



Universiteit
Leiden
The Netherlands

Abnormal mutations: evolution strategies don't require Gaussianity

Nobel, J.P. de; Vermetten, D.L.; Wang, H.; Kononova, A.V.; Rudolph, G.; Bäck, T.H.W.; Ochoa, G.

Citation

Nobel, J. P. de, Vermetten, D. L., Wang, H., Kononova, A. V., Rudolph, G., & Bäck, T. H. W. (2025). Abnormal mutations: evolution strategies don't require Gaussianity. *Gecco '25*, 701-709. doi:10.1145/3712256.3726479

Version: Publisher's Version

License: [Creative Commons CC BY 4.0 license](#)

Downloaded from: <https://hdl.handle.net/1887/4294909>

Note: To cite this publication please use the final published version (if applicable).



PDF Download
3712256.3726479.pdf
18 February 2026
Total Citations: 0
Total Downloads: 823

Latest updates: <https://dl.acm.org/doi/10.1145/3712256.3726479>

RESEARCH-ARTICLE

Abnormal Mutations: Evolution Strategies Don't Require Gaussianity

JACOB DE NOBEL, Leiden University, Leiden, Zuid-Holland, Netherlands

DIEDERICK VERMETTEN, Leiden University, Leiden, Zuid-Holland, Netherlands

HAO WANG, Leiden University, Leiden, Zuid-Holland, Netherlands

ANNA V KONONOVA, Leiden University, Leiden, Zuid-Holland, Netherlands

GÜNTER RUDOLPH

THOMAS H W BÄCK, Leiden University, Leiden, Zuid-Holland, Netherlands

Open Access Support provided by:

Leiden University

Published: 14 July 2025

[Citation in BibTeX format](#)

GECCO '25: Genetic and Evolutionary
Computation Conference
July 14 - 18, 2025
Malaga, Spain

Conference Sponsors:
SIGEVO

Abnormal Mutations

Evolution Strategies Don't Require Gaussianity

Jacob de Nobel
Leiden Institute for Advanced
Computer Science
Leiden, The Netherlands
j.p.de.nobel@liacs.leidenuniv.nl

Diederick Vermetten
Leiden Institute for Advanced
Computer Science
Leiden, The Netherlands

Hao Wang
Leiden Institute for Advanced
Computer Science
Leiden, The Netherlands

Anna V. Kononova
Leiden Institute for Advanced
Computer Science
Leiden, The Netherlands

Thomas Bäck
Leiden Institute for Advanced
Computer Science
Leiden, The Netherlands

Günter Rudolph
TU Dortmund University
Department of Computer Science
Dortmund, Germany

ABSTRACT

The mutation process in evolution strategies has been interlinked with the normal distribution since its inception. Many lines of reasoning have been given for this strong dependency, ranging from maximum entropy arguments to the need for isotropy. However, some theoretical results suggest that other distributions might lead to similar local convergence properties. This paper empirically shows that a wide range of evolutionary strategies, from the $(1+1)$ -ES to CMA-ES, show comparable optimization performance when using a mutation distribution other than the standard Gaussian. Replacing it with, e.g., uniformly distributed mutations, does not deteriorate the performance of ES, when using the default adaptation mechanism for the strategy parameters. We observe that these results hold not only for the sphere model but also for a wider range of benchmark problems.

KEYWORDS

Evolution Strategies, Mutation distributions, Gaussianity, Benchmarking, CMA-ES

ACM Reference Format:

Jacob de Nobel, Diederick Vermetten, Hao Wang, Anna V. Kononova, Thomas Bäck, and Günter Rudolph. 2025. Abnormal Mutations: Evolution Strategies Don't Require Gaussianity. In *Genetic and Evolutionary Computation Conference (GECCO '25)*, July 14–18, 2025, Málaga, Spain. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3712256.3726479>

1 INTRODUCTION

Evolution strategies (ES) have traditionally relied on the normal distribution to sample mutation vectors for continuous search problems, which has been central to the algorithm since its first appearance [23, 28]. Several arguments exist for this choice of mutation distribution, ranging from biological analogies: “small mutations should be more likely than large mutations” [7] to more rigorous arguments referencing the *maximum entropy principle* [25]. Consequently, much of the developed theory is based on the standard

Gaussian (see e.g. [1, 2, 6]). A notable exception is the Cauchy distribution that has a super-Gaussian tail: $P(|X| > t) \in O(1/t)$. It has been proposed to increase the robustness of ES by allowing the sampling of rare large mutations [18]. The local convergence rates of the $(1+1)$ -ES and $(1,\lambda)$ -ES have been theoretically studied by [27] for the Cauchy distribution, which indicates a potential benefit for multimodal problems. This was confirmed empirically for the (μ, λ) -ES by [33], where it was shown that replacing the Gaussian with a Cauchy distribution improves the strategy on a set of multimodal benchmark functions. Other distributions, such as the simple uniform distribution, have not been studied empirically in ES. For several other distributions, including the logistic and Laplace distributions, the local convergence rates have been studied for simple evolution strategies [26]. There, it was found that all factorizing distributions that have their finite absolute moments defined up to order four offer an almost equally fast local convergence. Notably, these studies have only considered ES without recombination. Other modifications to the sampling distribution include the use of deterministic low-discrepancy sequences [8, 30] and mirrored sampling strategies [3, 32]. These works contributed to reducing the sampling errors from the standard Gaussian instead of replacing the mutation distribution completely. While these types of modifications show that changes to mutation distribution can yield improved performance, they don't fundamentally change any core properties of mutation in ES. According to Beyer [6], the mutation operator of an ES needs to fulfill the following four properties:

- (1) **Reachability**: Any point in the search space should be reachable by the mutation operator. Namely, there is a nonzero probability to hit any other point $\mathbf{x}' \in S$ starting from an arbitrary point $\mathbf{x} \in S$.
- (2) **Scalability**: The length of the mutation steps should be *tuneable* for a locally optimal mutation strength.
- (3) **Absence of biases**: The mutation distribution should be unbiased.
- (4) **Symmetry**: Requires the mutations to be isotropic around the origin.

From this, it follows naturally that the Gaussian distribution is favorable, as it conforms to all requirements and is the continuous distribution with the maximum entropy for a specified mean and variance [25]. As only the first requirement is strictly necessary for an ES to work, the question can be raised as to whether these



This work is licensed under a Creative Commons Attribution 4.0 International License. *GECCO '25*, July 14–18, 2025, Málaga, Spain
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1465-8/2025/07
<https://doi.org/10.1145/3712256.3726479>

requirements were conceived with the Gaussian distribution in mind. Moreover, the choice of Gaussian distribution has another advantage: it is a stable distribution [21]. This makes the design and analysis of algorithms more straightforward [13].

In this work, we aim to assess empirically whether the results from [26] also hold for several common continuous probability distributions in ES with recombination. In addition, we study the effects of changing the mutation distribution in the contemporary CMA-ES algorithm. In general, we are interested in measuring whether there is an observable benefit of using Gaussian mutations over other distributions. We analyze the $(1+1)$ -ES with different distributions on the sphere model and provide detailed benchmarking results for several types of ES on BBOB.

Note that all code, data, and comments on reproducing the results shown throughout this paper are available on our Zenodo repository [9].

2 PRELIMINARIES

2.1 Sampling in ES

Practically, in an ES with global intermediate recombination, which uses a multivariate Gaussian distribution, we sample in a three-stage process:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (1)$$

$$\mathbf{y} = \mathbf{A}\mathbf{z} \quad (2)$$

$$\mathbf{x} = \mathbf{m} + \sigma\mathbf{y} \quad (3)$$

Here $\mathbf{m}$ stands for the current mean of the search population, σ for the global step-size, and the matrix $\mathbf{A}$ (of full rank) for a linear transformation, which in the case of the multivariate Gaussian is the square root of the covariance $\mathbf{C}$ of the mutation distribution.

Practically, each component of the mutation vector $\mathbf{z}$ is generated independently. This can be done by the inverse transformation sampling: first we generate a number $u_i \sim \mathcal{U}(0, 1)$, and then use the percent-point function (PPF) of the Gaussian distribution $Q_{gauss}(p)$ to generate the required standard normal variable: $z_i = Q_{gauss}(u_i) \sim \mathcal{N}(0, 1)$. Conversely, using the PPF of another distribution would result in a random variable of that specific distribution. For example, replacing Q_{gauss} in the aforementioned example with $Q_{laplace}$ would result in a mutation vector $\mathbf{z}$, which has each component consisting of independent random variables that follow a Laplacian distribution.

2.2 Considered distributions

While not all random distributions are suitable for the mutation operator, several alternatives can still be considered. This work considers all distributions summarized in Table 1. These distributions have been known in the statistical literature for a long time, except the *double Weibull distribution* that has been introduced in [5] and applied for ES in [19] for the first time.

Figure 1 shows the probability density functions for each distribution. Note that all distributions have been shifted to be symmetrical around zero and parameterized such that their variances are one; see Table 1. The exception is the Cauchy distribution, which has no finite second moment. However, the scale of the distribution can be controlled via the parameter η , which we set to 1 to produce the standard Cauchy distribution. If we consider the probability

density functions (definitions provided in Table 1), we can observe that most distributions are unimodal. The double Weibull distribution is the only bimodal distribution, as for $\beta > 1$, its PDF has two distinct peaks. For the parameterization used here, $\beta = 2$, the peaks are at $-\sqrt{\frac{1}{2}}$ and $\sqrt{\frac{1}{2}}$. Additionally, note that the support for most distributions includes all $x \in \mathbb{R}$, except the uniform distribution, which only has support for $-\sqrt{3} \leq x \leq \sqrt{3}$.

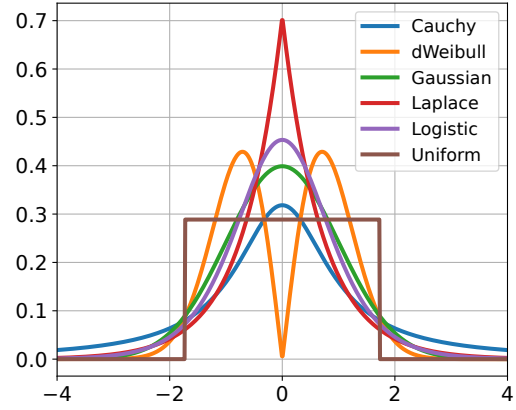


Figure 1: Probability density function for the Cauchy, double Weibull, Laplace, logistics, and uniform distributions.

2.2.1 Scalability. Any continuous distribution with finite mean and variance follows the scalability principle. While the Cauchy distribution does not satisfy this condition, it is still possible to control its location and scale via the scaling constant η .

2.2.2 Bias & Entropy. Specifically, the entropy H of a Gaussian distribution is $\frac{1}{2} \ln(2\pi e \sigma^2)$. Given a mean of 0 and variance of 1, we have $H \approx 1.42$. For the other distributions, this is given in Table 1. While the Gaussian distribution is indeed the maximum entropy distribution [24] for distributions with specified mean and variance, numerically, the differences between the considered distributions are relatively minor in the univariate case.

2.2.3 Symmetry. All of the considered distributions are symmetric around zero. Considering the multivariate case, only the Gaussian distribution (with identity covariance matrix) is strictly isotropic. To generate multivariate samples from these considered distributions, note that we are sampling each coordinate independently, resulting in non-spherical multivariate versions of each distribution [27].

3 ANALYZING DISTRIBUTIONS

To better understand the interaction between sampling distributions and the mutation process within ES, we investigate two core aspects of the mutation distribution: the effective step length and the isotropy.

3.1 Effective Step Length

As mentioned previously, all considered distributions follow the principle of scalability, ensuring that effective step length can be controlled. The effective step length is calculated using the L_2 -norm

Table 1: The definitions of the probability density functions (PDF), percent point functions (PPF), and the used parameterizations for each of the distributions are given. The entropy and the first four moments are also given, i.e., the mean, variance, skewness, and kurtosis. Note that $\gamma \approx 0.577$ denotes the Euler-Mascheroni constant.

Distribution	Cauchy	Double Weibull	Gaussian (Normal)	Laplace	logistic	uniform
Parameters	$x_0 = 0, \eta = 1$	$\alpha = 1, \beta = 2$	$\mu = 0, \sigma = 1$	$\mu = 0, b = 1/\sqrt{2}$	$\mu = 0, s = \sqrt{3}/\pi$	$a = -\sqrt{3}, b = \sqrt{3}$
PDF $f(x)$	$\frac{1}{\pi\eta} \frac{1}{1 + (\frac{x-x_0}{\eta})^2}$	$\frac{\beta}{2\alpha} x ^{\beta-1} \exp\left(-\left(\frac{ x }{\alpha}\right)^\beta\right)$	$\frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\frac{1}{2b} \exp\left(-\frac{ x-\mu }{b}\right)$	$\frac{\exp(-(x-\mu)/s)}{s[1+\exp(-(x-\mu)/s)]^2}$	$\begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$
PPF $Q(p)$	$x_0 + \eta \tan\left(\pi\left(p - \frac{1}{2}\right)\right)$	$\text{sign}(p-0.5) \alpha [-\ln(2 p-0.5)]^{1/\beta}$	$\mu + \sigma \sqrt{2} \text{erf}^{-1}(2p-1)$	$\mu + b \text{sign}(p-0.5) \ln\left(\frac{1}{2 p-0.5 }\right)$	$\mu + s \ln\left(\frac{p}{1-p}\right)$	$a + (b-a)p$
Moments	$\infty, \infty, \infty, \infty$	$0, \alpha^2 \Gamma(1 + \frac{2}{\beta}), 0, \frac{\Gamma(1+4/\beta)}{\Gamma(1+2/\beta)^2} - 3$	$\mu, \sigma^2, 0, 0$	$\mu, 2b^2, 0, 3$	$\mu, \frac{s^2 \pi^2}{3}, 0, \frac{6}{5}$	$\frac{a+b}{2}, \frac{(b-a)^2}{12}, 0, \frac{-6}{5}$
Entropy H	$\ln(4\pi\eta) \approx 2.531$	$-\gamma/\beta - \ln(\beta) + \gamma + 1 - \ln(\frac{1}{2}) \approx 1.289$	$\frac{1}{2} \ln(2\pi e \sigma^2) \approx 1.417$	$\ln(2b) + 1 \approx 1.347$	$\ln s + 2 \approx 1.405$	$\ln(b-a) \approx 1.242$

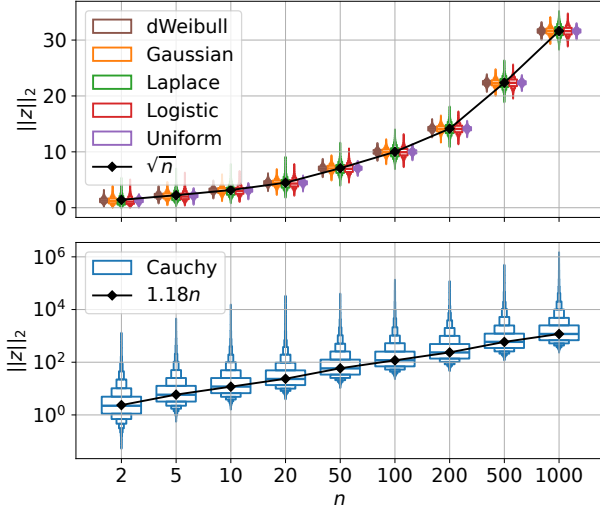


Figure 2: Effective step length L_2 -norm for each sampler type, parameterized according to Table 1, for increasing dimensionalities n . The distributions for which $\|z\|_2$ scales proportionally to $\sqrt{n}$ are shown in the top figure; Cauchy is shown separately. Note the log-scaling of the y-axis for the bottom figure.

and is denoted by $\|z\|_2$. It measures the size of the mutation and is an important quantity in ES, as it is used to parameterize *self-adaptation*. Since we propose replacing the mutation distribution in this work with something other than a standard Gaussian, we must ensure we can still correctly parameterize the algorithm. Considering the differences between probability density functions, one might expect similar differences in each distribution’s effective step length distributions.

For the standard Gaussian distribution, the expected value of $\|z\|_2$ scales proportionally to $\sqrt{n}$, and the variance remains constant with n , as illustrated in Figure 2. As can be seen from the figure, this scaling with dimensionality n holds for the other considered distributions, except for the Cauchy distribution. However, while $\frac{E\|z\|_2}{\sqrt{n}}$ converges to 1 as $n \rightarrow \infty$, we note that for smaller values of n , i.e. $n < 20$, $\sqrt{n}$ is a slight overestimation for $E\|z\|_2$ for all but the uniform distribution. Specifically, for the standard normal distribution, we know that $E\|z\|_2$ follows the square root of a χ^2 distribution with n -degrees of freedom [12], which is $\sqrt{2} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \leq \sqrt{n}$. Since

the expected value and variance for the Cauchy distribution remain undefined, we interpret its median and inter-quartile ranges (IQR) of $\|z\|_2$ in relation to n . This is made visible in the bottom panel of Figure 2, where we can observe that the median scales proportionally with $\approx 1.18n$. Additionally, we note that the IQR range of $\|z\|_2$ is not independent of dimensionality for the Cauchy distribution since this also increases linearly with n .

Finally, we can see a clear ranking in the sample variance of $\|z\|_2$ for each distribution in the top panel of Figure 2. We observe that the uniform and the double Weibull distributions are more condensed than the standard normal distribution. In contrast, the sample variance of $\|z\|_2$ for the logistic and Laplacian distributions is higher.

3.2 Angle Isotropy

Apart from effective step sizes, isotropy is an often-discussed property of mutation distributions [22, 27]. In the context of a probability distribution, isotropy refers to the property that a distribution exhibits the same statistical behavior in all directions. Specifically, it details whether the distribution is invariant under rotations and distance-preserving transformations. It consequently expresses equal variance in all directions, which can be geometrically interpreted as the distribution having spherical iso-contour lines in a multivariate density plot. Theoretically, isotropic distributions are more straightforward to model [13] since they can be naturally generalized toward multiple dimensions. From the distributions considered here, we know that only the standard Gaussian, with an identity covariance matrix $C = I$, satisfies this property.

To gain insight into the degree to which the other distributions are non-isotropic, we visualize the angle θ of a large set of samples drawn from a given distribution and a vector of all ones, $\mathbf{1}^n$, in two dimensions (see Figure 3). Since all distributions are symmetric, we observe a recurring pattern in each quadrant. Indeed, only the Gaussian has a uniform angle distribution, as the probability of sampling a vector with a given direction is equally likely for all directions. The other distributions all show some degree of anisotropy. For example, the rectangular shape of the uniform distribution makes it more likely to sample vectors that are aligned with or perpendicular to the $\mathbf{1}^n$ vector. Contrastingly, the Laplacian, logistic, and Cauchy distributions all have a higher probability of sampling vectors parallel to the axis. This is especially true for the Cauchy distribution, as its infinite variance makes it very likely to sample vectors parallel to the coordinate system. Notably, the angle distribution of the double Weibull has the same period as the

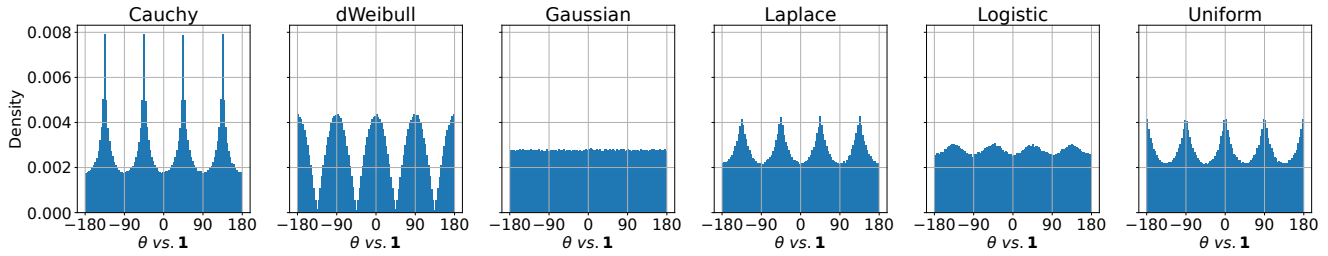


Figure 3: Normalized angle distribution of 10^5 sampled points in dimensionality $n = 2$ versus a vector of all ones, i.e., 1^n , for each probability distribution, parameterized according to Table 1.

uniform distribution, but the likelihood of sampling axis-parallel vectors is almost zero.

4 LOCAL CONVERGENCE OF THE $(1+1)$ -ES

Our experiments start with the $(1+1)$ -ES with a 1/5th success rule. This algorithm has been well-studied, yielding some of the earlier proofs for ES in continuous domains [17]. Specifically, the running time of this algorithm on the sphere function, i.e., $f(\mathbf{x}) = \mathbf{x}'\mathbf{x}$, has been analyzed extensively [4, 11] as a model of local convergence. We use the $(1+1)$ -ES described in [13], with pseudocode provided in Algorithm 1. Note our addition of the $Q(\mathbf{p})$ parameter, which maps a vector $\mathbf{p} \in [0, 1]^n$ to a sample of a given probability distribution, using the parameterization and percent point functions from Table 1 (see Section 2.1). This allows us to change the algorithm to use a different mutation distribution while keeping the rest unchanged. In this sense, we can use the self-adaptation rules as intended for the normal distribution with any other distribution we choose. This modification is highlighted in the pseudocode.

Algorithm 1 $(1+1)$ -ES with 1/5th success rule

Require: Initial step size $\sigma_0 \in \mathbb{R}$, initial point $\mathbf{x}_0 \in \mathbb{R}^n$, $n \in \mathbb{N}_+$, PPF $Q(\mathbf{p}): [0, 1]^n \rightarrow \mathbb{R}^n$

```

1: procedure  $(1+1)$ -ES
2:    $d \leftarrow \sqrt{n+1}$ 
3:    $\mathbf{m} \leftarrow \mathbf{x}_0$ 
4:   repeat
5:      $\mathbf{u} \sim \mathcal{U}^n(0, 1)$ 
6:      $\mathbf{z} \leftarrow Q(\mathbf{u})$ 
7:      $\mathbf{x} \leftarrow \mathbf{m} + \sigma \mathbf{z}$ 
8:      $\sigma \leftarrow \sigma \cdot \exp^{1/d}(\mathbb{1}_{(f(\mathbf{x}) \leq f(\mathbf{m}))} - 1/5)$ 
9:     if  $f(\mathbf{x}) \leq f(\mathbf{m})$  then
10:        $\mathbf{m} \leftarrow \mathbf{x}$ 
11:   until convergence

```

Mutation rate. We analyze the applicability of the 1/5th success rule for different mutation distributions in Figure 4. We show the evolution of the mutation rate averaged over 1000 runs on the sphere model for different dimensionalities. From the figure, it can be seen that the adaptation of σ is unaffected by the choice of mutation distribution. The exception is the Cauchy distribution,

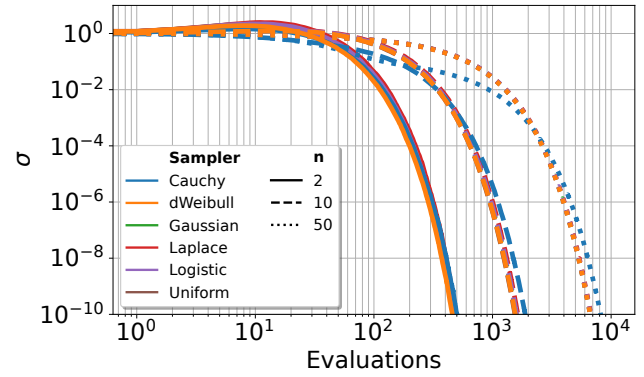


Figure 4: Evolution of the mutation rate σ of the $(1+1)$ -ES with 1/5th success rule on the sphere model $f(\mathbf{x}) = \mathbf{x}'\mathbf{x}$, averaged over 1000 runs, for dimensionalities $n \in \{2, 10, 50\}$ for different mutation distributions.

which causes σ to be adapted more slowly, especially for higher dimensionalities. Nevertheless, these differences are relatively minor, indicating that the local convergence speed, measured on the sphere model, is similar.

Expected Running Time. In this experiment, we use the definition of the sphere function from BBOB [14]. This means that the optimum is distributed uniformly at random in $[-4, 4]^n$ (with the recommended domain being $[-5, 5]^n$). For each mutation distribution, we perform one run on 1000 instances of the sphere problem for dimensionalities $n \in \{2, 10, 50\}$ and initialize the $(1+1)$ -ES in the center of the domain for each run. The left-most panel in Figure 5 shows the distribution of hitting times for target precision 10^{-8} . The distribution of hitting times shows a similar figure as the evolution of the mutation rate. Again, we observe only minor relative differences between the mutation distributions, with only the Cauchy distribution having notably higher hitting times, which becomes more pronounced with increasing n . However, we must point out that while the difference between Cauchy and the other distributions is noticeable, the absolute difference in hitting times is still relatively small.

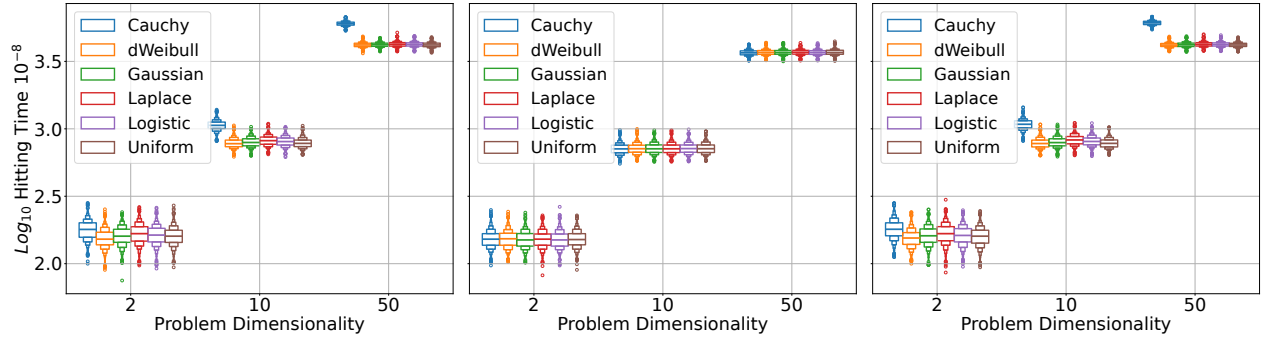


Figure 5: Hitting times of target precision 10^{-8} for $(1+1)$ -ES with a 1/5-success rule on sphere model, for different sampling distributions to determine step-size. Left: using the standard distribution. Center: Normalized mutation vectors to isolate the effect of isotropy. Right: using sphered versions of the distributions to isolate the effects of effective step size. Distributions are all over 1000 instances of the sphere model.

4.1 Isolating Effects

While we observe only minor differences in the performance of the $(1+1)$ -ES when changing mutation distributions, we would like to identify what causes these differences. Specifically, we would like to investigate whether this is caused by the differences in the angle distributions of each mutation distribution (isotropy) or the differences in the effective step size $\|z\|_2$ or a combination thereof. For this purpose, we run two experiments to isolate their respective effects on the hitting time of the $(1+1)$ -ES, using the same setup as before, using 1000 instances of the sphere model in BBOB for a target value 10^{-8} .

Isotropy. First, we study the effect of the directionality of the sampled mutation vectors in isolation. This can be achieved by normalizing each z -vector to a unit vector ($z/\|z\|_2$), which makes the effective step size $\|z\|_2$ of every sample identical. In this setting, the only differences we get between distributions are the directions of our mutations, and all sampled mutation vectors are located on the unit sphere. The results of this experiment are shown in the middle panel of Figure 5. From this figure, we can see that all the differences in hitting time disappear. Even the distributions with a very concentrated angle distribution, such as the double Weibull or Cauchy, perform identically in this situation.

Effective step size. Based on the previous experiment, we might conclude that the minor differences in hitting time are only due to differences in the effective size of the mutations. To validate this, we perform another experiment to ensure that our mutations are isotropic, i.e., the direction is drawn from a normal distribution, but each respective distribution controls the mutation scale. Specifically, we modify the sampling by first drawing a random vector $v \sim \mathcal{N}(0, I)$, which we then normalize and scale by the size of the original mutation vector z , i.e., $v/\|v\|_2 \cdot \|z\|_2$. This effectively samples a uniform direction for the mutations while the distribution of effective step sizes matches the target distribution. Hitting times for this scenario are shown on the right panel of Figure 5. In this figure, we see that the ordering between the distributions as they were in the leftmost panel is retrieved. This seems to indicate that any differences between sampling distributions for the $(1+1)$ -ES on a

sphere model result from the changes to the scale of the mutations, i.e., $\|z\|_2$, rather than due to the effect of isotropy.

5 BENCHMARKING MULTIPLE ES VARIANTS ON BBOB

As the sphere model shows relatively small differences in the performance of $(1+1)$ -ES based on the sampling distribution, we now extend our experimental setup to a broader range of optimization problems, performing a complete benchmarking study on BBOB. Additionally, we investigate the impact of the used sampling distribution within more complex evolution strategies. We use a multi-membered self-adaptive evolution strategy, as defined in Algorithm 2 in [13], and investigate the effect on both the standard and elitist versions of the *CMA-ES* algorithm [16]. In summary, we collect benchmarking data for the following algorithms:

- $(1+1)$ -ES, with 1/5th success rule, as described in the previous section.
- $(\mu/\mu, \lambda)$ - σ SA-ES: A population-based ES with self-adaptive step sizes and global recombination.
- $(\mu/\mu_w, \lambda)$ -CMA-ES: Canonical version of the *CMA-ES*, as introduced in [16], without any restart mechanisms.
- $(\mu/\mu_w + \lambda)$ -CMA-ES: Elitist version of the *CMA-ES*, where both the parent and offspring populations are considered for selection.

For each of these algorithms, the sampling procedure is modified in the same fashion as for the $(1+1)$ -ES in the previous section. The complete pseudocode for each algorithm can be found in the supplementary material (provided on Zenodo [9]), from which it can be seen that each algorithm accepts a parameterized PPF such that it can be modified to use a selected mutation distribution (see Section 2.1).

Path length normalization. For the *CMA-ES*, the assumption that mutations are drawn from a standard normal distribution is used directly in the parameter update. Namely, in the cumulative step-size adaptation (CSA) procedure, the expected effective step size of the standard Gaussian, i.e., $E\|\mathcal{N}(0, I)\|_2$, is used to normalize the

evolution path:

$$\sigma = \sigma \exp \left(\frac{c_\sigma}{d_\sigma} \left(\frac{\|\mathbf{p}_\sigma\|_2}{\mathbb{E}[\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|_2]} - 1 \right) \right) \quad (4)$$

As mentioned in Section 3.1, this expectation converges towards $\sqrt{n}$ for increasing dimensionalities $n \in \mathbb{N}$, and can be more precisely estimated by $\sqrt{2} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)}$ for the Gaussian distribution. When changing sampling distributions, we have to normalize the evolution path with a value suitable for the modified distribution. However, as seen in Figure 2, $\sqrt{n}$ is a reliable estimate for this value for all but the Cauchy distribution. For the Cauchy distribution, we set this normalization constant (denoted by ρ in the supplementary material) to $1.18n$, proportional to the median $\|z\|_2$ as seen in the bottom panel of Figure 2.

Covariance Matrix Adaptation. The CMA-ES is formulated as a variable metric approach that adapts the parameters of a multivariate normal distribution $\mathcal{N}(\mathbf{m}, \sigma\mathbf{C})$ such that the likelihood of successful mutations is increased [16]. Learning the covariance matrix $\mathbf{C}$, specifically, allows the method to capture correlations between object variables and, therefore, to become invariant against arbitrary rotations of the search space. However, this can also be viewed as learning an appropriate scaling σ and rotation $\mathbf{A}$ of the mutation distribution [29]. For the Gaussian, this changes the shape of the distribution from an isotropic sphere into an arbitrarily scaled hyper-ellipsoid. While this seems specific to the Gaussian, we can, in fact, use the same method with any other distribution, aiming to learn a proper rotation and scaling. Figure 6 shows an example of this, optimizing the sphere model with both a Gaussian (left) and uniform (right) distribution.

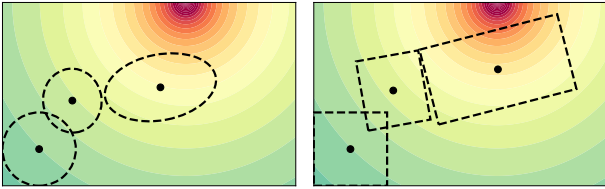


Figure 6: Isocontour lines for the mutation distribution when optimizing the sphere model $f(\mathbf{x}) = \mathbf{x}^T \mathbf{x}$ in $n = 2$ dimensions. Three consecutive generations are shown. The left figure uses a standard Gaussian mutation distribution, and the right figure uses a uniform mutation distribution.

Experimental Setup. For our benchmark problems, we use the well-known BBOB suite of 24 noiseless, single-objective problems [14], implemented in IOHexperimenter [10]. Even though the original problems are unconstrained, we add a bound violation penalty for solution vectors that exceed the suggested domain of $[-5, 5]^n$. We set this penalty to $v \cdot 10^{20}$, where v denotes the amount of boundary violation, measured by the Euclidean distance to the closest bound. This is done primarily to disallow the algorithms from providing better-than-intended solutions by sampling outside the domain. For example, this can happen for the Linear Slope function f_5 , potentially providing an unfair advantage to the Cauchy-based mutations. We use the standard problem dimensionalities of

$\{2, 3, 5, 10, 20, 40\}$ and perform one run on each of the first 100 problem instances. As mentioned in Section 4, each instance provides a new random global optimum location, allowing us to initialize the algorithms in the center of the search domain. For both variants of the CMA-ES and the $(1+1)$ -ES, we set the initial step size σ_0 to 2, and we set each element of σ_0 to $10^{\frac{1}{4}}$ for the $(\mu/\mu, \lambda)$ - σ SA-ES.

Empirical Attainment. To aggregate performance across functions, we use Empirical Cumulative Distribution Functions (ECDF). Instead of the target-based version, we use ECDFs based on the Empirical Attainment Function (EAF), which corresponds to the ECDF with infinite targets between the chosen bounds [20]. For the EAF, we set the upper and lower bounds on the precision to 10^8 and 10^{-8} , respectively.

On the left side of Figure 7, we show the EAF-based ECDF for every combination of ES variant and sampling distribution, aggregated over all 24 BBOB functions in dimensionality 10. The area under this curve for each line shown is given in the right panel of Figure 7, providing a summarizing view of the data. From these figures, we can see that each algorithm forms a group, with all the sampling distributions performing roughly similar to that of the other distributions for that algorithm. This is especially true for the $(1+1)$ -ES, where all sampling distributions show almost identical empirical performance. Note again that this figure shows an aggregated view over 24 different functions. This means that while the Cauchy distribution was observably worse for the sphere model, on average, over a broader benchmark, it is not. While it can be seen from the figure that the Cauchy distribution for the $(1+1)$ -ES is slower to converge, it reaches more targets than any of the other distributions, which results in the AUC being slightly higher. This is primarily because on multimodal functions, such as the Rastrigin-based functions f_3 and f_4 , the large mutations incurred by the Cauchy distribution allow the $(1+1)$ -ES to escape local optima more often. For the other distributions, differences in performance are negligible, as indicated by the completely overlapping EAFs for the $(1+1)$ -ES. This is similarly true for the $(\mu/\mu, \lambda)$ - σ SA-ES, where all distributions are almost indistinguishable from each other, except for Cauchy, which shows to be very much hampering performance in this algorithm. We expect this is due to the combination of non-elitism (i.e., comma selection) and the global recombination operator. Naturally, the potential of incorporating mutations with infinite variance into the update $\mathbf{m}$ has the chance of moving $\mathbf{m}$ detrimentally far. The elitist $(1+1)$ -ES does not suffer from this problem since the $+$ strategy would never select such mutations. We can observe something similar when comparing the $(\mu/\mu_w, \lambda)$ -CMA-ES with the $(\mu/\mu_w + \lambda)$ -CMA-ES. While the average performance of both algorithms is decreased by using the Cauchy distribution, this seems to affect the $(\mu/\mu_w, \lambda)$ -CMA-ES much more greatly than the $(\mu/\mu_w + \lambda)$ -CMA-ES. For the $(\mu/\mu_w, \lambda)$ -CMA-ES, the Gaussian distribution shows the highest empirical performance. Note that the AUC for the Gaussian distribution is only better by a tiny margin ($\approx 10^{-3}$) from the uniform distribution. For the $(\mu/\mu_w + \lambda)$ -CMA-ES, we can observe that the double Weibull distribution actually attains the highest AUC. More figures for different dimensionalities and individual algorithms are provided in the supplementary material on Zenodo [9]. From those figures, we can observe that while the general trend remains that Cauchy-based mutations are

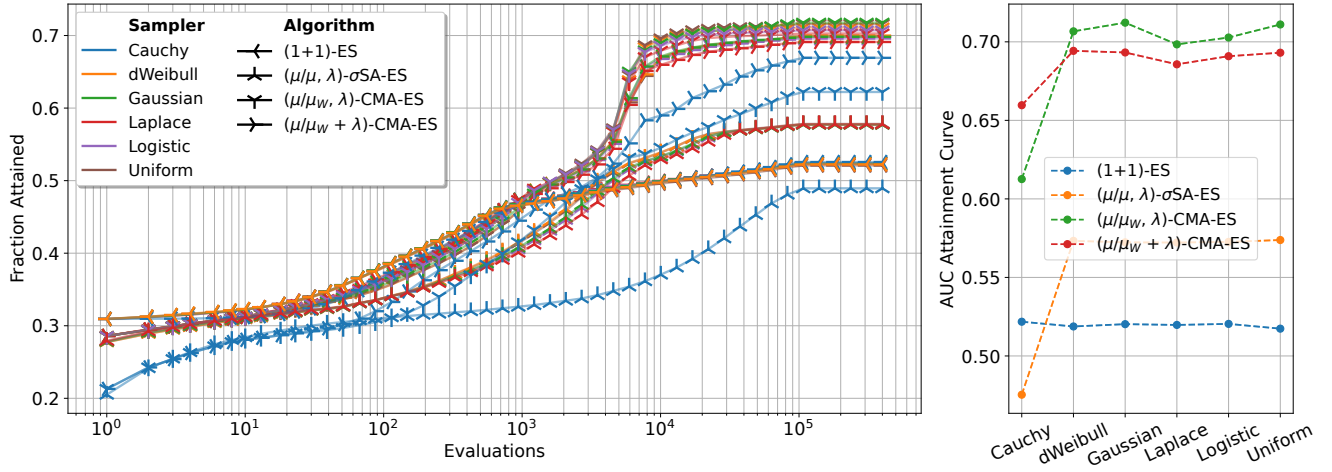


Figure 7: The left panel shows the EAF-based ECDF (bounds 10^8 and 10^{-8}) for the $(1+1)$ -ES, the $(\mu/\mu, \lambda)$ - σ SA-ES, the $(\mu/\mu_w, \lambda)$ -CMA-ES and the $(\mu/\mu_w + \lambda)$ -CMA-ES with different sampling methods. Aggregated over 100 instances of all 24 BBOB problems in dimensionality $d = 10$. Note that results and figures for each individual function and dimensionality can be found in our Zenodo [9] repository.

detrimental to performance, there are cases where this distribution provides a considerable speedup. These include (low-dimensional) multi-modal functions, such as f_3 , or functions with neutrality (f_7). Similarly, the individual differences between the other distributions are minor.

Function groups. The aggregate results presented in the previous paragraph show remarkably few differences between the sampling distributions for the tested ES. More can be seen when looking at the BBOB function groups individually. The BBOB functions can be categorized into five functional groups:

- (1) Separable functions
- (2) Unimodal functions with low or moderate conditioning
- (3) Unimodal functions with high conditioning
- (4) Multimodal functions with adequate global structure
- (5) Multimodal functions without adequate global structure

Figure 8 shows the same area under the EAF-based ECDF as was shown in the right side of Figure 7 for each function group specifically. Notably, it can be seen that while the CMA-ES algorithms are the best overall, for the separable functions (group 1), they are outperformed by the $(\mu/\mu, \lambda)$ - σ SA-ES. Interestingly, for the unimodal function group two, the $(1+1)$ -ES outperforms the $(\mu/\mu, \lambda)$ - σ SA-ES with all different sampling strategies, while the $(\mu/\mu, \lambda)$ - σ SA-ES performs better than $(1+1)$ -ES aggregated across the entire benchmark. It can also be observed here that there are several groups for which using a Cauchy distribution is strictly better than any of the other distributions for an algorithm. This is the case for the $(1+1)$ -ES on function group one and for both CMA-ES variants for function group five. In fact, for the $(\mu/\mu_w + \lambda)$ -CMA-ES, using Cauchy-based mutations results in $\sim 24\%$ higher AUC for function group five. For the other distributions, the differences in AUC are again less pronounced. The double Weibull and uniform distributions seem to be a bit better for $(\mu/\mu, \lambda)$ - σ SA-ES on function group one, and

the Laplace distribution is slightly worse than the others for the CMA-ES variants on function groups one and two.

6 DISCUSSION

In this paper, we have compared six continuous probability distributions to independently sample the components of the mutation vectors in different types of ES. We have analyzed the classical $(1+1)$ -ES in detail on the sphere model and found that any differences between the tested distributions are most likely due to variances in the effective step size $\|z\|_2$. Symmetry along each axis, on the other hand, remains a requirement. Naturally, a distribution with its center of mass not centralized at zero would lead to a severely biased mutation operator. The fact that different distributions lead to differences in the sampled directions (angles) of the mutations seems to have little impact.

Regarding local convergence, only the Cauchy distribution noticeably slows down the $(1+1)$ -ES. For the other distributions, differences in performance appear to be minimal, which matches the results from [26] as these all have defined moments up to order 4. This result translates to more complex benchmarks and ES. For all but the Cauchy distribution, the choice of sampling distribution has little impact on the empirical performance. In fact, from a practical viewpoint, we observed that all tested ES are remarkably stable w.r.t. the selected sampling distribution and that there seems to be no particular performance benefit to using the standard Gaussian. Even untypical distributions, such as the bimodal double Weibull distribution, perform equally well to the standard Gaussian distribution on BBOB. With all else being equal, we would like to point out that from a computational perspective, it is considerably faster to generate uniform random numbers than to generate normally distributed ones (see Figure 9). Care should be taken when using mutation vectors taken from the Cauchy distribution. As was already hinted at by [27], local convergence for the $(1+1)$ -ES is slower

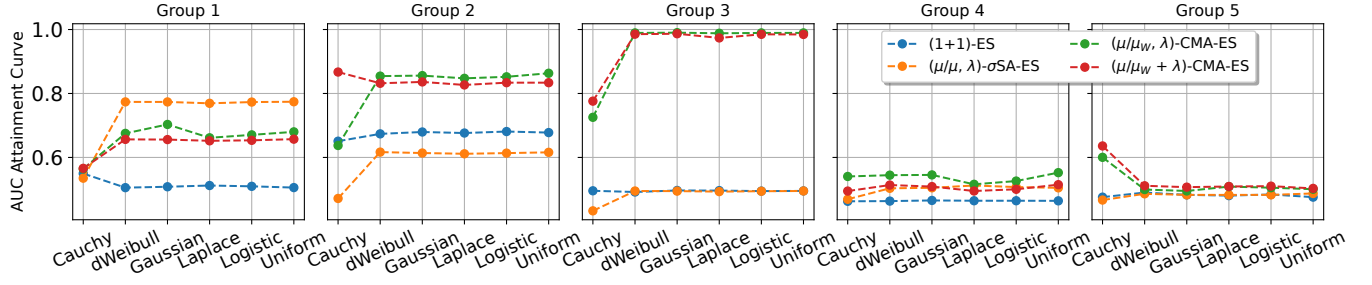


Figure 8: Area under the EAF-based ECDF (bounds 10^8 and 10^{-8}) for the $(1+1)$ -ES, the $(\mu/\mu, \lambda)$ - σ SA-ES, the $(\mu/\mu_w, \lambda)$ -CMA-ES and the $(\mu/\mu_w + \lambda)$ -CMA-ES with different sampling methods. Aggregated over 100 instances in dimensionality $d = 10$, grouped by the five BBOB function groups. Results for other dimensionalities can be found in our Zenodo repository [9].

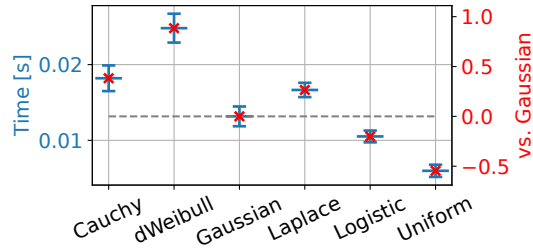


Figure 9: Timing comparison of generating 10^6 random samples using a given distribution, using SciPy [31]. The left axis shows the mean time over 10^4 trials in seconds, with one std. dev. indicated by the bars. The right axis shows the ratio of time saved vs. a Gaussian distribution.

when using this distribution. For this algorithm, we only observe benefits to using Cauchy mutations on *separable* multi-modal functions. This was similarly observed by [15]. However, it also does not deteriorate the algorithm on other function groups. Since the second moment of the Cauchy distribution is not finite, and the median of $\|z\|_2$ scales with n rather than $\sqrt{n}$, proper care should be taken when using parameters designed for ES with Gaussian mutations. Additionally, we note that recombination in combination with a non-elitist selection strategy can lead to problematic behavior when using the Cauchy distribution. Even though the local convergence using Cauchy mutations is considerably slower for all tested ES, we find cases where this distribution is preferable over the default mutation distribution. These include highly multimodal functions or functions with neutrality, where the large mutations incurred by the Cauchy distribution can help avoid stagnation (this can be observed on e.g. $f7$ and $f21$, figures available in [9]). This aligns with the findings from fast ES [33]. In fact, we observe considerable improvements for the $(\mu/\mu_w + \lambda)$ -CMA-ES on multimodal functions when using Cauchy-based mutations, even for non-separable problems (group 5). In addition to a CPU time argument, another benefit of using non-Gaussian mutations could be initialization and warm-starting. Currently, when using the Gaussian mutations in a box-constrained context, the hypersphere of the standard Gaussian cannot adequately cover the space. It must either

be configured in a sphere-in-a-box manner or a box-in-a-sphere manner. The uniform distribution, however, can be configured to match a box-constrained domain perfectly. Consequently, this could make constraint handling more manageable when using bounded uniform mutations. However, if we have an unbounded space with only a known starting point, having an isotropic distribution might be preferable [22] to properly explore around that point.

7 CONCLUSIONS & FUTURE WORK

We have shown that the Gaussianity of the mutation distribution, which has been central to Evolution Strategies (ES) since their inception [23, 28], is not a strong prerequisite. As long as the mutation distribution is properly scalable and symmetrical within each dimension/axis, the differences in empirical performance between mutation distributions are marginal. These results allow us to be confident that ES perform well with distributions that are not necessarily maximum entropy or completely isotropic. This opens the door to further experimentation with non-Gaussian mutations, as our results indicate that this is not a requirement to a functional ES. Even using a Cauchy distribution, with its infinite variance, can be a useful mutation distribution to prevent premature convergence in multimodal problems, although this comes with a tradeoff of worse performance on other function groups, especially when using a non-elitist strategy with recombination. In future work, we could integrate our findings in a dynamic switching context, adaptively changing the mutation distribution to enforce exploration. While we observed competitive performance using the parameters and learning rate constants as intended for ES with Gaussian mutation, we might want to explore more specific parameter settings for each distribution in future work. Here, only the path length normalization constant was considered for the CMA-ES, but several other parameters might be optimized further to improve empirical performance. This is similarly true for the $(\mu/\mu, \lambda)$ - σ SA-ES, where the learning rate parameters τ and τ_i might be tweaked for each distribution. Another research direction could be low discrepancy sequences [30]. Since these sequences are optimized to be evenly spread within an n -dimensional hypercube, using such points with a uniform distribution might be preferable over the Gaussian transformation. This might be especially useful when only a few such points are used deterministically [8].

REFERENCES

- [1] Alexandru Agapie, Mircea Agapie, Günter Rudolph, and Gheorghita Zbaganu. 2013. Convergence of Evolutionary Algorithms on the n -Dimensional Continuous Space. *IEEE Transactions on Cybernetics* 43, 5 (2013), 1462–1472.
- [2] Dirk V. Arnold. 2002. *Noisy optimization with evolution strategies*. Kluwer.
- [3] Anne Auger, Dimo Brockhoff, and Nikolaus Hansen. 2011. Mirrored sampling in evolution strategies with weighted recombination. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation* (Dublin, Ireland) (GECCO '11). Association for Computing Machinery, New York, NY, USA, 861–868. <https://doi.org/10.1145/2001576.2001694>
- [4] Anne Auger and Nikolaus Hansen. 2013. Linear convergence on positively homogeneous functions of a comparison based step-size adaptive randomized search: the $(1+1)$ ES with generalized one-fifth success rule. *arXiv preprint arXiv:1310.8397* (2013).
- [5] N. Balakrishnan and S. Kocherlakota. 1985. On the Double Weibull Distribution: Order Statistics and Estimation. *Sankhyā: The Indian Journal of Statistics, Series B* 47, 2 (1985), 161–178.
- [6] Hans-Georg Beyer. 2001. *The theory of evolution strategies*. Springer.
- [7] Hans-Georg Beyer. 2023. What You Always Wanted to Know About Evolution Strategies, But Never Dared to Ask. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation* (Lisbon, Portugal) (GECCO '23 Companion). Association for Computing Machinery, New York, NY, USA, 878–894. <https://doi.org/10.1145/3583133.3595041>
- [8] Jacob de Nobel, Diederick Vermetten, Thomas Bäck, and Anna Kononova. 2024. Sampling in CMA-ES: Low Numbers of Low Discrepancy Points. In *Proceedings of the 16th International Joint Conference on Computational Intelligence - ECTA. INSTICC, SciTePress*, 120–126. <https://doi.org/10.5220/0013000900003837>
- [9] J. de Nobel, D. Vermetten, H. Wang, A.V. Kononova, T. Baack, and G. Rudolph. 2025. Abnormal Mutations - Reproducibility Files and Additional Figures. (Jan 2025). <https://doi.org/10.5281/zenodo.15189142>
- [10] Jacob de Nobel, Furong Ye, Diederick Vermetten, Hao Wang, Carola Doerr, and Thomas Bäck. 2024. Iohexperimenter: Benchmarking platform for iterative optimization heuristics. *Evolutionary Computation* (2024), 1–6.
- [11] Tobias Glasmachers. 2017. Global Convergence of the $(1+1)$ Evolution Strategy. *arXiv preprint arXiv:1706.02887* (2017).
- [12] Nikolaus Hansen. 2023. The CMA Evolution Strategy: A Tutorial. *arXiv:1604.00772 [cs.LG]* <https://arxiv.org/abs/1604.00772>
- [13] Nikolaus Hansen, Dirk V. Arnold, and Anne Auger. 2015. *Evolution Strategies*. Springer Berlin Heidelberg, Berlin, Heidelberg, 871–898. https://doi.org/10.1007/978-3-662-43505-2_44
- [14] Nikolaus Hansen, Steffen Finck, Raymond Ros, and Anne Auger. 2009. *Real-Parameter Black-Box Optimization Benchmarking 2009: Noiseless Functions Definitions*. Research Report RR-6829. INRIA. <https://hal.inria.fr/inria-00362633/document>
- [15] Nikolaus Hansen, Fabian Gemperle, Anne Auger, and Petros Koumoutsakos. 2006. When do heavy-tail distributions help?. In *International Conference on Parallel Problem Solving from Nature*. Springer, 62–71.
- [16] Nikolaus Hansen and Andreas Ostermeier. 2001. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation* 9, 2 (2001), 159–195.
- [17] Jens Jägersküpper. 2003. Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces. In *International Colloquium on Automata, Languages, and Programming*. Springer, 1068–1079.
- [18] Cornelia Kappler. 1996. Are Evolutionary Algorithms Improved by Large Mutations?. In *Parallel Problem Solving from Nature*. <https://api.semanticscholar.org/CorpusID:11946748>
- [19] Frank Kursawe. 1990. A Variant of Evolution Strategies for Vector Optimization. In *Parallel Problem Solving from Nature, 1st Workshop, PPSN I, Dortmund, Germany, October 1-3, 1990, Proceedings*, Hans-Paul Schwefel and Reinhard Männer (Eds.). Springer, 193–197.
- [20] Manuel López-Ibáñez, Diederick Vermetten, Johann Dréo, and Carola Doerr. 2024. Using the Empirical Attainment Function for Analyzing Single-objective Black-box Optimization Algorithms. *CoRR* abs/2404.02031 (2024). <https://doi.org/10.48550/ARXIV.2404.02031> *arXiv:2404.02031*
- [21] John P. Nolan. 2020. *Univariate Stable Distributions*. Springer.
- [22] Andrzej Obuchowicz. 2019. *Stable Mutations for Evolutionary Algorithms*. Studies in Computational Intelligence, Vol. 797. Springer. <https://doi.org/10.1007/978-3-030-01548-0>
- [23] Ingo Rechenberg. 1965. Cybernetic solution path of an experimental problem. *Royal Aircraft Establishment Library Translation* 1122 (1965).
- [24] RD Rosenkrantz. 1989. Where do we stand on maximum entropy?(1978). In *ET Jaynes: Papers on Probability, Statistics and Statistical Physics*. Springer, 210–314.
- [25] Günter Rudolph. 1994. An evolutionary algorithm for integer programming. In *International Conference on Parallel Problem Solving from Nature*. Springer, 139–148.
- [26] Günter Rudolph. 1997. Asymptotical convergence rates of simple evolutionary algorithms under factorizing mutation distributions. In *European Conference on Artificial Evolution*. Springer, 223–233.
- [27] G. Rudolph. 1997. Local convergence rates of simple evolutionary algorithms with Cauchy mutations. *IEEE Transactions on Evolutionary Computation* 1, 4 (1997), 249–258. <https://doi.org/10.1109/4235.687885>
- [28] H-P Schwefel. 1965. Kybernetische Evolution als Strategie der experimentellen Forschung in der Stromungstechnik. *Diploma thesis, Technical Univ. of Berlin* (1965).
- [29] Thorsten Suttrop, Nikolaus Hansen, and Christian Igel. 2009. Efficient covariance matrix update for variable metric evolution strategies. *Machine Learning* 75 (2009), 167–197.
- [30] Olivier Teytaud and Sylvain Gelly. 2007. DCMA: yet another derandomization in covariance-matrix-adaptation. In *Genetic and Evolutionary Computation Conference, GECCO 2007, Proceedings, London, England, UK, July 7-11, 2007*, Hod Lipson (Ed.). ACM, 955–963. <https://doi.org/10.1145/1276958.1277150>
- [31] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- [32] Hao Wang, Michael Emmerich, and Thomas Bäck. 2014. Mirrored orthogonal sampling with pairwise selection in evolution strategies. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing* (Gyeongju, Republic of Korea) (SAC '14). Association for Computing Machinery, New York, NY, USA, 154–156. <https://doi.org/10.1145/2554850.2555089>
- [33] Xin Yao and Yong Liu. 1997. Fast evolution strategies. In *International conference on evolutionary programming*. Springer, 149–161.