When Do Heavy-Tail Distributions Help?

Nikolaus Hansen[1], Fabian Gemperle[2], Anne Auger[1], and Petros Koumoutsakos[1,2]

[1] Computational Laboratory, ETH Zurich, CH-8092, Switzerland
[2] Institute of Computational Science, ETH Zurich, CH-8092, Switzerland

Abstract. We examine the evidence for the widespread belief that heavy tail distributions enhance the search for minima on multimodal objective functions. We analyze isotropic and anisotropic heavy-tail Cauchy distributions and investigate the probability to sample a better solution, depending on the step length and the dimensionality of the search space. The probability decreases fast with increasing step length for isotropic Cauchy distributions and moderate search space dimension. The anisotropic Cauchy distribution maintains a large probability for sampling large steps along the coordinate axes, resulting in an exceptionally good performance on the separable multimodal Rastrigin function. In contrast, on a non-separable rotated Rastrigin function or for the isotropic Cauchy distribution the performance difference to a Gaussian search distribution is negligible.

1 Introduction

The optimization of multimodal objective functions is recognized as a fundamental problem in several areas of science and engineering. Stochastic search procedures such as Simulated Annealing or Evolutionary Algorithms are well-established methods to optimize multimodal objective functions. New candidate solutions are often sampled from isotropic multivariate Gaussian distributions. The choice of Gaussian distributions has several reasons. Isotropic Gaussian distributions do not favor any direction in the search space. Gaussian distributions are amenable to mathematical analysis because they are the only stable distribution—where the sum of iid variates has the same type of distribution as its summands—with finite variance. For a given variance, the Gaussian distribution has the maximal entropy, which can be interpreted in that the distribution shape contains the least additional assumptions on the objective function to be optimized. Finally, Gaussian distributions suggest themselves for bioinspired algorithms as they are widely observed in nature as for example in the distribution of phenotypic traits.

On the other hand, it is a common belief that, when employed for the optimization of *multimodal* objective functions, the exponentially decreasing tails of Gaussians are ineffective [8]. Instead, it is argued that heavy tails, such as those of the Cauchy distribution, are more appropriate, as long jumps occasionally lead to better solutions, that eventually lie within the attraction region of a better (local) optimum. Long jumps that produce worse solutions should be disregarded in general.[3] In this context several

[3] A search strategy is highly susceptible to divergence, if worse solutions from long jumps are accepted. The danger of divergence is way smaller, if an accepted worse solution is originated from a short step. Alternatively, worse solutions from long jumps can be exploited with local optimization, which is not considered in this paper.

search strategies that apply heavy tail distributions have been investigated, including Fast Simulated Annealing [6] and Fast Evolution Strategies [8]. In these strategies one key difference concerns the use of isotropic [6] and anisotropic [8] heavy tail distributions. The importance of the anisotropy of the coordinate-wise iid multivariate Cauchy distribution was already recognized in [5, 2]. Obuchowicz [2] observed a degradation of performance of the anisotropic distribution when rotating the search space. He proposed isotropic Gauss and Cauchy distributions with norms distributed as their one-dimensional counter parts with mixed results.

Rowe and Hidovic [4] investigated the use of a scale free distribution that allowed searching simultaneously on a given range of scales. In one-dimensional problems, the scale free distribution is uniformly distributed on the log scale in that $\Pr(x \in [a, b]) \propto \log b - \log a$, given a and b are in the supported range. We found the n-dimensional version of this scale free distribution to be highly anisotropic (similar to Fig. 3, lower right). Surprisingly, even with a (1+1)-selection scheme the scale free distribution shows exceptional performance on the multimodal Rastrigin function and this is explained with the advantage of long jumps. We summarize the common hypothesis.

Hypothesis 1 *Long jumps, attributed to sampling from heavy-tail or scale free distributions, occasionally lead to better solutions. They are therefore helpful for searching multimodal objective functions.*

On the other hand, for the *unimodal* sphere model, where $f(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2$, theoretical investigations and experiments show that compared to the Cauchy distribution the Gaussian consistently leads to faster convergence of the $(1, \lambda)$-evolution strategy, regardless of the choice of λ [5].

This paper investigates why and when heavy tails can help for global optimization. The goal is (a) to *quantify* the *possible* effect of heavy tails and (b) to separate the effects of the heavy tail and the anisotropy of the search distribution in a carefully chosen experimental set-up. The paper is structured as follows: in Sect. 2 the relation between step length and search space volume is discussed. In Sect. 3 search distributions are introduced. Their characteristics and potential impact are investigated in Sect. 4. In Sect. 5 simulations of an evolutionary algorithm are presented and Sect. 6 gives a short conclusion.

2 The Search Space Volume Phenomenon

The so-called *curse of dimensionality* casts doubt on Hypothesis 1: the search space volume increases exponentially fast with increasing dimension and large steps become more and more unsuccessful. Rechenberg [3, p.160ff] analyzes the situation for the 30-dimensional Rastrigin function. He finds only a narrow evolution window for jumps that can *initiate* successful new subpopulations. Here jumps are not expected to produce better solutions but to converge to better local optima in a local optimization. Smaller steps fall back into the originating local optimum, larger steps converge into worse optima.

The volume covered by a step of length r is given by the hypersphere surface area $S_n(r) = \frac{2\pi^{n/2}}{\Gamma(n/2)} r^{n-1}$, where r is the distance to the center, n is the dimension, and Γ is

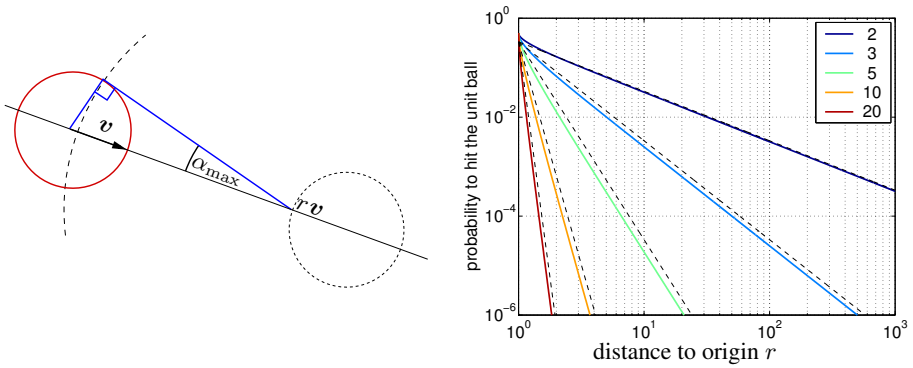

Fig. 1. Probability to hit the unit hyperball (solid) sampling from $r\boldsymbol{v}$ as mean with an optimal isotropic distribution, where $\boldsymbol{v} \in \mathbb{R}^n$ and $\|\boldsymbol{v}\| = 1$. The plots on the right show results for $n = 2, 3, 5, 10, 20$, from above to below. Dashed lines depict the approximation $\frac{1}{3\,r^{n-1}}$.

the Gamma function. The covered volume increases with r^{n-1} making it increasingly difficult to hit a particular area of given volume.

We investigate an idealized scenario for the probability to find *a better solution* by jumping into another region of attraction, as depicted in **Fig. 1** (left).

The arrow depicts a vector $\boldsymbol{v}$ with unit length. The starting point is $r\boldsymbol{v}$, located on the dotted circle on the right. Its closest local minimum is inside the dotted circle. A second volume of better solutions lies on the left, inside the unit hyperball around the coordinate system origin. We compute the probability to hit the unit hyperball by sampling around $r\boldsymbol{v}$ isotropically. We assume an optimal step-length distribution, where all steps lie on the hypersphere surface, corresponding to the dashed arc on the left. To hit the unit hyperball the angle between the sampled vector and $-\boldsymbol{v}$ has to be smaller than $\alpha_{\max} = \arcsin(1/r)$. Using the cumulative distribution function of the angle between a reference vector and a random vector uniformly distributed on the unit hypersphere [1, Theorem 9] we deduce the probability to hit the unit hyperball as

$$\frac{1}{2} - \frac{1}{2}\frac{\Gamma(n/2)}{\Gamma(1/2)\Gamma((n-1)/2)} \int_0^{1-\frac{1}{r^2}} t^{-\frac{1}{2}}(1-t)^{(n-3)/2} dt \ . \tag{1}$$

The probability is plotted as a function of r in **Fig. 1** (right) using Matlab's function `betainc`. Dashed lines depict $\frac{1}{3\,r^{n-1}}$, resembling the dependency of the hypersphere surface area on r, which turns out to be a reasonable approximation of (1). Even for moderate dimensions the probability drops fast with increasing r and becomes 10^{-4} for $r = 6, 2, 1.5$ and $n = 5, 10, 20$ respectively. For $r = 1$ the scenario resembles the sphere function and the success probability is 0.5 (for infinitesimally small step length). Our observations are summarized in an alternative hypothesis.

Hypothesis 2 *Long jumps virtually never lead to better solutions in high dimensional search spaces, because they get lost in the huge search space volume.*

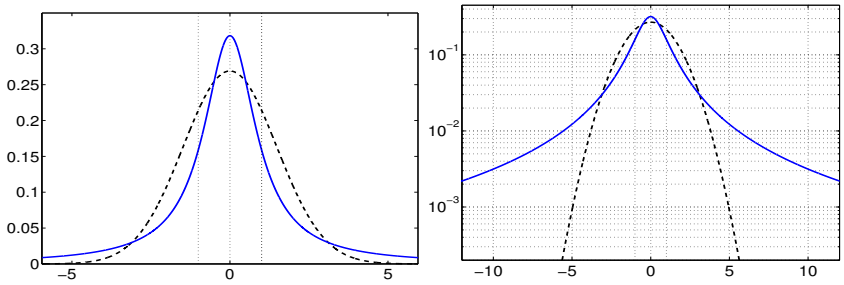

Fig. 2. Densities of the univariate normal (Gaussian) distribution (dashed) and the standard Cauchy distribution (solid) in a linear and a semi-log plot. The standard deviation of the normal distribution $\sigma = 1.4826$ is chosen such that the quartile values equal $-1, 0, 1$ (vertical dotted lines) as for the Cauchy distribution.

3 Search Distributions

3.1 Univariate Gaussian and Cauchy Distribution

The distributions that will be used in this paper are derived from the univariate standard normal and Cauchy distribution. The univariate normal distribution with zero mean and variance σ^2 obeys the density

$$f_{\mathcal{N}(0,\sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad . \tag{2}$$

The univariate Cauchy distribution with median zero and upper quartile τ obeys

$$f_{\mathcal{C}(0,\tau)}(x) = \frac{1}{\tau\,\pi}\,\frac{1}{x^2/\tau^2 + 1} = \frac{1}{\pi}\,\frac{\tau}{x^2 + \tau^2} \quad . \tag{3}$$

A standard Cauchy distributed number, where $\tau = 1$ can be sampled by dividing two independent, standard normally distributed random numbers. Furthermore $\mathcal{C}(0,\tau) \sim \tau\mathcal{C}(0,1)$, and $\mathcal{N}(0,\sigma^2) \sim \sigma\mathcal{N}(0,1)$. **Figure 2** shows the densities of both univariate distributions.

3.2 Multivariate Distributions

We consider both isotropic and anisotropic distributions [4, 7, 8], and as isotropic distributions we consider a heavy-tail distribution and a distribution with exponentially fast decreasing tail.

We use $\mathcal{G}_n$ to denote an n-dimensional Gaussian (normally) distributed random vector with zero mean and identity covariance matrix. The distribution $\mathcal{G}_n$ can be sampled by sampling independent standard $(0,1)$-normally distributed random numbers from Eq. 2 for each component of a vector. Furthermore, let $\mathcal{U}_n$ denote a uniform distribution on the n-dimensional unit hypersphere, where $\Pr(\|\mathcal{U}_n\| = 1) = 1$. The distribution $\mathcal{U}_n$

can be sampled by sampling $\mathcal{G}_n$ and normalizing the resulting vector to length one, *i.e.* $\mathcal{U}_n = \mathcal{G}_n / \|\mathcal{G}_n\|$.

The following search (mutation) distributions are used.

$\mathcal{C}_n \in \mathbb{R}^n$, an (anisotropic) n-dimensional Cauchy distribution, where each coordinate is independent standard $(0, 1)$-Cauchy distributed. This distribution is used, for example, in Fast Evolution Strategies [8] and Fast Evolutionary Programming [7].

$\mathcal{C}_n^{\text{iso}} \sim \|\mathcal{C}_n\| \times \mathcal{U}_n$, an *isotropic* n-dimensional distribution with the norm distributed as for $\mathcal{C}_n$.

$\mathcal{G}_n \sim \|\mathcal{G}_n\| \times \mathcal{U}_n$, the n-dimensional Gaussian (normal) distribution which is widely used in Evolutionary Algorithms such as Evolution Strategies or Evolutionary Programming. The distribution is isotropic (spherical), its norm is χ_n-distributed.

The distributions $\mathcal{C}_n$ and $\mathcal{C}_n^{\text{iso}}$ have polynomially decreasing (heavy) tails. The distributions $\mathcal{C}_n^{\text{iso}}$ and $\mathcal{G}_n$ are isotropic (spherical), and can be sampled by a product of a random number, i.e. a scalar representing the norm, and $\mathcal{U}_n$.

4 Characteristics of the Distributions

Figure 3 shows 10000 sampled points of $\mathcal{C}_2$ and $\mathcal{G}_2$ visualizing the characteristics of the distributions in 2D. For values between -3 and 3 the results of the Gaussian (first row) and the Cauchy distribution (second row) are comparable. While the Gaussian rarely realizes steps larger than five, the Cauchy distribution reveals a surprising picture. Zooming out further the distribution starts to resemble a cross parallel to the coordinate system (third row). That means, the distribution comes close to coordinate-wise sampling on the large scale.

Figure 4 presents data in the 10-dimensional case. Shown are densities of the vector norms (left) and densities along $r\boldsymbol{v} \in \mathbb{R}^{10}$, where r is a scalar and $\boldsymbol{v}$ is fixed, $\|\boldsymbol{v}\| = 1$ (right). The density for the norm of $\mathcal{C}_{10}^{\text{iso}}$ was obtained by Monte-Carlo simulations (about 10^9 samples), the respective density on the right by dividing with the hypersphere surface area $S_n(r) = \frac{2\pi^{n/2}}{\Gamma(n/2)}\, r^{n-1}$. The remaining densities are well-known or can be easily obtained analytically.

Comparing the lower and the upper bold graph in the right figure, again a striking difference between diagonal and coordinate axis parallel density can be recognized for $\mathcal{C}_n$. As can be derived from (3) (the multivariate density derives from a product of the univariate) the coordinate axis parallel density drops proportional to r^2, while the diagonal drops proportional to r^{2n}, for large r.

Two Gaussian densities along $r\boldsymbol{v}$ are shown. First $3.8 \times \mathcal{G}_{10}$ (dashed graph), where the median of the norm corresponds to the one of the Cauchy distributions. Second $1.25 \times \mathcal{G}_{10}$, where the density for small r compares to the one of the Cauchy distributions.

We compare the Gaussians with the isotropic Cauchy distribution (middle bold graph). In one case the density of the Gaussian drops below the Cauchy density for r larger than about 5.6. In the other case only for r between about 8 and 17 the Gaussian reveals a larger density than the isotropic Cauchy distribution $\mathcal{C}_n^{\text{iso}}$. For larger r Gaussian and Cauchy densities drop fast: for $\mathcal{C}_n^{\text{iso}}$ the slope is approximately r^{-10}. For example,

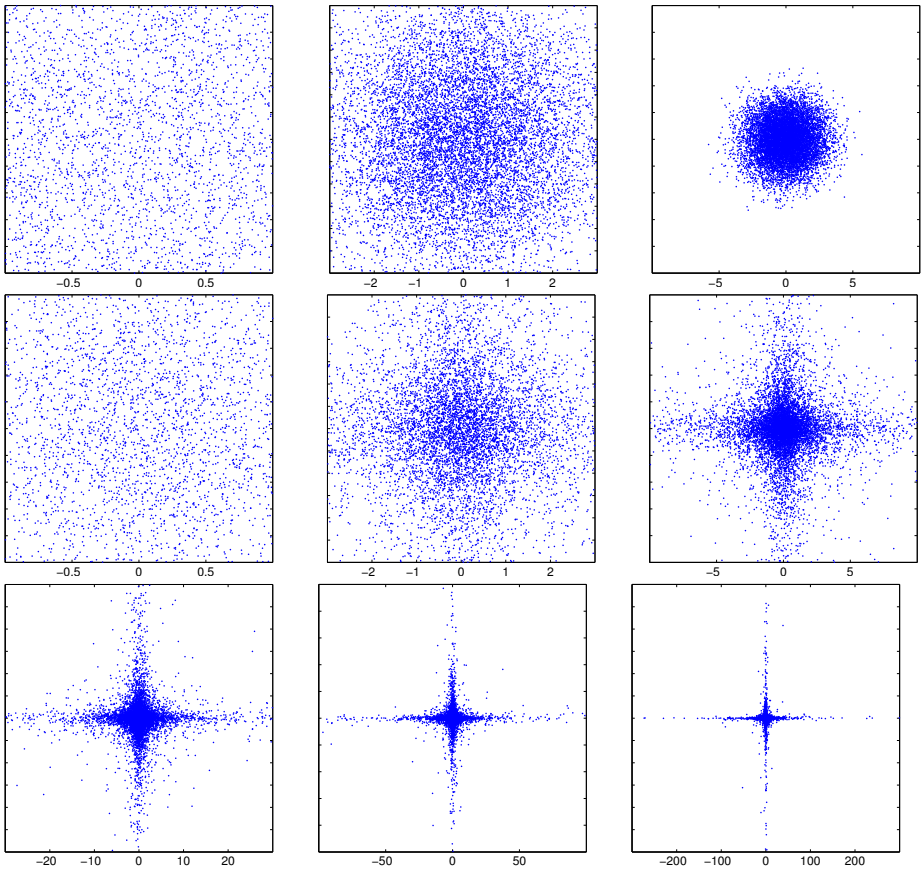

Fig. 3. Ten thousand 2D sample points from the Gaussian distribution $1.4826 \times \mathcal{G}_2$ (upper row) and the Cauchy distribution $\mathcal{C}_2$ (middle and lower row). Shown are the same sampled points on different scales ($\pm 1, \pm 3, \pm 10, \pm 30, \ldots$). The clippings contain $26, 91$, and 100% of the points for $\mathcal{G}_2$ and $25, 67, 88, 96, 98.72$, and 99.55% of the points for $\mathcal{C}_2$. For the larger scales $\mathcal{C}_n$ becomes mainly coordinate-wise sampling.

the probability to hit a volume in a distance of $60 = 3 \times 20$ is about $3^{10} \approx 10^5$ times lower than to hit the same volume in distance 20, a distance where $\mathcal{C}_n^{\text{iso}}$ and $3.8 \times \mathcal{G}_n$ have comparable densities. The other way around, the volume that can be found with a comparable probability by steps being three times longer needs to be 10^5 times larger. In contrast, for the coordinate axis direction the density drops slowly and volumes far away have a considerable probability of being reached.

We can draw two conclusions from these figures. First, the anisotropy of the Cauchy distribution might have a considerable effect on the search behavior. Second, compared to the Gaussian distribution that operates on a reasonable scale of search, the heavy tails should not be of great help. Both conclusions are confirmed in our experimental results.

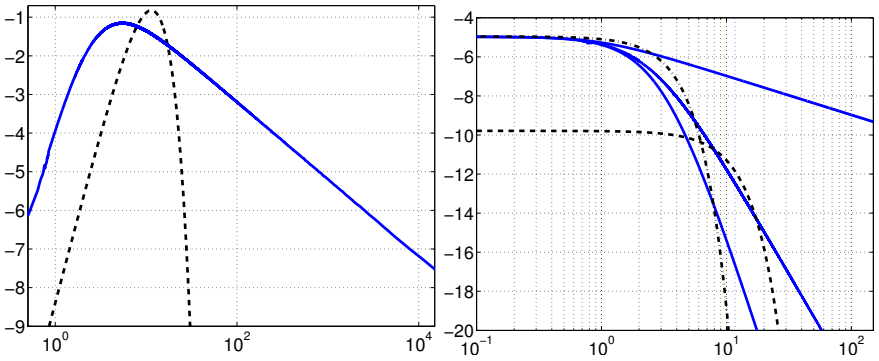

Fig. 4. Densities for $n = 10$ on the $\log_{10}$ scale. **Left**: Density of norms, $\|\mathcal{C}_{10}\|$ and $\|\mathcal{C}_{10}^{\mathrm{iso}}\|$ (same solid graph), and $3.8 \times \|\mathcal{G}_{10}\|$ (dashed), where the factor is chosen such that the median equals to 11.7 as for $\|\mathcal{C}_{10}\|$. **Right**: densities along $r\boldsymbol{v} \in \mathbb{R}^n$ versus r, where $\|\boldsymbol{v}\| = 1$. For $\mathcal{C}_n$ (solid) in coordinate axis direction ($\boldsymbol{v} = (1, 0, \ldots, 0)^{\mathrm{T}}$, upper graph) and in diagonal direction ($\boldsymbol{v} = (1, \ldots, 1)^{\mathrm{T}}/\sqrt{10}$, lower graph), for $\mathcal{C}_n^{\mathrm{iso}}$ (middle solid graph), for $3.8 \times \mathcal{G}_n$ (dashed), and for $1.25 \times \mathcal{G}_n$ (dashed dotted).

5 Simulation Results for the (1+1)-EA

5.1 The Test Functions and Evolutionary Algorithm

We use the highly multimodal Rastrigin function

$$f_{\mathrm{Rastrigin}} : \boldsymbol{x} \mapsto 10n + \sum_{i=1}^{n} y_i^2 + 10\cos(2\pi y_i) \ ,$$

where $\boldsymbol{y} = \boldsymbol{M}\boldsymbol{x}$ and $\boldsymbol{M}$ is an orthogonal matrix ($\boldsymbol{M}^{-1} = \boldsymbol{M}^{\mathrm{T}}$). We investigate two situations. First, the axis parallel Rastrigin function $f_{\mathrm{Rastrigin}}$, where $\boldsymbol{M} = \boldsymbol{I}$ is the identity matrix. The axis parallel Rastrigin function is separable and can therefore be solved by n one-dimensional optimization procedures *parallel to the coordinate axes*. Second, we consider the rotated Rastrigin function, with a randomly chosen $\boldsymbol{M}$, where all columns of $\boldsymbol{M}$ are uniformly distributed on the unit hypersphere and orthogonal, achieved by Gram-Schmidt orthogonalization of $\mathcal{G}_n$-distributed vectors. In the relevant region for $\boldsymbol{x} \in [-5, 5]^n$, the local optima of the Rastrigin function have function values that are close to integer values, which makes the integer bin centers used for the frequency histograms below particularly meaningful.

We apply the (1+1) evolutionary algorithm (EA) as depicted in **Fig. 5 (left)** in order to address the question whether and how the heavy tails can influence the global search performance. If not stated otherwise, we choose $\alpha = 10^{\frac{1.2}{10^4 n}}$, $\theta_{\mathrm{final}} = 10^{-3}$, and the initial $\theta_{\mathrm{start}} = 10^3$, leading to $50\,000 \times n$ iteration steps, and initial $\boldsymbol{x} = \boldsymbol{M}^{-1}(5, \ldots, 5)^{\mathrm{T}}$. The values for α result into $\alpha \approx 1.0000921, 1.0000553, 1.0000276$, for $n = 3, 5, 10$, all smaller than $1 + 10^{-4}$.

Neither (self-)adaptation nor a large population is applied so as to not interfere with the effects of the search distribution. *Adaptation* of distribution parameters, like the

The Algorithm

choose $\mathcal{D}_n, \theta_{\mathrm{start}}, \theta_{\mathrm{final}}, \alpha$
initialize $\boldsymbol{x}$, $\theta = \theta_{\mathrm{start}}$
while $\theta > \theta_{\mathrm{final}}$
  $\boldsymbol{x}' = \boldsymbol{x} + \theta \times \mathcal{D}_n$
  if $f(\boldsymbol{x}') \leq f(\boldsymbol{x})$
    $\boldsymbol{x} = \boldsymbol{x}'$
  $\theta \leftarrow \theta/\alpha$

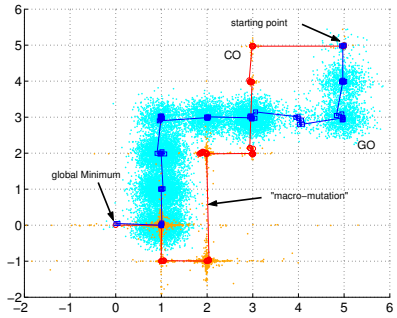

Fig. 5. The Evolutionary Algorithm (left), and paths and sampled points of two runs in 2D, where $\mathcal{D}_n = \mathcal{G}_n, \theta = 0.25$ (square marks □), and $\mathcal{D}_n = \mathcal{C}_n, \theta = 0.01$ (circle marks ○). The marks denote realized steps, where $f(\boldsymbol{x}') \leq f(\boldsymbol{x})$. The optima lie on an axis parallel grid allowing the Cauchy distribution to reach the vicinity of the global optimum about ten times faster.

step-size θ, is not expected to improve the global search performance, as it usually drops step lengths much faster than the given schedule. Large populations, and eventually recombination, will usually improve the performance, but this should be true for all distributions applied. The rationale behind this set-up is to slowly move through all scales and to allow any scale, in case, to conduct the search successfully. It takes about $2500n$ iterations to reduce θ by a factor of two.

Figure 5 shows two runs *on the axis parallel* Rastrigin function, where $n = 2$ and $\alpha = 1$, one run with $\mathcal{D}_n = \mathcal{G}_n$ and $\theta = 0.25$, one run with $\mathcal{D}_n = \mathcal{C}_n$ and $\theta = 0.01$. In both cases θ is chosen much too small. While the Cauchy distribution needs about 9000 iterations, the Gaussian needs about $80\,000$ iterations to approach the global optimum. Having in mind the 2D image of the Cauchy distribution $\mathcal{C}_n$ the result and the resulting picture are not surprising.

5.2 Results

Methods We conducted experiments on the axis parallel and the rotated Rastrigin function for dimensions $n = 3, 5, 10$, performing in each case 50 runs. We judge *performance* in terms of reached final function value and success rate to reach the global optimum with a precision of 10^{-2}. We compared success rates with the χ^2-test and the median final function values with the rank sum test.

Results The final distribution of function values for $n = 3, 5$, and 10 is shown in **Fig. 6**. For $n = 3$ the global optimum is found in most cases for all experimental conditions. On the axis parallel function $\mathcal{C}_n$ achieves a success probability of 100% and is slightly better than $\mathcal{C}_n^{\mathrm{iso}}$ and $\mathcal{G}_n$. For $n = 5$ the difference becomes much more pronounced. While the success probability drops to about five percent for $\mathcal{C}_n^{\mathrm{iso}}$ and $\mathcal{G}_n$, on the axis parallel function $\mathcal{C}_n$ has still a success probability of 100%. For $n = 10$ (Fig. 6, right) the success probability drops to zero in all cases but for $\mathcal{C}_n$ on the coordinate axis parallel function, where it is still one. The distributions in the five other cases are statistically

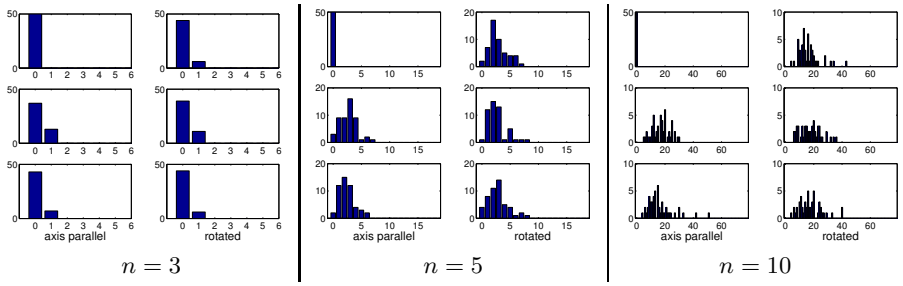

Fig. 6. Frequency of the final function value for, from above to below, $\mathcal{C}_n$, $\mathcal{C}_n^{\mathrm{iso}}$, and $\mathcal{G}_n$. Left: $n = 3$, Middle: $n = 5$, Right: $n = 10$. For $n = 3$ on the coordinate axis parallel function $\mathcal{C}_n$ has a significantly higher success probability than $\mathcal{C}_n^{\mathrm{iso}}$ ($p < 1.3 \times 10^{-4}$) and $\mathcal{G}_n$ ($p < 10^{-2}$). For $n = 5$ and $n = 10$ the difference regarding distribution median and success probability between $\mathcal{C}_n$ on the coordinate axis parallel function and all other cases is highly significant ($p < 10^{-15}$).

indistinguishable and the best final function value is close to four. In all dimensions all distributions perform virtually identical on the rotated function, and only $\mathcal{C}_n$ performs significantly different from the other distributions in the coordinate axis parallel case while the performance of $\mathcal{G}_n$ and $\mathcal{C}_n^{\mathrm{iso}}$ is invariant under rotation of the search space.

Validation of the Annealing Scheme To investigate the influence of the choices of θ_{start} and θ_{final} we ran simulations for all combinations of values $\theta_{\mathrm{start}} = 10^{10}, 10^{9}, \ldots, 10^{-5}$ and $\theta_{\mathrm{final}} = 10^{5}, 10^{4}, \ldots, 10^{-10}$ for $n = 5$, where $\theta_{\mathrm{start}} \geq \theta_{\mathrm{final}}$ and the number of iterations are $50\,000 \times 5$, choosing α respectively. The best result is obtained with $\theta_{\mathrm{start}} = 1, \theta_{\mathrm{final}} = 0.1$ for $\mathcal{C}_5$ and $\theta_{\mathrm{start}} = \theta_{\mathrm{final}} = 1$ for $\mathcal{G}_5$. The respective average final function values are 1.7 and 1.6, compared to 2.8, and 2.4 for the set-up chosen in the last section. The results confirm that the annealing schedule is reasonably chosen and does not dominate the outcome.

6 Summary and Conclusion

We analyzed densities of isotropic and anisotropic heavy-tail Cauchy distributions with respect to their effectiveness when employed in searching for optima in multimodal objective functions. The densities are determined to a great extent by the volume of the hypersphere surface area. Consequently, for *isotropic* search distributions the density (i.e. the probability to hit a given volume) must decrease faster than r^{-n}, where r is the distance to the distribution center.[4] For Gaussian distributions the density decreases exponentially fast with r, for the investigated isotropic Cauchy distribution the dependency is r^{-2n}. Even for moderate dimensions ($n = 5$ to 10), the relevance between polynomial and exponential decrease on the search performance becomes questionable and cannot be observed in our experiments.

In contrast, the effect of anisotropy of a search distribution on the search performance can be tremendous, in particular in higher dimensions ($n \geq 10$). The Cauchy

[4] Otherwise the density is not integrable for $r \to \infty$.

distribution, where coordinates are sampled independently, is highly anisotropic in that large steps occur most often close to the coordinate axes (see e.g. Fig. 3). Hence, it can perform exceptionally well on separable functions, like any algorithm performing coordinate-wise search. Therefore, the anisotropy of heavy-tail distributions is the most likely explanation for remarkable performance improvements on separable functions, e.g. of Fast Evolution Strategies [8] and of the so-called scale-free distribution [4]. If the coordinate system is rotated or the distribution is modified to become isotropic—keeping the distribution of the vector norm unchanged—the performance becomes indistinguishable from the Gaussian distribution in our experiments.

We believe that our result can be generalized beyond the specifically chosen set-up stating the following conjecture: *heavy tails are useful on multimodal objective functions (for global optimization) only if the large variations take place mainly in a low dimensional (sub-)space and the low dimensional space contains the better optima.* This is the case, for example, either if the search space by itself is low dimensional ($n \not\gg 3$), or if the search distribution is highly anisotropic with respect to the coordinate system and the objective function is separable. A challenging question arising from our conjecture is whether and how low dimensional subspaces can be found, such that the exceptional performance of the anisotropic Cauchy distribution on *separable* functions can be carried over to *non-separable* functions.

References

1. G. Frahm and M. Junker. Generalized elliptical distributions: Models and estimation. Caesar preprint, 2003. IDL 0037.
2. A. Obuchowicz. Multidimensional mutations in evolutionary algorithms based on real-valued representation. *International Journal of Systems Science*, 34(7):469–483, 2003.
3. Ingo Rechenberg. *Evolutionsstrategie '94*. Frommann-Holzboog, Stuttgart, Germany, 1994.
4. J.E. Rowe and D. Hidovic. An evolution strategy using a continuous version of the gray-code neighbourhood distribution. In K. Deb et al., editor, *Lecture Notes in Computer Science, proceedings of GECCO 2004*, volume 3102, pages 725–736. Springer-Verlag, 2004.
5. G. Rudolph. Local convergence rates of simple evolutionary algorithms with cauchy mutations. *IEEE Trans. Evolutionary Computation*, 1(4):249–258, 1997.
6. H. Szu and R. Hartley. Fast simulated annealing. *Phys. Lett. A*, 122(3,4):157–162, 1987.
7. X. Yao and Y. Liu. Evolutionary programming made faster. *IEEE Transactions on Evolutionary Computation*, 3:82–102, 1997.
8. X. Yao and Y. Liu. Fast evolution strategies. *Control and Cybernetics*, 26(3):467–496, 1997.