Singularities in a Teacup

Good Mathematics from Bad Lenses

Rajaram Nityananda

Standard presentations of optics concentrate on ideal systems made for imaging which bring all rays from a point source to one focus. But, in Nature, or in realistic optical systems with defects, rays do not behave precisely in this way. Rather than the focus simply being blurred, the rays, after reflection or refraction, form beautiful and rather universal patterns of bright lines known as caustics. Mathematically speaking, a family of rays is best viewed as a surface in a higher-dimensional space where we keep track of both the position and direction of rays. The intensity enhancement on approaching the caustic line is a singularity, arising from projection of a smooth surface from higher dimensions to lower dimensions. The universal features of such singularities, which arise in many contexts beyond optics, formed a major theme of Vladimir Arnold's work after 1965, when he was exposed to René Thom's vision of 'catastrophe theory'. Arnold and his school made seminal contributions to singularity theory.

Rajaram Nityananda works at the School of Liberal Studies, Azim Premji University, Bangalore. Earlier, he spent a decade at the National Centre for Radio Astrophysics in Pune, and more than two decades at the Raman Research Institute in Bangalore. He has taught physics and astronomy at the Indian Institute of Science in Bangalore and the IISERs in Pune and Mohali. He is keenly interested in optics in a broad sense, and more generally, problems, puzzles, paradoxes, analogies, and their role in teaching and understanding physics.

One of the standard topics we study in school is the action of a spherical mirror. *Figure* 1 shows a set of parallel rays all coming to a focus. We can also think of the family of rays as perpendicular to a surface, the so-called 'wavefront'. We can then say that the mirror converts a plane wavefront to a spherical wavefront, both shown in *Figure* 1.

Let us remember that this kind of focusing is an approximation, even for a spherical mirror. One is taught that parallel rays would focus at a point F distant half the radius from the centre, but this is not the whole truth. This statement is true only for rays for which the angle of incidence is small, as in *Figure* 1. The full picture is shown in *Figure* 2 (right), drawn to show large angles of incidence. We see from simple geometry that when the angle

Keywords
Caustics, fold and cusp catastrophes, singularity theory.

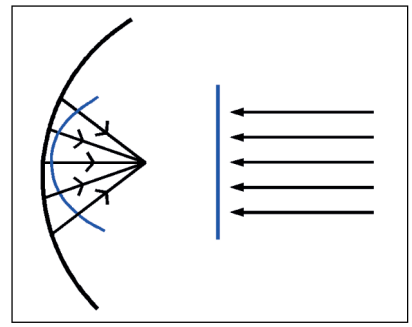


Figure 1. A family of parallel rays falls on a spherical mirror, and converges to a focus. The incident rays actually reach the mirror, but are shortened in the diagram to make the focussing action clearer. The coloured lines show the plane incident wavefront and the reflected, converging spherical wavefront. This focussing by a spherical mirror works only for rays whose angle of incidence is small.

In fact, many of you may have noticed this cusp shape when you look at the bottom of your empty teacup, either in sunlight or indoors with a bulb or compact fluorescent lamp.

of incidence is 60 degrees, the reflected ray meets the axis at the pole, P of the mirror, well away from the usual focal point F. This family of rays is seen to be tangent to a curve which forms a sharp cusp at F. This can be checked without elaborate calculations, by simple observation. In fact, many of you may have noticed this cusp shape when you look at the bottom of your empty teacup, either in sunlight or indoors with a bulb or compact fluorescent lamp. The rays reflected from the inside of the cup form a characteristic pattern (*Figure* 2, left). This does not require a perfect circular cross section of the tea cup, because cusps are seen much more widely. Everyone who wears glasses on which raindrops have fallen, and has looked at a distant street light or vehicle headlight through the drop, has seen cusped patterns. Clearly, some general mathematical principle is at work, making such a cusp shape universal.

Now let us think of another situation – sunlight shining down on a swimming pool, which has waves on it. We can think of parts of the surface as imperfect convex lenses, but other parts would look

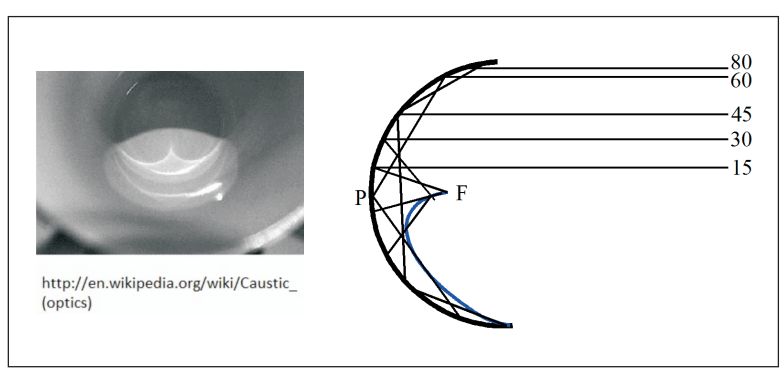


Figure 2. The drawing to the right shows the rays reflected off a semicircular mirror, including those at large angles of incidence. It is seen that they do not pass through the usual focal point at half the radius. (Strictly speaking, only the one along the axis does!) All the rays are tangent to a curve which has a cusped shape near the focus. The picture on the left, from Wikipedia, is a demonstration of the cusp by reflection off a teacup. The additional bright lines are due to multiple reflections.

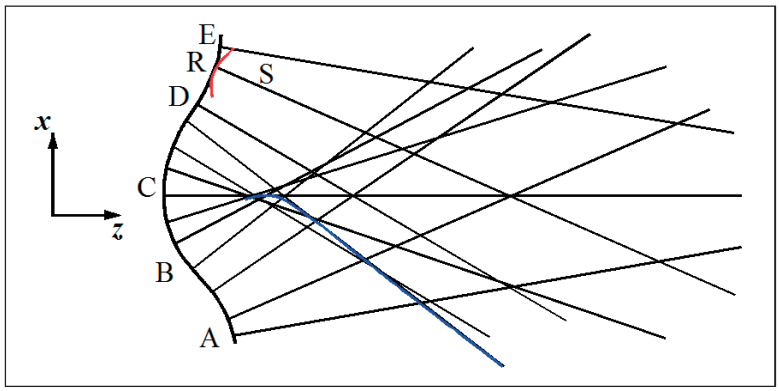


Figure 3. ABCDE is a small piece of the kind of wavefront which might be produced by an imperfect lens, such as a wave on a swimming pool. While rays near C focus, rays further away are seen to be crowded on a caustic line (shown in blue). Only half of the caustic, formed by rays between C and E, is shown for clarity. Mathematicians call such a line, to which the family of rays are tangent, the 'envelope' of the family. Notice that we have three rays passing through a given point inside the caustic, but only one outside.

like concave lenses. To understand how the rays behave after they undergo refraction at such a surface, we show a small piece of a wavefront which is not spherical (*Figure* 3). A first guess would be that the rays would not focus properly, and hence one would get some kind of blur. What actually happens is more interesting. The rays do not indeed pass through a single point, but are seen to be crowded along lines near which the intensity is high. These lines can be understood as follows. A pair of neighbouring rays cross at a point, but this point itself changes as we consider different rays, unlike the case of perfect focusing. The set of these 'foci' is a curve which is tangent to these rays, and is called a 'caustic'. The same word is used for chemicals, or verbal remarks, which burn. I would guess that early attempts to burn objects with lenses gave rise to this terminology for caustics in optics. Our figure is two dimensional so the caustics are lines. In three dimensions the caustics are surfaces. The surfaces intersect the bottom of the pool along bright lines which is what we see in *Figure* 4.

Figure 4. Intensity pattern at the bottom of a sunlit swimming pool, from the 'sketchucation.com' website. Note the bright lines which are a network of caustics, caused by refraction at the wavy surface of the pool.

The appropriate tool for analysing the behaviour of families of light rays is the notion of 'phase space', invented by the Irish genius William Rowan Hamilton about two hundred years ago. In fact, one of his papers is called 'Theory of systems of rays'. For simplicity, we describe the idea in two-dimensional space, instead of three (*Figure* 5). The initial wavefront is described by the thick curve on the left of the figure. This surface is compared to a reference plane wavefront, also shown. The z-axis is chosen perpendicular to this reference plane. The shape

of the wavefront is described by the function $S(x)$. The coordinate x specifies a point on the wavefront, and S is the path delay at x (with reference to the plane wavefront). The rays are drawn normal to the wavefront. Because the wavefront is not parallel to the reference plane, the rays are tilted with respect to the z axis. The tilt angle at the point x is denoted by $p(x)$. This tilt is related to the slope of the wavefront, as illustrated in *Figure* 5.

Now we would like to know how the family of rays behaves as we allow it to move forward. We use the symbol z for the distance travelled perpendicular to our reference plane, from the initial point. From the geometry of the figure, it is clear that a ray with positive p will move upwards, i.e., to larger x, while one with negative p will move downwards, i.e., x will decrease. This behaviour is contained in the simple equation, $x(z) = x(0) + pz$. In writing this, we have made the approximation that the angle p is small, so that the tangent of the angle is replaced by the angle itself, in radians of course. Also, z is being measured from the wavefront. Note that the value of p associated with a given ray does not change, because light travels in straight lines once inside the swimming pool.

This is a very simple equation, but it has very interesting consequences for the family of rays. To see this, we represent the wavefront in *Figure* 2 by the curve ABCDE in phase space, i.e., a plot of $p(x)$ versus x, the blue curve in *Figure* 6. Note that the value of p is positive for negative x and negative for positive x, to start with at $z = 0$. As z increases, our equation tells us that we have to move points with positive p to the right (increasing x) and points with negative p to the left (decreasing x). (The point at $x = 0$,

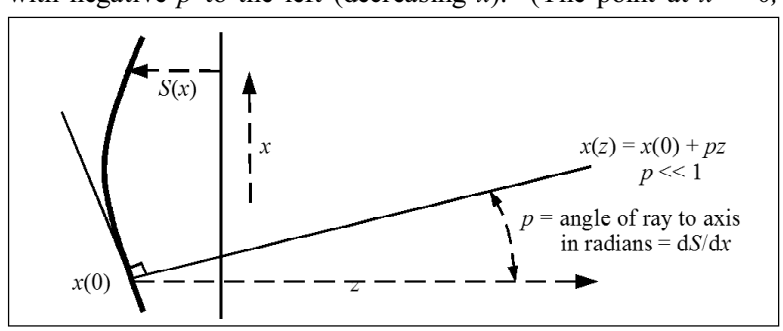


Figure 5. The concept of phase space illustrated with a curved wavefront. The shape of the wavefront is given by the function $S(x)$ which measures the path delay with respect to a reference plane. The slope of this wavefront also gives the angle made by the ray to the z-axis perpendicular to this reference surface. The angle is denoted by p. As a single ray moves along z, the transverse co-ordinate x changes, proportional to the angle p.

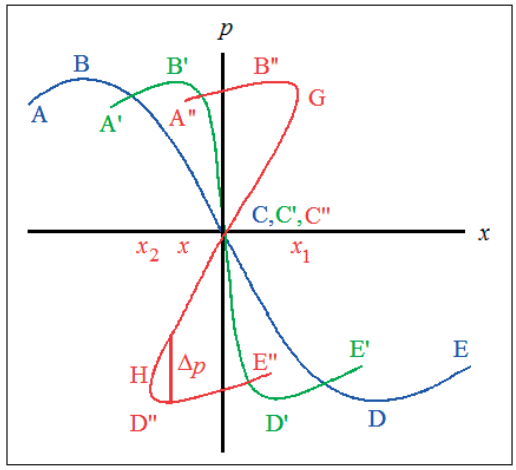


Figure 6. A phase space representation of the family of rays shown in *Figure* 3. The blue curve represents the situation at the initial wavefront, the green curve when the caustic just begins to form (the tip of the cusp), and the red curve the behaviour at a value of z to the right of the cusp. The three points C C' and C" coincide with the origin of the x–p plane and correspond to a ray parallel to the z-axis.

$p = 0$ remains at $x = 0$, the origin of the x–p plane). The green curve shows a situation when the tangent of the curve at the origin has become vertical. For larger values of z, the red curve A″ B″ C″ D″ E″ bends over. It is still perfectly smooth, but there are now three values of p for values of x falling between x_1 and x_2.

What does this mean? Going back to *Figure* 3, we see that for sufficiently large z (i.e., to the right side of the diagram), there are points where three rays intersect, so one indeed has three values of p for values of x in some range. At the boundary of this region, (the points G and H on the red curve in *Figure* 6) the $p(x)$ curve has a vertical tangent, i.e., locally many values of p fall into a small region of x. Put more simply, the rays crowd together near one value of x. This is the caustic. Near the points G and H, the curve can be approximated by a parabola, and hence the properties such as the behaviour of intensity can be worked out and will be the same in all such situations. Because it arose from the curve 'folding over', G and H are called 'fold caustics'. In space, the caustics are located at x_1 (projection of G onto the x-axis) and x_2 (the projection of H onto the x-axis). One can see from the parabolic shape that the angle Δp between the two rays meeting at a point like x just inside the fold caustic will grow as the square root of the distance from the caustic, i.e., proportional to $(x - x_2)^{1/2}$ (*Figure* 6). Also, let us try and understand the intensity distribution near the caustic. Assume that the energy contained in a range

Put more simply, the rays crowd together near one value of x. This is the caustic. Near the points G and H, the curve can be approximated by a parabola, and hence the properties such as the behaviour of intensity can be worked out and will be the same in all such situations. Because it arose from the curve 'folding over', G and H are called 'fold caustics'.

Qualitatively, we can see that many values of p pile up near the same value of x (either x_1 or x_2) and hence the intensity is very high near and inside the caustic. This is the non-mathematical explanation of the crowding of the rays as in *Figure* 3.

dp is a smooth function of p. Then one sees that because of the vertical tangent at G and H, the energy per unit x will not be a smooth function of x. Qualitatively, we can see that many values of p pile up near the same value of x (either x_1 or x_2) and hence the intensity is very high near and inside the caustic. This is the non-mathematical explanation of the crowding of the rays as in *Figure* 3. Quantitatively, the intensity, I, is given by the following calculation $\mathrm{d}I/\mathrm{d}x = (\mathrm{d}I/\mathrm{d}p)(\mathrm{d}p/\mathrm{d}x)$. The factor $\mathrm{d}p/\mathrm{d}x$ behaves like $(x_1 - x)^{-1/2}$. The other factor depends only on the distribution of intensity in the initial wavefront, since p does not change as z increases. This is assumed to be a smooth function.

To understand the cusped shape of the caustic in *Figure* 3, first choose a value of z corresponding to the red curve, i.e., go far enough from the initial wavefront so that the rays cross (as shown by the folding over of the curve in phase space).

Let us now decrease z so that the two points F and G both approach C. The limiting case is the green curve, when they merge, let us say this happens at $z = z_c$. We can see that near the point C, the phase curve has a 'point of inflection'. The first and second derivatives vanish here and hence its shape is given by the equation x proportional to p^3, neglecting higher terms. This allows us to work out what the separation between x_1 and x_2 shrinks to zero as $z - z_0$ tends to zero from above. This calculation is given in *Box* 1, and tells us that $(x_1 - x_2)$ is proportional to $(z - z_c)^{32}$. This explains the cusp shape of the caustic.

Now we can stand back and see that the results did not depend on some special property of the original wavefront. In fact, it was important that it was *not* a special wavefront like a small piece of a sphere, or a plane. The mathematical idea behind this is sometimes stated in the following words: "Consider a wavefront in general position". This implies that any small disturbance to the shape of the wavefront may move the caustics, and the location of the cusp, but not change the square root law for the angle between the rays near a fold, the inverse square root law for the intensity near a fold caustic, or the 3/2 power law for the separation of the

Now we can stand back and see that the results did not depend on some special property of the original wavefront. In fact, it was important that it was *not* a special wavefront like a small piece of a sphere, or a plane. The mathematical idea behind this is sometimes stated in the following words: "Consider a wavefront in general position".

Box 1. Calculating the Shape of the Cusp

The crucial step to understanding the cusp is to express x as a function of p instead of the other way around. Denoting $(x_1 - x_2)$ by X and $(z - z_c)$ by Z, the family of phase space curves in the vicinity of the cusp, i.e., $Z = 0$, is described by $X = ap^3 - bZp$ with both a and b greater than zero. This form can be made plausible by sketching the curves. One can see that if $Z > 0$, then there can be three roots for p for a given x, while for $Z < 0$, there is always only one root. However, the actual proof, crucial to singularity theory, that a general smooth case can be reduced to this form requires the kind of higher mathematics pioneered by Whitney and Thom. One can now see that the points G and H are the points where the tangent is horizontal (we earlier said vertical, but remember, we are now plotting x as a function of p!).

Differentiating X with respect to p and equating to zero for a horizontal tangent, we get $p = (bZ/3a)^{1/2}$. Substituting back into the equation for X in terms of p, we get X proportional to $Z^{3/2}$ which is the shape of the cusp near the singular point, $X = 0$, $Z = 0$.

two-fold branches near the cusp. Such properties are called 'generic' and constitute a rather basic idea in singularity theory. The broad idea of 'generic' vs. 'special', of course, needs a precise mathematical definition which is more elaborate, and not given here.

We can now try and put together the phase space pictures for all values of z. This will clearly be a three-dimensional object. Think of the x–z plane as horizontal, like a table, and plot the value of $p(x,z)$ in the vertical direction. You can imagine this surface by stacking the different coloured curves of *Figure* 6 (only three are shown but one should use the intermediate values of z as well). The resulting surface is shown in *Figure* 7, and goes by the name of 'cusp catastrophe'.

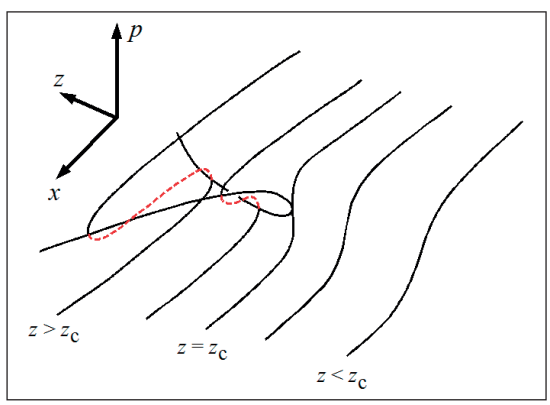


Figure 7. Combining the phase space curves of *Figure* 6 for different z into a single surface in three dimensions. The looping line connects the points like G and H where the slope dp/dx is infinite and is a smooth curve in three dimensions. However, it projects onto the cusp caustic in the x–z plane. The portion of each x–p curve between these two points is shown dotted. (Adapted from the Wikipedia article on Catastrophe theory.)

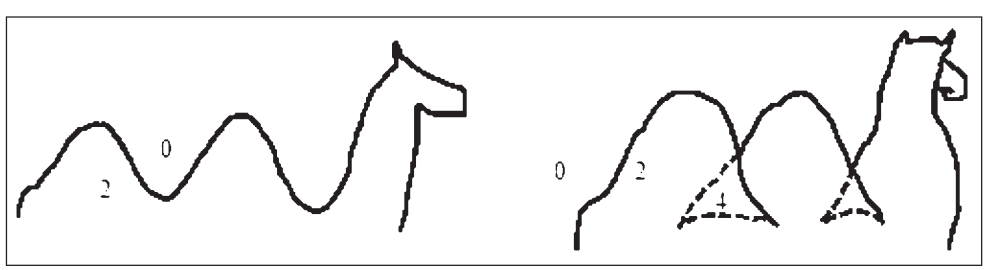
One famous example goes back to Euler – if we have a straight rod-like ruler, and compress it along its length from both sides, it buckles to one side at a critical value of load. Buckling of a structural element could indeed be a catastrophe in the ordinary sense of the word!

The reason for this name becomes clearer if we go into the history of this subject. There are many instances known when a system undergoes an apparently sudden change even though the external influences on it change gradually. One famous example goes back to Euler – if we have a straight rod-like a ruler, and compress it along its length from both sides, it buckles to one side at a critical value of load. Buckling of a structural element could indeed be a catastrophe in the ordinary sense of the word!

In our optical example, as we move from outside the caustic to inside, crossing a fold, two extra rays make their appearance. (Imagine x moving from less than x_2 to greater than x_2 in *Figure* 6). Similarly, in the buckling example, when we exceed the critical load, two new equilibrium configurations, with the ruler buckled on either side, appear over and above the one with the ruler straight. Hence the other name – 'bifurcation theory', which had been used earlier. In many such cases, the different states of the system were given by the minima of a function (e.g., the potential energy). Even in our optical example (*Figure* 3), the point R on the initial wavefront which sends a ray to a given point S is the one which is closest to S, i.e., minimises the path length between S and the wavefront. This minimum principle is even older than Hamilton's work and is named after Fermat (early 17th century), though a less general form of it was stated by Heron of Alexandria (1st century). The French mathematician René Thom, proposed a general theory of such phenomena. Considering the maxima and minima of a potential function, with up to three parameters being varied, he could classify the different kinds of surfaces on which the number of equilibrium states (stable and unstable) changes. He coined the term'Catastrophe theory', which was an instant hit. (We need to vary only one parameter to reach a fold, and two to reach a cusp). Thom also proposed many speculative applications of such a theory to natural phenomena, and others, notably the English mathematician E C Zeeman, went even further.

Arnold spent a year in Paris in 1965 and was excited and deeply influenced by Thom's ideas. However, he was also severely critical both of the lack of rigour in many of the applications, and of the fact that earlier work was not properly cited. This included much work by Russian mathematicians, but also by the American mathematician Whitney, who had established the generic nature of the cusp earlier. Fascinated by the universal nature of these singularities of smooth maps, and the beauty of the mathematics needed to understand them, Arnold devoted a major portion of his efforts after 1965 to this area. In the characteristic Russian style, he conducted a weekly seminar which lasted more than two hours, where problems in this area were discussed. With colleagues and students, the classification went up to 14 dimensions! His 'popular' book entitled *Catastrophe Theory* gives a feel for his vision of the subject. I put 'popular' in quotes because it has passages which would tax even well-trained mathematicians and physicists. But the book also offers beautiful insights, and here is an example which occurs very early. Readers would have heard of Schrödinger's cat (a thought experiment in quantum mechanics proposed by one of its founders), and in fact Arnold also has a cat named after him from a different area, viz., dynamics. I would like to introduce another animal, Arnold's camel, which he used to illustrate the projection of smooth surfaces resulting in cusps (Whitney's theorem). The left side of *Figure* 8 shows the two-humped camel in a side view, and the outline of the humps is a smooth curve. Now the camel decides to turn left, so that one hump obscures the other. If we had X-ray eyes, and could see

Figure 8. As explained in the text, the outline of a camel – assumed to be a smooth surface in three dimensions! – can develop cusps as the viewing angle is changed. The example is taken from Arnold's book *Catastrophe Theory*.



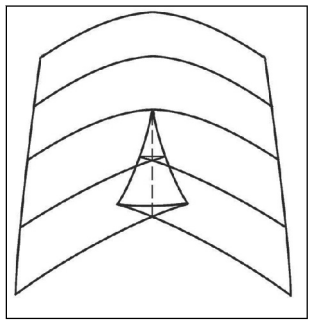


Figure 9. A diagram from the Wikipedia article on catastrophe theory, showing how two cusps can emerge from a smooth part of the outline of a three-dimensional object as it evolves. This is the 'Swallowtail catastrophe'.

Address for Correspondence
Rajaram NItyananda
Azim Premji University,
PES Institute of Technology Campus,
Pixel Park, B Block,
Electronics City, Hosur Road
(Beside NICE Road)
Bangalore – 560100
Email:
rajaram.nityananda@gmail.com

partially through the camel, the outline after the left turn would look like the right side of *Figure* 8. The dotted lines represent the part of the outline which is not visible without X-rays. The outline is now defined as the curve at which the number of intersections of our line of sight with the skin of the camel changes. On the left, we have only zero or two intersections, but in the right-hand figure, you can readily imagine that there are lines which will intersect the skin of the camel four times, as shown in *Figure* 8.

We can now put the camel to some more use. Let us concentrate on the space between the humps, and combine the views which we got from different directions as the camel turns. We see the outline evolving from a smooth curve of folds, to one with two cusps. This clearly has to be viewed in three dimensions, since it is obtained by combining a sequence of two-dimensional pictures. The birth of the two cusps is shown in *Figure* 8, and the entire figure goes by the name of the 'Swallowtail catastrophe' (*Figure* 9). It was one of Thom's original list, and our purpose in exhibiting it is to give a glimpse of the exciting world explored by Arnold and his school, going far beyond folds and cusps.

Suggested Reading (advanced)

[1] **V I Arnold, *Catastrophe Theory*, Springer, 2004.**

[2] **T Poston and I Stewart, *Catastrophe theory and its applications*, Dover, 2012.**

[3] **M V Berry and C Upstill, Catastrophe Optics, *Progress in Optics*, Vol.18, pp.257–346, 1980.**