Local optimization strategies to escape from poor local minima

Florian Bociort, Alexander Serebriakov and Joseph Braat

Optics Research Group, Delft University of Technology
Lorentzweg 1, NL - 2628 CJ Delft, The Netherlands,
E-mail: bociort@tn.tudelft.nl

ABSTRACT

Three local optimization strategies to escape from a local minimum are discussed. The first strategy is to radically modify the error function. The old and new error function should both tend to zero for ideal systems but must differ sufficiently from another. The second strategy is to temporarily over-design the system, i. e. to make available for optimization more system parameters than a designer would normally use for the given aperture and field specifications. Finally, it is shown that, with a small enhancement of the local optimization algorithm, it is possible to move from one local minimum into a neighboring one by locating a saddle point between them.

OCIS Code: (080.2740) Geometrical optics, optical system design
Keywords: optimization, local minima, merit function, saddle points

1. INTRODUCTION

The presence of multiple local minima during optimization is one of the major challenges of optical system design. Recent progress in global optimization gives the optical designer a very powerful tool to escape from poor local minima [1-4]. These methods tend however to be very time-consuming and therefore designers often still attempt to improve solutions trapped in unsatisfactory local minima by using local optimization strategies. In this work we will discuss three strategies of this type (sections 2,3 and 4).

2. MODIFYING THE MERIT FUNCTION

One of the empirical strategies to "tunnel" out of bad local minima (or to escape from stagnation) is to modify the conditions under which local optimization algorithms operate. This can be achieved by changing, for instance, the system parameters, some parameters of the local optimization algorithm, or the merit function. In this section we take a fresh look at the well-known strategy of switching to a different type of merit function.

Many computer programs allow an easy switch between merit functions based on transverse and wavefront aberrations. Although occasionally successful, this strategy is limited by the fact that when system parameters change the behavior of these two merit functions is often very similar. Our experience shows that the chances of success increase when the "old" and "new" merit function both tend to zero for ideal systems but differ sufficiently from another.

A new type of merit function that may be a useful switch partner for the standard merit functions in intermediate stages of optimization can be defined as follows: Consider an arbitrary ray ($A'I'$ in Fig. 1) that has in the image space the direction cosines L and M with respect to the x and y axes, respectively. The two components of the transverse aberration of the ray (defined with respect to the chief ray) are denoted by δx and δy. The ray intersects in I' the image plane and in A' a sphere centered in the intersection point I of the chief ray with the image plane. The radius R of the sphere should be chosen larger, but still of the same order of magnitude as the length $I\,I'$ of the transverse aberration vector. The length R' of the segment $A'I'$ is then given by

$$R'/R = a + \sqrt{1-b} = a + 1 - b/2 - b^2/8...$$

where we have used the abbreviations

$$a = (L\delta x + M\delta y)/R$$

and

$$b = \left[(\delta x^2 + \delta y^2) - (L\delta x + M\delta y)^2\right]/R^2 .$$

Since for ideal imaging R' tends to R , the quantity $R'/R-1$, averaged over all rays for a given field point, can be used as a intermediate stage merit function (the "radius" merit function). The power series expression should be used instead of the exact expression in order to avoid abnormal termination when for certain rays that have large aberrations b becomes larger than 1.

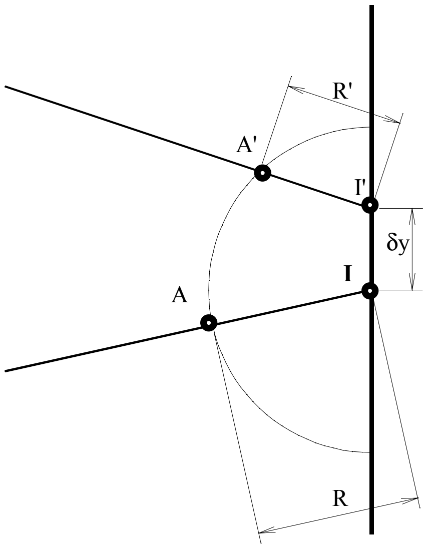


Figure 1. The definition of the "radius" merit function

The "radius" merit function has been implemented as a user defined merit function in CODE V. Figure 2 shows the evolution of the "radius" merit function as well as the root mean square (RMS) spot size and RMS wavefront during an optimization driven by the "radius" merit function. It can be observed that, while the "radius" always decreases, the other two merit functions actually increase beyond point *A*. Moreover, the RMS spot size and RMS wavefront appear to be strongly correlated i.e. they increase or decrease at the same time. In many other tests we have found that the behavior of the transverse aberration along trajectories in the parameter space is much stronger correlated with that of the wavefront that with that of the "radius". Therefore, the "radius" may be a more successful switch partner for the RMS spot size than the RMS wavefront.

In fact, switching back and forth between the RMS spot size and the "radius" merit functions can sometimes lead to good quality solutions even when we start from very poor initial configurations. For instance, a configuration very close to the well-known Double Gauss was obtained by starting from plane-parallel plates (that have a solve on the last surface for controlling the focal length).

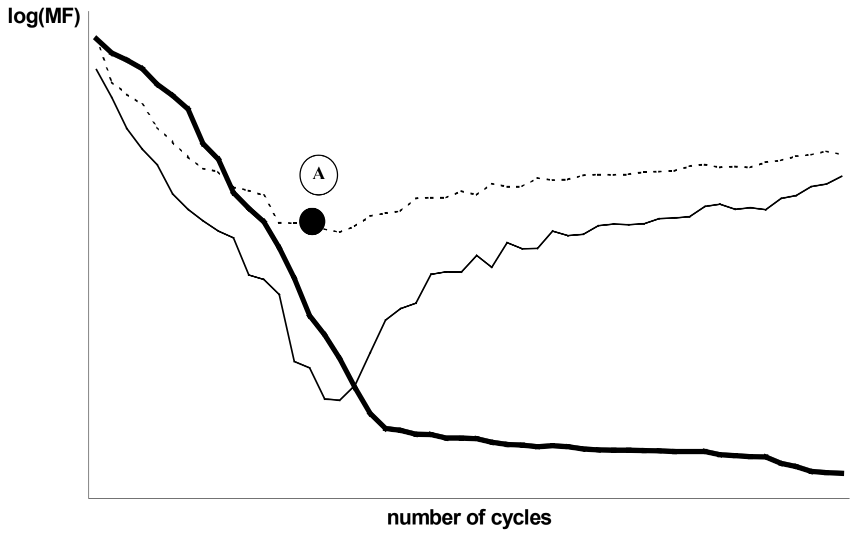


Figure 2. The evolution during optimization of the merit functions based on the "radius" (thick line), transverse aberration (thin line) and wavefront aberration (dashed).

3. OVER-DESIGNING

A different potentially useful strategy to escape from poor local minima is to temporarily over-design the system, i. e. to make available for optimization more system parameters than a designer would normally use for the given aperture and field specifications.

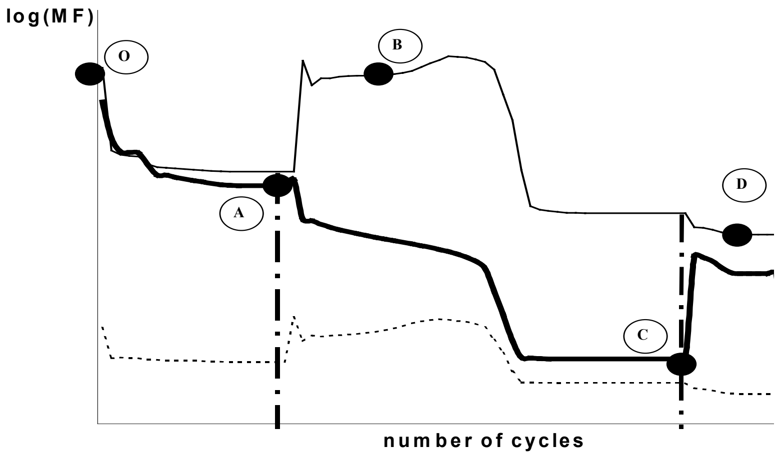


Figure 3. Evolution of the standard merit functions based on transverse aberrations (thin line) and wavefront aberrations (dashed) as well as the evolution of a merit function with very small off-axis field weights (thick line), in the process of obtaining the Cooke Triplet. Between *O* and *A* and between *C* and *D* the driving merit function is the standard transverse aberration. Between *A* and *C* the driving merit function is the one with reduced field weights. Note that along the entire trajectory the standard merit functions based on transverse and wavefront aberrations are strongly correlated.

A simple example is shown in Fig. 3, where the Cooke Triplet is obtained by starting from plane-parallel plates. Using the default CODE V RMS spot size and starting in point *O* (plane parallel plates with a solve on the last curvature to enforce the desired focal length) the solution reaches the local minimum *A*, where the quality is less than optimal. In this point the merit function is modified by reducing the off-axis field weights to values close to zero. Basically, between the points *A* and *C* the optimization uses all variables only to correct on-axis aberrations of different orders (over-designing). Because off-axis aberrations are (almost) neglected by the optimalization algorithm, the spot size in point *B* becomes very large. Despite this fact, if in point *C* the merit function is switched back to default, the system reaches the standard Cooke Triplet shape (*D*) where the imaging quality is considerably better than in point *A*.

This strategy turns out to work well also for poor local minima of more complex systems. In more general situations, if the system should have only spherical surfaces, a similar over-designing effect is sometimes achieved by temporarily making the surfaces aspheric, including the aspheric coefficients in the optimization, and then reducing them gradually back to zero. The temporary increase of the dimensionality of the variable space often creates new opportunities for descent directions at the location of the poor local minimum.

4. LOCATING SADDLE POINTS

The two strategies presented above suffer from the drawback that their success depends to a large extend on serendipity. It is very hard to predict when they will work and when not. Therefore, in this section we will discuss a very different type of strategy, one that also offers a deeper insight into the peculiarities of the specific design situation. We will see that, with a small enhancement of the local optimization algorithm, it is possible to move from one local minimum into a neighboring one by locating a saddle point between them.

An important issue to be examined is the following. If we are in the vicinity of a given local minimum (in any case, we are still in the basin of attraction of the given minimum) can we somehow feel the presence of a neighboring local minimum?

Let us first start with a simple mountain-landscape analogy. A tourist wants to walk from one valley of the landscape to a neighboring one, and spend the least amount of effort in doing so. The optimal path would then be one that passes through a mountain pass. A mountain pass is in fact an intuitive example of a saddle point. Returning to optimization, the strategy we propose is the following: for finding a neighboring local minimum we find first a saddle point that leads to it.

Figure 4 shows a two-dimensional saddle point. As for local minima the gradient of the merit function vanishes at the saddle point. However, while in one direction the saddle point is a minimum, in the other direction it is a maximum.

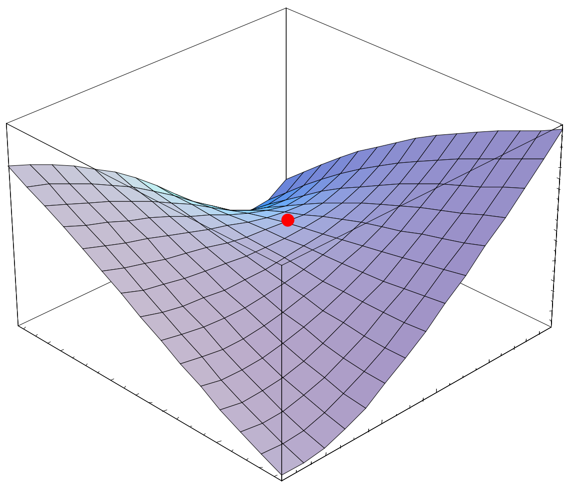


Figure 4. A two-dimensional saddle point.

In order to illustrate our method for locating saddle points, let us first examine the equimagnitude contours (i.e. the contours along which the function has a constant value) of the two-dimensional function shown in Figure 5. This function has three local minima, *A*, *B* and *C*. It can be generally shown that in a very small neighborhood of a local minimum the equimagnitude contours are always ellipses (or, more generally, ellipsoids, if the parameter space is multidimensional). We will call this ellipse the "reference ellipse" for the given minimum. Its half-axes and orientation can be computed with various standard techniques. If some scale parameter is used (e.g. one of the half-axes) it is possible to rescale the reference ellipse to any desired size, by keeping its shape unchanged. The dashed lines in Fig. 5 show the reference ellipse of the minimum *A*, rescaled to four different sizes.

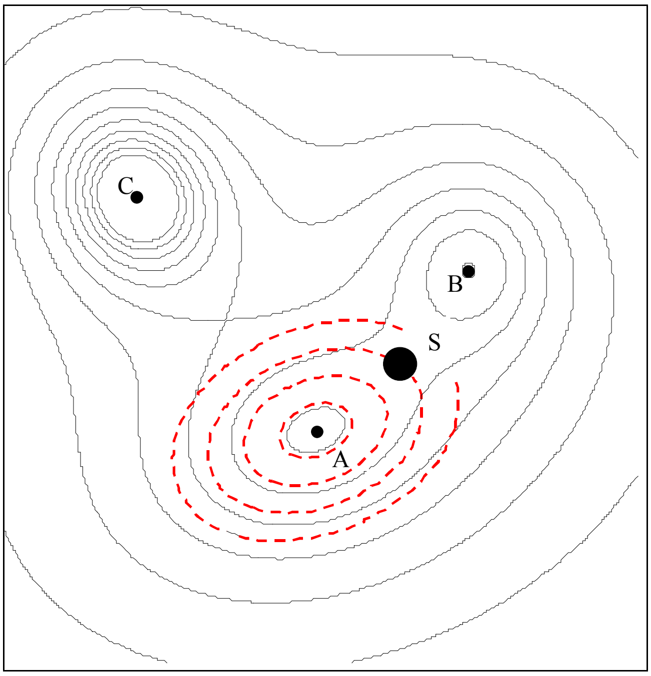


Figure 5. The deviation of the shape of the equimagnitude contours (continuous lines) from (properly rescaled) reference ellipses (dashed lines) is maximal in the direction of saddle points (*S*).

Let us now compare the shapes of the real equimagnitude contours with those of the reference ellipses. The point *S* in Fig. 5 is a saddle point. We observe that far away from saddle points the shapes of the equimagnitude contours are close to those of the ellipses. Close to the saddle point however, the shapes of the equimagnitude contours deviate strongly from the ellipses. Our working assumption will then be that the deviation of the shape of a equimagnitude contour from the reference ellipse is maximal in the direction of a saddle point. A closer inspection of Fig. 5 reveals that there where this deviation of shape is maximal, the merit function along the reference ellipse has a minimum. Therefore, our method to sense the direction of a saddle point is to compute the minimum of the merit function along rescaled reference ellipses. This means basically to do local optimization by imposing the reference ellipse as an equality constraint.

If we are in the local minimum *A* and compute the merit function along the smallest dashed ellipse in Fig.5 we obtain the results shown in Fig.6. Since we are on an ellipse, for each value of one Cartesian coordinate (e.g. x) we obtain two different values of the other coordinate (y) and therefore two values for the merit function. The two merit function curves form together the closed curve shown in Fig.6. Note that we have a minimum of the merit function in the right part of the lower curve. The pair of coordinates that corresponds to this local minimum is plotted in Fig.7 as the right large dot. We observe that this dot points, as expected, in the direction of point *B*.

It is sometimes possible to find several local minima around reference ellipses, which then point towards different saddle points. For instance, by slightly increasing the size of the reference ellipse around point *A* in Fig. 7, it becomes possible to sense the presence of point *C* (left large dot). Similarly, when we started in the local minima *B* and *C* and optimized around (sufficiently large) reference ellipses we could sense the presence of the two other minima, while still being in the basin of attraction of the start minimum.

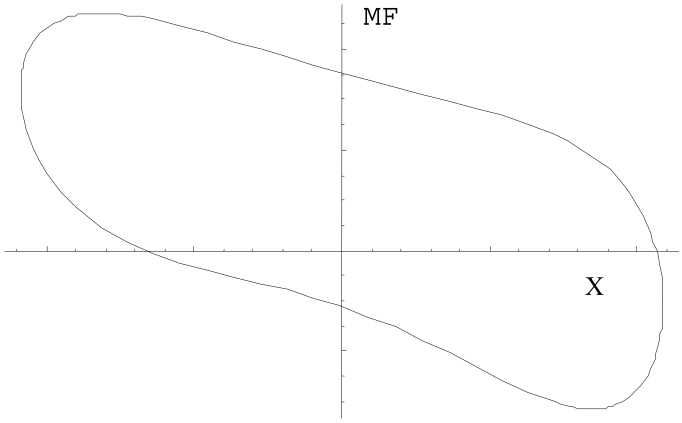


Figure 6. The two merit function curves corresponding to the upper and lower half ellipse form together a closed curve (see text).

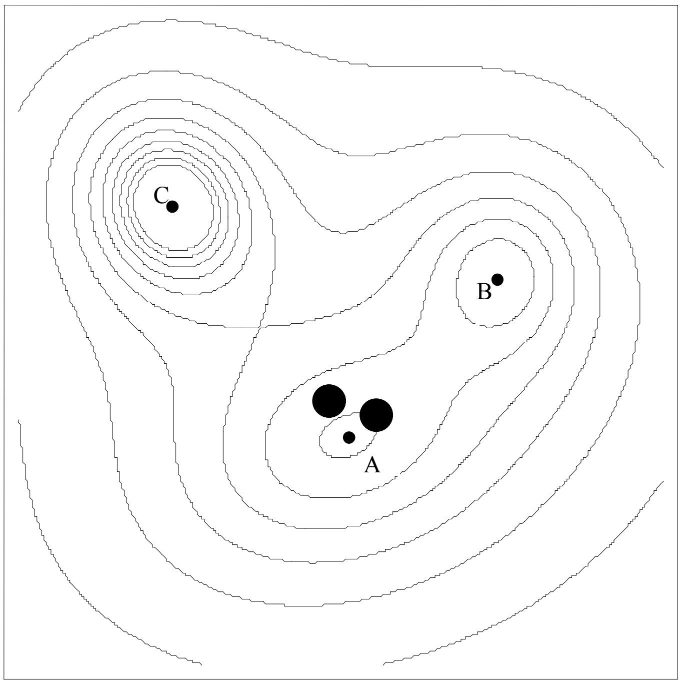
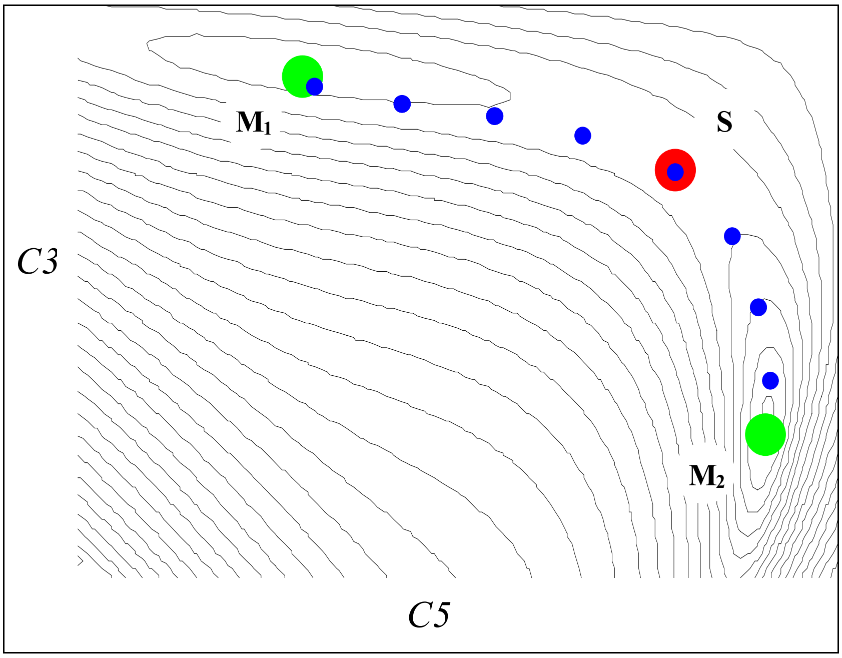
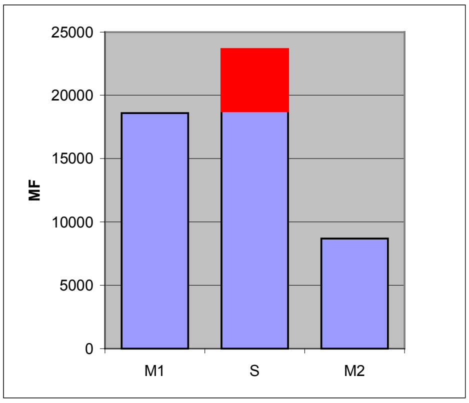


Figure 7. Sensing the directions of the saddle points that lead to the local minima *B* and *C* from the neighborhood of the local minimum *A*.

Once the direction of the saddle point has been determined, the saddle point itself can easily be found. Figure 8a shows a two-dimensional landscape of the default CODE V merit function (which is based on the transverse aberration). The local minimum M_1 corresponds to the triplet shown in Fig. 9a. The optimization variables are the curvatures of the 3rd and 5th surfaces. We determine first the direction of the saddle point as discussed previously (the small dot closest to M_1). Then we track this minimum by gradually increasing the size of the reference ellipse and reoptimizing after each step (the succession of small dots between M_1 and S). At each step (small dot) we compare the value of the minimal merit function with the minimal merit function at the previous step. As long as the minimal merit function increases from one step to another, we keep increasing the ellipse. When the merit function starts to decrease, we have passed the saddle point. Then we are already in the basin of attraction of the new local minimum and if we remove the ellipse constraint and optimize as usual we obtain the new local minimum M_2, which in this case is the well-known Cooke Triplet (Fig. 9c). The system corresponding to the saddle point S is shown in Fig. 9b.

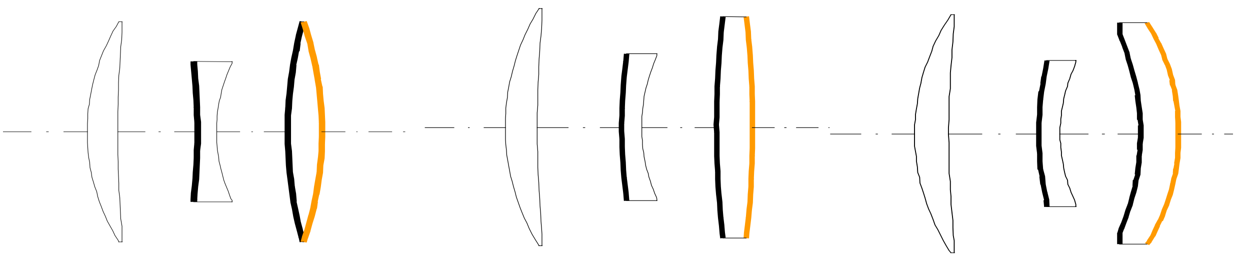


a



b

Figure 8. a. Two-dimensional merit function landscape for a triplet, with two local minima M_1 and M_2 and the saddle point S between them. The two variables are the curvatures of the 3rd and 5th surfaces. b. Bar chart indicating the values of the merit function for M_1 , M_2 and S.



a b c

Figure 9. The triplets corresponding in Fig. 8a to the points M_1 (a), S (b) and M_2 (c). The 3rd and 5th surfaces, whose curvatures are the optimization variables, are shown with thick black lines. The 6th surface, which is controlled by a solve, also changes and is shown with a thick gray line.

In the bar chart shown in Fig. 8b the bars represent the default CODE V merit function for the starting local minimum, the saddle point and the final Cooke Triplet. The difference between the heights of the first two bars (shown in dark) indicates the height of the merit function barrier that must be tunneled. Perhaps this type of information could also be useful for improving the efficiency of some of the global optimization methods.

The same strategy can also be extended to a multidimensional parameter space. At the time of this writing our experience is limited to simple systems only. For triplet configurations for instance, we can find saddle points and the corresponding neighboring minima in four dimensions.

5. CONCLUSIONS

In this paper we have discussed three strategies to escape from a given local minimum. The first two strategies produce sometimes very good results, but suffer, as so many other empirical tools in the arsenal of optical design, from the drawback that their success depends largely on serendipity. The third strategy is different. Even when we did not obtain the expected result, we have often discovered some new and unexpected features of the merit function topography. We believe that this new insight is part of a very valuable learning process. It is too early to say whether this particular method will be successful in the case of complex designs. However, we do believe that optical system design would benefit very much if we had some methods that not only produce practical results, but also provide insight in the complexity of the merit function landscape. Our third method is an attempt in this direction. It is also to early to decide whether asking questions about saddle points will be useful or not, but we do believe that asking *new* questions is essential for improving our understanding of optical design.

6. REFERENCES

1. T. G. Kuper, T. I. Harris, Global optimization for lens design - an emerging technology, Proc. SPIE **1780**, 14-28, 1992
2. G. W. Forbes, A. E. W. Jones, Towards global optimization with adaptive simulated annealing. Proc. SPIE **1354**, 144-151, 1991
3. M. Isshiki, H. Ono, K. Hiraga, J. Ishikawa, S. Nakadate, Lens design: Global optimization with Escape Function, Optical Review (Japan), ***6***, 463-470, 1995
4. K. E. Moore, Algorithm for global optimization of optical systems based on genetic competition, Proc. SPIE **3780**, 40-47, 1999