Asymptotical Convergence Rates of Simple Evolutionary Algorithms under Factorizing Mutation Distributions

Günter Rudolph

Universität Dortmund, Fachbereich Informatik, LS XI, D–44221 Dortmund

Abstract. The standard choice for mutating an individual of an evolutionary algorithm with continuous variables is the normal distribution. It is shown that there is a broad class of alternative mutation distributions offering local convergence rates being asymptotical equal to the convergence rates achieved with normally distributed mutations. Such mutation distributions must be factorizing and the absolute fourth moments must be finite. Under these conditions an asymptotical theory of the convergence rates of simple evolutionary algorithms can be established for the entire class of distributions.

1 Introduction

The standard choice to represent mutations in evolutionary models dealing with continuous quantities is the normal distribution. This choice is usually justified by the central limit theorem: Since mutations in nature are caused by a variety of physical and chemical influences that are not identifiable or measurable to a degree that allows for a deterministic model, these influences are considered as independent random perturbations whose normed sum approaches a normal random variable in the limit, provided that the first two absolute moments of the distributions of these random perturbations are finite and that the so–called Lindeberg condition is obeyed. Therefore it is not surprising that evolutionary algorithms with continuous search space model mutations by normally distributed random variables as well. But the biological original needs not necessarily be the best choice when mutations play the role of an exploration operator—as it is the case in evolutionary algorithms (EAs). It was noted several times [1, 2, 3] that mutation distributions with slowly (i.e., not exponentially) decreasing tails should offer a larger probability to escape from local optima (also see fig. 1 & 2). Although this claim is certainly correct if the variance is held fixed, it is still an open question whether this theoretical property carries over to practical EAs employing an auto–adaptive adjustment of the variances. But this question will not be addressed here. Instead, it is investigated to which extent non–normal mutation distributions may affect the local convergence behavior of evolutionary algorithms.

Two simple evolutionary algorithms will be studied here: The $(1+1)$–EA and the $(1, \lambda)$–EA. The first one generates a single offspring by mutation and accepts the offspring only if it is better than the parent, whereas the latter one generates

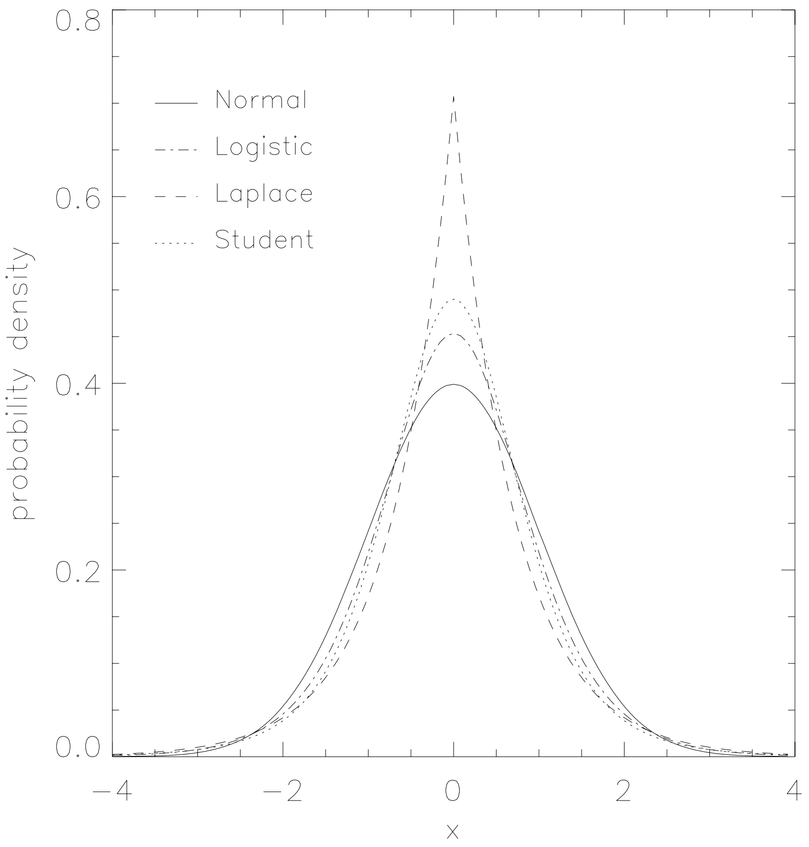

Fig. 1. The shape of some symmetrical univariate distributions with zero mean and unit variance.

distribution	density function
Normal	$\exp(-x^2/2)/\sqrt{2\,\pi}$
Logistic	$\pi\,\mathrm{sech}^2(\pi\,x/\sqrt{12})/\sqrt{48}$
Laplace	$\exp(-\lvert x\rvert\,\sqrt{2})/\sqrt{2}$
Student (d=5)	$8\,(1+x^2/3)^{-3}/\sqrt{27\,\pi^2}$

Table 1. Probability density functions of some symmetrical univariate distributions with zero mean and unit variance.

$\lambda \geq 2$ offspring independently with the same mutation distribution and chooses the best offspring among the λ offspring to serve as new parent (regardless of

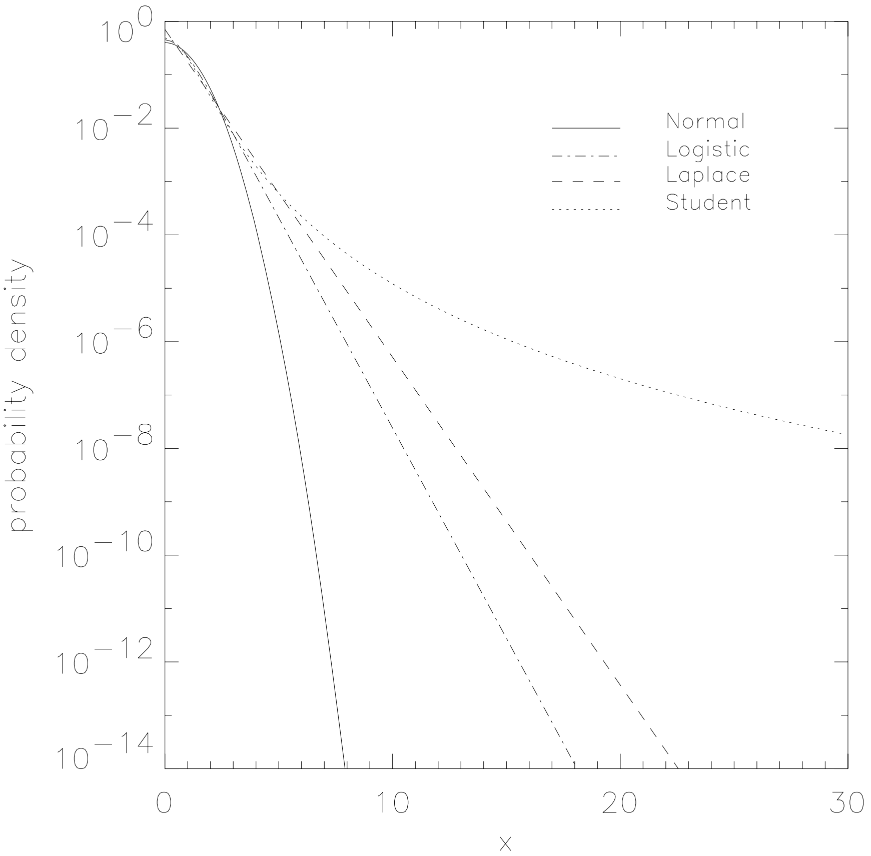

Fig. 2. The decay of the right tails of the symmetric Normal, Logistic, Laplace, and Student distribution. The tails of the first three distributions decline exponentially whereas the tail of the Student distribution (with 5 degrees of freedom) follows a power law.

the quality of the old parent). If $\theta \in \mathbb{R}^n$ denotes the current position of the EA in the search space, then a mutation is modeled by adding a random vector $\mathbf{Z}$ that must fulfill some conditions (details will follow shortly). Thus, an offspring $\mathbf{X}$ is represented by the random variable $\mathbf{X} = \theta + \mathbf{Z}$.

The test problem is the minimization of the objective function $f(\mathbf{x}) = \mathbf{x}'\,\mathbf{x}$ with $\mathbf{x} \in \mathbb{R}^n$. It will be assumed that n is large ($n \geq 100$). This test function reflects to some extent the case of a local optimum, and it is usually used to assess the local convergence behavior of evolutionary algorithms. To be comparable to previous work, this common practice is followed here.

2 Asymptotical Results

The *fundamental assumption* made in the remainder is that it will be postulated that the product moments of random vector $\mathbf{Z}$ do exist up to order 4. Further conditions on $\mathbf{Z}$ are given in the definition below which specifies the distribution class of random vector $\mathbf{Z}$.

Definition 1. The distribution of random vector $\mathbf{Z}$ is termed a *mutation distribution* if $\mathsf{E}[\,\mathbf{Z}\,] = \mathbf{0}$. In this case, random vector $\mathbf{Z}$ is called a *mutation vector*. A mutation distribution is said to be *factorizing* if the joint probability density function of the mutation vector $\mathbf{Z}$ can be written as

$$f_{\mathbf{Z}}(z_1, \ldots, z_n) = \prod_{i=1}^{n} f_{Z_i}(z_i)$$

with $f_{Z_1}(\cdot) = \ldots = f_{Z_n}(\cdot)$ where n denotes the dimension. □

Let $\mathbf{Z}$ possess a factorizing mutation distribution. Since the random objective function value of an offspring is given by

$$f(\theta + \mathbf{Z}) = \sum_{i=1}^{n} (\theta_i + Z_i)^2$$

each of the summands above is mutually independent to the remaining ones. As a consequence, the objective function value is representable by a sum of independent random variables. If such a sum is appropriately normed, then its distribution converges to some limit distribution as $n \to \infty$. This fact will be exploited to develop an asymptotical theory with regard to the convergence rates. In order to obtain the desired norming constants some preparatory results are necessary.

Lemma 2. Let Z be a symmetrical random variable with $\mathsf{E}[\,Z^{2k-1}\,] = 0$ for $k \in \mathbb{N}$ and set $X = \theta + Z$ with $\theta \in \mathbb{R}$. Then $\mathsf{E}[\,X^2\,] = \theta^2 + \mathsf{E}[\,Z^2\,]$ and $\mathsf{V}[\,X^2\,] = 4\,\theta^2\,\mathsf{E}[\,Z^2\,] + \mathsf{V}[\,Z^2\,]$. □

The proof of this lemma is trivial and therefore omitted while the next result is an immediate consequence of the lemma above.

Proposition 3. Let $Z_1, \ldots, Z_n$ be independent and identically distributed symmetrical random variables with $\mathsf{E}[\,Z_i^{2k-1}\,] = 0$ for $i = 1, \ldots, n$ and $k \in \mathbb{N}$. If $X_i = \theta_i + Z_i$ and $S_n = \sum_{i=1}^{n} X_i^2$ then

$$\begin{aligned} \mathsf{E}[\,S_n\,] &= \|\theta\|^2 + n\,\mathsf{E}[\,Z_1^2\,] \\ \mathsf{V}[\,S_n\,] &= 4\,\|\theta\|^2\,\mathsf{E}[\,Z_1^2\,] + n\,\mathsf{V}[\,Z_1^2\,] \end{aligned}$$

where $\theta \in \mathbb{R}^n$ and $\|\cdot\|$ denotes the Euclidean norm. □

The central limit theorem (see [4], p. 262) ensures that the distribution of the appropriately normed random scalar product $S_n = \mathbf{X}'\,\mathbf{X} = \|\mathbf{X}\|^2$ converges weakly to the standard normal distribution. Thus, since

$$\frac{S_n - \mathsf{E}[\,S_n\,]}{\mathsf{V}[\,S_n\,]^{1/2}} \;\longrightarrow\; N \sim N(0,1)$$

as $n \to \infty$ one obtains

$$\begin{aligned} S_n &\approx \mathsf{E}[\,S_n\,] + \mathsf{V}[\,S_n\,]^{1/2} \cdot N \\ &= \|\theta\|^2 + n\,\mathsf{E}[\,Z^2\,] + (4\,\|\theta\|^2\,\mathsf{E}[\,Z^2\,] + n\,\mathsf{V}[\,Z^2\,]\,)^{1/2} \cdot N. \end{aligned} \tag{1}$$

Let $\eta^2 = \mathsf{V}[\,Z\,] = \mathsf{E}[\,Z^2\,]$ and suppose that $\mathsf{V}[\,Z^2\,] = a\,\eta^4 = a\,\mathsf{V}[\,Z\,]^2$ for some $a > 0$. Then the random variable $S_n/\|\theta\|^2$ can be written as

$$\frac{S_n}{\|\theta\|^2} \approx 1 + \frac{1}{n}\left[\gamma^2 + \left(4\,\gamma^2 + \frac{a\,\gamma^4}{n}\right)^{1/2} \cdot N\right] \tag{2}$$

where $\gamma = n\,\eta/\|\theta\|$. After having established this approximation one can begin to calculate the expected asymptotical progress rates for the $(1+1)$–EA and the $(1,\lambda)$–EA provided that the objective function is $f(\mathbf{x}) = \|\mathbf{x}\|^2$. At first consider the $(1+1)$–EA. Assume the current position is $\theta \in \mathbb{R}^n$. Since the $(1+1)$–EA only accepts improvements the relative progress is given by

$$\mathsf{E}\left[\max\left\{\frac{\|\theta\|^2 - \|\theta + \mathbf{Z}\|^2}{\|\theta\|^2},\, 0\right\}\right] = \mathsf{E}\left[\max\left\{1 - \frac{S_n}{\|\theta\|^2},\, 0\right\}\right].$$

It will be useful to normalize the relative progress by the dimension n. This quantity will be called *normalized* progress. Owing to eqn. (2) one obtains the normalized progress

$$\mathsf{E}\left[\max\left\{n\,\left(1 - \frac{S_n}{\|\theta\|^2}\right),\, 0\right\}\right] \approx \mathsf{E}[\,\max\{-\gamma^2 \,+ \gamma\,\sqrt{4 + a\,\gamma^2/n} \cdot N, 0\}\,]\,. \tag{3}$$

Proposition 4. Let $\eta^2 = \mathsf{V}[\,Z\,] = \mathsf{E}[\,Z^2\,]$ and suppose that $\mathsf{V}[\,Z^2\,] = a\,\eta^4 = a\,\mathsf{V}[\,Z\,]^2$ for some $a > 0$. If $n \gg 1$ then the expected normalized progress rate of the $(1+1)$–EA is asymptotically given by

$$h(\gamma, a, n) = 2\,\gamma\,\sqrt{1 + \frac{a\,\gamma^2}{4\,n}}\;\varphi\left(\frac{\gamma}{2}\sqrt{\frac{4\,n}{4\,n + a\,\gamma^2}}\right) - \gamma^2\,\Phi\left(-\frac{\gamma}{2}\sqrt{\frac{4\,n}{4\,n + a\,\gamma^2}}\right)$$

with $\gamma = n\,\eta/\|\theta\|$ and where $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the probability density and distribution function of the standard normal distribution, respectively.

Proof: Let $W = -\gamma^2 \,+ \gamma\,\sqrt{4 + a\,\gamma^2/n} \cdot N$ with $N \sim N(0,1)$. The expected normalized progress as given in eqn. (3) becomes $\mathsf{E}[\,\max\{W, 0\}\,]$. Since $\max\{W, 0\} = W \cdot 1_{(0,\infty)}(W)$, where $1_A(x)$ is the indicator function of set A, one obtains

$$\mathsf{E}[\,\max\{W, 0\}\,] = \mathsf{E}[\,W \cdot 1_{(0,\infty)}(W)\,] = \int_0^\infty \frac{w}{\gamma\,\sqrt{4 + a\,\gamma^2/n}}\;\varphi\left(\frac{w + \gamma^2}{\gamma\,\sqrt{4 + a\,\gamma^2/n}}\right)\,dw$$

where $\varphi(\cdot)$ is the probability density function of the standard normal distribution. The determination of the integral yields the desired result. □

In principle, the same kind of approximation was presented in [5] for the special case of normally distributed mutations. Additionally, it was argued that the term $a\,\gamma^2/n$ in eqn. (3) becomes small for large n so that this term can be neglected. As a consequence, the random variable W reduces to $\widetilde{W} = -\gamma^2 + 2\,\gamma \cdot N$ and the expected normalized progress becomes

$$\tilde{h}(\gamma) = 2\,\gamma \cdot \varphi(\gamma/2) - \gamma^2 \cdot \Phi(-\gamma/2)$$

attaining its maximum $\tilde{h}(\gamma^*) = 0.404913$ at $\gamma^* = 1.224$ which is exactly the same result established 20 years earlier by Rechenberg [6]. Since all factorizing mutation distributions (with finite absolute moments) in Proposition 4 only distinguish from each other by the constant a, an analogous argumentation for an arbitrary factorizing mutation distribution leads to the result that the normalized improvement is asymptotically equal for all factorizing mutation distributions. Evidently, this kind of approximation is too rough to permit a sound comparison of the progress offered by different factorizing mutation distributions.

distribution	a	γ^*	$h(\gamma^*, a, 100)$
Normal	2	1.24389	0.40801
Logistic	16/5	1.25648	0.40992
Laplace	5	1.27639	0.41289
Student ($d = 5$)	8	1.31273	0.41811

Table 2. Optimal expected normalized progress rates for the (1 + 1)–EA for some factorizing mutation distributions in case of dimension $n = 100$ under the assumption $\mathsf{E}[\max\{n\,(1 - S_n/\|\theta\|^2), 0\}\,] \equiv h(\gamma, a, n)$.

Table 2 summarizes the optimal expected normalized progress rates for some factorizing mutation distributions under the assumption that the approximation of Proposition 4 is exact. The surprising observation which can be made from Table 2 is that the normal distribution is identified as yielding the least progress compared to the other distributions, provided that the assumption $h(\gamma, a, n) \equiv \mathsf{E}[\max\{n\,(1 - S_n/\|\theta\|^2), 0\}\,]$ holds true. The validity of this assumption, however, deserves careful scrutiny since the norming constants $a_n = \mathsf{E}[\,S_n\,]$ and $b_n^2 = \mathsf{V}[\,S_n\,]$ used in the central limit theorem do not necessarily represent the best choice for a rapid approach to the normal distribution. In fact, there may exist constants α_n, β_n obeying $\beta_n \sim b_n$ and $\alpha_n - a_n = o(b_n)$ that lead much faster to the limit [7, p. 262]. As a consequence, it may happen that the ranking of the distributions in Table 2 is reversed after using these (unknown) constants. Thus, unless the error of the approximation of Proposition 4 has been quantified, this kind of approximation is also too rough to permit a sound ranking of the mutation distributions. Nevertheless, the small differences in Table 2 provide evidence that (at least for $n \geq 100$) every factorizing mutation distribution offers a local convergence rate being comparable to that of a normal distribution.

The quality of the approximation in Proposition 4 can be checked in case of normally distributed mutations. As shown in [5], the random variable $V_n = S_n / ||\theta||^2$ follows a noncentral χ^2 distribution with probability density function

$$f_{V_n}(v;\delta) = \frac{\delta^2}{2}\, v^{(n-2)/4} \exp\left(-\frac{\delta^2\,(v+1)}{2}\right)\, I_{n/2-1}(\delta^2\sqrt{v}) \cdot 1_{(0,\infty)}(v)$$

where $I_m(\cdot)$ denotes the mth order modified Bessel function of the first kind and where $\delta = ||\theta||/\eta$ is the noncentrality parameter. Since $V_n > 0$ one obtains $\max\{n\,(1-V_n),0\} = n\,(1-V_n)\cdot 1_{(0,1)}(V_n)$ and hence

$$g(n,\delta) = \mathsf{E}[\,\max\{n\,(1-V_n),0\}\,] = n \int_0^1 (1-v)\, f_{V_n}(v;\delta)\, dv\,. \tag{4}$$

This integral can be evaluated numerically for any given n and δ. Since $\delta = ||\theta||/\eta$ and $\gamma = n\,\eta/||\theta||$ it remains to maximize the function $g(n,\delta) = g(n,n/\gamma)$ with respect to $\gamma > 0$. For example, in case of $n = 100$ a numerical optimization leads to $\gamma^* = 1.224$ with $g(n,n/\gamma^*) = 0.4049$. Figures 3 & 4 show that the optimal variance factor γ^* and the optimal normalized progress $g(n,n/\gamma^*)$ quickly stabilizes for increasing dimension n. In fact, the theoretical limits are almost reached for $n = 30$.

A similar investigation might be made for other mutation vectors $\mathbf{Z}$ with factorizing mutation distributions, if the distribution of $S_n = \sum_{i=1}^n (\theta_i - Z_i)^2$ were to be known. But this does not seem to be the case. For this reason and realizing that the knowledge of the true limits is of no practical importance, it is refrained from taking the burden of determining the density of S_n for other mutation vectors.

Even numerical simulations do not easily lead to a statistically supported ranking: Although the average of the outcomes of random variable

$$Y = \max\{n\,(1 - S_n/||\theta||^2),0\}$$

is an unbiased point estimator of the expectation, there is neither a standard parametric nor standard nonparametric test permitting a statistically supported decision which mean is the largest among the random variables Y generated from different mutation distributions. For example, the parametric t–test presupposes at least approximative normality of Y whereas the nonparametric tests require the continuity of the distribution function of Y. Neither of these requirements is fulfilled, so that it would be necessary to develop a specialized test for this kind of random variables. This is certainly beyond the scope of this paper.

Instead, the attention is devoted to the expected progress rates of the $(1,\lambda)$–EA. Since this EA generates $\lambda \geq 2$ offspring independently with the same distribution and accepts the best among them, the expected progress is simply

$$\mathsf{E}[\max_{i=1,\ldots,\lambda}\{||\theta||^2 - ||\theta + \mathbf{Z}_i||^2\}\,]\,.$$

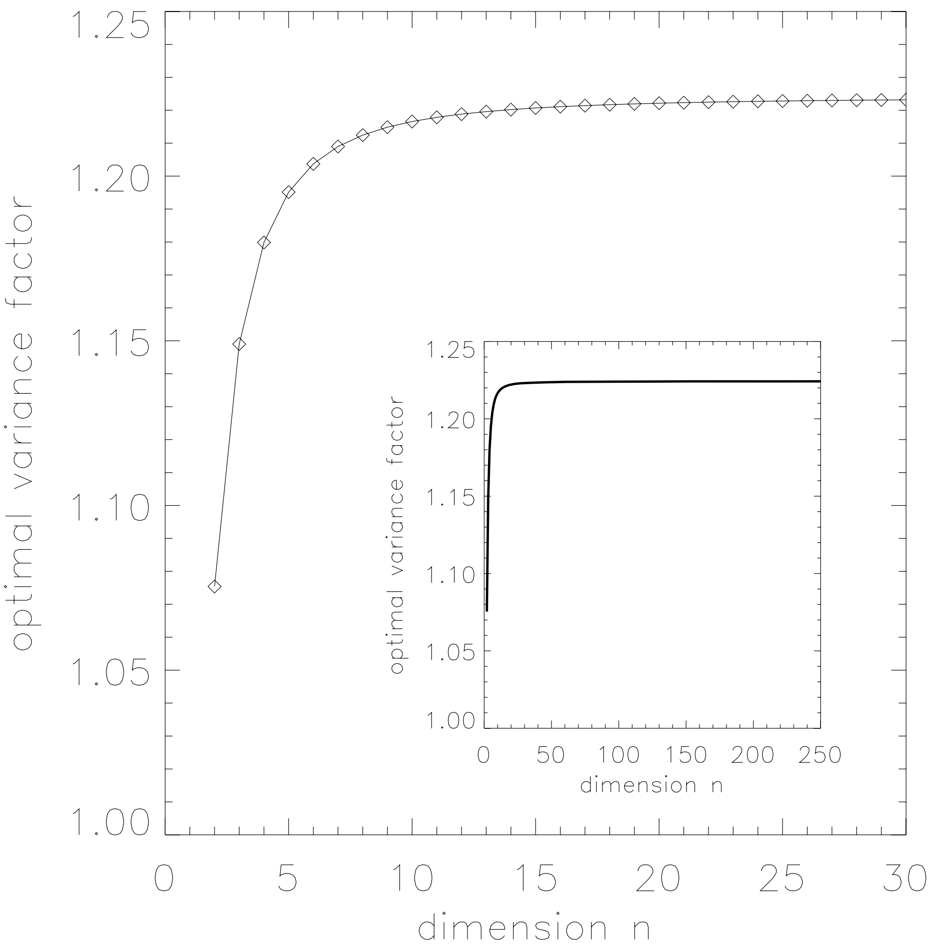

Fig. 3. The optimal variance factor γ^* in case of normal mutation vectors for increasing dimension n.

Following the lines of Proposition 4 and owing to eqn. (2) the normalized expected progress is approximately

$$h(\gamma, a, n) = -\gamma^2 + \gamma\,\sqrt{4 + a\,\gamma^2/n}\cdot \mathsf{E}[\,N_{\lambda:\lambda}\,] \tag{5}$$

where $N_{\lambda:\lambda}$ denotes the maximum of λ independent and identically distributed standard normal random variables. Let $c_\lambda = \mathsf{E}[\,N_{\lambda:\lambda}\,]$. Then the optimal expected normalized progress rate of the $(1,\lambda)$–EA is attained at

$$\gamma^* = \left(\frac{2\,c_\lambda^2}{1 - a\,c_\lambda^2/n + \sqrt{1 - a\,c_\lambda^2/n}}\right)^{1/2}$$

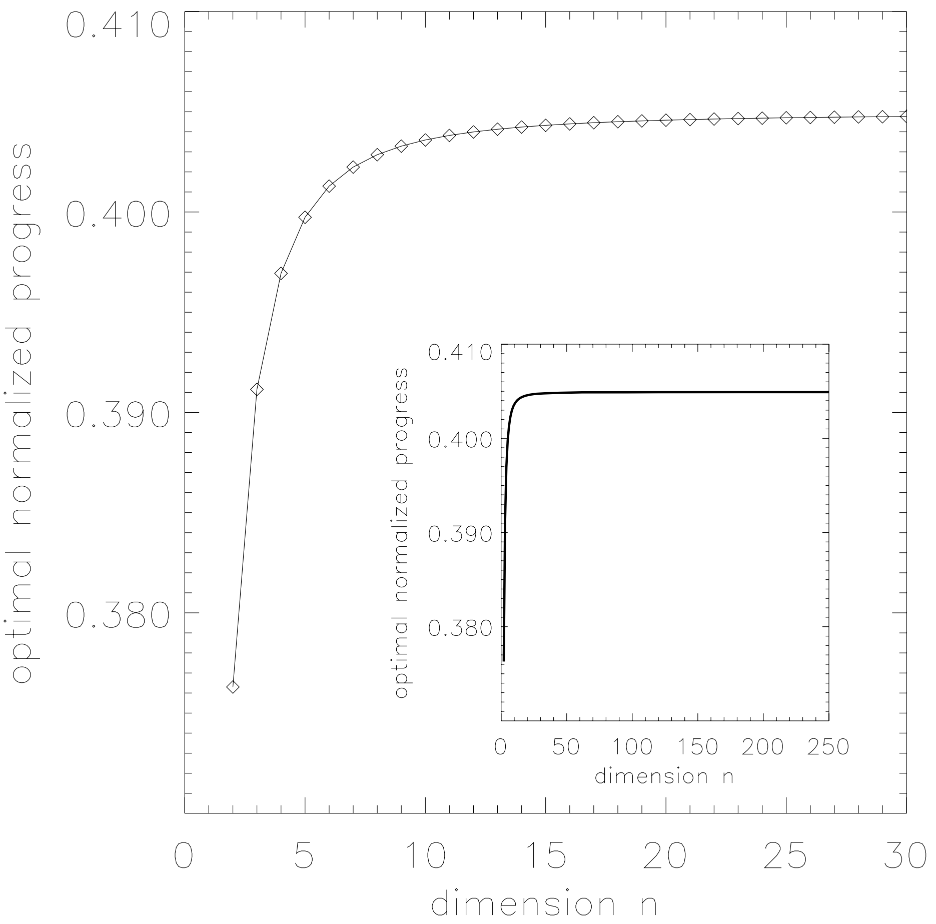

Fig. 4. The optimal normalized progress $g(n, n/\gamma^*)$ in case of normal mutation vectors for increasing dimension n.

which reduces to $\tilde{\gamma}^* = c_\lambda$ as $n \to \infty$. In general, the relation $h(\gamma^*, a, n) > c_\lambda^2$ is valid. Moreover, $h(\gamma, a+\epsilon, n) > h(\gamma, a, n)$ for arbitrary $\gamma > 0$ and $\epsilon > 0$ which follows easily from eqn. (5). Consequently, the expected progress becomes larger for increasing $a > 0$, provided that the approximation given in (2) holds with equality. But it has been seen in case of the $(1+1)$–EA that this approximation does not permit a sound ranking of the distributions. At this point there might arise the question for which purpose the approximations presented in this paper are good for at all. The answer is given in the next section.

3 Conclusions

Under the conditions of the central limit theorem an asymptotical theory of the expected progress rates of simple evolutionary algorithms has been established. If the mutation distributions are factorizing and possess finite absolute moments up to order 4, then each of these distributions offer an almost equally fast approach to the (local) optimum. The optimal variance adjustment w.r.t. fast local convergence is of the type $\eta_k = \gamma \, \|\mathbf{X}_k - \mathbf{x}^*\|/n$ for each of the distributions considered here. This implies that the self–adaptive adjustment of the "step sizes" originally developed for normal distributions needs not be modified in case of other factorizing mutation distributions. In the light of the theory developed in [8] it may be conjectured that these results carry over to population–based EAs *without* crossover or recombination.

Finally, notice that Student's t–distribution with d degrees of freedom converges weakly to the normal distribution as $d \to \infty$ whereas it is called the Cauchy distribution for $d = 1$. All results remain valid for $d \geq 5$. Lower values of d cannot be investigated within the framework presented here, since it was presupposed that the absolute moments of $\mathbf{Z}$ are finite up to order 4. If these moments do not exist the central limit theorem does not hold true. Rather, then there emerges an entire class of limit distributions [9] as already mentioned in [1]. But this case is beyond the scope of this paper and it remains for future research.

Acknowledgment

Besides my thanks to all anonymous reviewers whose suggestions led to several improvements, special thanks must be addressed to one of the reviewers who detected a severe flaw in my argumentation following Proposition 4. Therefore, the part after Proposition 4 was completely revised. As a result, the message of this paper is completely different from the original (wrong) one.

Finally, it should be mentioned that this work is a result of the *Collaborative Research Center "Computational Intelligence" (SFB 531)* supported by the German National Science Foundation (DFG).

References

1. C. Kappler. Are evolutionary algorithms improved by large mutations? In H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel, editors, *Parallel Problem Solving From Nature—PPSN IV*, pages 346–355. Springer, Berlin, 1996.
2. X. Yao and Y. Liu. Fast evolutionary programming. In L. J. Fogel, P. J. Angeline, and T. Bäck, editors, *Proceedings of the Fifth Annual Conference on Evolutionary Programming*, pages 451–460. MIT Press, Cambridge (MA), 1996.
3. X. Yao and Y. Liu. Fast evolution strategies. In P. J. Angeline, R. G. Reynolds, J. R. McDonnell, and R. Eberhart, editors, *Proceedings of the Sixth Annual Conference on Evolutionary Programming*, pages 151–161. Springer, Berlin, 1997.

4. W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. 2.* Wiley, New York, 2nd edition, 1971.
5. T. Bäck, G. Rudolph, and H.-P. Schwefel. Evolutionary programming and evolution strategies: Similarities and differences. In D. B. Fogel and W. Atmar, editors, *Proceedings of the 2nd Annual Conference on Evolutionary Programming*, pages 11–22. Evolutionary Programming Society, La Jolla (CA), 1993.
6. I. Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution.* Frommann–Holzboog Verlag, Stuttgart, 1973.
7. Y. S. Chow and H. Teicher. *Probability Theory.* Springer, New York, 1978.
8. G. Rudolph. *Convergence Properties of Evolutionary Algorithms.* Kovač, Hamburg, 1997.
9. B. V. Gnedenko and A. N. Kolmogorov. *Limit Distributions for Sums of Independent Random Variables.* Addison–Wesley, Reading (MA), revised edition, 1968.

This article was processed using the LaTeX macro package with LLNCS style