Conic Sections in the Double-Slit Experiment

Joe Hughes and Frederic Liebrand
Walla Walla University
204 S. College Avenue
College Place, Washington 99324 USA

ABSTRACT

The use of conic sections in obtaining the locations of double-slit maxima is absent in many undergraduate treatments [1-3]. However, their inclusion is not conceptually difficult and increases the understanding not only of interference phenomena, but also of the underlying geometry. This paper illustrates how elliptic and hyperbolic conic sections arise naturally in a conceptual analysis of the double slit.

I.   INTRODUCTION

The standard approach to the double-slit experiment involves a plane wave that passes through the slits (labeled A and B in Figure 1) coherently. The wavelength is smaller than the slit spacing (*d)* and slit width, and smaller by a few orders of magnitude than the distance to the observing screen (L). This allows the two angles, θ_1 and θ_2 to be taken as equal (Figure 1). This assumption makes BD and CD equal, and the path length difference between the two rays is then the section AC, which is $d\sin\theta$. For constructive interference, the path length difference must be an integer number of wavelengths, giving the usual formula for the angle θ_n at which the nth bright fringe occurs as

$$d \cdot \sin(\theta_n) = n\lambda \qquad (1)$$

It is not obvious from the above derivation that all maxima lie on ellipses with the two slits as foci. However, an approach emphasizing path length criteria for constructive interference clearly shows this to be the case.

In cross-section, the wave fronts emanating from each slit become circles with centers at the slits, as shown in Figure 2. Travelling from a slit to the nth concentric circle, or wave front, a ray crosses a distance of n wavelengths. The same will be true for the ray travelling m wavelengths
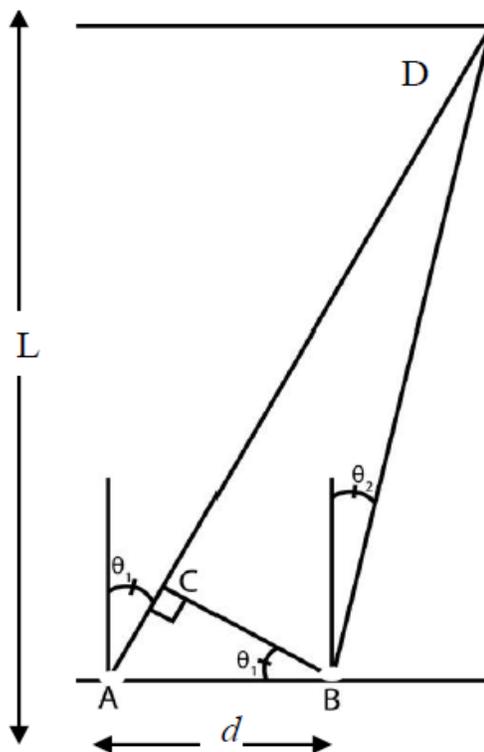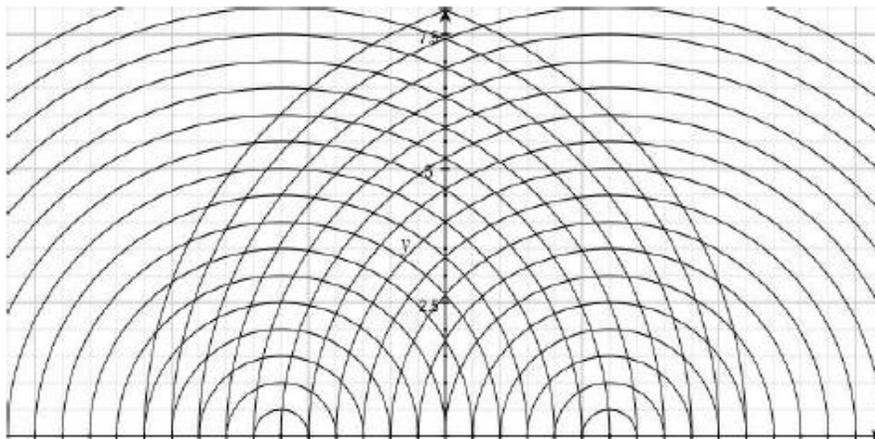

Figure 1. The standard double-slit arrangement.

from the other slit to its mth wave front. Since one ray travels a path of length $m\lambda$ and the other a path of length $n\lambda$, the path length difference between the two is itself an integer number of wavelengths—the condition for constructive interference.

Figure 2. Wavefronts from a double slit.

Graphically, constructive interference occurs where the circles cross.

For a proof that all constructive interferences lie on confocal ellipses, consider a point on the nth circle from the left foci and on the mth circle from the right foci (see Figure 3). If we go to the right one point (tracing from A to C) then we will be n+1 circles from the left foci and the m-1 circles from the right. Since the total distance, in this case, is always the same, those points *must* lie on the same ellipse [4]. The families of ellipses are superimposed on the concentric circles in Figure 4, reflected on both sides of the slits to emphasize the geometry.

To connect these conics to the standard textbook equation, we need to find the angles ($θn$) along which constructive interference occurs. If we traced from D to B instead of A to C, we would find hyperbolas (Figure 5), which is not apparent from the standard derivation. We will use the hyperbolas because unlike ellipses, they have asymptotes, which should be at the same angles given by the standard derivation. Because maxima lie continuously along these hyperbolas, we call them antinodal curves.[1]
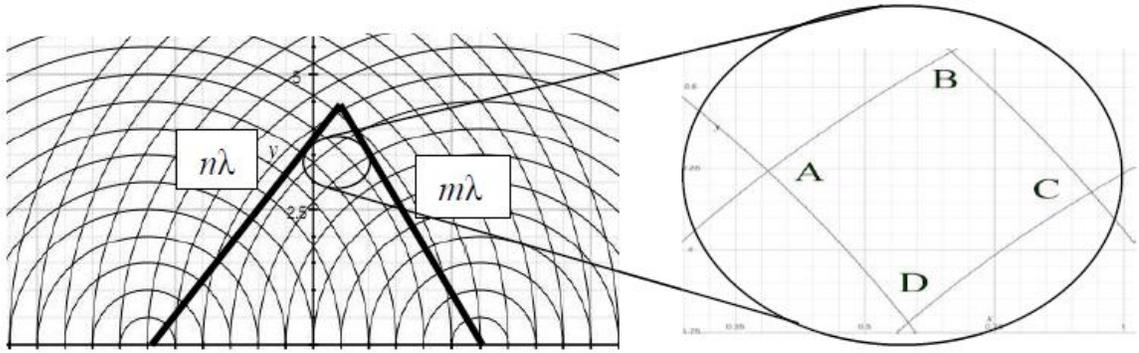
Hyperbolas are curves defined as all points such that their distance to one focus differs from their distance to the other by a fixed amount. For proof that maxima also lie on hyperbolas, consider a point on the jth circle from the left foci and on the kth circle from the right foci. A point above it (tracing from D to B) will be a distance $(j+1)λ$ from the left foci and $(k+1)λ$ from the right foci. Since the distance to one focus differs from the distance to the other by a fixed amount, in this case $(j-k)λ$, the two points *must* lie on the same hyperbola [5].

So the problem of connecting the conics to the standard textbook formula, $d \cdot \sin(\theta_n) = n\lambda$, becomes the problem of finding asymptotes to each consecutive hyperbola. The asymptotes pass through the midpoint between the two slits, and the angle is measured from the normal at that same point. Let's set up a coordinate system as shown in Figure 6. The asymptotes for constructive interference are the dark gray lines passing through solid points. The two foci, or slits, are at (x, y) = (±d/2, 0). The equations of the asymptotes are therefore given by
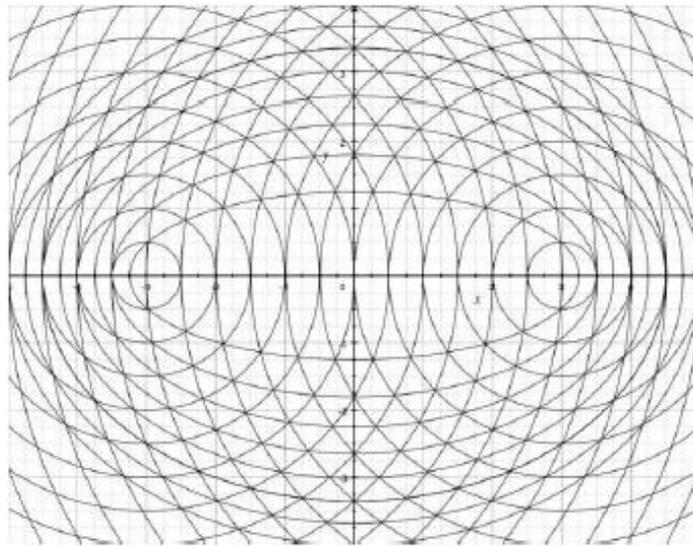
$$x = y \cdot \tan(\theta_n) \qquad (2)$$

If we concern ourselves mainly with small angles, tangent and the sine are approximately equal. Thus the asymptotic line becomes

$$x = y \cdot \sin(\theta_n) \qquad (3)$$

---

[1] In Figure 35.3 (p. 1166) of *Sears and Zemansky's University Physics, 13th Ed.*, the antinodal curves are illustrated, but neither in the figure nor in the text are they identified as being hyperbolas. The text wording does emphasize their determination in terms of path length differences corresponding to Equation (4) of this article, although not in the same detail.

Figures 3. The image is an enlargement of the "diamond" formed by the two rays.

Figure 4. Ellipses pass through maxima.
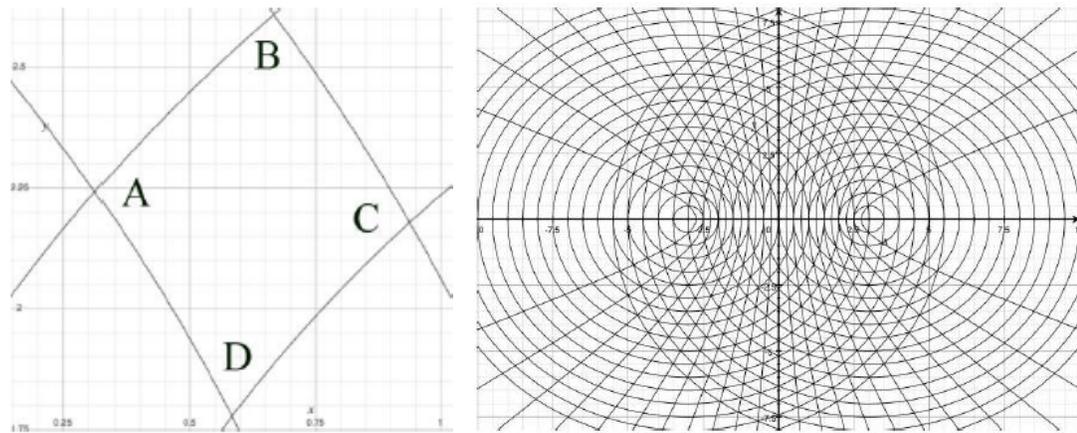
Figure 5. The right portion shows the hyperbolas along which continual maxima occur. At left is an enlargement.
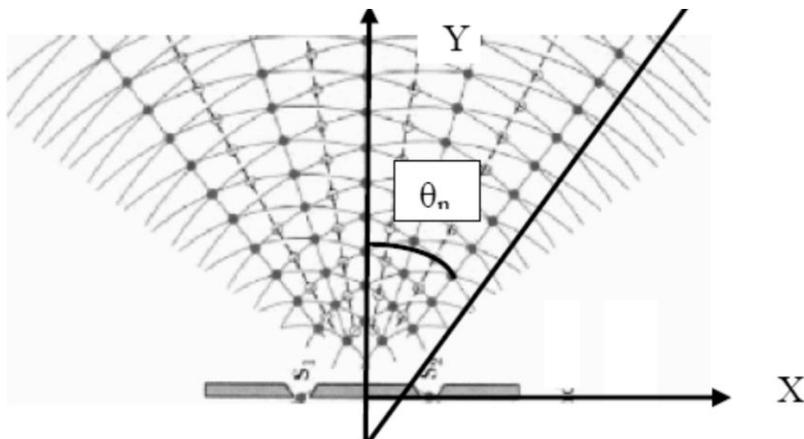
Figure 6. Setting up a coordinate system.

We already know that the hyperbolas for maxima satisfy the requirement that the differences the distances between the maxima at (x, y) and the foci at (±d/2, 0) is an integer number of wavelengths. This gives us

$$\sqrt{y^2 + \left(x + \tfrac{d}{2}\right)^2} - \sqrt{y^2 + \left(x - \tfrac{d}{2}\right)^2} = n\lambda \tag{4}$$

Let's now use the fact that y is large, and that d is small compared to x. This allows us to write, using standard binomial approximations,

$$\sqrt{y^2 \pm \left(x\,\frac{d}{2}\right)^2} = y\sqrt{1 + \frac{x^2}{y^2}\left(1 \pm \frac{d}{2x}\right)^2}$$

$$\approx y\left\{1 + \frac{1}{2}\frac{x^2}{y^2}\left(1 \pm \frac{d}{2x}\right)^2\right\}$$

$$\approx y\left\{1 + \frac{1}{2}\frac{x^2}{y^2}\left(1 \pm 2\frac{d}{2x}\right)\right\}$$

$$= y\left\{1 + \frac{1}{2}\frac{x^2}{y^2}\right\} \pm \frac{dx}{2y}$$

Thus equation (4), which is the exact equation of the hyperbola, becomes the asymptotic expression.

$$y = \frac{d}{n\lambda}\,x \tag{5}$$

Comparing with equation (3) and remembering that sine is roughly equal to tangent for small angles and that θ_n is made with the y-axis, we find that we have reproduced the standard formula (our equation (1)):

$$d \cdot \sin(\theta_n) = n\lambda$$

The beautiful thing about deriving the equation from conics is that not only are they exact, but they also give us a more complete picture of what is actually happening. A derivation in standard algebraic form of the conic sections is given as an appendix.
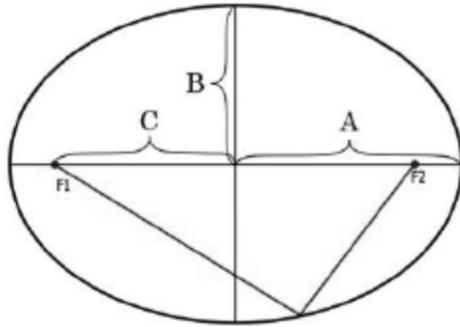
APPENDIX: THE STANDARD FORMS OF THE CONIC SECTIONS

The standard form of an ellipse is

$$\frac{x^2}{A^2} + \frac{y^2}{B^2} = 1,$$

where A is the distance from the center to the vertex along the major axis, B is the distance from the center to the edge along the minor axis, and C is the distance from the center to the foci (Figure 7) [6]. If we call λ the wavelength, d the distance between the two foci, and n the number of circles from the fixed point, then we can calculate A and B and describe our ellipses in terms of d, λ, and n.

Since the center of the ellipse lies halfway between the fixed points, and the distance between the fixed points is d, C

Figure 7. Standard form of an Ellipse.

must be d/2. A is C plus whatever is still needed to get to the vertex. We can express this as $\frac{d}{2} + \frac{n\lambda}{2}$. Once we have A and C, we can find B using a simple relation that is true for all ellipses, $C^2 = A^2 - B^2$. Solving for B and putting it all together, the elliptic equation is

$$\frac{x^2}{\left(\frac{d}{2}+\frac{n\lambda}{2}\right)^2} - \frac{4y^2}{2dn\lambda+n^2\lambda^2} = 1 \qquad (6)$$

The standard form of a hyperbola is
$$\frac{x^2}{A^2} + \frac{y^2}{B^2} = 1,$$
where A is the distance from the center to the curve measured along the x-axis, C is the distance from the center to the foci, and B is the height of the box that the asymptotes go through (Figure 8) [7].

To put this equation in terms of d, λ, and n we recognize C as d/2, and A, instead of $\frac{d}{2} + \frac{n\lambda}{2}$ would be $\frac{d}{2} - \frac{n\lambda}{2}$ because the foci are farther from the center than the vertices. We can find B using the hyperbolic relation $C^2 = A^2 + B^2$. This gives us the complete hyperbolic equation as

$$\frac{x^2}{\frac{1}{4}(d-\lambda n)^2} - \frac{y^2}{\frac{1}{4}(2d\lambda n-(\lambda n)^2)} = 1 \qquad (7)$$

To derive the slope of lines on which constructive interference will occur in the standard derivation, we must refer back to Equation (1). Note that if $d\sin(\theta n)=n\lambda$, then θ_n is $\sin^{-1}(n\lambda/d)$. To go from θ_n to the slope of a Cartesian line we can take the tangent of that angle, so the slope is $\tan(\sin^{-1}(n\lambda/d)$. To simplify this equation we will draw a triangle (Figure 9).

By trigonometry, we know the opposite side must be $n\lambda$ and the hypotenuse must be d. Once we find the third side using the Pythagorean Theorem, we can take the tangent and invert (the angle is made with the y-axis), leaving us with the equation of the standard approximation lines as

$$y = \frac{\sqrt{d^2-(\lambda n)^2}}{\lambda n}\, x \qquad (8)$$
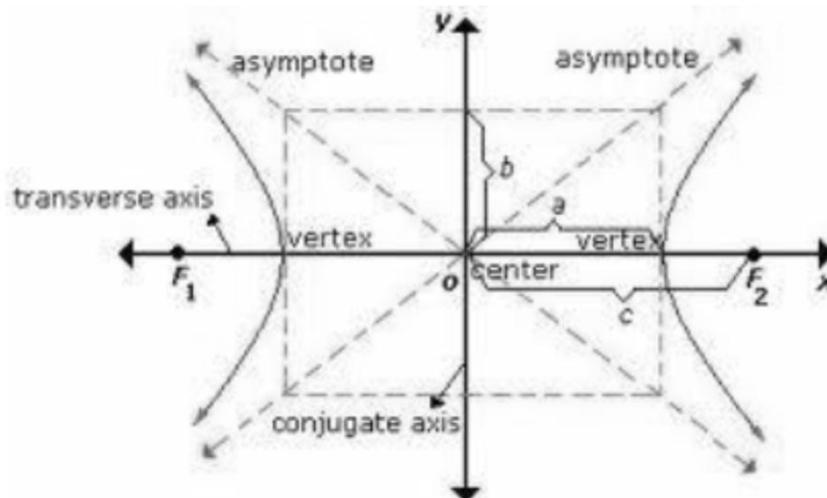
Figure 8. The asymptotes of a hyperbola. (Courtesy of icoachmath, http://www.icoachmath.com/SiteMap/Hyperbola.html.)
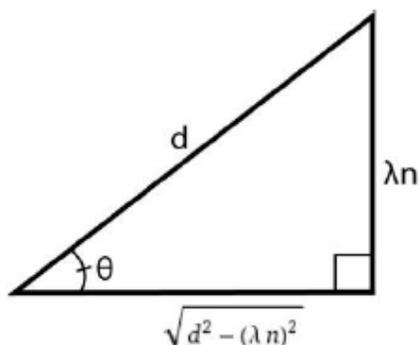
Figure 9. Reference Triangle.

The asymptotes to the hyperbolas can also be used to derive the approximation lines in the standard formula. The formula for an asymptote line for a given hyperbola is $y = \frac{B}{A} x$. Plugging in for A and B gives us

$$y = \frac{\sqrt{2d\lambda n - (\lambda n)^2}}{d - \lambda n} x \qquad (9)$$

This equation differs from Equation (8) because we are looking at a different $n\lambda$. One slope increases as n increases, and the other decreases as n increases. Since n can be any integer, this does not change the validity of the equation, only the form. Since

both equations are undefined when n gets larger than d/λ, we can infer that n only exists between 0 and d/λ. Substituting $\frac{d}{\lambda} - n$ for n in one equation will transform it into the other, showing that the two equations are indeed the same, and that both approaches yield the same results.

REFERENCES

1. Halliday, Resnick and Walker, *Fundamentals of Physics, 6th Edition* (John Wiley and Sons, Inc., New York, 2003) pp. 866-870.
2. Giancoli, *Physics for Scientists & Engineers, 4th Edition* (Pearson-Prentice Hall, New Jersey, 2008) pp. 903-909.
3. Young and Freedman, *Sears and Zemansky's University Physics, 13th Edition* (Peason Education publishing as Addison-Wesley, San Francisco CA, 2012) pp. 1164-1170.
4. Stein, Sherman K, *Calculus and Analytic Geometry, 4th Edition* (McGraw-Hill, New York, 1987) p.S51.
5. Stein, ibid, p. S53.
6. Stein, ibid, p. S52.
7. Stein, ibid, p. S54.