Network structure of the set of local minima in optical system optimization

Florian Bociort*, Eco van Driel, Alexander Serebriakov

Optics Research Group, Delft University of Technology
Lorentzweg 1, NL - 2628 CJ Delft, The Netherlands

ABSTRACT

We discuss a surprising new feature of the merit function landscape in optical system design. When certain conditions are satisfied, the set of local minima forms a network in which all nodes are connected. Each link between two neighboring minima contains a special type of saddle point (more precisely, a saddle point having a Morse index 1). On this basis, a new global optimization method that takes advantage of this feature is proposed. The central component of the new method, the algorithm for saddle point detection, works in a parameter space of arbitrary dimensionality, and uses only the local optimization engine of the optical design program. For a simple global optimization search (the symmetric Cooke triplet) the network of the corresponding set of local minima is presented.

Keywords: global optimization, saddle point, Morse index, network, optical system design

1. INTRODUCTION

The presence of multiple local minima during optimization is one of the major challenges in optical system design, as well as in many other fields. Over the past decades, the considerable amount of effort spent for developing and improving global optimization methods has resulted in a very large number of publications, both in mathematical literature (for an overview see e.g. Ref. 1) and in literature dedicated to specific types of applications. Successful algorithms such as simulated annealing[2], global synthesis[3], genetic algorithms[4], or the escape function method[5], are implemented in commercial optical design programs and have a major impact on modern optical design methodology. A limitation of present-day global optimization methods is however that local minima are given only as isolated points in the parameter space of the system, with no (or very little) information about the merit function topography around the individual local minima or in the space between them.

In this paper we show that, when certain quite general conditions are satisfied, the merit function landscape has a remarkable property, which we could not find mentioned in earlier literature. The local minima form then a network in which all nodes are connected via links that contain a special type of saddle point. It is known for several decades that a way to find a new local minimum is to identify a saddle point on the boundary of its region of attraction[6,7]. As shown in Section 2 however, not all saddle points are equally important: for finding new local minima it is sufficient to detect only the saddle points that have a Morse index of 1. In Section 3 we describe our first attempt to develop a new type of global optimization based on this network structure. We will show that the links between nodes can be found with an algorithm that uses only the local optimization engine of optical design software. In Section 4 we present as an example the network corresponding to the symmetric Cooke triplet global search.

* *E-mail: bociort@tn.tudelft.nl, Fax: +31 15 2788105*

2. CONNECTED NETWORKS OF LOCAL MINIMA

In this paper we consider for simplicity a global optimization problem with continuous variables, having no constraints or only equality constraints (the case of inequality constraints will be discussed in a future paper), and assume a merit function of the form

$$f(\mathbf{x}) = \sqrt{\frac{\sum w_i \left(a_i(\mathbf{x}) - \tilde{a}_i\right)^2}{\sum w_i}} \tag{1}$$

where a_i are image defects computed with ray tracing, w_i the corresponding weights and the tilde denotes the target values for the corresponding a_i. A point in the solution space is described by the vector $\mathbf{x} = (x_1, x_2, ..., x_N)$ whose components are the N optimization variables. The critical points in the N-dimensional solution space are those points for which the gradient of f vanishes

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_N}\right) = 0 \tag{2}$$

In this section we will focus on the behavior of the equimagnitude surfaces of f, which are $N-1$ dimensional hypersurfaces in the solution space along which f is constant. In a small neighborhood around a critical point, the equimagnitude surfaces are given by

$$\sum A_{ij} \hat{x}_i \hat{x}_j = const \tag{3}$$

where the circumflex denotes the values of the optimization variables in a translated coordinate system that has its origin at the critical point, and where A is the matrix of the second-order derivatives, computed at the critical point

$$A_{ij} = \frac{\partial^2 f}{\partial \hat{x}_i \partial \hat{x}_j} \tag{4}$$

As known from linear algebra, the coodinate system can be rotated in such a way that the quadratic form on the left-hand side of Eq.(3) contains only squares of the variables (denoted below by a bar) in the new coordinate system. The equimagnitude surfaces around the critical point now become

$$\sum \lambda_i \bar{x}_i^2 = const \tag{5}$$

The axes of the new coordinate system are then oriented along the eigenvectors of A, and the factors λ_i in Eq.(5) are the corresponding eigenvalues. (See Fig.1.) The way to perform this rotation of the coordinate system will be discussed in more detail in the next section.

Assuming that all eigenvalues are nonzero, the Morse index (MI) of the critical point is defined as the number of negative eigenvalues in Eq.(5)[8]. A negative eigenvalue means that along the direction defined by the corresponding eigenvector the critical point is a maximum. Thus, local minima have $MI=0$ (all eigenvalues are positive), i.e. they are minima in all directions, whereas local maxima have $MI=N$. As shown in Fig. 1, in both cases the equimagnitude surfaces around these points are ellipsoids with axes oriented along the eigenvectors. For saddle points, the Morse index has values between 1 and $N-1$. Figure 2 shows the merit function in the neighborhood of a two-dimensional saddle point and the enlarged detail of Fig. 3 shows the corresponding equimagnitude contours. Close to the saddle point, these contours are hyperbolas, whereas for the value of the merit function corresponding to the saddle point itself the equimagnitude contours degenerate into a pair of straight lines (the asymptotes of the hyperbolas).

As will be shown below, if we are interested in the detection of local minima in a solution space with $N>1$, the saddle points with $MI=1$, which are maxima in one direction and minima in $N-1$ directions, play a special role. Note first that equimagnitude surfaces having some value $f=f_0$ of the merit function encircle regions in the solution space for which $f<f_0$. If for instance f_0 is the merit function of a local minimum then the equimagnitude surface (or the part of it situated near the minimum) reduces to one point, the minimum itself. For slightly larger values of f_0 (a part of) the equimagnitude surface encircles a small ellipsoidal region around the local minimum. If the value of f_0 continues to increase then the encircled volume also increases.

We will now show in an intuitive way that the local minima within an arbitrary equimagnitude surface form a connected network, i.e. that there is a well-defined path from any local minimum to any other local minimum in the solution space. Consider first the situation shown in Fig. 3, two local minima in an N-dimensional solution space. We assume the existence of a surface with $f_0 = f_a$ that encircles both minima (the thick dashed curve). Then, for a sufficiently small value $f_0 = f_b < f_a$ (for instance for f_b slightly larger than the largest of the two merit function values corresponding to the local minima) the equimagnitude surface consists of two separate parts (the thick dotted curves), whereas for $f_0 = f_a$ we have only one encircling surface (thick dashed). Assuming that the merit function landscape is free of pathologies, for some value f_S with $f_b < f_S < f_a$ we will encounter the limiting case when the two separate parts of the encircling surface will touch each other in one point S. We now show that the split point S is in fact a saddle point with $MI=1$. If we consider a value of the merit function $f_B = f_{B'}$ slightly lower than f_S , the corresponding equimagnitude surface will be split. Let then B and B' be the two points on the separate parts of the equimagnitude surface for which the length of the segment BB' is minimal. (See enlarged detail in Fig. 3.) Obviously, along the line BB' the split point S is a maximum. We now consider an equimagnitude surface with a merit function $f_A = f_{A'}$ slightly larger than f_S . Since this equimagnitude surface encircles the one with $f_0 = f_S$, any line perpendicular to BB' and passing through S will intersect it in two points, denoted by A and A' in Fig. 3. Along AA' the point S is then a minimum. Since this is valid for any choice of the line AA' in a $N-1$ dimensional hyperplane orthogonal to BB', S is a minimum in $N-1$ directions, and is thus a saddle point with $MI=1$. If we now chose the points B and B' as starting points, local optimization will generate two paths in the solution space that will lead to the two minima. Together with the saddle point, these two paths form the link between the two local minima.

Assume now that we have an equimagnitude surface with some (large) value of $f_0 = f_a$ that encircles an arbitrary number p of local minima. (In Fig 4. where $p=3$, this is the outermost contour.) If we decrease f_0, at some value $f_{S1} < f_a$ the encircling surface will split into two surfaces that will now encircle p_1 and $p-p_1$ local minima, respectively. (In Fig 4. we have $p_1=1$.) Using the same reasoning as above, it can be seen that the point S_1 in Fig. 4 is also a saddle point with $MI=1$. By starting local optimizations at a pair of points obtained be slightly perturbing the saddle point on both sides along the eigenvector with negative eigenvalue, we obtain a link between one local minimum in the group of p_1 encircled local minima, and one in the group of $p-p_1$ local minima. By further decreasing f_0, we obtain successive splits of the encircling surfaces. Each such split generates an additional link between two local minima situated in the two different groups resulting from the split. When f_0 has reached a value that is lower than the merit function of the lowest $MI=1$ saddle point (S_2 in Fig.4) all local minima encircled by the equimagnitude surface with $f_0 = f_a$ are linked together in a network via links that contain each a $MI=1$ saddle point.

We have thus shown that the local minima encircled by an arbitrary equimagnitude surface form a connected network[9]. For our purposes, it is important to know whether in typical situations occurring during optical system optimization we can always find such equimagnitude surfaces that encircle all (useful) local minima. At the time of this writing, we have examined only a limited number of situations and further research is certainly necessary. Our present results make us believe however that either this desirable property of the landscape of the merit function (1) is satisfied automatically, or that it can be achieved by modifying the optimization problem adequately. It is well known to optical designers that outside some useful regions in the solution space the optical system configurations tend to suffer from ray failure because some rays either miss surfaces or suffer from total internal reflection. Close to ray-failure situations, the incidence angles of those rays at the critical surfaces are large, therefore the aberrations and the merit function (1) of the given optical system configuration tend to be large. Therefore, close to the ray failure borders we can expect in the solution space equimagnitude surfaces having a large value of the merit function. The local minima encircled by these surfaces form then a network. The possibility of enforcing the desirable properties of the merit function landscape when these properties are not automatically satisfied will be discussed in a future paper.

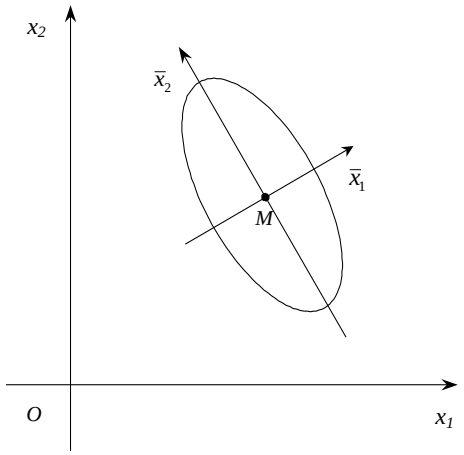


Figure 1. In a small neighborhood around a local minimum or maximum, the equimagnitude surfaces are ellipsoids having their axes oriented along the eigenvectors of the matrix *A*.

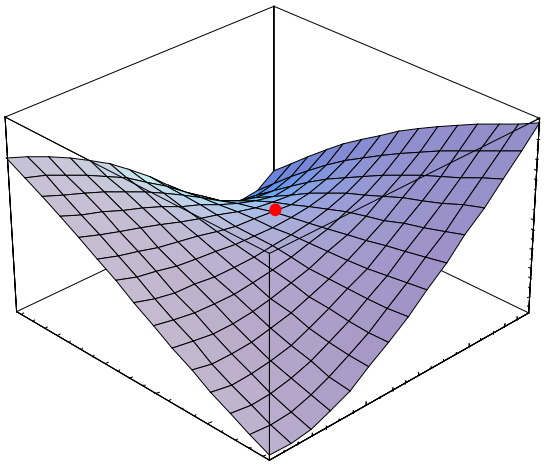


Figure 2. In a two-dimensional solution space, all saddle points have MI=1. At the saddle point, the merit function has a minimum in one direction and a maximum in the direction perpendicular to the first one.

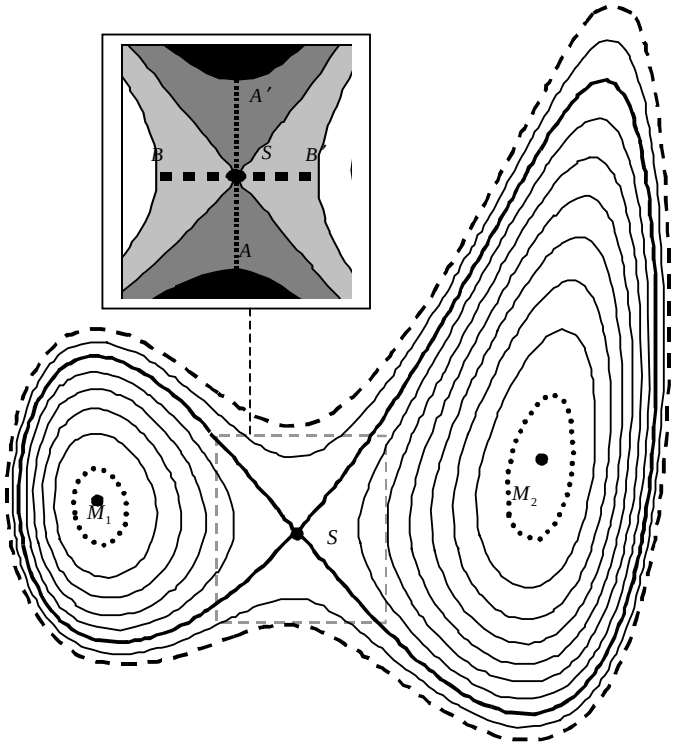


Figure 3. Two local minima and a saddle point. In the enlarged detail, light gray indicates lower and dark gray indicates higher values of the merit function.

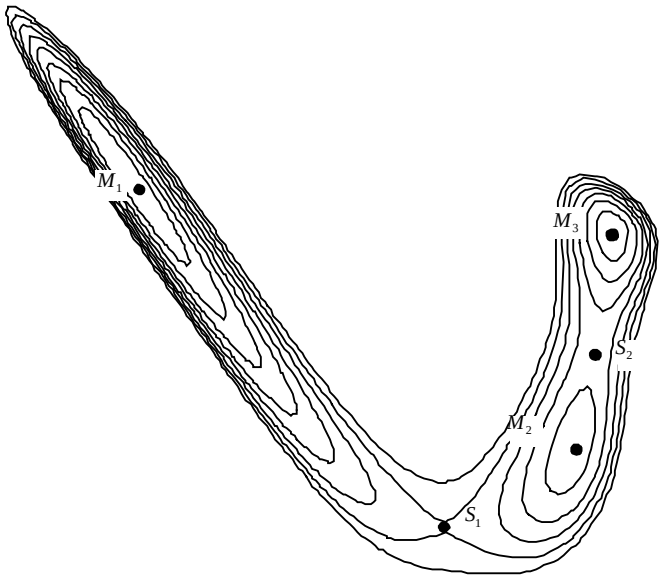


Figure 4. Several local minima and the saddle points between them. Figures 3 and 4 are two-dimensional cuts through the 5-dimensional merit function landscape of a Cooke triplet global search. (See Sec. 4.)

3. GLOBAL OPTIMIZATION BASED ON SADDLE-POINT DETECTION

A first practical utility of the network structure discussed in the previous section is for global optimization. This generally very difficult task can now be divided in three separate steps:

i) for a given local minimum, detect the *MI=1* saddle points that connect it with the neighboring local minima,

ii) starting from some arbitrary local minimum, find the rest of the network and

iii) when the network is known, select the best solution(s) or identify entire branches along which the imaging performances of the nodes are satisfactory.

In this section, we will briefly discuss our first attempt to develop a global optimization method based on this strategy. Additional details will be given in a subsequent paper. In what follows we will focus on the first step, the saddle point detection. Fortunately, we have to detect only the saddle points with *MI=1*, while those with a higher Morse index, which are more difficult to detect, can be safely ignored for the present purpose. For shortness, in the rest of this paper a *MI=1* saddle point will be referred to as a "saddle point".

The first step of the detection process is to compute for the given local minimum the eigenvectors of the matrix *A* given by Eq.(4), i.e. to find the orientation of the axes of the ellipsoid shown in Fig.1. Since this goal must be achieved within an optical design program, we have chosen a technique based on local optimization. We have therefore transformed into a computer algorithm a mathematical idea that is usually used to describe the rotation of axes (shown in Fig. 1) that diagonalizes the matrix A[10].

Consider around the local minimum M a hypersphere whose radius $r = MQ$ is sufficiently small so that the equimagnitude surfaces that intersect it are ellipsoids given by Eq. (3). (See Fig. 5.) We first compute the direction of the eigenvector that has the smallest eigenvalue, i.e. the direction of the longest axis of the ellipsoids. Since smaller ellipsoids have smaller values of f_0, for the merit functions of the points P , P_1 and P_2 in Fig. 5 we can write $f(P_2)>f(P_1)>f(P)$. Therefore, a local minimization of the merit function, constrained on the hypersphere of radius r, will produce as a result one of the two points in which the inscribed ellipsoid (thick curve) touches the hypersphere. Each of these two points can be used to define the direction *MP* of the eigenvector. In order to compute the remaining eigenvectors, we use the fact that they are all orthogonal to each other. We will now use a coordinate system having its origin in M. Thus, if P has the coordinates $\hat{x}_1^1, \hat{x}_2^1, \ldots \hat{x}_N^1$, the other eigenvectors must be situated in the *N-1* dimensional hyperplane orthogonal to *MP*, which is given by

$$\hat{x}_1^1 \hat{x}_1 + \hat{x}_2^1 \hat{x}_2 + \ldots + \hat{x}_N^1 \hat{x}_N = 0 \qquad (6)$$

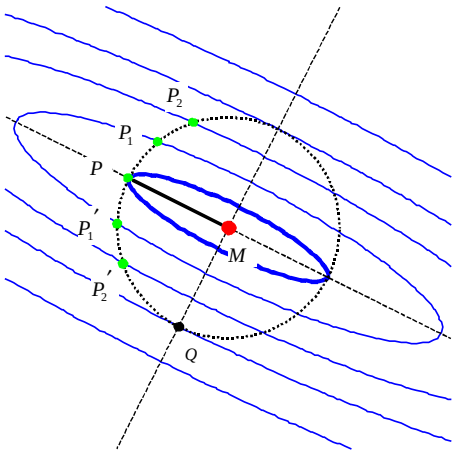


Figure 5. Computation of the eigenvectors based on local minimization (see text).

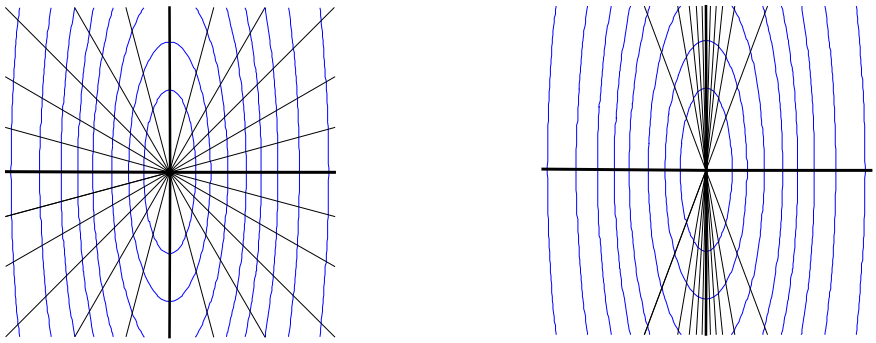
(In two dimensions, Eq.(6) is the equation for the direction *MQ* in Fig.5.) Adding Eq.(6) as a second constraint and reoptimizing along the hypersphere we obtain the direction of the second eigenvector. By adding for each newly found eigenvector an additional constraint similar to Eq.(6) and reoptimizing, all eigenvectors are found one after the other.

In order to detect the saddle points that connect the local minimum M with neighboring local minima, we first define a set of directions, each characterized by an unit direction vector **s**. For each such direction, we consider the set of hyperplanes orthogonal to **s**

$$s_1 \hat{x}_1 + s_2 \hat{x}_2 + \ldots + s_N \hat{x}_N = t \tag{7}$$

where *t* gives the distance between the hyperplane and M along the normal that passes through M. For a given value of *t* we compute the constrained minimum of *f* in the hyperplane (7). Let $\hat{\mathbf{x}}_\mathbf{s}(t)$ and $F_\mathbf{s}(t)$ be the position vector and the value of *f* corresponding to this minimum, respectively. For a given direction, if we start from the position of M at *t=0* and increase *t* gradually, at the beginning $F_\mathbf{s}(t)$ increases. A neighboring saddle point is detected when $F_\mathbf{s}(t)$ reaches a maximum for some value $t_{\max}$, provided that $\hat{\mathbf{x}}_\mathbf{s}(t)$ is continuous for $0 < t < t_{\max}$ (i.e. no jumps have been observed during the gradual increase of *t*). It is important to note that not all such searches lead to saddle points; some of them will terminate in dead ends.

Since the equimagnitude ellipsoids in the immediate vicinity of the point M are often strongly elongated, a useful set of search directions **s** can be determined on the basis of the eigenvectors computed at M. As will be shown in more detail in a future paper, if we start with an uniformly distributed set of directions **s** (Fig. 6a), for most directions **s** the corresponding vectors $\hat{\mathbf{x}}_\mathbf{s}(t)$ will then be concentrated in a narrow cone around the first eigenvector (Fig. 6b). However, if the directions **s** are oriented along eigenvectors the algorithm escapes from this cone and can explore different regions of the solution space. At present we use in our algorithm two searches in opposite directions for each eigenvector, i.e. a total of *2N* independent searches for each local minimum. For the global searches performed up to now, this number of searches seemed to be sufficient. In fact, many saddle points have been detected several times.



a) b)

Figure 6. a) Uniform angular spread of search directions **s** around a local minimum. b) The corresponding search trajectories $\hat{\mathbf{x}}_\mathbf{s}(t)$ for small values of *t* in the case of elongated equimagnitude surfaces. The search trajectories are oriented in the direction of the corresponding **s** only if **s** if oriented along eigenvectors (thick lines), otherwise they tend to be concentrated in a narrow cone along the longest axis of the ellipsoid.

4. RESULTS

We have implemented the algorithm described in the preceding section in the macro language of the commercial optical design program CODE V and have tested it in several simple cases. Figure 7 shows the results of our global search in the case of a triplet where the optimization variables were the six curvatures of the surfaces. We have found 19 local minima (drawn within thick-line boxes), connected via 23 saddle points (thin-line boxes). The first 17 of our local minima are identical with the 17 local minima found with Global Synthesis (the global optimization algorithm of CODE V). Our last two local minima, which have large values of the merit function, are not listed in the output of Global Synthesis. Interestingly, the saddle-point configurations s_{i-j} can be viewed as intermediate stages in a continuous transformation of the local minimum m_i into the minimum m_j.

For testing the reliability of our network detection, in this example we have chosen the specifications (distances between surfaces, glass types) to be rigorously symmetric with respect to the aperture stop. For this purpose, the central lens has been split by a fictitious stop surface (not shown in Fig.7). The image plane was placed at its paraxial position and the position of the object plane was controlled such that the transverse magnification was kept equal to -1. Because of an additional equality constraint (the distance between object and image was also kept constant) the search space was effectively 5-dimensional. The merit function used was the default merit function of CODE V, for which the image defects in Eq.(1) are transverse ray aberrations computed with respect to the chief ray.

As expected, the detected network is almost perfectly symmetric. With one exception, the saddle point s_{8-5}, the configurations in Fig.7 are either symmetric with respect to the stop (m_9, s_{1-2} , and s_{19-18}) or they have mirror images. For clarity, the pairs in which one configuration is (almost) the mirror image of the other have been grouped together in the same box. Moreover, with the exception of the two dashed links in the lower right part of Fig.7, the detected links display the same symmetry: if a saddle point links two minima, then the mirror of the saddle point will link the mirrors of the same minima. The minor deviations of the network from perfect symmetry are not surprising since the aberrations that affect the ray tracing results perturb to some extend the symmetry between object and image. At this early stage, we cannot guarantee that our algorithm has detected the entire network. However, the symmetry detected as expected increases our confidence in the potential of this type of global optimization algorithms.

The best two local minima of this search, m_1 and m_2, have the well-known shape of the Cooke triplet. Interestingly, for a numerical aperture of 0.055 (the value of this search), both are slightly asymmetric and form a mirror pair, whereas the saddle point s_{1-2} between them is symmetric. However, if we increase the numerical aperture, beyond the value of 0.075 these two minima will merge into a single symmetric one.

5. CONCLUSIONS

We have examined the properties of the optimization solution space from what mathematicians might call a topological perspective. Because this high-dimensional space is usually very complicated, it is unmanageable without focussing on particular features. By analyzing the splitting or merging of the equimagnitude surfaces when the corresponding merit function value changes, we have selected those features which are relevant for our present purpose and have ignored the rest. Based on the idea that the local minima form a network, a new type of global optimization algorithm has been proposed. Our simple but nontrivial example shows that algorithms based on this idea could in principle not only reproduce the results of presently known global optimization algorithms, but also provide additional insight into the topography of the merit function landscape.

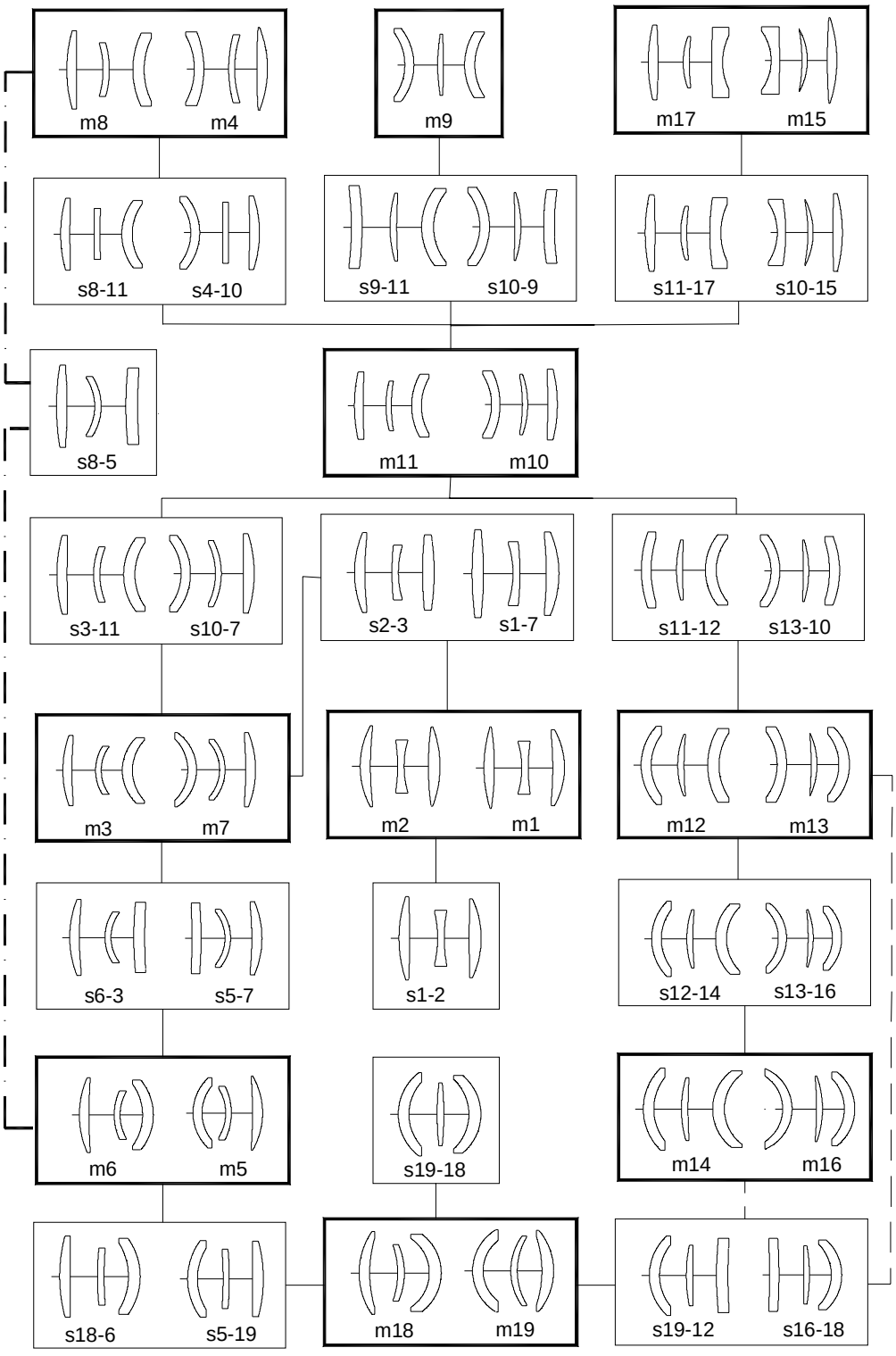


Figure 7. Network of the global search corresponding to the symmetric Cooke triplet

ACKNOWLEDGMENTS

We would like to thank Joseph Braat and Paul Urbach for stimulating discussions and Bas Swinkels for a useful suggestion. We also gratefully acknowledge the use of an educational license of CODE V.

REFERENCES

1. H. Reiner and P. Pardalos (editors), *Handbook of Global Optimization*, Kluwer, Dordrecht, 1995

2. G. W. Forbes and A. E. W. Jones, "Towards global optimization with adaptive simulated annealing", Proc. SPIE **1354**, 144-151, 1991

3. T. G. Kuper and T. I. Harris, "Global optimization for lens design - an emerging technology", Proc. SPIE **1780**, 14-28, 1992

4. K. E. Moore, "Algorithm for global optimization of optical systems based on genetic competition", Proc. SPIE **3780**, 40-47, 1999

5. M. Isshiki, H. Ono, K. Hiraga, J. Ishikawa, S. Nakadate, "Lens design: Global optimization with Escape Function", Optical Review (Japan), ***6***, 463-470, 1995

6. G. Treccani, L. Trabattoni and G. P. Szego, "A numerical method for the isolation of minima", in *Minimisation Algorithms, Mathematical Theories and Computer Results*, G. P. Szego (editor), p 239-255, Academic Press, 1972

7. C. R. Corles, "The use of regions of attraction to identify global minima", in *Towards global optimisation*, L.C.W. Dixon and G.P.Szego (editors), p55-95, North - Holland/Elsevier, Amsterdam, 1975

8. J. C. Hart, "Morse Theory for Implicit Surface Modeling" in *Mathematical Visualization*, H-C Hege and K. Polthier (editors), pp. 257-268, Springer-Verlag, Berlin, 1998

9. In some special situations, it may be useful to modify this statement. If we have for instance a positive merit function that decreases to zero when any variable (or linear combination of them) tends to infinity, then, in addition to the "usual" local minima we also have a continuum of minima at infinity. For practical purposes we may be interested in a network in which the links between "usual" local minima do not pass through infinity. Such a network is formed for instance when a pair of equimagnitude surfaces exists such that both surfaces encircle the "usual" local minima, and when the surface having the larger value of the merit function also encircles the other one.

10. R. Courant and D. Hilbert, *Methods of mathematical physics*, p23-26, John Wiley, New York, 1989