

# How Comma Selection Helps with the Escape from Local Optima<sup>\*</sup>

Jens Jägersküpper and Tobias Storch

Dortmund University, Informatik 2, 44221 Dortmund, Germany  
{JJ|Storch}@Ls2.cs.uni-dortmund.de

**Abstract.** We investigate  $(1,\lambda)$  ESs using isotropic mutations for optimization in  $\mathbb{R}^n$  by means of a theoretical runtime analysis. In particular, a constant offspring-population size  $\lambda$  will be of interest.

We start off by considering an adaptation-less  $(1,2)$  ES minimizing a linear function. Subsequently, a piecewise linear function with a jump/cliff is considered, where a  $(1+\lambda)$  ES gets trapped, i. e., (at least) an exponential (in  $n$ ) number of steps are necessary to escape the local-optimum region. The  $(1,2)$  ES, however, manages to overcome the cliff in an almost unnoticeable number of steps.

Finally, we outline (because of the page limit) how the reasoning and the calculations can be extended to the scenario where a  $(1,\lambda)$  ES using Gaussian mutations minimizes CLIFF, a bimodal, spherically symmetric function already considered in the literature, which is merely SPHERE with a jump in the function value at a certain distance from the minimum. For  $\lambda$  a constant large enough, the  $(1,\lambda)$  ES manages to conquer the global-optimum region – in contrast to  $(1+\lambda)$  ESs which get trapped.

## 1 Introduction

Since Schwefel has introduced the comma selection in the late 1960s (cf. Schwefel (1995)), every now and then there have been long debates about whether to favor elitist or comma selection. Unlike for the discrete search space  $\{0,1\}^n$  where according to Jansen et al. (2005, p. 415) “the difference between an elitist  $(1+\lambda)$  EA and a non-elitist  $(1,\lambda)$  EA is less important”, for optimization in the continuous domain  $\mathbb{R}^n$  this difference can be crucial. It seems common knowledge that comma selection should be auxiliary when a multi-modal function is to be optimized or when noise makes the function to appear multi-modal to the evolution strategy (ES) (cf. Arnold (2002)). On the other hand, it seems clear that on a smooth unimodal function elitist selection will always outperform comma selection – provided that an adequate adaptation is used.

The insights about the optimization of multimodal functions, however, base on intuition and a huge number of experimental investigations of the performance of a large variety of ESs – rather than on theoretical investigations. One reason

---

<sup>\*</sup> supported by the German Research Foundation (DFG) through the collaborative research center “Computational Intelligence” (SFB 531) resp. grant We 1066/11

Copyright: Springer-Verlag, 2006, LNCS 4193, pp. 52-61  
Parallel Problem Solving from Nature 9 (PPSN IX)

for this may be that the common progress-rate approach is unapplicable for these kinds of scenarios since it (implicitly) demands the progress to become stationary (possibly using some kind of normalization, for instance w. r. t. the distance from the optimum and/or the search space dimension). Jägersküpfer (2005) at least proves that elitist selection is no good choice when the fitness landscape shows “cliffs” or “gaps”; the more challenging question whether comma selection would do better is not tackled.

The present paper tackles this question. Namely, we follow this approach and contribute to the debates by investigations that base on probabilistic runtime analysis know from the classical field of the analysis of randomized algorithms in theoretical computer science.

## 2 The simplest scenario

We consider the linear function  $\text{SUM}_n : \mathbb{R}^n \rightarrow \mathbb{R}$  defined by

$$\text{SUM}_n(\mathbf{x}) := \sum_{i=1}^n x_i$$

which is also called ONEMAX when  $\mathbf{x} \in \{0, 1\}^n$ . For a given function-value  $a \in \mathbb{R}$  let  $H_{\text{SUM}=a}$  denote the hyper-plane  $\{\mathbf{x} \mid \text{SUM}(\mathbf{x}) = a\} \subset \mathbb{R}^n$ . Obviously,  $H_{\text{SUM}=a}$  and  $H_{\text{SUM}=b}$  are parallel, and it is easy to see that the distance between the two hyper-planes equals  $|a - b|/\sqrt{n}$ . Furthermore, for a search point  $\mathbf{c} \in \mathbb{R}^n$  let  $H_{\mathbf{c}}$  abbreviate  $H_{\text{SUM}=\text{SUM}(\mathbf{c})}$ , i. e.  $H_{\mathbf{c}} = \{\mathbf{x} \mid \text{SUM}(\mathbf{x}) = \text{SUM}(\mathbf{c})\}$ . Thus, for instance, a mutation of the current search point  $\mathbf{c}$  corresponds to a SUM-gain of 1 (we consider minimization!) iff the mutant  $\mathbf{c}' = \mathbf{c} + \mathbf{m}$  lies in  $H_{\text{SUM}=\text{SUM}(\mathbf{c})-1}$ , implying that  $\text{dist}(\mathbf{c}', H_{\mathbf{c}}) = 1/\sqrt{n}$ , where “dist” denotes to the Euclidean distance – as we minimize in Euclidean  $n$ -space. Furthermore, we focus on the function (class)  $\text{LINC}L\text{IFF}_n^\Delta : \mathbb{R}^n \rightarrow \mathbb{R}$  with  $\Delta : \mathbb{N} \rightarrow \mathbb{R}_{>0}$  defined by

$$\text{LINC}L\text{IFF}_n^\Delta := \begin{cases} \text{SUM}_n(\mathbf{x}) & \text{for } \text{SUM}_n(\mathbf{x}) \geq 0 \\ \text{SUM}_n(\mathbf{x}) + \sqrt{n} \cdot \Delta(n) & \text{for } \text{SUM}_n(\mathbf{x}) < 0 \end{cases}$$

As we minimize, all points  $\mathbf{x}$  with  $\text{SUM}(\mathbf{x}) = 0$  are local optima with function value 0 (there is no global optimum); namely, the hyper-plane  $H_{\text{SUM}=0}$  contains all local optima. For  $\mathbf{x}$  with negative SUM-value a “penalty” of  $\sqrt{n} \cdot \Delta$  is added, where  $\Delta$  might depend on  $n$ . Thus, there are two different hyper-planes with LINC LIFF-value 0: one is  $H_{\text{SUM}=0}$ , which contains all local optima, and the other one is  $H_{\text{SUM}=-\sqrt{n}\Delta}$ . Recall that the distance between these two hyper-planes equals  $\Delta$ .

When talking about “the gain” of a mutation or a step, we mean the *spatial gain* of a mutation/step (unless we explicitly state “SUM-gain”, of course). The change in the SUM-value is merely used as an indicator whether the mutant of  $\mathbf{c}$  lies in the one half-space w. r. t. the hyper-plane  $H_{\mathbf{c}}$  or in the other.

As we focus on isotropically distributed mutation vectors, the larger the length of  $\mathbf{m}$ , the larger the expected distance between the mutant  $\mathbf{c}'$  and  $H_{\mathbf{c}}$

(and the larger the expected SUM-gain). To focus on the core of the reasoning, for the present we consider unit isotropic mutations, i. e. isotropic mutations the lengths of which are not random but concentrated at 1 (so that the mutation vector  $\mathbf{m}$  is uniformly distributed upon the unit hyper-sphere). Later we show how to extend the calculations to (scaled) Gaussian mutations, the length of which follows a (scaled)  $\chi$ -distribution. So, the random spatial gain

$$G := \begin{cases} \text{dist}(\mathbf{c}', H_{\mathbf{c}}) & \text{if } \text{SUM}(\mathbf{c}') < \text{SUM}(\mathbf{c}) \\ -\text{dist}(\mathbf{c}', H_{\mathbf{c}}) & \text{if } \text{SUM}(\mathbf{c}') \geq \text{SUM}(\mathbf{c}) \end{cases}$$

corresponds to the “signed distance” of the mutant from the hyper-plane containing its parent. Jägersküpper (2003) shows that the density of  $G$  at  $g \in [-1, 1]$  equals  $(1 - g^2)^{(n-3)/2} / \Psi$  for  $n \geq 4$ , where  $\Psi := \int_{-1}^1 (1 - g^2)^{(n-3)/2} dg$  lies in the interval  $\sqrt{2\pi} / \sqrt{n - [1.5 \pm 0.5]}$  (normalization), giving a symmetric bell-shaped function with inflection points at  $\pm 1/\sqrt{n-4}$  for  $n \geq 6$ .

When the (1+1) ES minimizes SUM, the expected gain of a step, which consists of a (unit isotropic) mutation *and* selection, equals the expectation of the random variable (r.v.)  $G^+ := G \cdot \mathbb{1}_{\{G \geq 0\}}$  since the indicator variable “ $\mathbb{1}_{\{G \geq 0\}}$ ” implements elitist selection (in this case). We have

$$\bar{g} := \mathbb{E}[G^+] = \int_0^1 g \cdot (1 - g^2)^{(n-3)/2} dg / \Psi = (n-1)^{-1} / \Psi \in \left[ \frac{0.3989}{\sqrt{n+1}}, \frac{0.4}{\sqrt{n}} \right].$$

For the (1, $\lambda$ ) ES, however,  $G_{\lambda;\lambda}$ , the maximum of  $\lambda$  independent copies of  $G$ , equals the gain of a step. The following general property of the second-order statistic of a symmetric r.v. tells us that the expected one-step gain of the (1,2) ES (when optimizing SUM) is at least as large as the one of the (1+1) ES (cf. the appendix for a proof).

**Proposition 1.** *Let the r.v.  $X$  be symmetric, i. e.,  $\mathbb{P}\{X \geq g\} = \mathbb{P}\{X \leq -g\}$  for  $g \in \mathbb{R}$ . Then  $\mathbb{E}[X_{2;2}] \geq \mathbb{E}[X \cdot \mathbb{1}_{\{X \geq 0\}}]$  ( $= \mathbb{E}[X \mid X \geq 0]/2$ ).*

Hence, also the expected total gain of  $i$  steps of the (1,2) ES is at least as large as the expected  $i$ -step gain of the (1+1) ES. There is a crucial difference, though: Unlike for the (1+1) ES, for the (1,2) ES the total gain  $G_{2;2}^{[i]}$  of  $i$  steps, which is formally the sum of  $i$  independent copies of  $G_{2;2}$ , can be negative, i. e., the evolving search point may visit the half-space consisting of all points with a larger SUM-value than the initial search point. Note that  $G_{2;2}^{[i]}$  is a generalized random walk.

We are interested in the r.v.  $G_{2;2}^{\text{inf}} := \inf_{i \geq 0} G_{2;2}^{[i]}$ , the maximum loss compared to the starting point. In particular, we'd like to know  $\mathbb{P}\{G_{2;2}^{\text{inf}} \geq 0\}$ , the probability that the evolving search point is never (i. e. even when running the (1,2) ES ad infinitum) worse than the initial one. (As the very first step yields a negative gain with probability 1/4, obviously  $\mathbb{P}\{G_{2;2}^{\text{inf}} \geq 0\} \leq 3/4$ .)

**Lemma 2.**  $\mathbb{P}\{G_{2:2}^{\text{inf}} \geq 0\} = \Omega(1)$ .

*Proof.* Recall that  $\mathbb{E}[G_{2:2}] \geq \bar{g}$  ( $= \mathbb{E}[G^+]$ ). Consider the partition of  $\mathbb{R}_{\geq 0}$  given by the intervals  $P_i = [\bar{g} \cdot i^3; \bar{g} \cdot (i+1)^3]$  for  $i \in \mathbb{N}_0$ . Note that the width of  $P_i$  equals  $w_i := \bar{g} \cdot (3i^2 + 3i + 1)$ . We identify the current search point with the corresponding total (spatial) gain. Then we are interested in the probability of getting from  $P_i$  to  $P_{>i} := \cup_{j>i} P_j$  without hitting  $\mathbb{R}_{<0}$ . In fact, we want to prove that, when starting in  $P_i$ , the probability of hitting  $\mathbb{R}_{<0}$  before hitting  $P_{>i}$  is  $e^{-\Omega(i)}$ . Since, for  $k$  a constant large enough,  $\sum_{i \geq k} e^{-\Omega(i)} \leq 1/2$ , we would know that once the current individual has made it into  $P_k$ , then with probability at least  $1/2$  it would never again visit the half-space corresponding to a negative total gain. On the other hand, since  $\mathbb{P}\{G_{2:2} \geq \bar{g}\} \geq \mathbb{P}\{G \geq \bar{g}\} = \Omega(1)$ , with probability  $\mathbb{P}\{G_{2:2} \geq \bar{g}\}^{k^3} = \Omega(1)$  each of the first  $k^3$  steps yields a gain of at least  $\bar{g}$ , implying that  $P_k$  is hit without visiting  $\mathbb{R}_{<0}$ . All in all, we'd have shown that  $\mathbb{R}_{<0}$  is never visited right from the start with probability  $\Omega(1) \cdot 1/2 = \Omega(1)$ .

It remains to show that the probability of hitting  $\mathbb{R}_{<0}$  before  $P_{>i}$  when starting in  $P_i$  is in fact bounded by  $e^{-\Omega(i)}$ . Therefore, recall that the width of  $P_i$  equals  $w_i = \bar{g} \cdot (3i^2 + \Theta(i))$ . Thus, the expected number of steps necessary to get from  $\bar{g} \cdot i^3$  ( $= \min P_i$ ) into  $P_{>i}$  (possibly including a visit to  $\mathbb{R}_{<0}$ ) is at most  $w_i/\bar{g} = 3i^2 + \Theta(i)$  (by using a modification of Wald's equation). As  $P_i$  is at distance  $\bar{g} \cdot i^3$  from  $\mathbb{R}_{<0}$ , one may already intuit that the probability of a visit to  $\mathbb{R}_{<0}$  becomes smaller and smaller as  $i$  increases.

Formally, we want to prove that this probability is  $e^{-\Omega(i)}$ . Therefore, consider the period starting (ending) with the first visit to  $P_i$  (resp.  $P_{>i}$ ). Assume that in each mutation in this period  $|G|$  was at most  $\sqrt{i} \cdot \bar{g}$ . Then in each step  $G_{2:2} \geq -\sqrt{i} \cdot \bar{g}$ , and thus, more than  $\bar{g} \cdot i^3 / (\sqrt{i} \cdot \bar{g}) = i^{2.5}$  steps would be necessary for a visit to  $\mathbb{R}_{<0}$  to be at all possible. For  $i$  large enough, the expected *conditional* one-step gain (under the condition  $|G| \leq \sqrt{i} \bar{g}$ ) is at least  $\bar{g}/2$  (see appendix), and hence, the expected number of necessary steps (under the condition on  $|G|$ ) is at most  $2 \cdot (3i^2 + \Theta(i)) = 6i^2 + \Theta(i)$ . By Hoeffding's bound, for  $i$  large enough,  $9i^2$  steps do *not* suffice with a probability of  $e^{-\Omega(i)}$  (see appendix). As the condition on  $|G|$  is *not* met also with probability  $e^{-\Omega(i)}$  (see appendix), the total failure probability (of *not* getting from  $P_i$  into  $P_{>i}$  within  $9i^2$  steps such that in each of these steps  $|G| \leq \sqrt{i} \bar{g}$  for both mutations) is upper bounded by  $e^{-\Omega(i)} + 2 \cdot 9i^2 \cdot e^{-\Omega(i)} = e^{-\Omega(i)}$ . Finally note that (under the condition on  $|G|$  and for  $i$  large enough)  $\mathbb{R}_{<0}$  cannot be reached in  $9i^2$  steps as we have already seen. In short, with probability  $1 - e^{-\Omega(i)}$  the search gets from  $P_i$  (in particular from  $\bar{g} \cdot i^3 = \min P_i$ ) into  $P_{>i}$  without visiting  $\mathbb{R}_{<0}$  in at most  $9i^2$  steps.  $\square$

As " $G_{2:2}^{\text{inf}} \geq 0$ " implies that  $\mathbb{R}_{<0}$  is never visited, the probability of observing  $v > 0$  visits to  $\mathbb{R}_{<0}$  is bounded above by  $(1 - \Omega(1))^v = e^{-\Omega(v)}$ . Thus, the search drops behind the hyper-plane containing the initial search point at most  $n^\varepsilon$  times w. o. p., where we can choose the positive constant  $\varepsilon$  arbitrarily small.

Now consider the minimization of  $\text{LINCLIFF}_n^\Delta$  where  $\Delta > 0$ . Recall that there are two different hyper-planes with  $\text{LINCLIFF}$ -value 0:  $H_{\text{SUM}=0}$ , which contains all local optima, and  $H_{\text{SUM}=-\sqrt{n}\Delta}$ . The distance between these two hyper-planes

equals  $\Delta$ . Call the half-space  $H_{\text{SUM} \geq 0} = \{\mathbf{x} \mid \text{SUM}(\mathbf{x}) \geq 0\}$  local-optimum region. Then a mutant  $\mathbf{c}'$  of  $\mathbf{c} \in H_{\text{SUM} \geq 0}$  that hits  $H_{\text{SUM} < 0}$  (i.e., it leaves the local-optimum region) such that  $\text{LINCLIFF}_n^\Delta(\mathbf{c}') \leq \text{LINCLIFF}_n^\Delta(\mathbf{c})$  must necessarily yield a spatial gain of at least  $\Delta$ . Then  $\mathbb{P}\{G \geq \Delta\}$  equals the corresponding probability of such a successful mutation. For unit isotropic mutations, the elitist  $(1+\lambda)$  ES cannot overcome the cliff if  $\Delta \geq 1$ , of course. Jägersküpper (2005) investigates how the chances of  $(1+\lambda)$  ES (using isotropic mutations) to get over cliffs/gaps depends on how the size of the cliff relates to the step length/mutation strength. Note that, unlike for the spherical symmetric function  $\text{CLIFF}_n^\Delta$  considered therein, for  $\text{LINCLIFF}_n^\Delta$  there is always a good chance of getting over the cliff if only the step length is made appropriately large.

In the present paper, however, we show that a (1,2) ES manages to overcome the cliff in a “short” time *independently* of how large  $\Delta$  is. The challenge is to show that drop-backs to  $H_{\text{SUM} \geq 0}$  become more and more unlikely with the number of escapes and, in particular, to prove an upper bound on the number of steps necessary to get that far away from the local-optimum region such that there is w. o. p. no drop-back. The next result tells us that, if the current search point is “close to the cliff” in the local-optimum region, then with a “considerable” probability the local-optimum region is left in the next step once and for all.

**Lemma 3.** *Let the (1,2) ES minimize  $\text{LINCLIFF}_n^\Delta$  using unit isotropic mutations. Assume that after  $t$  steps the current search point  $\mathbf{c}^{[t]}$  lies in the half-space  $H_{\text{SUM} \geq 0}$  such that  $\mathbb{P}\{\mathbf{c}^{[t]} + \mathbf{m} \in H_{\text{SUM} < 0}\} = \Omega(1)$ . Then, independently of  $\Delta$ ,  $\mathbb{P}\{\mathbf{c}^{[t+j]} \in H_{\text{SUM} < 0} \text{ for } j \in \mathbb{N}\} = \Omega(1)$ .*

*Proof.* Obviously, we will follow the proof of Lemma 2. With a probability of  $\mathbb{P}\{\mathbf{c}^{[t]} + \mathbf{m} \in H_{\text{SUM} < 0}\}^2 = \Omega(1)$  both mutants of  $\mathbf{c}^{[t]}$  generated in the next step lie in  $H_{\text{SUM} < 0}$  so that one of them becomes  $\mathbf{c}^{[t+1]}$ . Subsequently, with a probability of  $(\mathbb{P}\{G \geq \bar{g}\} \cdot 1/2)^{k^3} = \Omega(1)$  for the constant  $k$  from the proof of Lemma 2, in each of the  $k^3$  following steps both mutants yield positive gains such that one of them is at least  $\bar{g}$ . Then a drop-back to  $H_{\text{SUM} \geq 0}$  is precluded within these steps, and moreover, the distance from  $H_{\text{SUM} \geq 0}$  is at least  $k^3 \bar{g}$  after these steps. From here on (when  $i \geq k$ ), exactly the same reasoning about getting from  $P_i$  into  $P_{>i}$  without ever dropping behind  $H_{\text{SUM} = 0}$  as in the proof of Lemma 2 applies.  $\square$

As a consequence, w. o. p. we observe at most  $n^\varepsilon$  drop-backs, where the constant  $\varepsilon > 0$  can be chosen arbitrarily small. The question is how many steps it takes the (1,2) ES until this has happened. Therefore, we must show first that, when in  $H_{\text{SUM} \geq 0}$ , the search gets close enough to the cliff  $H_{\text{SUM} = 0}$  for  $\mathbb{P}\{\mathbf{c} + \mathbf{m} \in H_{\text{SUM} < 0}\}$  to be  $\Omega(1)$ . Note that (as Jägersküpper (2003) shows) in fact  $\mathbb{P}\{\mathbf{c} + \mathbf{m} \in H_{\text{SUM} < 0}\} = \Omega(1) \iff \text{dist}(\mathbf{c}, H_{\text{SUM} < 0}) = O(\mathbb{E}[G^+])$ . The next result tells us that, when the search approaches the cliff, as long as the distance from the cliff is at least four times the (stationary one-step) drift on SUM, the drift towards the cliff is at least a quarter of this drift.

**Lemma 4.** *Let the (1,2) ES minimize  $\text{LINCLIFF}_n^\Delta$  in  $\mathbb{R}^n$  using unit isotropic mutations. If the search point  $\mathbf{c}$  lies in the local-optimum region  $H_{\text{SUM} \geq 0}$  such that  $\text{dist}(\mathbf{c}, H_{\text{SUM}=0}) \geq 4\mathbb{E}[G^+]$  then  $\mathbb{E}[G_{2:2} \cdot \mathbf{1}_{\{G_1, G_2 \leq \text{dist}(\mathbf{c}, H_{\text{SUM}=0})\}}] \geq \mathbb{E}[G^+]/4$ .*

*Proof.* Recall  $\bar{g} := \mathbb{E}[G^+]$ . The appendix shows  $\mathbb{E}[G^+ \cdot \mathbf{1}_{\{G \leq \sqrt{2/n}\}}] \geq \bar{g}/2$  as well as  $4\bar{g} \geq \sqrt{2/n}$ , and why this implies  $\mathbb{E}[G_{2:2} \cdot \mathbf{1}_{\{G_1, G_2 \leq 4\bar{g}\}}] \geq \mathbb{E}[G^+]/4$ .  $\square$

As a consequence, we merely get an additional factor of 4 in upper bounds on the number of steps necessary for the distance from  $H_{\text{SUM} < 0}$  to drop below  $4\bar{g}$ .

**Theorem 5.** *Let the (1,2) ES minimize  $\text{LINCLIFF}_n^\Delta$  in  $\mathbb{R}^n$  using unit isotropic mutations. Assume that the current search point  $\mathbf{c}$  lies in  $H_{\text{SUM} \geq 0}$  such that  $\text{dist}(\mathbf{c}, H_{\text{SUM}=0}) = O(\mathbb{E}[G^+])$ . Then, independently of  $\Delta$ , after  $3n^{0.4}$  steps w. o. p.  $H_{\text{SUM} \geq 0}$  has been left once and for all.*

*Proof.* Let  $\delta := \text{dist}(\mathbf{c}, H_{\text{SUM} \geq 0})$  in this proof and notice that  $\delta > 0$  implies  $\mathbf{c} \in H_{\text{SUM} < 0}$ . The proof of Lemma 2 directly implies (by choosing  $i = n^{0.1}$ , i. e.  $i^3 = n^{0.3}$ ) that once  $\delta$  has exceeded  $n^{0.3}\bar{g}$ , the local-optimum region  $H_{\text{SUM} \geq 0}$  is never visited again w. o. p., namely with probability  $1 - e^{-\Omega(n^{0.1})}$ . Using a pigeonhole-principle-like argument, we will show that, if  $\delta$  does not exceed  $\bar{g}n^{0.3}$  within at most  $3n^{0.4}$  steps, then w. o. p. there must be at least  $n^{0.1}$  drop-backs (from  $H_{\text{SUM} < 0}$  back into  $H_{\text{SUM} \geq 0}$ ). Consequently, there would also be  $n^{0.1}$  transitions from  $H_{\text{SUM} \geq 0}$  into  $H_{\text{SUM} < 0}$ , and since for each of those there is a  $\Omega(1)$  probability of never dropping back (Lemma 3), those  $n^{0.1}$  drop-backs happen only with probability  $e^{-\Omega(n^{0.1})}$ . Thus, since our assumption “ $\delta$  does not exceed  $\bar{g}n^{0.3}$  within  $3n^{0.4}$  steps” implies the occurrence of an event which does *not* happen w. o. p., this assumption does *not* hold true w. o. p. In other words, w. o. p.  $\delta$  does exceed  $\bar{g}n^{0.3}$  in at most  $n$  steps, finally implying the theorem.

Consider  $2n^{0.3}$  steps, namely the r.v.  $S$  defined as the sum of  $2n^{0.3}$  independent copies of  $G_{2:2}$ . A straight forward application of Hoeffding’s bound (just like the one in the appendix) shows that w. o. p.  $S$  exceeds  $\mathbb{E}[S]/2 = n^{0.3}\mathbb{E}[G_{2:2}] \geq n^{0.3}\bar{g}$ . Thus, right after a step in which  $H_{\text{SUM} \geq 0}$  was left, w. o. p. within at most  $2n^{0.3}$  steps either there is a drop-back or  $d$  exceeds  $n^{0.3}\bar{g}$ . In the latter case we are done; if there is a drop-back, however, the question arises how many steps it takes until the next transition from  $H_{\text{SUM} \geq 0}$  into  $H_{\text{SUM} < 0}$  takes place w. o. p.

Therefore note that  $\mathbf{c}$ ’s distance from  $H_{\text{SUM} < 0}$  right after a drop-back is at most  $n^{0.1}\bar{g}$  w. o. p. Thus, the number of steps until the distance from the cliff drops below  $4\bar{g}$  again is upper bounded by  $4 \cdot 2n^{0.1}$  w. o. p. (a rather loose bound; the factor “4” stems from the lemma preceding the theorem, the factor “2” from considering twice the number of steps that would suffice in expectation to apply Hoeffding’s bound again). Recall that  $\text{dist}(\mathbf{c}, H_{\text{SUM} < 0}) = O(\bar{g})$  implies  $\mathbb{P}\{\mathbf{c} + \mathbf{m} \in H_{\text{SUM} < 0}\} = \Omega(1)$ . Thus, w. o. p. within at most  $n^{0.2}$  steps after a drop-back,  $H_{\text{SUM} \geq 0}$  is left anew (again a rather loose bound since one of  $n^\varepsilon$  trials succeeds already w. o. p.). After this leave it takes w. o. p. at most another  $2n^{0.3}$  steps until either a drop-back occurs again or  $\delta > n^{0.3}\bar{g}$ , and so on. Hence, our initial assumption “ $\delta \leq n^{0.3}\bar{g}$  for  $3n^{0.4}$  steps” finally implies that w. o. p. at least  $3n^{0.4}/(2n^{0.3} + n^{0.2}) \geq n^{0.1}$  drop-backs take place. This was to be shown.  $\square$

We note that the theorem remains true if we substitute “ $3n^{0.4}$ ” by “ $n^\varepsilon, \varepsilon \in \mathbb{R}_{>0}$ ”. Recall that a  $(1+\lambda)$  ES (using unit isotropic mutations) is incapable of conquering the cliff for  $\Delta := 1$ , for instance. It would stay in  $H_{\text{SUM} \geq 0}$  forever and keep on converging towards  $H_{\text{SUM}=0}$  at a declining rate – what a noticeable difference.

### 3 Extension to CLIFF and Gaussians (Extended Outline)

As already noted, when  $\text{LINCLIFF}_n^\Delta$  is minimized, for a fixed  $\Delta$  we can always choose a step length such that also a  $(1+\lambda)$  ES can overcome the cliff in a short time. On the other hand, for a fixed length of an isotropic mutation, there is always a choice for  $\Delta$  disabling a  $(1+\lambda)$  ES from conquering the cliff. One may argue that commonly the length of an isotropic mutation is also random. For instance, the length of a Gaussian mutation  $\widetilde{\mathbf{m}} \in \mathbb{R}^n$  (each component of which is independently standard-normal distributed) follows a  $\chi$ -distribution with  $n$  degrees of freedom. Then arbitrary large lengths are possible. However, since the density of  $|\widetilde{\mathbf{m}}| = \ell$  equals  $\ell^{n-1} \cdot e^{-\ell^2/2} \cdot 2^{1-n/2} / \Gamma(n/2)$  (a unimodal distribution having its mode at  $\sqrt{n-1}$  and inflection points at  $\sqrt{n-1/2 \pm \sqrt{2n-7/4}}$ ), the probability that the length exceeds  $\ell$  drops exponentially for  $\ell \geq \sqrt{3n}$ . In short, the length of a Gaussian mutation is too concentrated, and hence, if  $\Delta$  is by a factor of  $n^\varepsilon, \varepsilon \in \mathbb{R}_{>0}$ , larger than the expected length of a Gaussian mutation, then the probability that a mutation conquers the cliff is exponentially small. An ad hoc solution to this problem could be to choose a different distribution for the length of a mutation to make large step lengths more probable, e. g. a Cauchy distribution. If the lower level sets are bounded (which is *not* the case for  $\text{LINCLIFF}$ ), however, all this is pointless: Steps with immoderate length are vain anyway (they fail to hit the lower level set with high probability).

Therefore, consider the spherically symmetric function  $\text{CLIFF}_n^\Delta: \mathbb{R}^n \rightarrow \mathbb{R}$

$$\text{CLIFF}_n^\Delta(\mathbf{x}) := \begin{cases} |\mathbf{x}| + \Delta(n) & \text{if } |\mathbf{x}| < 1 - \Delta(n), \\ |\mathbf{x}| & \text{otherwise,} \end{cases}$$

where  $\Delta: \mathbb{N} \rightarrow (0, 0.3]$ , introduced by Jägersküpper and Witt (2005). All points in the hyper-sphere  $\{\mathbf{x} \mid |\mathbf{x}| = 1 - \Delta\} \subset \mathbb{R}^n$  are local, non-global optima. The best chances to get over the cliff, however, are at unit distance from the optimum; cf. Jägersküpper (2005). There the ratio of the gain necessary to overcome the cliff (of  $\Delta$  towards the optimum/origin  $\mathbf{o} \in \mathbb{R}^n$ ) to distance from  $\mathbf{o}$  is minimal.

Consider the well-known SPHERE-function ( $\text{SPHERE}(\mathbf{x}) = |\mathbf{x}|^2 = \sum_{i=1}^n x_i^2$ ). For any  $(1+\lambda)$  ES using isotropic mutations there is a distinct normalized (here w. r. t. to the distance from the origin/optimum, *not(!)* w. r. t. to  $n$ ) length of an isotropic mutation resulting in maximum *expected* one-step gain. As we are interested in the number of function evaluations – which equals  $\lambda$  times the number of steps –, we are particularly interested in *constant*  $\lambda$ , i. e.  $\lambda$  is *not* a function of  $n$ . Then the optimum expected one-step gain (progress rate) is  $O(d/n)$  where  $d := |\mathbf{c}|$  equals the distance from the global optimum (d.g.o.). For the  $(1+\lambda)$  ES on SPHERE, an isotropic mutation of length  $\ell = \Theta(d/\sqrt{n})$  results in

an expected gain of  $\Theta(d/n)$ . A (1,2) ES (using isotropic mutations) is incapable of realizing an expected one-step gain of  $\Omega(d/n)$  for SPHERE. However, a straight forward calculation (Jägersküpfer, 2006) shows:

1) For the  $(1,\lambda^*)$  ES with  $\lambda^*$  a constant large enough, isotropic mutations with a length of  $\Theta(d/\sqrt{n})$  result in an expected one-step gain of  $\Theta(d/n)$  on SPHERE.

Now we can follow the reasoning for “the simplest scenario”. Namely, we’d show:

2) For the  $(1,\lambda^*)$  ES using isotropic mutations of fixed length  $\ell := \Theta(d^{[0]}/\sqrt{n})$  there is a  $\Omega(1)$  probability that the d.g.o. never exceeds  $d^{[0]}$ , the initial one.

3) For  $d^{[0]} \in [1 - \Delta; 1 - \Delta + \ell/\sqrt{n}]$  there is a  $\Omega(1)$  probability that the first step conquers the cliff and that the search never drops back to the local optimum region afterwards, i. e.  $\mathbb{P}\{d^{[i]} < 1 - \Delta \text{ for } i \in \mathbb{N}\} = \Omega(1)$ .

4) We’d show that 1), 2), 3) remain true when using Gaussian mutations scaled by a mutation strength  $\sigma \in \mathbb{R}_{>0}$  that is  $\Theta(d^{[0]}/n)$  (we would utilize the concentration of the  $\chi$ -distribution already mentioned at the beginning of this section).

5) When started at a distance, say,  $d^{[0]} \in [1.2, 1.3]$  then w. o. p. after  $t = O(n)$  steps  $d^{[t]} \in [1 - \Delta; 1 - \Delta + \sigma]$  such that 3) applies. After at most  $n^{0.1}$  trials of conquering the cliff within at most  $3n^{0.4}$  steps, the global-optimum region  $\{\mathbf{x} \mid |\mathbf{x}| < 1 - \Delta\} \subset \mathbb{R}^n$  is conquered such that it is never left again w. o. p.

After another  $O(n)$  steps, w. o. p.  $d$  drops below  $0.6 \leq 1 - \Delta - 0.1$ , implying the following result:

**Theorem 6.** *Let a  $(1,\lambda)$  ES minimize  $\text{CLIFF}_n^\Delta$  using Gaussian mutations scaled by a fixed  $\sigma$ . Assume that after initialization  $|c^{[0]}| \in [1.2, 1.3]$  and  $\sigma = \Theta(|c^{[0]}|/n)$ . Then, independently of  $\Delta$ , for  $\lambda$  a constant large enough, the number of steps  $t$  until  $|c^{[t]}| \leq 0.6$  (i. e. the distance from the optimum is halved) is  $O(n)$  w. o. p.*

Since  $\lambda$  is a constant, the  $(1,\lambda)$  ES gets by with  $O(n)$  function evaluation to halve the d.g.o. Finally, compare this with the (1+1) ES on SPHERE: It needs w. o. p.  $\Omega(n)$  function evaluations to halve the d.g.o. even if the length of isotropic mutations would be adapted perfectly in each step! Thus, indeed, the cliff does not keep the  $(1,\lambda)$  ES from halving the d.g.o. within the asymptotically smallest possible number of function evaluations, which is  $\Theta(n)$ .

Since the 1/5-rule (*non-endogenous  $\sigma$ -adaptation*) uses an observation phase of  $\Theta(n)$  steps, and since conquering the cliff takes place in a sub-linear number of steps, we are even able to extend the theorem: When the 1/5-rule is used, the number of CLIFF-evaluations to reduce the d.g.o. to a  $2^{-b}$ -fraction of the initial one is  $O(b \cdot n)$  w. o. p. – wherever the initial starting point lies (given that  $1 \leq b = \text{poly}(n)$  and  $\sigma^{[0]} = \Theta(|c^{[0]}|/n)$ , though). And that’s it.

## References

Arnold, D. (2002): *Noisy Optimization with Evolution Strategies*. Springer.

- Hoeffding, W. (1963): *Probability inequalities for sums of bounded random variables*. American Statistical Association Journal, 58(301):13–30.
- Jägersküpper, J. (2003): *Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces*. In *Proceedings of the 30<sup>th</sup> Int'l Colloquium on Automata, Languages and Programming (ICALP)*, volume 2719 of LNCS, 1068–79, Springer.
- Jägersküpper, J. (2005): *On the complexity of overcoming gaps with isotropic mutations and elitist selection*. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC)*, 206–213, IEEE Press.
- Jägersküpper, J. (2006): *Probabilistic runtime analysis of  $(1+\lambda)$  ES using isotropic mutations*. Accepted for the Genetic and Evolutionary Computation Conference (GECCO).
- Jägersküpper, J., Witt, C. (2005): *Rigorous runtime analysis of a  $(\mu+1)$  ES for the sphere function*. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, 849–856, ACM Press.
- Jansen, T., De Jong, K. A., Wegener, I. (2005): *On the choice of the offspring population size in evolutionary algorithms*. Evolutionary Computation, 13(4):413–440.
- Schwefel, H.-P. (1995): *Evolution and Optimum Seeking*. Wiley, New York.

## Appendix

**Proof of Proposition 1.** Note that  $P\{X \geq 0\} = P\{X \leq 0\} \geq 1/2$  due to the symmetry. As  $X_{2:2} = \max\{X_1, X_2\}$ , where  $X_1, X_2$  are independent copies of  $X$ ,

$$\begin{aligned} E[X_{2:2}] &= E[X_{2:2} \cdot \mathbf{1}_{\{X_1, X_2 \geq 0\}}] + E[X_{2:2} \cdot \mathbf{1}_{\{X_1 \geq 0, X_2 \leq 0\}}] \\ &\quad + E[X_{2:2} \cdot \mathbf{1}_{\{X_1, X_2 \leq 0\}}] + E[X_{2:2} \cdot \mathbf{1}_{\{X_1 \leq 0, X_2 \geq 0\}}]. \end{aligned}$$

The first summand can be bounded from below by

$$\begin{aligned} E[X_{2:2} \cdot \mathbf{1}_{\{X_1, X_2 \geq 0\}}] &\geq E[X_1 \cdot \mathbf{1}_{\{X_1, X_2 \geq 0\}}] \\ &= E[X_1 \cdot \mathbf{1}_{\{X_1 \geq 0\}}] \cdot P\{X_2 \geq 0\} \\ &\geq E[X_1 \cdot \mathbf{1}_{\{X_1 \geq 0\}}] \cdot 1/2. \end{aligned}$$

Analogously, one obtains  $E[X_{2:2} \cdot \mathbf{1}_{\{X_1, X_2 \leq 0\}}] \geq E[X_1 \cdot \mathbf{1}_{\{X_1 \leq 0\}}]/2$  as well as  $E[X_{2:2} \cdot \mathbf{1}_{\{X_i \geq 0, X_{3-i} \leq 0\}}] \geq E[X_i \cdot \mathbf{1}_{\{X_i \geq 0\}}]/2$  for  $i \in \{1, 2\}$ . Altogether,

$$E[X^{2:2}] \geq 3 \cdot E[X \cdot \mathbf{1}_{\{X \geq 0\}}]/2 + E[X \cdot \mathbf{1}_{\{X \leq 0\}}]/2 = E[X \cdot \mathbf{1}_{\{X \geq 0\}}]$$

since  $E[X \cdot \mathbf{1}_{\{X \leq 0\}}] = -E[X \cdot \mathbf{1}_{\{X \geq 0\}}]$  because of the symmetry.  $\square$

Moreover, if  $u > 0$  such that  $E[X \cdot \mathbf{1}_{\{u \geq X \geq 0\}}] \geq E[X \cdot \mathbf{1}_{\{X \geq 0\}}]/2$ , then

$$E[X_{2:2} \cdot \mathbf{1}_{\{X_1, X_2 \leq u\}}] \geq 3 \cdot \frac{E[X \cdot \mathbf{1}_{\{X \geq 0\}}]}{2} / 2 - E[X \cdot \mathbf{1}_{\{X \geq 0\}}]/2 = E[X \cdot \mathbf{1}_{\{X \geq 0\}}]/4.$$

**Additional Calculations for the Proof of Lemma 2.** Recall that here  $G$  corresponds to the spatial gain of a unit isotropic mutation. The r.v. “ $G \cdot \mathbb{1}_{\{|G| \leq \sqrt{i} \cdot \mathbb{E}[G^+] \}}$ ” is also symmetric, and thus, Proposition 1 applies, that is,  $\mathbb{E}[\max\{G_1, G_2\} \cdot \mathbb{1}_{\{|G_1|, |G_2| \leq u\}}] \geq \mathbb{E}[G^+ \cdot \mathbb{1}_{\{|G| \leq u\}}]$ . Thus, it suffices to show that

1)  $\mathbb{E}[G^+ \cdot \mathbb{1}_{\{|G| \leq \sqrt{i} \cdot \mathbb{E}[G^+] \}}] \geq \mathbb{E}[G^+]/2$  for  $i$  large enough.

Recall that the density of  $G$  at  $g \in [-1, 1]$  equals  $(1-g^2)^{(n-3)/2} \cdot \sqrt{n} \cdot (1-\Theta(1/n))$  (for  $n \geq 4$ ). We use  $(1-t/n)^n \leq e^{-t}$  for  $0 \leq t \leq n$ . Then for  $i \in [0, n]$

$$\begin{aligned} \mathbb{E}[G^+ \cdot \mathbb{1}_{\{|G| \leq \sqrt{i/n}\}}] &= \int_0^{\sqrt{i/n}} g \cdot (1-g^2)^{(n-3)/2} dg \cdot 1/\Psi = \\ &= \left[ \frac{(1-x^2)^{(n-1)/2}}{-(n-1)} \right]_0^{\sqrt{i/n}} \cdot 1/\Psi = \left( \frac{1}{n-1} - \frac{(1-(i/n))^{(n-1)/2}}{n-1} \right) \cdot 1/\Psi \\ &= \left( 1 - \underbrace{(1-(i/n))^{(n-1)/2}}_{\leq e^{-(i/n)(n-1)/2}} \right) \cdot \underbrace{\frac{1}{n-1}}_{= \mathbb{E}[G^+]} \cdot 1/\Psi \end{aligned}$$

and  $e^{-(i/n)(n-1)/2} \leq e^{-i \cdot 3/8} < 1/2$  for  $i \geq 2$  (yet  $i \leq n \geq 4$ ; for  $i > n$  the indicator variable becomes meaningless). Thus  $\mathbb{E}[G^+ \cdot \mathbb{1}_{\{|G| \leq \sqrt{2/n}\}}] > \mathbb{E}[G^+]/2$  and, hence, finally  $\mathbb{E}[G^+ \cdot \mathbb{1}_{\{|G| \leq 4\mathbb{E}[G^+] \}}] > \mathbb{E}[G^+]/2$  (since  $\mathbb{E}[G^+] \geq 0.3989/\sqrt{n+1}$ )

2) We want  $\mathbb{P}\{|G| > \sqrt{i/n}\} = e^{-\Omega(i)}$ .

We assume (solely for better legibility) that  $\sqrt{i}$  as well as  $\sqrt{n}$  are integral.

$$\begin{aligned} \mathbb{P}\{|G| > \sqrt{i/n}\} &= 2\sqrt{n} \cdot (1-\Theta(1/n)) \cdot \int_{\sqrt{i/n}}^1 (1-g^2)^{(n-3)/2} dg \\ &\leq 2\sqrt{n} \sum_{k=\sqrt{i}}^{\sqrt{n}} (1-k^2/n)^{(n-3)/2} \cdot \frac{1}{\sqrt{n}} \leq 2 \sum_{k=\sqrt{i}}^{\sqrt{n}} e^{-(k^2/n)(n-3)/2} < 2 \sum_{k=\sqrt{i}}^{\infty} e^{-k^2/8} \end{aligned}$$

Since  $e^{-(k+1)^2/8}/e^{-k^2/8} = e^{-(2k+1)/8} < 1/2$  for  $k \geq 3$ , for  $i \geq 3^2$  we obtain  $\mathbb{P}\{|G| > \sqrt{i/n}\} \leq 2 \cdot 2 \cdot e^{-i/8} = e^{-\Omega(i)}$ .

3) The application of Hoeffding’s bound ...

... to obtain a probability of  $e^{-\Omega(i)}$  that  $9i^2$  steps do *not* suffice to get from  $P_i$  into  $P_{>i}$  (given that in each mutation  $|G| \leq \sqrt{i} \cdot \bar{g}$ , where  $i$  is large enough such that the expected conditional one-step gain is at least  $\bar{g}/2$ ).

Hoeffding (1963, Theorem 2) tells us that for the r.v.  $S$  defined as the sum  $X_1 + \dots + X_k$  of  $k$  independent r.v.s  $X_j \in [a_j, b_j]$  for  $j \in \{1, \dots, k\}$  we have  $\mathbb{P}\{S \leq \mathbb{E}[S] - t\} \leq e^{-2 \cdot t^2 / \sum_{j=1}^k (b_j - a_j)^2}$  for  $t \geq 0$ . In our case,  $k := 9i^2$  so that  $\mathbb{E}[S] \geq 4.5 i^2 \bar{g}$ , and furthermore,  $a_j = -\sqrt{i} \cdot \bar{g}$  and  $b_j = \sqrt{i} \cdot \bar{g}$ . Since the necessary gain is at most  $w_i = \bar{g} \cdot (3i^2 + \Theta(i)) \leq 4 i^2 \bar{g}$  for  $i$  large enough, we can choose  $t := 0.5 i^2 \bar{g}$ . Thus, the exponent becomes  $-2 \cdot (0.5 i^2 \bar{g})^2 / \sum_{j=1}^{9i^2} (2\sqrt{i} \bar{g})^2 = -i/72$ .