Convergence Rates of Evolutionary Algorithms for a Class of Convex Objective Functions

Günter Rudolph
Universität Dortmund
Fachbereich Informatik, LS XI
D–44221 Dortmund / Germany
rudolph@icd.de

April 21, 1997

Abstract

Probabilistic optimization algorithms that mimic the process of biological evolution are usually subsumed under the term 'evolutionary algorithms.' This work extends the convergence theory of evolutionary algorithms by presenting a sufficient convergence condition for those evolutionary algorithms that do not necessarily generate a sequence of feasible points such that the associated objective function values decrease monotonically to the global minimum. Moreover, it is investigated how fast the sequence of objective function values generated by an evolutionary algorithm approaches the minimum of strongly convex functions in a probabilistic sense. The theoretical analysis presented here distinguishes from related studies in three points: First, it does not require advanced calculus. Second, only the first partial derivatives of the objective function are assumed to exist. Third, one obtains sharp bounds on the convergence rates for a class of functions being a superset of the class of quadratic functions with positive definite Hessian matrix.

1 Introduction

Evolutionary algorithms (EAs) belong to the class of probabilistic optimization algorithms whose design is inspired by principles of biological evolution. A population of individuals—each of them representing a feasible solution of an optimization problem—repeatedly undergoes a cycle of random variation and selection which leads in many cases to practically acceptable solutions and sometimes even to globally optimal solutions.
The typical field of application of EAs are difficult optimization problems for which specialized methods are not available or traditional methods fail for reasons whichever. Here, it is analyzed how fast a specific subclass of EAs approaches the minimum of a convex function. Although convex objective functions are not an appropriate domain for EAs (such problems can be solved by deterministic optimization methods more efficiently) it is not useless to consider them, since an optimization method that is intended to tackle just the most difficult problems also ought to be "sufficiently efficient" for simple problems.
Investigations in this direction have a long list of predecessors. Early publications (Rastrigin 1963; Schumer and Steiglitz 1968; Rechenberg 1973) considered algorithms, later classified as $(1+1)$–EA, that may be interpreted as the simplest form of an evolutionary

process: A single individual is randomly mutated and the worse of the original and the new point is selected to "die." The objective function under consideration was the sum of squares of n real–valued variables. A considerable extension of these results is presented in Rappl (1989) who investigated the performance of the same algorithm for a class of objective functions that is essentially identical to the class considered here. Another avenue of extension was entered in Schwefel (1977), pp. 150–157: The objective function was again the sum of squares but a single individual now generates $\lambda \geq 2$ offspring by random mutations and the best of the offspring and the parent becomes the parent of the next generation. Alternatively, the new parent is chosen solely among the λ offspring. Although this petty modification might seem to be neglectable it has the theoretically significant effect that the new parent may be worse than the old one. This algorithm, known as $(1, \lambda)$–EA, was investigated in case of quadratic objective functions (interpreted as second order Taylor expansion of the original objective function) in Rechenberg (1994), pp. 51–60, by exploiting the principal axis theorem, and in Beyer (1994), pp. 64–67, via Riemannian differential geometry. Compared to these investigations the approach taken here has three advantages: First, the analysis does not require advanced calculus. Second, only the first partial derivatives of the objective function are assumed to exist. Third, one obtains sharp bounds on the convergence rates for a class of functions being a superset of the class of quadratic functions with positive definite Hessian matrix. This is shown in section 3.

But prior to these calculations it is useful to clarify the underlying meaning of stochastic convergence. Moreover, it is not obvious whether a $(1, \lambda)$–EA will converge (in a sense whichever) to the optimum or not. Those questions are addressed in section 2. The main result actually is an extremely simplified version of the supermartingale approach presented in Rudolph (1994). An alternative route is proposed in Yin, Rudolph, and Schwefel (1996) via tools developed for the analysis of stochastic approximation methods in continuous time. The work presented here, however, will concentrate on evolutionary algorithms with discrete time.

The bounds on the convergence rates developed in section 3 are involved with a constant that depends in a nonlinear manner on the problem dimension n and the number of offspring λ. In principle, this constant can be determined for any specific pair (n, λ) but the efforts required—especially for large n and λ—do not pay for the utility of knowing the exact values. Therefore section 4 is devoted to the development of asymptotical expressions. Finally, some conclusions are drawn in section 5.

2 A Sufficient Convergence Condition

Since the state transitions of an evolutionary algorithm are of stochastic nature the deterministic concept of the "convergence to the optimum" is not appropriate. In order to clarify the exact semantic of a phrase like "the EA converges to the global optimum" one has at first to distinguish between the various modes of stochastic convergence.

DEFINITION 1

Let $Z, Z_0, Z_1, \ldots$ be random variables defined on a probability space $(\Omega, \mathcal{A}, \mathsf{P})$. The sequence $(Z_k : k \geq 0)$ is said to *converge with probability* 1 (w.p.1) or *almost surely* (a.s.) to random variable Z if $\mathsf{P}\{ \lim_{k\to\infty} | Z_k - Z | = 0 \} = 1$, to *converge in probability*

to Z if $\mathsf{P}\{\,|\,Z_k - Z\,| > \epsilon\,\} = o(1)$ as $k \to \infty$ for any $\epsilon > 0$, and to *converge in mean* to Z if $\mathsf{E}[\,|\,Z_k - Z\,|\,] = o(1)$ as $k \to \infty$. □

Both convergence with probability 1 and convergence in mean implies convergence in probability whereas the converse is wrong in general (Lukacs 1975, pp. 33–36). With the definitions above one can assign a rigorous meaning to the notion of the convergence of an evolutionary algorithm.

DEFINITION 2
Let $(X_k : k \geq 0)$ be the sequence of populations generated by some evolutionary algorithm and let $F_k^* = \min\{f(X_{k,1}), \ldots, f(X_{k,\mu})\}$ denote the best objective function value of the population of size $\mu < \infty$ at generation $k \geq 0$. An evolutionary algorithms is said to converge in mean (in probability, with probability 1) to the global minimum $f^* = \min\{f(x) : x \in \mathbb{R}^n\}$ of objective function $f : \mathbb{R}^n \to \mathbb{R}$ if the nonnegative random sequence $(Z_k : k \geq 0)$ with $Z_k = F_k^* - f^*$ converges in mean (in probability, with probability 1) to zero. □

The convergence theory of probabilistic optimization methods resembling a $(1+1)$–EA was established in Devroye (1976), Oppel and Hohenbichler (1978), Born (1978), Solis and Wets (1981), Pintér (1984), and others. The proofs in each of these publications exploited the algorithms' property that the parent of the next generation cannot be worse than the current one, i.e., it is guaranteed by the construction of the algorithms that the stochastic sequence $(Z_k : k \geq 0)$ is *monotonically* decreasing. The result presented below only requires the weaker precondition that the sequence $(Z_k : k \geq 0)$ decreases monotonically *on average*. As a consequence, the objective function value of the best parent may be worse than that of the best parent of the previous generation—as it may happen for the sequence $(Z_k : k \geq 0)$ generated by a $(1, \lambda)$–EA.

THEOREM 1
Let $(X_k : k \geq 0)$ be the sequence of populations generated by some evolutionary algorithm and let $F_k^* = \min\{f(X_{k,1}), \ldots, f(X_{k,\mu})\}$ denote the best objective function value of the population at generation $k \geq 0$. If $\mathsf{E}[\,Z_k\,] < \infty$ and

$$\mathsf{E}[\,Z_{k+1}\,|\,X_k, X_{k-1}, \ldots, X_0\,] \leq c_k\, Z_k \qquad \text{a.s.} \tag{1}$$

where $Z_k = F_k^* - f^*$ and $c_k \in [\,0, 1)$ for all $k \geq 0$ such that the infinite product of the c_k converges to zero, then the evolutionary algorithm converges in mean and with probability 1 to the global minimum of the objective function $f(\cdot)$.
PROOF:
Taking expectations on both sides of inequality (1) yields

$$\mathsf{E}[\,Z_{k+1}\,] = \mathsf{E}[\,\mathsf{E}[\,Z_{k+1}\,|\,X_k, X_{k-1}, \ldots, X_0\,]\,] \leq c_k\, \mathsf{E}[\,Z_k\,]$$

for all $k \geq 0$. This implies

$$\mathsf{E}[\,Z_k\,] \leq \mathsf{E}[\,Z_0\,] \prod_{i=0}^{k-1} c_i \longrightarrow 0$$

as $k \to \infty$ since the infinite product of the c_i converges to zero and $\mathsf{E}[\,Z_0\,] < \infty$ by the preconditions of the theorem. Thus, the sequence $(Z_k : k \geq 0)$ converges in mean to zero.

As for convergence with probability 1, notice that inequality (1) implies that the nonnegative sequence $(Z_k : k \geq 0)$ is a nonnegative supermartingale that converges w.p.1 to a random variable $Z < \infty$ (Neveu 1975, p. 26). This ensures that $(Z_k : k \geq 0)$ converges in probability to Z. But since $(Z_k : k \geq 0)$ also converges in probability to zero by the first part of the proof, and since the limits are unique (Lukacs 1975, p. 39), one may conclude that $Z \equiv 0$. Consequently, the random sequence $(Z_k : k \geq 0)$ converges w.p.1 to zero. □

3 Convergence Rates for Strongly Convex Functions

The notion of the 'convergence rate' of an iterative optimization method is well established is the field of deterministic optimization. It serves as a measure of how fast the deterministic sequence of objective function values approaches the global optimum. For example, let $(x_k : k \geq 0)$ be the sequence of points generated by some deterministic minimization method and $\epsilon_k = f(x_k) - f^*$. The method is said to converge geometrically fast if there exists an index k_0, a constant $A > 0$ and a constant $c \in [0, 1)$ such that $\epsilon_k \leq A\, c^k$ for all $k \geq k_0$. Here c is termed the convergence rate. Following this definition a related concept for stochastic sequences is given below.

DEFINITION 3
Let $(Z_k : k \geq 0)$ be a nonnegative random sequence defined by $Z_k = F_k^* - f^*$ where F_k^* is the best objective function value of a population of some evolutionary algorithm at generation $k \geq 0$. The evolutionary algorithm is said to converge geometrically fast in mean (in probability, w.p.1) to the global minimum if there exists a constant $q > 1$ such that the sequence $(q^k\, Z_k : k \geq 0)$ converges in mean (in probability, w.p.1) to zero. Let $q^* > 1$ be supremum of all constants $q > 1$ such that geometrically fast convergence is still guaranteed. Then $c = 1/q$ is called the convergence rate. □

Let $\tilde{Z}_k = q^k\, Z_k$ with $q > 1$ and assume that $\mathsf{E}[\, Z_{k+1} \,|\, X_k \,] \leq c_k\, Z_k$ for all $k \geq 0$ in the sense of Theorem 1. Since

$$\mathsf{E}[\, \tilde{Z}_{k+1} \,|\, X_t \,] = q^{k+1}\, \mathsf{E}[\, Z_{k+1} \,|\, X_k \,] \leq q^{k+1}\, c_k\, Z_k = q\, c_k\, \tilde{Z}_k \qquad \text{a.s.}$$

for all $k \geq 0$, it suffices to find a constant $c \in (0, 1)$ with $c_k \leq c$ to ensure geometrically fast convergence to the optimum with probability 1 and in mean. For example, one may set $q = 2/(c + 1) > 1$ to guarantee that $c\, q \in (0, 1)$. Thus, it was proven:

THEOREM 2
Let $(X_k : k \geq 0)$ be the sequence of populations generated by some evolutionary algorithm and let $F_k^* = \min\{f(X_{k,1}), \ldots, f(X_{k,\mu})\}$ denote the best objective function value of the population at generation $k \geq 0$. If $\mathsf{E}[\, Z_k \,] < \infty$ and

$$\mathsf{E}[\, Z_{k+1} \,|\, X_k, X_{k-1}, \ldots, X_0 \,] \leq c\, Z_k \qquad \text{a.s.}$$

where $Z_k = F_k^* - f^*$ and $c \in (0, 1)$ for all $k \geq 0$ then the evolutionary algorithm converges with probability 1 and in mean geometrically fast to the optimum of the objective function $f(\cdot)$. □

Evidently, it cannot be expected that an evolutionary algorithm converges geometrically fast to the optimum for arbitrary objective functions. Rather, this property is likely to be restricted to a tiny subset of the set of all possible objective functions. As will be shown in the sequel, objective functions of the type introduced below are included in this subset.

DEFINITION 4
Let $f : S \subseteq \mathbb{R}^n \to \mathbb{R}$. Then f is called (K, Q)–*strongly convex* if for all $x, y \in S$ and for each $\theta \in [\,0, 1\,]$ the inequalities

$$\frac{K}{2}\,\theta\,(1-\theta)\,\|\,x-y\,\|^2 \le \theta\cdot f(x) + (1-\theta)\cdot f(y) - f(\theta\,x + (1-\theta)\,y) \le \frac{L}{2}\,\theta\,(1-\theta)\,\|\,x-y\,\|^2$$

with $0 < K \le L := K \cdot Q < \infty$ are valid. □

For example, every quadratic function $f(x) = x'Ax + b'x + c$ is (K, Q)–strongly convex if the Hessian matrix $\nabla^2 f(x) = 2\,A$ is positive definite. Another example is the function

$$f(x_1, x_2) = 2\,x_1^2 + 4\,x_2^2 + 2\,x_1 - 2\,x_2 + 2\,x_1 \arctan x_1 - \log(x_1^2 + 1) + 4\,\cos x_2\,. \qquad (2)$$

In case of twice differentiable functions, Nemirovsky and Yudin (1983), p. 255, have offered a simple condition to verify the (K, Q)–strong convexity of some function $f(\cdot)$. Let ν_1 be the smallest and ν_2 be the largest eigenvalue of the Hessian matrix. If there exist positive constants K and L such that $0 < K \le \nu_1 \le \nu_2 \le L < \infty$ for all $x \in S$ then the function $f(x)$ is (K, Q)–strongly convex with $Q = L/K$. Owing to this condition one easily finds $K = 4$ and $L = 12$ for the function given in equation (2). Alternatively, the same result can be obtained from the result below which presupposes only the availability of the gradient $\nabla f(x)$ of the function under consideration.

THEOREM 3 (Göpfert 1973, pp. 170–173)
Let $f : S \subseteq \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and S an open convex set. Then the following statements are equivalent:

(a) f is (K, Q)–strongly convex.

(b) $K\,\|\,x-y\,\|^2/2 \;\le\; f(x) - f(y) - \nabla f(y)'(x-y) \;\le\; L\,\|\,x-y\,\|^2/2$ for all $x, y \in S$.

(c) $K\,\|\,x-y\,\|^2 \;\le\; (\nabla f(x) - \nabla f(y))'(x-y) \;\le\; L\,\|\,x-y\,\|^2$ for all $x, y \in S$. □

These characterizations lead to a result that will be useful later on.

LEMMA 1
If $f : S \subseteq \mathbb{R}^\ell \to \mathbb{R}$ is differentiable and (K, Q)–strongly convex then for all $x \in S$

$$\frac{\|\,\nabla f(x)\,\|^2}{2\,L} \;\ge\; \frac{f(x) - f(x^*)}{Q^2} \qquad (3)$$

where $x^* \in S$ denotes the global minimum point of $f(\cdot)$.
PROOF:
Since $\nabla f(x^*) = 0$ for the optimum the setting $y = x^*$ in Theorem 3(c) leads to the

inequality $K \| x - x^* \|^2 \leq \nabla f(x)'(x - x^*)$ that can be further bounded by the Cauchy–Schwarz inequality yielding $K \| x - x^* \|^2 \leq \nabla f(x)'(x - x^*) \leq \| \nabla f(x) \| \cdot \| x - x^* \|$. If $\| x - x^* \| > 0$, which may be presupposed, one obtains

$$\| \nabla f(x) \|^2 \geq K^2 \| x - x^* \|^2 . \tag{4}$$

Insertion of $y = x^*$ in Theorem 3(b) delivers $\| x - x^* \|^2 \geq 2 [f(x) - f(x^*)]/L$ and together with inequality (4) one finally obtains the desired result. □

Consider a $(1, \lambda)$–EA and let the objective function $f : \mathbb{R}^n \to \mathbb{R}$ be (K, Q)–strongly convex. The current parent X_k is mutated via $X_k + r_k U_k$ where $r_k > 0$ and U_k is a random vector uniformly distributed on the boundary of the unit hyperball of dimension $n \geq 2$. Owing to Theorem 3(b) the random objective function value $f(X_k + r_k U)$ can be bounded by

$$f(X_k + r_k U) \leq f(X_k) + r_k \nabla f(X_k)' U + r_k^2 U' U \cdot L/2 . \tag{5}$$

Notice that the Euclidean length of vector U is $\|U\| = 1$ with probability 1. It follows that $U'U = \|U\|^2 = 1$ and inequality (5) reduces to

$$f(X_k + r_k U) \leq f(X_k) + r_k \nabla f(X_k)' U + r_k^2 L/2 . \tag{6}$$

For further simplifications we need the following result.

LEMMA 2 (Yin, Rudolph, and Schwefel 1996, p. 479)
If U is a random vector uniformly distributed on the boundary of the unit hyperball of dimension $n \geq 2$ and $x \in \mathbb{R}^n$ with $\|x\| = 1$, then the random scalar product $B = -x'U$ possesses a Beta distribution with probability density function

$$p(x) = \frac{2^{2-n} (1 - x^2)^{(n-3)/2}}{B(\frac{n-1}{2}, \frac{n-1}{2})} \cdot 1_{(-1,1)}(x)$$

where $B(\cdot, \cdot)$ denotes the complete Beta function and $1_A(x)$ is the indicator function of some set A. The mean, mode, and median of B is zero while the variance is $1/n$. □

Thus, inequality (6) is equivalent to

$$f(X_k + r_k U_k) \leq f(X_k) - r_k \|\nabla f(X_k)\| B + r_k^2 L/2 \tag{7}$$

where B is a Beta random variable as specified in Lemma 2. Since the $(1, \lambda)$–EA generates λ offspring and chooses the best among them to serve as the new parent, the random objective function value of the new parent is equivalent to the value of the best offspring and it can be bounded via

$$f(X_{k+1}) = \min\{f(X_k + r_k U_i) : i = 1, \ldots, \lambda\} \leq f(X_k) - r_k \|\nabla f(X_k)\| B_{\lambda:\lambda} + r_k^2 L/2 \tag{8}$$

by taking into account inequality (7) and where $B_{\lambda:\lambda}$ denotes the maximum of λ independent and identically distributed Beta random variables. Taking conditional expectations on both sides of (8) yields

$$\mathsf{E}[f(X_{k+1}) \,|\, X_k] \leq f(X_k) - r_k \|\nabla f(X_k)\| \, \mathsf{E}[B_{\lambda:\lambda}] + r_k^2 L/2. \tag{9}$$

Differentiation of (9) with respect to r_k leads to the optimal choice

$$r_k^* = \frac{\|\nabla f(X_k)\|\, M_\lambda}{L} \tag{10}$$

where $M_\lambda = \mathsf{E}[\, B_{\lambda:\lambda}\,]$. After insertion of r_k^* into inequality (9) and subtraction of f^* on both sides, inequality (9) becomes

$$\begin{aligned}
\mathsf{E}[\, f(X_{k+1}) - f^* \,|\, X_k\,] &\leq f(X_k) - f^* - \frac{\|\nabla f(X_k)\|^2\, M_\lambda^2}{2\,L} && (11)\\
&\leq f(X_k) - f^* - \frac{(f(X_k) - f^*)\, M_\lambda^2}{Q^2} \\
&= \left(1 - \frac{M_\lambda^2}{Q^2}\right) \cdot (f(X_k) - f^*) && (12)
\end{aligned}$$

by inserting inequality (3) given in Lemma 1 into to the r.h.s. of inequality (11). Owing to Theorem 1 and inequality (12) it is guaranteed that the $(1,\lambda)$–EA will converge with probability 1 and in mean to the optimum, provided that $M_\lambda > 0$. Moreover, since $c = 1 - M_\lambda^2/Q^2 \in (0,1)$ for $M_\lambda > 0$ it is guaranteed by Theorem 2 that the rate of approach to the optimum is geometric in mean and with probability 1. To show that $M_\lambda > 0$ for $\lambda \geq 2$ the result below is useful.

LEMMA 3 (David 1970, p. 8)
Let $Y_1, \ldots, Y_\lambda$ be independent and identically distributed continuous random variables with probability density function $p(\cdot)$ and distribution function $P(\cdot)$. If these random variables are ordered such that $Y_{1:\lambda} \leq Y_{2:\lambda} \leq \ldots \leq Y_{\lambda:\lambda}$ then the probability density function of $Y_{i:\lambda}$ is

$$p_{i:\lambda}(x) = \frac{p(x)\, P^{i-1}(x)\, [\,1 - P(x)\,]^{\lambda - i}}{B(i, \lambda - i + 1)}$$

where $B(\cdot,\cdot)$ denotes the complete Beta function. □

Since the probability density function of random variable B is symmetrical with respect to zero the identities $p(-x) = p(x)$ and $P(-x) = 1 - P(x)$ are valid. It follows that $P(x) > 1/2$ for $x > 0$ and hence

$$M_2 = \mathsf{E}[\, B_{2:2}\,] = \int_{-\infty}^{\infty} x\, p_{2:2}(x)\, dx = 2 \int_{-1}^{1} x\, p(x)\, P(x)\, dx = 2 \int_{0}^{1} x\, p(x)\, [\,2\,P(x) - 1\,]\, dx > 0$$

(because of the positivity of the integrand for $x > 0$) and finally $M_\lambda \geq M_2 > 0$ for $\lambda \geq 2$. Notice that the actual values of M_λ also nonlinearly depend on the dimension n. Despite this fact there is — in principle — no problem to calculate M_λ for each $n \geq 2$. For example,

$$M_\lambda = \begin{cases} \displaystyle\sum_{i=1}^{\lfloor \lambda/2 \rfloor} \frac{(2\,i)!}{\pi^{2\,i}} \binom{\lambda}{2\,i} (-1)^{i+1} + (1 - \lambda \bmod 2)\, (-1)^{\lambda/2+1}\, \frac{\lambda!}{\pi^\lambda} & \text{if } n = 2, \\ \dfrac{\lambda - 1}{\lambda + 1} & \text{if } n = 3 \end{cases}$$

with $\lambda \in \mathbb{N}$. Apart from few exceptional cases, however, the resulting expressions become more and more complicated the larger is the value of n. Therefore, the next section is devoted to the investigation of the asymptotics of constant $M_{\lambda,n}$.

4 Asymptotics

Since $M_{\lambda,n}$ depends on two parameters one has to investigate two subcases: First, the asymptotics of $M_{\lambda,n}$ for fixed λ and $n \to \infty$. Second, the asymptotics for fixed n and $\lambda \to \infty$. The basic technique rests on the idea to approximate the distributions of the random variables by *appropriate* limit distributions. For this purpose one needs the notion of the *weak convergence* of probability measures.

DEFINITION 5
Let $\{P(x), P_i(x) : i \in \mathbb{N}\}$ be a collection of distribution functions of the random variables $\{Y, Y_i : i \in \mathbb{N}\}$ on some probability space. If $P_i(x) \to P(x)$ as $i \to \infty$ for every continuity point x of $P(\cdot)$, then the sequence $P_i(\cdot)$ of distribution functions is said to *converge weakly* to $P(\cdot)$, denoted as $P_i \stackrel{w}{\to} P$. In such an event, the sequence of random variables Y_i is said to *converge in distribution* to Y, denoted as $Y_i \stackrel{d}{\to} Y$. □

Convergence in distribution is implied by convergence in probability whereas the converse is wrong in general (Lukacs 1975, p. 33). If the random variables are continuous and the sequence of probability density functions converge to the limit variable's p.d.f. for each of its continuity points then a theorem in Scheffé (1947) ensures weak convergence of the associated distributions functions. This result may be used for the case with fixed $\lambda \in \mathbb{N}$ and $n \to \infty$.

LEMMA 4
Let B_n be a Beta random variable parametrized by $n \geq 2$ as specified in Lemma 2. If $Y_n = \sqrt{n}\, B_n$ then $Y_n \stackrel{d}{\to} Y \sim N(0,1)$ as $n \to \infty$, where $N(0,1)$ denotes the standard normal distribution with zero mean and unit variance. □
PROOF:
As mentioned previously it suffices to show that the probability density functions $q_n(\cdot)$ of the random variables Y_n converge to the probability density function $q(\cdot)$ of the standard normal random variable Y for every continuity point of $q(\cdot)$ as $n \to \infty$. Let $p_n(\cdot)$ denote the p.d.f. of random variable B_n. Since $Y_n = \sqrt{n}\, B_n$ the probability density transformation rule leads to the p.d.f.

$$q_n(x) = \frac{1}{\sqrt{n}}\, p_n\left(\frac{x}{\sqrt{n}}\right) = \frac{1}{\sqrt{n}}\, \frac{2^{2-n}}{B((n-1)/2, (n-1)/2)} \left(1 - \frac{x^2}{n}\right)^{(n-3)/2} 1_{(-\sqrt{n},\sqrt{n})}(x).$$

Note that

$$\begin{aligned}
\frac{1}{\sqrt{n}}\, \frac{2^{2-n}}{B((n-1)/2, (n-1)/2)} &\to \frac{1}{\sqrt{2\,\pi}} \\
\left(1 - \frac{x^2}{n}\right)^{(n-3)/2} &\to \exp\left(-\frac{x^2}{2}\right) \\
1_{(-\sqrt{n},\sqrt{n})}(x) &\to 1_{(-\infty,\infty)}(x)
\end{aligned}$$

as $n \to \infty$ for every fixed $x \in \mathbb{R}$. Thus, one obtains

$$\lim_{n\to\infty} q_n(x) = \frac{1}{\sqrt{2\,\pi}} \exp\left(-\frac{x^2}{2}\right) 1_{(-\infty,\infty)}(x) \tag{13}$$

for every continuity point $x \in \mathbb{R}$ of $q(\cdot)$. Since the r.h.s. of equation (13) is the p.d.f. of a standard normal random variable it has been shown that $\sqrt{n}\, B_n \stackrel{d}{\to} Y \sim N(0,1)$ as $n \to \infty$. □

An immediate consequence of this lemma is the result below:

THEOREM 4
Let $B_{\lambda:\lambda}(n)$ be the maximum of λ independent and identically distributed (i.i.d.) Beta random variables as in Lemma 4 (parametrized by $n \geq 2$). If $Y_{\lambda:\lambda}$ denotes the maximum of λ i.i.d. standard normal random variables, then $\sqrt{n}\, B_{\lambda:\lambda}(n) \stackrel{d}{\to} Y_{\lambda:\lambda}$ as $n \to \infty$.
PROOF:
Let $q_n(\cdot)$ and $Q_n(\cdot)$ be the probability density function and the distribution function of random variable $\sqrt{n}\, B(n)$, respectively. Recall from Lemma 4 that $q_n \stackrel{w}{\to} q$ as well as $Q_n \stackrel{w}{\to} Q$ as $n \to \infty$, where $q(\cdot)$ and $Q(\cdot)$ denote the p.d.f. respective distribution function of a standard normal random variable. Owing to this fact and Lemma 3 one obtains for *fixed* $\lambda \in \mathbb{N}$

$$q_{\lambda,\lambda;n}(x) \;=\; \lambda\, q_n(x)\, Q_n^{\lambda-1}(x) \;\;\longrightarrow\;\; \lambda\, q(x)\, Q^{\lambda-1}(x) \;=\; q_{\lambda,\lambda}(x)$$

for every continuity point of $q_{\lambda,\lambda}(x)$ as $n \to \infty$. This ensures that the distribution functions associated with random variables $\sqrt{n}\, B_{\lambda:\lambda}(n)$ converges weakly to the distribution function of the maximum of λ i.i.d. standard normal random variables as $n \to \infty$. □

This result reveals that the distribution of random variable $B_{\lambda:\lambda}(n)$ is approximately equal to the distribution of $n^{-1/2}\, Y_{\lambda:\lambda}$ for large n and hence

$$M_{\lambda,n} \;=\; \mathsf{E}[\, B_{\lambda:\lambda}(n)\,] \;\approx\; n^{-1/2}\, \mathsf{E}[\, Y_{\lambda:\lambda}\,] \;=\; n^{-1/2}\, C_\lambda$$

where $C_\lambda = \mathsf{E}[\, Y_{\lambda:\lambda}\,]$ denotes the mean of λ standard normal random variables. The actual values of C_λ can be analytically determined and expressed in terms of elementary functions up to $\lambda = 5$ (see David 1970, pp. 30–34). In general, the bounds

$$\Phi^{-1}\left(1 - \frac{1}{\lambda}\right) \;\leq\; C_\lambda \;\leq\; \Phi^{-1}\left(1 - \frac{1}{2\,\lambda}\right) \tag{14}$$

are valid (David 1970, p. 64), where $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal distribution function. Since $C_2 = \pi^{-1/2} \approx 0.5642$, $C_{1000} < 3.2415$, and C_λ must increase monotonically, it is obvious that the rate of increase must decline considerably for increasing λ. In fact, taking into account that the value of C_λ can be bracketed as noted in (14) it can be shown (David 1970, p. 209) that roughly $C_\lambda \approx (2 \log \lambda)^{1/2}$ for sufficiently large λ. Accurate approximations of C_λ can be easily obtained by numerical integration. They are tabulated, for example, in Rechenberg (1994), pp. 236–240, up to $\lambda = 1000$. Figure 1 reveals that the approximation $M_{\lambda,n} \approx n^{-1/2}\, C_\lambda$ becomes more accurate the larger is the value of $n \geq 2$. For example, even for the relatively low dimension $n = 31$ the relative error is less than 3 % for $\lambda \leq 80$.
As for the second subcase with fixed $n \geq 2$ and $\lambda \to \infty$, some results from the asymptotical theory of extreme order statistics are needed.

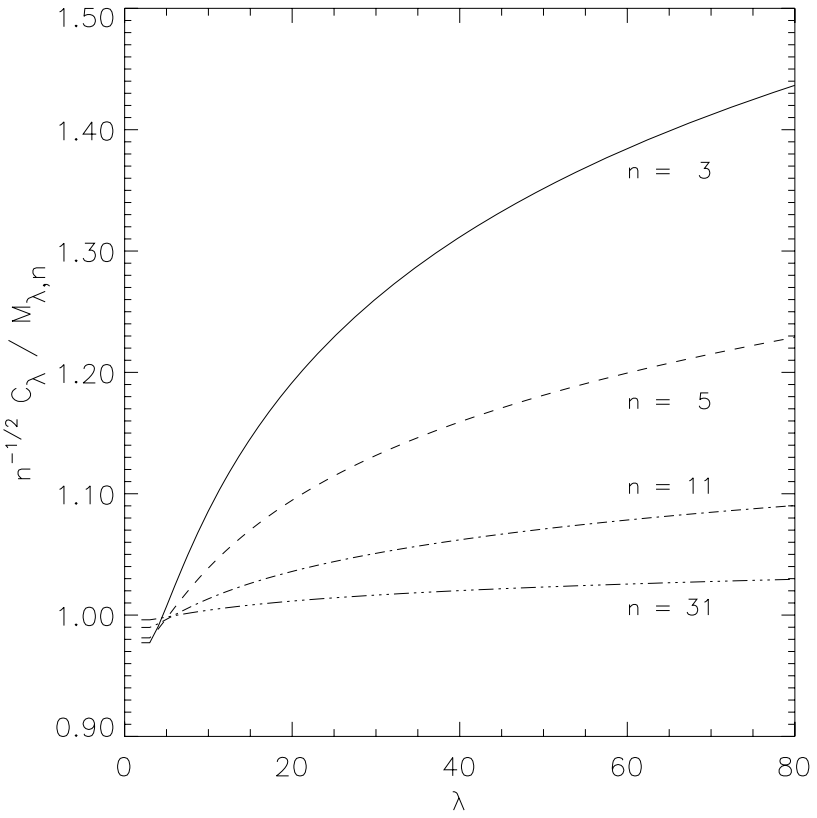

Figure 1: Ratio between the approximation $n^{-1/2}\,C_\lambda$ and the exact values of $M_{\lambda,n}$.

THEOREM 5 (Leadbetter et al. 1983, Chapter 1)
Let $\hat{x} = \sup\{x \in \mathbb{R} : P(x) < 1\} < \infty$, $\alpha > 0$, and $Y_{\lambda:\lambda}$ be the maximum of λ independent and identically distributed continuous random variables which possess distribution function $P(\cdot)$. The following statements are equivalent:

$$\text{(a)} \qquad \lim_{h \to 0^+} \frac{1 - P(\hat{x} - x\,h)}{1 - P(\hat{x} - h)} = x^\alpha \qquad \text{for all } x > 0\,.$$

$$\text{(b)} \qquad \mathsf{P}\{\, a_\lambda\,(Y_{\lambda:\lambda} - b_\lambda) \le x\,\} \;\xrightarrow{w}\; G_\alpha(x) \;=\; \begin{cases} \exp(-(-x)^\alpha) & , x < 0 \\ 1 & , x \ge 0 \end{cases}$$

where $a_\lambda = (\hat{x} - \gamma_\lambda)^{-1}$, $b_\lambda = \hat{x}$, and $\gamma_\lambda = \inf\{x \in \mathbb{R} : P(x) \ge 1 - 1/\lambda\}$. □

Let W be the random variable with distribution function $G_\alpha(\cdot)$ as given above and suppose that part (a) of Theorem 5 holds true for Beta random variables with a distribution function as specified in Lemma 2. Then part (b) of the theorem reveals that $B_{\lambda:\lambda}$ has approximately the same distribution as $a_\lambda^{-1}\,W + b_\lambda$ for large λ and one may approximate the mean via $\mathsf{E}[B_{\lambda:\lambda}] \approx a_\lambda^{-1}\,\mathsf{E}[W] + b_\lambda$. Since $\hat{x} = 1$ and $-W$ is Weibull distributed

with $\mathsf{E}[-W] = \Gamma(1+1/\alpha)$ one obtains

$$\mathsf{E}[\,B_{\lambda:\lambda}\,] \approx 1 - (1-\gamma_\lambda)\cdot\Gamma(1+\alpha^{-1})\,.$$

Notice that part (a) of Theorem 5 indeed holds true since

$$\lim_{h\to 0+}\frac{1-P_n(1-x\,h)}{1-P_n(1-h)} \;=\; \lim_{h\to 0+}\frac{x\,p_n(1-x\,h)}{p_n(1-h)} \;=\; \lim_{h\to 0+}\;x\,\left(\frac{2\,x-x^2\,h}{2-h}\right)^{\frac{n-3}{2}} \;=\; x^{(n-1)/2}$$

where $P_n(\cdot)$ and $p_n(\cdot)$ are the distribution respective probability density function of the Beta random variable parametrized by $n \geq 2$. Thus, $\alpha = (n-1)/2$. It remains to find an expression for γ_λ. For this purpose one has to find at least an asymptotical solution of the equation $P_n(\gamma_\lambda) = 1-\lambda^{-1}$ for $\lambda \to \infty$. Since the p.d.f. is symmetrical with respect to zero an equivalent condition is

$$P_n(-\gamma_\lambda) = 1/\lambda\,. \tag{15}$$

It is clear that necessarily $\gamma_\lambda \to 1$ as $\lambda \to \infty$. Notice that

$$P_n(x) = \mathsf{P}\{\,B < x\,\} = \mathsf{P}\{\,2\,\tilde{B}-1 < x\,\} = \mathsf{P}\left\{\,\tilde{B} < \frac{x+1}{2}\,\right\} = \tilde{P}_n\left(\frac{x+1}{2}\right)$$

where $\tilde{B}$ is a Beta random variable with probability density function

$$\tilde{p}_n(x) = \frac{x^{(n-3)/2}\,(1-x)^{(n-3)/2}}{B(\frac{n-1}{2},\frac{n-1}{2})}\;\cdot\,1_{(0,1)}(x)\,.$$

Using the relationships above, condition (15) changes to

$$P_n(-\gamma_\lambda) = \tilde{P}_n(\tilde{\gamma}_\lambda) = 1/\lambda \tag{16}$$

where $\gamma_\lambda = 1-2\,\tilde{\gamma}_\lambda$. Entry 26.5.23 in Abramowitz and Stegun (1965) reveals that the distribution function of $\tilde{B}_n$ can be expressed by the Gauss hypergeometric series

$$\tilde{P}(x) = \frac{2\,x^{(n-1)/2}}{(n-1)\,B(\frac{n-1}{2},\frac{n-1}{2})}\;{}_2F_1\left(\frac{n-1}{2},\,-\frac{n-3}{2};\,\frac{n+1}{2};\,x\right)\,,$$

where ${}_2F_1(\cdot)$ stands for the Gauss series as defined in entry 15.1.1 in Abramowitz and Stegun (1965), that reduces to a polynomial in x for odd $n \geq 3$. It suffices to consider this special case. As a result, the condition (16) becomes

$$\lambda\,\tilde{P}(\tilde{\gamma}_\lambda) = \frac{2\,\lambda\,\tilde{\gamma}_\lambda^{(n-1)/2}}{(n-1)\,B(\frac{n-1}{2},\frac{n-1}{2})}\;\left[1+\sum_{i=1}^{\frac{n-3}{2}}\binom{\frac{n-3}{2}}{i}\,(-1)^i\,\frac{n-1}{n-1+2\,i}\,\tilde{\gamma}_\lambda^{\,i}\right] = 1\,. \tag{17}$$

Notice that necessarily $\tilde{\gamma}_\lambda \to 0$ since $\gamma_\lambda \to 1$ as $\lambda \to \infty$. In this case the term in the brackets of equation (17) converges to 1 because each term in the sum converges to zero for $\tilde{\gamma} \to 0$. The term left to the term in the brackets describes the asymptotics of the entire expression since it contains the least power of $\tilde{\gamma}_\lambda$, namely of order $(n-1)/2$,

whereas all other terms are of higher order converging faster to zero than $\tilde{\gamma}_\lambda^{(n-1)/2}$. Therefore one may approximate condition (17) by the asymptotical condition

$$\frac{2\,\lambda\,\tilde{\gamma}_\lambda^{(n-1)/2}}{(n-1)\,B(\frac{n-1}{2},\frac{n-1}{2})} = 1\,.$$

The solution of this equation is

$$\tilde{\gamma}_\lambda \;=\; \left[\frac{n-1}{2}\,B\left(\frac{n-1}{2},\,\frac{n-1}{2}\right)\right]^{\frac{2}{n-1}} \cdot \lambda^{-2/(n-1)}$$

which leads—after several resubstitutions—to the final asymptotical expression

$$M_{\lambda,n} \approx \widehat{M}_{\lambda,n} := 1 - 2\left[\frac{n-1}{2}\,B\left(\frac{n-1}{2},\,\frac{n-1}{2}\right)\right]^{\frac{2}{n-1}} \cdot \Gamma\left(1+\frac{2}{n-1}\right) \cdot \lambda^{-2/(n-1)}$$

for large λ and fixed $n \geq 2$. The quality of the approximation may be expressed by the ratio $\widehat{M}_{\lambda,n}/M_{\lambda,n}$. To avoid waste of computing time only few ratios have been calculated[1]. They are summarized in Table 1 below.

$n \setminus \lambda$	5	11	31	50	100
5	1.040	1.019	1.007	1.004	1.002
9	1.289	1.140	1.064	1.046	1.029
15	1.617	1.337	1.181	1.141	1.101
21	1.902	1.522	1.302	1.244	1.184
35	2.456	1.899	1.565	1.473	1.376

Tab. 1: Ratios $\widehat{M}_{\lambda,n}/M_{\lambda,n}$.

Two observations can be made. First, the ratio approaches 1 for increasing λ. Second, the asymptotical expression $\widehat{M}_{\lambda,n}$ is closer to the true value $M_{\lambda,n}$ the smaller is the value of n. The latter observation is not surprising because the neglected finite sum in the brackets of condition (17) contains products in which n appears in type of binomial coefficients. As a consequence, the larger is the value of n the larger must be the value of λ such that the sum becomes sufficiently small.

5 Conclusions

In the course of section 3 it was tacitly presupposed that the evolutionary algorithm has access to a subroutine that returns the Euclidean length of the gradient—an assumption that is usually not justified in practice. But it can be shown (Rudolph 1997, pp. 191–192) that it is sufficient to estimate the gradients' length up to a relative error of 99.9 % to ensure geometrical convergence rates in case of (K,Q)–strongly convex functions. In real world evolutionary algorithms this task is accomplished by a mechanism termed 'auto–adaptation' (see e.g. Bäck and Schwefel 1993), but a mathematically rigorous proof of this property is still pending.

[1]Note that $M_{\lambda,n}$ is a rational number for odd $n \geq 3$. It can be exactly calculated but the costs to do so are not neglectable. For example, both the nominator and denominator of $M_{100,35}$ are integers with 1367 digits each and it took more than 30 hours CPU time to obtain them.

Several results presented here can be sharpened. Theorem 2 remains valid if 'convergence with probability 1' is replaced by the stronger property of 'complete convergence' (this concept was introduced in Hsu and Robbins 1947). The proofs of Lemma 4 and Theorem 4 show the convergence of the probability density functions for each continuity point of the limit random variable's probability density function. This is actually stronger than the weak convergence of the distribution functions. But the demonstration of these subtle differences was omitted in favor of an easy presentation.
Finally, it should be noticed that the convergence rates derived for the $(1, \lambda)$–EA are also valid for the practically more relevant (μ, λ)–EA. Here, each of the μ parents generates $m = \lambda/\mu \in \mathbb{N}$ offspring ($m \geq 2$). Under the assumption that the differences between parents and offspring are solely caused by mutations, then the convergence rate of the (μ, λ)–EA can be bounded by $c \leq 1 - M_{m,n}^2/Q^2 \in (0, 1)$ for (K, Q)–strongly convex functions.

Acknowledgments

The major part of the work presented was supported by the German Federal Ministry of Education, Science, Research, and Technology (BMB+F), grant 01 IB 403 A, while the author was with the Informatik Centrum Dortmund (ICD), Germany. The opportunity to complete the work was offered by the Research Center "Computational Intelligence" (SFB 531) at the University of Dortmund. Financial support by the German Research Foundation (DFG) is gratefully acknowledged.

References

Abramowitz, M. and I. A. Stegun (Eds.) (1965). *Handbook of Mathematical Functions.* New York: Dover Publications.

Bäck, T. and H.-P. Schwefel (1993). An overview of evolutionary algorithms for parameter optimization. *Evolutionary Computation 1*(1), 1–23.

Beyer, H.-G. (1994). Towards a theory of 'evolution strategies': Results for $(1 \dot{+} \lambda)$–strategies on (nearly) arbitrary fitness functions. In Y. Davidor, H.-P. Schwefel, and R. Männer (Eds.), *Parallel Problem Solving from Nature—PPSN III*, pp. 58–67. Berlin: Springer.

Born, J. (1978). *Evolutionsstrategien zur numerischen Lösung von Adaptationsaufgaben.* Dissertation A, Humboldt-Universität, Berlin.

David, H. A. (1970). *Order Statistics.* New York: Wiley.

Devroye, L. P. (1976). On the convergence of statistical search. *IEEE Transactions on Systems, Man, and Cybernetics 6*(1), 46–56.

Göpfert, A. (1973). *Mathematische Optimierung in allgemeinen Vektorräumen.* Leipzig: Teubner.

Hsu, P. L. and H. Robbins (1947). Complete convergence and the law of large numbers. *Proceedings of the National Academy of Sciences of the U.S.A. 33*, 25–31.

Leadbetter, M. R., G. Lindgren, and H. Rootzén (1983). *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer.

Lukacs, E. (1975). *Stochastic Convergence* (2nd ed.). New York: Academic Press.

Nemirovsky, A. S. and D. B. Yudin (1983). *Problem complexity and method efficiency in optimization*. Chichester: Wiley.

Neveu, J. (1975). *Discrete–Parameter Martingales*. Amsterdam and Oxford: North Holland.

Oppel, U. G. and M. Hohenbichler (1978). Auf der Zufallssuche basierende Evolutions–prozesse. In B. Schneider and U. Ranft (Eds.), *Simulationsmethoden in der Medizin und Biologie*, pp. 130–155. Berlin: Springer.

Pintér, J. (1984). Convergence properties of stochastic optimization procedures. *Mathematische Operationsforschung und Statistik, Series Optimization 15*, 405–427.

Rappl, G. (1989). On linear convergence of a class of random search algorithms. *Zeitschrift für angewandte Mathematik und Mechanik (ZAMM) 69*(1), 37–45.

Rastrigin, L. A. (1963). The convergence of the random search method in the extremal control of a many-parameter system. *Automation and Remote Control 24*, 1337–1342.

Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Stuttgart: Frommann–Holzboog Verlag.

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

Rudolph, G. (1994). Convergence of non–elitist strategies. In *Proceedings of the First IEEE Conference on Computational Intelligence, Vol. 1*, pp. 63–66. IEEE Press.

Rudolph, G. (1997). *Convergence Properties of Evolutionary Algorithms*. Hamburg: Kovač.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics 18*(3), 434–438.

Schumer, M. A. and K. Steiglitz (1968). Adaptive step size random search. *IEEE Transactions on Automatic Control 13*, 270–276.

Schwefel, H.-P. (1977). *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*. Basel: Birkhäuser.

Solis, F. J. and R. J.-B. Wets (1981). Minimization by random search techniques. *Mathematics of Operations Research 6*, 19–30.

Yin, G., G. Rudolph, and H.-P. Schwefel (1996). Analyzing $(1, \lambda)$ evolution strategy via stochastic approximation methods. *Evolutionary Computation 3*(4), 473–489.