An Elementary Method to Compute the Distance From an Ellipse to a Line

Andrew Paul

Introduction

Finding the shortest distance between an ellipse (an oval or a circle scaled along two perpendicular axes) and a line is a tedious exercise in calculus. The tedious nature of the solution to this problem led me to wonder if there was a way to simplify the problem to something more manageable. As it turns out, ellipses are closely related to circles. Furthermore, the shortest distance between a circle and line can easily be found with an elementary approach on a rectangular coordinate system using analytic geometry. I discovered that I could solve the problem by reducing the ellipse into a circle by adjusting the coordinate system through scaling and then applying techniques which can only be applied on circles. We will be using the following lemma:

Lemma: *The shortest distance between a circle and a line that does not intercept it is equal to the difference between the distance between the center of the circle and the line and the radius of the circle.*

Consider a circle Ω centered at O with radius r, and a line ℓ that does not intercept the circle. Let the altitude from O to ℓ have a foot F and intercept Ω at A. Define P to be on the semicircular subset of Ω that is closest to ℓ. Let the foot of the altitude from P to ℓ be T. Let $\theta = \angle AOP$ and $\alpha = \angle PFT$. Let $AF = x$ and $FP = y$.

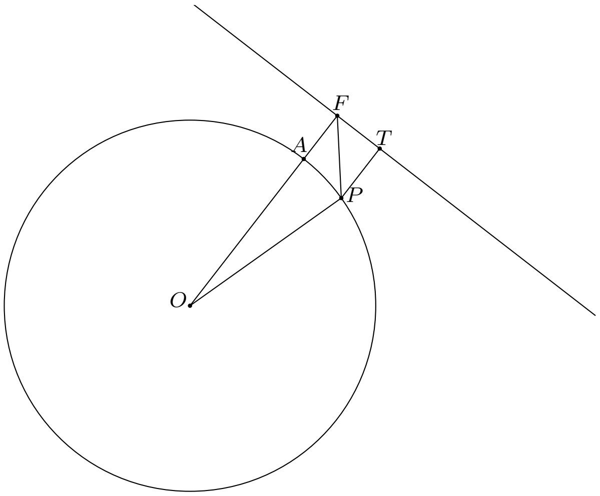

Figure 1: The configuration

Using the Law of Cosines on $\triangle FOP$, we have:

$$y^2 = (r+x)^2 + r^2 - 2r(r+x)\cos\theta \tag{1}$$

Now we repeat the Law of Cosines on $\triangle FOP$, using $\angle PFO = 90^\circ - \alpha$ instead:

$$r^2 = y^2 + (r+x)^2 - 2y(r+x)\cos(90^\circ - \alpha) \tag{2}$$

Observe that sine and cosine are cofunctions, hence (2) becomes:

$$r^2 = y^2 + (r+x)^2 - 2y(r+x)\sin\alpha \tag{3}$$

Rearranging to solve for $\sin\alpha$:

$$\sin\alpha = \frac{y^2 + (r+x)^2 - r^2}{2y(r+x)} \tag{4}$$

Since $\triangle FTP$ is a right triangle, we have $\sin\alpha = \frac{PT}{y} \Rightarrow PT = y\sin\alpha$. Substituting for $\sin\alpha$ from (4), we have:

$$PT = \frac{y^2 + (r+x)^2 - r^2}{2(r+x)} \tag{5}$$

Substituting (1) into (5):

$$PT = r + x - r\cos\theta \tag{6}$$

r and x are constants, so (6) is of the form:

$$PT = c_0 - c_1\cos\theta \tag{7}$$

For positive constants c_0 and c_1. To minimize this function, we take the derivative with respect to θ and set it equal to 0:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}PT = c_1\sin\theta = 0 \Rightarrow \theta = 0$$

We take the second derivative to test the critical point:

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}PT\right|_{\theta=0} = \left.c_1\cos\theta\right|_{\theta=0} = c_1 > 0$$

Hence, $\theta = 0$ corresponds to a relative minimum by the second derivative test. $\theta = 0$ also corresponds with P coinciding with A. The result follows: $PT_{\text{min}} = x$. □

Remarks: The equation $PT = r + x - r\cos\theta$ is relatively simple, suggesting a more efficient derivation. A quicker derivation follows from simply dropping an altitude from P to $\overline{OA}$. Letting the foot of the altitude be G, $OG = r\cos\theta$, and the result follows.

This lemma is in fact a weak form of a stronger theorem that relates the point on a differentiable function closest to a line to the derivatives of the function and the line. It is important to note that the line cannot intercept the function (or else the closest point is simply the intersection point) and the line cannot be an asymptote of the function (or else there is no closest point).

Theorem: *The point on the graph of a differentiable function $f(x)$ that is closest to a line ℓ that does not intercept or asymptotically approach f is the point of tangency for a tangent line that* *is parallel to ℓ.*

To prove this result, we must first find an expression yielding the distance between a point $(x_0, f(x_0))$ and a line in the form of $ax + by = c$. Let P be the point with coordinate $(x_0, f(x_0))$, and let S be a point on the line, which we will call ℓ. Observe that for the distance PS to be minimized, the segment $\overline{PS}$ must form a right angle with ℓ. If S is not the foot of the altitude from P to ℓ, let F be the foot. $\triangle PFS$ is then a right triangle with hypotenuse PS. The hypotenuse is minimized in the degenerate case when $S = F$, in which case the hypotenuse converges to one of the legs.

Since $\overline{PS}$ is perpendicular to ℓ, it is a subset of the line that is perpendicular to ℓ through P. The equation of this line can easily be found. The slope of ℓ is $-\frac{a}{b}$. Therefore, any perpendicular line will have a slope of $\frac{b}{a}$. Since the line also passes through P, it has the equation:

$$y = \frac{b}{a}x - \frac{b}{a}x_0 + y_0$$

To find S, we solve the system of equations between ℓ and $\overleftrightarrow{PS}$:

$$\begin{cases} ax + by = c \\ y = \frac{b}{a}x - \frac{b}{a}x_0 + y_0 \end{cases}$$

Solving this system of equations, we find that:

$$x = \frac{ac - aby_0 + b^2x_0}{a^2 + b^2}$$

$$y = \frac{bc - abx_0 + a^2y_0}{a^2 + b^2}$$

Now we apply the distance formula to obtain the distance PS:

$$\begin{aligned} PS &= \sqrt{\left(\frac{ac - aby_0 + b^2x_0}{a^2 + b^2} - x_0\right)^2 + \left(\frac{bc - abx_0 + a^2y_0}{a^2 + b^2} - y_0\right)^2} \\ &= \sqrt{\frac{(ac - aby_0 - a^2x_0)^2 + (bc - abx_0 - b^2y_0)^2}{(a^2 + b^2)^2}} \\ &= \frac{\sqrt{a^2\left(c - by_0 - ax_0\right)^2 + b^2\left(c - ax_0 - by_0\right)^2}}{a^2 + b^2} \\ &= \frac{|c - by_0 - ax_0|\sqrt{a^2 + b^2}}{a^2 + b^2} \\ &= \frac{|c - by_0 - ax_0|}{\sqrt{a^2 + b^2}} \end{aligned}$$

In fact, this formula is well-known. We obtain its common form by factoring out -1 in the numerator and absorbing it in the absolute value:

$$PS = \frac{|ax_0 + by_0 - c|}{\sqrt{a^2 + b^2}}$$

Now we consider x_0 to be variable, letting it run across the domain of f. Under this interpretation, PS is a function of x_0 and $f(x_0)$:

$$PS(x_0) = \frac{|ax_0 + bf(x_0) - c|}{\sqrt{a^2 + b^2}}$$

We will now differentiate PS with respect to x_0. We deal with the absolute value by splitting it into two cases – one for each possible parity.

Case 1: Suppose $ax_0 + bf(x_0) - c > 0$. Then:

$$PS(x_0) = \frac{ax_0 + bf(x_0) - c}{\sqrt{a^2 + b^2}}$$

Differentiating with respect x_0:

$$PS'(x_0) = \frac{a + bf'(x_0)}{\sqrt{a^2 + b^2}}$$

To minimize PS, we set the derivative equal to 0. Solving for $f'(x_0)$, we have:

$$f'(x_0) = -\frac{a}{b}$$

Case 2: Suppose $ax_0 + bf(x_0) - c < 0$. Then:

$$PS(x_0) = \frac{c - ax_0 - bf(x_0)}{\sqrt{a^2 + b^2}}$$

Differentiating with respect x_0:

$$PS'(x_0) = -\frac{a + bf'(x_0)}{\sqrt{a^2 + b^2}}$$

To minimize PS, we set the derivative equal to 0. Solving for $f'(x_0)$, we have:

$$f'(x_0) = -\frac{a}{b}$$

We arrive at $f'(x_0) = -\frac{a}{b}$ in either case. This is the slope of ℓ, completing the proof. □

The lemma is connected to this theorem by the fact that tangent line to a circle is perpendicular to the radius at the point of tangency.

An ellipse is a conic section that results from scaling a circle along perpendicular axes that intersect at the circle's center. A circle is thus a special case of an ellipse. The equation of an ellipse is:

$$\frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1$$

It is apparent from this equation that this is simply the equation of a unit circle centered at (h, k) scaled horizontally by a factor of a and vertically by a factor of b. The graph of the ellipse $\frac{(x-1)^2}{4} + (y-2)^2 = 1$ is shown:

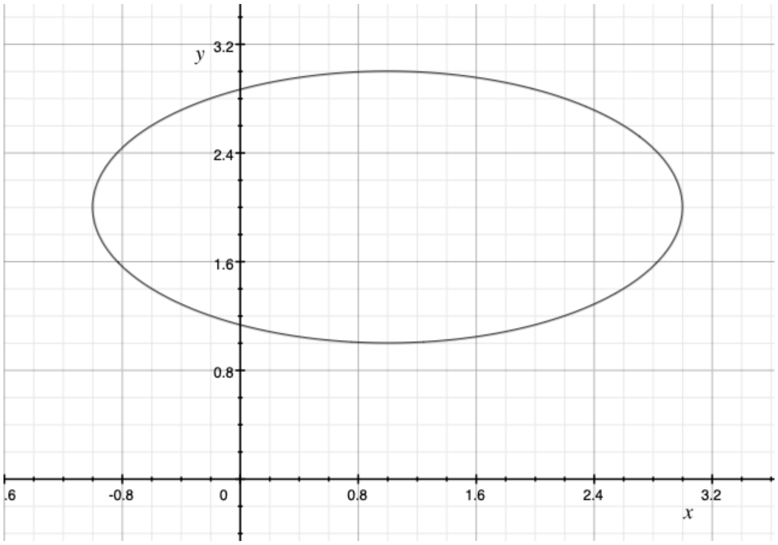

Figure 2: $\frac{(x-1)^2}{4} + (y-2)^2 = 1$

In the ellipse above, the center is $(1, 2)$. The semi-major axis is the largest diameter of the ellipse. In this case, it has length $2\sqrt{4} = 4$. The semi-minor axis, the smallest diameter of the ellipse, has a length of $2\sqrt{1} = 2$.

While the ellipse is intimately related to the circle, it is unclear if the method used to determine the shortest distance between an ellipse and a line resembles the analytic method used to determine the shortest distance between a circle and a line. In this paper, we will exploit the relationship between circles and ellipses to show that every ellipse-to-line distance problem can be reduced to a circle-to-line distance problem on a scaled coordinate system.

Defining C_1: Translating to the Origin

We begin by defining the ellipse $\mathcal{E}$ and line $\mathcal{L}$ as:

$$\mathcal{E} : \frac{(x-h)^2}{a^2} + \frac{(y-k)^2}{b^2} = 1$$

And:

$$\mathcal{L} : px + qy = r$$

Respectively. We will first translate the center of the ellipse to the origin. Since translation is isometric, the distance between the ellipse and the line is preserved. This step is natural considering that when we scale the axes, the scaling is *centered at the origin.* The simplest scaling substitutions can be found with an ellipse that is centered at the origin.

We will define our own notation. Let C_0 be the standard Cartesian coordinate system, and let $C_1(\mathcal{E})$, which we will abbreviate as C_1, be the coordinate system defined by the following transformation on C_0:

$$(x, y) \to (x + h, y + k)$$

Therefore, $\mathcal{E}$ in C_1 is given by:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

After some rearrangement, $\mathcal{L}$ in C_1 is given by:

$$px + qy = r - ph - qk$$

Defining C_2: Scaling the Axes

We define our final coordinate system, $C_2(\mathcal{E})$, which we will abbreviate as C_2. The transformation mapping C_1 to C_2 is given by:

$$(x, y) \to (ax, by)$$

Now, $\mathcal{E}$ in C_2 is given by:

$$x^2 + y^2 = 1$$

Note that all ellipses are mapped to the unit circle in C_2. On the other hand, in C_2, $\mathcal{L}$ is given by:

$$apx + bqy = r - ph - qk$$

We have now simplified our problem to a circle-to-line problem (but only in C_2). The bulk of the work now occurs in C_2.

Working in C_2: Unit Circle to Line

To find the shortest distance between $\mathcal{E}$ and $\mathcal{L}$ in C_2, we use the lemma discussed in the Introduction. The lemma implies that the point on $\mathcal{E}$ closest to $\mathcal{L}$ in C_2 is the intersection point between the segment perpendicular to $\mathcal{L}$ through the origin and the unit circle in C_2. The segment is a subset of the line with equation:

$$y = \frac{bq}{ap}x$$

We treat the following equations as a system to find the aforementioned intersection point:

$$\begin{cases} y = \frac{bq}{ap}x \\ x^2 + y^2 = 1 \end{cases}$$

The solutions to this system are:

$$\left(\pm \frac{1}{\sqrt{1 + \left(\frac{bq}{ap}\right)^2}}, \pm \frac{1}{\sqrt{1 + \left(\frac{ap}{bq}\right)^2}} \right)$$

More tedious computations can be performed to find out which combination of pluses and minuses yields the correct intersection point. We will call the point on $\mathcal{E}$ that is closest to $\mathcal{L}$ the *approach* and its property of being the closest point on $\mathcal{E}$ to $\mathcal{L}$ as its *closeness*.

Our next step is our leap of faith. We figure that reversing the $C_1 \to C_2$ transformation preserves the closeness of the approach. Naturally, if this is true, then we can easily find the approach in C_0 because the $C_0 \to C_1$ transformation is a translation which is isometric. In fact, going all the way back to C_0 is pointless if all we want to do is find the shortest distance between $\mathcal{E}$ and $\mathcal{L}$ for this reason.

Proving That Closeness is Preserved from C_1 to C_2

Reversing the $C_1 \to C_2$ transformation, we find that the coordinates of the C_2 approach is:

$$\left(\pm \frac{1}{a\sqrt{1+\left(\frac{bq}{ap}\right)^2}}, \pm \frac{1}{b\sqrt{1+\left(\frac{ap}{bq}\right)^2}} \right)$$

If we can show that the slope of $\mathcal{E}$ at this point is equal to the slope of $\mathcal{L}$ in C_1, then we are done. We first implicitly differentiate the equation of $\mathcal{E}$ in C_1:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x^2}{a^2}+\frac{y^2}{b^2}\right) = 0 \Rightarrow \frac{2x}{a^2}+\frac{2y}{b^2}y' = 0 \Rightarrow y' = -\frac{xb^2}{ya^2}$$

It looks like we will have to substitute the approach coordinate into the derivative expression. Displeased, we search for a different solution, and we find one. Recall that the approach coordinate in C_2 satisfies the equation:

$$y = \frac{bq}{ap}x$$

This rearranges to:

$$\frac{x}{y} = \frac{ap}{bq}$$

Reversing the $C_1 \to C_2$ transformation, we deduce that the following equation must be true in C_1:

$$\frac{\frac{x}{a}}{\frac{y}{b}} = \frac{ap}{bq} \Rightarrow \frac{x}{y} = \frac{a^2p}{b^2q}$$

Now the path is clear. We substitute $\frac{x}{y}$ into the implicit derivative of $\mathcal{E}$ in C_1:

$$y' = -\frac{xb^2}{ya^2} = -\frac{a^2p}{b^2q}\cdot\frac{b^2}{a^2} = -\frac{p}{q}$$

Surely enough, we observe that $\mathcal{L}$ has a slope of $-\frac{p}{q}$ in C_1 as well.

We are almost done. We complete using our theorem from the introduction. The final hurdle is that the converse of our theorem is not necessarily true. Consider a function f that is differentiable over its domain of $\mathbb{R}$. Suppose the function is piecewise such that it is sinusoidal for some interval $(-\infty, c)$ for some $c \in \mathbb{R}$. Let the function diverge nonasymptotically outside of this interval. Let a line exist such that its slope matches the slope of the sinusoidal portion of f over a single period.

According to our theorem, f has a slope equal to that of the line at the point on f that is the closest to the line. However, given a point on f such that f has a slope equal to that of the line at that point, we cannot ascertain that that particular point is the closest point on f to the line. This is because the line's slope is equal to the slope of f at an infinite number of points due to the infinite number of periods in f and the fact that the line shares the slope of f in its sinusoidal interval.

We can, however, modify our theorem so that the converse becomes true. To do this, we add a restriction on functions f that we discuss. The crux of the problem with the converse is that the derivative of f is not invertible. If we restrict our theorem to functions f that had invertible derivatives, then the converse becomes true. This is because there would then only be one possible point where the slope of the line can match up with the slope of the function, and so that point *must* be the closest point on the function to the line.

If we consider the half of $\mathcal{E}$ that is the closest to $\mathcal{L}$ to be a function, we find that its derivative is in fact invertible because the top and bottom halves of an ellipse are concave down and concave up respectively, so the same slope is never repeated.

Therefore, by the converse of our modified theorem, since the slope of $\mathcal{E}$ at the C_2 approach in C_1 is equal to the slope of $\mathcal{L}$ in C_1, the closeness of the approach is preserved. In other words, if we reverse the $C_1 \to C_2$ transformation, the C_2 approach does in fact become the C_1 approach (and vice versa, depending on which direction you look at it from). The shortest distance between $\mathcal{E}$ and $\mathcal{L}$ is then simply the distance between the approach and the line in C_1.

Conclusion

The distance between an ellipse and a line can be determined by analytic means. It is possible to make clever substitutions (transformation) that map any ellipse to the unit circle. This reduces the problem to a circle-to-line problem which is substantially easier to solve.

By stumbling upon this result, I am able to see the connections between various topics, including geometry, analytic geometry, and calculus. Every time I had an obstacle, I was able to resort to a seemingly unrelated technique that cleared up the issue. The broad array of mathematical techniques I used makes me appreciate the interconnected nature of mathematical disciplines. Every time I had an obstacle, I was able to resort to a seemingly unrelated technique that cleared up the issue.