Handling Sharp Ridges with Local Supremum Transformations

Tobias Glasmachers
Institut für Neuroinformatik
Ruhr-Universität Bochum, Germany
tobias.glasmachers@ini.rub.de

ABSTRACT

A particular strength of many evolution strategies is their invariance against strictly monotonic and therefore rank-preserving transformations of the objective function. Their view onto a continuous fitness landscape is therefore completely determined by the shapes of the level sets. Most modern algorithms can cope well with diverse shapes as long as these are sufficiently smooth. In contrast, the sharp angles found in level sets of ridge functions can cause premature convergence to a non-optimal point. We propose a simple and generic family of transformation of the fitness function to avoid this effect. This allows general purpose evolution strategies to solve even extremely sharp ridge problems.

Categories and Subject Descriptors

[**Evolution Strategies and Evolutionary Programming**]

General Terms

Algorithms

Keywords

Evolution strategies, Robust Optimization, Ridge Functions

1. INTRODUCTION

In principle, evolution strategies (ES) are applicable to an extremely wide variety of optimization problems: with their direct search paradigm they interface the objective function as a black-box, and they do not presume any specific structure or even smoothness. Nowadays there is a rather modular repertoire of techniques available for handling different types of difficulties by means of online adaptation. The most prominent of these techniques is covariance matrix adaptation [8] which remedies otherwise slow convergence into ill-conditioned optima. Other challenges posed by multi-modality [2], fitness noise [7], and black-box constraints [1] have been approached with good success.

Here we focus on the challenge of non-smooth objectives. Although capable of handling non-smooth and even discontinuous objectives in principle it is well known that ES may fail on such problems. This was demonstrated recently for the extremely difficult "HappyCat" problem [4]. Since we are far from understanding the complete picture of which cases can and cannot be handled successfully by evolution strategies present analysis is limited to prototypical test cases such as ridge functions. We follow this approach in the present paper.

Our aim is to advance the optimization capabilities of evolution strategies (and possibly other optimizers) when facing difficult, non-smooth objectives. To this end we propose a rather generic technique that is based on performing a specific transformation to the objective function. This transformation alleviates the difficulty of sharp angles found in the shapes of non-smooth level sets. We make sure that the optimum of the untransformed objective can be located with arbitrarily good precision.

Ridges.

Ridge functions such as

$$f^{\text{ridge}}_{\alpha,d} : \mathbb{R}^N \to \mathbb{R}\,; \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \mapsto x_1 + d \cdot \left(\sum_{i=2}^{N} x_i^2 \right)^{\alpha} \quad (1)$$

are a class or rather well investigated problems (see [9] for a systematic investigation for evolution strategies with isotropic search distributions, as well as [4] and references therein for an up-to-date discussion). They have been of some interest since they can be demonstrated to systematically misguide search strategy adaptation rules to the point where optimization breaks down completely. It should be noted that many gradient-based search schemes are facing similar problems, however, they are often considered inapplicable to non-smooth problems in the first place while ES remain a viable alternative.

In particular on sharp[1] ridges with $\alpha < 1/2$ the core mechanism of step size control is susceptible to cause premature convergence. This behavior can be understood qualitatively from the fact that rank-based algorithms "perceive"

[1]We refer to the cases $\alpha < 1/2$, $\alpha = 1/2$, and $\alpha > 1/2$ as sharp, linear, and smooth ridges, respectively. Note that in some previous work the case $\alpha = 1/2$ is already considered "sharp".

GECCO'14, July 12–16 2014, Vancouver, BC, Canada.

ACM 978-1-4503-2662-9/14/07 ...$15.00.
http://dx.doi.org/10.1145/2576768.2598215.

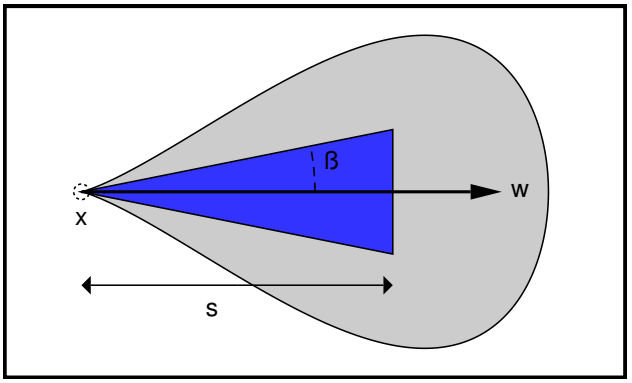

Figure 1: Illustration of the definition of the angle-based hardness measure for (sub-)level sets. The figure shows a non-smooth, connected level set with the corresponding sub-level set shaded in gray. The dark blue cone realizes the largest opening angle in the corner.

the objective landscape in terms of (sub-)level sets only, ignoring absolute function values.

Level sets and sub-level sets of ridge functions are connected but non-smooth. Their problem hardness or deceptiveness can be measured, e.g., by the inner angle of sub-level sets. We define the smallest inner angle of a sub-level set as the smallest opening angle of an infinitesimal cone that fits into the set. More precisely, let $w \in \mathbb{R}^N$ be a unit vector, $s > 0$, $\beta \in (0, \pi/2)$ and $x \in \mathbb{R}^N$ then we define the cone

$$C_{x,s,w,\beta} = \left\{ x + p \;\middle|\; p \in \mathbb{R}^N, \frac{\|p - (w^T p)w\|}{\tan(\beta)} \leq w^T p \leq s \right\}$$

at x with principal axis w, length s and opening angle β. We define the smallest opening angle of the set M as

$$\measuredangle(M) = \inf_{x \in M} \lim_{s \to 0} \sup_{w \in S^{N-1}} \left\{ \beta \in (0, \pi/2) \;\middle|\; C_{x,s,w,\beta} \subset M \right\}$$

where S^{N-1} denotes the unit sphere in $\mathbb{R}^N$, and for the positivity of β we define the infimum of an empty set as zero. Consequently we define the (maximal) hardness of an objective function $f : \mathbb{R}^N \to \mathbb{R}$ (to be minimized) as

$$\measuredangle(f) = \inf_{v \in \mathbb{R}} \measuredangle\big(\{x \in \mathbb{R}^N \mid f(x) < v\}\big) \ .$$

Figure 1 illustrates the concept.

For example, it holds $\measuredangle(f^{\text{ridge}}_{\alpha,d}) = \pi/2$ for smooth ridges with $\alpha > 1/2$, $0 < \measuredangle(f^{\text{ridge}}_{\alpha,d}) = \tan^{-1}(1/d) < \pi/2$ for linear ridges with $\alpha = 1/2$, and $\measuredangle(f^{\text{ridge}}_{\alpha,d}) = 0$ for sharp ridges with $\alpha < 1/2$.

Any positive angle can be opened up as widely as necessary by means of a linear transformation and therefore essentially by a CMA mechanism. Thus, an ES with CMA may handle ridges with $\alpha = 1/2$ for any value of d when given sufficient time to adapt the covariance matrix.

The same approach does not solve the problem for the arbitrarily sharp angles of level sets for $\alpha < 1/2$. The probability of sampling better offspring quickly decays to zero as the search distribution approaches the ridge. Consequently success-based step size adaptation (e.g., by means of the classic 1/5 rule [10]) results in premature convergence. At the same time there is a trend for the most successful offspring to be sampled close to the ridge. This effect can result in convergence of estimation of distribution style approaches, including the cumulative step size control mechanisms used in many modern non-elitist ES (see e.g. [8]).

Furthermore, the problem can be made even harder by bending the ridge and by replacing the linear trend term x_1 along the ridge with a quadratic term (e.g., x_1^2). Then the relative impact of the trend along the ridge vanishes when approaching the optimum not only towards but also along the ridge. These two additional challenges have been combined in the so-called HappyCat benchmark problem [4] that turns out notoriously hard to solve for a number of well-established direct search algorithms.

Smoothing Difficult Level Sets.

In this paper we propose an approach to handling difficult shapes of level sets by optimizing a parameterized family of approximate objectives with more regularly shaped level sets. Since the question of what makes shapes of level sets difficult for optimization is hard to answer in general we focuses our analysis on the important aspect of ridges, which covers a large share of the difficulties arising in continuous, non-smooth objective functions.

Our approach does not tackle the problem by proposing novel search strategies tailored to the handling of ridges. In particular we do not touch existing step size and covariance matrix adaptation mechanisms. Instead our method can be understood as a family of transformations of the objective function, making the method generically applicable in principle to nearly every existing search algorithm. The only addition required to this search algorithm is a control mechanism for the parameter of the family of transformations, which can be set up as an additional outer loop.

In the following section we propose the aforementioned family of transformations and establish general properties that are of importance for optimization in general and for ES in particular. Then we investigate the impact of the transformations on different types of level sets, with a focus on ridge functions. We turn these insights into a meta optimization algorithm that can embrace many different types of optimizers as a module. Its performance is then evaluated on a number of standard benchmarks as well as on ridge functions of varying difficulty. We demonstrate how our approach enables an ES with CMA to successfully solve not only "standard" ridges of arbitrary difficulty but also more difficult variants. We close with our conclusions.

2. LOCAL SUPREMUM TRANSFORMS

In this section we introduce a number of non-linear transformations of objective (or fitness) functions. Let $f : \mathbb{R}^N \to \mathbb{R}$ be the objective function to be optimized. Without loss of generality we consider minimization in the following. For non-empty $\Delta \subset \mathbb{R}^N$ we define the new function

$$F_{f,\Delta}(x) = \sup \left\{ f(x + \delta) \;\middle|\; \delta \in \Delta \right\} \ .$$

We call $F_{f,\Delta}$ a *supremum transform* of the original fitness f. For maximization of f we replace the supremum with the infimum. Of course, in special cases such as for finite Δ or for continuous f and compact Δ the supremum is guaranteed to be attained and the supremum actually becomes a maximum.

The function $F_{f,\Delta}$ has a simple interpretation: it computes the worst case fitness over the set $x+\Delta = \{x+\delta \,|\, \delta \in \Delta\}$. For $\Delta = \{0\}$ it holds $F_{f,\Delta} = f$. If Δ is a small neighborhood of the origin then we can think of $F_{f,\Delta}$ as a "pessimistic" approximation to f, in the sense that minimizing $F_{f,\Delta}$ reveals the solution with best worst case behavior over its Δ-shaped neighborhood.

It is intuitive that for small enough Δ and under certain assumptions optimizing f and $F_{f,\Delta}$ will give similar results. The quality of the approximation will typically increase as Δ shrinks. We are interested in approximating an optimum of f, which can be achieved by optimizing $F_{f,\Delta}$ for a sequence of sets Δ with increasing concentration around the origin. For the time being we fix this approach and postpone the discussion why this may make sense to section 3. The simplest way to make the proceeding explicit is by means of an additional scaling parameter. For a fixed, bounded neighborhood Δ of the origin, all scaled versions $s \cdot \Delta$, $s > 0$, are also bounded neighborhoods of the origin. We define the family

$$F_{f,\Delta,s}(x) = \sup\left\{ f(x + s \cdot \delta) \,\middle|\, \delta \in \Delta \right\} \ .$$

of *local supremum transforms (LSTs)* of f. We consider the set Δ fixed and vary only the scale parameter $s \geq 0$. Similar to filters in topology the emphasis in on the fact that for $s \to 0$ the sets $x + s \cdot \Delta$ over which the suprema are taken become arbitrarily concentrated around x.

Such families of LSTs have a number of properties with relevance in an optimization context. In the following we assume that the set Δ is bounded and contains the origin. Then it holds:

1. For a scale of $s = 0$ the transformation reduces to the original objective: $F_{f,\Delta,0} = f$.

2. For a continuous function f it holds

   $$\lim_{s \to 0} F_{f,\Delta,s}(x) = f(x) \ ,$$

   for all $x \in \mathbb{R}^N$, i.e., the family $F_{f,\Delta,s}$ of functions converges to the original objective in a point-wise manner. Furthermore, each $F_{f,\Delta,s}$ is a continuous function.

3. For a uniformly continuous function f it holds

   $$\lim_{s \to 0} F_{f,\Delta,s} = f$$

   with convergence in the maximum norm topology, i.e., the family $F_{f,\Delta,s}$ of functions converges uniformly to the original objective. Furthermore, each $F_{f,\Delta,s}$ is uniformly continuous. Note that this property holds automatically for continuous fitness f restricted to a compact subset of $\mathbb{R}^N$.

4. Assume that f is continuous and unimodal and that all of its level sets are bounded. Then the set X^* of optimizers of f as well as all sets $X^*_{f,\Delta,s}$ of optimizers of $F_{f,\Delta,s}$ are non-empty. Let $d(\cdot,\cdot)$ denote Euclidean distance between compact sets (with points as a special case). Then it holds

   $$\lim_{s \to 0} \ \max\left\{ d(x, X^*) \,\middle|\, x \in X^*_{f,\Delta,s} \right\} = 0 \ .$$

5. For a convex function f also all functions $F_{f,\Delta,s}$ are convex.

6. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a strictly monotonically increasing function. Then it holds

   $$F_{\phi \circ f,\Delta,s} = \phi \circ F_{f,\Delta,s} \ .$$

Proofs of these elementary properties are found in the supplementary material. The first property is of limited value when applying $F_{f,\Delta,s}$ for optimization. However, properties 2 and 3 indicate that shrinking s to zero over the course of the optimization turns the objective function gradually into the objective of original interest. This is useful when we are interested in solutions of approximately optimal quality. It is property 4 that ensures that minimization of LSTs of f results in convergence to a minimizer of f. Property 5 is a trivial consequence of convexity. Its importance lies in the fact that the LSTs do not destroy the polynomial time complexity of a convex optimization problem.

The last property is relevant in the context of rank-based algorithms, including most evolution strategies. It ensures that $F_{f,\Delta,s}$ is compatible with rank-based selection since it is invariant under rank-preserving transformations.

The supremum transformed fitness $F_{f,\Delta}$ is loosely connected to robust optimization [3] where the task is to find the solution with the best worst case behavior over a region of parameters that result, e.g., from parameter uncertainty. The uncertainty is often of a more complex nature than a Δ-shaped region, e.g., when parameters of constraints are uncertain. Our situation is quite different. We assume that values of f are reliable, thus we are not at all interested into optimizing worst case behavior. This allows us to drive s to zero, which does not make any sense in a robust optimization setting since uncertainty is not reduced during the optimization run.

3. ANALYSIS OF LEVEL SETS

The view of a direct search optimizer with invariance under rank-preserving transformations of objective values on the fitness landscape is completely determined by the shapes of the level sets. In this section we will investigate the effect of the (local) supremum transform on the shape of level sets. We will analyze three different models, namely the smooth, convex level sets of the ellipsoid function and the level sets of the ridge function (1) for $\alpha = 1/2$ and for $\alpha < 1/2$. For simplicity we assume that Δ is the closed unit ball $\Delta = \left\{\delta \in \mathbb{R}^N \,\middle|\, \|\delta\| \leq 1\right\}$. Geometrically this means that the level sets of $F_{f,\Delta,s}$ originate from level sets of f but "inwards shifted" towards the optimum by a distance of s and possibly cut off appropriately. This will become more clear in the examples below.

Ellipsoid.

The standard ellipsoid function takes the form

$$f^{\text{elli}}_{\alpha}(x) = \sum_{i=1}^{N} \alpha^{\frac{i-1}{N-1}} x_i^2 \ .$$

For $\alpha = 1$ we obtain the sphere function $f^{\text{sphere}}(x) = \|x\|^2$. It is easy to see that it holds $F_{f,\Delta,s}(x) = (\|x\| + s)^2$ by means of which the level sets of $F_{f,\Delta,s}$ turn out to be also spheres around the origin. They originate from moving the level sets "inwards" by s. In other words even for fixed $s > 0$ optimization of f and $F_{f,\Delta,s}$ are equivalent.

The situation changes as we change the parameter α and with it the difficulty of the problem. A standard choice

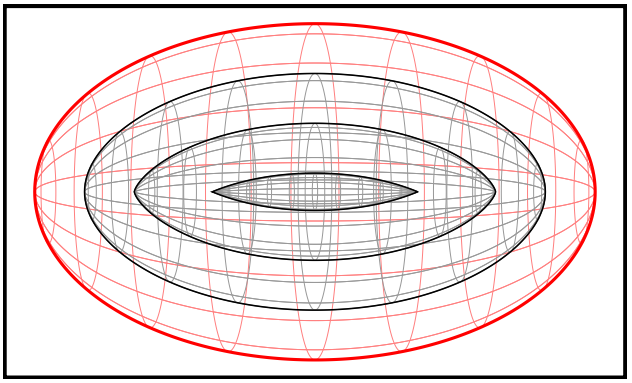

Figure 2: A prototypical level set of the ellipse function (outermost, red), with three inwards shifted level sets of LSTs. The angle $\measuredangle$ decreases from outer to inner (growing s), corresponding to increasing problem hardness.

for benchmarking is $\alpha = 10^{-6}$, resulting in ellipsoidal level sets with shortest principal axis 1000 times shorter than the longest principal axis. This is a linear transformation of the sphere that can be corrected for with a CMA mechanism. However, moving such a level sets inwards by a radius of s first results in ellipsoid-like shapes with even worse ratio of longest and shortest principal axis, and from some point on the outline becomes non-smooth. This effect is illustrated in figure 2. A level set of f with smallest principal axis length of $s + \epsilon$ shows up as a corresponding level set of $F_{f,\Delta,s}$ with shorted axis length of only ϵ. For $\epsilon \to 0$ the level set becomes arbitrarily eccentric, with increasing difficulty $\measuredangle \to 0$. Thus for a fixed value of s, when minimizing $F_{f,\Delta,s}$ we cannot expect to get systematically much closer to the optimum than reaching a level set with a smallest principal axis of about s. Of course, shrinking s to zero will allow a reasonable optimizer to make further progress and to locate the optimum with arbitrary precision.

Ridge.

For $\alpha > 1/2$ the level sets of the ridge function (1) are smooth. We do not consider this situation further since it is essentially covered by the ellipse case. For $\alpha = 1/2$ the level sets turn into cones with a singularity right on the ridge. The parameter d controls the opening angle $\measuredangle$ of these cones. For large d the angle decays towards zero which makes following the linear trend along the ridge increasingly hard. Since for fixed d the angle is always positive it can be opened up with a simple linear transformation which makes CMA mechanisms suitable for solving this type of problem.

Similar to the sphere, shifting the cone-shaped level sets towards the ridge does not change their shapes, see also figure 3. Therefore optimization of f and $F_{f,\Delta,s}$ are equivalent.

Sharp Ridge.

For $\alpha < 1/2$ the level sets of the ridge undergo a decisive change of shape: when approaching the ridge the opening angle quickly approaches zero. Figure 4 illustrates this property. In this situation CMA mechanisms can turn out to be insufficient if the search approaches the ridge faster

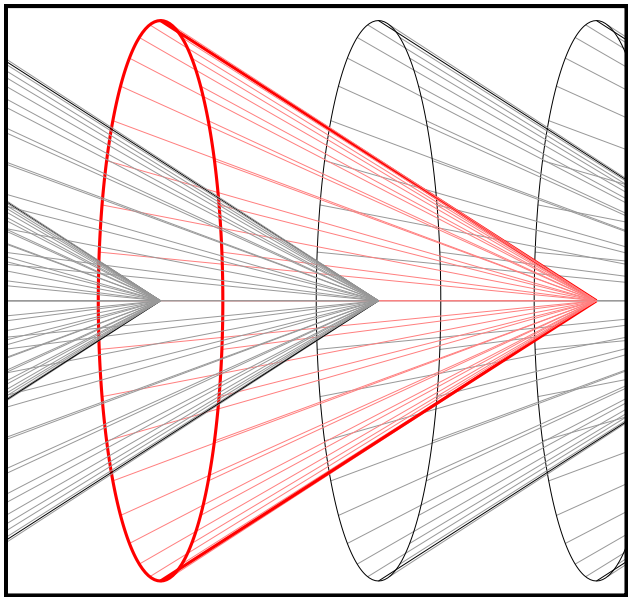

Figure 3: Level sets of the ridge function $f^{\mathbf{ridge}}_{\alpha,d}$ for $\alpha = 1/2$ and $d = 1$. The level sets are cone shaped. The angle of the cone singularity becomes smaller for increasing values of the problem hardness parameter d. The level sets of all LSTs coincide with those of the ridge function.

than CMA can open up the angle, possibly resulting in premature convergence.

Interestingly, shifting the level sets of the sharp ridge inwards by any fixed distance $s > 0$ leaves us with a positive opening angle which can easily be opened up further with CMA, see figure 5. It is obvious that this property makes supremum transformations extremely valuable for the optimization of ridge functions.

On the sharp ridge the family of LSTs has a regularizing effect in a sense similar to, e.g., Tikhonov regularization: the problem of optimizing f is "ill-posed" (degenerate level sets), while for each fixed $s > 0$ optimizing $F_{f,\Delta,s}$ is "well-posed" (positive opening angle of the level set), and the solution of the ill-posed problem can be found by taking the limit $s \to 0$ of the solutions of the well-posed problems. Importantly, taking limits on the levels of problems and solutions commutes.

4. A META OPTIMIZATION STRATEGY

In this section we discuss a practical implementation of an optimization strategy based on LSTs. This is a rather conceptual algorithm, leaving many opportunities for improvement. At this point we are mostly interested in the difference between premature convergence and convergence to the optimum, which is a binary observable. The algorithm is by no means tuned for efficiency in terms of minimal number of queries to the black box fitness f.

The first question is for the shape of the set Δ. A ball shape is attractive for a number of reasons, including mathematical simplicity and invariance properties. However, this renders the practical evaluation of $F_{f,\Delta,s}$ intractable. Therefore we resort to a finite approximation. A reasonable set Δ should satisfy a few minimal requirements: its convex

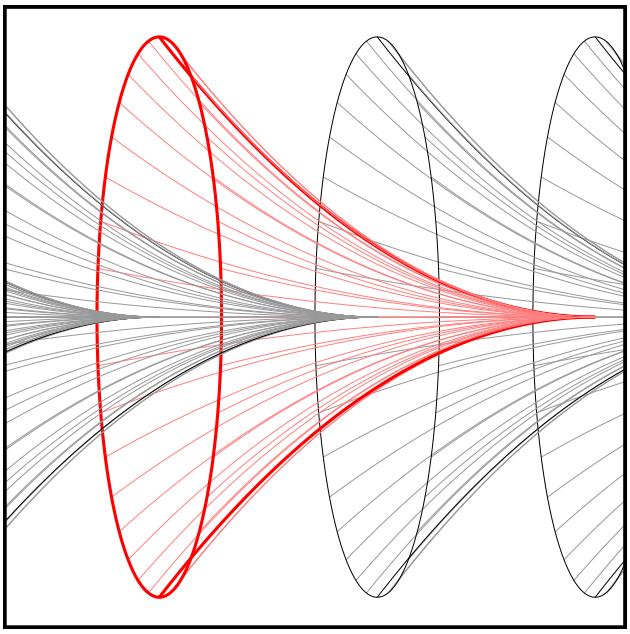

Figure 4: Level sets of the sharp ridge with parameter $\alpha = 1/4 < 1/2$. The sets originate (for $N = 3$) from rotating one branch of a parabula around the x_1 axis. Thus the opening angle of the sub-level set in the singularity becomes arbitrarily small when approaching the ridge. It is not possible to change this property by means of a linear transformation. Thus CMA techniques cannot fully resolve this issue.

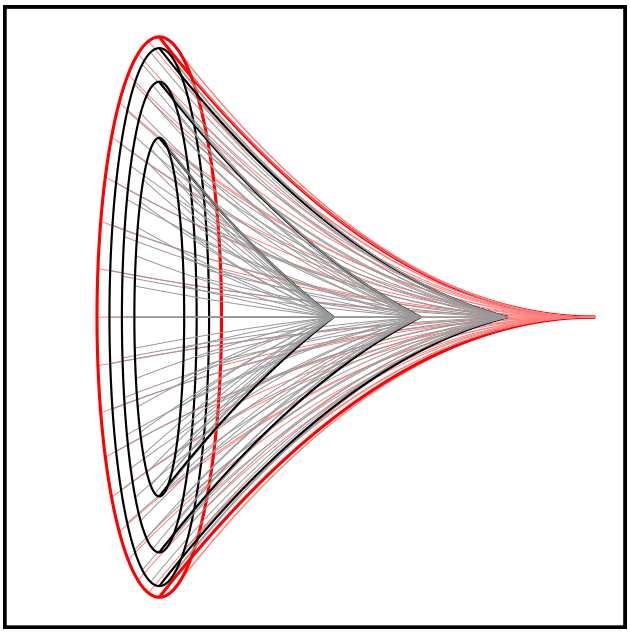

Figure 5: Level set of the sharp ridge with parameter $\alpha = 1/4$ (outermost, red), as well as three level sets of LSTs. Notice that the opening angle $\measuredangle$ of the singular point is positive for all $s > 0$. It increases (decreasing problem hardness) for growing s.

hull should be of full dimension, and it should contain the origin. Any such set is of size at least $N + 1$. One may further require symmetry of some sort as well as centering in some sense, e.g., with the origin forming the center of mass. Our intuition is that with these requirements fulfilled the exact choice of Δ should not play too much of a role. A simple choice fulfilling these requirements is

$$\Delta = \{0\} \cup \bigcup_{i=1}^{N} \{e_i, -e_i\}$$

where $e_i = (0, \ldots, 1, \ldots, 0) \in \mathbb{R}^N$ denotes the i-th unit vector. Due to $|\Delta| = 2N + 1$, the complexity of evaluating $F_{f,\Delta,s}$ is increased by a factor of $2N + 1 \in \mathcal{O}(N)$ as compared to the original fitness function f.

After fixing Δ it is straightforward to apply any black box optimizer A to the supremum transformed objective $F_{f,\Delta,s}$ for some fixed $s > 0$. Even if successful this will not result in the exact optimizer of f. Instead we need to drive s towards zero during the optimization run. This can be achieved trivially in an outer loop around the black box optimizer that shrinks s according to a predefined schedule. Starting with a rather large s and halving it in every iteration until some target accuracy is reached gives a straightforward implementation. Starting with large s is in accordance with a large initial variance of the search distribution so as to initially "cover" a large enough part of the search space.

It is worth noting that this proceeding requires the definition of a stopping criterion for the black box optimizer A after which the scale s can be reduced and the search can start over, typically starting from the position of the best-so-far solution. In many cases it is possible to leave some state variables of A in place when transitioning to a smaller value of s while other variables will need updating. This wrapper approach is formalized in algorithm 1.

Algorithm 1: Wrapper algorithm for LST optimization.

Input: $s_0 > s_{\min} > 0$, search algorithm A
$s = s_0$
repeat
  update/reset state of A for fitness $F_{f,\Delta,s}$
  repeat
    perform step of A with fitness $F_{f,\Delta,s}$
  until *stopping criterion of A is met*
  $s \leftarrow s/2$
until $s < s_{min}$

At this point we specialize our considerations to an evolution strategy with adaptive step size parameter σ. This ES may additionally employ a CMA mechanism. We stop the inner loop as soon as the step size σ drops below then s-dependent threshold $c \cdot s$. Setting $c = 10^{-5}$ has proven sufficient in our experiments. Then the optimizer is reinitialized. Its position parameter (e.g., the parent in a (1+1)-ES) stays as is. Also the covariance matrix can be kept (or it may be regularized in order to undo adaptation for fine tuning). The step size σ is reset to the new value of s. Elitist strategies must reevaluate all parents since the fitness function has effectively changed. It is obvious that the concrete measures depend on the underlying algorithm and that many variations are possible.

N	2	4	8	16	32
xNES	339	1,064	3,985	16,374	70,084
LST wrapper	480	1,608	5,250	19,092	75,852

Table 1: Number of function evaluations for xNES with plain fitness and the xNES wrapper algorithm using local supremum transforms (LST) for the Ellipsoid function in different dimensions N. Note that supremum evaluations correspond to $2d+1$ black box fitness evaluations each. The optimum was successfully located within the target accuracy in all cases.

5. EXPERIMENTS

In this section we evaluate LST optimization empirically. We base our evaluation on the cases that have already been investigated in section 3. However, in this section we replace the ball-shaped set Δ with the finite version defined in the previous section. In addition we consider the extremely difficult ridge function known as HappyCat [4].

We implemented the wrapper algorithm described in the previous section with the xNES evolution strategy [6] as its inner loop optimization algorithm. We compare this approach to plain xNES without LST. xNES shares many properties with the well-known CMA-ES algorithm [8] with rank-μ update, but without evolution paths. In particular it adapts the covariance of its Gaussian search distribution to the problem at hand.

Our setup is as follows. Both optimizers are initialized with an isotropic search distribution with unit covariance matrix. The center of the distribution is randomly sampled in each run the from standard normal distribution.[2] For the wrapper algorithm the scale parameter is initialized to the same value as the global step size of xNES, in other words to $s_0 = 1$.

The algorithms are stopped as soon as a (problem dependent) target fitness value is reached. Premature convergence was detected by monitoring the smallest eigenvalue of the covariance matrix. If this value falls below the numerical accuracy of IEEE 64bit floating point ("double precition") numbers that were used for all experiments then the algorithm has converged prematurely.

All experiments were repeated 100 times. All numbers reported in this sections are medians over 100 independent runs.

5.1 Ellipsoid

It was shown in section 3 that for convex (sub-)level sets the supremum transform can increase problem difficulty. Here we investigate the severity of this effect. Therefore we have run the above algorithm on the ellipsoid problem with $\alpha = 10^{-6}$ in dimensions $N \in \{2, 4, 8, 16, 32\}$. The target fitness was set to 10^{-10}. The number of function evaluations is reported in table 1.

The difference between the two algorithms is surprisingly small. The LST wrapper requires between 10% to 75% more individuals to locate the optimum.[3] This is despite the fact that it can runs the inner optimizer several times for different scales s. We can conclude that the irregular level sets of the supremum transformed ellipsoid problem do not make the problem significantly more difficult.

[2]For the HappyCat problem (see below) the distribution refers to the original problem formulation. In other words it is shifted by $(1, \ldots, 1)^T$ in our experiments.

[3]The difference in number of black box queries grows linearly with problem dimension since the size of the set Δ does.

d	10^3	10^4	10^5	10^6	10^7	10^8	10^9
xNES	98	86	53	53	18	18	1
LST wrapper	100	100	100	100	100	100	100

Table 2: Number of successful runs out of 100 for both algorithms for the sharp ridge problem with $\alpha = 1/8$ in $N = 10$ dimensions for growing values of the parameter d, corresponding to increasing problem difficulty.

5.2 Linear Ridge

For the ridge function (1) with $\alpha = 1/2$ both algorithms reliably manage to reach the target fitness value of -10^6 for virtually any value of the parameter $d > 0$. We have tested values up to one billion, where both algorithms are still capable of escaping the drag of the ridge in all 100 out of 100 runs.

5.3 Sharp Ridge

We perform a test similar to the previous one with the ridge function for $\alpha = 1/8$. The resulting ridge is sharp in the sense that the opening angles of the level sets become arbitrarily small when approaching the ridge. In successful runs the algorithms manages to quickly reach a target value of -10^6, while prematurely convergent runs usually get stuck far before reaching the target value. The number of successful runs out of 100 is reported in table 2.

The performance of default xNES slowly degrades with growing problem difficulty. Due to the analysis of section 3 and with the results on the linear ridge it is not surprising that the LST wrapper algorithm manages to solve the sharp ridge problem reliably even for extremely difficult instances with d as large as a billion. In other words, the theoretically predicted advantage is fully observable in practice.

5.4 Ridge with Corner

Until now both algorithms did profit from the symmetry of the test problems. For ellipse and ridge the optima of all supremum transforms for sufficiently symmetric Δ coincide with the optimum of the fitness, making it in principle unnecessary to shrink the scale parameter s (of course, the shape of the level sets of the ellipse function creates a necessity from a practical point of view). Now we consider less symmetric case for which the optima of $F_{d,\Delta,s}$ are indeed shifted away from the optimum of f

Further on, the above test problems do require an adaptation of the covariance matrix, but once the adaptation phase was successful no further adaptation is required. Next we test both algorithms on more difficult problems that require continuous adaptation of the covariance matrix.

We introduce a 90° corner into the ridge:

$$f^{\text{corner}}_{\alpha,d}(x) = x_1 + d \cdot \left((x_2 - |x_1|)^2 + \sum_{i=3}^{N} x_i^2 \right)^{\alpha}$$

Figure 6 illustrates this problem.

For this benchmark we sample the starting point with unit variance around $(10, 10, 0, \ldots, 0) \in \mathbb{R}^N$ (i.e., close to

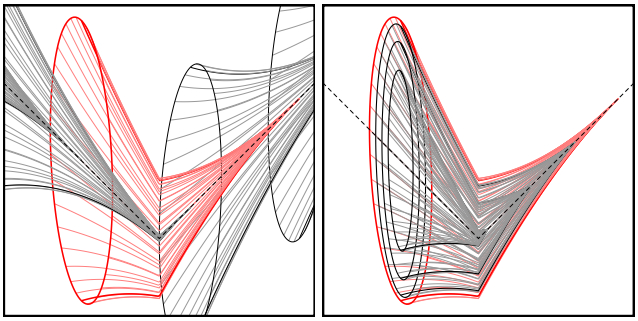

Figure 6: Left: level sets of the sharp ridge with corner with parameter $\alpha = 1/4$. Right: level sets of corresponding LSTs. The ridge is indicated by the dashed line.

d	10^1	10^2	10^3	10^4	10^5	10^6	10^7
xNES	100	1	0	0	0	0	0
LST wrapper	100	27	78	75	17	5	0

Table 3: Number of successful runs out of 100 for each of the algorithm for the sharp ridge with corner with $\alpha = 1/8$ in $N = 10$ dimensions for growing values of the parameter d, corresponding to increasing problem difficulty.

the ridge on the "worse" side of the corner). The additional difficulty of this problem is that a covariance matrix that is well adapted to the ridge hampers sampling a successful offspring around the corner. This effect becomes more pronounced with increasing eccentricity of the search distribution. Thus CMA algorithms may get stuck at the corner.

The results for both algorithms are found in table 3. The problem turns out to be much harder than without corner. Plain xNES breaks down already for rather low values of d. Compared to the sharp ridge without corner the performance of the LST wrapper algorithm is far less stable, but obviously the LST technique helps a lot also in this case. However, for extremely high values of d it fails completely.

5.5 HappyCat

The benchmark problem named HappyCat [4]

$$f_\alpha^{\mathrm{hc}}(x) = \left[\left(\|x\|^2 - N\right)^2\right]^\alpha + \frac{1}{N}\left(\frac{1}{2}\|x\|^2 + \sum_{i=1}^{N} x_i\right) + \frac{1}{2}$$

is composed of a spherical ridge and a quadratic trend along this ridge. The naming of the function is due to its level lines for dimension $N = 2$, see figure 7. This is a unimodal function with unique optimum at $x^* = (-1, \ldots, -1)$, with optimal value $f^* = f(x^*) = 0$. For $\alpha = 1$ the function is a fourth order polynomial, but for the default value of $\alpha = 1/8$ the ridge is rather sharp, while the quadratic trend within the ridge vanishes when approaching the optimum.

The visible trend towards the optimum decays while the impact of the ridge remains constant. Thus, this benchmark effectively tests the ability to solve the sharp ridge problem continuously for ever increasing values of d. At the same time the ridge is curved which requires a continuous adaptation of the covariance matrix. Both effects are active at the same time, making the problem extremely hard for most direct search algorithms.

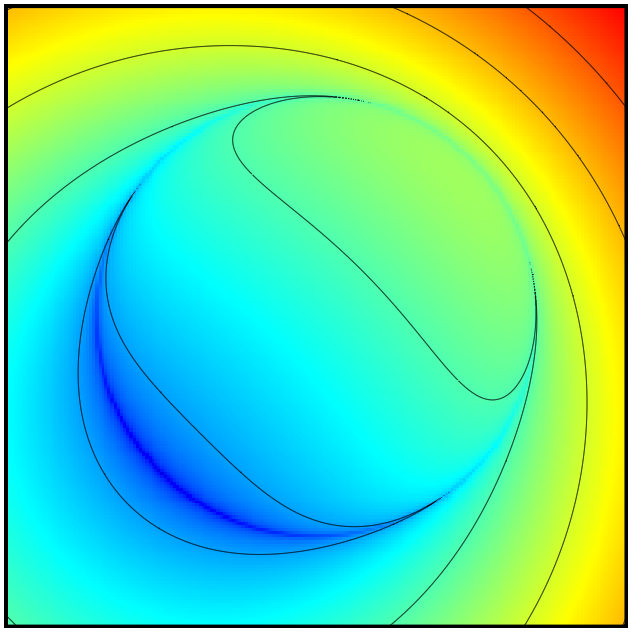

Figure 7: Level lines of the HappyCat benchmark function for $\alpha = 1/8$ in the area $[-2, +2]^2 \subset \mathbb{R}^2$. The picture arising from the level lines explains the name. The lines for levels $f = 1$ and $f = 2$ exhibit sharp angles. They remain nearly parallel while the curvature of the ridge becomes already visible. This gives an idea of the hardness of the problem once the optimizer has arrived at the ridge.

In the original work [4] on the HappyCat function an ES-style algorithm based on line search (the so-called "Ray-ES") has been proposed as a possible conceptual solution. However, this algorithm is obviously tailored to the task of following a narrow valley or ridge while its performance on other objective was not evaluated. Here we test whether the supremum transform in conjunction with a general purpose ES can serve as an alternative algorithm for solving the HappyCat problem.

With the above formulation of the HappyCat function one quickly observes numerical problems. Assume some x close to x^* is represented with standard 64bit IEEE double precision floating point numbers (52 bits mantissa, 11 bits exponent, one sign bit). Then a distance of, say, $10^{-20} \approx 2^{-66}$ of x from the ridge is already far below the numerical resolution. Due to the small value of $\alpha = 1/8$ such a deviation results in a considerable cost of about $(10^{-20})^{2\alpha} = 10^{-5}$. For $N = 5$ the same penalty is received with the quadratic function in a distance of about $\sqrt{2N \cdot 10^{-5}} = 10^{-2}$ from the optimum. An ES with Gaussian sampling is essentially unable to avoid the minimal deviation of 10^{-20} while making steps in the order of 10^{-2} along the ridge, again because numerical accuracy limits the conditioning (multiplicative difference between eigen values) of the covariance matrix to about 10^{15}. Thus, with an ES with Gaussian sampling we cannot expect to obtain solutions systematically much closer than about 10^{-2} to the optimum, which is a poor resolution.

To alleviate the issue we shift the problem by $(1, \ldots, 1)$, moving the optimum x^* into the origin. The shifted func-

N	2	3	5	10
(1+1)-xNES	$1.89 \cdot 10^{-1}$	$2.38 \cdot 10^{-1}$	$4.49 \cdot 10^{-1}$	$7.75 \cdot 10^{-1}$
LST wrapper	$2.11 \cdot 10^{-8}$	$1.52 \cdot 10^{-9}$	$1.21 \cdot 10^{-5}$	$2.14 \cdot 10^{-1}$

Table 4: Distance of the best search point from the optimum for the HappyCat problem with $\alpha = 1/8$.

tion can be expressed compactly as

$$f_\alpha^{\mathrm{hc}}(x) = \left[Q^2 - 4QS + 4S^2\right]^\alpha + \frac{Q}{2N} \ .$$

with $Q = \|x\|^2$ and $S = \sum_{i=1}^N x_i$. This formulation can make good use of the increased precision of IEEE floating point numbers close to the origin. In theory this should enable solutions of extremely high precision. However, this arguments holds only for points very close to the origin, while in practice one has to follow the ridge already in some distance from the origin.

We have tested both algorithms on the HappyCat function for a number of different dimensions N. It turns out that the inner xNES instance of the LST wrapper regularly experiences numerical difficulties in its covariance update (due to its need for a matrix exponential, which turns out to be problematic in this case). Therefore we replace this algorithm with its (1+1)-ES counterpart (see [5]). The rank one updates of this variant can be computed in exponential form without the need for a matrix decomposition. For fairness of comparison we use this numerically stable algorithm also as a baseline method.

The algorithms were run until convergence and the distance of the best point to the optimum was recorded. The results are presented in table 4.

The results are clearly in favor of the LST wrapper approach. Plain (1+1)-xNES fails to locate the optimum to a satisfactory precision. Local supremum transforms do obviously help significantly. The algorithm manages to locate the optimum nearly exactly in low dimensions. However, for $N = 10$ performance degrades significantly. We conclude that the LST wrapper algorithm with (1+1)-xNES solves the HappyCat problem only in low dimensions.

6. CONCLUSION

We have introduced local supremum transforms, a family of parameterized transformations of the objective function. The transformations are compatible with the black box model since in practice they amount to simple maximization over a finite set of function evaluations. They interact nicely with rank-based algorithms since they are themselves invariant under rank-preserving transformations. The family of local supremum transforms has desirable properties that make it suitable as a replacement of the original optimization objective.

We have analyzed the impact of the supremum transformation on a class of objective functions that is notoriously difficult for black box direct search algorithms, namely sharp ridge functions. It was shown that the supremum transformed fitness is significantly easier to optimize than the ridge function itself. This enables an ES with covariance matrix adaptation to solve problems involving even extremely sharp ridges. In low search space dimensions the approach succeeds even on the HappyCat problem – a ridge-based benchmark problem that was designed to demonstrate the limitations of many direct search procedures.

We argue that the supremum transform is a principled approach to the otherwise problematic handling of sharp ridge functions with ES since it is completely agnostic to the underlying optimizer. Its disadvantages are a significant increase in the number of function evaluations and the need for an explicit mechanism controlling the scale parameter. Both shortcomings are worth addressing in future work, e.g., by means of online adaptation of the scale parameter and by adaptive switching between the optimization of plain fitness and its local supremum transformations.

7. REFERENCES

[1] D. V. Arnold and N. Hansen. A (1+1)-CMA-ES for constrained optimisation. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2012)*, pages 297–304. ACM, 2012.

[2] A. Auger and N. Hansen. A restart CMA evolution strategy with increasing population size. In B. McKay et al., editors, *The 2005 IEEE International Congress on Evolutionary Computation (CEC'05)*, volume 2, pages 1769–1776, 2005.

[3] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[4] H.-G. Beyer and S. Finck. HappyCat – A Simple Function Class Where Well-Known Direct Search Algorithms Do Fail. In *Parallel Problem Solving from Nature (PPSN) XII*, pages 367–376. Springer, 2012.

[5] T. Glasmachers, T. Schaul, and J. Schmidhuber. A Natural Evolution Strategy for Multi-Objective Optimization. In *Parallel Problem Solving from Nature (PPSN) XI*, 2010.

[6] T. Glasmachers, T. Schaul, Y. Sun, D. Wierstra, and J. Schmidhuber. Exponential Natural Evolution Strategies. In *Genetic and Evolutionary Computation Conference (GECCO)*, Portland, OR, 2010.

[7] N. Hansen, S. P. N. Niederberger, L. Guzzella, and P. Koumoutsakos. A Method for Handling Uncertainty in Evolutionary Optimization with an Application to Feedback Control of Combustion. *IEEE Transactions on Evolutionary Computation*, 13(1):180–197, 2009.

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

[9] A. I. Oyman. *Convergence Behavior of Evolution Strategies on Ridge Functions*. PhD thesis, Universität Dortmund, 1999.

[10] I. Rechenberg. *Evolutionsstrategie: Optimierung technisher Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog, 1973.