)$ is given by the mixture

$$p(r) = \frac{1}{\mu}\sum_{m=1}^{\mu} p(r|R_m). \tag{45}$$

We can now proceed with the calculation of the first moment M and further with D^2 and M_3. M is defined by (36) which gives with (44/45)

$$M = \int_0^\infty r\,\frac{1}{\mu}\sum_{m=1}^{\mu}\frac{1}{\sqrt{2\pi}\tilde{\sigma}_m}\exp -\frac{1}{2}\left(\frac{r - \sqrt{R_m^2 + \sigma^2 N}}{\tilde{\sigma}_m}\right)^2 dr.$$

After the substitution $r = \tilde{\sigma}_m t + \sqrt{R_m^2 + \sigma^2 N}$ one obtains

$$M = \frac{1}{\mu}\sum_{m=1}^{\mu}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left(\tilde{\sigma}_m t + \sqrt{R_m^2 + \sigma^2 N}\right) e^{-\frac{1}{2}t^2}\, dt.$$

Here we have extended the lower limit to $-\infty$ bringing the same arguments into play as in the case of the $(1, \lambda)$ integral (15). The result is

$$M = \frac{1}{\mu}\sum_{m=1}^{\mu}\sqrt{R_m^2 + \sigma^2 N}. \tag{46}$$

For the central moments M_k of order k

$$M_k = \int_0^\infty (r - M)^k p(r)\, dr$$

one finds by a similar calculation

$$M_k = \frac{1}{\mu}\sum_{m=1}^{\mu}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[\tilde{\sigma}_m t + \left(\sqrt{R_m^2 + \sigma^2 N} - M\right)\right]^k e^{-\frac{1}{2}t^2}\, dt.$$

This gives for M_2

$$D^2 = M_2 = \frac{1}{\mu}\sum_{m=1}^{\mu} \tilde{\sigma}_m^2 + \frac{1}{\mu}\sum_{m=1}^{\mu}\left(\sqrt{R_m^2+\sigma^2 N} - M\right)^2 \qquad (47)$$

and for M_3

$$M_3 = \frac{3}{\mu}\sum_{m=1}^{\mu} \tilde{\sigma}_m^2 \left(\sqrt{R_m^2+\sigma^2 N} - M\right) + \frac{1}{\mu}\sum_{m=1}^{\mu}\left(\sqrt{R_m^2+\sigma^2 N} - M\right)^3. \qquad (48)$$

At this stage we are able to express M, $D = \sqrt{M_2}$, and $\gamma = M_3 D^{-3}/6$ by the R_m. However, these R_m are random variates themselves produced by the order statistics of the $p(r)$ from the preceding generation (M, M_2, and M_3 are *conditional* moments of the order statistics)! Therefore, one has to calculate the expectancies $\overline{M}$, $\overline{D} = \overline{\sqrt{M_2}}$, and $\overline{\gamma} = \overline{M_3 D^{-3}}/6$ using the order statistics with $p(r)$ from the preceding generation. Because M, D, and γ are functions $f(R_1, R_2, \dots R_\mu)$ of the first μ order statistics the expectancies $\overline{f}$ are given by μ-fold integrals over the ordered domain satisfying $R_1 \le R_2 \le \dots \le R_\mu$

$$\overline{f} = \underset{R_1 \,\le\, R_2 \dots \le\, R_\mu}{\int\int\cdots\int} f(R_1, R_2, \dots R_\mu)\, p_o(R_1, R_2, \dots R_\mu)\, dR_\mu \cdots dR_1 \qquad (49)$$

with the joint density p_o

$$p_o(R_1, R_2, \dots R_\mu) = \frac{\lambda!}{(\lambda-\mu)!}\,[1 - P(R_\mu)]^{\lambda-\mu} \prod_{m=1}^{\mu} p(R_m). \qquad (50)$$

This cries for further simplifications. First, the sums of square roots containing R_m in (46, 47, 48) will be simplified. Secondly, in section 3.1.6, $\overline{D}$ and $\overline{\gamma}$ will be stochastically approximated.

3.1.5 Simplification of M, M_2, and M_3

The simplification of (46), (47), and (48) is mainly based on the assumption that the variates R_m are concentrated around some average $\langle R\rangle$ (cf. (26)). By defining the deviation ΔR_m

$$\Delta R_m := R_m - \langle R\rangle,$$

this assumption can be expressed by

$$|\Delta R_m| \ll \langle R\rangle \quad \Longrightarrow \quad \frac{|\Delta R_m|}{\langle R\rangle} \ll 1 \quad \Longrightarrow \quad \frac{R_m}{\langle R\rangle} = \frac{\Delta R_m + \langle R\rangle}{\langle R\rangle} \approx 1. \qquad (51)$$

For the treatment of M the square root in (46) is expanded in a Taylor series

$$\begin{aligned}
\sqrt{R_m^2+\sigma^2 N} &= \sqrt{(\langle R\rangle + \Delta R_m)^2 + \sigma^2 N} \\
&= \langle R\rangle \sqrt{1 + \frac{\sigma^2 N}{\langle R\rangle^2} + 2\left(\frac{\Delta R_m}{\langle R\rangle}\right) + \left(\frac{\Delta R_m}{\langle R\rangle}\right)^2} \\
&= \langle R\rangle \left[\sqrt{1+\frac{\sigma^2 N}{\langle R\rangle^2}} + \frac{1}{\sqrt{1+\frac{\sigma^2 N}{\langle R\rangle^2}}}\left(\frac{\Delta R_m}{\langle R\rangle}\right)\right. \\
&\quad \left. + \frac{1}{2}\frac{\frac{\sigma^2 N}{\langle R\rangle^2}}{\left(\sqrt{1+\frac{\sigma^2 N}{\langle R\rangle^2}}\right)^3}\left(\frac{\Delta R_m}{\langle R\rangle}\right)^2 - O\left(\left(\frac{\Delta R_m}{\langle R\rangle}\right)^3\right)\right].
\end{aligned} \qquad (52)$$

One obtains for (46)

$$M = \frac{1}{\mu} \sum_{m=1}^{\mu} \sqrt{R_m^2 + \sigma^2 N} \approx \langle R \rangle \sqrt{1 + \frac{\sigma^2 N}{\langle R \rangle^2}}. \tag{53}$$

Note, the linear term of (52) vanishes because

$$\sum_{m=1}^{\mu} \Delta R_m = 0 \tag{54}$$

holds.
Analogously to the $(1, \lambda)$ theory (cp. 19) normalized quantities will be introduced

$$\varphi^{\star}_{\mu,\lambda} := \varphi_{\mu,\lambda} \frac{N}{\langle R \rangle} \qquad \text{and} \qquad \sigma^{\star} := \sigma \frac{N}{\langle R \rangle}. \tag{55}$$

This is possible because $\langle R \rangle$, actually a random variate, fluctuates only slightly about its expectancy $\overline{\langle R \rangle}$. The fluctuations are of the order $\frac{1}{\sqrt{\mu}} \frac{\Delta R_m}{\langle R \rangle}$, therefore $\langle R \rangle$ can be treated as a constant.[5]
Using (55) and the abbreviation ζ one gets from (53)

$$M = \langle R \rangle \, \zeta \qquad \text{with} \qquad \zeta := \sqrt{1 + \frac{(\sigma^{\star})^2}{N}}. \tag{56}$$

The simplification of M_2, eq. (47), needs two steps. First, $\tilde{\sigma}_m^2$ will be tackled. By setting $x = x_0 + h = \langle R \rangle + \Delta R_m$ and $a = \sigma^2 N/2$ in (44) $\tilde{\sigma}_m^2$ is expanded in a Taylor series about $x_0 = \langle R \rangle$

$$\begin{aligned}
\frac{\tilde{\sigma}_m^2}{\sigma^2} &= \frac{x^2 + a}{x^2 + 2a} = 1 - \frac{a}{x^2 + 2a} \\
&= \frac{x_0^2 + a}{x_0^2 + 2a} + \frac{2 x_0 a}{(x_0^2 + 2a)^2} h - \frac{3 a x_0^2 - 2a^2}{(x_0^2 + 2a)^3} h^2 + \cdots \\
&= \frac{1 + \frac{\sigma^2 N}{2\langle R \rangle^2}}{1 + \frac{\sigma^2 N}{\langle R \rangle^2}} + \frac{\frac{\sigma^2 N}{\langle R \rangle^2}}{\left(1 + \frac{\sigma^2 N}{\langle R \rangle^2}\right)^2} \left(\frac{\Delta R_m}{\langle R \rangle} \right) - \frac{1}{2} \frac{3 \frac{\sigma^2 N}{\langle R \rangle^2} - \left(\frac{\sigma^2 N}{\langle R \rangle^2}\right)^2}{\left(1 + \frac{\sigma^2 N}{\langle R \rangle^2}\right)^3} \left(\frac{\Delta R_m}{\langle R \rangle} \right)^2 + \cdots .
\end{aligned} \tag{57}$$

The sum over $\tilde{\sigma}_m^2$ reads with (54)

$$\frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\sigma}_m^2 = \sigma^2 \frac{1 + \frac{\sigma^2 N}{2\langle R \rangle^2}}{1 + \frac{\sigma^2 N}{\langle R \rangle^2}} - \sum_m O\left(\left(\frac{\Delta R_m}{\langle R \rangle} \right)^2 \right). \tag{58}$$

For further derivations (58) will be written as

$$\frac{1}{\mu} \sum_{m=1}^{\mu} \tilde{\sigma}_m^2 = \tilde{\sigma}^2 \quad \text{with} \quad \tilde{\sigma} = \sigma \sqrt{\frac{1 + \frac{\sigma^2 N}{2\langle R \rangle^2}}{1 + \frac{\sigma^2 N}{\langle R \rangle^2}}} = \sigma \sqrt{\frac{1 + \frac{(\sigma^{\star})^2}{2N}}{1 + \frac{(\sigma^{\star})^2}{N}}}. \tag{59}$$

[5]Due to (53) this holds for M, $\overline{M}$ as well.

The parentheses in the second sum of (47) and (48), respectively, can be written with (52/53/56)

$$\begin{aligned}\sqrt{R_m^2+\sigma^2 N}-M &= \frac{\Delta R_m}{\zeta}\left(1+\frac{\zeta^2-1}{2\zeta^2}\left(\frac{\Delta R_m}{\langle R\rangle}\right)\right)\\ &\qquad -\frac{\zeta^2-1}{2\zeta^3}\,\frac{1}{\mu}\sum_{m'=1}^{\mu}\Delta R_{m'}\left(\frac{\Delta R_{m'}}{\langle R\rangle}\right)+\cdots \end{aligned} \tag{60}$$

Taking the assumption (51) into account, (60) reduces to

$$\sqrt{R_m^2+\sigma^2 N}-M=\frac{\Delta R_m}{\zeta}. \tag{61}$$

Thus, we have

$$\frac{1}{\mu}\sum_{m=1}^{\mu}\left(\sqrt{R_m^2+\sigma^2 N}-M\right)^k=\frac{1}{\zeta^k}\frac{1}{\mu}\sum_{m=1}^{\mu}(\Delta R_m)^k=:\frac{1}{\zeta^k}\langle(\Delta R)^k\rangle.$$

This gives for (47) with (58)

$$\begin{aligned} M_2 &= \tilde{\sigma}^2+\frac{1}{\zeta^2}\frac{1}{\mu}\sum_{m=1}^{\mu}(R_m-\langle R\rangle)^2\\ M_2 &= \tilde{\sigma}^2+\frac{1}{\zeta^2}\left[\frac{1}{\mu}\sum_{m=1}^{\mu}R_m^2-\left(\frac{1}{\mu}\sum_{m=1}^{\mu}R_m\right)^2\right]. \end{aligned} \tag{62}$$

Similarly, we find for the second sum of (48)

$$\begin{aligned}\frac{1}{\mu}\sum_{m=1}^{\mu}\left(\sqrt{R_m^2+\sigma^2 N}-M\right)^3 &= \frac{1}{\zeta^3}\frac{1}{\mu}\sum_{m=1}^{\mu}(R_m-\langle R\rangle)^3\\ &= \frac{1}{\zeta^3}\frac{1}{\mu}\sum_{m=1}^{\mu}\left(R_m^3-3R_m^2\langle R\rangle+3R_m\langle R\rangle^2-\langle R\rangle^3\right)\\ &= \frac{1}{\zeta^3}\left[\frac{1}{\mu}\sum_{m=1}^{\mu}R_m^3-3\left(\frac{1}{\mu}\sum_{m=1}^{\mu}R_m^2\right)\left(\frac{1}{\mu}\sum_{m=1}^{\mu}R_m\right)+2\left(\frac{1}{\mu}\sum_{m=1}^{\mu}R_m\right)^3\right].\end{aligned}$$

It is interesting to note, that the first sum in (48) vanishes due to the approximations (51, 57, 61) used

$$\sum_{m=1}^{\mu}\tilde{\sigma}_m^2\left(\sqrt{R_m^2+\sigma^2 N}-M\right)=\sum_{m=1}^{\mu}\left(\tilde{\sigma}^2+O\left(\frac{\Delta R_m}{\langle R\rangle}\right)\right)\frac{\Delta R_m}{\zeta}\approx\sum_{m=1}^{\mu}\tilde{\sigma}^2\frac{\Delta R_m}{\zeta}=0.$$

Thus, we finally obtain

$$M_3=\frac{1}{\zeta^3}\left[\frac{1}{\mu}\sum_{m=1}^{\mu}R_m^3-3\left(\frac{1}{\mu}\sum_{m=1}^{\mu}R_m^2\right)\left(\frac{1}{\mu}\sum_{m=1}^{\mu}R_m\right)+2\left(\frac{1}{\mu}\sum_{m=1}^{\mu}R_m\right)^3\right]=\frac{1}{\zeta^3}\langle(\Delta R)^3\rangle. \tag{63}$$

3.1.6 Stochastical Approximation for $\overline{D}$ and $\overline{\gamma}$

D is given by the square root of M_2, eq. (62). If one interprets the square brackets in (62) as a random variate x

$$x := \frac{1}{\mu}\sum_{m=1}^{\mu} R_m^2 - \left(\frac{1}{\mu}\sum_{m=1}^{\mu} R_m\right)^2 = \langle(\Delta R)^2\rangle \tag{64}$$

that fluctuates only slightly about $\overline{x}$, then D can be expanded in a Taylor series

$$D = \sqrt{\tilde{\sigma}^2 + \frac{\overline{x}}{\zeta^2} + \frac{x-\overline{x}}{\zeta^2}} = \sqrt{\tilde{\sigma}^2 + \frac{\overline{x}}{\zeta^2}} \;+\; \frac{1}{2\zeta^2}\frac{x-\overline{x}}{\sqrt{\tilde{\sigma}^2+\frac{\overline{x}}{\zeta^2}}} \;-\; \frac{1}{8\zeta^4}\frac{(x-\overline{x})^2}{\left(\sqrt{\tilde{\sigma}^2+\frac{\overline{x}}{\zeta^2}}\right)^3} \;+\cdots.$$

Hence, the expected value $\overline{D}$ reads

$$\overline{D} \approx \sqrt{\tilde{\sigma}^2 + \frac{\overline{x}}{\zeta^2}} \;-\; \frac{1}{8\zeta^4}\frac{\overline{(x-\overline{x})^2}}{\left(\sqrt{\tilde{\sigma}^2+\frac{\overline{x}}{\zeta^2}}\right)^3}\,.$$

$\overline{x}$ and $\overline{(x-\overline{x})^2}$ are the expectations of the order statistics defined by (49, 50). Even this approximation is still to difficult for further operation. We therefore make the radical - and somehow cruel - simplification

$$\overline{D} = \sqrt{\tilde{\sigma}^2 + \frac{\overline{\langle(\Delta R)^2\rangle}}{\zeta^2}}. \tag{65}$$

Thus, the determination of $\overline{D}$ is mainly concentrated on that of

$$\overline{\langle(\Delta R)^2\rangle} = \frac{1}{\mu}\sum_{m=1}^{\mu}\overline{R_m^2} - \frac{1}{\mu^2}\overline{\left(\sum R_m\right)^2}. \tag{66}$$

For the calculation of $\overline{\gamma}$ we proceed in the same manner. From the definition (42) it follows

$$\overline{\gamma} = \frac{1}{6}\,\overline{M_3/D^3}.$$

The simplest stochastical approximation is therefore

$$\overline{\gamma} = \frac{1}{6}\frac{\overline{M_3}}{\left(\overline{D}\right)^3} = \frac{1}{6}\frac{1}{\zeta^3}\frac{\overline{\langle(\Delta R)^3\rangle}}{\left(\overline{D}\right)^3} \tag{67}$$

with (cp. (63))

$$\overline{\langle(\Delta R)^3\rangle} = \frac{1}{\mu}\sum\overline{R_m^3} \;-\; 3\,\overline{\left(\frac{1}{\mu}\sum R_m^2\right)\left(\frac{1}{\mu}\sum R_m\right)} \;+\; 2\,\overline{\left(\frac{1}{\mu}\sum R_m\right)^3}. \tag{68}$$

3.1.7 Approaching the Steady-State - the Self-Consistent Method

Assume, the population is observed at generation $(g+1)$. The expected offspring distribution may be $p^{(g+1)}(r)$ approximated by (42)

$$p^{(g+1)}(r) = \frac{1}{\sqrt{2\pi}\,\overline{D}^{(g+1)}}\;\exp -\frac{1}{2}\left(\frac{r-M^{(g+1)}}{\overline{D}^{(g+1)}}\right)^2\;\left[1+\overline{\gamma}^{(g+1)}He_3\left(\frac{r-M^{(g+1)}}{\overline{D}^{(g+1)}}\right)\right]. \tag{69}$$

$M^{(g+1)}$, $\overline{D}^{(g+1)}$, and $\overline{\gamma}^{(g+1)}$ are functionals of the distribution $p^{(g)}(r)$ at generation (g) (because of (56), (65), and (67))

$$\begin{aligned} M^{(g+1)} &= \mathcal{F}_M\left[p(M^{(g)}, \overline{D}^{(g)}, \overline{\gamma}^{(g)})\right] , \\ \overline{D}^{(g+1)} &= \mathcal{F}_D\left[p(M^{(g)}, \overline{D}^{(g)}, \overline{\gamma}^{(g)})\right] , \\ \overline{\gamma}^{(g+1)} &= \mathcal{F}_\gamma\left[p(M^{(g)}, \overline{D}^{(g)}, \overline{\gamma}^{(g)})\right] . \end{aligned}$$

Approaching the steady-state ($g \to \infty$) one will expect that the 'shape' of $p(r)$ does not change anylonger.[6] Due to the approximation (69) this 'shape' is determined by $\overline{D}$ and $\overline{\gamma}$. Hence, for the steady state $\overline{D}^{(g+1)} = \overline{D}^{(g)} = \overline{D}$ and $\overline{\gamma}^{(g+1)} = \overline{\gamma}^{(g)} = \overline{\gamma}$ should hold. This leads to the *self-consistent equations*

$$\overline{D} = \mathcal{F}_D\left[p(\overline{D}, \overline{\gamma})\right] , \tag{70}$$

$$\overline{\gamma} = \mathcal{F}_\gamma\left[p(\overline{D}, \overline{\gamma})\right] \tag{71}$$

to be solved. Note, since $\overline{D}$ and $\overline{\gamma}$ are directly connected to the second and third central moments, respectively, they are independent from M (translation invariance). Therefore, for the solution of the nonlinear functional system (70/71) the *ansatz*

$$p(r) = \frac{1}{\sqrt{2\pi}\,\overline{D}} \exp\left(-\frac{1}{2}\left(\frac{r}{\overline{D}}\right)^2\right) \left[1 + \overline{\gamma} He_3\left(\frac{r}{\overline{D}}\right)\right] \tag{72}$$

suffices, whereas for the final $\varphi_{\mu,\lambda}$ calculation (33) the full form (69) with (56) will be used.

3.2 Implementation of the (μ, λ)-Theory

The theory developed so far connects M, $\overline{D}$, and $\overline{\gamma}$ to moments of sums of order statistics. A glance at the joint distribution (50) taking (42/43) into account reveals that the self-consistent equations (70/71) as well as the progress rate integral (33) are nonlinear in $\overline{\gamma}$. These nonlinearities are (luckily) multiplicative ones forming series with $\overline{\gamma}^k$. Since $\overline{\gamma}$ (cp. (67)) is a measure for the skewness of the offspring distribution we can expect that it will be sufficiently small. Indeed, for $\mu = 2$ and the normal approximation (44)

$$\mu = 2 \quad \Longrightarrow \quad \overline{\gamma} = 0$$

nearly holds. This is due to the fact that mixing two normal distributions with equal standard deviations will result in a symmetrical distribution (no skewness), hence $\overline{\gamma} = 0$. For the cases with more than two parents ($\mu > 2$) the skewness increases, but will still be small enough to yield usable results from the linearized theory, i. e. an approximation dropping all $\overline{\gamma}^\beta$ terms for $\beta > 1$. This has been tested by implementation of the nonlinear theory on a computer algebra system performing the integrations and the root finding by numerical routines. The differences are small compared to the approximation errors stemming from the series break off in (38/42/43) and the stochastical approximations (65/67). *Subsequently, we will always use the linearized theory.*

[6] Strictly speaking, there remains a slight change of the $p(r)$-shape from (g) to $(g+1)$ due to the necessity to adjust the mutation strength σ in order to obtain a maximum progress rate. However, this change is small and will be neglected here.

3.2.1 The Approximation of the Progress Rate Integral

The general expression for the progress rate $\varphi_{\mu,\lambda}$ is given by (33) with the $p(t)$ approximation (69)[7]

$$p(t) = \frac{1}{\sqrt{2\pi}\,D} e^{-\frac{1}{2}\left(\frac{t-M}{D}\right)^2} \left[1 + \gamma\, He_3\left(\frac{t-M}{D}\right)\right]. \tag{73}$$

The substitutions $(t-M)/D = x$ and $(r-M)/D = y$ are applied to (33). One obtains

$$\varphi_{\mu,\lambda} \;=\; \langle R\rangle - (\lambda-\mu)\binom{\lambda}{\mu} \int_{x=-\infty}^{x=\infty} p_s(x)[1-P_s(x)]^{\lambda-\mu-1}[P_s(x)]^{\mu-1} \int_{y=-\infty}^{y=x} (Dy+M)p_s(y)\,dy\,dx, \tag{74}$$

with

$$p_s(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2}\, [1+\gamma He_3(z)], \tag{75}$$

$$P_s(z) = \frac{1}{2} + \Phi_0(z) - \frac{\gamma}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2}\, He_2(z). \tag{76}$$

Note, we have shifted the lower limits $x = y = -M/D$ to $-\infty$ because most of the probability mass of $p_s(z)$ is concentrated around $x = 0$ (standardization!) whereas $M \gg D$ holds. The integration over y gives

$$\int_{y=-\infty}^{y=x} (Dy+M)p_s(y)\,dy = MP_s(x) - \frac{D}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}(1+\gamma x^3). \tag{77}$$

If (77) is inserted into (74), then the first term of (77) produces an integral that can be simplified by (29)

$$\begin{aligned}
&(\lambda-\mu)\binom{\lambda}{\mu} \int_{x=-\infty}^{x=\infty} p_s(x)[1-P_s(x)]^{\lambda-\mu-1}[P_s(x)]^{\mu}\,dx \\
&= \frac{\lambda!}{((\mu+1)-1)!\,(\lambda-(\mu+1))!} \int_{-\infty}^{\infty} p_s(x)[P_s(x)]^{(\mu+1)-1}[1-P_s(x)]^{\lambda-(\mu+1)} dx \\
&= \int_{-\infty}^{\infty} p_{(\mu+1):\lambda}(x)\,dx = 1
\end{aligned}$$

and the progress rate reads

$$\varphi_{\mu,\lambda} = \langle R\rangle - M + D\,(\lambda-\mu)\binom{\lambda}{\mu} \int_{x=-\infty}^{x=\infty} p_s(x)[1-P_s(x)]^{\lambda-\mu-1}[P_s(x)]^{\mu-1} \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}\,(1+\gamma x^3)\,dx. \tag{78}$$

The integrand in (78) is nonlinear in γ. As explained in the introduction to section 3.2 we will drop all higher order terms in γ ($|\gamma| \ll 1$). We therefore obtain from (76)

$$[P_s(x)]^{\mu-\alpha} \approx \left(\tfrac{1}{2} + \Phi_0(x)\right)^{\mu-\alpha} - \tfrac{\gamma}{\sqrt{2\pi}}(\mu-\alpha)\left(\tfrac{1}{2} + \Phi_0(x)\right)^{\mu-\alpha-1} e^{-\frac{1}{2}x^2}\, He_2(x) \tag{79}$$

[7] The generation counter and the expectation symbol for D and γ have been omitted, the latter will be re-used if the self-consistent equations will be derived.

and

$$[1 - P_s(x)]^{\lambda-\mu-1} \approx \left(\frac{1}{2} - \Phi_0(x)\right)^{\lambda-\mu-1} + \frac{\gamma}{\sqrt{2\pi}}(\lambda-\mu-1)\left(\frac{1}{2} - \Phi_0(x)\right)^{\lambda-\mu-2} e^{-\frac{1}{2}x^2}\, He_2(x). \tag{80}$$

Now we can substitute (75), (79), and (80) into the $\varphi_{\mu,\lambda}$-integral (78)

$$\begin{aligned}
\varphi_{\mu,\lambda} \;=\; & \langle R\rangle - M + D\,(\lambda-\mu)\binom{\lambda}{\mu}\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2}\left(\frac{1}{2} - \Phi_0(x)\right)^{\lambda-\mu-1}\left(\frac{1}{2} + \Phi_0(x)\right)^{\mu-1} dx\right. \\
& + \gamma\,\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2}\left(\frac{1}{2} - \Phi_0(x)\right)^{\lambda-\mu-1}\left(\frac{1}{2} + \Phi_0(x)\right)^{\mu-1}(x^3-3x)\,dx \\
& + \gamma\,\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2}\left(\frac{1}{2} - \Phi_0(x)\right)^{\lambda-\mu-1}\left(\frac{1}{2} + \Phi_0(x)\right)^{\mu-1} x^3\,dx \\
& + \gamma\,\frac{(\lambda-\mu-1)}{\left(\sqrt{2\pi}\right)^3}\int_{-\infty}^{\infty} e^{-\frac{3}{2}x^2}\left(\frac{1}{2} - \Phi_0(x)\right)^{\lambda-\mu-2}\left(\frac{1}{2} + \Phi_0(x)\right)^{\mu-1}(x^2-1)\,dx \\
& \left. -\; \gamma\,\frac{(\mu-1)}{\left(\sqrt{2\pi}\right)^3}\int_{-\infty}^{\infty} e^{-\frac{3}{2}x^2}\left(\frac{1}{2} - \Phi_0(x)\right)^{\lambda-\mu-1}\left(\frac{1}{2} + \Phi_0(x)\right)^{\mu-2}(x^2-1)\,dx\right].
\end{aligned}$$

As have already been pointed out, we have neglected higher order terms in γ. After some straightforward integrations by parts and the substitution $x = -t$ one obtains

$$\begin{aligned}
\varphi_{\mu,\lambda} \;=\; & \langle R\rangle - M + D\,(\lambda-\mu)\binom{\lambda}{\mu}\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-t^2}\left(\frac{1}{2} + \Phi_0(t)\right)^{\lambda-\mu-1}\left(\frac{1}{2} - \Phi_0(t)\right)^{\mu-1} dt\right. \\
& \left. -\; \gamma\,\frac{1}{2\pi}\int_{-\infty}^{\infty} t\, e^{-t^2}\left(\frac{1}{2} + \Phi_0(t)\right)^{\lambda-\mu-1}\left(\frac{1}{2} - \Phi_0(t)\right)^{\mu-1} dt\right].
\end{aligned} \tag{81}$$

These integrals cannot be further simplified. In order to avoid writing long integral constants the author introduces <u>*generalized progress coefficients*</u> $e^{\alpha,\beta}_{\mu,\lambda}$

$$\boxed{e^{\alpha,\beta}_{\mu,\lambda} := \frac{\lambda-\mu}{\left(\sqrt{2\pi}\right)^{\alpha+1}}\binom{\lambda}{\mu}\int_{-\infty}^{\infty} t^{\beta}\, e^{-\frac{\alpha+1}{2}t^2}\left(\frac{1}{2} + \Phi_0(t)\right)^{\lambda-\mu-1}\left(\frac{1}{2} - \Phi_0(t)\right)^{\mu-\alpha} dt. \qquad (82)}$$

E. g., the progress coefficient $c_{1,\lambda}$ of the $(1,\lambda)$-ES (cf. (18)) can be expressed as

$$c_{1,\lambda} = e^{0,1}_{0,\lambda} = e^{1,0}_{1,\lambda}. \tag{83}$$

The progress rate formula (81) becomes with (82)

$$\varphi_{\mu,\lambda} = \langle R\rangle - M + D\left[e^{1,0}_{\mu,\lambda} - \gamma\, e^{1,1}_{\mu,\lambda}\right].$$

Substitution of M by (56) and the normalization (55) lead us to

$$\varphi^{\star}_{\mu,\lambda} = N\left(1 - \sqrt{1 + \frac{(\sigma^{\star})^2}{N}}\right) + \frac{N}{\langle R\rangle}\, D\left[e^{1,0}_{\mu,\lambda} - \gamma\, e^{1,1}_{\mu,\lambda}\right]. \tag{84}$$

Note, the $(1,\lambda)$ progress rate formula (20) can be recovered from (84) because $\gamma = 0$ (only one normally distributed parent) and $D = \tilde{\sigma}$ hold. For $\mu > 1$, D, and $\mu > 2$, γ are still unknown. The determination by the self-consistent method (cp. 3.1.7) will be performed in 3.2.4.

3.2.2 Derivation of $\overline{D}$

For the derivation of $\overline{D}$ approximation (65) is used. The main task will be the calculation of $\overline{\langle(\Delta R)^2\rangle}$ defined by (66). With (49, 50) a μ-fold integral had to be determined. There is a simpler way omitting the μ-fold integral. $\overline{\langle(\Delta R)^2\rangle}$ may be expressed by an *ordered sum*. From (64) we obtain

$$\begin{aligned}
\langle(\Delta R)^2\rangle &= \frac{1}{\mu}\sum_{m=1}^{\mu} R_m^2 - \left(\frac{1}{\mu}\sum_{m=1}^{\mu} R_m\right)^2 = \frac{1}{\mu}\sum_{m=1}^{\mu} R_m^2 - \frac{1}{\mu^2}\sum_{l=1}^{\mu}\sum_{k=1}^{\mu} R_k R_l \\
\langle(\Delta R)^2\rangle &= \frac{\mu-1}{\mu}\frac{1}{\mu}\sum_{m=1}^{\mu} R_m^2 - \frac{1}{\mu^2}\sum_{l=1}^{\mu}\sum_{k\neq l} R_k R_l \\
\langle(\Delta R)^2\rangle &= \frac{\mu-1}{\mu}\frac{1}{\mu}\sum_{m=1}^{\mu} R_m^2 - \frac{2}{\mu^2}\sum_{l=2}^{\mu}\sum_{k=1}^{l-1} R_k R_l \,.
\end{aligned}$$

Thus, $\langle(\Delta R)^2\rangle$ is expressed by ordered expressions obeying the order relations $R_k \le R_l$ and $k < l$. The expectation reads

$$\overline{\langle(\Delta R)^2\rangle} = \frac{\mu-1}{\mu}\frac{1}{\mu}\sum_{m=1}^{\mu} \overline{R_m^2} - \frac{2}{\mu^2}\sum_{l=2}^{\mu}\sum_{k=1}^{l-1} \overline{R_k R_l}\,. \qquad (85)$$

$\overline{R_m^2}$ is the expectation of the second moment of the m-th order statistics sampled from the offspring distribution $p(r)$ (72). More generally, the a-th moment reads

$$\overline{R_m^a} = \int r^a p_{m:\lambda}(r)\, dr.$$

Using (29) the sum over m becomes

$$\frac{1}{\mu}\sum_{m=1}^{\mu} \overline{R_m^a} = \frac{\lambda}{\mu}\int_0^{\infty} r^a p(r) \sum_{m=1}^{\mu} \frac{(\lambda-1)!}{(m-1)!\,(\lambda-m)!}\,[P(r)]^{m-1}[1-P(r)]^{\lambda-m} dr.$$

By the same calculation used to derive (33) from (30) one obtains

$$\frac{1}{\mu}\sum_{m=1}^{\mu} \overline{R_m^a} = (\lambda-\mu)\binom{\lambda}{\mu}\int_{t=0}^{t=\infty} p(t)[1-P(t)]^{\lambda-\mu-1}[P(t)]^{\mu-1}\int_{r=0}^{r=t} r^a p(r)\, dr\, dt. \qquad (86)$$

$\overline{R_k R_l}$ is a product moment to be built from the joint density $p_{k\,l:\lambda}(r_k, r_l)$ of the ordered variates $r_k \le r_l$. For the general case $\overline{R_k^a R_l^b}$ one has

$$\overline{R_k^a R_l^b} = \int_{r_l=0}^{r_l=\infty}\int_{r_k=0}^{r_k=r_l} r_k^a\, r_l^b\, p_{k\,l:\lambda}(r_k, r_l)\, dr_k\, dr_l. \qquad (87)$$

The joint distribution $p_{k\,l:\lambda}(r_k, r_l)$ can be easily obtained by considering the 'ordered event scale' of the R_m-variates

$$\underbrace{R_{1:\lambda} \ldots R_{k-1:\lambda}}_{(k-1)} \quad \underbrace{\le R_{k:\lambda}}_{(1)} \quad \underbrace{\le R_{k+1:\lambda} \ldots R_{l-1:\lambda}}_{(l-k-1)} \quad \underbrace{\le R_{l:\lambda}}_{(1)} \quad \underbrace{\le R_{l+1:\lambda} \ldots R_{\lambda:\lambda}}_{(\lambda-l)}$$
$$[P(R \le r_k)]^{k-1} \quad p(r_k) \quad [P(r_k \le R \le r_l)]^{l-k-1} \quad p(r_l) \quad [P(R > r_l)]^{\lambda-l}$$

From this ordering one reads the joint density

$$p_{k\,l:\lambda}(r_k, r_l) = \lambda!\, \frac{[P(r_k)]^{k-1}\, p(r_k)\, [P(r_l) - P(r_k)]^{l-k-1}\, p(r_l)\, [1 - P(r_l)]^{\lambda-l}}{(k-1)!\, 1!\, (l-k-1)!\, 1!\, (\lambda-l)!}$$

taking into account the different equivalent orderings. With (87) one obtains ($u := r_k$, $v := r_l$)

$$\sum_{l=2}^{\mu} \sum_{k=1}^{l-1} \overline{R_k^a R_l^b} = \int_{v=0}^{v=\infty} v^b p(v) \int_{u=0}^{u=v} u^a p(u)\, \lambda! \sum_{l=2}^{\mu} \frac{[1 - P(v)]^{\lambda-l}}{(\lambda-l)!} \times \sum_{k=1}^{l-1} \frac{[P(u)]^{k-1}[P(v) - P(u)]^{l-k-1}}{(k-1)!\,(l-k-1)!}\, du\, dv. \quad (88)$$

The sum over k can be expressed by the binomial theorem

$$\sum_{k=1}^{l-1} \frac{[P(u)]^{k-1}[P(v) - P(u)]^{l-k-1}}{(k-1)!\,(l-k-1)!} = \sum_{s=0}^{l-2} \frac{[P(u)]^{s}[P(v) - P(u)]^{l-2-s}}{s!\,(l-2-s)!}$$
$$= \frac{1}{(l-2)!} \sum_{s=0}^{l-2} \frac{(l-2)!}{s!\,(l-2-s)!} [P(u)]^{s}[P(v) - P(u)]^{l-2-s} = \frac{1}{(l-2)!} [P(v)]^{l-2},$$

and the sum over l in (88) becomes

$$\sum_{l=2}^{\mu} \frac{[P(v)]^{l-2}[1 - P(v)]^{\lambda-l}}{(l-2)!\,(\lambda-l)!} = \sum_{m=1}^{\mu-1} \frac{[P(v)]^{m-1}[1 - P(v)]^{\lambda-1-m}}{(m-1)!\,(\lambda-1-m)!}. \quad (89)$$

The comparison with (31) shows that the sum can be expressed by

$$\sum_{m=1}^{\mu-1} \frac{[P(v)]^{m-1}[1 - P(v)]^{\lambda-1-m}}{(m-1)!\,(\lambda-1-m)!} = \frac{1}{(\lambda-\mu-1)!\,(\mu-2)!} \int_{x=0}^{x=1-P(v)} x^{\lambda-\mu-1}(1-x)^{\mu-2} dx.$$

Therefore, from (89) it follows

$$\lambda! \sum_{l=2}^{\mu} \frac{[P(v)]^{l-2}[1 - P(v)]^{\lambda-l}}{(l-2)!\,(\lambda-l)!} = \mu(\mu-1)(\lambda-\mu) \binom{\lambda}{\mu} \int_{x=0}^{x=1-P(v)} x^{\lambda-\mu-1}(1-x)^{\mu-2} dx$$

and (88) becomes

$$\sum_{l=2}^{\mu} \sum_{k=1}^{l-1} \overline{R_k^a R_l^b} = \mu(\mu-1)(\lambda-\mu) \binom{\lambda}{\mu} \int_{v=0}^{v=\infty} \int_{x=0}^{x=1-P(v)} v^b p(v) \times \int_{u=0}^{u=v} u^a p(u)\, du\, x^{\lambda-\mu-1}(1-x)^{\mu-2} dx\, dv.$$

Performing the same transformation applied to (32) in order to get (33) one finds (let $r = v$)

$$\sum_{l=2}^{\mu}\sum_{k=1}^{l-1}\overline{R_k^a R_l^b} = \mu(\mu-1)(\lambda-\mu)\binom{\lambda}{\mu}\int_{t=0}^{t=\infty} p(t)[1-P(t)]^{\lambda-\mu-1}[P(t)]^{\mu-2} \times \int_{r=0}^{r=t} r^b p(r)\left(\int_{u=0}^{u=r} u^a p(u)\,du\right)\,dr\,dt. \quad (90)$$

Now we can assemble the integral representation for (85) with $a = 2$ in (86), and $a = b = 1$ in (90), respectively

$$\overline{\langle(\Delta R)^2\rangle} = \frac{\mu-1}{\mu}(\lambda-\mu)\binom{\lambda}{\mu}\left\{\int_{t=0}^{t=\infty} p(t)[1-P(t)]^{\lambda-\mu-1}[P(t)]^{\mu-1}\int_{r=0}^{r=t} r^2 p(r)\,dr\,dt - 2\int_{t=0}^{t=\infty} p(t)[1-P(t)]^{\lambda-\mu-1}[P(t)]^{\mu-2}\int_{r=0}^{r=t} r\,p(r)\left(\int_{u=0}^{u=r} u\,p(u)\,du\right)\,dr\,dt\right\}. \quad (91)$$

The next step would be the approximation of (91) using (73). The derivation is similar to that in section 3.2.1 and will not be presented here (for details cp. Beyer, 1994b). As the result one finds

$$\overline{\langle(\Delta R)^2\rangle} = \frac{\mu-1}{\mu}\,D^2\,\left[1 + e^{1,1}_{\mu,\lambda} - e^{2,0}_{\mu,\lambda} - 2\gamma\left(e^{1,0}_{\mu,\lambda} + e^{1,2}_{\mu,\lambda} - e^{2,1}_{\mu,\lambda}\right)\right]. \quad (92)$$

Finally we obtain from (65)

$$\overline{D} = \sqrt{\tilde{\sigma}^2 + \frac{D^2}{\zeta^2}\,\frac{\mu-1}{\mu}\,\left[1 + e^{1,1}_{\mu,\lambda} - e^{2,0}_{\mu,\lambda} - 2\gamma\left(e^{1,0}_{\mu,\lambda} + e^{1,2}_{\mu,\lambda} - e^{2,1}_{\mu,\lambda}\right)\right]}. \quad (93)$$

3.2.3 Derivation of $\overline{\gamma}$

In order to approximate $\overline{\gamma}$, given by (67), the $\overline{\langle(\Delta R)^3\rangle}$-expression (68) has to be calculated. Since the derivations are very long, they are omitted here (detailed derivations can be found in Beyer, 1994b). The course to $\overline{\gamma}$ is as follows.

First, $\overline{\langle(\Delta R)^3\rangle}$ (cf. (68)) is to express by sums of ordered products of the R_m-variates obeying $R_k \le R_l \le R_m$ and $k < l < m$

$$\overline{\langle(\Delta R)^3\rangle} = \left(1 - \frac{3}{\mu} + \frac{2}{\mu^2}\right)\frac{1}{\mu}\sum_{m=1}^{\mu}\overline{R_m^3} - \left(3 - \frac{6}{\mu}\right)\frac{1}{\mu^2}\sum_{l=2}^{\mu}\sum_{k=1}^{l-1}\left(\overline{R_k^2 R_l} + \overline{R_k R_l^2}\right) + \frac{12}{\mu^3}\sum_{m=3}^{\mu}\sum_{l=2}^{m-1}\sum_{k=1}^{l-1}\overline{R_k R_l R_m}\,.$$

In the second step, integral representations for the sums are to derive, similar to section 3.2.2. As the result one obtains

$$\overline{\langle(\Delta R)^3\rangle} = \frac{\mu-1}{\mu}\,\frac{\mu-2}{\mu}(\lambda-\mu)\binom{\lambda}{\mu}\left\{\int_{t=0}^{t=\infty} p(t)[1-P(t)]^{\lambda-\mu-1}[P(t)]^{\mu-1}\int_{r=0}^{r=t} r^3 p(r)\,dr\,dt - 3\int_{t=0}^{t=\infty} p(t)[1-P(t)]^{\lambda-\mu-1}[P(t)]^{\mu-2}\int_{r=0}^{r=t}\int_{u=0}^{u=r}(ru^2 + r^2u)\,p(u)\,p(r)\,du\,dr\,dt + 12\int_{t=0}^{t=\infty} p(t)[1-P(t)]^{\lambda-\mu-1}[P(t)]^{\mu-3}\right.$$

$$\times \int_{w=0}^{w=t} \int_{v=0}^{v=w} \int_{u=0}^{u=v} wvu\, p(u)\, p(v)\, p(w)\, du\, dv\, dw\, dt \Bigg\} . \tag{94}$$

Third, the integrals in (94) are approximated using (73) for p. Finally one arrives at

$$\begin{aligned} \overline{\langle (\Delta R)^3 \rangle} \quad = \quad & \frac{\mu-1}{\mu} \frac{\mu-2}{\mu} D^3 \Big[\Big(e^{1,0}_{\mu,\lambda} - e^{1,2}_{\mu,\lambda} + 3e^{2,1}_{\mu,\lambda} - 2e^{3,0}_{\mu,\lambda} \Big) \\ & + \gamma \Big(6 + 3e^{1,1}_{\mu,\lambda} + 3e^{1,3}_{\mu,\lambda} - 6e^{2,0}_{\mu,\lambda} - 9e^{2,2}_{\mu,\lambda} + 6e^{3,1}_{\mu,\lambda} \Big) \Big] . \end{aligned}$$

Therefore, the $\overline{\gamma}$-expression (67) becomes

$$\begin{aligned} \overline{\gamma} \quad = \quad & \frac{1}{\zeta^3} \frac{\mu-1}{\mu} \frac{\mu-2}{\mu} \frac{1}{6} \Big[\Big(e^{1,0}_{\mu,\lambda} - e^{1,2}_{\mu,\lambda} + 3e^{2,1}_{\mu,\lambda} - 2e^{3,0}_{\mu,\lambda} \Big) \\ & + \gamma \Big(6 + 3e^{1,1}_{\mu,\lambda} + 3e^{1,3}_{\mu,\lambda} - 6e^{2,0}_{\mu,\lambda} - 9e^{2,2}_{\mu,\lambda} + 6e^{3,1}_{\mu,\lambda} \Big) \Big] . \end{aligned} \tag{95}$$

3.2.4 The Self-Consistent Determination of the Progress Rate

Now, we are able to perform the self-consistent method described in 3.1.7. The self-consistent equation (71), equating γ and $\overline{\gamma}$, can be easily solved for $\overline{\gamma}$. Let $\gamma_{\mu,\lambda} := \overline{\gamma}$, we obtain from (95) the self-consistent solution

$$\gamma_{\mu,\lambda} = \frac{e^{1,0}_{\mu,\lambda} - e^{1,2}_{\mu,\lambda} + 3e^{2,1}_{\mu,\lambda} - 2e^{3,0}_{\mu,\lambda}}{\frac{6\mu^2}{(\mu-1)(\mu-2)} \left(\sqrt{1 + \frac{(\sigma^\star)^2}{N}} \right)^3 - \Big(6 + 3e^{1,1}_{\mu,\lambda} + 3e^{1,3}_{\mu,\lambda} - 6e^{2,0}_{\mu,\lambda} - 9e^{2,2}_{\mu,\lambda} + 6e^{3,1}_{\mu,\lambda} \Big)} . \tag{96}$$

Note, for $\mu = 1$ and $\mu = 2$ eq. (96) is not applicable. However, from (95) one gets

$$\gamma_{1,\lambda} = \gamma_{2,\lambda} = 0.$$

The self-consistent equation (70) for determining $\overline{D}$ reads with (93)

$$\overline{D} = \sqrt{ \tilde{\sigma}^2 + \frac{\overline{D}^2}{\zeta^2} \frac{\mu-1}{\mu} \Big[1 + e^{1,1}_{\mu,\lambda} - e^{2,0}_{\mu,\lambda} - 2\overline{\gamma} \Big(e^{1,0}_{\mu,\lambda} + e^{1,2}_{\mu,\lambda} - e^{2,1}_{\mu,\lambda} \Big) \Big] }.$$

The solution to $\overline{D}$ is therefore (NB, $\overline{\gamma} = \gamma_{\mu,\lambda}$)

$$\overline{D} = \frac{\tilde{\sigma}}{\sqrt{1 - \left(\sqrt{1 + \frac{(\sigma^\star)^2}{N}} \right)^{-2} \frac{\mu-1}{\mu} \Big[1 + e^{1,1}_{\mu,\lambda} - e^{2,0}_{\mu,\lambda} - 2\gamma_{\mu,\lambda} \Big(e^{1,0}_{\mu,\lambda} + e^{1,2}_{\mu,\lambda} - e^{2,1}_{\mu,\lambda} \Big) \Big]}} . \tag{97}$$

With (84) and (55/59) we finally arrive at the <u>*progress rate formula of the (μ, λ)-strategy*</u>

$$\boxed{\begin{aligned} \varphi^\star_{\mu,\lambda}(\sigma^\star) \quad = \quad & N \left(1 - \sqrt{1 + \frac{(\sigma^\star)^2}{N}} \right) \\ & + \sigma^\star \frac{\sqrt{1 + \frac{(\sigma^\star)^2}{2N}} \Big(e^{1,0}_{\mu,\lambda} - \gamma_{\mu,\lambda} e^{1,1}_{\mu,\lambda} \Big)}{\sqrt{1 + \frac{(\sigma^\star)^2}{N} - \frac{\mu-1}{\mu} \Big[1 + e^{1,1}_{\mu,\lambda} - e^{2,0}_{\mu,\lambda} - 2\gamma_{\mu,\lambda} \Big(e^{1,0}_{\mu,\lambda} + e^{1,2}_{\mu,\lambda} - e^{2,1}_{\mu,\lambda} \Big) \Big]}} . \end{aligned}} \tag{98}$$

This formula depends on nine generalized progress coefficients (82) (taking $\gamma_{\mu,\lambda}$ into account). If (96) is inserted, there are still five independent coefficients built up from the nine $e^{\alpha,\beta}_{\mu,\lambda}$-coefficients. For numerical calculations this is not a serious problem, but it may be the case for theoretical investigations. Therefore, in 3.4 we will derive a simpler formula sufficient for most of the applications. In the next section we will give some comparisons of (98) with experiments. At this point we will only discuss the different levels of approximation included in (98/96).

The lowest level of approximation may be the assumption that all parents are located at one point. However, re-interpreted algorithmically, *this would not be a (μ, λ) strategy at all!* Neglecting this important fact we can derive the rather poor formula from (98) or (84) by letting $\overline{D} = \tilde{\sigma}$ (no parental spreading) and $\gamma_{\mu,\lambda} = 0$ (no skewness because of Gaußian mutations)

$$\varphi^\star_{\mu,\lambda} = N\left(1 - \sqrt{1 + \frac{(\sigma^\star)^2}{N}}\right) \; + \; e^{1,0}_{\mu,\lambda}\,\sigma^\star \sqrt{\frac{1 + \frac{(\sigma^\star)^2}{2N}}{1 + \frac{(\sigma^\star)^2}{N}}} \qquad \texttt{(don't use it!)}.$$

The next approximation level yields better results by the assumption that the parental spreading may be described by a normal distribution with standard deviation $\overline{D}$. Because there is no skewness, we have $\gamma_{\mu,\lambda} = 0$, hence, from (98) it follows

$$\varphi^\star_{\mu,\lambda}(\sigma^\star) = N\left(1 - \sqrt{1 + \frac{(\sigma^\star)^2}{N}}\right) \; + \; e^{1,0}_{\mu,\lambda}\,\sigma^\star \, \frac{\sqrt{1 + \frac{(\sigma^\star)^2}{2N}}}{\sqrt{1 + \frac{(\sigma^\star)^2}{N} - \frac{\mu-1}{\mu}\left(1 + e^{1,1}_{\mu,\lambda} - e^{2,0}_{\mu,\lambda}\right)}}. \qquad (99)$$

The third approximation level is given by (98) itself. Here, skewness is incorporated in the distribution of the parents.

The fourth approximation would be the introduction of an excess parameter being the fifth coefficient in the expansion (38). However, this is beyond the scope of this paper. As will be verified by experiments, the approximation (98) works well.

3.3 Comparison with Experiments

The predictions of the N-dependent progress rate formula (98/96) have been excessively tested. Simulating the progress rate of (μ, λ)-strategies for $\mu > 1$ is somewhat involved. The reconstruction of φ from a single ES run with constant mutation strength σ (not $\sigma^\star$) is much too inexact due to the stochastic fluctuations.

One possibility may be averaging over a bunch of runs. However, this produces random samples of $(\varphi^\star, \sigma^\star)$-pairs to be sorted in an adequate number of $\sigma^\star$-classes. To get reliable class averages one has to perform many ES runs.

Facing the problem of estimating $(\varphi^\star, \sigma^\star)$ in classes, the author has developed a somehow artificial but efficacious simulation method: The ES is tested for constant σ *and* $\sigma^\star$, i. e. the parental $\langle R \rangle = R_0$ is held constant from generation to generation. This is accomplished by re-scaling the selected μ offspring such that the average distance of the new parents to the optimum is always R_0. Starting from an arbitrary parent distribution (usually, all parents may be concentrated in one point) a steady state behavior will be reached after a number G of generations. For the experiments to be presented $G = 100$ has been chosen. In Fig. 3 results from the $(\mu, 8)$ and $(\mu, 80)$-ES with $N = 100$ are depicted. The dots display the simulation results and the curves are graphs from (98). For each data point (dot) 20,000 ($\lambda = 8$) and 10,000 ($\lambda = 80$) generations,respectively, have been performed. One observes a good agreement between theory and experiment.

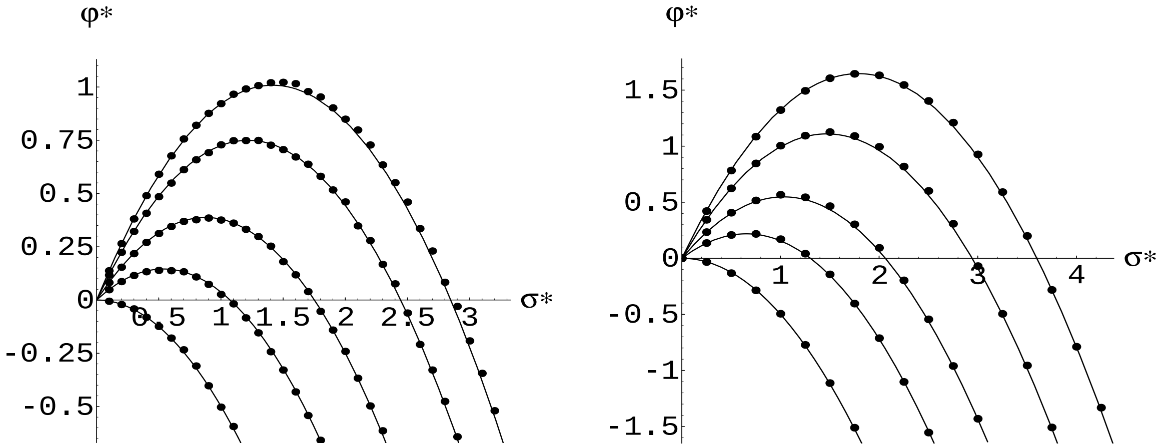

Figure 3: Comparison between experiment (dots) and theory (curves) on the progress rate $\varphi^\star$ of (μ,λ)-strategies. Left picture: $(\mu, 8)$-strategies with $\mu = 8,\ 6,\ 4,\ 2,\ 1$ (from the left curve to the right one). Right picture: $(\mu, 80)$-strategies with $\mu = 80,\ 60,\ 40,\ 20,\ 10$ (from the left curve to the right curve).

3.4 A Simplified Progress Rate Formula and Further Experiments

Since the progress rate formula (98/96) is rather unwieldy, we would like to find a simpler one. The simplified formula should contain only one progress coefficient $c_{\mu,\lambda}$ which does not depend on N and $\sigma^\star$

$$c_{\mu,\lambda} \neq f(\sigma^\star, N). \tag{100}$$

Further, we postulate a correspondence to the findings of section 2.2: The N-dependence observed in equation (20) should hold for the new approximative $\varphi^\star_{\mu,\lambda}$-formula in the case $\mu = 1$, too. Furthermore, $c_{\mu,\lambda}$ should be given by (18), if $\mu = 1$.
Equation (98) can be rewritten taking (84, 59, 55, 56, 97) into account

$$\begin{aligned}\varphi^\star_{\mu,\lambda}(\sigma^\star) \quad &= \quad N\left(1 - \sqrt{1 + \frac{(\sigma^\star)^2}{N}}\right) \\ &\quad + \underbrace{\frac{\left(e^{1,0}_{\mu,\lambda} - \gamma_{\mu,\lambda} e^{1,1}_{\mu,\lambda}\right)}{\sqrt{1 - \frac{\mu-1}{\zeta^2\,\mu}\left[1 + e^{1,1}_{\mu,\lambda} - e^{2,0}_{\mu,\lambda} - 2\gamma_{\mu,\lambda}\left(e^{1,0}_{\mu,\lambda} + e^{1,2}_{\mu,\lambda} - e^{2,1}_{\mu,\lambda}\right)\right]}}}_{\overset{!}{=} c_{\mu,\lambda}} \,\sigma^\star \sqrt{\frac{1 + \frac{(\sigma^\star)^2}{2N}}{1 + \frac{(\sigma^\star)^2}{N}}}.\end{aligned}$$

By comparison with (20) it becomes clear that postulating a $c_{\mu,\lambda}$-coefficient is equivalent to let $\zeta = 1$ in the 'underbraced' fraction. Thus, we obtain the progress rate formula

$$\boxed{\varphi^\star_{\mu,\lambda}(\sigma^\star) = N\left(1 - \sqrt{1 + \frac{(\sigma^\star)^2}{N}}\right) \; + \; c_{\mu,\lambda}\,\sigma^\star \sqrt{\frac{1 + \frac{(\sigma^\star)^2}{2N}}{1 + \frac{(\sigma^\star)^2}{N}}} \qquad (101)}$$

with the <u>*progress coefficient*</u> $c_{\mu,\lambda}$

$$c_{\mu,\lambda} = \frac{e^{1,0}_{\mu,\lambda} - \gamma^{\infty}_{\mu,\lambda} e^{1,1}_{\mu,\lambda}}{\sqrt{1 - \frac{\mu-1}{\mu}\left[1 + e^{1,1}_{\mu,\lambda} - e^{2,0}_{\mu,\lambda} - 2\gamma^{\infty}_{\mu,\lambda}\left(e^{1,0}_{\mu,\lambda} + e^{1,2}_{\mu,\lambda} - e^{2,1}_{\mu,\lambda}\right)\right]}} \tag{102}$$

and

$$\gamma^{\infty}_{\mu,\lambda} = \frac{e^{1,0}_{\mu,\lambda} - e^{1,2}_{\mu,\lambda} + 3e^{2,1}_{\mu,\lambda} - 2e^{3,0}_{\mu,\lambda}}{\frac{6\mu^2}{(\mu-1)(\mu-2)} - \left(6 + 3e^{1,1}_{\mu,\lambda} + 3e^{1,3}_{\mu,\lambda} - 6e^{2,0}_{\mu,\lambda} - 9e^{2,2}_{\mu,\lambda} + 6e^{3,1}_{\mu,\lambda}\right)}. \tag{103}$$

The $c_{\mu,\lambda}$ progress coefficients can be easily computed by numerical integration techniques applied to (82). Table 3.4 displays $c_{\mu,\lambda}$-values over a wide range. Between these values interpolation is a good means to obtain further $c_{\mu,\lambda}$ values.

$\mu \setminus \lambda$	5	10	20	30	40	50	100	150	200	250	300
1	1.16	1.54	1.87	2.04	2.16	2.25	2.51	2.65	2.75	2.82	2.88
2	0.92	1.36	1.72	1.91	2.04	2.13	2.40	2.55	2.65	2.72	2.79
3	0.68	1.20	1.60	1.80	1.93	2.03	2.32	2.47	2.57	2.65	2.71
4	0.41	1.05	1.49	1.70	1.84	1.95	2.24	2.40	2.51	2.59	2.65
5	0.00	0.91	1.39	1.62	1.77	1.87	2.18	2.34	2.45	2.53	2.60
10	-	0.00	0.99	1.28	1.46	1.59	1.94	2.12	2.24	2.33	2.40
20	-	-	0.00	0.76	1.03	1.20	1.63	1.84	1.97	2.07	2.15
30	-	-	-	0.00	0.65	0.89	1.41	1.64	1.79	1.90	1.99
40	-	-	-	-	0.00	0.57	1.22	1.49	1.65	1.77	1.86
50	-	-	-	-	-	0.00	1.06	1.35	1.53	1.65	1.75
100	-	-	-	-	-	-	0.00	0.81	1.07	1.24	1.36

Table 1: Selected $c_{\mu,\lambda}$-coefficients.

From (101), the progress rate of the hypersphere, one can easily derive the linear model '*hyperplane*'. It can be obtained by the assumption of small mutations compared to the average radius of curvature. The fitness landscape appears for such small mutations as a flat space, i. e. the ES climbs up an inclined (hyper)plane. Expanding (101) in a Taylor series about $\sigma^\star = 0$ and breaking off after the linear term yields with (55)

$$\varphi(\sigma) = c_{\mu,\lambda}\sigma. \tag{104}$$

Thus, we have gained the progress formula for the model 'inclined plane'

Eq. (104) gives a good starting point for testing the $c_{\mu,\lambda}$ values by experiments. If $\sigma = 1$, i. e. for standard normal mutations, $c_{\mu,\lambda}$ can be easily computed by an one-dimensional simulation of the (μ, λ)-ES:

1. Start at $x = 0$,
2. produce λ offspring by $N(0,1)$-mutations,
3. select the largest μ x-values and compute their average $\langle x \rangle$,
4. determine the distance $\Delta\langle x \rangle$ to the preceding generation,

5. `go to 2.`

After a number of generations G_1 the $\Delta\langle x\rangle$-values are collected over the next G_2 generations. Their average is an estimate for $\varphi = c_{\mu,\lambda}$. In Figure 4 results from such experiments with $G_1 = 1,000$, $G_2 = 1,000,000$ are displayed by dots. The curves are graphs from $c_{\mu,\lambda} = f_\lambda(\mu)$ treating μ as a real variable and λ as a (fixed) parameter. As can be seen, there is a satisfactory agreement between theory and experiment.

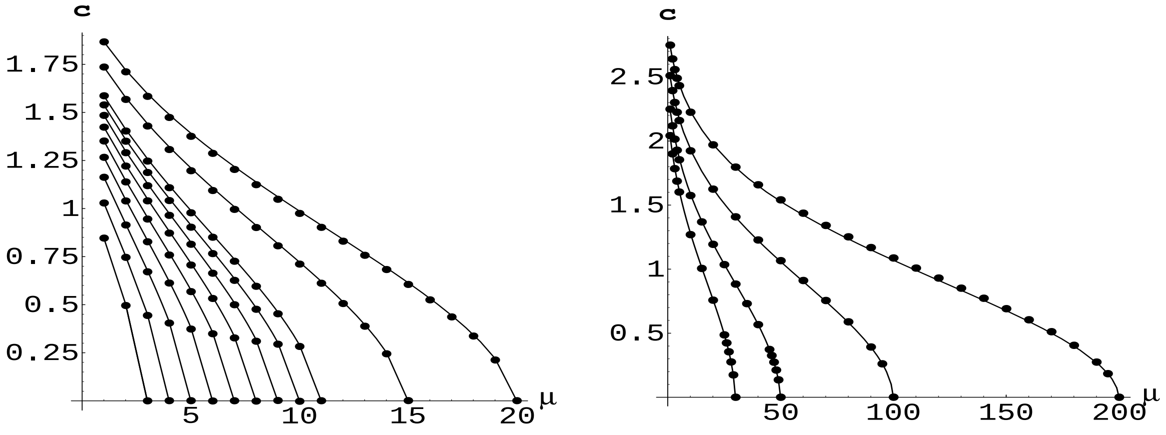

Figure 4: The progress coefficients $c_{\mu,\lambda}$ in theory (curves) and experiment (dots). Left picture: $c_{\mu,\lambda}$ for $\lambda = 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10,\ 11,\ 15,\ 20$, right picture: $c_{\mu,\lambda}$ for $\lambda = 30,\ 50,\ 100,\ 200$.

4 Conclusions and Outlook

In this paper an N-dependent progress rate theory for multi-membered Evolution Strategies has been developed. The starting point was the well-known $(1,\lambda)$ spherical model that has been revised to incorporate the influence of the parameter space dimension N. The new result of this revision demonstrates the limits of the quadratic progress rate formula for larger mutations.

The analysis of the (μ,λ)-ES has been performed in section 3. The main problem to be dealt with was the self-consistent determination of the offspring distribution in the parameter space. Due to the spherical symmetry the problem reduces to an one-dimensional one in radius direction. Even this problem was still too difficult to be solved without further simplifications. After expressing the offspring distribution by a series of Hermite polynomials, stochastic approximation techniques, and application of the self-consistent method we arrived at a progress rate formula which is in good agreement with experiments. Neglecting partially the N-dependence in this formula, it was possible to derive a progress rate expression similar to the $(1,\lambda)$ theory containing only one progress coefficient $c_{\mu,\lambda}$. This formula should suffice for most of the applications.

The method presented can be extended. One possibility is the development of progress rate formulas for strategies on real world optimizations including noisy fitness data. The combination of the techniques used in (Beyer, 1993) and the methods presented in this article should provide an approach to this problem.

Another research direction deals with recombination. It is interesting to notice that strategies using *multi-recombination* exhibit large progress rates. Such strategies denoted by $(\mu/\mu, \lambda)$ are analyzed in (Beyer, 1995).

Acknowledgments

The author is grateful to Joachim Born, Technical University Berlin, for helpful remarks. This work was funded by the Deutsche Forschungsgemeinschaft, DFG grant Be 1578/1-1.

References

Abramowitz, M., & Stegun, I. A. (1984). *Pocketbook of Mathematical Functions.* Thun, Germany: Verlag Harri Deutsch.

Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1992). *A First Course in Order Statistics.* New York: Wiley.

Beyer, H.-G. (1993). Toward a Theory of Evolution Strategies: Some Asymptotical Results from the $(1 \overset{+}{,} \lambda)$-Theory. *Evolutionary Computation*, 1(2), 165–188.

Beyer, H.-G. (1994a). Towards a Theory of 'Evolution Strategies': Progress Rates and Quality Gain for $(1 \overset{+}{,} \lambda)$-Strategies on (Nearly) Arbitrary Fitness Functions. In Y. Davidor, R. Männer, & H.-P. Schwefel (Eds.), *Parallel Problem Solving from Nature, 3*, 58–67. Heidelberg, Germany: Springer Verlag.

Beyer, H.-G. (1994b). 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, Department of Computer Science, University of Dortmund.

Beyer, H.-G. (1995). Toward a Theory of Evolution Strategies: On the Benefit of Sex - the $(\mu/\mu, \lambda)$-Theory. To appear in *Evolutionary Computation*, 3(1).

Fisz, M. (1971). *Wahrscheinlichkeitsrechnung und mathematische Statistik.* Berlin, Germany: VEB Deutscher Verlag der Wissenschaften.

Mühlenbein, H., & Schlierkamp-Voosen, D. (1993). Predictive models for the Breeder Genetic Algorithm. *Evolutionary Computation*, 1(1), 25–49.

Rechenberg, I. (1989). Artificial evolution and artificial intelligence. In R. Forsyth (Ed.) *Machine Learning: Principles and Techniques*, 83–103. London, UK: Chapman and Hall.

Rechenberg, I. (1994). *Evolutionsstrategie '94.* Stuttgart, Germany: Frommann–Holzboog Verlag.

Schwefel, H.-P. (1981). *Numerical Optimization of Computer Models.* Chichester: Wiley.