The Maximal Deflection on an Ellipse

Dan Kalman

Dan Kalman (kalman@american.edu) received his Ph.D. from the University of Wisconsin in 1980. Before joining the mathematics faculty at American University in 1993, he worked for eight years in the aerospace industry in Southern California. He is a former associate executive director of the MAA, the author of a book, *Elementary Mathematical Models*, published by the MAA, and a frequent contributor to MAA journals. He delights in puns and word play of all kinds, and is an avid fan of Douglas Adams, J. R. R. Tolkien, and Gilbert and Sullivan.

"Give me more problems like that!" insisted Mickey. He was an unusual student, a senior history major who signed up for Calculus 3 as an elective just because he was interested in mathematics. He did not care for routine drill problems. He wanted problems that posed a challenge, problems that required him to bring together many different parts of the course. After class on the day we covered Lagrange multipliers, Mickey pointed out a problem from an earlier assignment, and asked that I give him more of the same. It would be a shame to disappoint such a student.

Under this stimulus, what I thought would be an interesting application of Lagrange multipliers came to mind: find the maximal deflection between the radial direction and the normal direction at a point of an ellipse. The normal direction would be something Mickey could find using properties of gradients, which we had recently studied. The angle between vectors could be formulated using dot products. And Lagrange multipliers was an obvious method for maximizing a function over the points of an ellipse.

As it turns out, this problem is *not* particularly well suited for Lagrange multipliers. But it *is* a good problem, and has an interesting answer. It can be attacked from a variety of viewpoints, each of which adds a little insight. It even has some applied significance. Best of all, there is a nice generalization to higher dimensions, with a little bit of a twist. Sharing the details behind all these assertions is the purpose of this paper.

The problem

For reference, here is a careful formulation of the problem. We consider an ellipse, centered at the origin, with semi-major axis a, semi-minor axis b, and with these axes along the x- and y-axes respectively. The equation of the ellipse is

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

The vector from the origin to any point of the ellipse defines the radial direction for that point. A vector perpendicular to the tangent line defines the normal direction. We know that at the x- and y-intercepts, the radial and normal directions coincide, so at these points the deflection δ, defined as the angle between normal and radial vectors, is 0. At any other point of the ellipse, the deflection will be greater than 0. Our goal is to

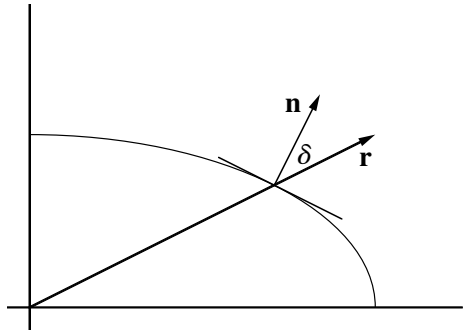

Figure 1. Radial and normal directions for an ellipse.

find the maximal deflection and where it occurs. With no loss of generality, we confine our attention to the part of the ellipse in the first quadrant. In Figure 1 a representative ellipse is shown, with the deflection δ, the normal direction $\mathbf{n}$, and the radial direction $\mathbf{r}$ at a point in the first quadrant.

The answer

When I posed the problem, I had no idea whether there would be any simple geometric description for the solution. I figured it was even odds that the answer would come out to be a root of some algebraic equation, with no further significance. So I was delighted to find that both the magnitude of the maximal deflection, and the point where it occurs, have simple geometric constructions. Indeed, there are several compact formulations for the maximal deflection:

$$\begin{aligned} \delta_{\max} &= \frac{\pi}{2} - 2\arctan\frac{b}{a} \\ &= \arctan\frac{a}{b} - \arctan\frac{b}{a} \\ &= \arctan\frac{a^2 - b^2}{2ab}. \end{aligned} \tag{1}$$

The maximal deflection occurs where the ellipse meets the line from the origin to (a, b). In other words, if you inscribe the ellipse in a rectangle with sides parallel to the axes, and if you draw a line from the center of the ellipse to one of the corners of the rectangle, the line's intersection with the ellipse locates the maximal deflection between the radial and normal directions.

These results, which are illustrated in Figure 2, are formalized for future reference in the following theorem.

Theorem 1. *Let E be an ellipse with center C, semi-major axis a and semi-minor axis b. At any point $P \in E$, let δ be the angle between the normal vector at P and the vector CP. Then the maximum value of δ over E is given by* (1)*. This value is assumed at the points where E intersects the diagonals of the circumscribed rectangle whose sides are parallel to the major and minor axes.*

Note the first expression for $\delta_{\max}$ in (1) shows that it increases monotonically with a/b. This is as expected: the more eccentric the ellipse, the greater the maximal de-

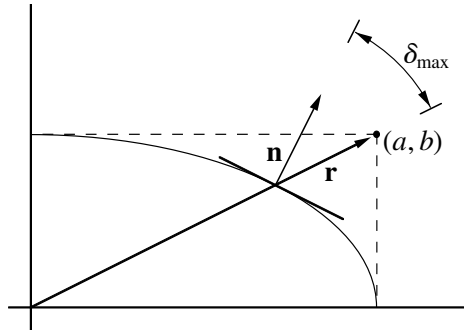

Figure 2. Maximal deflection point.

flection should be. Conversely, the closer the ellipse is to a circle, the closer the radial and normal directions will be, and so the smaller the maximal deflection will be. The monotonicity of the maximal deflection will be important in the n-dimensional case.

The location of $\delta_{\max}$ given in Theorem 1 suggests an obvious conjecture for n dimensions. As I will explain at the end of the paper, this obvious conjecture is wrong. Happily, there is a beautiful extension to the general case, and in fact, at this point in the discussion you already have enough information to deduce what it is. If you like puzzles and have good geometric intuition, you may want to work out the n-dimensional case before you read the last section.

Constructing the maximal deflection

Once you know where the maximal deflection occurs, it is easy to construct. The intersection between the ellipse and the line through the origin and the point (a, b) is given by $(a, b)/\sqrt{2}$. At this point, it is readily shown that the angle between the tangent and radial lines is bisected by a horizontal line. This is illustrated in Figure 3, where the normal, radial, horizontal, and tangent lines are marked N, R, H, and T. That H bisects the angle between R and T is indicated by the equality of the angles marked α in the figure. We see at once that $\alpha = \arctan(b/a)$, and also that δ, the angle between R and N, is $\pi/2 - 2\alpha$.

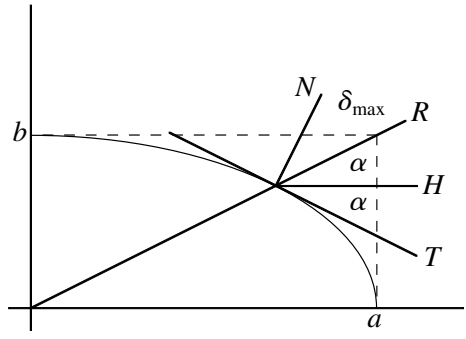

Figure 3. Geometric construction.

In fact, this gives a purely geometric construction of tangent and normal lines at this point. Draw the radial line R and the horizontal line H. Duplicate the angle between R and H to construct T. Construct the normal to T to define N.

There is an alternative construction. At the same point $(a, b)/\sqrt{2}$, a line with slope 1 bisects the angle between R and N. Indeed, the slope of R is b/a while the slope of N is a/b. Accordingly, the angles between these lines and the x-axis are, respectively, $\arctan(a/b)$ and $\arctan(b/a)$. This gives for δ the alternate expression $\arctan(a/b) - \arctan(b/a)$. It also shows how to construct N and T: draw line R; where it intersects the ellipse, construct a line with slope 1; reflect R across this line to define N; and construct the perpendicular to N to define T.

Both of these constructions depend only on considering the special point $(a, b)/\sqrt{2}$ on the ellipse. We have not yet shown that this is the point where the deflection is maximized. To do so is the object of the original optimization problem. As mentioned, my original intention was that this would be an application of Lagrange multipliers. We will see next how this method leads to the point $(a, b)/\sqrt{2}$. Subsequent sections will show a couple of alternative derivations.

Lagrange multipliers. In order to apply Lagrange multipliers, we need to express the problem in terms of an objective function to be maximized and a constraint. Because we consider only points of the ellipse, its equation defines the constraint. Accordingly, we define the function

$$g(x, y) = \frac{x^2}{a^2} + \frac{y^2}{b^2},$$

and understand the constraint to require $g(x, y) = 1$.

For the objective function, we want the angle δ between the normal and the radial vectors at a point (x, y) of the ellipse. We may take $\mathbf{r} = (x, y)$ as the radial vector. For the normal vector, take $\mathbf{n} = (x/a^2, y/b^2)$, which is one-half of the gradient of g. Then, δ is determined by the equation

$$\cos\delta = \frac{\mathbf{r} \cdot \mathbf{n}}{|\mathbf{r}||\mathbf{n}|}.$$

Now observe that $\mathbf{r} \cdot \mathbf{n} = g(x, y) = 1$ for any point on the ellipse. Accordingly, we simplify matters by inverting and squaring to obtain

$$\sec^2\delta = |\mathbf{r}|^2|\mathbf{n}|^2 = (x^2 + y^2)\left(\frac{x^2}{a^4} + \frac{y^2}{b^4}\right).$$

We define this to be our objective function. That is,

$$f(x, y) = (x^2 + y^2)\left(\frac{x^2}{a^4} + \frac{y^2}{b^4}\right).$$

For (x, y) in the first quadrant and on the ellipse, we know that δ is between 0 and $\pi/2$. On this interval, $\sec^2\delta$ is an increasing function. Therefore, δ is maximized where f is.

Our problem now is to maximize f subject to the constraint $g = 1$. The solution must occur at a point where ∇f and ∇g are parallel. Using the fact that vectors (u, v) and (p, q) are parallel just when $uq = pv$, this leads to the single equation

$$\frac{\partial f}{\partial x}\frac{\partial g}{\partial y} = \frac{\partial f}{\partial y}\frac{\partial g}{\partial x}.$$

From this equation, it is a straightforward (if slightly complicated) matter to derive

$$\frac{y}{x} = \pm\frac{b}{a}. \tag{2}$$

As a first step, compute the partial derivatives

$$\frac{\partial f}{\partial x} = \frac{2x[2b^4x^2 + (a^4 + b^4)y^2]}{a^4b^4}$$
$$\frac{\partial f}{\partial y} = \frac{2y[2a^4y^2 + (a^4 + b^4)x^2]}{a^4b^4}$$
$$\frac{\partial g}{\partial x} = \frac{2x}{a^2}$$

and

$$\frac{\partial g}{\partial y} = \frac{2y}{b^2}.$$

Combining these leads to

$$\frac{\partial f}{\partial x}\frac{\partial g}{\partial y} = \frac{4xy}{a^4b^4} \cdot \frac{2b^4x^2 + (a^4 + b^4)y^2}{b^2}$$
$$\frac{\partial f}{\partial y}\frac{\partial g}{\partial x} = \frac{4xy}{a^4b^4} \cdot \frac{2a^4y^2 + (a^4 + b^4)x^2}{a^2},$$

These expressions are equal if and only if

$$2a^2b^4x^2 + (a^4 + b^4)a^2y^2 = 2a^4b^2y^2 + (a^4 + b^4)b^2x^2.$$

One more rearrangement now produces

$$a^2y^2(a^4 - 2a^2b^2 + b^4) = b^2x^2(a^4 - 2a^2b^2 + b^4),$$

from which (2) is apparent.

This shows that in the first quadrant, the solution to our optimization problem must lie on the line joining the origin to (a, b). Although somewhat protracted, the algebra in the preceding derivation it is not too complicated to complete by hand. Nevertheless, it is sufficiently involved to discourage the typical calculus student.

Direct parameterization. The standard parameterization of the ellipse, namely,

$$(x, y) = (a\cos t, b\sin t),$$

provides an alternative to Lagrange multipliers. Substitution in the objective function leads to

$$F(t) = f(a\cos t, b\sin t) = \cos^4 t + \cos^2 t \sin^2 t \left(\frac{a^2}{b^2} + \frac{b^2}{a^2}\right) + \sin^4 t$$

as a function of a single variable. We wish to find the maximum value of this function for $0 \le t \le \pi/2$.

Before carrying out the optimization, let us simplify the expression for F. After squaring the Pythagorean identity $\cos^2 t + \sin^2 t = 1$, we find that F reduces to

$$F(t) = 1 + \left(\frac{a^2}{b^2} + \frac{b^2}{a^2} - 2\right)\cos^2 t \sin^2 t.$$

Proceeding further, the double angle identity for sine now leads to

$$F(t) = 1 + \frac{1}{4}\left(\frac{a^2}{b^2} + \frac{b^2}{a^2} - 2\right)\sin^2(2t).$$

At this point, we can almost find the extreme point by inspection. It remains only to determine the sign of the coefficient of $\sin^2(2t)$. After one final algebraic simplification, we arrive at

$$F(t) = 1 + \frac{1}{4}\left(\frac{a^2 - b^2}{ab}\right)^2 \sin^2(2t).$$

This shows that the maximum value of F must occur when $|\sin(2t)| = 1$, and for $0 \le t \le \pi/2$, this implies that the maximum occurs when $t = \pi/4$. Returning to the parameterization of the ellipse, that yields $(x, y) = (a, b)/\sqrt{2}$.

Note that this approach not only tells us where the maximum value of F occurs, but also what the maximum value is:

$$F_{\max} = 1 + \frac{1}{4}\left(\frac{a^2 - b^2}{ab}\right)^2.$$

Retracing our steps back to $f(x, y) = \sec^2 \delta$, we ultimately arrive at

$$\delta_{\max} = \arctan \frac{a^2 - b^2}{2ab}.$$

This should be recognized as one of the expressions in (1).

Among our students, it is a common misconception that the variable t in the standard parameterization of an ellipse is equal to the polar angle θ. A moment's reflection shows that this is incorrect. For example, when $t = \pi/4$ the vector $(a\cos t, b\sin t)$ has slope b/a, not 1. Generally speaking, the difference α between t and the polar angle for $(a\cos t, b\sin t)$ is not 0, although this difference does vanish on the axes. The problem of maximizing α is similar to maximizing δ, and has another interesting solution. This is left as an exercise for the interested reader.

Focusing once more on δ, the fact that the maximal deflection occurs at the midpoint of the parameter domain is striking, and suggests searching for a geometric interpretation. In fact, there is a kind of symmetry that makes the solution point at $\pi/4$ very natural. This idea will be discussed after briefly considering another solution of the optimization problem.

Using slopes. While the direct parameterization worked out pretty nicely, there is an alternate approach worthy of consideration. It expresses everything in terms of slopes, and uses only ideas from the first calculus course. To begin, we consider a point (x, y) on the ellipse, and observe that the slope of the radial line there is $m = y/x$. Next, compute m_N, the slope of the normal line, as follows: by implicit differentiation, the slope of the tangent line to the ellipse is $-b^2x/a^2y$, so $m_N = (a^2/b^2)m$.

Given these slopes, we can compute the tangent of the angle between the lines from the formula

$$\tan(\delta) = \frac{m_N - m}{1 + m_N m}.$$

After some algebra, this leads to

$$\tan(\delta) = \frac{m(a^2 - b^2)}{b^2 + a^2 m^2}.$$

The maximum value of δ occurs at the same point as the maximum of $f(m) = \tan(\delta)/(a^2 - b^2) = m/(b^2 + a^2 m^2)$. Differentiation gives us

$$f'(m) = \frac{b^2 - a^2 m^2}{(b^2 + a^2 m^2)^2}.$$

This shows immediately that f assumes a unique maximum for positive m when $m = b/a$. This is, of course, consistent with what we found before.

Symmetry. As previously mentioned, there is a symmetry that makes the location of the point of maximal deflection very natural. Recall the standard parameterization of the ellipse, $\mathbf{r} = (a\cos t, b\sin t)$. At any point on the ellipse, we may take $\mathbf{r}$ itself as a radial vector. Differentiating with respect to t produces the tangent vector $(-a\sin t, b\cos t)$, from which we obtain the outward normal vector $\mathbf{n} = (b\cos t, a\sin t)$.

These give rise to a revealing geometric interpretation. The key ideas are illustrated in Figure 4, which depicts circles of radius b and a centered at the origin. Rectangle $PQRS$ has vertices given parametrically by

$$P = (b\cos t, b\sin t)$$
$$R = (a\cos t, a\sin t)$$
$$Q = (a\cos t, b\sin t)$$
$$S = (b\cos t, a\sin t).$$

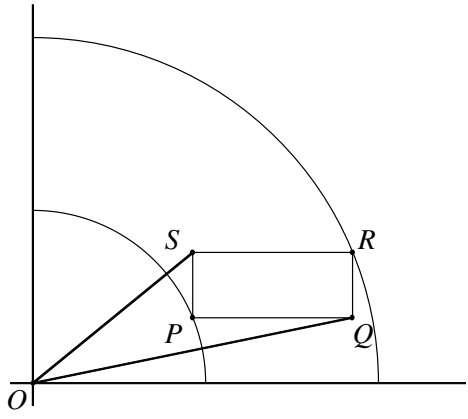

Figure 4. Parametric rectangle.

Observe that as t varies, P and R trace out the circles, and the vectors OP and OR are parallel with polar angle t. At the same time, OQ is the vector $\mathbf{r}$, and so Q traces our ellipse. Now we see that S has two interpretations. On the one hand, S traces a second ellipse, with horizontal semi-minor axis b and vertical semi-major axis a. This is the reflection of the first ellipse in the line $y = x$. On the other hand, OS is the vector $\mathbf{n}$, normal to the original ellipse at Q. This view reveals an elegant geometric relationship between the parameter t and the radial and normal vectors to point Q on the ellipse. In particular, the difference between the polar angle for Q and the parameter t is represented by $\angle POQ$. Moreover, the deflection angle δ between the radial and normal vectors appears as $\angle SOQ$. As the parameter t varies from 0 to $\pi/2$, OR sweeps around the outer circle at a constant rate, while $\square PQRS$ evolves continuously from a horizontal segment, through a progression of rectangles, to a vertical segment. In the process, $\angle SOQ$ portrays the variation of δ.

This is where symmetry enters the picture. Consider Figure 5 which shows $\square PQRS$ and $\square P'Q'R'S'$ corresponding to parameter values t and $t' = \pi/2 - t$. Since $t + t' = \pi/2$, the two rectangles are mirror images in the line $y = x$. Indeed, reflection in this line preserves the identities of the P and R points, while interchanging Q and S. But note particularly that the $\angle SOQ$ and $\angle S'OQ'$ are the same. This shows that δ assumes equal values for t and t'. That is, the function $\delta(t)$ is symmetric with respect to $t = \pi/4$.

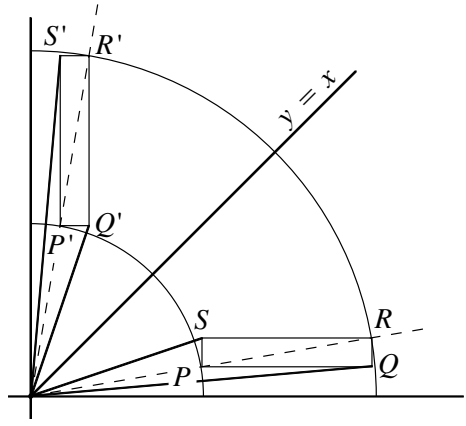

Figure 5. Symmetric rectangles.

In itself, this is not sufficient to tell us that the maximum value of δ occurs at $t = \pi/4$. But if it occurs elsewhere, there must be two solution points which are symmetric about $\pi/4$. We readily observe that δ is 0 for $t = 0$ or $\pi/2$. And it is plausible that δ increases from 0 to a unique maximum, and then decreases back to 0 as t goes from 0 to $\pi/2$. Given that assumption, that there is a unique maximum, symmetry shows that it must occur at $t = \pi/4$. (For a related discussion of symmetry in optimization problems, see [**3**].)

An application

Although conceived purely as a calculus exercise, the question of maximal deflection has an application. It concerns the ellipsoidal model of the Earth, and two ways to define latitude. The following discussion of this application is based on [**1**, p. 94].

On a spherical globe, the latitude at a point is the angle between the equatorial plane and the position vector from the center of the sphere. This latter vector defines the local vertical direction, which is also the normal to the local tangent plane and the (opposite of the) direction of the gravitational force at the point.

For careful astronomical measurements, however, the spherical model is insufficiently accurate. Instead, it is customary to use an ellipsoidal model, also referred to as an oblate sphereoid, with circular cross sections parallel to the equatorial plane and elliptical cross sections at a fixed eccentricity perpendicular to the equatorial plane. For such a model, the local vertical direction, indicated by a hanging plumb bob, does not point in the direction of the center of the Earth. Rather, it is normal to the surface of the ellipsoid (which is ideally a level surface with respect to the combination of gravity and the centrifugal acceleration induced by the Earth's rotation). Because the local vertical direction is much easier to measure than the direction of the center of the Earth, it is a convenient reference for defining latitude. Indeed, the angle between the local vertical direction and the equatorial plane gives what is called the *geodetic* latitude ϕ_d. It is geometrically distinct from the more familiar *geocentric* latitude ϕ_c, defined as in the spherical model to be the angle between the radial direction and the equatorial plane. An exaggerated version of the geometry is shown in Figure 6.

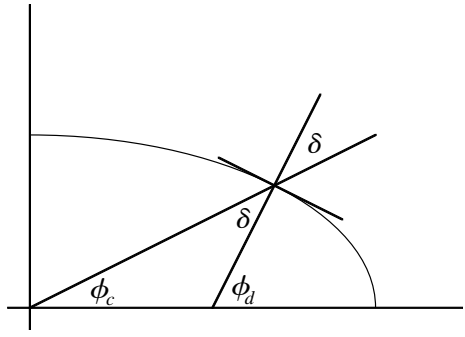

Figure 6. Geodetic and geocentric latitude.

In reference to the figure, it is apparent that what I have called the deflection, δ, is the difference $\phi_d - \phi_c$. In this setting, Theorem 1 reveals how far apart the geodetic and geocentric latitudes are at the worst case, and where on the globe that occurs.

The distinction between geodetic and geocentric latitudes is important, for example, in determining the locations of celestial objects. Most local observations are made relative to the local vertical, or plumb bob direction. In order to reconcile observations from different points on the globe, or to register them in a global geospatial model, we have to take into account the deviation between the radial and plumb bob directions.

Generalization to R^n. As mentioned at the start of the paper, there is a nice generalization of the problem of maximal deflection to the n-dimensional case. We consider the ellipsoid with semi-axes $a_1, a_2, \cdots a_n$, whose equation is

$$\frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} + \cdots + \frac{x_n^2}{a_n^2} = 1 \tag{3}$$

and we wish to locate the point where the angle between the normal and radial vectors is greatest. By analogy with the 2-dimensional case, it is tempting to conjecture that the solution lies on the segment joining the origin to $(a_1, a_2, \ldots, a_n)$. Call this point A.

Unfortunately, none of the solution methods presented above lends itself particularly well to a solution in n dimensions. The simplest method in the plane formulated everything in terms of slope, which does not extend in an obvious way to higher dimensions. Direct parameterization is a possibility, using for example n-dimensional spherical coordinates, but the algebra quickly gets out of hand. Lagrange multipliers is the method that extends most easily to n dimensions, at least as far as formulating the necessary condition for an extreme point. However, it is not obvious how to solve the Lagrange equations. On the other hand, it is easy enough to check that the coordinates of A do *not* satisfy the Lagrange conditions. So at least we can determine that the obvious conjecture is false.

It turns out that in spite of these apparent difficulties, the n-dimensional case is amazingly simple, once you have the right geometric insight. For me, that insight occurred when a colleague produced an illustration of the geometry in three-dimensions. Using Mathwright, the software he created, James White implemented a routine to color-code the surface of any three-dimensional ellipsoid according to the size of the deflection δ, and to view the result from any angle. (For more about White and Mathwright, see [**2**].) One sample of this coloring appears in Figure 7. The points that are brightest correspond to the largest values of δ. At the ends of the axes of the ellipsoid, where δ is 0, the shading is darkest.

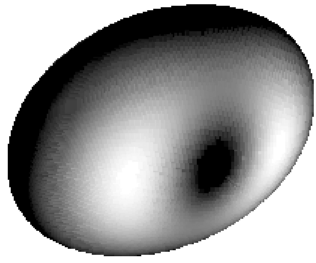

Figure 7. Shaded ellipsoid.

As suggested by the figure, δ achieves its maximum values in a plane cross section of the ellipsoid, corresponding to the greatest and least semi-axis. This is a consequence of the monotonicity of $\delta_{\max}$ with eccentricity, and leads to the following theorem.

Theorem 2. *On the ellipsoid* (3)*, the maximum value of* δ *can be found as follows: Consider the plane cross section of the ellipsoid determined by the axes corresponding to the greatest and least values of* a_i*. This cross section is an ellipse. The maximum value of* δ *on this ellipse is also the maximum of* δ *over the entire ellipsoid.*

Proof. Let a^* be the maximum of the a_i, and let b^* be the minimum of the a_i. Without loss of generality, we may assume that $a^* = a_1$ and $b^* = a_n$. In the plane determined by the x_1 and x_n axes, the ellipsoid's cross section is an ellipse E^* with semi-major axis a^* and semi-minor axis b^*. By Theorem 1, the maximum value of δ on E^* is given by $\delta^* = \pi/2 - 2\arctan(b^*/a^*)$. We wish to show that this is the maximal deflection over the entire ellipsoid.

Now consider an arbitrary point P on the ellipsoid. At P the radial and normal vectors determine a plane that intersects the ellipsoid in an ellipse through P. Call this ellipse E; let a be its semi-major axis and b its semi-minor axis. Invoking Theorem 1 again, the maximal deflection over E is $\pi/2 - 2\arctan(b/a)$, and this must be greater than or equal to δ_P, the deflection at P.

To complete the proof, compare E^* with E. Note that no point of the ellipsoid is further from the center than a^* nor closer to the center than b^*. Therefore, $a \le a^*$ and $b \ge b^*$. This shows that $a/b \le a^*/b^*$, so the maximum value of δ on E is less than or equal to the maximum on E^*. Together, these considerations imply that

$$\delta_P \le \max_E \delta \le \max_{E^*} \delta = \delta^*.$$

Therefore, δ^* is the maximum value of δ over the entire ellipsoid. ■

Conclusion

So ends my account of the surprisingly fertile question: Where on an ellipse is the angle between the radial and normal vectors the greatest? We have seen several methods of solution, an interesting appearance of symmetry, extensions to higher dimensions, and even an application. It is a shame that so entertaining an investigation must now come to an end. Unless.... I wonder what happens when the origin is at a focus of the ellipse, instead of at its center?

References

1. Roger R. Bate, Donald D. Mueller, and Jerry E. White, *Fundamentals of Astrodynamics,* Dover, 1971.
2. Dan Kalman, Virtual Empirical Investigation: Concept Formation and Theory Justification, *Amer. Math. Monthly* 112 (2005), 786–798.
3. William C. Waterhouse, Do Symmetric Problems have Symmetric Solutions?, *Amer. Math. Monthly* 90 (1983), 378–387.

Names in Boxes Puzzle

From Peter Winkler (peter.winkler@dartmouth.edu), Dartmouth College, Hanover, NH 03755

A. The Problem The names of one hundred prisoners are placed in one hundred wooden boxes, one name to a box, and the boxes are lined up on a table in a room. One by one, the prisoners are led into the room; they may look into up to fifty of the boxes to try to find their own name, but must leave the room exactly as it was. They are permitted no further communication after leaving the room. The prisoners do have a chance to plot a strategy in advance, and they are going to need it, because unless *they all find their own names* they will all be executed. There is a strategy that has a probability of success exceeding thirty per cent—find it.

Comment. If each prisoner examines fifty boxes at random, their probability of survival is an unenviable $1/2^{100}$, or about

$$0.00000000000000000000000000000008.$$

Thirty per cent seems ridiculously out of reach—but yes, you heard the problem correctly!

For the strategy, turn to page 285.