Sliding to the Global Optimum: How to Benefit from Non-Global Optima in Multimodal Multi-Objective Optimization

Christian Grimme[1], Pascal Kerschke[1], Michael T. M. Emmerich[2], Mike Preuss[1], André H. Deutz[2] and Heike Trautmann[1]

[1]*Information Systems and Statistics Group, University of Münster, Germany*
{christian.grimme, kerschke, mike.preuss, trautmann}@uni-muenster.de
[2]*LIACS, Leiden University, The Netherlands*
{a.h.deutz, m.t.m.emmerich}@liacs.leidenuniv.nl

Abstract. There is a range of phenomena in continuous, global multi-objective optimization, that cannot occur in single-objective optimization. For instance, in some multi-objective optimization problems it is possible to follow continuous paths of gradients of straightforward weighted scalarization functions, starting from locally efficient solutions, in order to reach globally Pareto optimal solutions. This paper seeks to better characterize multimodal multi-objective landscapes and to better understand the transitions from local optima to global optima in simple, path-oriented search procedures.

INTRODUCTION

Let us consider a vector valued function $f : \mathcal{X} \to \mathbb{R}^m$ and define x dominates $x' \Leftrightarrow f(x) \leq f(x') \wedge f(x) \neq f(x')$. The aim of multi-objective optimization is to find all points in $\mathcal{X}$ that are not dominated by other points in $\mathcal{X}$. This set is called the efficient set $\mathcal{X}^*$ and its image $f(\mathcal{X}^*)$ is called the Pareto front. In the following, we are considering unconstrained continuous minimization problems, that is $\mathcal{X} = \mathbb{R}^d$ for some dimension d.

Let us define two cases for local optimality, of which the first case is the usual definition with $B_\varepsilon(x) = \{x' \in \mathbb{R}^d | \, \|x - x'\| \leq \varepsilon\}$.

- A *local efficient point* $x \in \mathbb{R}^d$ is a point for which exists $\varepsilon > 0$ such that for all $x' \in B_\varepsilon(x)$ it holds that x' does not dominate x.
- A *strictly local efficient point* $x \in \mathbb{R}^d$ is a point for which exists $\varepsilon > 0$ such that for all $x' \in B_\varepsilon(x) \setminus \{x\}$ it holds that x dominates x'.

Typically, in single-objective optimization, local optima are considered to be traps for local search procedures, such as gradient descent methods. Here we extend the definition by denoting points that are locally efficient but not globally efficient as (strict) *locally-non-globally efficient* (LnGE) points.

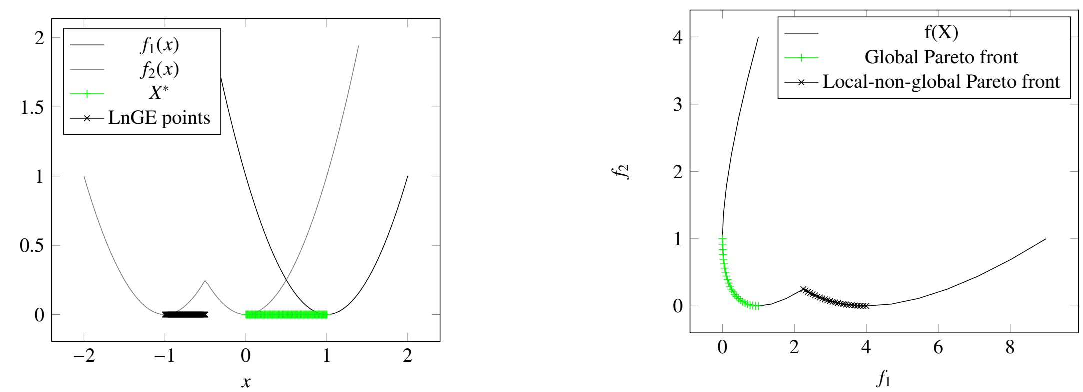
In the following, we investigate to what extent and in which cases LnGE points are traps for local search algorithms, e.g., gradient descent methods or methods that follow paths of dominated points but are also allowed to explore directions to neighboring incomparable (mutually nondominated) solutions. It will be shown, that different categories of local optima should be distinguished in multi-objective optimization, when categorizing problem difficulty and creating benchmark problems. Moreover, unlike in single-objective optimization, local optimality does not necessarily mean that local search gets trapped, unless it only accepts dominating solutions as improvements.

LOCAL OPTIMA AND GRADIENT DESCENT

Let us denote a linearly weighted scalarization function with $F_w(x) = \sum w_i f_i(x)$, for some non-negative weights $w = (w_1, ..., w_m)$. Given F_w is differentiable, its gradient $\nabla F_w(x)$ is defined as $\sum_i w_i \nabla f_i(x)$. We now investigate the question, whether following the gradient path starting from a LnGE point x_s, defined by the initial value problem $\dot{x} = \nabla F_w(x)$, $x(0) = x_s$, can lead to global optimal points. In single-objective optimization the gradient is zero in local optima, and so is $\nabla F_w(x_s) = w_1 \nabla f(x_s)$. Therefore, x_s is a stationary point. Note, that when viewing single-objective optimization as a special case of multi-objective optimization with $m = 1$, w_1 would be the only weight in a weighted scalarization approach and it is only introduced here to make the generalization explicit.

In multi-objective optimization with $m > 1$ it may, however, occur that $\nabla F_w(x_s)$ is non-zero and on a path to a globally efficient point. Let us consider a differentiable f that is locally convex and differentiable around x_s and satisfies the Karush Kuhn Tucker conditions with $\lambda_1 \nabla f_1(x) + \lambda_2 \nabla f_2(x) = 0$, with $\lambda \geq 0$ and $\lambda \neq 0$. Then, given that $\nabla f_1(x) \neq 0 \wedge \nabla f_2(x) \neq 0$, for any weight vector (w_1, w_2) that is not collinear to (λ_1, λ_2) the gradient $\nabla F_w(x)$ is non-zero and will point in the direction of steepest descent of $F_w(x)$. For non-extremal solutions (i.e., for solutions that are not optimal w.r.t. any of the objectives) it is always possible to choose a direction $w_1 = 1, w_2 = 0$ or $w_1 = 0, w_2 = 1$ to move closer to the extremes of the Pareto front. Without loss of generality, assume $w_1 = 1$, $w_2 = 0$. The move along $x_{t+dt} \leftarrow x_t + d\nabla F_w(x)$ for some infinitesimal d will generate an incomparable solution, because $f_1(x_{t+dt})$ will improve $f_1(x_t)$ and necessarily $f_2(x_{t+dt})$ will get worse than $f_2(x_t)$. Note, that if $f_2(x_{t+dt})$ would improve $f_2(x_t)$ this would contradict the assumption that x_t is non-dominated. It is hence possible, to move to an incomparable solution by means of a move in the gradient direction.

A simple example is given by the functions $f_1(x) = (x-1)^2$ and $f_2(x) = \min(x^2, (x+1)^2)$ where the line segment $[-1, -0.5)$ consists of only LnGE points and the line segment $[0, 1]$ is also globally efficient. Following the gradient of the weighted scalarization $F_w(x)$, with w_1, w_2 would yield a path from the LnGE points to optimal points.



A VISUAL CASE STUDY

For a better understanding of our perspective on multi-objective multimodality, we visualize two exemplary multi-objective problems: (1) a simple bi-objective mixed-sphere minimization problem, which consists of two mixed-peak functions and (2) a problem based on Lamé super-spheres [1] and the layered construction approach of Deb et al. [2].

The first bi-objective problem F^{MPM} is defined using unconstrained functions $f_i : \mathbb{R}^2 \to \mathbb{R}$ with

$$
\begin{aligned}
f_i^{MPM}(x) &= 1 - \max_{1 \leq j \leq N_i} \{g_{i,j}(x)\}, \qquad i \in \{1, 2\} \\
g_{i,j}(x) &= h_{i,j} \cdot \left(1 + 4\left((x_1 - c_1^{(i,j)})^2 + (x_2 - c_2^{(i,j)})^2\right)\right)^{-1}, \qquad j = 1, \dots, N_i
\end{aligned}
$$

where N_i is the number of peaks defined for each objective i. The peaks are defined at $c^{(1,1)} = (1, 0.75)^T$, $c^{(2,1)} = (0, 0)^T$, $c^{(2,2)} = (1/3, 2/3)^T$ and $c^{(2,3)} = (1, 1)^T$, and the heights are set to $h_{1,1} = h_{2,1} = 1$, $h_{2,2} = 2$ and $h_{2,3} = 3$. This results in the composition of two functions, a unimodal and a multimodal one, with one global and two local optima.

The second problem F^{ED} is constructed using the following explicit formulation:

$$
f_1^{ED}(\mathbf{x}) = \frac{1}{F_{natmin}(|x_2|) + 1} \cdot \cos\left(\frac{x_1 + 2}{4} \cdot \frac{\pi}{2}\right) \qquad f_2^{ED}(\mathbf{x}) = \frac{1}{F_{natmin}(|x_2|) + 1} \cdot \sin\left(\frac{x_1 + 2}{4} \cdot \frac{\pi}{2}\right)
$$

with $F_{\text{natmin}}(r) = b + (r - a) + 0.5 + 0.5 \cdot \cos(2\pi \cdot (r - a) + \pi)$, where $r = |x_2|$, $a \approx 0.051373$, $b \approx 0.0253235$, and $\mathbf{x} = (x_1, x_2)^T \in [-2, 2]^2$. This problem provides multiple (local) efficient sets and exposes a concave Pareto-front.

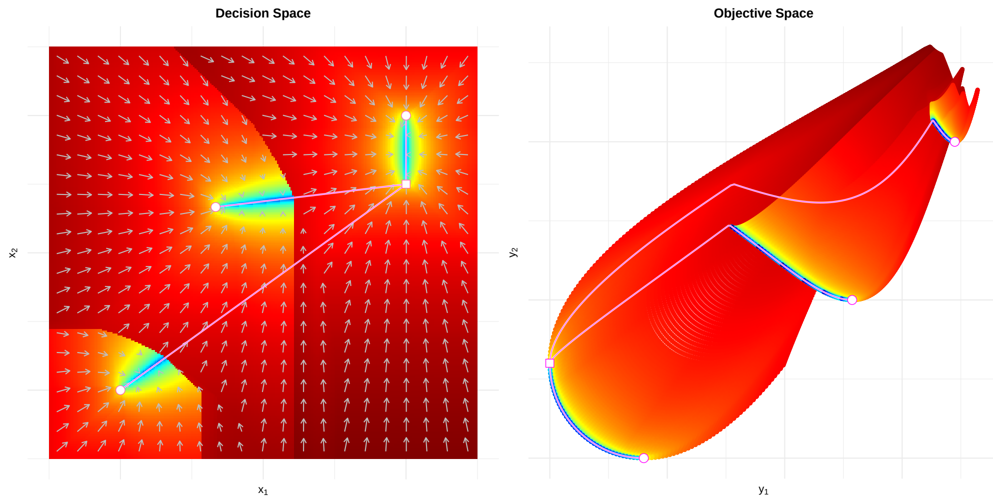


FIGURE 1. Visualization of the *gradient field heatmap* in the decision space (left), and its corresponding image in the objective space (right) for the problem F^{MPM}. Blue areas show local efficient sets, pink lines (connections of optima pairs from both objectives) indicate their potential locations, and local optima per objective are depicted as squares and circles, respectively. The arrows in the left image point in the direction of the respective combined gradient, and their lengths represent the gradients' "steepness".

The gradients of the two single-objective functions f_1 and f_2 are estimated using numerical differentiation, i.e.,

$$\nabla f_i(x) \approx \lim_{\varepsilon \to 0} \frac{f_i(x+\varepsilon) - f_i(x-\varepsilon)}{||\varepsilon||}, \qquad i \in \{1, 2\}.$$

By summing up both normalized objective-wise gradients

$$\nabla f(x) = \frac{\nabla f_1(x)}{||\nabla f_1(x)||} + \frac{\nabla f_2(x)}{||\nabla f_2(x)||}$$

we define the corresponding bi-objective gradient $\nabla f(x)$ of the joint function $f : \mathbb{R}^2 \to \mathbb{R}^2$ as $f(x) = (f_1(x), f_2(x))^T$. Due to numerical issues caused by the discretization of the decision space (see below), we consider a point $x \in \mathbb{R}^2$ to be (approximately) locally efficient, if $||\nabla f(x)|| < \xi$ or the angle $\angle(-\nabla f_1(x), \nabla f_2(x)) < \psi$. Within our experiments, we used $\varepsilon := (10^{-6}, 10^{-6})^T$, $\xi := 10^{-3}$ and $\psi := 10^{-6}$.

To get an impression of the decision space "landscape", we compute a *gradient field heatmap*. Therefore, the decision space is discretized, and for each point of the grid we follow the path of bi-objective gradients from one cell to the respective adjacent cell, until a locally efficient point is reached (there, the bi-objective gradient is zero). For each point of the grid, the respective heatmap's utility function value is defined as the cumulated length of all bi-objective gradients along the path towards its current "attractor". Sets of connected local efficient points are shown as blue valleys within the heatmap [3].

Figures 1 and 2 show the decision (left) and objective (right) spaces for F^{ED} and F^{MPM}, respectively. In Figure 1, we clearly find two LnGE sets that are cut by the basin of attraction of another efficient set. Independent of the starting point for a search, a multi-objective gradient descent will lead to an efficient set. If we have met a LnGE set we will eventually step beyond the ridge and the gradient descent method will lead us to a (locally) efficient point in the entered basin. This can be repeated until we meet (possibly only a part of) the in this case global efficient set. From here, we will no longer be able to leave the basin by combined gradient descent as its ridge-free neighborhood is completely surrounded by dominated solutions. For the problem depicted in Figure 2, efficient sets can also be reached by gradient descent. However, depending on the starting point of the search and also as consequence of walking along incomparable solutions, the search will be diverted to the boundary of the decision space. Note that gradient descent will not be able to leave these locations any more.

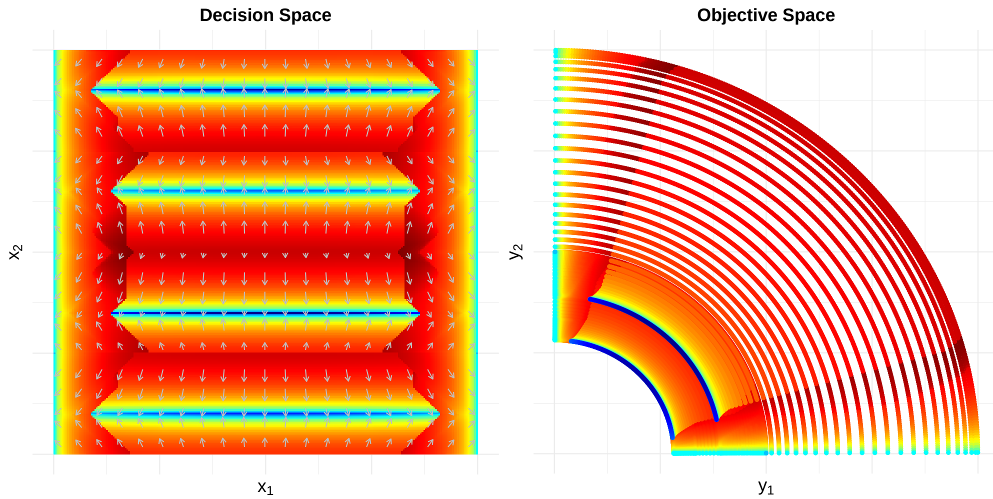


FIGURE 2. Visualization of the *gradient field heatmap* in the decision space (left), along with its corresponding image in the objective space (right) for the problem F^{ED}.

CONCLUSIONS

In this work, we have shown that in multi-objective optimization, locally efficient points are not necessarily traps for local search using gradient methods. As a matter of fact, gradient methods might be used to slide along the efficient set and discover new non-dominated solutions. They can even be used to discover globally optimal solutions, starting from locally-non-globally efficient points. These properties are very different to those found in single-objective optimization, and could be used to construct global optimization methods that utilize gradient descent methods. An example has been discussed where a method that moves along the gradient of a linear scalarization function can lead from a LnGE to a globally optimal point. We suspect that many multimodal benchmark problems in the literature belong to this class and that the discussed phenomenon explains why some local search methods perform surprisingly well on these. As such, our findings complement the conjectures on regularity made by [4] and open up perspectives for algorithm design. We have also provided an example of a function where a linear scalarization gradient cannot be used to escape from a local optimum, hence, featuring local efficient set components that are isolated from the global efficient components by "true" barriers. For the design of benchmark problems, this can be considered as a more difficult case of multi-objective multimodal optimization.

ACKNOWLEDGMENTS

The authors acknowledge support by the *European Research Center for Information Systems* (ERCIS).

REFERENCES

[1] M. T. M. Emmerich and A. H. Deutz, "Test Problems Based on Lamé Superspheres," in *Proceedings of the 4th International Conference on Evolutionary Multi-Criterion Optimization (EMO)*, LNCS, Vol. 4403 (Springer, 2007), pp. 922 – 936.

[2] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler, "Scalable Test Problems for Evolutionary Multiobjective Optimization," in *Evolutionary Multiobjective Optimization*, AI & KP (Springer, 2005), pp. 105 – 145.

[3] P. Kerschke and C. Grimme, "An Expedition to Multimodal Multi-Objective Optimization Landscapes," in *Proceedings of the 9th International Conference on Evolutionary Multi-Criterion Optimization (EMO)*, LNCS, Vol. 10173 (Springer, 2017), pp. 329 – 343.

[4] Y. Jin and B. Sendhoff, "Connectedness, Regularity and the Success of Local Search in Evolutionary Multi-Objective Optimization," in *The 2003 Congress on Evolutionary Computation (CEC)*, Vol. 3 (IEEE, 2003), pp. 1910 – 1917.