A Median Success Rule for Non-Elitist Evolution Strategies: Study of Feasibility

Ouassim Ait ElHara
TAO Team, INRIA
Saclay–Île-de-France, LRI,
Paris Sud University, France
firstname.ait_elhara@inria.fr

Anne Auger
TAO Team, INRIA
Saclay–Île-de-France, LRI,
Paris Sud University, France
lastname@lri.fr

Nikolaus Hansen
TAO Team, INRIA
Saclay–Île-de-France, LRI,
Paris Sud University, France
lastname@lri.fr

ABSTRACT

Success rule based step-size adaptation, namely the one-fifth success rule, has shown to be effective for single parent evolution strategies (ES), e.g. the (1+1)-ES. The success rule remains feasible in non-elitist single parent strategies, where the target success rate must be roughly inversely proportional to the population size. This success rule is, however, not easily applicable to multi-parent strategies. In this paper, we introduce the *median success rule* for step-size adaptation, applicable to non-elitist multi-recombinant evolution strategies. The median success rule compares the median fitness of the population to a fitness from the previous iteration. The comparison fitness is chosen to achieve a target success rate of 1/2, thereby a deviation from the target can be measured reliably in comparatively few iteration steps. As a prerequisite for feasibility of the median success rule, we studied the way the fitness comparison quantile depends on the search space dimension, the population size, the parent number, the recombination weights and the objective function. The findings are encouraging: the choice of the comparison quantile appears to be relatively uncritical and experiments on a variety of functions, also in combination with CMA, reveal reasonable behavior.

Categories and Subject Descriptors

I.2.8 [**Problem Solving, Control Methods, and Search**]; I.2.6 [**Learning**]: Parameter Learning; G.1.6 [**Optimization**]

Keywords

Evolution strategies; step-size control; adaptation; median success rule

1. INTRODUCTION

In this paper we consider the problem of minimizing a function defined on a continuous domain $f : \mathbb{R}^n \mapsto \mathbb{R}$ in a black-box scenario. That is, no derivatives of f are available and algorithms can only use objective function values to update their different parameters.

Evolution Strategies address black-box numerical optimization problems by sampling points in $\mathbb{R}^n$ using multivariate normal distributions. We consider the case where at iteration t all solutions are sampled from the same multivariate normal distribution parametrized by a mean vector, $\mathbf{X}_t$, that represents the favorite solution, and a covariance matrix. This covariance matrix determines the overall shape of the distribution and is usually decomposed into a scale parameter called step-size σ_t and a symmetric positive definite matrix $\mathbf{C}_t$. The covariance matrix of the overall sample distribution equals $\sigma_t^2 \mathbf{C}_t$ then.

Adaptation of σ_t, referred to as step-size adaptation, is crucial. Broadly speaking, σ_t determines the speed of convergence and its proper adaptation leads to linear convergence on a relatively large class of functions [5]. A number of methods for the adaptation of the step-size in evolution strategies exist, namely the 1/5-th success rule [12], self-adaptation [13], and cumulative step-size adaptation (CSA) [10]. These rules work comparatively well with a small population size and in particular with a small number of parents, however they often perform suboptimally with large or huge population sizes. While CSA is the standard method in combination with covariance matrix adaptation (CMA) [7], the method heavily relies on properties of the sample distribution and, in combination with a variable metric approach like CMA, on a proper coordinate system transformation into isotropic coordinates. The objective of this paper is to explore a less demanding control mechanism that can achieve close to optimal step-size and fast convergence rates independently of the population size.

The method explored in this paper is inspired from the one-fifth success rule where the probability of success is tracked to determine if the step-size should increase or decrease. This step-size rule was elaborated for a $(1 + 1)$-ES and $1/5$ is a compromise for an optimal asymptotic probability of success on the sphere function and on the corridor model [12]. As we will explain in Section 3, the rule is however not directly applicable to the $(\mu/\mu_w, \lambda)$-ES, where we have non-elitist selection and possibly $\mu \gg 1$.

The *median success rule*, introduced in Section 3, compares the median fitness of the current population, roughly speaking, to the median fitness of the better half of the previous population. Specifically, the fitness comparison quantile in the previous population is chosen such that the success

GECCO'13, July 6-10, 2013, Amsterdam, The Netherlands.

probability with maximal progress is 1/2. Then, larger success probabilities indicate smaller step-sizes and vice versa.

2. CONVERGENCE RATE ON THE SPHERE

The investigations of the new success based rule are carried out on three models, the linear function, the sphere function and a ridge function. In this section we recall useful theoretical results on the sphere. First, we introduce some notations. Given, at iteration t, the mean vector and step-size $(\mathbf{X}_t, \sigma_t)$, λ new solutions are created as

$$\mathbf{X}_t^i = \mathbf{X}_t + \sigma_t \mathbf{N}^i \qquad i = 1, \ldots, \lambda \,, \tag{1}$$

where the $\mathbf{N}^i$ are i.i.d. following a standard multivariate normal distribution and consequently $\mathbf{X}_t^i$ follows the multivariate normal distribution $\mathcal{N}(\mathbf{X}_t, \sigma_t^2 I_n)$ where I_n denotes the identity matrix. These λ solutions are evaluated on the objective function and ranked, i.e.

$$f(\mathbf{X}_t^{1:\lambda}) \leq f(\mathbf{X}_t^{2:\lambda}) \leq \ldots \leq f(\mathbf{X}_t^{\lambda:\lambda}) \,\,,$$

where $i{:}\lambda$ denotes the index of the i^{th} best individual. The new mean vector results from weighted recombination of μ vectors and reads

$$\mathbf{X}_{t+1} = \mathbf{X}_t + \sigma_t \sum_{i=1}^{\mu} w_i \mathbf{N}^{i:\lambda} \,\,, \tag{2}$$

where $w_i \in \mathbb{R}$ are weights satisfying $w_1 \geq \ldots \geq w_\mu > 0$ and summing to one, i.e. $\sum_{i=1}^{\mu} |w_i| = 1$. The step-size is then adapted only using ranking information (and not the exact function value).

We denote an algorithm following (1) and (2) a $(\mu/\mu_w, \lambda)$-ES. If a proper step-size adaptation mechanism is used, linear convergence will be observed on a wide class of functions, which has been proven on spherical and some convex quadratic functions for $\mu = 1$ [5, 8]. This linear convergence is characterized by

$$\lim_{t\to\infty} \frac{1}{\lambda} \frac{1}{t} \ln \frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} = -\,\mathrm{CR} \,\,,$$

where CR is the convergence rate of the algorithm. With a step-size adaptive rule, σ_t is a sequence of random variables. On spherical functions where w.l.o.g. the optimum is in zero (but more generally on so-called scaling-invariant functions) the sequence $\mathbf{X}_t/\sigma_t$ is a homogeneous Markov chain whose study can lead to proofs of linear convergence [5].

An important theoretical algorithm is built by assuming that $\|\mathbf{X}_t\|/\sigma_t$ is constant, or in other words the step-size σ_t is proportional to the distance to the optimum, $\sigma_t = c\,\|\mathbf{X}_t\|$. For a $(1+1)$-ES or $(1, \lambda)$-ES for a certain choice of constant c (see below), this artificial algorithm achieves on the sphere function, the optimal convergence rate that the respective algorithm can achieve on all possible functions [5]. For the $(\mu/\mu_w, \lambda)$-ES, there is no such general proof while we conjecture the result to be true.

Let us now remind some theoretical results for the $(\mu/\mu_w, \lambda)$-ES with distance proportional step-size.

PROPOSITION 1. *Consider the $(\mu/\mu_w, \lambda)$-ES with a step-size proportional to the distance to the optimum, $\sigma_t = \frac{\sigma^\star}{n}\|\mathbf{X}_t\|$ and $\sum |w_i| = 1$ minimizing the spherical functions $f(\mathbf{x}) = g(\|\mathbf{x}\|)$ with g a strictly increasing function. Then the algorithm converges linearly and satisfies more precisely*

$$\lim_{t\to\infty} \frac{1}{\lambda} \frac{1}{t} \ln \frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} = \frac{1}{\lambda} E \ln \|\mathbf{e}_1 + \frac{\sigma^\star}{n} \sum_{i=1}^{\mu} w_i \mathbf{N}^{i:\lambda}\| =: -\,\mathrm{CR} \tag{3}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)$ and $\mathbf{N}^{i:\lambda}$ are obtained after ranking standard i.i.d. multivariate normal vectors according to

$$\|\mathbf{e}_1 + \frac{\sigma^\star}{n}\mathbf{N}^{1:\lambda}\| \leq \|\mathbf{e}_1 + \frac{\sigma^\star}{n}\mathbf{N}^{2:\lambda}\| \leq \ldots \leq \|\mathbf{e}_1 + \frac{\sigma^\star}{n}\mathbf{N}^{\lambda:\lambda}\| \,\,.$$

In addition, for all t

$$\frac{1}{\lambda} E\left[\ln \frac{\|\mathbf{X}_{t+1}\|}{\|\mathbf{X}_t\|}\right] = -\,\mathrm{CR} \,\,. \tag{4}$$

The proof can be found for instance in [4].

The limit of the convergence rate in (3) can be derived to give an asymptotic formula for the convergence rate. This limit is well known and was computed in [2].

PROPOSITION 2 (ASYMPTOTIC CONVERGENCE RATE). *When the dimension n goes to infinity, the convergence rate of the $(\mu/\mu_w, \lambda)$-ES on unimodal spherical functions satisfies*

$$\lim_{n\to\infty} n \times \mathrm{CR} = -\frac{\sigma^\star}{\lambda}\left(\sum_{i=1}^{\mu} w_i E[\mathcal{N}^{i:\lambda}] + \frac{1}{2}\sigma^\star \sum_{i=1}^{\mu} w_i^2\right) \tag{5}$$

where $\mathcal{N}^{i:\lambda}$ are order statistics of standard one-dimensional normal distributions.

A formal proof can be found in [4]. The coefficient $\sum_{i=1}^{\mu} w_i^2$ is sometimes denoted as $1/\mu_{\text{eff}}$, where for $\sum_i |w_i| = 1$ the $\mu_{\text{eff}} \geq 1$ is the variance effective selection mass and equals to the number of selected point, μ, when all recombination weights are equal. The coefficient $-\sum_{i=1}^{\mu} w_i E[\mathcal{N}^{i:\lambda}] =: c_w$ is the progress coefficient and (5) is also called progress rate [6]. The asymptotic convergence rate reads

$$\lim_{n\to\infty} n \times \mathrm{CR} = \frac{\sigma^\star}{\lambda}\left(c_w - \frac{\sigma^\star}{2\mu_{\text{eff}}}\right) \,\,.$$

Convergence takes place for CR > 0 which is the case i ff 0$<$ $\sigma^\star < 2c_w\mu_{\text{eff}}$. The step-size that maximizes the asymptotic convergence rate or progress rate satisfies

$$\sigma^*_{\text{opt}} = \frac{-\sum_{i=1}^{\mu} w_i E[\mathcal{N}^{i:\lambda}]}{\sum_{i=1}^{\mu} w_i^2} = -\mu_{\text{eff}} \sum_{i=1}^{\mu} w_i E[\mathcal{N}^{i:\lambda}] \tag{6}$$

and we refer to this step-size as optimal step-size. Accordingly, for finite dimension, optimal step-size refers to the step-size with the largest convergence rate (see for instance Figure 2). The expression for the optimal step-size depends on the normalization of the weights and we find the relation $\mathrm{CR}(\sigma^\star, \mathbf{w}) = \mathrm{CR}(\alpha\sigma^\star, \mathbf{w}/\alpha)$, hence a normalization of the weights by α implies an α times larger optimal step-size.[1]

Also, from this asymptotic expression of the convergence or progress rate, optimal weights can be derived and are proportional to the expected order statistics [2]. More precisely optimal normalized weights equal

$$w_i^{\text{opt}} = \frac{-E[\mathcal{N}^{i:\lambda}]}{\sum_{k=1}^{\mu} |E[\mathcal{N}^{k:\lambda}]|} \qquad i = 1, \ldots, \mu \,\,. \tag{7}$$

Hence if $i < \lambda \not{2} + 1/2$, the weight w_i^{opt} is positive and negative if i is larger than $\lambda/2 + 1/2$.

[1] This holds because for dimension to infinity, the selection does not depend on the step-size. In finite dimension the step-size matters for selection, because $\|\mathbf{N}^i\|$ varies.

3. DESIGN OF A SUCCESS RULE FOR THE $(\mu/\mu_w, \lambda)$-ES

To design a success-based adaptation rule for the step-size, we need to define success. For the (1+1)-ES there is not much of a choice: success means that the newly sampled point is better than the current mean vector (the parent). However, in the $(\mu/\mu_w, \lambda)$-ES, the definition becomes ambiguous and various choices are available. For instance, we can define success as the i-th best individual being better than the current mean vector, $f(\mathbf{X}_{t+1}^{i:\lambda}) \leq f(\mathbf{X}_t)$. With $i = 1$, this is in accordance with the one-fifth success rule and works reasonably well if μ is small. Alternatively, we can generalize the one-fifth success rule by defining success as the new mean vector being better than the current mean, $f(\mathbf{X}_{t+1}) \leq f(\mathbf{X}_t)$, or as improvement of the i-th best individual, i.e., $f(\mathbf{X}_{t+1}^{i:\lambda}) \leq f(\mathbf{X}_t^{i:\lambda})$.

However in the $(\mu/\mu_w, \lambda)$-ES, the target success rate (for optimal step-size) in all these definitions tends with increasing μ either to zero or to one, because (i) the optimal step-size increases with increasing μ_{eff}, see (6), and (ii) the order statistics of the fitness values become more and more concentrated around their respective expectation.

To adapt the step-size in practice, a success rate is measured (serving as estimate of the true probability of success). For the most reliable estimate of the deviation of the success rate from its target value in the least number of iterations, the target success probability must be close to 1/2. As single measurements are Bernoulli distributed, the number of iterations needed to get a somewhat reliable estimate is larger than $p^{-1} \vee (1-p)^{-1}$ and target values of $0.1 \leq p \leq 0.9$ seem admissible. None of the above success definitions meets this requirement for large values of μ.

We solve this dilemma by comparing the median fitness of the population to a previous *different* fitness percentile. The question of what is the optimal success rate then becomes the question of what is the optimal comparison percentile in the following. Therefore we investigate the previous population index $j(\lambda, n)$ for which the probability of success of the current median, the *median success probability*, becomes 1/2 with optimal step-size,

$$\boxed{\Pr\left[f\left(\mathbf{X}_{t+1}^{m(\lambda)}\right) \leq f\left(\mathbf{X}_t^{j(\lambda,n):\lambda}\right)\right] \approx 1/2} \quad , \tag{8}$$

where $m(\lambda)$ is the index of the offspring with the *median* fitness in iteration $t+1$. In general, the index $j(\lambda, n)$ also depends on μ, on the recombination weights and on the objective function, that is $j(\lambda, n) = j(\lambda, n, \mathbf{w}, f)$. Note that when λ is even, the definition of the median *index* is ambiguous. Then, we implement $f(\mathbf{X}_{t+1}^{m(\lambda)})$ as being $f(\mathbf{X}_{t+1}^{\lambda/2:\lambda})$ or $f(\mathbf{X}_{t+1}^{\lambda/2+1:\lambda})$, each with probability 1/2. We take the expected value of the outcome if invariance to monotonic f-transformations is preserved.

Algorithm implementation.

In order to measure the success of the median individual, compared to the $j(\lambda, n)$-th individual from the previous iteration in practice, we may first count the number of successful individuals in the population,

$$K_{\text{succ}} = \sum_{i=1}^{\lambda} \mathbb{1}_{\left\{f(\mathbf{X}_{t+1}^{i}) \leq f(\mathbf{X}_t^{j(\lambda,n):\lambda})\right\}} \quad . \tag{9}$$

From the definitions follows that $K_{\text{succ}} \geq (\lambda+1)/2 \iff f(\mathbf{X}_{t+1}^{m(\lambda)}) \leq f(\mathbf{X}_t^{j(\lambda,n):\lambda})$ and we define a normalized measurement

$$z = \frac{2}{\lambda}\left(K_{\text{succ}} - \frac{\lambda+1}{2}\right) \tag{10}$$

such that $z \geq 0$ iff the median individual was successful. The target z-value is zero and the range between minimum and maximum z-values is two. We formulate the adaptation in a standardized way including smoothing as

$$s \leftarrow (1 - c_\sigma)s + c_\sigma z \tag{11}$$

$$\sigma \leftarrow \sigma \exp\left(\frac{s}{d_\sigma}\right) \tag{12}$$

In order to implement the algorithm, we must set $j(\lambda, n)$ in (9). To identify appropriate values we conduct some theoretical and empirical investigations on linear, sphere and ridge functions in the following. Generally, the result for $j(\lambda, n) \in \mathbb{R}$ will not be an integer value, that is, $j \notin \mathbb{Z}$. Then we implement (9) as a weighted average of the two closest integer values, $j^- = \lfloor j \rfloor$ and $j^+ = \lceil j \rceil$, in the form $K_{\text{succ}}(j) = p^- K_{\text{succ}}(j^-) + p^+ K_{\text{succ}}(j^+)$ with $K_{\text{succ}}(j^\pm)$ as in (9) and $p^\pm = 1 - |j^\pm - j|$.[2]

4. THEORETICAL DERIVATIONS

We derive in this section the probability of success defined in the LHS of (8), as a function of j, on linear functions, on the sphere function and on a ridge function. For the linear and sphere functions, we assume that during two consecutive iterations t and $t+1$ the step-size σ_t is fixed and equals σ, while it is assumed constant for all t for the ridge function. To avoid ambiguity, we use a time subscript for the standard multivariate normal vectors, i.e. for instance the recombined vector at iteration $t+1$ equals $\mathbf{X}_{t+1} = \mathbf{X}_t + \sigma \sum_{i=1}^{\mu} w_i \mathbf{N}_t^{i:\lambda}$.

4.1 Probability of success on linear functions

We now derive the success probability, as defined in the LHS of (8), on linear functions. Without loss of generality we assume that $f(\mathbf{x}) = [\mathbf{x}]_1$ where $[\mathbf{x}]_1$ denotes the first coordinate of a vector $\mathbf{x}$. Hence, the selection that affects the distribution of the selected steps $\mathbf{N}_t^{i:\lambda}$ boils down to selecting only along the first coordinate of the standard multivariate normal distribution and $([\mathbf{N}_t^{i:\lambda}]_1)_i$ are hence following order statistics of standard normal distributions while all the other components remain distributed as standard normal distributions.

The relation $f(\mathbf{X}_{t+1}^{m(\lambda)}) \leq f(\mathbf{X}_t^{j:\lambda})$ becomes on the linear function

$$[\mathbf{X}_{t+1}]_1 + \sigma[\mathbf{N}_{t+1}^{m(\lambda)}]_1 \leq [\mathbf{X}_t]_1 + \sigma[\mathbf{N}_t^{j:\lambda}]_1 \tag{13}$$

and replacing $\mathbf{X}_{t+1}$ by $\mathbf{X}_t + \sigma \sum_{i=1}^{\mu} w_i \mathbf{N}_t^{i:\lambda}$, the relation simplifies into

$$\sum_{i=1}^{\mu} w_i [\mathbf{N}_t^{i:\lambda}]_1 + [\mathbf{N}_{t+1}^{m(\lambda)}]_1 \leq [\mathbf{N}_t^{j:\lambda}]_1 \quad , \tag{14}$$

where $[\mathbf{N}_{t+1}^{m(\lambda)}]_1$ is the median of the order statistics of λ i.i.d. standard normal distributions and $([\mathbf{N}_t^{i:\lambda}]_1)_{1 \leq i \leq \lambda}$ are independent order statistics of λ standard normal distributions.

[2] Each time a non integer index is encountered, this method is used to simulate its attributes.

Overall we have sketched the expression of the probability of success on linear functions:

PROPOSITION 3 (SUCCESS PROBABILITY ON LINEAR FCTS.). *On linear functions, for a $(\mu/\mu_w,\lambda)$-ES with constant step-size $\sigma > 0$ between iterations t and $t+1$, the probability of success defined in the LHS of (8) equals for all $j = 1,\dots,\lambda$*

$$\boxed{\Pr\left(\tilde{\mathcal{N}}^{m(\lambda)} \leq \mathcal{N}^{j:\lambda} - \sum_{i=1}^{\mu} w_i \mathcal{N}^{i:\lambda}\right)} \tag{15}$$

where $\tilde{\mathcal{N}}^{m(\lambda)}$ is the median of λ order statistics of standard normal distributions $\tilde{\mathcal{N}}^i$ and $\mathcal{N}^{i:\lambda}$ are λ order statistics of standard normal distribution independent of $\tilde{\mathcal{N}}^i$.

The probability in (15) is independent of the dimension and the step-size. For the hypothetical setting $w_j = 1$ and $w_i = 0$ for $i \neq j$ the probability is 1/2. We aim for $j(\lambda, \mathbf{w})$ with clearly a larger median success probability on linear functions (which is the case if $\sum_{i=j+1}^{\mu} w_i$ is small).

4.2 Probability of success on the sphere

We compute the median success probability on unimodal spherical functions assuming a constant step-size σ. This probability of success between iterations t and $t+1$, i.e. $\Pr(\|\mathbf{X}_{t+1}^{m(\lambda)}\| \leq \|\mathbf{X}_t^{j:\lambda}\|)$, equals

$$\Pr(\|\mathbf{X}_{t+1} + \sigma \mathbf{N}_{t+1}^{m(\lambda)}\| \leq \|\mathbf{X}_t + \sigma \mathbf{N}_t^{j:\lambda}\|) \ ,$$

where $\mathbf{X}_{t+1} = \mathbf{X}_t + \sigma \sum_{i=1}^{\mu} w_i \mathbf{N}_t^{i:\lambda}$. This probability of success depends, a priori, on $\mathbf{X}_t$ and σ.

To compute this probability, we first condition with respect to $\mathbf{X}_t$ and use the tower property, $E[X] = E[E[X|Y]]$. The conditional probability of success is a function of $\mathbf{X}_t$ and σ,

$$\Pr\left(\|\mathbf{X}_{t+1} + \sigma \mathbf{N}_{t+1}^{m(\lambda)}\| \leq \|\mathbf{X}_t + \sigma \mathbf{N}_t^{j:\lambda}\| \mid \mathbf{X}_t\right) =: \mathcal{G}(\mathbf{X}_t, \sigma) \ .$$

Let us now consider the function $\mathcal{G}(\mathbf{x}, \sigma)$ for any (deterministic) vector $\mathbf{x}$ of $\mathbb{R}^n$ and σ

$$\mathcal{G}(\mathbf{x}, \sigma) = \Pr\left(\|\mathbf{x} + \sigma \mathbf{N}_t^{j:\lambda}\| \geq \|\mathbf{x} + \sigma \sum_{i=1}^{\mu} w_i \mathbf{N}_t^{i:\lambda} + \sigma \mathbf{N}_{t+1}^{m(\lambda)}\|\right)$$

We define $\sigma^\star = \sigma n / \|\mathbf{x}\|$ and dividing by $\|\mathbf{x}\|$ on both sides inside the probabiliy, we find that $\mathcal{G}(\mathbf{x}, \sigma)$ equals

$$\Pr\left(\|\frac{\mathbf{x}}{\|\mathbf{x}\|} + \frac{\sigma^\star}{n} \mathbf{N}_t^{j:\lambda}\| \geq \|\frac{\mathbf{x}}{\|\mathbf{x}\|} + \frac{\sigma^\star}{n} \sum_{i=1}^{\mu} w_i \mathbf{N}_t^{i:\lambda} + \frac{\sigma^\star}{n} \mathbf{N}_{t+1}^{m(\lambda)}\|\right) .$$

Isotropy of the multivariate normal distribution and of the sphere function implies that this probability is invariant if $\mathbf{x}/\|\mathbf{x}\|$ is replaced by any vector of unit length, w.l.o.g. we will use the vector $\mathbf{e}_1$ and thus $\mathcal{G}(\mathbf{x}, \sigma) = \mathcal{G}(\mathbf{e}_1, \sigma^\star / n)$. Overall we have proven the following proposition.

PROPOSITION 4 (SUCCESS PROBABILITY ON THE SPHERE). *Let us consider a $(\mu/\mu_w, \lambda)$-ES optimizing the sphere function. Let us assume $\mathbf{X}_t = \mathbf{x}$ and a constant step-size between iterations t and $t+1$, i.e. $\sigma_t = \sigma_{t+1} = \sigma$. We define $\sigma^\star = \sigma n / \|\mathbf{x}\|$. Then the success probability defined in the LHS of (8) equals for all $j = 1, \dots, \lambda$*

$$\Pr\left[\|\mathbf{e}_1 + \frac{\sigma^\star}{n} \mathbf{N}^{j:\lambda}\| \geq \|\mathbf{e}_1 + \frac{\sigma^\star}{n} \sum_{i=1}^{\mu} w_i \mathbf{N}^{i:\lambda} + \frac{\sigma^\star}{n} \tilde{\mathbf{N}}^{m(\lambda)}\|\right] , \tag{16}$$

where $\mathbf{N}^{j:\lambda}$ are determined from ranking $\mathbf{N}^i$ i.i.d. standard multivariate normal variables according to the values $\|\mathbf{e}_1 + \frac{\sigma^\star}{n} \mathbf{N}^i\|$ and $\tilde{\mathbf{N}}^{m(\lambda)}$ is the median of the ranking of

$$\|\mathbf{y} + \frac{\sigma^\star}{n} \tilde{\mathbf{N}}^{1:\lambda}\| \leq \ldots \leq \|\mathbf{y} + \frac{\sigma^\star}{n} \tilde{\mathbf{N}}^{\lambda:\lambda}\| \tag{17}$$

where $\mathbf{y} = \mathbf{e}_1 + \frac{\sigma^\star}{n} \sum_{i=1}^{\mu} w_i \mathbf{N}^{i:\lambda}$ and $\tilde{\mathbf{N}}^i$ are i.i.d. standard multivariate normal variables independent of $\mathbf{N}^i$ and the ranking in (17) is conditionally to $(\mathbf{N}^i)$, i.e. $\mathbf{y}$ can be thought as deterministic.

We now compute the asymptotic limit of the probability of success when the dimension n goes to infinity.

PROPOSITION 5 (ASYMPTOTIC SUCCESS PROBABILITY). *Assume that $2[\sum_{i=1}^{\mu} w_i \mathbf{N}^{i:\lambda} + \tilde{\mathbf{N}}^{m(\lambda)}]_1 + \frac{\sigma^\star}{n} \| \sum_{i=1}^{\mu} w_i \mathbf{N}^{i:\lambda} + \tilde{\mathbf{N}}^{m(\lambda)}\|^2$ converges to $2[\sum w_i \mathcal{N}^{i:\lambda} + \tilde{\mathcal{N}}^{m(\lambda)}] + \sigma^\star [\sum w_i^2 + 1]$ almost surely when n goes to infinity. Then the probability of success in (16) for each $j = 1, \dots, \lambda$ converges when n goes to infinity to*

$$\boxed{\Pr\left(\tilde{\mathcal{N}}^{m(\lambda)} \leq \mathcal{N}^{j:\lambda} - \sum_{i=1}^{\mu} w_i \mathcal{N}^{i:\lambda} - \frac{1}{2}\sigma^\star \sum_{i=1}^{\mu} w_i^2\right)} \tag{18}$$

where $(\mathcal{N}^{j:\lambda})_{1 \leq j \leq \lambda}$ are order statistics of λ standard normal variables and $\tilde{\mathcal{N}}^{m(\lambda)}$ is the median of λ independent order statistics $(\tilde{\mathcal{N}}^{j:\lambda})_{1 \leq j \leq \lambda}$ of λ standard normal variables.

For the proof, see appendix in [1]. For step-size $\sigma^\star$ going to zero in (18), we recover the expression for the success probability on the linear function (15), as to be expected.

Interestingly enough, assuming distance-proportional, instead of constant, step-size leads to the same asymptotic probability of success (18). We omit the proof due to space reasons.

We will need later an expression for the median success probability with optimal step-size given in (6) which reads

$$\boxed{\Pr\left(\tilde{\mathcal{N}}^{m(\lambda)} \leq \mathcal{N}^{j:\lambda} - \sum_{i=1}^{\mu} w_i (\mathcal{N}^{i:\lambda} - \frac{1}{2} E[\mathcal{N}^{i:\lambda}])\right)} \ . \tag{19}$$

4.3 Ridge function

The previous theoretical derivations are helpful to later identify $j(\lambda, n)$ with simulations on the sphere function. Similar to the derivations for the 1/5th success rule for the $(1+1)$-ES carried out on the sphere function and the corridor function [12], we consider an other model where we can estimate the function $j(\lambda, n)$ such that (8) is satisfied approximately. We consider thus the following ridge function

$$f(\mathbf{x}) = x_1 + \beta \left(\sum_{i=2}^{n} x_i^2\right)^{\alpha/2} \tag{20}$$

with $\alpha = 4$, and $\beta = 1$. We introduce the notation $[\mathbf{x}]_{2\ldots n}$ to denote the vector $(x_2, \dots, x_n)$ and remind that $[\mathbf{x}]_1$ is the first coordinate of $\mathbf{x}$. Hence the ridge function writes $f(\mathbf{x}) = [\mathbf{x}]_1 + \beta \|[\mathbf{x}]_{2\ldots n}\|^\alpha$ where $\|.\|$ denotes the euclidian norm.

The choice of the parameter α is motivated by the fact that for $\alpha > 2$ the optimal step-size is finite [3]. To reduce the f-value on the ridge function, either the distance to the ridge can be reduced or progress along the ridge can

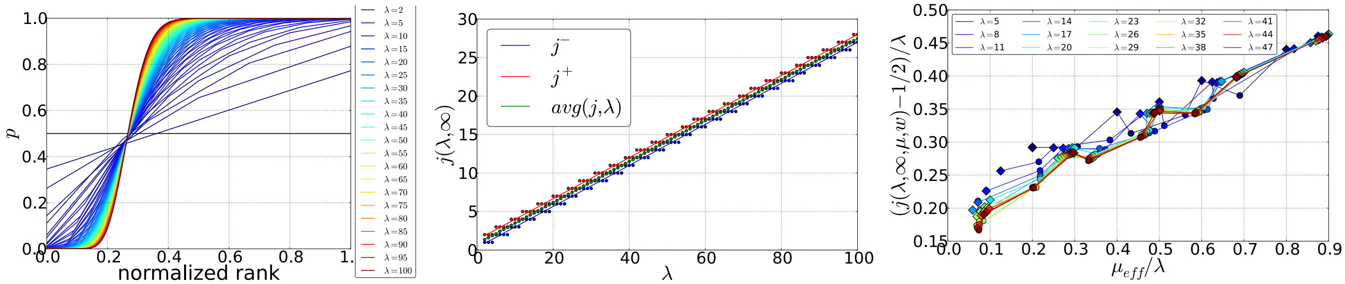

Figure 1: Median success probabilities and optimal comparison quantiles on the ∞-dimensional sphere function. Left: Success probability, p, of the median offspring, given optimal step-size, versus the rank of the previous iteration offspring to which the median is compared to, normalized as $(\text{rank}-1)/(\lambda-1)$, for different population sizes. Center: comparison index $j(\lambda,\infty)$ versus population size λ. Blue (resp. red) points are j^- (resp. j^+) and represent the indices that give, compared to $1/2$, a smaller (resp. larger) but closest success probability; green points are the estimated j, $avg(j,\lambda)$ (see Eq. (23)). Right: optimal comparison quantiles for different values of μ/λ and two weighting schemes: "$\circ$" weighted recombination, Eq. (7), and "$\diamond$" intermediate recombination.

be achieved. While the first objective requires diminution of the step-size, the second one requires that the step-size does not converge to zero. Investigations on ridges were carried out in [3] where the authors arrive to asymptotic progress estimates. However they rely on several assumptions (see [3, Eq. 6] in particular) that remain to be proven rigorously. We focus instead on finite dimensional results and assume a constant step-size σ. Mathematically speaking, $\Phi_t = ([\mathbf{X}_t]_1 - [\mathbf{X}_{t-1}]_1, [\mathbf{X}_t]_{2\ldots n})$ is a homogeneous Markov chain as formalized below.

LEMMA 1. *Let a $(\mu/\mu_w,\lambda)$-ES with constant step-size σ minimize the ridge function (20). Then*

$$\Phi_t = ([\mathbf{X}_t]_1 - [\mathbf{X}_{t-1}]_1, [\mathbf{X}_t]_{2\ldots n}) \tag{21}$$

is an homogeneous Markov chain.

The proof of this Lemma is in the appendix in [1]. Simulations of the Markov Chain$\Phi_{\ t}$ suggest a stable behavior, i.e. convergence towards a stationary distribution. We believe it is possible to prove the stability using standard tools like drift conditions ([9]) but leave the proof for future work and assume it. More precisely, we assume that$\Phi_{\ t}$ is φ-irreducible and aperiodic and admits a stationary distribution π which is a probability measure and is Harris-recurrent [9] so that a Law of Large Numbers can be applied. Under those assumptions we can prove that $\frac{1}{t-t_0}([\mathbf{X}_t]_1 - [\mathbf{X}_{t_0}]_1)$ converges to a constant that coincides with $E_\pi\left[[\mathbf{X}_{t+1}]_1 - [\mathbf{X}_t]_1\right]$. This later quantity coincides with the so-called progress rate [3, Eq. (4)] [11].

PROPOSITION 6. *Consider the $(\mu/\mu_w,\lambda)$-ES with constant step-size σ minimizing the ridge function (20). Assume that the homogeneous Markov Chain Φ_t is φ-irreducible, aperiodic, admits a probability stationary measure π and is Harris recurrent. Then*

$$\boxed{\lim_{t\to\infty}\frac{1}{t-t_0}\left([\mathbf{X}_t]_1 - [\mathbf{X}_{t_0}]_1\right) = \sigma E_\pi\left[\sum_{i=1}^{\mu} w_i[\mathbf{N}^{i:\lambda}]_1\right] =: \varphi(\sigma)} \tag{22}$$

where $\mathbf{N}^{i:\lambda}$ are the selected steps on the ridge starting from $\Phi_t \sim \pi$.

The proof of this Proposition is in the appendix in [1].

5. SIMULATION RESULTS

In this section we perform different numerical experiments to investigate the feasibility of the median success rule. All the results are obtained via Monte-Carlo simulations, except for the asymptotic progress rate on the sphere. Unless otherwise mentioned, we use $\mu = \lfloor\lambda/2\rfloor$ and positive optimal weights from (7).

5.1 Comparison index approximation

In this section we investigate the comparison index $j(\lambda, n)$ using experiments on the sphere function.

Asymptotic results on the sphere.

We start by investigating the index $j(\lambda, n)$ for $n = \infty$, i.e. the asymptotic case of infinite dimension. Figure 1 (left) shows the average success probability of the median offspring for optimal step-size (see (6)), over a set of 10^6 data points, as expressed in (19), versus all possibles comparison indices j, $1 \le j \le \lambda$, with $\lambda \in [2, 100]$. The probability is necessarily monotonous in j. For larger values of λ, the probability reaches values close to zero and one for $j = 1$ and $j = \lambda$ respectively. The j/λ-quantile yielding a success probability of $1/2$ is decreasing with increasing λ and approaches a value close to 0.27 for large λ.

Figure 1 (center) shows the comparison indices j^+ and j^- (red and blue points respectively) that yield the median success probability closest to $1/2$ (compare (8) and end of Section 3). The green points represent the estimated, real-valued, $j(\lambda,\infty)$:

$$avg(j,\lambda) = j^- + \frac{p_s^+ - 0.5}{p_s^+ - p_s^-} \tag{23}$$

where p_s^+ (resp. p_s^-) is the success probability on the LHS of (8) considering λ with $j = j^+$ (resp. $j = j^-$). The index $j(\lambda,\infty)$ appears to depend almost perfectly linearly on λ, the linear regression for j yields

$$j(\lambda,\infty) \simeq 0.27\lambda + 0.83\ . \tag{24}$$

The two plots on the left of Figure 2, show the asymptotic median success probability expressed in (18) versus the step-size for all indices j and for $j(\lambda,\infty)$ taken from (24). For optimal step-size, the comparison index just above $j = \lambda/4$ closely assumes $p \approx 1/2$. The limits to the left for step-size

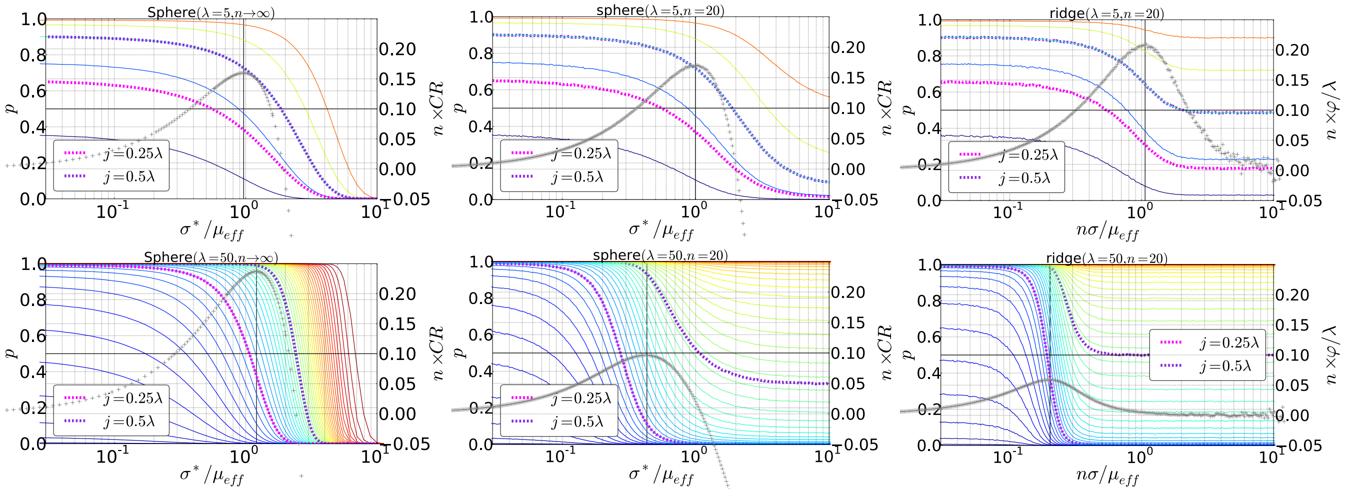

Figure 2: Median success probability, p, for all comparison indices $j = 1, \ldots, \lambda$, increasing from left to right and blue to red, and normalized convergence rate or progress rate respectively on the sphere or ridge function (+) plotted versus $\sigma^\star/\mu_{\text{eff}}$ for the sphere and $\sigma n/\mu_{\text{eff}}$ for the ridge. The dashed lines indicate the 25 and 50%-tile j-value using the weighted average measure of success probability for $j \notin \mathbb{N}$.

to zero correspond to the success probabilities on the linear function whose theoretical expression is given in (15). They show a remarkable dependency on the population size λ.

Variables μ/λ and weights.

We now investigate the effect of changing the number of parents, μ, and the recombination weights. We use selection ratios $r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ setting $\mu = 1 \vee \lfloor r \times \lambda \rfloor$ and consider, in addition to weighted recombination, intermediate recombination with $w_i = 1/\mu$. For weighted recombination we use w_i^{opt} from (7) taking $\lambda = 2 \times \mu$ to compute the weights. Each point is estimated by 10^5 samples.

Figure 1 (right) suggests that the comparison quantile, $(j(\lambda, \infty, \mu, w) - 1/2))/\lambda$, which assumes a median success probability of 1/2 on the ∞-dimensional sphere function with optimal step-size, can be approximated by $\beta \times \mu_{\text{eff}}/\lambda + \alpha$ with $\alpha = 0.15 + 1/(4\lambda)$ and $\beta = 0.35 - 1/(4\lambda)$. The dependency on the different weighting schemes is comparatively small.

5.2 Sphere Function

Figure 2 shows convergence rate and success probabilities on the sphere function versus $\sigma^\star/\mu_{\text{eff}}$ for different λ and n as well as for $n = \infty$. The convergence rate CR for finite dimension is computed by using the expression given in the RHS of (3) averaging over 10^5 realizations[3]. For the asymptotic convergence rate we use the expression given in the RHS of (5) where the coefficients $E[\mathcal{N}^{i:\lambda}]$ are computed by numerical integration. Finite dimensional success probability is computed using (16) (where the previous footnote applies) averaging over 10^5 realizations. The optimal step-size, emphasized on the plots of Figure 2 by a vertical solid line, is obtained (except for $n \to \infty$ for which (5) provides an exact formula) from the empiric data by taking all values that give a progress no less than 90% of the best empirically observed progress and fitting a second degree polynomial to these selected data.

Figure 2 shows some of the results obtained for specific values of λ and n. The behavior of the success probability in finite dimension for an index k resembles that of the same index on the asymptotic case (leftmost plots) when k is close to $j(\lambda, \infty)$. However, it differs significantly when k strays from $j(\lambda, \infty)$.

The limit of the probability of success for step-size to zero (linear case) does not depend on the dimension, compare (15).

5.3 Ridge function

For the ridge function, we consider the case of constant step size and measure the progress made in the direction of the ridge, as we assume a steady state distribution for the distance to the ridge. The ridge function is defined in (20) and its parameters are indicated in Section 4.3. The progress is computed as $\frac{1}{t-t_0}([\mathbf{X}_t]_1 - [\mathbf{X}_{t_0}]_1)$, see Proposition 6, with $t_0 = t - t_0 = 2 \times 10^4$. Here, t_0 is the burn-in time to reach the stationary distribution. Furthermore, the starting point of the simulation, $\mathbf{X}_0 = \mathbf{x}_0$, is chosen close to the stationary state, setting $[\mathbf{x}_0]_1$ to zero and $\|\mathbf{x}_0\|$ to R of Eq. (12) of [3].

The graphs for success probability and progress on the ridge function in Figure 2, right, look qualitatively similar to those for the sphere function (middle and left). For large step-size however the progress flattens out without becoming negative. Consequently, the progress window appears to be wider. On the downside seems the dependency of the optimal j-index on the population size more pronounced than on the sphere function.

5.4 Comparison quantile for optimal step-size

We consider the behavior of $j(\lambda, n)$ in finite dimension n on both, sphere and ridge functions, the same way we did for infinite dimension, when investigating the impact of $\mathbf{w}$ and μ/λ. The data are obtained as for the experiments in Sections 5.2 and 5.3.

Figure 3 depicts the optimal fitness comparison quantile $(j(\lambda, n) - 1/2)/\lambda$ versus dimension for different λ. The optimal comparison percentile decreases with increasing n

[3] Instead of starting from $\mathbf{e}_1$ we start–due to historical reasons–from the vector $(1/\sqrt{n}, \ldots, 1/\sqrt{n})$.

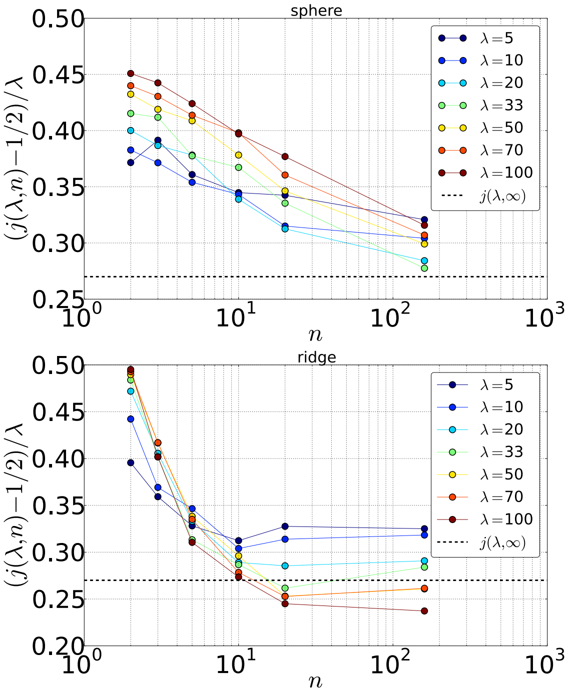

Figure 3: Fitness comparison quantile $(j(\lambda,n)-1/2)/\lambda$ to achieve optimal progress versus dimension on the sphere function (above) and the ridge (below) for $\mu=\lfloor\lambda/2\rfloor$, optimal weights and different λ-values. The dashed line represents the result for infinite dimension on the sphere function.

and approaches the asymptotic value found previously for $n\to\infty$ (dashed lines). The behavior is similar for both functions, while the influence of the dimension on the comparison percentile is slightly more pronounced on the ridge function.

5.5 Single runs

Based on the results from the previous sections and on additional experiments with the *median success rule* (MSR), (10)–(12), applied to the $(\mu/\mu_w,\lambda)$-ES and the $(\mu/\mu_w,\lambda)$-CMA-ES, we identify a feasible preliminary parameter setting: $j(\lambda,n)=0.3\lambda$, $c_\sigma=0.3$, and $d_\sigma=2(n-1)/n$. Figure 4 shows single runs on the sphere function with initial $\mathbf{x}_0=(1,\ldots,1)$ and $\sigma_0=10^{-4}$. The initial setting demands to increase the step-size by about three orders of magnitudes and tests the behavior in a virtually linear environment. Afterwards, convergence on the simplest quadratic function is required. The MSR is compared to CSA-ES (respectively upper left) and CMA-ES with its default step-size adaptation CSA (respectively upper right).

For $n=20$ and $\lambda=10$ we observe the typical behavior on the sphere function, with and without CMA. All four strategies perform comparable. The median success rule, MSR, increases the step-size, from 10^{-4} to 10^{-1}, about two time faster than CSA. The convergence speeds afterwards are almost identical.

Rugged lines show the input signals used in CSA and MSR taken to the power of ten. The value 10^0 marks the distinction between step-size decrease and increase (compare (12)). The norm of the evolution path in CSA clearly distinguishes between both scenarios and consequently step-size increment and decrement are smooth. For MSR, the displayed z-value from (10) appears to be comparatively noisy, in particular during convergence. The smoothing in (11) leads, however, to an acceptably smooth course of σ and an almost indistinguishable course of f compared to CSA.

For $\lambda=1000\gg n=5$ (lower four subfigures) the MSR is overall about 2.5 times faster than CSA (left subfigures), because the latter does not allow for a fast step-size decrement. In both cases, CMA improves the performance remarkably, for CSA even by a factor of ten. Mainly because CSA increases the step-size 50% faster than MSR, the latter is slower in combination with CMA: the covariance ma-

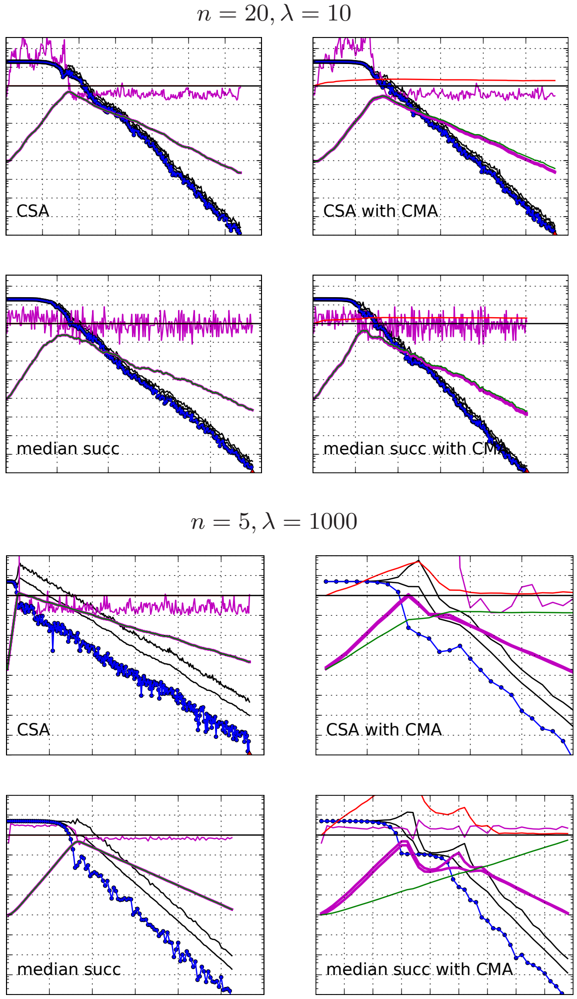

Figure 4: Single runs of the $(\mu/\mu_w,\lambda)$-ES on the sphere function, comparing the median success rule, (10)–(12), with cumulative step-size adaptation (CSA), where the x-axis shows iterations. Line with dots: best f-value of the iteration, median and worst displayed as thinner lines; lines starting from 10^{-4}: σ (green) and largest and smallest coordinate-wise standard deviation of the sample distribution (without CMA they all coincide); rugged line: 10^z with z from (10) and $10^{\|\mathbf{p}_\sigma\|^2/n-1}$ from CSA respectively; line with values ≥ 1 starting from one (red, only CMA): square root of the condition number of the covariance matrix.

trix becomes ill-conditioned in the beginning and needs to be readjusted afterwards (however it should be easily possible to correct this behavior). In combination with CMA, both algorithms reveal comparable convergence speed. The course of σ during the convergence phase is remarkable in either case. For CSA, σ remains virtually constant; for MSR, σ even continuously increases. In both cases, convergence takes place only because the covariance matrix in CMA becomes small. It remains to be seen whether this behavior shall be interpreted as a bug or as a feature.

In further experiments on a number of uni- and multimodal functions no malfunctioning of the median success adaptation were found (see appendix in [1]).

6. DISCUSSION

We have introduced a success based step-size adaptation rule, the *median success rule* (MSR). The *median success rule* compares the offspring with median fitness to an individual from the previous iteration. A small success rate indicates a too large step-size and vice versa. Instead of comparing to the previous median and identifying the step-size-optimal target success rate, we assume a target success rate of 1/2 and identify the step-size-optimal comparison percentile. We find two advantages in this approach: (i) a success rate of 1/2 is optimal to generate a reliable measurement in the least number of iterations, as for (much) larger or smaller success rates the change of success necessarily flattens out under changes of the step-size (compare Figure 2). (ii) in contrast to the target success probability, the target comparison quantile is not strongly affected by the number of parents μ.

Given optimal step-size, the current median is typically better than the previous median. Compared to the median of the previously *selected* population, the current median is typically worse, because additional variance is introduced, unless on the linear function, where additional variance does not affect the median and this success probability is close to 1/2. Consequently, the comparison quantile to achieve median success probability of 1/2 must lie between the median of the selected and the median of the entire previous population, that is, roughly between the $(\mu_{\text{eff}}/\lambda) \times 50\%$- and 50%-tiles. For typical values of $\mu_{\text{eff}}/\lambda \approx 0.3$, comparing to the 30%-tile best individual of the previous iteration therefore turns out to be a reasonable choice. As the optimal quantile depends on the selection ratio μ_{eff}/λ and the dimension n, a more accurate choice might be the $(1+\mu_{\text{eff}}/\lambda+1/n) \times 20\%$-tile.

Unlike cumulative step-size adaptation, as used by default in the $(\mu/\mu_w, \lambda)$-CMA-ES, the median success rule does not explicitly exploit specific properties of the sample distribution. Therefore it is more likely to be broader applicable and, e.g., to be compatible with constraint handling. While the rule does not reveal apparent weaknesses in simulations, it yet relies, like any success-based rule, on an "internal model": the success rate of the median offspring, compared to the 30% best previous offspring, must decrease monotonically with increasing step-size σ and assume $p < 1/2$ for $\sigma \to \infty$ and $p > 1/2$ for $\sigma \to 0$ *on any function*.

Our implementation counts the number of successful offspring, whereby quantifying the step-size misalignment. This trick is conveniently available for the median while it remains ineffective if, e.g., the best offspring is compared. The method retains all invariance properties, namely to order-preserving f-transformations and scale-invariance. The median success rule is, to the best of our knowledge, the first success-based adaptation rule that can be reasonably applied to the $(\mu/\mu_w, \lambda)$-ES even for large μ-values.

Acknowledgments.

This work was supported by the grants ANR-2010-COSI-002 (SIMINOLE) and the ANR-12-MONU-0009 (NumBBO) of the French National Research Agency.

7. REFERENCES

[1] O. Ait Elhara, A. Auger, and N. Hansen. A median success rule for non-elitist evolution strategies: Study of feasibility. Inria, 2013. Preprint with appendix on http://hal.inria.fr/hal-00801414.

[2] D. V. Arnold. Weighted multirecombination evolution strategies. *Theoretical Computer Science*, 361(1):18–37, 2006.

[3] Dirk V. Arnold and Alexander MacLeod. Step length adaptation on ridge functions. *Evol. Comput.*, 16(2):151–184, June 2008.

[4] A. Auger, D. Brockhoff, and N. Hansen. Mirrored sampling in evolution strategies with weighted recombination. In *Genetic and Evolutionary Computation Conference (GECCO 2011)*, pages 861–868. ACM Press, 2011.

[5] A. Auger and N. Hansen. Theory of evolution strategies: A new perspective. In A. Auger and B. Doerr, editors, *Theory of Randomized Search Heuristics: Foundations and Recent Developments*, chapter 10, pages 289–325. World Scientific Publishing Company, 2011.

[6] H.-G. Beyer. *The Theory of Evolution Strategies.* Springer Verlag, 2001.

[7] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[8] Jens Jägersküpper. Algorithmic analysis of a basic evolutionary algorithm for continuous optimization. *Theoretical Computer Science*, 379(3):329–347, 2007.

[9] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, New York, 1993.

[10] A. Ostermeier, A. Gawelczyk, and N. Hansen. Step-size adaptation based on non-local use of selection information. In Y. Davidor et al., editors, *Parallel Problem Solving from Nature (PPSN III)*, volume 866 of *Lecture Notes in Computer Science*, pages 189–198. Springer Verlag, 1994.

[11] A. I. Oyman, H.-G. Beyer, and H.-P Schwefel. Where elitists start limping: Evolution strategies at ridge functions. In et al. Eiben, A. E., editor, *Parallel Problem Solving from Nature (PPSN V)*, Lecture Notes in Computer Science, pages 109–118. Springer Verlag, 1998.

[12] I. Rechenberg. *Evolutionsstrategie '94.* Frommann-Holzboog Verlag, 1994.

[13] H.-P. Schwefel. *Evolution and Optimum Seeking.* Wiley, 1995.