Global Convergence of the (1+1) Evolution Strategy

Tobias Glasmachers
Institute for Neural Computation
Ruhr-University Bochum, Germany
tobias.glasmachers@ini.rub.de

Abstract

We establish global convergence of the (1+1)-ES algorithm, i.e., convergence to a critical point independent of the initial state.

The analysis is based on two ingredients. We establish a sufficient decrease condition for elitist, rank-based evolutionary algorithms, formulated for an essentially monotonically transformed variant of the objective function. This tool is of general value, and it is therefore formulated for general search spaces. To make it applicable to the (1+1)-ES, we show that the algorithm state is found infinitely often in a regime where step size and success rate are simultaneously bounded away from zero, with full probability.

The main result is proven by combining both statements. Under minimal technical preconditions, the theorem ensures that the sequence of iterates has a limit point that cannot be improved in the limit of vanishing step size, a generalization of the notion of critical points of smooth functions. Importantly, our analysis reflects the actual dynamics of the algorithm and hence supports our understanding of its mechanisms, in particular success-based step size control.

We apply the theorem to the analysis of the optimization behavior of the (1+1)-ES on various problems ranging from the smooth (non-convex) cases over different types of saddle points and ridge functions to discontinuous and extremely rugged problems.

1 Introduction

Global convergence of an optimization algorithm refers to convergence of the iterates to a critical point independent of the initial state—in contrast to local convergence, which guarantees this property only for initial iterates in the vicinity of a critical point.[1] For example, many first order methods enjoy this property (Gilbert and Nocedal, 1992), while Newton's method does not. In the realm of direct search algorithms, mesh adaptive search algorithms are known to be globally convergent (Torczon, 1997).

Evolution strategies (ES) are a class of randomized search heuristics for direct search in continuous domains. The covariance matrix adaptation evolution strategy (CMA-ES) by Hansen and Ostermeier (2001) and its many variants mark the state-of-the-art. The algorithm maintains a multivariate Gaussian search distribution parameterized by mean and covariance matrix, which allows for adaptation to linear distortions of the search space, effectively resembling second order methods.

[1] Some authors refer to global convergence as convergence to a global optimum. We do not use the term in this sense.

A "global step size" parameter is maintained in addition. Adaptation of the step size is crucial, since it enables linear convergence on scale invariant problems (e.g., convex quadratic objective functions), while algorithms without step size adaptation converge as slowly as pure random search. Furthermore, being rank-based methods, ESs are invariant to strictly monotonic transformations of objective values. ESs tend to be robust and suitable for solving difficult problems (rugged and multimodal fitness landscapes), a capacity that is often attributed to the above invariance properties.

To date, convergence guarantees for ESs are scarce. Some results exist for convex quadratic problems, which essentially implies local convergence on twice continuously differentiable functions.

Jägersküpper (2003, 2005, 2006a,b) analyzed the (1+1)-ES[2] on the sphere function as well as general convex quadratic functions. His analysis ensures linear convergence with overwhelming probability, i.e., with a probability of $1 - \exp\left(\Omega(d^\varepsilon)\right)$ for some $\varepsilon > 0$, where d is the problem dimension. In other words, the analysis is asymptotic in the sense $d \to \infty$, and for fixed (finite) dimension $d \in \mathbb{N}$, no concrete value or bound is attributed to this probability. A dimension-dependent convergence rate of $\Theta(1/d)$ is obtained.

The analysis by Auger (2005) is based on the stability of the Markov chain defined by the normalized state m/σ. Since the chain is shown to converge to a stationary distribution and the problem is scale-invariant, linear convergence or divergence is obtained, with full probability. There exists sufficient empirical evidence for convergence, however, this is not ensured by the result.

The global convergence analyzed by Akimoto et al. (2010) is closest to the present paper. The analysis is extremely general in the sense that it covers a broad range of algorithms including CMA-ES, with the sole condition of successful divergence on a linear function. Its main restriction is that it applies only to continuously differentiable functions, a problem class where gradient-based methods are often preferable.

In this paper we rigorously prove global convergence of the (1+1)-ES, with an emphasis on covering the widest possible class of problems. To this end we introduce a novel regularity condition ensuring proper function of success-based step-size control, which is much weaker than continuous differentiability.

The paper and the proofs are organized as follows. In the next section we establish a sufficient decrease condition for rank-based elitist algorithms. This condition is extremely general, and it is in no way tied to continuous search spaces and the (1+1)-ES. Its role in the global convergence proof is to ensure a sufficient rate of optimization progress as long as the step size is well adapted and the progress rate is bounded away from zero. In section 3 we define the (1+1)-ES and introduce the regularity condition. Based on this condition we show that the step size returns infinitely often to a range where non-trivial progress can be concluded from the sufficient decrease theorem. Based on these achievements we establish a global convergence theorem in section 4, essentially stating that there exists a sub-sequence of iterates converging to a critical point, the exact notion of which is defined in section 3. We also establish a negative result, showing that a non-optimal critical point results in premature convergence with positive probability, which excludes global convergence. In section 5 we apply the analysis to a variety of settings and demonstrate their implications. We close with conclusions and open questions.

[2] Jägersküpper analyzed a different step size adaptation rule. However, it exhibits essentially the same dynamics as Algorithm 1.

2 Sufficient Decrease

In this section, we establish a general sufficient decrease condition for randomized rank-based elitist algorithms. We consider a general search space X. This space is equipped with a σ-algebra and a reference measure denoted Λ. The usual choice of the reference measure is the counting measure for discrete spaces and the Lebesgue measure for continuous spaces. The objective function $f : X \to \mathbb{R}$, to be minimized, is assumed to be measurable. The parent selection and variation operations of the search algorithm are also assumed to be measurable; indeed we assume that these operators give rise to a distribution from which the offspring is sampled, and this distribution has a density with respect to Λ.

A rank-based optimization algorithm ignores the numerical fitness scores (f-values), and instead relies solely on pairwise comparisons, resulting in exactly one of the relations $f(x) < f(x')$, $f(x) = f(x')$, or $f(x) > f(x')$. This property renders it invariant to strictly monotonically increasing (rank preserving) transformations of the objective values. Therefore it "perceives" the objective function only in terms of its level sets, not in terms of the actual function values. For $f : X \to \mathbb{R}$ let

$$L_f(y) = \Big\{ x \in X \;\Big|\; f(x) = y \Big\}$$
$$S_f^<(y) = \Big\{ x \in X \;\Big|\; f(x) < y \Big\}$$
$$S_f^\leq(y) = \Big\{ x \in X \;\Big|\; f(x) \leq y \Big\}$$

denote the level set of f, and the sub-level sets below and including level $y \in \mathbb{R}$. For $m \in X$ we define the short notations $L_f^<(m) := L_f^<\big(f(m)\big)$, $S_f^<(m) := S_f^<\big(f(m)\big)$ and $S_f^\leq(m) := S_f^\leq\big(f(m)\big)$.

Due to the assumption that the offspring generation distribution is Λ-measurable, with full probability, the algorithm is invariant to the values of the objective function restricted to zero sets. The following definition captures these properties. It encodes the "essential" level set structure of an objective function.

Definition 1. *We call two measurable functions $f, g : X \to \mathbb{R}$ equivalent and write*

$$f \sim g$$

if there exists a zero set $Z \subset X$ and a strictly monotonic function $\phi : f(X) \to g(X)$ such that $g(x) = \phi\big(f(x)\big)$ for all $x \in X \setminus Z$. Here $f(X)$ and $g(X)$ denote the images of f and g, respectively. We denote the corresponding equivalence class in the set of measurable functions by $[f] := f/\!\sim$.

It follows immediately from the definition that the sub-level sets of equivalent objective functions $f \sim g$ coincide outside a zero set.

In the next step we construct a canonical representative for each equivalence class, which we can think of as a "normal form" of an objective function.

Definition 2. *For $f : X \to \mathbb{R}$ we define the spatial suboptimality functions*

$$\widehat{f}_\Lambda^< : X \to \mathbb{R} \cup \{\infty\}, \quad x \mapsto \Lambda\big(S_f^<(x)\big)$$
$$\widehat{f}_\Lambda^\leq : X \to \mathbb{R} \cup \{\infty\}, \quad x \mapsto \Lambda\big(S_f^\leq(x)\big).$$

If $\widehat{f}_\Lambda^<$ and $\widehat{f}_\Lambda^\leq$ coincide then we drop the upper index and simply denote the spatial suboptimality function by $\widehat{f}_\Lambda$.

In the following, $m \in X$ will denote the elite (or parent) point, and $m^{(t)}$ is the elite point in iteration $t \in \mathbb{N}$ of an iterative algorithm, i.e., an evolutionary algorithm with $(1+\lambda)$ selection. For two very different reasons, namely 1) to avoid divergence of the algorithm in the case of unbounded search spaces, and 2) for simplicity of the technical arguments in the proofs, we restrict ourselves to the case that the sub-level set $S_f^\leq\big(m^{(0)}\big)$ of the initial iterate $m^{(0)}$ is bounded and has finite spatial sub-optimality. For most reasonable reference measures, boundedness implies finite spatial sub-optimality. For $X = \mathbb{R}^d$ equipped with the Lebesgue measure this is equivalently to its topological closure of $S_f^\leq\big(m^{(0)}\big)$ being compact. The assumptions immediately imply that $S_f^<(y)$ and $S_f^\leq(y)$ are bounded for all $y \leq f\big(m^{(0)}\big)$, and that restricted to $S_f^\leq(m^{(0)})$ the functions $\widehat{f}_\Lambda^<$ and $\widehat{f}_\Lambda^\leq$ take values in the bounded range $\big[0, \widehat{f}_\Lambda\big(m^{(0)}\big)\big]$. Since an elitist algorithm never accepts points outside $S_f^\leq(m^{(0)})$, we will from here on ignore the issue of infinite $\widehat{f}_\Lambda$-values.[3]

The canonical representation in terms of spatial suboptimality is convenient for defining essential optima (minima) of f, and actually of the whole class $[f]$.

Definition 3. *Consider a measurable function $f : X \to \mathbb{R}$. The set $(\widehat{f}_\Lambda^\leq)^{-1}(0)$ is called the set of essential global optima of f. A point $x \in X$ is called an essential local optimum of f if it is a local optimum of $\widehat{f}_\Lambda^\leq$, i.e., if it is minimal restricted to an (open) neighborhood $N \subset \mathbb{R}^d$ of x.*

In the continuous case, a plateau is a level set of positive Lebesgue measure. Note that according to this definition, an inner point of a plateau is a local optimum, which may not always be intended. Anyway, when analyzing the (1+1)-ES we will not handle plateaus and instead assume that level sets of f are zero sets. This also implies that $\widehat{f}_\Lambda^\leq$ and $\widehat{f}_\Lambda^<$ agree.

Lemma 4. *Let $f : X \to \mathbb{R}$ be measurable. If $\widehat{f}_\Lambda^\leq(x)$ is finite for all $x \in X$ then it holds $\widehat{f}_\Lambda^\leq \sim f \sim \widehat{f}_\Lambda^<$.*

The proof is found in the appendix. We use $\widehat{f}_\Lambda^\leq$ and $\widehat{f}_\Lambda^<$ (or simply $\widehat{f}_\Lambda$ if possible) as a "canonical" representative of its equivalence class (if the function values are finite, but see the discussion above). These functions have the remarkable property

$$\widehat{f}_\Lambda^\leq(x) = \Lambda\big(S_{\widehat{f}_\Lambda^\leq}(x)\big) \qquad \widehat{f}_\Lambda^<(x) = \Lambda\big(S_{\widehat{f}_\Lambda^<}(x)\big)$$

i.e., $\widehat{f}_\Lambda$ encodes the Lebesgue measure of its own sub-level sets. We will measure optimization progress in terms of $\widehat{f}_\Lambda$-values. Decreasing the spatial suboptimality $\widehat{f}_\Lambda$ by $\delta > 0$ amounts to reducing the volume of better points by δ.

Due to the rank-based nature of the algorithms under study we cannot expect to obtain a sufficient decrease condition based on f-values. Instead, the following theorem establishes a sufficient decrease condition measured in terms of the spatial suboptimality function $\widehat{f}_\Lambda$. It gets

[3] An alternative approach to avoiding infinite values is to apply a bounded reference measure with full support, e.g., a Gaussian on $\mathbb{R}^d$. In the absence of a uniform distribution on X, the price to pay for a bounded and everywhere positive reference measure is a non-uniform measure, which complicates matters. These technical complications seem to outweigh the slightly increased generality of the results.

around the problem of inconclusive values in objective space (which, in case of single-objective optimization, is just the real line) by considering a quantity in *search space*, namely the reference measure of the sub-level set.

The algorithm is randomized, hence the decrease follows a distribution. The following theorem controls the expected value as well as the quantiles of the decrease distribution.

Theorem 5. Let P denote a probability distribution on X with a bounded density with respect to Λ. Define the upper bound

$$u = \sup \left\{ \frac{P(A)}{\Lambda(A)} \;\middle|\; A \subset X \text{ measurable with } \Lambda(A) > 0 \right\}$$

on the density. Consider a measurable objective function $f : X \to \mathbb{R}$, as well as a sample $x \sim P$. Define the functions

$$\begin{aligned} r^{<} &: \mathbb{R} \to [0,1], \qquad z \mapsto P\big(f(x) < z\big) \\ r^{\leq} &: \mathbb{R} \to [0,1], \qquad z \mapsto P\big(f(x) \leq z\big). \end{aligned}$$

We define $s : [0,1] \to \mathbb{R}$ as a measurable "inverse" function fulfilling $r^{<}\big(s(q)\big) \leq q \leq r^{\leq}\big(s(q)\big)$ for all $q \in [0,1]$. We collect the discontinuities of $r^{<}$ and $r^{\leq}$ in the set $Z = \big\{ z \in \mathbb{R} \,\big|\, r^{<}(z) < r^{\leq}(z) \big\}$ and define the sum

$$\zeta = \sum_{z \in Z} \Big(r^{\leq}(z) - r^{<}(z) \Big)^2$$

of squared jumps.

Fix a reference point $m \in X$ and let $p = r^{<}\big(f(m)\big)$ denote the probability of strict improvement of x over m. Then for each $q \in [0,p]$, the q-quantile of the $\widehat{f}_\Lambda^{<}$-decrease is bounded from below by

$$\Pr\left(\widehat{f}_\Lambda^{<}(m) - \widehat{f}_\Lambda^{<}(x) \geq \frac{p - r^{<}\big(s(q)\big)}{u} \right) \geq q,$$

and the the q-quantile of the $\widehat{f}_\Lambda^{\leq}$-decrease is bounded from below by

$$\Pr\left(\widehat{f}_\Lambda^{\leq}(m) - \widehat{f}_\Lambda^{\leq}(x) \geq \frac{p - r^{\leq}\big(s(q)\big)}{u} + \Lambda\big(L_f(m)\big) \right) \geq q,$$

The expected $\widehat{f}_\Lambda^{<}$-decrease is bounded from below by

$$\mathbb{E}\Big[\max\big\{0, \widehat{f}_\Lambda^{<}(m) - \widehat{f}_\Lambda^{<}(x)\big\} \Big] \geq \frac{p^2 + \zeta}{2u},$$

and the expected $\widehat{f}_\Lambda^{\leq}$-decrease is bounded from below by

$$\mathbb{E}\Big[\max\big\{0, \widehat{f}_\Lambda^{\leq}(m) - \widehat{f}_\Lambda^{\leq}(x)\big\} \Big] \geq \frac{p^2 - \zeta}{2u} + \Lambda\big(L_f(m)\big).$$

Proof By construction it holds $P\big(L_f(z)\big) = r^{\leq}(z) - r^{<}(z)$ for all $z \in \mathbb{R}$.

For fixed q, define the level $y_q = s(q)$. Consider the distribution of $f(x)$. If $f(x)$ happens to lie in the left tail of mass p, then $f(x)$ is better than $f(m)$. However, if the left tail is restricted to mass $q \leq p$, then $f(x)$ is better than y_p and has "overjumped" a probability mass of $p - q$. The same holds for the distribution of x: if $f(x) < y_q$, then $\widehat{f}_\Lambda^{<}$ has overjumped an area of P-mass of at least $p - q$, and even of mass $p - r^{<}\big(s(q)\big)$. Similarly, if $f(x) \leq y_q$, then $\widehat{f}_\Lambda^{\leq}$ has overjumped an area of P-mass of at least $p - r^{\leq}\big(s(q)\big)$. Due to the very definition of u, this probability mass corresponds to a Λ-measure which is larger by a factor of at least $1/u$, and due to the fact that a strict improvement was sampled, $\widehat{f}_\Lambda^{<}$ is also reduced by $\Lambda\big(L_f(m)\big)$. This proves the lower bounds on the q-quantiles.

The expected $\widehat{f}_\Lambda^{<}$ improvement is lower bounded by

$$\begin{aligned}
\mathbb{E}\Big[\max\big\{0, \widehat{f}_\Lambda^{<}(m) - \widehat{f}_\Lambda^{<}(x)\big\}\Big] &\leq \int_0^p \frac{p - r^{<}\big(s(q)\big)}{u}\, dq \\
&= \int_0^p \frac{p-q}{u}\, dq + \int_0^p \frac{q - r^{<}\big(s(q)\big)}{u}\, dq \\
&= \int_0^p \frac{p-q}{u}\, dq + \sum_{z \in Z} \int_{r^{<}(z)}^{r^{\leq}(z)} \frac{q - r^{<}(z)}{u}\, dq \\
&= \frac{p^2}{2u} + \sum_{z \in Z} \frac{\big(r^{\leq}(z) - r^{<}(z)\big)^2}{2u} \\
&= \frac{p^2 + \zeta}{2u}.
\end{aligned}$$

The proof of the expected $\widehat{f}_\Lambda^{\leq}$ improvement is analogous. ∎

In a $(1 + \lambda)$ evolutionary algorithm, the reference point m is usually the current best iterate (parent), and the distribution P is the search distribution, from which the offspring are sampled. Our main application is the (1+1)-ES introduced in the next section, where x corresponds to the offspring point sampled from a Gaussian centered on m.

Due to the term $\Lambda\big(L_f(m)\big)$ in the decrease of $\widehat{f}_\Lambda^{\leq}$, the theorem covers the fitness-level method (Droste et al., 2002; Wegener, 2003). However, in particular for search distributions spreading their probability mass over many level sets, the theorem is considerably stronger.

In the continuous case, in the absence of plateaus, the statement can be simplified considerably:

Corollary 6. Under the assumptions and with the notation of theorem 5 we assume in addition that all level sets of f have measure zero. Then for each $q \in [0, p]$, the q-quantile of the $\widehat{f}_\Lambda$-decrease is bounded from below by

$$\Pr\left(\widehat{f}_\Lambda(m) - \widehat{f}_\Lambda(x) \geq \frac{p-q}{u}\right) \geq q,$$

and the expected $\widehat{f}_\Lambda$-decrease is bounded from below by

$$\mathbb{E}\Big[\max\big\{0, \widehat{f}_\Lambda(m) - \widehat{f}_\Lambda(x)\big\}\Big] \geq \frac{p^2}{2u}.$$

The following corollary is a broken down version for Gaussian search distributions $\mathcal{N}(m, C)$ with mean m and covariance matrix C, which has the density

$$p(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(C)}} \exp\left(-\frac{1}{2}(x-m)^T C^{-1}(x-m)\right).$$

Corollary 7. Consider the search space $\mathbb{R}^d$ and the Lebesgue measure Λ. Let $f : \mathbb{R}^d \to \mathbb{R}$ denote a measurable objective function with level sets of measure zero. Consider a normally distributed sample $x \sim \mathcal{N}(m, C)$. Under the assumptions and with the notation of theorem 5, for each $q \in [0, p]$, the q-quantile of the $\widehat{f}_\Lambda$-decrease is bounded from below by

$$\Pr\left(\widehat{f}_\Lambda(m) - \widehat{f}_\Lambda(x) > (2\pi)^{d/2} \cdot \sqrt{\det(C)} \cdot (p-q)\right) \geq q,$$

and the expected $\widehat{f}_\Lambda$-decrease is bounded from below by

$$\mathbb{E}\Big[\max\left\{0, \widehat{f}_\Lambda(m) - \widehat{f}_\Lambda(x)\right\}\Big] > (2\pi)^{d/2} \cdot \sqrt{\det(C)} \cdot \frac{p^2}{2}.$$

An isotropic distributions with component-wise standard deviation (step size) $\sigma > 0$ corresponds has covariance matrix $C = \sigma^2 I$, where $I \in \mathbb{R}^{d\times d}$ is the identity matrix; hence we have $\det(C) = \sigma^d$. In the context of continuous search spaces, Jägersküpper (2003) refers to $\widehat{f}_\Lambda$-progress as "spatial gain". He analyzes in detail the gain distribution of an isotropic search distribution on the sphere model. This result is much less general than the previous corollary, since we can deal with *arbitrary* objective functions, which are characterized (locally) only by a single number, the success probability. For the special case of a Gaussian mutation and the sphere function, Jägersküpper's computation of the spatial gain is more exact, since it is tightly tailored to the geometry of the case, in contrast to being based on a general bound. Still, we lose only a multiplicative factor of the gain, which does not impact the analysis significantly.

The above statements apply immediately to evolutionary algorithms with (1+1) selection. Generalizing them to the best out of λ samples is rather straightforward (based on well-known statements on the distribution of the minimum of λ i.i.d. samples), resulting in bounds for the one-step behavior of elitist algorithms with $(1 + \lambda)$ selection. This is not done here because we are not primarily interested in the scaling with λ, which is usually non-essential for the question whether a $(1 + \lambda)$-ES converges to an essential (local) optimum.

3 Success-bases Step Size Control in the (1+1)-ES

In this section we introduce the (1+1)-ES algorithm and provide an analysis of its success-based step size adaptation rule that will allow us to derive global convergence theorems. To this end we introduce a non-standard regularity property.

From here on, we consider the search space $X = \mathbb{R}^d$, equipped with the standard Borel σ-algebra, and Λ denotes the Lebesgue measure. We denote the d-dimensional (multivariate) isotropic normal distribution with expectation m and (component-wise) standard deviation σ by $\mathcal{N}(m, \sigma^2)$. The covariance matrix is $\sigma^2 I$, where I denotes the identity matrix in $\mathbb{R}^d$.

In each iteration $t \in \mathbb{N}$, the state of the (1+1)-ES is given by $(m^{(t)}, \sigma^{(t)}) \in \mathbb{R}^d \times \mathbb{R}^+$. It samples one candidate offspring from the isotropic normal distribution $x^{(t)} \sim \mathcal{N}(m^{(t)}, (\sigma^{(t)})^2)$.

The parent is replaced by successful offspring, meaning that the offspring must perform at least as good as the parent.

The goal of success-based step size adaptation is to maintain a stable distribution of the success rate, for example concentrated around 1/5. This can be achieved with a number of different mechanisms. Here we consider the maybe simplest such mechanism, namely immediate adaptation based on "success" or "failure" of each sample. Pseudocode for the full algorithm is provided in Algorithm 1.

The constants $c_- < 0$ and $c_+ > 0$ in Algorithm 1 control the change of $\log(\sigma)$ in case of failure and success, respectively. They are parameters of the method. For $c_+ + 4 \cdot c_- = 0$ we obtain an implementation of Rechenberg's classic 1/5-rule (Rechenberg, 1973). We call $\tau = \frac{c_-}{c_- - c_+}$ the target success probability of the algorithm, which is always assumed to be strictly less than 1/2. This is equivalent to $c_+ > -c_-$. A reasonable parameter setting is $c_-, c_+ \in \Omega\left(\frac{1}{d}\right)$.

Algorithm 1 (1+1)-ES

input $m^{(0)} \in \mathbb{R}^d$, $\sigma^{(0)} > 0$, i.i.d. random vectors $\left(z^{(t)}\right)_{t\in\mathbb{N}} \sim \mathcal{N}(0,1)$
$t \leftarrow 0$
repeat
  $x^{(t)} \leftarrow m^{(t)} + \sigma^{(t)} \cdot z^{(t)}$
  if $f\left(x^{(t)}\right) \leq f\left(m^{(t)}\right)$ **then**
    $m^{(t+1)} \leftarrow x^{(t)}$
    $\sigma^{(t+1)} \leftarrow \sigma^{(t)} \cdot e^{c_+}$
  else
    $m^{(t+1)} \leftarrow m^{(t)}$
    $\sigma^{(t+1)} \leftarrow \sigma^{(t)} \cdot e^{c_-}$
  $t \leftarrow t + 1$
until *stopping criterion is met*

Two properties of the algorithm are central for our analysis: it is rank-based and it performs elitist selection, ensuring that the best-so-far solution is never lost and the sequence $f(m^{(t)})$ is monotonically decreasing.

Since step-size control depends crucially on the concept of successful offspring, we define the success probability of the algorithm, which is the probability of a sampled point outperforming the parent in the search distribution center.

Definition 8. *For a measurable function $f : \mathbb{R}^d \to \mathbb{R}$, we define the "success probability" functions*

$$p_f^< : \mathbb{R}^d \times \mathbb{R}^+ \to [0,1], \quad (m,\sigma) \mapsto \int_{S_f^<(m)} \frac{1}{(2\pi)^{d/2}\sigma^d} \cdot \exp\left(-\frac{\|x-m\|^2}{2\sigma^2}\right) dx,$$

$$p_f^\leq : \mathbb{R}^d \times \mathbb{R}^+ \to [0,1], \quad (m,\sigma) \mapsto \int_{S_f^\leq(m)} \frac{1}{(2\pi)^{d/2}\sigma^d} \cdot \exp\left(-\frac{\|x-m\|^2}{2\sigma^2}\right) dx.$$

The function $p_f^\leq$ computes the probability of sampling a point at least as good as m, while $p_f^<$ computes the probability of sampling a strictly better point. If $p_f^<$ and $p_f^\leq$ coincide (i.e., if there

are no plateaus), then we write p_f. A nice property of the success probability is that it does not drop too quickly when increasing the step size:

Lemma 9. *For all $m \in \mathbb{R}^d$, $\sigma > 0$ and $c \geq 1$ it holds*

$$p_f^{\leq}(m, c \cdot \sigma) \geq \frac{1}{c^d} \cdot p_f^{\leq}(m, \sigma),$$
$$p_{\overline{f}}^{\leq}(m, c \cdot \sigma) \geq \frac{1}{c^d} \cdot p_{\overline{f}}^{\leq}(m, \sigma).$$

The proof is found in the appendix; this is the case for a number of technical lemmas in this section. The next step is to define a plausible range for the step size.

Definition 10. *For $p \in [0, 1]$ and a measurable function $f : \mathbb{R}^d \to \mathbb{R}$, we define upper and lower bounds*

$$\xi_p^f(m) = \inf\Big\{\sigma \in \mathbb{R}^+ \,\Big|\, p_f^{\leq}(m, \sigma) \leq p\Big\}$$
$$\eta_p^f(m) = \sup\Big\{\sigma \in \mathbb{R}^+ \,\Big|\, p_{\overline{f}}^{\leq}(m, \sigma) \geq p\Big\}$$

on the step size guaranteeing lower and upper bounds on the probability of (strict) improvement.

We think of $\xi_p^f(m)$ with $p > \tau$ as a "too small" step size at m. Similarly, for $p < \tau$, $\eta_p^d(m)$ is a "too large" step size at m. We aim to establish that if the step size is outside this range then step size adaptation will push it back into the range. The main complication is that the range for σ depends on the point m.

The following lemma establishes a gap between lower and upper step size bound, i.e., a lower bound on the size of the step size range.

Lemma 11. *For $0 \leq p_H \leq p_T \leq 1$ it holds $\sqrt[d]{p_H} \cdot \xi_{p_T}^f(x) \leq \sqrt[d]{p_T} \cdot \eta_{p_H}^f(x)$ for all $x \in \mathbb{R}^d$.*

The following definition is central. It captures the ability of the (1+1)-ES to recover from a state with a far too small step size. This property is needed to avoid premature convergence.

Definition 12. *For $p > 0$, a function $f : \mathbb{R}^d \to \mathbb{R}$ is called p-improvable in $x \in \mathbb{R}^d$ if $\xi_p^f(x)$ is positive. The function is called p-improvable on $Y \subset \mathbb{R}^d$ if $\xi_p^f\big|_Y$ is lower bounded by a positive, lower semi-continuous function $\tilde{\xi}_p^f : Y \to (0, 1]$. A point $x \in \mathbb{R}^d$ is called p-critical if it is not p-improvable for any $p > 0$.*

The property of p-improvability is a non-standard regularity condition. The concept applies to measurable functions, hence we do not need to restrict ourselves to smooth or continuous objectives. On the one hand side, the property excludes many measurable and even smooth functions. On the other hand, it is far less restrictive than continuity and smoothness, since it allows the objective function to jump and the level sets to have kinks. Intuitively, in the two-dimensional case, if for each point the sub-level set opens up in an angle of more than $2\pi p$, then the function is p-improvable. This is the case for many discontinuous functions, however, not for all smooth ones. The degree three polynomial $f(x_1, x_2) = x_1^3 + x_2^2$ can serve as a counter example, since every point of the form $(0, x_2)$ is p-critical. All of its contour lines form cuspidal cubics. Local optima are always p-critical, but many critical points of smooth functions are

not (see below). The above example demonstrates that some saddle points share this property, however, if x is p-critical but not essentially locally optimal then $p_f^<(x,\sigma) > 0$ for all $\sigma > 0$. I.e., such a point can be improved with positive probability for each choice of the step size, but in the limit $\sigma \to 0$ the probability of improvement tends to zero.

We should stress the difference between point-wise p-improvability, which simply demands that ξ_p^f is positive, and set-wise p-improvability, which in addition demands that ξ_p^f is lower bounded by a lower semi-continuous positive function. The latter property ensures the existence of a positive lower bound for ξ_p^f on a compact set. In this sense, set-wise p-improvability is uniform on compact sets. In sections 5.5 and 5.6 we will see examples where this makes a decisive difference.

Intuitively, the value of p of a p-improvable function is critical: if it is below τ then the algorithms may be endangered to systematically decrease its step size while it should better do the contrary. However, with a covariance matrix adaptation mechanism in place, a small angle representing the region of improvement can be opened up, which can bring the success probability arbitrarily close to 1/2. Hence, for many (but not for all) cases this implies that with a CMA-style mechanism in place the exact value of p does not matter, as long as a point is not p-critical.

The next lemma establishes that smooth functions are p-improvable in all regular points, and also in most saddle points.

Lemma 13. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable.*

1. *For a regular point $x \in \mathbb{R}^d$, f is p-improvable in x for all $p < \frac{1}{2}$.*
2. *Let Y denote the set of all regular points of f, then f is p-improvable on Y, for all $p < \frac{1}{2}$.*
3. *Let $x \in \mathbb{R}^d$ denote a critical point of f, let f be twice continuously differentiable in a neighborhood of x, and let $H = \nabla^2 f(x)$ denote the Hessian matrix. If H has at least one negative eigen value, then x is not p-critical.*

Similarly, we need to ensure that the step size does not diverge to ∞. This is easy, since the spatial suboptimality is finite:

Lemma 14. *Consider the state (m, σ) of the (1+1)-ES. For each $p \in (0, 1)$, if*

$$\sigma \geq \sqrt[d]{\frac{\widehat{f}_\Lambda(m)}{p \cdot (2\pi)^{d/2}}}$$

then $p_f^<(m, \sigma) \leq p$.

Applying the above inequality to $p < \tau$ implies that for large enough step size σ, the expected change of $\log(\sigma)$ is negative.

The following lemma is elementary. It is used multiple times in proofs, with the interpretation of the event "1" meaning that a statement holds true. It has a similar role as drift theorems in an analysis of the expected or high-probability behavior (Lehre and Witt, 2013; Lengler and Steger, 2016; Correa et al., 2016), however, here we aim for almost sure results.

Lemma 15. *Let $X^{(t)} \in \{0, 1\}$ denote a sequence of independent binary random variables. If there exists a uniform lower bound $\Pr(X^{(t)} = 1) \geq p > 0$, then almost surely there exists an infinite sub-sequence $(t_k)_{k\in\mathbb{N}}$ so that $X^{(t_k)} = 1$ for all $k \in \mathbb{N}$.*

In applications of the lemma, the events of interest are not necessarily independent, however, they can be "made independent" by considering a sequence of independent events with imply the events of interest. This is always the case if the events of actual interest hold with probability of at least p; then an i.i.d. sequence of of Bernoulli events implying corresponding sub-events sub-events with probability of exactly p does the job. We imply this construction in all applications of the lemma.

The following lemma establishes, under a number of technical conditions, that the step size control rule succeeds in keeping the step size stable. If the prerequisites are fulfilled then the result yields an impossible fact, namely that the overall reduction of the spatial suboptimality is unbounded. So the lemma is designed with proofs by contradiction in mind.

Lemma 16. *Let $\left(m^{(t)}, \sigma^{(t)}\right)$ denote the sequence of states of the (1+1)-ES on a measurable objective function $f : \mathbb{R}^d \to \mathbb{R}$. Let $p_T, p_H \in (0,1)$ denote probabilities fulfilling $p_H < \tau < p_T$ and $\frac{p_H}{p_T} \leq e^{d \cdot c_-}$, and assume the existence of constants $0 < b_T < b_H$ such that*

$$b_T \leq \xi^f_{p_T}(m^{(t)}) \quad \text{and} \quad e^{c_+} \cdot \eta^f_{p_H}(m^{(t)}) \leq b_H$$

for all $t \in \mathbb{N}$. Then, with full probability, there exists an infinite sub-sequence $(t_k)_{k \in \mathbb{N}}$ of iterations fulfilling

$$\sigma^{(t_k)} \in \left[\xi^f_{p_T}\left(m^{(t_k)}\right), \eta^f_{p_H}\left(m^{(t_k)}\right)\right] \tag{1}$$

for all $k \in \mathbb{N}$.

4 Global Convergence

In this section we establish our main result. The theorem ensures the existence of a limit point of the sequence $m^{(t)}$ in a subset of desirable locations. In many cases this amounts to convergence of the algorithm to an essential (local) optimum.

Theorem 17. Consider a measurable objective function $f : \mathbb{R}^d \to \mathbb{R}$ with level sets of measure zero. Assume that $K_0 = \overline{S^{\leq}_f\left(m^{(0)}\right)}$ is compact, and let $K_1 \subset K_0$ denote a closed subset. If f is p-improvable on $K_0 \setminus K_1$ for some $p > \tau$, then the sequence $\left(m^{(t)}\right)_{t \in \mathbb{N}}$ has a limit point in K_1.

Proof Lemma 14 ensures the existence of $0 < p_H < e^{-d \cdot c_-} \cdot \tau$ and

$$b_H = \sqrt[d]{\frac{\widehat{f}_\Lambda\left(m^{(0)}\right)}{p_H \cdot (2\pi)^{d/2}}}$$

such that it holds $\eta^f_{p_H}(x) \leq b_H$ uniformly for all $x \in K_0$. In particular, b_H is an upper bound on $\eta^f_{p_H}$.

Let $B(x,r)$ denote the open ball of radius $r > 0$ around $x \in \mathbb{R}^d$ and define the compact set

$$K(r) = K_0 \Big\setminus \bigcup_{x \in K_1} B(x,r).$$

It holds $K(r) \subset K_0 \setminus K_1$ and $\bigcup_{r>0} K(r) = K_0 \setminus K_1$; hence $K(r)$ is a compact exhaustion of $K_0 \setminus K_1$.

Fix $r > 0$, and assume for the sake of contradiction that all points $m^{(t)}$, $t > t_0$, are contained in $K(r)$. We set $p_T = p$. Let $\tilde{\xi}^f_{p_T}$ denote the lower semi-continuous lower bound on $\xi^f_{p_T}$, which is guaranteed to exist due to the p-improvability of f. We define

$$b_T = \min \left\{ \tilde{\xi}^f_{p_T}(m) \,\middle|\, m \in K(r) \right\} > 0$$

and apply lemma 16 to obtain an infinite sub-sequence of states with step size lower bounded by $\sigma^{(t)} \geq b_T > 0$. According to lemma 9, the success probability is lower bounded by $p_f\big(m^{(t)}, \sigma^{(t)}\big) \geq p_I$, $p_I = (b_T/b_H)^d \cdot p_T > 0$, for all $m \in K(r)$ and $\sigma \in [b_T, b_H]$.

Corollary 7 ensures that in each such state the probability to decrease the $\widehat{f}_\Lambda$-value by at least $(2\pi)^{d/2} \cdot b_T^d \cdot p_I/2$ is lower bounded by $p_I/2 > 0$. Then lemma 15 implies that the overall $\widehat{f}_\Lambda$-decrease is almost surely infinite, which contradicts the fact that $\widehat{f}_\Lambda(m^{(0)})$ is finite and $\widehat{f}_\Lambda$ is lower bounded by zero. Hence, the sequence $m^{(t)}$ leaves $K(r)$ after finitely many steps, almost surely. For $r = 1/n$, let t_n denote an iteration fulfilling $m^{(t_n)} \notin K(r)$. The sequence $\big(m^{(t_n)}\big)_{n\in\mathbb{N}}$ does not have a limit point in $K_0 \setminus K_1$ (since that point would be contained in $K(r)$ for some $r > 0$), however, due to the Bolzano-Weierstraß theorem it has at least one limit point in K_0, which must therefore be located in K_1. ∎

The above theorem is of primary interest if K_1 is the set of essential (local) minima of f, or at least the set of critical or p-critical points. Due to the prerequisites of the theorem we always have

$$\overline{\left\{ x \in K_0 \,\middle|\, x \text{ is } p\text{-critical} \right\}} \subset K_1,$$

i.e., p-critical points are candidate limit points.

In accordance with Akimoto et al. (2010), the following corollary establishes convergence to a critical point for continuously differentiable functions.

Corollary 18. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function with level sets of measure zero. Assume that $K_0 = \overline{S^{\leq}_f\big(m^{(0)}\big)}$ is compact. Then the sequence $\big(m^{(t)}\big)_{t\in\mathbb{N}}$ has a critical limit point.

Proof Define $K_1 = \{x \in \mathbb{R}^d \,|\, \nabla f(x) = 0\}$ as the set of critical points. This set is closed. Lemma 13 ensures that f is p-improvable on $K_0 \setminus K_1$ for all $p < 1/2$. Then the claim follows immediately from theorem 17. ∎

Technically the above statements do not apply to problems with unbounded sub-level sets. However, due to the fast decay of the tails of Gaussian search distributions we can often approximate these problems by changing the function "very far away" from the initial search distribution, in order to make the sub-level sets bounded. We may then even apply the theorem with empty K_1, which implies that after a while the approximation becomes insufficient since the algorithm diverges. In this sense we can conclude divergence, e.g., on a linear function. We will use this argument several times in the next section, mainly to avoid unnecessary technical complications when defining saddle points and ridge functions.

We may ask whether p-improvability for $p > \tau$ is not only a sufficient but also a necessary condition for global convergence. This turns out to be wrong. The quadratic saddle point case discussed below in section 5.2 is a counter example, where the algorithm diverges reliably even if the success probability is far smaller than τ. In contrast, the ridge of p-critical saddle points analyzed in section 5.3 results in premature convergence, despite the fact that the critical points form a zero set, and this can even happen for a ridge of p-improvable points with $p < \tau$, see section 5.4. Drift analysis is a promising tool for handling all of these cases. Here we provide a rather simple result, which still suffices for many interesting cases.

Theorem 19. Consider a measurable objective function $f : \mathbb{R}^d \to \mathbb{R}$ with level sets of measure zero. Let $m \in \mathbb{R}^d$ be a p-critical point. If the success probability decays sufficiently quickly, i.e.,

$$\sum_{k=0}^{\infty} p_f^{\leq}(m, e^{k \cdot c_-}) < \infty$$

then for each given $p < 1$ there exists an initial condition such that the (1+1)-ES converges to m with probability of at least p.

Proof Define the zero sequence $S_K = \sum_{k=K}^{\infty} p_f^{\leq}(m, e^{k \cdot c_-})$. For given $p < 1$, there exists a K_0 such that $S_{K_0} < 1 - p$. By definition, the probability of never sampling a successful offspring when starting the algorithm in the initial state $m^{(0)} = m$, $\sigma^{(0)} = e^{K_0 \cdot c_-}$ is given by S_{K_0}; in this case we have $m^{(t)} = m$ for all $t \in \mathbb{N}$. ∎

The above theorem precludes global convergence to an essential (local) optimum in the presence of a suitable non-optimal p-critical point.

5 Case Studies

In this section we analyze various example problems with very different characteristics, by applying the above convergence analysis. We characterize the optimization behavior of the (1+1)-ES, giving either positive or negative results in terms of global convergence. We start with smooth functions and then turn to less regular cases of non-smooth and discontinuous functions. On the one hand side, we show that the theorem is applicable to interesting and non-trivial cases; on the other hand we explore its limits.

5.1 The Rosenbrock Function

The two-dimensional Rosenbrock function is given by

$$f(x_1, x_2) = 100(x_1^2 - x_2)^2 + (x_1 - 1)^2.$$

This is a degree four polynomial. The function is unimodal, but not convex. Moreover, it does not have critical points other than the global optimum $x^* = (1, 1)$.

The Rosenbrock function is a popular test problem since its solution requires a diverse set of optimization behaviors: the algorithm must descend into a parabolic valley, follow the valley while adapting to its curved shape, and finally converge into the global optimum, which is a smooth optimum with non-trivial (but still moderate) conditioning.

Corollary 18 immediately implies convergence of the (1+1)-ES into the global optimum. It does not say anything about the speed of convergence, however, Jägersküpper (2005) established linear convergence in the last phase with overwhelming probability, and since divergence is excluded, Auger (2005) proves linear convergence based on drift and stability of the Markov chain.

Taken together, these results give a rather complete picture of the optimization process: irrespective of the initial state we know that the algorithm manages to locate the global optimum without getting stuck on the way. Once the objective function starts to look quadratic in good enough approximation, the convergence is linear. The same analysis applies to all twice continuously differentiable unimodal function without critical points other than the optimum.

5.2 Saddle Points—The p-improvable Case

We consider the quadratic objective function

$$f(x_1, x_2) = a \cdot x_1^2 - x_2^2$$

with parameter $a > 0$. The origin is a saddle point. It is p-improvable for all $p < \cot^{-1}(a^2)/\pi < 1/2$. For small enough a, the success probability is larger than τ and corollary 18 applies, while for large values of a the success probability decays to zero and we lose all guarantees.

Simulations show that the ES overcomes the zero level set containing the saddle point without a problem, also for large values of a. We conclude that p-improvable saddle points are no obstacle for the algorithm, irrespective of the value of $p > 0$. However, this statement is based on an empirical observation, not on a rigorous proof.

5.3 Saddle Points—The p-critical Case

The cubic polynomial

$$f(x_1, x_2) = x_1^3 + x_2^2$$

has p-critical saddle points on the line $\{0\} \times \mathbb{R} \subset \mathbb{R}^2$, forming a ridge. For small σ we have $p_f^{\leq}(0, \sigma) \in \mathcal{O}(\sqrt{\sigma})$. Hence, the cumulative success probability

$$\sum_{t=0}^{\infty} p_f^{\leq}(0, e^{t \cdot c_-}) = \mathcal{O}\left(\sum_{t=0}^{\infty} e^{t \cdot c_-/2}\right) = \mathcal{O}\left(\frac{1}{1 - e^{c_-/2}}\right) = \mathcal{O}(1)$$

is finite and theorem 19 implies (premature) convergence with arbitrarily high probability.

5.4 Linear Ridge

Consider the linear ridge objective

$$f(x_1, x_2) = x_1 + a \cdot |x_2|$$

with parameter $a > 0$. The function is continuous, and its level sets contain a kink. Again, the line $\{0\} \times \mathbb{R}$ is critical; this is where the function is non-differentiable. The function is p-improvable for $p < \cot^{-1}(a)/\pi < 1/2$. For $a \to \infty$ the success probability decays to zero.

As long as $\cot^{-1}(a)/\pi > \tau$ we can conclude divergence of the algorithm (the intended behavior) from theorem 17. Otherwise we lose this property, and it is well known and easy to check with simulations that for large enough a the algorithm indeed converges prematurely.

5.5 Sphere with Jump

Our next example is an "essentially discontinuous" problem in the sense that in general no function in the equivalence class $[f]$ is continuous. We consider objective functions of the form

$$f(x) = \|x\|^2 + \mathbb{1}_S(x),$$

where $\mathbb{1}_S$ denotes the indicator function of a measurable set $S \subset \mathbb{R}^d$. If S has a sufficiently simple shape then this problem is similar to a constrained problem where S is the infeasible region (Arnold and Brauer, 2008), at least for $\sigma \ll 1$. As long as $m^{(t)} \in S$ the (1+1)-ES essentially optimizes the sphere function, and as soon as $m^{(t)} \notin S$ the (soft) constraint comes into play.

If S is the complement of a star-shaped open neighborhood of the origin then it is easy to see that the function is unimodal and p-improvable for all $p < 1/2$. Theorem 17 applied with $K_1 = \{0\}$ yields the existence of a sub-sequence converging to the origin, which implies convergence of the whole sequence due to monotonicity of $f\big(m^{(t)}\big)$. Again, the results of Jägersküpper and Auger imply linear convergence.

Other shapes of S give different results. For example, if S is a ball not containing the origin then the function is still unimodal. E.g., define S as the open ball of radius $1/2$ around the first unit vector $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$. Then at $m = 3/2 \cdot e_1$ we have $\xi_p^f(m) = 0$ for all $p > 0$, and according to theorem 19 the algorithm can converge prematurely if the step size is small. Alternatively, if S is the closed ball, then all points except the origin are p-improvable for all $p < 1/2$, however, there does not exist a positive lower semi-continuous lower bound on ξ_p^f in any neighborhood of $m = 3/2 \cdot e_1$, and again the algorithm can converge to this point, irrespective of the target success probability τ.

Now consider the strip $S = (a, \infty) \times (0, 1) \subset \mathbb{R}^2$ with parameter $a > 0$. An elementary calculation of the success rate at $m = (a+\varepsilon, 1)$ for $\sigma \to 0$ shows that the (1+1)-ES is guaranteed to converge to the optimum irrespective of the initial conditions if $(a^2+1)^{-1/2} < \cos(2\pi\tau)$, i.e., if a is large enough; otherwise the algorithm can converge prematurely to a point on the edge $(a, \infty) \times \{1\}$ of S.

5.6 Extremely Rugged Barrier

Let us drive the above discontinuous problem to the extreme. Consider the one-dimensional problem

$$f(x) = x + \mathbb{1}_S(x),$$

where $S \subset [-1, 0]$ is a Smith-Volterra-Cantor set, also known as a fat Cantor set. S is closed, has positive measure (usually chosen as $\Lambda(S) = 1/2$), but is nowhere dense. Counter-intuitively, the function is unimodal in the sense that no point is optimal restricted to an open neighborhood (which is what defines a local optimum). Still, intuitively, S should be able to act as a barrier blocking optimization progress with high probability.

The function is point-wise p-improbable everywhere. However, similar to the closed ball case in the previous section, there is no positive, lower semi-continuous lower bound on ξ_p^f. Therefore

theorem 17 does not apply. Indeed, unsurprisingly, simulations[4] show that the algorithm gets stuck with positive probability when initialized with $x^{(0)} > 0$, $x^{(0)} \ll 1$ and $\sigma \ll 1$. When removing 0 from C, then analogous to section 5.3 we obtain $p^{\leq}_{f}(m, \sigma) \in \mathcal{O}(\sqrt{\sigma}))$ for $m = 0$ and small σ, and hence theorem 19 applies.

In contrast, if C is a Cantor set of measure zero then the algorithm diverges successfully, since it ignores zero sets with full probability.

6 Conclusions and Future Work

We have established global convergence of the (1+1)-ES for an extremely wide range of problems. Importantly, with the exception of a few proof details, the analysis captures the actual dynamics of the algorithm and hence consolidates our understanding of its working principles.

Our analysis rests on two pillars. The first one is a probabilistic sufficient decrease condition for rank-based evolutionary algorithms with elitist selection. In its simplest form, it bounds the progress on problems without plateaus. It seems to be quite generally applicable, e.g., to runtime analysis and hence to the analysis of convergence speed.

The second ingredient is an analysis of success-based step size control. The current method barely suffices to show global convergence. It is not suitable for deducing stronger statements like linear convergence on scale invariant problems. Control of the step size on general problems therefore needs further work.

Many natural questions remain open, the most significant are listed in the following. These open points are left for future work.

- The approach does not directly yield results on the speed of convergence. However, the sufficient decrease condition of theorem 5 is a powerful tool for such an analysis. It can provide us with drift conditions and hence yield bounds on the expected runtime and on the tails of the runtime distribution. However, for that to be effective we need better tools for bounding the tails of the step size distribution. Here, again, drift is a promising tool.

- The current results are limited to step-size adaptive algorithms and do not include covariance matrix adaptation. One could hope to extend the proceeding to the (1+1)-CMA-ES algorithm (Igel et al., 2007), or to (1+1)-xNES (Glasmachers et al., 2010). Controlling the stability of the covariance matrix is expected to be challenging. It is not clear whether additional assumptions will be required. As an added benefit, it may be possible to relax the condition $p > \tau$ for p-improvability, by requiring it only after successful adaptation of the covariance matrix.

- Plateaus are currently not handled. Theorem 5 shows how they distort the distribution of the decrease. Worse, they affect step size adaptation, and they make it virtually impossible to obtain a lower bound on the one-step probability of a strict improvement. Therefore, proper handling of plateaus requires additional arguments.

[4]Special care must be taken when simulating this problem with floating point arithmetic. Our simulation is necessarily inexact, however, not beyond the usual limitations of floating point numbers. It does reflect the actual dynamics well. The fitness function is designed such that the most critical point for the simulation is zero, which is where standard floating point numbers have maximal precision.

- In the interest of generality, our convergence theorem only guarantees the existence of a limit point, not convergence of the sequence as a whole. We believe that convergence actually holds in most cases of interest (at least as long as there are no plateaus, see above). This is nearly trivial if the limit point is an isolated local optimum, however, it is unclear for a spatially extended optimum, e.g., a low-dimensional variety or a Cantor set.

- Our current result requires a saddle point to be p-improvable for some $p > \tau$, otherwise the theorem does not exclude convergence of the ES to the saddle point. We know from simulations that the (1+1)-ES overcomes p-improvable saddle points reliably, also for $p \ll \tau$. A proper analysis guaranteeing this behavior would allow to establish statements analogous to work on gradient-based algorithms that overcome saddle points quickly and reliably, see e.g. Dauphin et al. (2014). However, this is clearly beyond the scope of the present paper.

- We provide only a minimal negative result stating that the algorithm may indeed converge prematurely with positive probability if there exists a p-critical point for which the cumulative success probability does not sum to infinity. In section 5.5 it becomes apparent that this notion is rather weak, since the statement is not formally applicable to the case of a closed ball, which however differs from the open ball scenario only on a zero set. This makes clear that there is still a gap between positive results (global convergence) and negative results (premature convergence). Theorem 19 can for sure be strengthened, but the exact conditions remain to be explored. A single p-improvable point with $p < \tau$ is apparently insufficient. A p-critical point may be sufficient, but it is not necessary.

Acknowledgments

I would like to thank Anne Auger for helpful discussions, and I gratefully acknowledge support by Dagstuhl seminar 17191 "Theory of Randomized Search Heuristics".

References

Akimoto, Y., Nagata, Y., Ono, I., and Kobayashi, S. (2010). Theoretical analysis of evolutionary computation on continuously differentiable functions. In *Genetic and Evolutionary Computation Conference*, pages 1401–1408. ACM.

Arnold, D. and Brauer, D. (2008). On the behaviour of the (1+1)-es for a simple constrained problem. In *Parallel Problem Solving from Nature (PPSN)*, pages 1–10. Springer.

Auger, A. (2005). Convergence results for the $(1, \lambda)$-SA-ES using the theory of φ-irreducible Markov chains. *Theoretical Computer Science*, 334(1–3):35–69.

Correa, C., Wanner, E., and Fonseca, C. (2016). Lyapunov design of a simple step-size adaptation strategy based on success. In *Parallel Problem Solving from Nature (PPSN XIV)*, pages 101–110. Springer.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2933–2941. Curran Associates, Inc.

Droste, S., Jansen, T., and Wegener, I. (2002). On the analysis of the (1+ 1) evolutionary algorithm. *Theoretical Computer Science*, 276(1-2):51–81.

Gilbert, J. and Nocedal, J. (1992). Global Convergence Properties of Conjugate Gradient Methods for Optimization. *SIAM Journal on optimization*, 2(1):21–42.

Glasmachers, T., Schaul, T., and Schmidhuber, J. (2010). A Natural Evolution Strategy for Multi-Objective Optimization. In *Parallel Problem Solving from Nature (PPSN) XI*, pages 627–636. Springer.

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

Igel, C., Hansen, N., and Roth, S. (2007). Covariance matrix adaptation for multi-objective optimization. *Evolutionary Computation*, 15(1):1–28.

Jägersküpper, J. (2003). Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces. *Automata, Languages and Programming*, pages 188–188.

Jägersküpper, J. (2005). Rigorous runtime analysis of the (1+ 1) ES: 1/5-rule and ellipsoidal fitness landscapes. In *International Workshop on Foundations of Genetic Algorithms*, pages 260–281. Springer.

Jägersküpper, J. (2006a). How the (1+ 1) ES using isotropic mutations minimizes positive definite quadratic forms. *Theoretical Computer Science*, 361(1):38–56.

Jägersküpper, J. (2006b). Probabilistic runtime analysis of $(1+,\lambda)$, ES using isotropic mutations. In *Proceedings of the 8th annual Conference on Genetic and Evolutionary Computation (GECCO)*, pages 461–468. ACM.

Lehre, P. K. and Witt, C. (2013). General drift analysis with tail bounds. Technical Report arXiv:1307.2559.

Lengler, J. and Steger, A. (2016). Drift analysis and evolutionary algorithms revisited. Technical Report arXiv:1608.03226.

Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung technisher Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog.

Torczon, V. (1997). On the convergence of pattern search algorithms. *SIAM Journal on optimization*, 7(1):1–25.

Wegener, I. (2003). Methods for the analysis of evolutionary algorithms on pseudo-boolean functions. In *Evolutionary Optimization*, pages 349–369. Springer.

Appendix

Here we provide the proofs of technical lemmas that were omitted from the main text in the interest of readability.

Proof of lemma 4 It is immediately clear from definition 1 that the level sets of f are a refinement of the level sets of $\widehat{f}^{\leq}_{\Lambda}$ and $\widehat{f}^{<}_{\Lambda}$, i.e., $f(x) = f(x')$ implies $\widehat{f}^{\leq}_{\Lambda}(x) = \widehat{f}^{\leq}_{\Lambda}(x')$ and $\widehat{f}^{<}_{\Lambda}(x) = \widehat{f}^{<}_{\Lambda}(x')$, and $\widehat{f}^{\leq}_{\Lambda}(x) < \widehat{f}^{\leq}_{\Lambda}(x')$ and $\widehat{f}^{<}_{\Lambda}(x) < \widehat{f}^{<}_{\Lambda}(x')$ both imply $f(x) < f(x')$. It remains to be shown that $\widehat{f}^{\leq}_{\Lambda}$ and $\widehat{f}^{<}_{\Lambda}$ do not join f-level sets of positive measure. Let $y \in \mathbb{R}$ denote a level so that $Y = \big(\widehat{f}^{\leq}_{\Lambda}\big)^{-1}(y)$ has positive measure $\Lambda(Y) > 0$. We have to show that this measure (not necessarily the whole set) is covered by a single f-level set.

Assume the contrary, for the sake of contradiction. Then we find ourselves in one of the following situations:

1. There exist $x, x' \in Y$ fulfilling $a = f(x) < f(x') = a'$ and it holds $\Lambda\big(f^{-1}(a)\big) > 0$ and $\Lambda\big(f^{-1}(a')\big) > 0$. I.e., the mass of Y is split into at least two chunks of positive measure. This implies $\widehat{f}^{<}_{\Lambda}(x') - \widehat{f}^{<}_{\Lambda}(x) \geq \Lambda\big(f^{-1}(a)\big) > 0$, which contradicts the assumption that x and x' belong to the same $\widehat{f}^{<}_{\Lambda}$-level.

2. There exist $x, x' \in Y$ fulfilling $a = f(x) < f(x') = a'$ and it holds $\Lambda\big(f^{-1}(I)\big) > 0$ for the open interval $I = (a, a')$. I.e., a continuum of level sets of measure zero makes up Y. Again, this implies $\widehat{f}^{<}_{\Lambda}(x') - \widehat{f}^{<}_{\Lambda}(x) \geq \Lambda\big(f^{-1}(I)\big) > 0$, leading to the same contradiction as in the first case.

The argument for $\widehat{f}^{\leq}_{\Lambda}$ is exactly analogous. ∎

Proof of lemma 9 It holds

$$\begin{aligned} p^{<}_{f}(m, c\cdot\sigma) &= \int_{S^{<}_{f}(m)} \frac{1}{(2\pi)^{d/2} c^d \sigma^d} \cdot \exp\left(-\frac{\|x-m\|^2}{2c^2\sigma^2}\right) dx \\ &\geq \frac{1}{c^d} \cdot \int_{S^{<}_{f}(m)} \frac{1}{(2\pi)^{d/2} \sigma^d} \cdot \exp\left(-\frac{\|x-m\|^2}{2\sigma^2}\right) dx \\ &= \frac{1}{c^d} \cdot p^{<}_{f}(m, \sigma). \end{aligned}$$

The computation for $p^{\leq}_{f}$ is analogous. ∎

Proof of lemma 11 Fix x and define $\xi = \xi^{f}_{p_T}(x)$. For $c \geq 1$ it holds

$$\begin{aligned} p_T \leq \quad & \int_{S^{<}_{f}\left(f(x)\right)} \frac{1}{(2\pi)^{d/2}\xi^d} \exp\left(-\frac{\|x'-x\|^2}{2\xi^2}\right) dx' \\ = c^d \cdot & \int_{S^{<}_{f}\left(f(x)\right)} \frac{1}{(2\pi)^{d/2} c^d \xi^d} \exp\left(-\frac{\|x'-x\|^2}{2\xi^2}\right) dx' \\ \leq c^d \cdot & \int_{S^{<}_{f}\left(f(x)\right)} \frac{1}{(2\pi)^{d/2} c^d \xi^d} \exp\left(-\frac{\|x'-x\|^2}{2c^2\xi^2}\right) dx'. \end{aligned}$$

In other words, the success probability for step size $c \cdot \xi$ is at least p_T/c^d. Hence, in order to push the success probability below p_T/c^d, the step size must be at least $\xi \cdot c$, which therefore bounds $\eta^f_{p_T/c^d}(x)$ from below. The case $p_H = 0$ is trivial. Applying the above argument with $c = \sqrt[d]{p_T/p_H}$ completes the proof. ∎

Proof of lemma 13 In a small enough neighborhood of a regular point x the function f can be approximated arbitrarily well by a linear function, for which the probability of strict improvement is exactly $1/2$. Hence we have

$$\lim_{\sigma \to 0} p^<_f(x, \sigma) = \frac{1}{2},$$

which immediately implies the first statement.

To show the second statement, we claim that ξ^f_p itself is lower-semi-continuous. We have to show that $(\xi^f_p)^{-1}([0, r])$ is closed for all $r > 0$. Due to $\lim_{\sigma \to 0} p^<_f(x, \sigma) = 1/2 > p$, this is the case if $(p^<_f)^{-1}([0, q])$ is closed for all $q \in [0, 1/2]$. This again follows from the fact that for continuous f the function $p^<_f$ is lower semi-continuous, and even continuous in its second parameter (the step size).

To show the last statement, let v denote an eigen vector of H fulfilling $v^T H v < 0$. For $\sigma \to 0$, the objective function is well approximated by the quadratic Taylor expansion

$$f(x') \approx g(x') = f(x) + (x - x')^T H (x - x').$$

Hence, $S^<_f(x)$ is locally well approximated by $S^<_g(x)$, which is a cone centered on x. Whether a ray $x + \mathbb{R} \cdot z$ belongs to $S^<_g(x)$ or not depends on whether $z^T H z < 0$ or not. Now, the eigen vector v has this property, and due to continuity of g, the same holds for an open neighborhood $v \in N$. The cone $x + \mathbb{R} \cdot N$ is contained in $S^<_g(x)$ and has the same positive probability $s^<_g(x, \sigma) = p > 0$ under $\mathcal{N}(x, \sigma^2)$ for all $\sigma > 0$. We conclude

$$\lim_{\sigma \to 0} p^<_f(x, \sigma) = p > 0,$$

which completes the proof. ∎

Proof of lemma 14 Let $S = S^<_f(m)$ denote the area of improvement, with Lebesgue measure $\widehat{f}_\Lambda(m)$. The probability of sampling from this area is bounded by

$$\begin{aligned}
p^<_f(m) &= \int_S \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{\|x - m\|^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \frac{1}{(2\pi)^{d/2}\sigma^d} \int_S \exp\left(-\frac{\|x - m\|^2}{2\sigma^2}\right) \mathrm{d}x \\
&< \frac{1}{(2\pi)^{d/2}\sigma^d} \int_S \mathrm{d}x \\
&= \frac{\widehat{f}_\Lambda(m)}{(2\pi)^{d/2}\sigma^d} \\
&\leq p,
\end{aligned}$$

where the last inequality is equivalent to the assumption. ∎

Proof of lemma 15 Assume the contrary, for the sake of contradiction. Then

$$\sum_{t=1}^{\infty} X^{(t)} < \infty.$$

Fix $N \in \mathbb{N}$. Hoeffding's inequality applied with $\varepsilon = p/2$ and $n \geq \frac{2N}{p}$ yields

$$\Pr\left(\sum_{t=1}^{n} X^{(t)} \leq N\right) \leq \exp\left(-n \cdot \frac{p^2}{2}\right) \underset{n\to\infty}{\longrightarrow} 0.$$

Hence, for $n \to \infty$, with full probability the infinite sum exceeds N. Since N was arbitrary, we arrive at a contradiction. ∎

Proof of lemma 16 In each iteration, the step size σ is multiplied by either e^{c_-} or e^{c_+}. According to Lemma 11, the condition $\frac{p_H}{p_T} \leq e^{d \cdot c_-}$ yields

$$\frac{\eta^f_{p_H}\big(m^{(t_k)}\big)}{\xi^f_{p_T}\big(m^{(t_k)}\big)} \geq e^{-c_-}.$$

An unsuccessful step of the (1+1)-ES in iteration t result in a reduction of the step size by the factor $\frac{\sigma^{(t+1)}}{\sigma^{(t)}} = e^{c_-} < 1$ and leaves $m^{(t+1)} = m^{(t)}$ unchanged. We conclude that no such step can overjump the interval $\big[\xi^f_{p_T}\big(m^{(t)}\big), \eta^f_{p_H}\big(m^{(t)}\big)\big]$, in the sense of $\sigma^{(t)} \geq \eta^f_{p_H}\big(m^{(t)}\big)$ and $\sigma^{(t+1)} \leq \xi^f_{p_T}\big(m^{(t)}\big)$. The above property also implies $\frac{b_H}{b_T} \geq e^{-c_-}$.

First we show that there exists an infinite sub-sequence of iterations t fulfilling $\sigma^{(t)} \in [b_T, b_H]$. Assume for the sake of contradiction that there exists t_0 such that $\sigma^{(t)} \leq b_T$ for all $t \geq t_0$. The logarithmic step size change $\delta^{(t)} = \log(\sigma^{(t+1)}) - \log(\sigma^{(t)})$ takes the values $c_+ > 0$ with probability at least $p_T > \tau$ and $c_- < 0$ with probability at most $1 - p_T < 1 - \tau$, hence

$$\mathbb{E}\big[\delta^{(t)}\big] \geq \Delta := p_T \cdot c_+ + (1 - p_T) \cdot c_- > 0.$$

For $t_1 > t_0$ we consider the random variable $\log(\sigma^{(t_1)}) = \log(\sigma^{(t_0)}) + \sum_{t=t_0}^{t_1-1} \delta^{(t)}$. The variables $\delta^{(t)}$ are not independent. We create independent variables as follows. For each candidate state (m, σ) fulfilling $\sigma < b_T$ we fix a set $I(m, \sigma) \subset S^<_f(m)$ of improving steps with probability mass exactly p_T under the distribution $\mathcal{N}(m, \sigma^2)$. Let $\tilde{\delta}^{(t)}$ denote the step size change corresponding to $\delta^{(t)}$ for which the step size is increased only if the iterate $m^{(t+1)}$ is contained in $I(m, \sigma)$. Note that these hypothetical step size changes do not influence the actual sequence of algorithm states. Therefore the sequence is i.i.d., and it holds $\tilde{\delta}^{(t)} \leq \delta^{(t)}$. From Hoeffding's inequality applied with $\varepsilon = \Delta/2$ to $\sum_{t=t_0}^{t_1-1} \tilde{\delta}^{(t)} \leq \sum_{t=t_0}^{t_1-1} \delta^{(t)}$ we obtain

$$\Pr\left\{\log(\sigma^{(t_1)}) \leq \log(\sigma^{(t_0)}) + (t_1 - t_0) \cdot \frac{\Delta}{2}\right\} \leq \exp\left(-(t_1 - t_0) \cdot \frac{\Delta^2}{2(c_+ - c_-)^2}\right),$$

i.e., the probability that the log step size grows by less than $\Delta/2$ per iteration on average is exponentially small in $t_1 - t_0$. For $t_1 \gg t_0 + 2/\Delta \cdot \big(\log(b_T) - \log(\sigma^{(t_0)})\big)$ the probability becomes minuscule, and for $t_1 \to \infty$ it vanishes completely. Hence, with full probability, we arrive at a

contradiction. The same logic contradicts the assumption that $\sigma^{(t)} \geq b_H$ for all $t \geq t_0$. Hence, with full probability, sub-episodes of very small and very large step size are of finite length, and according to lemma 15 the sequence of step sizes returns infinitely often to the interval $[b_T, b_H]$.

Next we show that there exists an infinite sub-sequence of iterations fulfilling equation (1). Again, assume the contrary. We know already that $\sigma^{(t)}$ does not stay below b_T or above b_H for an infinite time. Hence, there must exist an infinite sub-sequence fulfilling either

$$\sigma^{(t)} \in \left[b_T, \xi^f_{p_T}\big(m^{(t)}\big)\right] \tag{2}$$

or

$$\sigma^{(t)} \in \left[\eta^f_{p_H}\big(m^{(t)}\big), b_H\right]. \tag{3}$$

Assume an infinite sub-sequence fulfilling equation (2). For each of these iterations, the success probability is lower bounded by p_T. Consider the case of consecutive successes. Until the event

$$\sigma^{(t)} \geq \xi^f_{p_T}\big(m^{(t)}\big) \tag{4}$$

the probability of success remains lower bounded by $p_T > 0$. The condition is fulfilled after at most $n^+ = \big(\log(b_H) - \log(b_T)\big)/c_+$ successes in a row, hence the probability of such an episode occurring is lower bounded by $p_T^{n^+} > 0$. Lemma 15 ensures the existence of an infinite sub-sequence of iterations with this property. Each such episode contains a point fulfilling either equation (1) or equation (4). By assumption, the former happens only finitely often, which implies that the latter happens infinitely often.

Hence, this case as well as the alternative assumption of an infinite sequence fulfilling equation (3) result in an infinite sub-sequence with the property

$$\sigma^{(t)} \in \left[\eta^f_{p_H}\big(m^{(t)}\big), e^{c_+} \cdot b_H\right].$$

Following the same line of arguments as above, as long as $\sigma^{(t)} \geq \eta^f_{p_H}\big(m^{(t)}\big)$, the probability of an unsuccessful step is lower bounded by $1-p_H > 0$. After at most $n^- = \big(\log(b_T)-\log(b_H)+e_+\big)/c_-$ unsuccessful steps in a row, the step size must have dropped below $b_T \leq \eta^f_{p_H}\big(m^{(t)}\big)$, hence the probability of such an episode occurring is lower bounded by $(1-p_H)^{n^-} > 0$. According to lemma 15, an infinite number of such episodes occurs. By construction, these episodes do not change $m^{(t)}$ and, and they cross the interval $\left[\xi^f_{p_T}\big(m^{(t)}\big), \eta^f_{p_H}\big(m^{(t)}\big)\right]$. Therefore there exists an infinite sub-sequence of iterations within this interval, in contradiction to the assumption. ∎