Local Convergence Rates of Simple Evolutionary Algorithms with Cauchy Mutations

Günter Rudolph
Universität Dortmund
Fachbereich Informatik, LS XI
D–44221 Dortmund, Germany
Rudolph@LS11.CS.Uni-Dortmund.de

April 27, 1998

Abstract

The standard choice for mutating an individual of an evolutionary algorithm with continuous variables is the normal distribution; however other distributions, especially some versions of the multivariate Cauchy distribution, have recently gained increased popularity in practical applications. Here the extent to which Cauchy mutation distributions may affect the local convergence behavior of evolutionary algorithms is analyzed. The results show that the order of local convergence is identical for Gaussian and spherical Cauchy distributions, whereas nonspherical Cauchy mutations lead to slower local convergence. As a by–product of the analysis some recommendations for the parametrization of the self–adaptive step size control mechanism can be derived.

1 Introduction

The Gaussian distribution is the predominant choice for a mutation distribution in evolutionary algorithms (EAs) with search space $\mathbb{R}^\ell$ [1, 2, 3, 4, 5]. 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 would permit a deterministic model, these influences are considered as independent random perturbations whose normed sum approaches a Gaussian 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 (see [6], p. 291). But the Gaussian distribution is not the only limit distribution for normed sums of random variables. If the underlying random variables are independent and identically distributed (i.i.d.) and have finite absolute moments at most of order k, then for $k \geq 2$ only the Gaussian distribution can arise as a limit, whereas if $0 < k \leq 2$, the limit laws are instances of a class called stable distributions (see [6], p. 436). A stable distribution F is characterized by the property that the distribution of the sum of two independent random variables with distributions of type F is also of type F. The only stable distributions besides the Gaussian distribution possessing finite variances are degenerate. Although each stable distribution (except the degenerate ones) has a unimodal and infinitely often differentiable probability density function (p.d.f.) these p.d.f.s can be given in explicit form only in exceptional cases (see [7], p. 366). Such an example is the p.d.f. of the Cauchy distribution

$$f_C(x) = \frac{1}{\pi} \frac{d}{d^2 + x^2} \tag{1}$$

whose absolute moments exist only for $0 < k < 1$. Probability distributions with infinite absolute moments appear in physics in various settings [8, 9]. Thus the Cauchy distribution may arise as a limit law describing the cumulative effect of independent random perturbations and therefore there is no reason to preclude this distribution from the set of candidate distributions playing the role of mutation distributions in evolutionary algorithms as models of natural systems.

But some care is necessary when comparing the performance of optimization algorithms with Cauchy and Gaussian mutations. Whereas the univariate Cauchy distribution has a unique definition, there exist at least two multivariate versions of the Cauchy distribution: the spherically symmetric Cauchy distribution and the Cauchy distribution with independent univariate Cauchy random variables in each dimension. The first version was employed as a search distribution in "simulated annealing" (SA) algorithms [10, 11]. Ingber [12] also considered the second version for SA but abandoned the idea for theoretical reasons. Recently, apparently inspired by these publications, some experimental results [13, 14, 15] concerning Cauchy-type mutations in evolutionary algorithms became available. These experiments employed the second version of the multivariate Cauchy distribution whereas the theoretical analysis presented by Kappler [16] rests on the first version.

The work in hand may be seen as a continuation of Kappler's effort. She calculated the expected convergence rate of a $(1+\lambda)$–EA for a two–dimensional problem in the case of Gaussian and spherical Cauchy mutations. The task to solve this optimization problem, called the "bounded inclined corridor problem," resembles the situation of finding the entrance to a small corridor in the search space leading to better solutions.

Here, we investigate the ability of simple EAs to locate a local minimum under the assumption that the EA has already entered the local optimum's basin of attraction. This situation may be studied by the problem to minimize the objective function $f(x) = x'x$ with $x \in \mathbb{R}^\ell$. In Section 2 the mutation distributions under consideration are spherically symmetric. This includes the Gaussian, spherical Student, and Cauchy distribution. Cauchy mutations with independent components are analyzed in Section 3. These results lead to implications for the self-adaptive mutation mechanism that is discussed in Section 4. Finally, the conclusions are drawn in Section 5.

2 Convergence Rates Under Spherically Symmetric Mutations

The convergence rates of simple evolutionary algorithms with different mutation distributions may be compared for the following problem: minimize $f(x) = x'x$ with $x \in \mathbb{R}^\ell$ and $\ell \geq 2$. The objective function $f : \mathbb{R}^\ell \to \mathbb{R}$ is a special instance from the class of quadratic functions with positive definite Hessian matrix.

The evolutionary algorithm under consideration is the $(1+1)$–EA. An individual $\theta \in \mathbb{R}^\ell$ is mutated by adding a random vector ηZ where parameter $\eta > 0$ controls the scale of the distribution. If the offspring $\theta + \eta Z$ has a better objective function value than its parent θ, i.e., if $f(\theta + \eta Z) < f(\theta)$, then the mutation is accepted and the offspring will serve as new parent in the next iteration. Otherwise, the mutation is rejected and the old parent will pass into the next iteration.

Usually, the random vector Z must fulfill some basic requirements. It is reasonable to postulate that—at least initially—no preference of a certain direction should be given. This request leads to the property that the random vector Z is spherically symmetric with respect to the origin $0 \in \mathbb{R}^\ell$.

2.1 Spherically Symmetric Distributions

There are several avenues to generalize a symmetrical univariate distribution to a multivariate version [17]. Here, the definition below will be used.

Definition 1

A random vector X of dimension $\ell \geq 2$ with location parameter $\theta \in \mathbb{R}^\ell$ is said to possess a *spherically symmetric distribution* if $(X - \theta) \stackrel{d}{=} T'(X - \theta)$ for every orthogonal matrix T with $T'T = I$, where I denotes the unit matrix. The operator $\stackrel{d}{=}$ indicates that the distributions of the random elements to its left- and right-hand side are identical. □

Spherically symmetric distributions possess many nice properties but only a few will be exploited here. The results summarized below are extracted from [17, Sect. 2.1].

Theorem 1

A random vector X of dimension $\ell \geq 2$ with location parameter $\theta \in \mathbb{R}^\ell$ is spherically symmetric if and only if it has the stochastic representation

$$X \stackrel{d}{=} \theta + r\,U$$

where r is a nonnegative random variable and U is a random vector uniformly distributed on the surface of a unit hyperball of dimension ℓ. Moreover, r and U are independent. If additionally $\mathsf{P}\{\, X = \theta \,\} = 0$ then $r \stackrel{d}{=} \|X - \theta\|$ and $U \stackrel{d}{=} (X - \theta)/\|X - \theta\|$ where $\|\cdot\|$ denotes the Euclidean norm.□

Let $Z \stackrel{d}{=} r\,U$ be a spherically symmetric random vector of dimension $\ell \geq 2$ with location parameter $\theta = 0$. If the random variable r is χ_ℓ distributed with ℓ degrees of freedom then random vector Z is normally distributed with zero mean and covariance matrix described by the unit matrix. If r/ℓ has F–distribution with ℓ and $s \in \mathbb{N}$ degrees of freedom, then Z has a multivariate spherical t_s–distribution with s degrees of freedom. In case of $s = 1$ the multivariate t_s–distribution is called the multivariate spherical Cauchy distribution. The next results are adapted from [18].

Theorem 2

A spherically symmetric random vector X with location parameter $\theta \in \mathbb{R}^\ell$ and scale parameter $\eta > 0$ has a p.d.f. $f_X(\cdot)$ if and only if there exists a nonnegative scalar function $g(\cdot)$ with

$$c = \int_0^\infty y^{\ell-1}\, g(y^2)\, dy < \infty$$

such that

$$f_X(x) = \frac{\Gamma(\ell/2)}{2\,\pi^{\ell/2}\,c}\,\eta^{-\ell/2}\, g\left(\frac{\|x - \theta\|^2}{\eta^2}\right) \tag{2}$$

where $\Gamma(\cdot)$ denotes the complete Gamma function. □

The function $g(\cdot)$ is termed the *density generator*. Its structural form determines to which class the distribution of X belongs. Two classes are presented below.

- Multivariate Kotz-type distributions:
  The density generator is $g(t) = t^{n-1}\exp(-r\,t^s)$ with $r, s > 0$, $2\,n + \ell > 2$, and constant $c = \Gamma(q)/(2\,s\,r^q)$ where $q = (2\,n + \ell - 2)/(2s)$. This class includes the multinormal distribution with $n = s = 2\,r = 1$.

- Multivariate Pearson-type VII distributions:
  The density generator is $g(t) = (1 + t/s)^{-n}$ with $n > \ell/2$, $s > 0$, and constant $c = B(\ell/2, n - \ell/2) \cdot s^{\ell/2}/2$ where $B(\cdot,\cdot)$ denotes the complete Beta function. This class includes the multivariate spherical versions of Student's t–distribution with s degrees of freedom with $n = (\ell + s)/2$ as well as the multivariate spherical Cauchy distribution with $n = (\ell + 1)/2$ and $s = 1$.

In the next subsection the distribution of the random scalar product $X'X$ needs to be known: Notice that the distribution of $X'X$ with $X = \theta + \eta\, Z$ will represent the distribution of the random offspring's objective function value. If X is normally distributed with location parameter $\theta \neq 0 \in \mathbb{R}^\ell$ and scale parameter $\eta = 1$, then it is known (see e.g. [19, p. 130]) that $X'X$ is noncentrally χ^2_ℓ distributed with ℓ degrees of freedom and noncentrality parameter $\delta = \|\theta\|$. The distributions resulting from the scalar product of other spherically symmetric random vectors do not seem to bear their own names. Nevertheless, their p.d.f.s are not difficult to obtain. Here, the result given in [18] is presented in slightly modified form.

THEOREM 3
Let X be a spherically symmetric random vector of dimension $\ell \geq 2$ with some p.d.f. as given in (2). If $\theta \neq 0 \in \mathbb{R}^\ell$ then the p.d.f. of $V = X'X/\theta'\theta$ is

$$f_V(v) = \frac{\delta^\ell\, v^{\ell/2-1}}{2\, c\, B((\ell-1)/2, 1/2)} \int_{-1}^{1} g(\delta^2 [\, v - 2\, t\, \sqrt{v} + 1\,])\, (1-t^2)^{(\ell-3)/2}\, dt \tag{3}$$

with noncentrality parameter $\delta = \|\theta\|/\eta$ and where $B(\cdot,\cdot)$ denotes the complete Beta function. $\square$

The random variable $V = X'X/\theta'\theta$ represents the *relative variation* of the offspring's objective function value. If $0 < V < 1$ then the offspring is better than its parent whereas it is worse than its parent if $V \geq 1$. An important property of the relative variation's distribution is revealed by (3). If the scale parameter η is proportional to $\|\theta\|$ then the distribution of V is only parametrized by the noncentrality parameter δ and the dimension ℓ. Thus, whatever the actual location θ, the relative variation's distribution is always the same. Notice that this is true for every spherically symmetric mutation distribution.
In the remainder, however, the investigation will be restricted to the multivariate spherical normal and Student's t_s distributions, including the Cauchy distribution. The t_s distribution is of particular interest because it may be seen as an intermediate form between the Gaussian and Cauchy distribution: If $s \to \infty$ then the t_s distribution converges weakly to the Gaussian distribution whereas it becomes the Cauchy distribution at the other extreme with $s = 1$.
The probability density functions of the associated relative variation can be derived via Theorem 3. Let $X = \theta + \eta\, Z$ be normally distributed. Then the p.d.f. of V is

$$f_V(v) = \frac{\delta^2}{2}\, v^{(\ell-2)/4} \exp\left(-\frac{\delta^2\,(v+1)}{2}\right) I_{\ell/2-1}(\delta^2\sqrt{v}) \cdot 1_{(0,\infty)}(v) \tag{4}$$

where $I_m(\cdot)$ denotes the modified Bessel function of the first kind and order m. If $X = \theta + \eta\, Z$ is t_s distributed one obtains

$$f_V(v) = \frac{s^{s/2}}{B(\ell/2, s/2)} \frac{\delta^\ell\, v^{\ell/2-1}}{[\, s + \delta^2\,(v+1)\,]^{(\ell+s)/2}}\ {}_2F_1\left(\frac{\ell+s}{4}, \frac{\ell+s}{4} + \frac{1}{2}; \frac{\ell}{2}, z^2\right) \cdot 1_{(0,\infty)}(v) \tag{5}$$

with $z = (2\,\delta^2\sqrt{v})/(s + \delta^2\,(v+1))$ and where ${}_2F_1(\cdot)$ denotes the Gauss hypergeometric series. These p.d.f.s are expressible by elementary functions if the dimension ℓ is odd with $\ell \geq 3$. The Bessel function may be reformulated by using entry 10.2.24 in [20], whereas the Gauss hypergeometric series can be brought down to a finite sum of rational polynomials and their logarithms via entries 15.1.4, 8–10 and repeated application of entries 15.2.12 and 15.3.3 [20]. For example, if $\ell = 3$ then (4) becomes

$$f_V(v) = \delta\, \exp(-\delta^2\,(v+1)/2)\, \sinh(\delta^2\sqrt{v})/\sqrt{2\,\pi} \cdot 1_{(0,\infty)}(v) \tag{6}$$

and (5) reduces to

$$f_V(v) = \frac{1}{\pi} \, \frac{2\,\delta^3\,\sqrt{v}}{1+\delta^4\,(v-1)^2 + 2\,\delta^2\,(v+1)} \cdot 1_{(0,\infty)}(v) \tag{7}$$

in case of $s = 1$.

2.2 Exact Convergence Rates of the (1+1)–EA in Dimension 3

Since the $(1+1)$–EA only accepts improvements the new objective function value is given by the random variable $\min\{\|\theta + \eta\,Z\|^2, \|\theta\|^2\}$. Therefore the expected convergence rate $c \in (0,1)$ is determined by the relation $\mathsf{E}[\,\min\{\|\theta+\eta\,Z\|^2, \|\theta\|^2\}\,|\,\theta\,] = c \cdot \|\theta\|^2$ which may be equivalently expressed as

$$\mathsf{E}\left[\,\min\left\{\frac{\|\theta+\eta\,Z\|^2}{\|\theta\|^2}, 1\right\}\,\middle|\,\theta\,\right] = \mathsf{E}[\,\min\{V,1\}\,|\,\theta\,] = c\,. \tag{8}$$

Notice that the convergence velocity increases with smaller $c \in (0,1)$. To see this let $\varepsilon_t = \mathsf{E}[\,\|\theta_t - \theta^*\|^2\,]$ be the expected error at iteration $t \ge 0$ (here, $\theta^* = 0 \in \mathbb{R}^\ell$). If there exists a constant $c \in (0,1)$ then $\varepsilon_{t+1} = c\,\varepsilon_t$ or $\varepsilon_t = \varepsilon_0\,c^t$ for $t \ge 0$. Elementary transformations of the latter equation leads to

$$t = \frac{\log_{10}(\varepsilon_t/\varepsilon_0)}{\log_{10}(c)} = -\frac{\Delta}{\log_{10}(c)} \tag{9}$$

where $\Delta > 0$ denotes the orders of magnitude the error is to be decreased. If Δ is fixed then the time t that is required to decrease the error by Δ orders of magnitude decreases as c decreases towards zero. To determine the constant c for the $(1+1)$-EA one must evaluate (8). Since V is nonnegative the relation $\min\{V,1\} = V \cdot 1_{(0,1)}(V) + 1_{[1,\infty)}(V)$ is valid. Thus,

$$\mathsf{E}[\,\min\{V,1\}\,] = \int_0^1 v\,f_V(v)\,dv + \int_1^\infty f_V(v)\,dv = 1 - \int_0^1 (1-v)\,f_V(v)\,dv\,. \tag{10}$$

At first, let $\ell = 3$. Insertion of (6) into (10) yields the convergence rate

$$c(\delta) = 1 - \sqrt{\frac{2}{\pi}}\,\frac{\delta^2 + 1 - \exp(-2\,\delta^2)}{\delta^3} + \frac{6\,\Phi(2\,\delta) - 3}{2\,\delta^2}$$

in case of Gaussian mutations. Similarly, insertion of (7) into (10) leads to

$$c(\delta) = 1 - \frac{1}{\delta\,\pi}\left[\frac{3}{\delta}\,\arctan(2\,\delta) + \frac{2\,\delta^2 - 1}{2\,\delta^2}\,\log(4\,\delta^2 + 1) - 4\right]$$

in case of Cauchy mutations. Figure 1 shows the convergence rates as a function of δ while Table 1 summarizes the optimal convergence rate $c^* = c(\delta^*)$ with optimal δ^* for several mutation distributions. Since the constant c^* for the Gaussian mutations is smaller than that for Cauchy mutations, it has been shown that Gaussian mutations lead to faster convergence than Cauchy mutations. But notice that the order of convergence is the same for both distributions.

After the analysis of the low-dimensional case one may inquire in the scaling behavior of the convergence rates if the dimension ℓ becomes large. Since solving the integral in (10) seems intractable for arbitrary ℓ, the subsequent analysis will be confined to asymptotic convergence rates ($\ell \gg 1$).

Distribution	s	c^*	δ^*
Cauchy	1	0.910032	4.31981
Student	2	0.892636	3.59588
Student	3	0.885364	3.30606
Student	4	0.881457	3.14775
Student	5	0.879050	3.04765
Student	10	0.874201	2.83535
Student	20	0.871844	2.72369
Student	30	0.871084	2.68600
Gauss	∞	0.869617	2.61093

Table 1: Optimal convergence rates $c^* \in (0,1)$ and noncentrality parameters $\delta^* > 0$ of the $(1+1)$-EA for Cauchy, Student, and Gaussian distribution for $\ell = 3$. The maximum convergence velocity increases (since c^* decreases) from Cauchy via Student to Gaussian mutations.

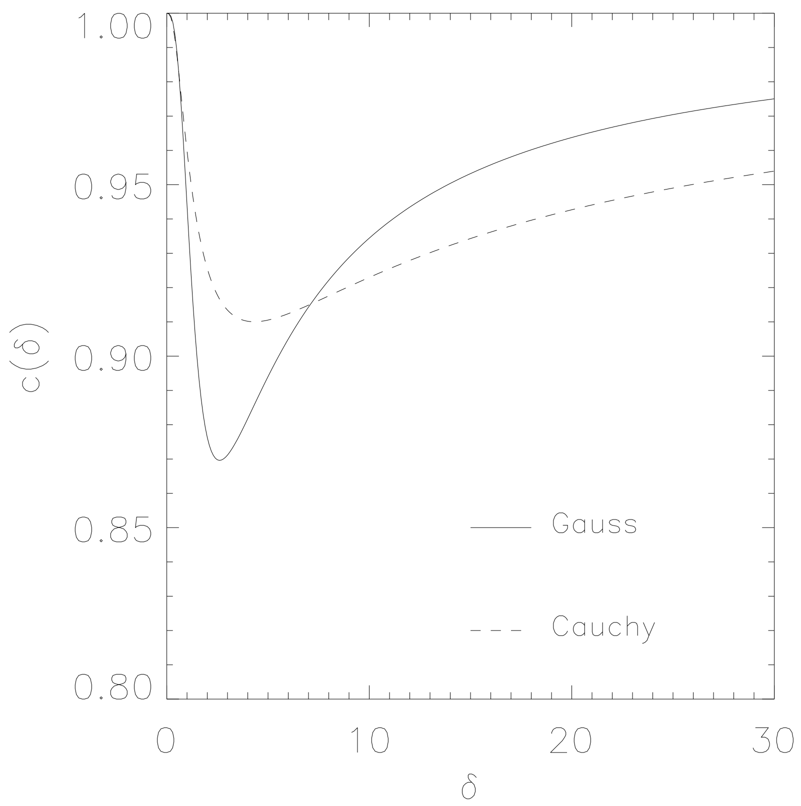

Figure 1: Convergence rate $c(\delta)$ of the $(1+1)$-EA as a function of noncentrality parameter δ in the case of Gaussian and Cauchy mutations with scale parameter $\eta = ||\theta||/\delta$ in dimension $\ell = 3$. The optima δ^* are less sensitive to a shift to the right than to the left.

2.3 Asymptotic Convergence Rates of the (1+1)–EA

The basic idea of the approach presented here is as follows: Since the relative variation V depends on ℓ, which is hereinafter emphasized by writing V_ℓ, it is necessary to determine the constants $a_\ell > 0$ and $b_\ell \in \mathbb{R}$ under which the normalized random variable $(V_\ell - b_\ell)/a_\ell$ converges in distribution to a nondegenerate limit random variable L whose distribution is independent from the dimension ℓ. If such a limit L exists then V_ℓ may be approximated by $V_\ell \approx a_\ell\, L + b_\ell$ provided that ℓ is sufficiently large. Suppose that $b_\ell \equiv 1$. In this case one obtains the approximation

$$\min\{V_\ell, 1\} = 1 - \max\{1 - V_\ell, 0\} \approx 1 - \max\{-a_\ell\, L, 0\} = 1 - a_\ell\, \max\{-L, 0\}\,.$$

Consequently, the convergence rate is given by

$$c = \mathsf{E}[\,\min\{V_\ell, 1\}\,] = 1 - a_\ell\, \mathsf{E}[\,\max\{-L, 0\}\,]\,. \tag{11}$$

In the following it is shown that this plan can be realized. At first, observe that the random objective function value $\|\theta + \eta\, Z\|^2$ has the stochastic decomposition

$$\|\theta + \eta\, Z\|^2 = (\theta + \eta\, Z)'(\theta + \eta\, Z) = \theta'\theta + 2\,\eta\,\theta' Z + \eta^2\, Z' Z = \|\theta\|^2 + 2\,\eta\,\theta' Z + \eta^2\,\|Z\|^2\,.$$

To proceed one needs the distribution of the stochastic scalar product $\theta' Z$. If Z is a standard Gaussian vector then the distribution is easily obtained. Since the sum of ℓ independent Gaussian random variables with zero mean and variances θ_i^2 is a Gaussian random variable with zero mean and variance $\sum_{i=1}^{\ell} \theta_i^2 = \|\theta\|^2$ it follows that

$$\theta' Z = \sum_{i=1}^{\ell} \theta_i\, Z_i \stackrel{d}{=} \|\theta\|\, Z_1\,,$$

that is, the random scalar product $\theta' Z$ has the same distribution as its marginal Z_1 multiplied by $\|\theta\|$. This remarkable property is characteristic not only for Gaussian random vectors but for all spherically symmetric random vectors.

Lemma 1 ([17, p. 31])
The random vector $Z = (Z_1, \ldots, Z_\ell)'$ with location parameter $0 \in \mathbb{R}^\ell$ is spherically symmetric if and only if $\theta' Z \stackrel{d}{=} \|\theta\|\, Z_1$ for every $\theta \in \mathbb{R}^\ell$. □

Recall from Theorem 3 that the scaling parameter of the mutation vector was set to $\eta = \|\theta\|/\delta$. Owing to Lemma 1 one may write

$$\|\theta + \eta\, Z\|^2 - \|\theta\|^2 \stackrel{d}{=} 2\,\eta\,\|\theta\|\, Z_1 + \eta^2\,\|Z\|^2 = 2\,\delta^{-1}\,\|\theta\|^2\, Z_1 + \delta^{-2}\,\|\theta\|^2\,\|Z\|^2\,.$$

Division by $\|\theta\|^2 \neq 0$ leads to

$$\frac{\|\theta + \eta\, Z\|^2}{\|\theta\|^2} - 1 \stackrel{d}{=} 2\,\delta^{-1}\, Z_1 + \delta^{-2}\,\|Z\|^2 = \frac{2\,\gamma}{\ell}\, Z_1 + \frac{\gamma^2}{\ell^2}\,\|Z\|^2$$

with $\delta = \ell/\gamma$, $\gamma > 0$. Multiplication by ℓ yields

$$\ell\,(V_\ell - 1) \stackrel{d}{=} 2\,\gamma\, Z_1 + \gamma^2\,\frac{1}{\ell}\sum_{i=1}^{\ell} Z_i^2\,. \tag{12}$$

This equation is valid for every spherically symmetric random vector Z. Now suppose that Z is standard multinormally distributed. With the result below, the limit of the random variable in (12) is easy to identify.

Lemma 2
Let $Z_1, Z_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) standard normal random variables. As $\ell \to \infty$ then

$$\frac{1}{\ell} \sum_{i=1}^{\ell} Z_i^2 \longrightarrow 1 \quad \text{with probability 1.}$$

Proof: Since $Z_1, Z_2, \ldots$ are i.i.d. random variables so are $Z_1^2, Z_2^2, \ldots$ with $\mathsf{E}[\, Z_1^2 \,] = 1$. Now the desired result follows immediately from the strong law of large numbers. Also see exercise 5.2.5 in [6, p. 131]. □

According to Lemma 2 the random variables in (12) converge in distribution to the limit $L = 2\,\gamma\, Z_1 + \gamma^2$. Thus, the limit L is normally distributed with mean γ^2 and variance $4\,\gamma^2$. The normalizing constants are $a_\ell = 1/\ell$ and, as required, $b_\ell \equiv 1$. It remains to calculate $\mathsf{E}[\max\{-L, 0\}\,]$. Since $\max\{x, 0\} = x \cdot 1_{(0,\infty)}(x)$ one obtains

$$\mathsf{E}[\max\{-L, 0\}\,] = \int_0^\infty x \, \frac{1}{2\,\gamma}\, \varphi\left(\frac{x+\gamma}{-2\,\gamma}\right) \, dx = 2\,\gamma \cdot \varphi(\gamma/2) - \gamma^2 \cdot \Phi(-\gamma/2) = g(\gamma)$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the probability density and cumulative distribution function of the standard normal distribution, respectively. Owing to (11) the convergence rate is approximately $c(\gamma) \approx 1 - g(\gamma)/\ell$. The optimal convergence rate can be obtained by maximizing $g(\gamma)$. Numerical optimization yields $\gamma^* = 1.224$ with $g(\gamma^*) = 0.404913$ which is the same result established more than 20 years earlier by Rechenberg [1] but with much more effort.
Now insist that Z has a multivariate spherical Cauchy distribution. A random vector with this distribution may be generated as follows: Let N be a standard normal vector and S_1 be a χ_1 distributed random variable with one degree of freedom, where N and S_1 are independent. Then $Z = N/S_1$ is multivariate spherically Cauchy distributed [17, p. 85]. Owing to Lemma 1 one obtains

$$\theta' Z \stackrel{d}{=} \theta' N / S_1 \stackrel{d}{=} \|\theta\| \, N_1 / S_1 \stackrel{d}{=} \|\theta\| \, C \tag{13}$$

where C is a standard Cauchy random variable with p.d.f. as given in (1) with $d = 1$. Notice that the distribution is independent from the dimension ℓ. Therefore it is sufficient to enter the previous analysis at (12) yielding

$$\ell\,(V_\ell - 1) \stackrel{d}{=} 2\,\gamma\, C + \gamma^2 \, \frac{1}{\ell} \sum_{i=1}^{\ell} Z_i^2 \tag{14}$$

under usage of (13). To proceed, one needs a result that parallels Lemma 2.

Lemma 3
Let $Z = (Z_1, \ldots, Z_\ell)'$ be a standard multivariate spherical Cauchy vector. As $\ell \to \infty$ then

$$\frac{1}{\ell} \sum_{i=1}^{\ell} Z_i^2 \longrightarrow G \quad \text{in distribution}$$

where G has p.d.f. $f_G(x) = x^{-3/2} \exp(-1/(2\,x))/\sqrt{2\,\pi} \cdot 1_{(0,\infty)}(x)$.
Proof: Since $Z \stackrel{d}{=} N/S_1$ it follows that

$$\frac{1}{\ell} \sum_{i=1}^{\ell} Z_i^2 \stackrel{d}{=} \frac{\ell^{-1} \sum_{i=1}^{\ell} N_i^2}{S_1^2}\,. \tag{15}$$

Notice that the random variables $N_1^2, N_2^2, \ldots$ are independent squared standard normally distributed random variables. Therefore Lemma 2 ensures that the numerator of the r.h.s. in (15) converges almost surely to unity. The distribution of S_1^2 is not affected by parameter ℓ. Owing to Slutsky's Theorem (see e.g. [21, p. 180]) one may conclude that the normalized sum on the l.h.s. of (15) converges in distribution to the random variable $G = 1/S_1^2$. Since S_1^2 possesses χ_1^2 distribution with one degree of freedom and probability density function $f_{S_1^2}(x) = (2\,\pi\,x)^{-1/2}\exp(-x/2)\cdot 1_{(0,\infty)}(x)$ the density transformation $f_G(x) = x^{-2}\,f_{S_1^2}(1/x)$ leads to the distribution of the limit G. □

Now let $\ell \to \infty$ in (14). Thanks to Lemma 3 one may conclude that $\ell\,(V_\ell - 1)$ converges in distribution to the limit random variable $L = 2\,\gamma\,C + \gamma^2\,G$ whose distribution only depends on γ. Again, the normalizing constants are $a_\ell = 1/\ell$ and, as required, $b_\ell \equiv 1$. It remains to determine $g(\gamma) = \mathsf{E}[\,\max\{-L, 0\}\,]$. The explicit distribution of the limit L is unknown yet—only its existence has been shown. A not necessarily successful route to obtain the limit distribution is as follows: Consider the random variable $W_\ell = \ell\,(V_\ell - 1)$. Its density is easily obtained via the transformation $f_{W_\ell}(x) = f_{V_\ell}(1 + x/\ell)/\ell$. If $f_{W_\ell}(x)$ converges to $f_{W_\infty}(x)$ for every continuity point as $\ell \to \infty$, then $f_{W_\infty}(\cdot)$ is the density of the limit L. This would follow from Scheffé's "useful convergence theorem" [22]. But notice that in general the densities need not converge even though the distribution functions converge weakly to a limit distribution function possessing a continuous density (see the instructive example in [23], p. 252). To see whether or not such a scenario is appropriate here set $s = 1$ and $\delta = \ell/\gamma$ with $\gamma > 0$ in (5) before applying the density transformation $f_{W_\ell}(x) = f_{V_\ell}(1 + x/\ell)/\ell$ for $x \in (-\ell, \infty)$. As can be seen from Figure 2 the density of W_ℓ quickly stabilizes for increasing ℓ. Thus, there is some evidence that the densities of W_ℓ will converge to the density of the limit L.
The limit operation on these densities, however, is difficult. This is primarily caused by the complicated limit behavior of the Gauss hypergeometric series when the first three parameters tend to infinity. As a consequence, the density of the limit L has not been found yet so that another method is required to derive the optimal convergence rate. A remedy to obtain the optimal values γ^* and $g(\gamma^*)$ might be as follows: Since it is known empirically from Figure 2 that the density of W_ℓ quickly stabilizes for increasing ℓ simply choose a large value for ℓ, set $s = 1$, and use $\delta = \ell/\gamma$ in the p.d.f. of V_ℓ given in (5). Insert this density into (10). Since the limits in the resulting integral are 0 and 1 there is no problem in evaluating the integral numerically for given γ and ℓ. Let $c(\gamma, \ell)$ be the result of the numerical integration. According to (11) and taking into account that $a_\ell = 1/\ell$ one finds that $g(\gamma) \approx \ell\,(1 - c(\gamma, \ell))$ for sufficiently large ℓ. Thus, the value of $g(\gamma)$ should become stable for increasing ℓ. Then the optimal value of γ can be approximated via univariate numerical optimization over γ with large fixed ℓ. Table 2 summarizes the results of this approach for the spherical Cauchy as well as the Gaussian mutation distribution. Evidently, the value of $\ell\,(1 - c(\gamma^*, \ell))$ already stabilizes for both distributions when $\ell \approx 100$. But even $\ell = 30$ yields reasonable results. In case of the Gaussian distribution there is a tiny discrepancy between the approximation in Table 2 and the theoretical values obtained previously.

2.4 Convergence Rates for the $(1,\lambda)$–EA

For the determination of the convergence rates of the $(1,\lambda)$–EA the theory of order statistics [24] has been proven successful. Let $Y_1, Y_2, \ldots, Y_\lambda$ be random variables. If they are rearranged in ascending order of magnitude, written as

$$Y_{1:\lambda} \le Y_{2:\lambda} \le \ldots Y_{\lambda:\lambda}\,,$$

then $Y_{i:\lambda}$ is called the ith order statistic ($i = 1, \ldots, \lambda$). In this terminology the $(1,\lambda)$–EA accepts that offspring having objective function value $Y_{1:\lambda}$, where $Y_1, Y_2, \ldots, Y_\lambda$ denote the unordered random

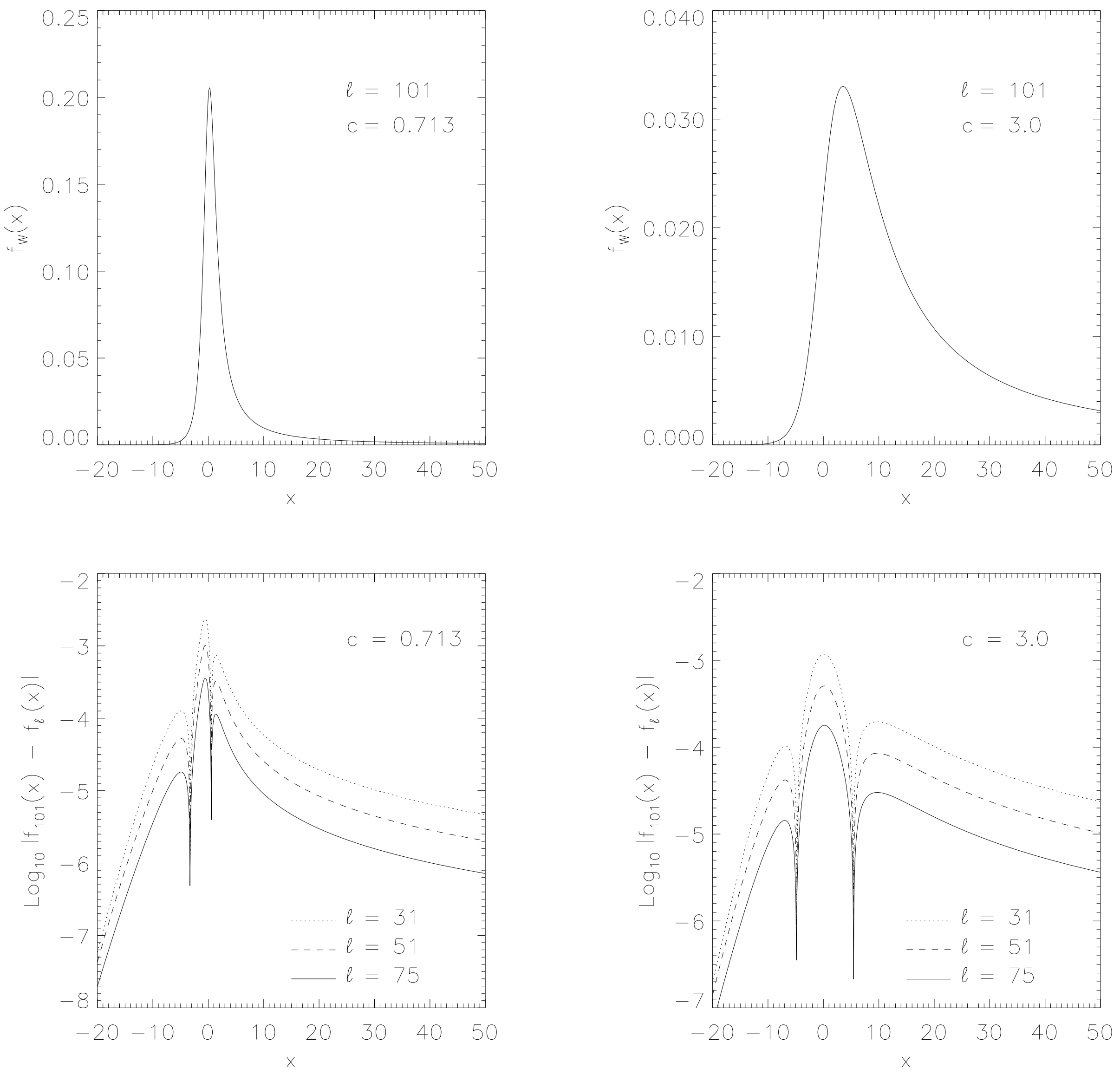

Figure 2: Top: The p.d.f. of the normalized relative variation W_ℓ for dimension $\ell = 101$ and scale parameters $c = 0.713$ (left) and $c = 3$ (right). Bottom: The absolute differences between the p.d.f.s of W_{101} and W_ℓ for $\ell = 31, 51, 75$. Evidently, the p.d.f. of W_ℓ converges weakly to a limit p.d.f. as $\ell \to \infty$.

objective function values of the λ offspring. Notice that the offspring are generated independently and with the same distribution. Therefore the probability density functions of the ith order statistic has a simple form.

Lemma 4 ([24, p. 8])
Suppose that $Y_1, Y_2, \ldots, Y_\lambda$ are i.i.d. random variables with p.d.f. $f_Y(\cdot)$ and distribution function $F_Y(\cdot)$.

	Cauchy mutations		Gaussian mutations	
ℓ	$\ell\,(1-c(\gamma^*,\ell))$	γ^*	$\ell\,(1-c(\gamma^*,\ell))$	γ^*
2	0.2639	0.680	0.3763	1.075
3	0.2699	0.695	0.3912	1.149
4	0.2723	0.700	0.3969	1.180
5	0.2734	0.703	0.3997	1.195
10	0.2750	0.707	0.4036	1.217
20	0.2754	0.708	0.4046	1.222
30	0.2755	0.713	0.4048	1.223
40	0.2755	0.713	0.4048	1.224
50	0.2755	0.713	0.4049	1.224
100	0.2756	0.713	0.4049	1.225
150	0.2756	0.713	0.4049	1.225
200	0.2756	0.713	0.4049	1.225

Table 2: Approximated optimal values for γ^* and $g(\gamma^*) \approx \ell\,(1-c(\gamma^*,\ell))$ in case of spherical Cauchy and Gaussian mutation vectors for increasing dimension ℓ. As conjectured, the optimal step size scaling parameters γ^* quickly stabilize as the dimension gets large.

Then the p.d.f. of the ith order statistic is

$$f_{Y_{i:\lambda}}(x) = \frac{1}{B(i, \lambda-i+1)}\, f_Y(x)\, F_Y^{i-1}\,[\,1-F_Y(x)\,]^{\lambda-i} \tag{16}$$

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

Recall from Theorem 3 that the random variable $Y = \|\theta + \eta\, Z\|^2$ with $\eta > 0$ and $\|\theta\| \neq 0$ can be represented by $Y = \|\theta\|^2\, V$, where random variable V only depends on the dimension ℓ and noncentrality parameter $\delta = \|\theta\|/\eta$. Suppose that $\mathsf{E}[\,V_{1:\lambda}\,]$ exists. Since

$$\mathsf{E}[\,Y_{1:\lambda}\,] = \mathsf{E}[\,\|\theta\|^2\, V_{1:\lambda}\,] = \|\theta\|^2\, \mathsf{E}[\,V_{1:\lambda}\,]$$

the convergence rate of the $(1,\lambda)$–EA is simply $\mathsf{E}[\,V_{1:\lambda}\,]$. If the mutation vector Z has standard Gaussian distribution then $\mathsf{E}[\,V_{1:\lambda}\,] \sim \lambda^{-2/\ell}$ for fixed ℓ and increasing λ, whereas $\mathsf{E}[\,V_{1:\lambda}\,] \sim 1 - 2\,\log(\lambda)/\ell$ for fixed λ and increasing ℓ (see [25], pp. 188–190). If Z is a spherical Cauchy vector then $\mathsf{E}[\,Z\,]$ does not exist. As a consequence, $\mathsf{E}[\,\|\theta+\eta\, Z\|^2\,]$ and hence $\mathsf{E}[\,V\,]$ does not exist as well. But this does not preclude that $\mathsf{E}[\,V_{1:\lambda}\,]$ may exist for sufficiently large λ.

Theorem 4

Let $Y = \|\theta + \eta\, Z\|^2$ with $\theta \neq 0 \in \mathbb{R}^\ell$, $\eta > 0$, and where Z is a spherical Student random vector with $s \in \mathbb{N}$ degrees of freedom and dimension ℓ. The kth moment of the ith order statistic $Y_{i:\lambda}$ with $1 \le i \le \lambda$ from a sample of λ i.i.d. random variables of type Y do exist if the relation $2\,k < s\,(\lambda - i + 1)$ is valid. In particular, if $s = 1$ then the expectation exists for $\lambda \ge 3$ and $1 \le i \le \lambda - 2$. If $s = 2$ then the expectation exists for $\lambda \ge 2$ and $1 \le i \le \lambda - 1$. If $s = 3$ then $\mathsf{E}[\,X\,]$ and hence the expectation of all order statistics do exist.

Proof:

It suffices to prove the theorem for random variable $V = Y/\|\theta\|^2$. Since V is nonnegative Lemma 4 reveals that the kth moment $\mathsf{E}[\,V_{i:\lambda}^k\,]$ does exist if and only if

$$\int_0^\infty x^k\, f_V(x)\, F_V^{i-1}(x)\,[\,1-F_V(x)\,]^{\lambda-i}\, dx < \infty\,. \tag{17}$$

Notice that $f_V(\cdot)$ is continuous on $(0,\infty)$ with $f(x) \to 0$ as $x \to 0$. Therefore $F_V(\cdot)$ and hence the entire integrand in (17) do not have singularities. Thus, the integral diverges if the integrand decays proportional to x^α with $\alpha \ge -1$ as $x \to \infty$. Let $g(\cdot)$ denote the integrand. It follows from the theory of regularly varying functions [23, p. 280] that

$$\int_0^\infty g(x)\,dx < \infty \quad \Leftrightarrow \quad \forall x > 0 : \lim_{h\to\infty} \frac{g(h\,x)}{g(h)} = x^\alpha$$

with $\alpha < -1$. Thus one has to consider the limit of the quotient

$$\frac{g(h\,x)}{g(h)} = x^k\,\frac{f_V(h\,x)}{f_V(h)}\left[\frac{F_V(h\,x)}{F_V(h)}\right]^{i-1}\left[\frac{1-F_V(h\,x)}{1-F_V(h)}\right]^{\lambda-i}.$$

Since $F_V(h\,x) \to 1$ and $F_V(h) \to 1$ as $h \to \infty$ it follows that $F_V(h\,x)/F_V(h) \to 1$ as $h \to \infty$. Taking into account the rule of l'Hospital one obtains

$$\lim_{h\to\infty}\frac{1-F_V(h\,x)}{1-F_V(h)} = \lim_{h\to\infty}\frac{x\,f_V(h\,x)}{f_V(h)}$$

revealing that it suffices to investigate the limit behavior of $f_V(h\,x)/f_V(h)$. Owing to (5) the limit is

$$\lim_{h\to\infty}\frac{f_V(h\,x)}{f_V(h)} = x^{-(s+2)/2}\lim_{h\to\infty}\frac{{}_2F_1(a,b;c;d(h\,x))}{{}_2F_1(a,b;c;d(h))} \tag{18}$$

with $a = (\ell+s)/4$, $b = a + 1/2$, $c = \ell/2$, and

$$d(y) = \frac{4\,\delta^4\,y}{(s+\delta^2+\delta^2\,y)^2} \to 0$$

as $y \to \infty$. Since entry 15.1.1 in [20] yields ${}_2F_1(a,b;c;d(y)) \to 1$ as $d(y) \to 0$ the rightmost limit in (18) converges to unity as $h \to \infty$. Putting everything together one finds that

$$\lim_{h\to\infty}\frac{g(h\,x)}{g(h)} = x^{k-s\,(\lambda-i+1)/2-1}.$$

Since the exponent must be smaller than -1, one finally arrives at the desired condition $2\,k < s\,(\lambda - i+1)$. □

Thus, the expected convergence rate $\mathsf{E}[\,V_{1:\lambda}\,]$ does also exist for spherical Cauchy mutations if $\lambda \ge 3$. This observation shows that it makes sense to derive an asymptotical expression for $\mathsf{E}[\,V_{1:\lambda}\,]$ in case of fixed ℓ and $\lambda \gg 1$. For this purpose, regularly varying functions also play an important role [26].

Lemma 5
Let V be a nonnegative continuous random variable with distribution function $F_V(\cdot)$. If for every $x > 0$

$$\lim_{h\to 0}\frac{F_V(x\,h)}{F_V(h)} = x^\alpha \qquad (\alpha > 0) \tag{19}$$

then $\mathsf{P}\{\,V_{1:\lambda}/a_\lambda \le x\,\}$ converges weakly to $[\,1-\exp(-x^\alpha)\,]\cdot 1_{(0,\infty)}(x)$ and conversely. A suitable choice the normalizing constants is $a_\lambda = F_V^{-1}(\lambda^{-1})$. □

The limit distribution in the lemma above is termed the Weibull distribution. Let W have Weibull distribution and assume that the condition (19) is fulfilled for random variable V. In this case one may conclude that $V_{1:\lambda} \approx a_\lambda W$ and hence $\mathsf{E}[V_{1:\lambda}] \approx a_\lambda \mathsf{E}[W]$ for sufficiently large λ. According to (5) and l'Hospital's rule one obtains

$$\lim_{h\to 0} \frac{F_V(x\,h)}{F_V(h)} = x \lim_{h\to 0} \frac{f_V(x\,h)}{f_V(h)} = x^{\ell/2} \lim_{h\to 0} \frac{{}_2F_1(a,b;c;d(h\,x))}{{}_2F_1(a,b;c;d(h))} \tag{20}$$

with $a = (\ell+s)/4$, $b = a + 1/2$, $c = \ell/2$, and

$$d(y) = \frac{4\,\delta^4\,y}{(s+\delta^2+\delta^2\,y)^2} \to 0$$

as $y \to 0$. Again, entry 15.1.1 in [20] yields ${}_2F_1(a,b;c;d(y)) \to 1$ as $d(y) \to 0$ so that the rightmost limit in (20) converges to unity. As a consequence, condition (19) is fulfilled with $\alpha = \ell/2 > 0$. Lemma 5 also implies that $F_V(x) \sim x^{1/\alpha}$ as $x \to 0$ which in turn implies that $a_\lambda = F_V^{-1}(\lambda^{-1}) \sim \lambda^{-1/\alpha} = \lambda^{-2/\ell}$. Since $\mathsf{E}[W] = \Gamma(1+2/\ell)$ one finally arrives at

$$\mathsf{E}[V_{1:\lambda}] \approx a_\lambda\,\mathsf{E}[W] \sim \lambda^{-2/\ell}\,\Gamma(1+2/\ell)$$

for large λ and fixed ℓ. Thus, the order of the convergence rate of the $(1,\lambda)$–EA with spherical Cauchy mutations is identical to the order in the case of Gaussian mutations. To decide which type of mutations actually lead to faster convergence it is necessary to determine the constants hidden by the asymptotical expression $O(\lambda^{-2/\ell})$. A first assessment of the differences can be gained from setting $\ell = 3$ and calculating $\mathsf{E}[V_{1:\lambda}]$ for varying $\lambda \geq 3$. Table 3 summarizes the results revealing that Gaussian mutations consistently lead to faster convergence than spherical Cauchy mutations regardless of the number of offspring $\lambda \geq 3$.

One might conjecture that this relation also holds in dimension $\ell > 3$. Actually, numerical integration and optimization reveals that this relation also holds for $\ell = 31$ but the computational effort to obtain the optimal values is not negligible. Moreover, the knowledge of the optimal values is of no practical interest—it should suffice to know that Gaussian mutations offer faster convergence than Cauchy mutations.

3 Convergence Rate Under Nonspherical Cauchy Mutations

Another multivariate version of the Cauchy distribution can be obtained by drawing a univariate standard Cauchy random number independently for each entry of the random vector. The resulting multivariate distribution is, however, not spherically symmetric. Therefore it cannot be expected that there is a uniform convergence rate being valid for all locations $\theta \in \mathbb{R}^\ell$. In fact, it will turn out that the convergence rate depends not only on the dimension but also on the ratio $\|\theta\|_1/\|\theta\|_2 \in [1, \sqrt{\ell}]$ with $\theta \in \mathbb{R}^\ell$. Here, $\|\cdot\|_1$ denotes the norm

$$\|\theta\|_1 = \sum_{i=1}^{\ell} |\,\theta_i\,|$$

whereas $\|\cdot\|_2$ is the usual Euclidean norm. Notice that the interval bounds for the ratio given above are sharp: If θ is located on some coordinate axis then $\|\theta\|_1 = \|\theta\|_2$, whereas $\|\theta\|_1 = \sqrt{\ell}\,\|\theta\|_2$ if all the entries of vector θ are identical. The result below reveals at which point the norm $\|\cdot\|_1$ enters the scene.

	Cauchy mutations		Gaussian mutations	
λ	c^*	γ^*	c^*	γ^*
3	0.8473	0.3116	0.7633	0.8791
4	0.7702	0.4637	0.6674	1.0398
5	0.7096	0.5776	0.5951	1.1439
6	0.6603	0.6668	0.5389	1.2171
7	0.6193	0.7393	0.4940	1.2716
8	0.5844	0.7998	0.4572	1.3140
9	0.5543	0.8513	0.4265	1.3481
10	0.5280	0.8958	0.4005	1.3761
15	0.4332	1.0531	0.3122	1.4654
20	0.3729	1.1508	0.2604	1.5144
30	0.2986	1.2695	0.2008	1.5681
40	0.2534	1.3409	0.1666	1.5978
50	0.2223	1.3895	0.1440	1.6170
60	0.1995	1.4252	0.1278	1.6305
70	0.1818	1.4528	0.1155	1.6407
80	0.1676	1.4748	0.1058	1.6487
90	0.1559	1.4929	0.0978	1.6561
100	0.1461	1.5081	0.0913	1.6605

Table 3: Optimal convergence rates $c^* \in (0,1)$ and scaling parameters $\gamma^* = \ell/\delta^*$ of the $(1,\lambda)$–EA in dimension $\ell = 3$. Since the convergence rate c^* for Gaussian mutations is consistently smaller than the rate for Cauchy mutation, the convergence velocity is fastest with Gaussian mutations regardless of the number of offspring $\lambda \geq 3$.

Lemma 6 ([17, p. 183])
Let $Z = (Z_1, \ldots, Z_\ell)'$ be a random vector where $Z_1, \ldots, Z_\ell$ are i.i.d. standard Cauchy random variables with p.d.f. as given in (1) with $d = 1$. For every $\theta \in \mathbb{R}$ it holds true that

$$\theta' Z \stackrel{d}{=} \|\theta\|_1 \, Z_1 \, . \qquad \square$$

Thus the decomposition of $\|\theta + \eta\, Z\|_2^2$ will be different from that of the preceding section. According to Lemma 6 one obtains

$$\|\theta + \eta\, Z\|_2^2 - \|\theta\|_2^2 \stackrel{d}{=} 2\,\eta\,\|\theta\|_1\, Z_1 + \eta^2\,\|Z\|_2^2 = \|\theta\|_1^2\,(2\,c\,Z_1 + c^2\,\|Z\|_2^2)$$

where $\eta = c\,\|\theta\|_1$, $c > 0$. Division by $\|\theta\|_1^2 \neq 0$ yields

$$\frac{\|\theta\|_2^2}{\|\theta\|_1^2}\,(V_\ell - 1) \stackrel{d}{=} 2\,c\,Z_1 + c^2 \sum_{i=1}^{\ell} Z_i^2 \, . \tag{21}$$

To proceed one needs to know under which normalization the sum of squares on the r.h.s. of (21) converges to a limit random variable whose distribution is independent from the dimension ℓ. The result below offers the desired information.

Lemma 7 ([27])
Let $Z_1, Z_2, \ldots$ be i.i.d. standard Cauchy random variables with p.d.f. as given in (1) with $d = 1$. As $\ell \to \infty$ then

$$\frac{1}{\ell^2} \sum_{i=1}^{\ell} Z_i^2 \longrightarrow G \quad \text{in distribution}$$

where G has the same distribution as in Lemma 3. □

Thus, one has to choose $c = \gamma/\ell^2$. Insertion into (21) and subsequent multiplication by ℓ^2 yields

$$\ell^2 \, \frac{||\theta||_2^2}{||\theta||_1^2} \, (V_\ell - 1) \stackrel{d}{=} 2\,\gamma\, Z_1 + \gamma^2 \, \frac{1}{\ell^2} \sum_{i=1}^{\ell} Z_i^2 \,. \tag{22}$$

Finally, Lemma 7 ensures that the l.h.s. of (22) converges in distribution to the limit random variable $L = 2\,\gamma\, Z_1 + \gamma^2 G$ as $\ell \to \infty$. Since Z_1 is a standard Cauchy random variable the limit L only depends on γ. Moreover, the limit distribution is identical to the limit distribution in case of *spherical* Cauchy mutation. But notice that the normalizing constants a_ℓ differ. At this point a cautionary remark is necessary: It is not guaranteed that the accuracy of the approximations for given ℓ is equally good for both types of Cauchy mutations.

But if the approximations are equally good (which is assumed for the moment) then the optimal choice for $g(\gamma) = \mathsf{E}[\,\min\{L, 1\}\,]$ may be taken from Table 2. Thus, $\gamma^* = 0.713$ with $g(\gamma^*) = 0.2756$ as in case of *spherical* Cauchy mutations. Since $||\theta||_1^2/||\theta||_2^2 \in [\,1, \ell\,]$ and hence

$$\frac{g(\gamma^*)}{\ell^2} \le \frac{g(\gamma^*)}{\ell^2} \cdot \frac{||\theta||_1^2}{||\theta||_2^2} \le \frac{g(\gamma^*)}{\ell}$$

it would follow that *spherical* Cauchy mutations generally lead to faster convergence than mutation vectors with independent Cauchy random variables.

Now assume that the approximations are not equally good for given ℓ. Then there exists a function $\tilde{g}(\gamma)$, attaining its minimum at $\tilde{\gamma}^*$, that replaces $g(\gamma)$ in case of i.i.d. Cauchy mutations. It may be expected that $\tilde{\gamma}^*$ and hence $\tilde{g}(\tilde{\gamma}^*)$ quickly stabilizes for increasing ℓ. Thus, $g(\gamma^*)$ as well as $\tilde{g}(\tilde{\gamma}^*)$ may be regarded as constants provided that ℓ is sufficiently large. If $\tilde{g}(\tilde{\gamma}^*) \le g(\gamma^*)$ then spherical Cauchy mutations lead to faster convergence than nonspherical Cauchy mutations. Even if $\tilde{g}(\tilde{\gamma}^*) > g(\gamma^*)$ then there exists an ℓ_0 such that in case of a fixed ratio $||\theta||_1^2/||\theta||_2^2$ the relation

$$\frac{\tilde{g}(\tilde{\gamma}^*)}{\ell^2} \le \frac{\tilde{g}(\tilde{\gamma}^*)}{\ell^2} \cdot \frac{||\theta||_1^2}{||\theta||_2^2} \le \frac{g(\gamma^*)}{\ell}$$

is valid for all $\ell > \ell_0$. This observation reveals that spherical Cauchy mutations offer faster convergence than nonspherical Cauchy mutations as $\ell \to \infty$.

As an illustration of the differences consider the following numerical experiment: The $(1+1)$-EA was run 100 times with starting point $\theta_0 = (1000, \ldots, 1000)'$. Owing to (9) it was measured how many iterations were necessary to reduce the error by $\Delta = 50$ orders of magnitude. The scaling parameter η for the mutation distribution was set to the optimal value

$$\eta^* = \begin{cases} \dfrac{1.224}{\ell}\, ||\theta||_2 & \text{for Gaussian mutations} \\[2ex] \dfrac{0.713}{\ell}\, ||\theta||_2 & \text{for spherical Cauchy mutations} \\[2ex] \dfrac{0.713}{\ell^2}\, ||\theta||_1 & \text{for nonspherical Cauchy mutations} \end{cases}$$

according to the preceding theoretical analysis (it was assumed that the approximations for spherical and nonspherical Cauchy mutations are equally good for given ℓ). Figure 3 reveals that the running time increases linearly in ℓ for spherical Cauchy and Gaussian mutations, but quadratically in ℓ for nonspherical Cauchy mutations.

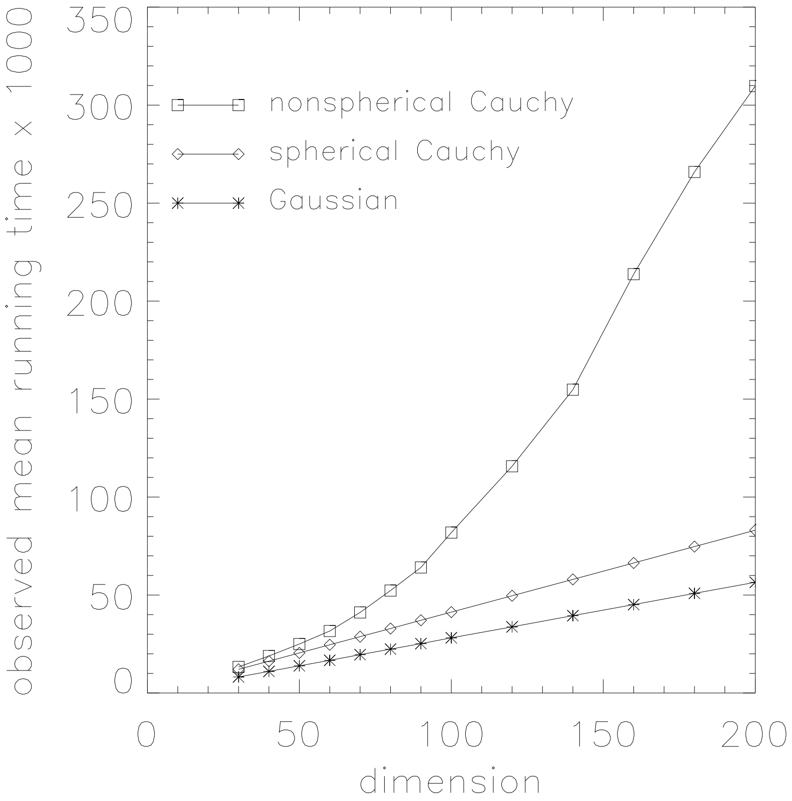

Figure 3: Observed mean running time of the $(1+1)$-EA to reduce the error by 50 orders of magnitude under Gaussian, spherical and nonspherical Cauchy mutations with optimal scaling parameter η^*. The running time increases linearly in ℓ for spherical Cauchy and Gaussian mutations, but quadratically in ℓ for nonspherical Cauchy mutations.

4 Implications for Self-Adaptive Mutation Mechanisms

In the analysis presented so far it was tacitly presupposed that the EA has knowledge about its Euclidean distance to the optimum in order to optimally adjust the mutation distributions—an assumption that is usually not justified in practice. In contemporary evolutionary algorithms with multiple offspring the task of adjusting the mutation distribution is accomplished by a mechanism termed "self-adaptation." The probably most popular version was introduced by Schwefel [2] and works as follows.

Consider a $(1,\lambda)$-EA with $\lambda \geq 2$ and a mutation distribution that is adjustable by a single parameter. The parent at iteration $t \geq 0$ consists of the pair (θ_t, η_t) where $\theta_t \in \mathbb{R}^\ell$ is the current position in the search space and η_t the scale parameter of the mutation distribution. An offspring $(\theta_t^{(i)}, \eta_t^{(i)})$ with $i = 1, \ldots, \lambda$ is produced according to

$$\begin{aligned} \eta_t^{(i)} &= \eta_t \cdot \exp(N) \\ \theta_t^{(i)} &= \theta_t + \eta_t^{(i)} \cdot Z \end{aligned}$$

where Z is a random vector with some fixed mutation distribution and N is a Gaussian random variable with zero mean and variance τ^2. Since the random variable $\exp(N)$ is lognormally distributed the probability of increasing the scale parameter η_t at least by factor $a > 1$ is equal to the probability of decreasing η_t at least by factor $1/a$. More specifically, if $b > a > 1$ then

$$\mathsf{P}\{\, a\,\eta_t \le \eta_t \exp(N) \le b\,\eta_t \,\} = \Phi\left(\frac{\log b}{\tau}\right) - \Phi\left(\frac{\log a}{\tau}\right) = \mathsf{P}\{\, \eta_t/b \le \eta_t \exp(N) \le \eta_t/a \,\}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard Gaussian random variable. The reason for permitting a potential enlargement of the scale parameter rests on the fact that the initial setting of η_0 may be too small. In this case the scale parameter must be increased until reaching a nearly optimal value. As soon as this has happened the scale parameter should decrease.
Needless to say, the crucial point in achieving this behavior is an appropriate setting of τ. It is clear that a theoretical argumentation must be based on the dynamics of the process. Beyer [28] has given a detailed treatise of this topic in the case of Gaussian random vectors Z. A similar consideration for Cauchy random vectors, however, is beyond the scope of this paper. Instead, a heuristic argumentation from a more static point of view is offered here.
Assume that the scale parameter η_t at iteration $t \ge 0$ is optimally adjusted. If Z is a Gaussian or spherically symmetric Cauchy random vector then $\eta_t^* = \gamma^* \, \|\theta_t\|_2/\ell$. To achieve an optimally adjusted scale parameter in the next iteration, the current scale parameter should be decreased by the factor

$$\frac{\eta_{t+1}^*}{\eta_t^*} = \frac{\|\theta_{t+1}\|_2}{\|\theta_t\|_2}\,.$$

Since $\mathsf{E}[\, \|\theta_{t+1}\|_2^2 \,|\, \theta_t \,] = c\,\|\theta_t\|_2^2$, where $c \in (0,1)$ is the convergence rate, Jensen's inequality yields $\mathsf{E}[\, \|\theta_{t+1}\|_2 \,|\, \theta_t \,] \le c^{1/2}\,\|\theta_t\|_2$ and hence $\eta_{t+1}^*/\eta_t^* \approx c^{1/2}$. It appears plausible that the realizations of the lognormal random variable $\exp(N)$ should be placed more frequently in the vicinity of $c^{1/2}$ than in the vicinity of any other point. This property can be achieved by adjusting the distribution of $\exp(N)$ such that its mode equals $c^{1/2}$, i.e., $\exp(-\tau^2) = c^{1/2}$. Recall from Section 2.4 that $c \sim \lambda^{-2/\ell}$ for Gaussian as well as spherical Cauchy mutations. This leads to

$$\exp(-\tau^2) = c^{1/2} \sim \lambda^{-1/\ell} = \exp(-\log(\lambda)/\ell) \quad \Leftrightarrow \quad \tau \sim \left(\frac{\log \lambda}{\ell}\right)^{1/2}$$

for large ℓ. Notice that this relationship was also established in [28] in the case of Gaussian random vectors.
Now let Z be a nonspherical Cauchy vector. The optimal scale parameter is $\eta_t^* = \gamma^* \, \|\theta_t\|_1/\ell$. As a consequence, the reduction factor should be

$$\frac{\eta_{t+1}^*}{\eta_t^*} = \frac{\|\theta_{t+1}\|_1}{\|\theta_t\|_1} = \frac{\alpha_{t+1}}{\alpha_t}\,\frac{\|\theta_{t+1}\|_2}{\|\theta_t\|_2}\,. \tag{23}$$

where $\alpha_t = \|\theta_t\|_1/\|\theta_t\|_2$ with $\alpha_t \in [1, \sqrt{\ell}]$. Recall from Section 2.4 that $\|\theta_{t+1}\|_2^2 = V_{1:\lambda}(\ell)\,\|\theta_t\|_2^2$. Owing to (22) one obtains

$$\frac{\ell}{\alpha^2}\,(V_{1:\lambda}(\ell) - 1) \overset{d}{\approx} L_{1:\lambda} \tag{24}$$

where $L_{1:\lambda}$ is the minimum of λ independent random variables possessing the distribution of the limit random variable L. Notice that $\mathsf{E}[\, L_{1:\lambda} \,] < 0$ for sufficiently large $\lambda \ge 3$. Let $h(\lambda) = |\, \mathsf{E}[\, L_{1:\lambda} \,] \,| > 0$. Under usage of (24) the reduction factor in (23) can be approximated by

$$\frac{\eta_{t+1}^*}{\eta_t^*} \approx \frac{\alpha_{t+1}}{\alpha_t}\sqrt{1 - \frac{\alpha_t^2\, h(\lambda)}{\ell^2}} \approx \frac{\alpha_{t+1}}{\alpha_t}\,\exp\left(-\frac{\alpha_t^2\, h(\lambda)}{2\,\ell^2}\right) \tag{25}$$

for large ℓ. Notice that the sequence $(\alpha_t : t \geq 0)$ changes its values only gradually. Therefore it may be assumed that $\alpha_{t+1}/\alpha_t \approx 1$. Again, insist that the mode of $\exp(N)$ should be approximately equal to the reduction factor. Owing to the rightmost approximation in (25) one finally obtains the relationship

$$\tau \approx \frac{\alpha_t \sqrt{h(\lambda)/2}}{\ell} .$$

Notice that τ depends on $\alpha_t = \|\theta_t\|_1/\|\theta_t\|_2 \in [1, \sqrt{\ell}]$. As a consequence, even a more rigorous theoretical analysis would not lead to an optimal *fixed* value for τ.

5 Conclusions

If fast local convergence is desirable, Gaussian mutations are preferable to spherical Cauchy mutations which are in turn preferable to nonspherical Cauchy mutations. If the problem dimension is fixed then each of the three mutation distributions leads to an exponentially fast approach to the local optimum. But the differences between the convergence velocities associated with these three distributions get larger as the problem dimension increases. Whereas the number of iterations required to reduce the objective function value by a certain amount under Gaussian or spherical Cauchy mutations increase as a linear function of the problem dimension, the number of iterations increase as a quadratic function of the problem dimension if nonspherical Cauchy mutations are used. But since fast local convergence enhances the danger that the evolutionary algorithm may be quickly trapped by local minima, these results may be interpreted as an advantage of nonspherical Cauchy mutations in the case of multimodal optimization problems.
From a practical point of view, nonspherical Cauchy mutations require another parametrization of the self-adaptation mechanism: the parameter τ^2 of the lognormal distribution should be proportional to ℓ^{-2} in lieu of ℓ^{-1}. This observation leads to the recommendation that the parametrization of the self-adaptation mechanism should be carefully reviewed whenever another mutation distribution than the Gaussian distribution is employed in an evolutionary algorithm.

Acknowledgment
This work is a result of the *Collaborative Research Center SFB 531* supported by the *Deutsche Forschungsgemeinschaft (DFG).*

References

[1] I. Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution.* Frommann-Holzboog Verlag, Stuttgart, 1973.

[2] H.-P. Schwefel. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie.* Birkhäuser, Basel, 1977.

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

[4] D. B. Fogel. *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence.* IEEE Press, New York, 1995.

[5] T. Bäck. *Evolutionary Algorithms in Theory and Practice.* Oxford University Press, New York, 1996.

[6] Y. S. Chow and H. Teicher. *Probability Theory.* Springer, New York, 1978.

[7] P. H. Müller, editor. *Wahrscheinlichkeitsrechnung und mathematische Statistik: Lexikon der Stochastik.* Akademie Verlag, Berlin, 5th edition, 1991.

[8] J.-P. Bouchaud and A. Georges. Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications. *Physics Reports*, 195(3 & 4):127–293, 1990.

[9] B. H. Lavenda. Limitations of Boltzmann's principle. *International Journal of Theoretical Physics*, 34(4):605–614, 1995.

[10] H. Szu and R. Hartley. Fast simulated annealing. *Physics Letters A*, 122(3/4):157–162, 1987.

[11] H. Szu and R. Hartley. Nonconvex optimization by fast simulated annealing. *Proceedings of the IEEE*, 75(11):1538–1540, 1987.

[12] L. Ingber. Very fast simulated re–annealing. *Mathematical and Computer Modelling*, 12(8):967–973, 1989.

[13] 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.

[14] 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.

[15] N. Saravanan and D. B. Fogel. Multi-operator evolutionary programming: A preliminary study on function optimization. In P. J. Angeline, R. G. Reynolds, J. R. McDonnell, and R. Eberhart, editors, *Proceedings of the Sixth Annual Conference on Evolutionary Programming*, pages 215–221. Springer, Berlin, 1997.

[16] 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.

[17] K.-T. Fang, S. Kotz, and K.-W. Ng. *Symmetric Multivariate and Related Distributions.* Chapman and Hall, London and New York, 1990.

[18] T. Cacoullos and M. Koutras. Quadratic forms in spherical random variables: Generalized noncentral χ^2 distribution. *Naval Research Logistics Quarterly*, 31:447–461, 1984.

[19] N. L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Distributions - 2.* Wiley, New York, 1970.

[20] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions.* Dover Publications, New York, 1965.

[21] A. Gut. *An Intermediate Course in Probability.* Springer, New York, 1995.

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

[23] W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. 2.* Wiley, New York, 2nd edition, 1971.

[24] H. A. David. *Order Statistics.* Wiley, New York, 1970.

[25] G. Rudolph. *Convergence Properties of Evolutionary Algorithms.* Kovač, Hamburg, 1997.

[26] S. L. Resnick. *Extreme values, regular variation, and point processes.* Springer, New York, 1987.

[27] F. Eicker. Sums of independent squared Cauchy variables grow quadratically: Applications. *Sankhyã A*, 47:133–140, 1985.

[28] H.-G. Beyer. Toward a theory of evolution strategies: Self–adaptation. *Evolutionary Computation*, 3(3):311–347, 1995.