Finding new local minima in lens design landscapes by constructing saddle points

Florian Bociort, MEMBER SPIE
Maarten van Turnhout
Delft University of Technology
Optics Research Group
Lorentzweg 1
Delft, 2628 CJ
The Netherlands
E-mail: f.bociort@tudelft.nl

Abstract. Finding good new local minima in the merit function landscape of optical system optimization is a difficult task, especially for complex design problems where many minima are present. Saddle-point construction (SPC) is a method that can facilitate this task. We prove that, if the dimensionality of the optimization problem is increased in a way that satisfies certain mathematical conditions (the existence of two independent transformations that leave the merit function unchanged), then a local minimum is transformed into a saddle point. With SPC, lenses are inserted in an existing design in such a way that subsequent optimizations on both sides of the saddle point result in two different system shapes, giving the designer two choices for further design. We present a simple and efficient version of the SPC method. In spite of theoretical novelty, the practical implementation of the method is very simple. We discuss three simple examples that illustrate the essence of the method, which can be used in essentially the same way for arbitrary systems.

Subject terms: saddle point; global optimization; optical system design

1 Introduction

Over the past two decades, impressive progress in global optimization of optical systems has resulted in powerful software tools. For optical designs for which the complexity is not too high, present-day global optimization algorithms are valuable tools for finding a good (perhaps even the best) solution among the many local minima that are found in the merit function landscape. For moving an optical configuration from one local minimum to another, these methods rely almost exclusively on generally applicable mathematical algorithms, rather than on specific optical properties of the design landscape. However, when the number of components is growing, even local optimization becomes time consuming, and it becomes increasingly difficult to apply such tools straightforwardly.

Our studies of the network structure of the set of local minima have shown that not only local minima, but also saddle points are useful for understanding the merit function landscape of optical systems. Minima, saddle points, and maxima are all critical points, i.e., the gradient of the merit function vanishes at these points. An important property of (nondegenerate) critical points is the so-called Morse index. (When critical points merge, they are called degenerate.) A more rigorous discussion of the Morse index and related properties is given in an earlier paper. Intuitively, one can think about a two-dimensional (2-D) saddle point (the surrounding surface has the shape of a horse saddle), which is a minimum along a certain direction and a maximum along the perpendicular direction. Similarly, critical points in an N-dimensional optimization problem have a set of mutually orthogonal directions; along some of these directions, the critical points are minima, along the other ones (called downward directions), they are maxima. The Morse index is the number of downward directions. Thus, for minima and maxima, the Morse index is 0 and N, respectively, and saddle points have a Morse index between 1 and N − 1.

For optimization problems, saddle points with Morse index 1 are of special interest. They are maxima in one direction, which one can visualize as the downward direction of a 2-D saddle point, and they are minima in the remaining N − 1 directions, which are all very similar to the upward direction in the 2-D case. As in a 2-D situation, choosing two points close to each other, but on opposite sides of the saddle and starting local optimizations at those points, will lead to two distinct minima. (An illustration will be given in Sec. 2) The saddle points are therefore special points on the boundary between the basins of attraction that correspond to the two adjacent local minima. (The basin of attraction is the set of starting points that, after local optimization, lead to the same minimum.)

As has been shown earlier, if a local minimum is known, new local minima can be found by detecting Morse index 1 saddle points in the vicinity of the known minimum and then by optimizing the configurations on the other side of these saddle points. A drawback of this method is that detecting Morse index 1 saddle points without a priori information about them is computationally more expensive than finding local minima.
As a computationally effective alternative to saddle point detection, we present here a method, which we call saddle-point construction (SPC), that can be used with success even in the case of very complex systems with many variables and constraints, because it can lead to new system shapes with only a small number of local optimizations.

With the SPC method, saddle points are created by inserting a lens into an existing optical configuration, which is already a minimum in its variable space. Lens designers frequently insert lenses into their designs, and in the traditional way, one new system shape results after optimization. However, when a lens is inserted with SPC, two distinct system shapes result, and for further design, one can choose the better one. By inserting lenses according to the SPC method, and then, if necessary, by extracting lenses, new local minima can be obtained effectively for optical design tasks of arbitrary complexity. Present experience suggests that, if added to the optical designer’s arsenal, this new tool can increase design productivity in certain situations.

Since we use in this paper mathematical ideas that are rather new in optical system design (such as that of a saddle point in a high-dimensional variable space), we will first give an intuitive description of the SPC method in Sec. 2. In Sec. 3, we will discuss SPC rigorously, and we will prove that a system constructed according to our receipt is a saddle point with Morse index 1. Three simple examples that illustrate the essence of the SPC method in Sec. 4 show that the practical implementation of the method is actually very easy. In Sec. 5, additional mathematical properties of SPC will be investigated.

## 2 SPC in a Nutshell

In this section, we give an intuitive description of the SPC method. This method can be used with a broad class of optical merit functions, e.g., with a merit function based on transverse aberrations (root-mean-square spot size), wavefront aberration, aberration coefficients, etc. (For optimizations that must include properties that are not frequently included as optimization operands, the applicability of SPC must be examined separately by using the same reasoning as shown in Sec. 3. If the mathematical conditions for SPC are not satisfied, such unusual operands can be omitted in the first stage. After the two local minima on both sides of the saddle point are obtained, those properties can be included in the merit function again.)

Figure 1 gives an intuitive illustration of the SPC method. We start with an optimized system with \( N \) surfaces. For clarity, only one variable is shown in dark gray in the upper left part of the figure. The original local minimum can have any number and type of variables (e.g., thicknesses, curvatures, etc.). In this system, we insert a meniscus lens with zero axial thickness and equal variable curvatures. Such a meniscus disappears physically and does not affect the path of any ray or the merit function of the system. Therefore, we call this meniscus a null-element. For simplicity, we discuss here the case when the element to be inserted is a lens with spherical surfaces, but the method also works with mirrors and aspherical surfaces.\(^{7,10}\) Although, when inserted, the null-element does physically nothing, it comes with two new variable curvatures, which, when changed during optimization, allow the merit function to decrease.

For some specific values for these two curvatures (see Sec. 3), the null-element transforms the original local minimum into a saddle point with Morse index 1 in the variable space with increased dimensionality. Because such a saddle point contains a null-element lens, we call this saddle point a null-element saddle point (NESP). The new downward direction, shown in Fig. 1 by the lighter gray curve passing through the saddle point, and a new upward direction (not shown, but similar to the dark gray curve through the saddle point) appear in the new variable space (with a dimension increased by 2). Along the downward direction, the new system is a maximum. Since, as will be shown in the next section, the Morse index is 1, despite the fact that typically we have much more than two variables, the surrounding surface of the saddle point resembles very much a 2-D horse saddle. If we choose two points close to the saddle, but situated on opposite sides, and then optimize them, the optimizations “roll down” from the saddle and arrive at two distinct local minima. (The optimization variables are those of the starting local minimum plus the two curvatures of the null-element.)

In the general case, when the insertion position and the glass of the null-element are arbitrary, the curvature of the null-element meniscus that makes the original local minimum a saddle point can be computed numerically.\(^{14}\) In this paper, a simple but still very efficient version of the method is presented.

## 3 SPC: Theory

In the special case, a certain restriction on glass and axial thicknesses is needed. The null-element is inserted in contact (i.e., at zero axial distance) from an existing surface (called a reference surface) in the original local minimum. This reference surface should have a variable curvature, and the glass of the null-element should be the same as the glass of the reference surface. (The conditions for insertion position and glass type are not as restrictive as they might seem, because once the two minima on both sides of the NESP have been obtained, the distances between surfaces and the glass of the lens resulting from the null-element can be changed as desired.) We will show in this section that when the curvatures of the null-element are
equal to the curvature of the reference surface, we obtain an NESP.

Assume that the reference surface is the $k$’th surface of an optimized $N$-dimensional system [Fig. 2(a)]. The value of the merit function is $MF_{\text{ref}}$, and the curvature of the surface in the starting minimum is $c_{\text{ref}}$. After the surface, we introduce a null-element in contact with the $k$’th surface (i.e., we have two consecutive zero axial thicknesses). When the curvatures $c_{k+1}$ and $c_{k+2}$ of the null-element lens are varied, but are kept equal, then the inserted meniscus remains a null-element, and the merit function of the new system with $N+2$ surfaces remains equal to $MF_{\text{ref}}$ [Fig. 2(b)].

Similarly, if the curvatures $c_k$ and $c_{k+1}$ are varied, but are kept equal, the air space before the lens becomes a null-element. When, in addition, we have $c_{k+2}=c_{\text{ref}}$, the merit function of the new system remains again unchanged by the insertion and equal to $MF_{\text{ref}}$ [Fig. 2(c)]. (Note that, although the curvature $c_k$ of the original minimum with $N$ surfaces is varied, the second curvature of the new lens takes its role and ensures that $MF_{\text{ref}}$ remains unchanged.)

When $c_k=c_{k+1}=c_{k+2}=c_{\text{ref}}$, we are in the special situation shown in Fig. 2(d). To examine this case, consider the following transformations:

$$c_k = c_{\text{ref}}, \quad c_{k+1} = c_{k+2} = u,$$

where in both transformations all other variables of the original minimum are kept unchanged. The transformations (1) and (2) describe two lines in the variable space of the new system with $N+2$ surfaces. The position of individual points along these lines is given by the parameters $u$ and $v$. The two lines intersect for $u=v=c_{\text{ref}}$, which gives us:

$$c_k = c_{k+1} = c_{k+2} = c_{\text{ref}},$$

where all other variables have the values of the original minimum. As shown earlier, along both lines, the merit function is invariant and equal to $MF_{\text{ref}}$. Therefore, the system with the property described by Eq. (3) has also a merit function value equal to $MF_{\text{ref}}$.

In the preceding analysis, we assumed that a thin lens with surfaces $k+1$ and $k+2$ is inserted after the $k$’th surface in an existing design that is a local minimum. However, Eqs. (1)-(3) are also valid if a thin lens with surfaces $k$ and $k+1$ is placed before the $k+2$’th surface in an existing minimum. In this case, the invariant lines (1) and (2) are related to the null-airspace lens and the null-glass lens that are formed, respectively.

In optical optimization problems, paraxial equality constraints (e.g., the effective focal length is kept constant) are frequently used. In the following, we analyze the properties of the intersection point (3) in the variable space of the new system with $N+2$ surfaces, and we show that this crossing point is a saddle point with Morse index 1 or 2 if no constraints are used, and with Morse index 1 when the same paraxial constraint is used both in the existing minimum, and during SPC.

For examining the crossing point (3) in the $N+2$-dimensional merit function landscape, it is sufficient to investigate a 3-D subspace of this landscape, defined by the variables $c_k, c_{k+1}$, and $c_{k+2}$ (Fig. 3). The coordinate system in Fig. 3(a) is given by:

$$x = c_k - c_{\text{ref}}, \quad y = c_{k+1} - c_{\text{ref}}, \quad z = c_{k+2} - c_{\text{ref}}.$$  

In the unit cube shown in Fig. 3(a), the invariant lines (1) and (2) are oriented along the vectors:

$$\mathbf{OA} = (1,1,0), \quad \mathbf{OB} = (0,1,1),$$

respectively. It is convenient to rotate the coordinate system so that the points in the plane $OAB$ can be parameterized by only two numbers (instead of three). Since $OA=OB=AB$, the angle between $\mathbf{OA}$ and $\mathbf{OB}$ is 60 deg, and the two lines cannot be used as axes in a rectangular coordinate system. However, an orthogonal axis system in the plane $OAB$ can be easily constructed. The new axes $x'$ and $y'$ are then oriented along the vectors:

$$\mathbf{OR} = \frac{1}{2}(\mathbf{OA} - \mathbf{OB}) = \frac{1}{2}(1,0,-1),$$

and
Bociort and van Turnhout: Finding new local minima in lens design landscapes

\[
OQ = \frac{1}{2}(OA + OB) = \frac{1}{2}(1,2,1),
\]

(7)

see Fig. 3(a). By computing the vector product \( OR \times OQ \), it can be seen that the axis \( z' \) orthogonal to the plane \( OAB \) (not shown) is then oriented along the vector \( v = (1,-1,1) \).

By taking unit-length vectors along \( OQ \) and \( OR \), the position of an arbitrary point in the plane \( OAB \) is given by:

\[
x^2 \frac{2}{\sqrt{6}} OR + y^2 \frac{2}{\sqrt{6}} OQ = \left( \frac{x'}{\sqrt{2}} + \frac{y'}{\sqrt{6}}, \sqrt{\frac{2}{3}} y' - \frac{x'}{\sqrt{2}} + \frac{y'}{\sqrt{6}} \right).
\]

The three curvatures for the points in the plane \( OAB \) are then given by:

\[
c_k = c_{\text{ref}} + \frac{x'}{\sqrt{2}} + \frac{y'}{\sqrt{6}},
\]

(9)

\[
c_{k+1} = c_{\text{ref}} + \frac{\sqrt{2}}{3} y',
\]

(10)

\[
c_{k+2} = c_{\text{ref}} - \frac{x'}{\sqrt{2}} + \frac{y'}{\sqrt{6}}.
\]

(11)

If in a plane, two lines, along which a function is constant, cross, then the crossing point is a 2-D saddle point of that function. In the plane \( OAB \), the two invariant lines (1) and (2) intersect in \( O \) [Fig. 3(b)]. We consider the case when the null-element lens with curvatures \( c_{k+1} \) and \( c_{k+2} \) is placed in contact with the reference surface \( k \) of the \( N \)-dimensional original local minimum, in which \( c_k = c_{\text{ref}} \).

We start by showing that \( O \) is a critical point in the plane \( OAB \). It is well known that the direction of the most rapid variation of a function is orthogonal to the direction along which the function is constant. Therefore, at each point along the invariant lines, the projection on \( OAB \) of the gradient of the merit function is orthogonal to the invariant lines. At the intersection point, the projection of the gradient must be zero, because it cannot point in two different directions. Assuming that \( O \) is nondegenerate, in the plane \( OAB \) the point \( O \) is then a 2-D saddle point. (For a maximum or a minimum, the equimagnitude contours for merit function (MF) values close to \( MF_{\text{ref}} \) are ellipses, which reduce to a point for \( MF = MF_{\text{ref}} \). The point \( O \) cannot be a 2-D maximum or minimum, because the equimagnitude contours are crossing straight lines. We will come back to the assumption of nondegeneracy in Sec. 5.) Along one of the orthogonal axes in the plane \( OAB \) (\( x' \) or \( y' \)), the point \( O \) is a minimum; along the other one, it is a maximum.

Because for the original local minimum, the merit function derivative with respect to \( c_k \) was already zero, the point \( O \) is also a minimum along the direction of \( Ox \), outside \( OAB \). (The preceding analysis is also valid if a thin lens with surfaces \( k \) and \( k+1 \) is placed before the \( k+2 \)th surface in an existing minimum. Then, the point \( O \) is a minimum along the direction of \( Oz \).) In the new coordinate system with axes \( x', y' \), and \( z' \), the direction of \( Oz \) is oriented along the vector \( r' = (r'_1,r'_2,r'_3) = (3,1,2) \). Since \( O \) is a minimum along the direction of \( Ox \), the derivative of \( MF \) with respect to \( x', y' \), and \( z' \) is zero along the direction of \( r' \):

\[
\nabla_r MF = \frac{\partial MF}{\partial x'} r'_1 + \frac{\partial MF}{\partial y'} r'_2 + \frac{\partial MF}{\partial z'} r'_3 = 0.
\]

(12)

In the plane \( OAB \), the partial derivatives \( \frac{\partial MF}{\partial x'} \) and \( \frac{\partial MF}{\partial y'} \) at \( O \) vanish. Because \( r'_3 \neq 0 \), the partial derivative \( \frac{\partial MF}{\partial z'} \) should be equal to zero as well to satisfy Eq. (12).

Since the partial derivatives with respect to \( x', y' \), and \( z' \) are zero, the merit function derivatives with respect to \( y \) and \( z \) must be both zero at \( O \). In addition, because the variables of the original local minimum other than \( c_k \) are kept unchanged, the merit function derivatives with respect to them remain zero as well. Therefore, all components of the gradient of the merit function vanish at \( O \), making \( O \) a critical point.

We have shown that \( O \) is a maximum in one direction in the plane \( OAB \), that it is a minimum in the orthogonal direction in that plane, and that it is also a minimum with respect to the variables of the original local minimum other than \( c_k \). The only direction that remains to be studied is that of \( OZ' \). If there are no equality constraints, along this direction \( O \) can be a minimum or a maximum, and the Morse index is then 1 or 2, respectively. (In case there are no constraints, it can be investigated numerically whether \( O \) is a minimum along the direction \( OZ' \). Otherwise, the Morse index will be 2, and the procedure for generating local minima from the saddle point must be adapted, but we have not encountered such situations yet.)

In the following analysis, we demonstrate that when we use paraxial constraints, \( OZ' \) cannot be an additional upward direction and the Morse index remains 1, because these constraints are violated along the direction of \( OZ' \). We first show that if the original minimum satisfies a paraxial constraint, then all points in the plane \( OAB \) satisfy this constraint, because for these points, paraxial ray paths at surfaces other than \( k, k+1 \), and \( k+2 \) are left unchanged. We can write the total power \( K_{\text{tot}} \) of surfaces \( k, k+1 \), and \( k+2 \) that are changed in Fig. 2 as:

\[
K_{\text{tot}} = K_k + K_{k+1} + K_{k+2},
\]

\[
= (n - 1)(-c_k + c_{k+1} - c_{k+2}),
\]

\[
= (n - 1)(-x + y - z - c_{\text{ref}}),
\]

\[
= (1 - n)(c_{\text{ref}} + xv_x + yv_y + zv_z),
\]

(13)

where \( n \) is the refractive index of the null-element meniscus and of the lens with which it is in contact, and \( v = (1, -1, 1) \) is the normal vector to the plane \( OAB \). Note, however, that

\[
xv_x + yv_y + zv_z = 0,
\]

(14)

is the equation for the plane passing through the origin, perpendicular to \( OZ' \) (which is oriented along the vector \( v \)), i.e., for the plane \( OAB \). The total power \( K_{\text{tot}} \) of surfaces \( k, k+1 \), and \( k+2 \) remains constant in the plane \( OAB \) and is
equal to the power \((1-n)c_{\text{ref}}\) of surface \(k\) of the original local minimum (in which \(c_k = c_{\text{ref}}\)).

Because of zero axial thicknesses between surfaces \(k\), \(k+1\), and \(k+2\), the ray heights at surfaces \(k+1\) and \(k+2\) of any paraxial ray are all equal to the ray height \(h_k\) at surface \(k\), and the angle \(u_{k+2}\) after the null-element is given by:

\[
u_{k+2} = nu_{k-1} - h_k K_{\text{tot}},\]

where \(u_{k-1}\) is the ray angle before refraction at surface \(k\). Note that if \(K_{\text{tot}}\) is kept constant, \(u_{k+2}\) remains constant as well, and the entire paraxial ray path remains unaffected.

The preceding analysis shows that the paraxial properties of the entire system remain invariant when \(c_k\), \(c_{k+1}\), and \(c_{k+2}\) are changed in the plane \(OAB\), and that the paraxial constraint is violated in the direction of \(OZ'\), normal to the plane \(OAB\). Therefore, \(O\) is an NESP with Morse index 1 in the paraxially constrained variable space. (If constraints on real rays are used instead of paraxial ones, it is not expected that the Morse index will change, because then one of the eigenvalues of the Hessian matrix should change from positive to negative. The real ray properties are not so different from the paraxial ray properties to expect a change in sign of an eigenvalue.)

Before inserting the meniscus lens, the original local minimum should be properly optimized so that the residual gradient of the merit function is sufficiently close to zero. By inserting the null-element in such a way that a saddle point is created, we again obtain a system that has zero gradient. For optimizing this system, we first construct two starting points on opposite sides of the saddle. This can be done for instance by slightly perturbing two consecutive surfaces, which are in contact in the NESP, with a small change ±\(\epsilon\) in the surface curvatures:

\[
c_k = c_{k+1} = c_{\text{ref}} \pm \epsilon, \quad c_k = c_{\text{ref}},\]

\[
\text{the points } P_1 \text{ and } P_2 \text{ in Fig. 3(b)}, \text{ or alternatively:}\]

\[
c_k = c_{k+1} = c_{\text{ref}} \pm \epsilon, \quad c_{k+2} = c_{\text{ref}},\]

\[
\text{the points } P_3 \text{ and } P_4 \text{ in Fig. 3(b)}, \text{ where } c_{\text{ref}} \text{ is the curvature of the reference surface in the starting local minimum. (The procedure remains valid if radii are used instead of curvatures.) Perturbing the saddle point in this way has the advantage that in both starting systems, we still have a (glass or air) null-element that does not affect the ray paths. Therefore, if the original local minimum already satisfies the required optimization constraints, the starting points will satisfy those constraints automatically.}\]

Since a saddle point has zero gradient, the two starting points have a small value of the gradient for small values of \(\epsilon\). In order to obtain the desired outcome of the subsequent local optimization, the residual gradient of the original local minimum must be significantly smaller than the \(\epsilon\)-dependent gradient at the two starting points.

By optimizing the two points obtained with either Eq. (16) or Eq. (17), we obtain two different local minima. If \(\epsilon\) is chosen in the correct range, and the basins of attraction are well behaved, the same pair of local minima is obtained from Eq. (16) or Eq. (17). (For this reason, it should be noted that this way of choosing starting points on both sides of the saddle can conflict, in certain situations, with choices that have been made in the software implementation of local optimization algorithms. Ideally, the two starting points should be points situated deeper within the two basins of attraction that correspond to the two adjacent local minima. However, the basin shapes depend on implementation details of the local optimization algorithm (e.g., damping method), and we have found examples where the two starting points are also close to the boundary between the two basins. In such cases, the outcome of local optimization started at those points can become less predictable. However, when this problem occurs, typically only one pair of points seems to be affected, either the one obtained with Eq. (16) or the one with Eq. (17). The other pair is still well-behaved and is adequate for SPC. Other choices of the pair of points on opposite sides of the saddle are also possible.) Using only one pair of points, either (16) or (17) is sufficient. Once the two minima on both sides of the NESP have been obtained, the distances between surfaces and the glass of the lens resulting from the null-element can be changed as desired. Technically, it is easier to increase thicknesses in the two resulting minima than in the saddle point itself. However, many NESPs continue to exist as saddle points in the merit function landscape when in them the thin-lens thickness is increased.\[17\]

### 4 SPC: Examples

In this section, we illustrate the special case of the SPC method with three examples. A detailed description of all necessary steps is given in an earlier publication.\[19\] For CODE V,\[20\] lens files for all examples and a macro that creates the NESP are available via our website.\[21\] All three examples can be executed in a very short time. By optimizing two points at opposite sides of the saddle (the points given by Eq. (16) or Eq. (17)), we obtain two different local minima with zero distances between surfaces inherited from the null-element. In these local minima, we gradually increase the zero distances to the desired values. The steps should be small enough to avoid jumps to other local minima. After each increment, we reoptimize the system.

The examples are independent of the optical design software used. We have tested all examples in CODE V and ZEMAX,\[22\] obtaining the same results. In these examples, we use a merit function that is based on transverse aberrations (root-mean-square spot size) with respect to the chief ray. In the merit function, all wavelengths and fields have a weight factor of unity. The object is placed at infinity, and the image is kept at its paraxial position. The control of edge thickness violation was disabled when optimizing the two starting points situated on both sides of the saddle. With two zero thicknesses, an edge thickness violation can easily appear but may disappear when the zero thicknesses are increased in the resulting local minima.

#### 4.1 Example 1: Generating Doublets from a Singlet

This very simple example serves two purposes: it illustrates the special case of the SPC method, which can be used in essentially the same way in all cases (including very complex systems), and also shows an advantage of SPC. When using SPC to insert a lens so that an NESP is created, two systems result after optimization. Inserting or splitting a lens in the traditional way results in a single system, which is not necessarily the better one. In this example, the better...
Bociort and van Turnhout: Finding new local minima in lens design landscapes…

The six resulting quintet minima, three of them (one resulting from each NESP) are identical, which illustrates the property we call convergence that we have observed frequently with SPC: the same final design can be obtained in several different ways.

As a starting system, we use an optimized monochromatic quartet \( f = 2 \text{ (number 2, field of 14 deg)} \), with the first seven curvatures used as variables [Fig. 6(a)]. The last surface is used to keep the effective focal length constant. All lenses have the same glass.

We first construct a quintet NESP by inserting a null-element (with variable curvatures and the same glass as the existing lenses) in contact with the second surface of the quartet. We perturb the saddle according to Eq. (16) or Eq. (17), and after optimization and increasing of thickness of the second lens of the quintet, we obtain the local minima shown in Figs. 6(b) and 6(c). [In Fig. 6(b), the axial distance between the first two lenses has also been increased in order to remove edge separation violation.]

Similarly, we construct two other quintet NESP’s by inserting a thin-lens meniscus at the first and third surface of the quartet, respectively. From the resulting four quintet local minima, two of these turn out to be identical to the one shown in Fig. 6(c). This is a very simple example of convergence to the same system via three different design routes. For configurations that have this property, if for any reason a design route that should be successful accidentally misses the goal (e.g., sometimes instabilities in local optimization influence the outcome), the same goal can be achieved via another design route of the same kind. For more complex examples of convergence, see Ref. 17.

4.3 Example 3: Obtaining a Double Gauss Design

Frequently, experienced designers observe that the shape of a design is probably not the best possible one. From the optimized system shown in Fig. 7 \( f = 3.33, \text{ field of 14 deg} \), we want to obtain a Double Gauss shape. Here we

---

**Fig. 6** SPC at the second surface of a quartet. (a) The starting quartet system; (b) and (c) the two resulting quintet local minima.

**Fig. 7** Starting system in example 3.
show how we can move from one local minimum to a better one first by inserting and then by extracting a lens. The optimization variables are all lens curvatures, except the curvatures of the two plane cemented surfaces, which are kept unchanged, and the last curvature, which is used to keep the effective focal length constant. First, we use the SPC method at the second surface of the starting system to construct an NESP.

On one side of the NESP, we optimize the point corresponding to point $P_1$ in Fig. 3(b) [i.e., we take the $+$-sign in Eq. (16)], and we obtain the local minimum shown in Fig. 8(a). [As mentioned earlier, edge thickness control is temporarily disabled. Therefore, the second lens, which in Fig. 8(a) seems to be a negative lens, has actually positive power.] Next, we gradually increase the thickness of the second lens to the same value as that of the first lens. The resulting lens system with 12 surfaces is shown in Fig. 8(b). Then, we remove the first lens in two steps: after gradually decreasing its thickness to zero [Fig. 8(c)] (and reoptimizing the system after each decrease in thickness), we make both curvatures of the first lens equal and reoptimize the system (with the first two curvatures fixed) [see Fig. 8(d)]. In this way, the first lens becomes a null-element, which can be removed without changing any ray path. The resulting system is a local minimum with the same number of lenses as the starting one, but with a much lower merit function value [Fig. 8(e)]. The final shape resembles the well-known Double Gauss design. Similar techniques have been applied successfully in the design of state-of-the-art lithographic objectives. Sometimes, after SPC insertion, some lens in the resulting system seems to have no role any more. Such a lens is then a good candidate for removal, even when it is situated farther away than in the preceding case from the position of SPC insertion.

5 Nondegenerate and Degenerate Merit Function Behavior during SPC

For a better understanding of the properties of SPC, Fig. 9(a) shows the equimagnitude contours (i.e., the contours along which the merit function is constant) close to the NESP constructed at the second surface in the previous example, computed numerically for points in the plane $OAB$ given by Eqs. (9)–(11). The NESP with the property $c_2=c_4$ is located exactly in the middle of the figure, where two equimagnitude lines cross. In the near vicinity of the crossing point $c_2=c_4$, the MF landscape in the plane $OAB$ has the shape of a 2-D saddle. The dark regions correspond to regions with low MF values. Figure 9(b) shows the MF values along the horizontal axis (thin line) and vertical axis (thick line) in Fig. 9(a).
the thick curve has two minima and a local maximum at the position of the NESP. The thick curve indicates the two sides of the saddle where the MF initially decreases. Further from the crossing point, the MF increases in all directions of the plane $OAB$.

This example of merit function behavior in the plane $OAB$ illustrates important properties of the SPC method. In Sec. 3, it was assumed that the saddle point $O$ in the plane $OAB$ is nondegenerate. In the absence of degeneracy, when two equimagnitude lines cross in a plane, the crossing point is a saddle point. For instance, for $f(x,y) = xy$ (the plot of this function is very similar to the saddle surface in Fig. 1), the crossing equimagnitude lines for $f(x,y) = 0$ are the two lines $x = 0$ and $y = 0$, which correspond to the two axes of the coordinate system. The origin $[f(0,0) = 0]$ is then a saddle point: it is a minimum along the line $x = y$, where $f(x,y) > 0$ for all nonzero $x$ and $y$ values and a maximum along the line $x = -y$ where $f(x,y) < 0$.

However, from a mathematical point of view, two crossing equimagnitude lines do not necessarily indicate the presence of a saddle point. For instance, for $g(x,y) = x^2y^2$ (see Fig. 10), the same two axes of the coordinate system are also equimagnitude lines for $g(x,y) = 0$, but the origin with $g(0,0) = 0$ cannot be a saddle point, because we have $g(x,y) > 0$ for all points that are not on the axes and there is no direction along which the origin is a maximum. In fact, all points on the two axes are critical points: they are minima in one direction and are flat (i.e., $g$ is constant) in the orthogonal one. Such critical points are called nonisolated because in arbitrarily small vicinities of any critical point, we can find other critical points. Because of the flatness in one direction, nonisolated critical points are degenerate, but mathematically, as Poston and Stewart write, they are “in a strong sense extremely uncommon, so for many purposes may be ignored.”

The remarkable property revealed by Fig. 9 is that, while close enough to the crossing point of the invariant lines given by Eqs. (1) and (2) the MF behavior is nondegenerate, far away from the crossing point the behavior of MF tends to become degenerate. In the latter case, the points on the invariant lines tend to become nonisolated critical points (i.e., perpendicular to the invariant lines they are minima, and, obviously, along the invariant lines these points are flat).

While from a mathematical point of view one may always expect a nondegenerate region around the point constructed with SPC, in practical situations one can encounter situations where the numerically observed behavior is practically degenerate. It is well known that, if a lens is inserted at an inadequate position in an optimized system, subsequent optimization cannot decrease the merit function significantly. The thick curve with two minima and a maximum in Fig. 9(b) gives an intuitive idea about what can be expected in the entire multidimensional variable space around the saddle point. We now want to understand SPC behavior at a position where the system does not “need” extra degrees of freedom for improvement. If the nondegenerate region close to the saddle point (not only in the $OAB$ plane, but in the entire variable space) is too small and the degenerate behavior is dominant, the positions and heights of the two minima on both sides of the saddle point can be so close to those of the saddle point itself that SPC becomes useless. Identifying the best insertion positions for SPC in an existing design is an important issue that requires further research. In the Double Gauss example, the insertion position for which SPC was successful was determined on the basis of previous knowledge about the desirable outcome.

6 Conclusion

One of the major difficulties in present-day global optimization is that the computing time increases significantly when the dimensionality of the optimization problem is increased. The saddle-point construction (SPC) method for finding new local minima suffers much less from this drawback.

In this paper, a simple and efficient version of the SPC method is presented. We prove that if the dimensionality of the optimization problem is increased in a way that satisfies certain mathematical conditions [the existence of two independent transformations, Eqs. (1) and (2), that leave the merit function unchanged], then a local minimum is transformed into a saddle point. In lens design, we transform a local minimum into a saddle point by adding a null-element meniscus (which does not affect the path of any ray or the merit function of the system) in contact with an existing lens. The null-element comes with two new variables (the two surface curvatures), and when the values for these curvatures are equal to the curvature of the contact surface (the reference surface), and it is made from the same glass as the lens at the reference surface, it transforms the local minimum into a null-element saddle point (NESP) in the variable space with increased dimensionality. After optimization, two new local minima result from the NESP.

Lens designers frequently insert lenses into their designs, and in the traditional way, one new system shape results after optimization. However, when a lens is inserted with SPC, two distinct system shapes result, and for further design one can choose the better one. By inserting lenses according to the SPC method, and then, if necessary, by extracting lenses, new local minima for optical systems of arbitrary complexity can be obtained very rapidly.

When lenses are inserted in the traditional way, the position of the new starting system within the basin of attraction of a given minimum is not known. Therefore, if several insertion trials are made, they may lead after optimization
to the same minimum or they may miss certain basins of attraction entirely. Since saddle points are points on the basin boundaries, changing the starting points for optimization with SPC increases the diversity of the results obtained after optimization. Also, for simple systems we have examined in detail all possible design shapes obtained with different global optimization algorithms. (Different local minima that correspond to closely resembling systems are considered to have the same design shape.) As will be shown in detail elsewhere, all design shapes we have encountered can be obtained with the special version of SPC discussed in this paper combined with the general SPC method.

In principle, SPC should also be applicable in other optimization problems, where it is possible to define a null element and to find two independent transformations, similar to Eqs. (1) and (2), that leave the merit function unchanged, e.g., in thin-film optimization. However, in applications other than lens design, more research is needed to investigate the practical utility of SPC. Work is presently done on describing SPC in a broader context, independently of lens design.

Acknowledgments

This research is supported by the Dutch Technology Foundation STW (Project No. 06817).

References

22. ZEMAX Development Corporation, ZEMAX, Bellevue, WA.

Florian Bociort teaches optics at Delft University of Technology, Delft, The Netherlands. His research interests include optimization, with special emphasis on the topology of high-dimensional merit function landscapes, and aberration theory. He holds a PhD in physics from the Technical University of Berlin, Germany.

Maarten van Turnhout is a PhD student in the Optics Research Group of Delft University of Technology. He received his physics diploma from the same university in 2004. He is interested in optical system optimization methods.