Computing Science Group

An Analogue Solution to the Problem of Factorization

Ed Blakey
edward.blakey@queens.ox.ac.uk

CS-RR-07-04

Oxford University Computing Laboratory
Wolfson Building, Parks Road, Oxford, OX1 3QD

Abstract. The task of factorizing a given integer is notoriously difficult, to the extent of rendering computationally infeasible the extraction of factors of numbers beyond a certain size. This infeasibility is what makes the RSA cryptographic system, for example, secure.
We describe an analogue method[1] of factorizing. Just as with traditional algorithms, there is a practical limit to the size of numbers that the method can factorize; in contrast with traditional algorithms, however, the method suffers no increase in calculation time as the input number approaches this limit.
The process described exploits a direct physical implementation of a geometric formulation of the problem of factorizing; this allows factors of numbers within the allowed range to be ascertained (or else primality guaranteed) virtually instantaneously.

1 Geometric Formulation

In this section, we reformulate as a geometric problem the numeric problem of factorization.

Proposition 1. *The task of finding factors of a given, positive, natural number N is equivalent to that of finding points that lie both in the integer grid $\mathbb{Z} \times \mathbb{Z}$ and on the curve $y = \frac{N}{x}$.*

Proof. A point (a, b) is on the curve $y = \frac{N}{x}$ if and only if $N = ab$; it is in the grid $\mathbb{Z} \times \mathbb{Z}$ if and only if $a, b \in \mathbb{Z}$. Hence, (a, b) is both on the curve and in the grid if and only if a and b offer a factorization, into two integers, of N. □

Note 1. The factorization of N corresponding to a point in the grid and on the curve is not necessarily—in fact, is rarely—a full decomposition of N into primes (it may even be no more informative than to demonstrate that $N = 1.N$). However, each prime factor p of N has a corresponding point $\left(p, \frac{N}{p}\right)$ in the grid and on the curve; thus, all prime factors are represented by at fewest one such point each.

Note 2. Since, by hypothesis, N is positive, the curve $y = \frac{N}{x}$ exists only in quadrants $x, y \geq 0$ and $x, y \leq 0$; further, since only positive factors of N (specifically, primes) are sought, only the former quadrant need be considered.

Similarly, by the symmetry of the curve and of the integer grid—specifically, because each is symmetric about the line $y = x$—only one octant within this quadrant need be considered (since (a, b) is both on the curve and in the grid if and only if (b, a) is, and both points correspond, due to commutativity of multiplication, to the same partial factorization: $N = ab$). Accordingly, we consider only the octant $0 \leq x \leq y$.

[1] The method is the subject of a pending US patent, applied for by IBM and with sole inventor Ed Blakey.

Proposition 2. *The curve $y = \frac{N}{x}$, $z = 0$ (which lies in three-space) can be expressed as the intersection of the (x, y)-plane and the cone[2] that consists of those lines that both pass through the point $\left(0, 0, \sqrt{2N}\right)$ and make an angle of $\frac{\pi}{4}$ of a radian with the line $y = x$, $z = \sqrt{2N}$.*

Proof. Let C be the cone made up of those lines that both pass through the point $P := \left(0, 0, \sqrt{2N}\right)$ (the tip of the cone) and make an angle of $\frac{\pi}{4}$ of a radian with the line $y = x$, $z = \sqrt{2N}$ (call this line L).

Let $A := (a, b, 0)$ be an arbitrary point both on the cone C and in the (x, y)-plane; let B be the point $\left(x_A, x_A, \sqrt{2N}\right)$, where $x_A = \frac{\left(a^2+b^2+2N\right)^{\frac{1}{2}}}{2}$. Note that

$$\begin{aligned} |BP| &= \left(x_A^2 + x_A^2 + 0\right)^{\frac{1}{2}} \\ &= x_A\sqrt{2} \\ &= \frac{\left(a^2 + b^2 + 2N\right)^{\frac{1}{2}}}{\sqrt{2}} \\ &= \frac{|AP|}{\sqrt{2}} \quad ; \end{aligned}$$

further, since A is on C and B on L, the angle APB is, by definition of C, $\frac{\pi}{4}$. Hence, ABP is a right angle, and $|BP| = |AB|$. So

$$\begin{aligned} \frac{\left(a^2 + b^2 + 2N\right)^{\frac{1}{2}}}{\sqrt{2}} &= |BP| \\ &= |AB| \\ &= \left(\left(a - x_A\right)^2 + \left(b - x_A\right)^2 + 2N\right)^{\frac{1}{2}} \\ &= \left(a^2 + b^2 - 2\left(a + b\right)x_A + 2x_A^2 + 2N\right)^{\frac{1}{2}} \quad . \end{aligned}$$

Multiplying each side by $\sqrt{2}$,

$$\left(a^2 + b^2 + 2N\right)^{\frac{1}{2}} = \left(2\left(a^2 + b^2 - 2\left(a + b\right)x_A + 2x_A^2 + 2N\right)\right)^{\frac{1}{2}} \quad .$$

Squaring,

$$a^2 + b^2 + 2N = 2\left(a^2 + b^2 - 2\left(a + b\right)x_A + 2x_A^2 + 2N\right) \quad .$$

Subtracting $a^2 + b^2 + 2N$ and recalling that $x_A = \frac{\left(a^2+b^2+2N\right)^{\frac{1}{2}}}{2}$,

$$\begin{aligned} 0 &= a^2 + b^2 - 4\left(a + b\right)x_A + 4x_A^2 + 2N \\ &= a^2 + b^2 - 2\left(a + b\right)\left(a^2 + b^2 + 2N\right)^{\frac{1}{2}} + \left(a^2 + b^2 + 2N\right) + 2N \\ &= 2a^2 + 2b^2 + 4N - 2\left(a + b\right)\left(a^2 + b^2 + 2N\right)^{\frac{1}{2}} \quad . \end{aligned}$$

[2] That $y = \frac{N}{x}$, $z = 0$ is the intersection of the (x, y)-plane and *some* cone is no surprise: $y = \frac{N}{x}$ is a hyperbola, and, hence, a conic section, which fact motivates the geometric formulation presented here.

Dividing by 2 and rearranging,

$$a^2 + b^2 + 2N = (a + b)\left(a^2 + b^2 + 2N\right)^{\frac{1}{2}} \quad .$$

Dividing by $\left(a^2 + b^2 + 2N\right)^{\frac{1}{2}}$ (valid since a, b and N are all positive),

$$\left(a^2 + b^2 + 2N\right)^{\frac{1}{2}} = a + b \ .$$

Squaring,

$$\begin{aligned} a^2 + b^2 + 2N &= (a + b)^2 \\ &= a^2 + b^2 + 2ab \ . \end{aligned}$$

Hence, $ab = N$, and so A, an arbitrary point both on the cone C and in the (x, y)-plane, is on the curve $y = \frac{N}{x}$, $z = 0$.

Conversely, let $A := \left(a, \frac{N}{a}, 0\right)$ be an arbitrary point on the curve $y = \frac{N}{x}$, $z = 0$. Let B be the point $\left(x_A, x_A, \sqrt{2N}\right)$, where $x_A = \frac{a^2+N}{2a}$. Then

$$\begin{aligned} |AB| &= \left(\left(a - \frac{a^2 + N}{2a}\right)^2 + \left(\frac{N}{a} - \frac{a^2 + N}{2a}\right)^2 + 2N\right)^{\frac{1}{2}} \\ &= \left(\left(\frac{a^2 - N}{2a}\right)^2 + \left(\frac{N - a^2}{2a}\right)^2 + 2N\right)^{\frac{1}{2}} \\ &= \left(2\left(\left(\frac{a^2 - N}{2a}\right)^2 + N\right)\right)^{\frac{1}{2}} \\ &= \left(2\left(\frac{a^2 + N}{2a}\right)^2\right)^{\frac{1}{2}} \\ &= \frac{a^2 + N}{a\sqrt{2}} \ ; \end{aligned}$$

$$\begin{aligned} |BP| &= \left(x_A^2 + x_A^2 + 0\right)^{\frac{1}{2}} \\ &= x_A\sqrt{2} \\ &= \frac{a^2 + N}{a\sqrt{2}} \ ; \end{aligned}$$

and

$$|AP| = \left(a^2 + \left(\frac{N}{a}\right)^2 + 2N\right)^{\frac{1}{2}}$$

$$= \left(\frac{a^4 + N^2 + 2Na^2}{a^2} \right)^{\frac{1}{2}}$$
$$= \frac{a^2 + N}{a} \ .$$

Hence, $|AB| = |BP| = \frac{1}{\sqrt{2}} |AP|$, so the angle between AP and BP (and, hence, the angle between AP and L, since B and P are on L) is $\frac{\pi}{4}$. Thus A, an arbitrary point on the curve $y = \frac{N}{x}$, $z = 0$, is both on the cone C and in the (x, y)-plane, as required. □

The physical implementation, discussed in the following section, of the factorization method that we describe exploits the facts, demonstrated above, (a) that factorization can be reformulated as the search for integer points on the curve $y = \frac{N}{x}$, and (b) that this curve can be expressed as the intersection of a cone and a plane.

2 Physical Implementation

2.1 Implementation of the Integer Grid

Definition 1 (in which we implement the integer grid).

1. *As before, let N be the natural number to be factorized. Assume that N is odd (see Remark 1).*
2. *Let ϵ be a small, positive, fixed real ($0 < \epsilon \ll 1$).*
3. *Let M_1 be a parabolic mirror, reflective on its concave side, occupying the curve $\left\{ \left(x, -\frac{1}{2(1+\epsilon)} x^2 + x + (1+\epsilon), 0 \right) : 0 \leq x \leq 1 \right\}$.*
4. *Let M_2 be a plane mirror, reflective on its $x < y$ side, occupying the line segment $\{(x, x, 0) : 0 \leq x \leq 1\}$.*
5. *Let M_3 be a plane mirror, reflective on its $x > 0$ side, occupying the line segment $\{(0, y, 0) : 0 \leq y \leq 1\}$.*
6. *Let S be a source at $(1 + \epsilon, 1 + \epsilon, 0)$ (this is the focus of the parabola of which M_1 is part) of electromagnetic radiation with wavelength $\lambda := \frac{2}{N}$; suppose that S is shielded such that its radiation stays close to the plane $z = 0$.*
7. *Let $B := \{(x, y, 0) : 1 \leq x \leq y < 1 + \epsilon\}$ be a blackbody that absorbs radiation arriving from S.*

Remark 1. We assume that the number N to be factorized is odd.[3] This is because, for ease of implementation, the reduced grid $\left\{(x, y) : x, y, \frac{x+y}{2} \in \mathbb{Z}\right\}$ (that is, pairs (x, y) of integers, where the parity of x is that of y) is implemented instead of the full grid $\mathbb{Z} \times \mathbb{Z} = \{(x, y) : x, y \in \mathbb{Z}\}$ mentioned in Sect. 1; then any factorization of N (which is odd) into integers x and y will be such that x and y are both odd, so this reduced grid suffices.

[3] Should a factorization be required of an even number, it is computationally trivial to iteratively divide by two until an odd number—which can be factorized as described here—is obtained.

Further, consideration need be made only of that part of the reduced grid with $0 \leq x \leq y \leq N$ (since no factor of N is greater than N, and by Note 2); only this part of the grid is implemented (see Remark 4).

Proposition 3. *Radiation incident on M_1 from S is reflected by M_1 as a beam of waves parallel to the y-axis, in the band $0 \leq x \leq 1$ (which is entirely spanned by such waves), and travelling in the direction of decreasing y.*

Proof. (Readers for whom it suffices to note that S sits at the focus of the parabola containing M_1, that this parabola is symmetric about a line parallel to the y-axis, and that the projection of M_1 onto the x-axis is the interval $[0,1]$ may skip this proof.)

Since the region $\left\{(x,y,0) : y < -\frac{1}{2(1+\epsilon)}x^2 + x + (1+\epsilon)\right\}$ under the parabola containing M_1 is convex, since S lies in this region, and since no point of B lies on a line between S and any point on M_1, there is radiation incident on each point of M_1.

Consider the radiation incident on an arbitrary point A of M_1; say $A = \left(a, -\frac{1}{2(1+\epsilon)}a^2 + a + (1+\epsilon), 0\right)$ with $0 \leq a \leq 1$.

The gradient of the curve $y = -\frac{1}{2(1+\epsilon)}x^2 + x + (1+\epsilon)$ is given by $y' = -\frac{1}{1+\epsilon}x + 1$, which at A is $1 - \frac{a}{1+\epsilon}$, so the tangent t_A at A to the curve has equation $y = \left(1 - \frac{a}{1+\epsilon}\right)x + \frac{a^2}{2(1+\epsilon)} + 1 + \epsilon$, $z = 0$ and the normal n_A at A to the curve has equation $y = \left(\frac{1+\epsilon}{a-1-\epsilon}\right)x + a + 1 + \epsilon + \frac{a(1+\epsilon)}{1+\epsilon-a} - \frac{a^2}{2(1+\epsilon)}$, $z = 0$. The radiation from S incident on A is reflected along the line passing through the reflections in n_A of S and A.

Let u_A be the line parallel to t_A and passing through S; this has equation $y = \left(1 - \frac{a}{1+\epsilon}\right)x + a$, $z = 0$. Let T be the point on n_A and u_A. By construction (and, specifically, since n_A and u_A are perpendicular and since u_A passes through S), T is the midpoint between S and the reflection of S (call this S') in n_A. In particular, letting P_1 denote the x-coordinate of a point P, T_1 is the mean of S_1 and S_1'; that is, $S_1' = 2T_1 - S_1$. S_1, recall, is $1+\epsilon$. T_1 is the value of x for which $y = \left(1 - \frac{a}{1+\epsilon}\right)x + a$, $z = 0$ (i.e. u_A) meets $y = \left(\frac{1+\epsilon}{a-1-\epsilon}\right)x + a + 1 + \epsilon + \frac{a(1+\epsilon)}{1+\epsilon-a} - \frac{a^2}{2(1+\epsilon)}$, $z = 0$ (i.e. n_A); that is, the value of x such that $\left(1 - \frac{a}{1+\epsilon}\right)x + a = \left(\frac{1+\epsilon}{a-1-\epsilon}\right)x + a + 1 + \epsilon + \frac{a(1+\epsilon)}{1+\epsilon-a} - \frac{a^2}{2(1+\epsilon)}$. So

$$
\begin{aligned}
T_1 &= \left(a + 1 + \epsilon + \frac{a\,(1+\epsilon)}{1+\epsilon-a} - \frac{a^2}{2\,(1+\epsilon)} - a\right) \div \left(1 - \frac{a}{1+\epsilon} - \frac{1+\epsilon}{a-1-\epsilon}\right) \\
&= \left(1 + \epsilon + \frac{a\,(1+\epsilon)}{1+\epsilon-a} - \frac{a^2}{2\,(1+\epsilon)}\right) \div \left(1 - \frac{a}{1+\epsilon} + \frac{1+\epsilon}{1+\epsilon-a}\right) \\
&= \frac{2\,(1+\epsilon-a)\,(1+\epsilon)^2 + 2a\,(1+\epsilon)^2 - a^2\,(1+\epsilon-a)}{2\,(1+\epsilon-a)\,(1+\epsilon)} \\
&\quad \cdot \frac{(1+\epsilon-a)\,(1+\epsilon)}{(1+\epsilon-a)\,(1+\epsilon) - a\,(1+\epsilon-a) + (1+\epsilon)^2}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{2\left(1+\epsilon-a\right)\left(1+\epsilon\right)^2 + 2a\left(1+\epsilon\right)^2 - a^2\left(1+\epsilon-a\right)}{2\left(1+\epsilon-a\right)\left(1+\epsilon\right) - 2a\left(1+\epsilon-a\right) + 2\left(1+\epsilon\right)^2} \\
&= \big(2 + 4\epsilon + 2\epsilon^2 + 2\epsilon + 4\epsilon^2 + 2\epsilon^3 - 2a - 4\epsilon a \\
&\quad - 2\epsilon^2 a + 2a + 4\epsilon a + 2\epsilon^2 a - a^2 - \epsilon a^2 + a^3\big) \\
&\quad \div \big(2 + 2\epsilon + 2\epsilon + 2\epsilon^2 - 2a - 2\epsilon a - 2a - 2\epsilon a + 2a^2 + 2 + 4\epsilon + 2\epsilon^2\big) \\
&= \frac{2 + 6\epsilon + 6\epsilon^2 + 2\epsilon^3 - a^2 - \epsilon a^2 + a^3}{2\left(2 + 4\epsilon + 2\epsilon^2 - 2a - 2\epsilon a + a^2\right)} \ .
\end{aligned}
$$

Hence,

$$
\begin{aligned}
S_1' &= 2T_1 - S_1 \\
&= \frac{2 + 6\epsilon + 6\epsilon^2 + 2\epsilon^3 - a^2 - \epsilon a^2 + a^3}{2 + 4\epsilon + 2\epsilon^2 - 2a - 2\epsilon a + a^2} - 1 - \epsilon \\
&= \big(2 + 6\epsilon + 6\epsilon^2 + 2\epsilon^3 - a^2 - \epsilon a^2 + a^3 - 2 - 4\epsilon - 2\epsilon^2 + 2a \\
&\quad + 2\epsilon a - a^2 - 2\epsilon - 4\epsilon^2 - 2\epsilon^3 + 2\epsilon a + 2\epsilon^2 a - \epsilon a^2\big) \\
&\quad \div \big(2 + 4\epsilon + 2\epsilon^2 - 2a - 2\epsilon a + a^2\big) \\
&= \frac{-2a^2 - 2\epsilon a^2 + a^3 + 2a + 4\epsilon a + 2\epsilon^2 a}{2 + 4\epsilon + 2\epsilon^2 - 2a - 2\epsilon a + a^2} \\
&= \frac{a\left(2 + 4\epsilon + 2\epsilon^2 - 2a - 2\epsilon a + a^2\right)}{2 + 4\epsilon + 2\epsilon^2 - 2a - 2\epsilon a + a^2} \\
&= a \ .
\end{aligned}
$$

So the line of the reflected radiation passes through points A and S' (which are distinct, being separated by the same distance as are A and S), and $S_1' = A_1 = a$. Hence, the reflected radiation passes along the line $x = a$, as claimed.

Further, since each point on M_1 has incident radiation, the band $0 \leq x \leq 1$ is entirely spanned by reflected waves. □

Remark 2. Radiation from S not incident on M_1 is not of interest here; it is either absorbed by B or completely leaves the apparatus.

Proposition 4. *The beam described in Proposition 3 is reflected by M_2 to form a beam parallel to the x-axis, in the band $0 \leq y \leq 1$ (which is entirely spanned by the reflected beam), and travelling in the direction of decreasing x.*

Proof. The part of the incoming beam incident on an arbitrary point $A := (a, a, 0)$ $(0 \leq a \leq 1)$ of M_2 travels to A along the line $x = a$, $z = 0$, with y decreasing. This is reflected by M_2, which sits at an angle of $\frac{\pi}{4}$ to the x- and y-axes, along the line $y = a$, $z = 0$, with x decreasing.

Further, the reflected beam spans the band $0 \leq y \leq 1$ since, by Proposition 3, the incoming beam spans $0 \leq x \leq 1$. □

Proposition 5. *Radiation incident on M_3 from S (via M_1 and M_2) is reflected by M_3 back along itself, producing a standing wave.*

Proof. By Proposition 4, radiation reaches a point $A := (0, a, 0)$ $(0 \leq a \leq 1)$ of M_3 by travelling along the line $y = a$, $z = 0$, with x decreasing. Since this line is parallel to the x-axis and M_3 to the y-axis, the incident ray is normal to the mirror and is reflected along itself. The nature of the standing wave thus produced is described in Proposition 6. □

Remark 3 (in which we summarize the preceding propositions). A ray from S that is of interest (that is, that falls on mirror M_1 rather than leaving the apparatus or hitting B) meets M_1 at the point $\left(a, -\frac{1}{2(1+\epsilon)}a^2 + a + (1+\epsilon), 0\right)$ for some $0 \leq a \leq 1$ (conversely, each such a has a corresponding ray). It is then reflected by M_1 vertically down to $(a, a, 0)$, where M_2 reflects it horizontally across to $(0, a, 0)$. M_3 then reflects the ray back along itself via M_2 and M_1 to S, setting up a standing wave, which is described below.

Proposition 6. *In the triangular region $R := \{(x, y, 0) : 0 \leq x \leq y \leq 1\}$, the interference pattern produced by the standing waves mentioned above is such that a point $(a, b, 0)$ is at maximum amplitude (specifically, four times the amplitude of the original radiation from S) if and only if Na and Nb are integers of the same parity.*

Proof. (For brevity, define $f : [0, 1] \to \mathbb{R}$ by $f : x \mapsto -\frac{x^2}{2(1+\epsilon)} + x + 1 + \epsilon$.) The interference pattern at a point $(a, b, 0)$ in R is influenced by only four rays from S:

1. the ray from S via $(a, f(a), 0)$ on M_1 to $(a, b, 0)$;
2. the ray from S via $(a, f(a), 0)$ on M_1, $(a, a, 0)$ on M_2, $(0, a, 0)$ on M_3 and $(a, a, 0)$ on M_2 to $(a, b, 0)$;
3. the ray from S via $(b, f(b), 0)$ on M_1 and $(b, b, 0)$ on M_2 to $(a, b, 0)$; and
4. the ray from S via $(b, f(b), 0)$ on M_1, $(b, b, 0)$ on M_2 and $(0, b, 0)$ on M_3 to $(a, b, 0)$.

The amplitude at the point $(a, b, 0)$ of each of these rays can be modelled by $\alpha \sin\left(\frac{2\pi}{\lambda}d + t\right)$, where α is the amplitude of the original radiation, λ its wavelength, d the total distance travelled by the ray from S to $(a, b, 0)$ and t a representation of time. Writing $g(x)$ for $\left((1+\epsilon-x)^2 + (1+\epsilon-f(x))^2\right)^{\frac{1}{2}} + f(x)$ $(0 \leq x \leq 1)$, the respective values of d for the four rays are $d_1 := g(a) - b$, $d_2 := g(a) + b$, $d_3 := g(b) - a$ and $d_4 := g(b) + a$.

Note that these expressions can be simplified since $g(x) = 2(1+\epsilon)$ for each $x \in [0, 1]$; this equality holds because

$$\begin{aligned} g(x) &= \left[2(1+\epsilon)^2 - 2(1+\epsilon)(x + f(x)) + x^2 + f(x)^2\right]^{\frac{1}{2}} + f(x) \\ &= \left[2(1+\epsilon)(1+\epsilon - f(x) - x) + x^2 + f(x)^2\right]^{\frac{1}{2}} + f(x) \\ &= \left[2x^2 - 4x(1+\epsilon) + f(x)^2\right]^{\frac{1}{2}} + f(x) \end{aligned}$$

$$
\begin{aligned}
&= \left[\frac{1}{4}x^4\left(1+\epsilon\right)^{-2} - x^3\left(1+\epsilon\right)^{-1} + 2x^2 - 2x\left(1+\epsilon\right) + \left(1+\epsilon\right)^2\right]^{\frac{1}{2}} + f(x) \\
&= \left[\left(1+\epsilon-x+\frac{x^2}{2\left(1+\epsilon\right)}\right)^2\right]^{\frac{1}{2}} + f(x) \\
&= 1+\epsilon-x+\frac{x^2}{2\left(1+\epsilon\right)} + f(x) \\
&= 2\left(1+\epsilon\right) \ .
\end{aligned}
$$

Recalling that $\sin\theta + \sin\phi = 2\sin\frac{\theta+\phi}{2}\cos\frac{\theta-\phi}{2}$ (and that $\cos\theta = \cos\left(-\theta\right)$), and that $\lambda = \frac{2}{N}$, the resultant amplitude $\sum_{i=1}^{4}\alpha\sin\left(\frac{2\pi}{\lambda}d_i + t\right)$ at $(a,b,0)$ can be written as

$$
\begin{aligned}
\sum_{i=1}^{4}\alpha\sin\left(\frac{2\pi}{\lambda}d_i+t\right) &= \sum_{i=1}^{2}\alpha\sin\left(\frac{2\pi}{\lambda}d_i+t\right) + \sum_{j=3}^{4}\alpha\sin\left(\frac{2\pi}{\lambda}d_j+t\right) \\
&= 2\alpha\sin\left(\frac{2\pi}{\lambda}\cdot\frac{d_1+d_2}{2}+t\right)\cos\left(\frac{2\pi}{\lambda}\cdot\frac{d_1-d_2}{2}\right) \\
&\quad + 2\alpha\sin\left(\frac{2\pi}{\lambda}\cdot\frac{d_3+d_4}{2}+t\right)\cos\left(\frac{2\pi}{\lambda}\cdot\frac{d_3-d_4}{2}\right) \\
&= 2\alpha\left(\sin\left(\frac{2\pi}{\lambda}g(a)+t\right)\cos\left(\frac{2\pi}{\lambda}b\right)\right. \\
&\quad \left. + \sin\left(\frac{2\pi}{\lambda}g(b)+t\right)\cos\left(\frac{2\pi}{\lambda}a\right)\right) \\
&= 2\alpha\left(\sin\left(2N\pi\left(1+\epsilon\right)+t\right)\cos\left(N\pi b\right)\right. \\
&\quad \left. + \sin\left(2N\pi\left(1+\epsilon\right)+t\right)\cos\left(N\pi a\right)\right) \ .
\end{aligned}
$$

Note that, since the sine and cosine functions both give values in the interval $[-1,1]$, this amplitude is in the interval $[-4\alpha, 4\alpha]$. The result to be proven is that a point $(a,b,0)$ in R is such that Na and Nb are integers of the same parity if and only if $2\alpha\left(\sin\left(2N\pi\left(1+\epsilon\right)+t\right)\cos\left(N\pi b\right) + \sin\left(2N\pi\left(1+\epsilon\right)+t\right)\cos\left(N\pi a\right)\right) = 4\alpha$ for some $t\in\mathbb{R}$.

Let $(a,b,0)$ be a point such that Na and Nb are integers of the same parity; suppose first that Na and Nb are even; let $t_0 = \pi\left(\frac{1}{2} - 2N\left(1+\epsilon\right)\right)$. Then $N\pi a$ and $N\pi b$ are even multiples of π, and so $\cos\left(N\pi a\right) = \cos\left(N\pi b\right) = 1$. Further,

$$
\begin{aligned}
\sin\left(2N\pi\left(1+\epsilon\right)+t_0\right) &= \sin\left(2N\pi\left(1+\epsilon\right) + \pi\left(\frac{1}{2} - 2N\left(1+\epsilon\right)\right)\right) \\
&= \sin\left(\frac{\pi}{2}\right) \\
&= 1 \ ,
\end{aligned}
$$

so

$$
2\alpha\left(\sin\left(2N\pi\left(1+\epsilon\right)+t_0\right)\cos\left(N\pi b\right)\right.
$$

$$
\begin{aligned}
+ \sin\left(2N\pi\left(1+\epsilon\right)+t_0\right)\cos\left(N\pi a\right)\big) &= 2\alpha\left(1.1+1.1\right)\\
&= 4\alpha \ ;
\end{aligned}
$$

that is, if Na and Nb are both even, then the amplitude at $(a,b,0)$ is 4α. Suppose instead that Na and Nb are both odd, and let $t_0 = \pi\left(-\frac{1}{2} - 2N\left(1+\epsilon\right)\right)$. Then $N\pi a$ and $N\pi b$ are odd multiples of π, and so $\cos\left(N\pi a\right) = \cos\left(N\pi b\right) = -1$. Further,

$$
\begin{aligned}
\sin\left(2N\pi\left(1+\epsilon\right)+t_0\right) &= \sin\left(2N\pi\left(1+\epsilon\right)+\pi\left(-\frac{1}{2}-2N\left(1+\epsilon\right)\right)\right)\\
&= \sin\left(-\frac{\pi}{2}\right)\\
&= -1 \ ,
\end{aligned}
$$

so

$$
\begin{aligned}
2\alpha\big(&\sin\left(2N\pi\left(1+\epsilon\right)+t_0\right)\cos\left(N\pi b\right)\\
&+ \sin\left(2N\pi\left(1+\epsilon\right)+t_0\right)\cos\left(N\pi a\right)\big) = 2\alpha\left((-1)^2+(-1)^2\right)\\
&= 4\alpha \ ;
\end{aligned}
$$

that is, if Na and Nb are both odd, then the amplitude at $(a,b,0)$ is 4α.

Conversely, if a point $(a,b,0)$ has amplitude 4α (say that t_0 is such that $2\alpha\left(\sin\left(2N\pi\left(1+\epsilon\right)+t_0\right)\cos\left(N\pi b\right)+\sin\left(2N\pi\left(1+\epsilon\right)+t_0\right)\cos\left(N\pi a\right)\right)$ is equal to 4α), then we have that $\sin\left(2N\pi\left(1+\epsilon\right)+t_0\right) = \cos\left(N\pi b\right) \in \{\pm 1\}$ and $\sin\left(2N\pi\left(1+\epsilon\right)+t_0\right) = \cos\left(N\pi a\right) \in \{\pm 1\}$. So, since $\cos\left(\pi Nb\right)$ and $\cos\left(\pi Na\right)$ are in $\{\pm 1\}$, Na and Nb are integers. Required is that Na and Nb have the same parity. Now

$$
\begin{aligned}
\cos\left(N\pi b\right) &= \sin\left(2N\pi\left(1+\epsilon\right)+t_0\right)\\
&= \cos\left(N\pi a\right)\\
&= \begin{cases} 1 & \text{if } Na \text{ is even}\\ -1 & \text{if } Na \text{ is odd} \ . \end{cases}
\end{aligned}
$$

Hence, Nb has the same parity as Na. □

Remark 4. The set of high-amplitude points of the interference pattern in R, described in Proposition 6, models the reduced grid described in Remark 1 as follows: a point $\left(\frac{a}{N},\frac{b}{N},0\right)$ in the former represents (a,b) in the latter. (In fact, the whole region R, of which the high-amplitude points are a subset, corresponds under the same transformation $\left(\left(\frac{x}{N},\frac{y}{N},0\right) \mapsto (x,y)\right)$ to the region $\{(x,y) : 0 \le x \le y \le N\}$, of which the reduced grid is a subset.) This change of scale, by a multiplicative factor of N, of each axis of the (x,y)-plane is carried out in order that the dimensions and layout of the apparatus described be independent of the choice of N, in practice allowing use of the same apparatus for different values of N.

2.2 Implementation of the Cone

Definition 2 (in which we implement the cone).

1. *Let P_N be a source at $\left(0,0,\sqrt{\frac{2}{N}}\right)$ of electromagnetic radiation.*
2. *Let C_N be a detector along the curve*

$$\left\{(x,2-x,z): \begin{array}{cc} & 2\left(x-1\right)^2+\left(z-\sqrt{\frac{2}{N}}\right)^2=2 \\ \wedge & z\leq\frac{1-N}{1+N}\sqrt{\frac{2}{N}} \\ \wedge & 2-x\geq 1 \end{array}\right\} .$$

Proposition 7. *The curve of C_N is the circular arc produced by projecting the curve $G_N := \left\{(x,y,0)\in R : \frac{1}{xy}=N\right\}$ onto the plane $y=2-x$ from P_N. Hence, radiation arriving from P_N at a point on C_N passes through the plane $z=0$ at a point $(x,y,0)$ such that $\frac{1}{xy}=N$.*

Proof. Let $(a,b,0)$ be an arbitrary point in G_N; then $\frac{1}{ab}=N$, so this point is $\left(a,\frac{1}{Na},0\right)$ (note that $a,b>0$). The line that passes through both $\left(a,\frac{1}{Na},0\right)$ and P_N is given by $\left\{\left(a,\frac{1}{Na},0\right)+\gamma\left(a,\frac{1}{Na},-\sqrt{\frac{2}{N}}\right):\gamma\in\mathbb{R}\right\}$; this is equal to $\left\{\left((\gamma+1)\,a,\frac{\gamma+1}{Na},-\gamma\sqrt{\frac{2}{N}}\right):\gamma\in\mathbb{R}\right\}$. This line meets the plane defined by $y=2-x$ at $\left(\frac{2Na^2}{1+Na^2},\frac{2}{1+Na^2},\frac{1+Na^2-2Na}{1+Na^2}\sqrt{\frac{2}{N}}\right)$ (that is, when $\gamma+1=\frac{2Na}{1+Na^2}$). Now

$$\begin{aligned} &2\left(\frac{2Na^2}{1+Na^2}-1\right)^2 \\ &+\left(\frac{1+Na^2-2Na}{1+Na^2}\sqrt{\frac{2}{N}}-\sqrt{\frac{2}{N}}\right)^2 &&= 2\left(\frac{2Na^2-1-Na^2}{1+Na^2}\right)^2 \\ &&&\quad+\frac{2}{N}\left(\frac{1+Na^2-2Na-1-Na^2}{1+Na^2}\right)^2 \\ &&&= \frac{2\left(\left(Na^2-1\right)^2+\frac{1}{N}\left(-2Na\right)^2\right)}{\left(1+Na^2\right)^2} \\ &&&= \frac{2\left(N^2a^4-2Na^2+1+4Na^2\right)}{\left(1+Na^2\right)^2} \\ &&&= 2\ , \end{aligned}$$

so the point $\left(\frac{2Na^2}{1+Na^2},\frac{2}{1+Na^2},\frac{1+Na^2-2Na}{1+Na^2}\sqrt{\frac{2}{N}}\right)$ satisfies the first of the three conditions in the definition of C_N (namely '$2\left(x-1\right)^2+\left(z-\sqrt{\frac{2}{N}}\right)^2=2$').

Further, the quadratic $q(x) := Nx^2-(N+1)\,x+1$ has a positive leading coefficient (namely N), and so, in the range $\frac{1}{N}\leq x\leq 1$, attains its maximum

at either $x = \frac{1}{N}$ or $x = 1$; so, since $0 < b \leq 1$ and $\frac{1}{ab} = N$ give that $a \geq \frac{1}{N}$, and since $a \leq 1$, $q(a) \leq \max\left\{q\left(\frac{1}{N}\right), q(1)\right\} = 0$. So

$$Na^2 - (N+1)a + 1 \leq 0 \ ;$$

multiplying throughout by $2N$ and rearranging,

$$-2Na + N + N^2a^2 - 2N^2a \leq -N - N^2a^2 \ .$$

So

$$\begin{aligned}\left(1 + Na^2 - 2Na\right)(1+N) &= 1 + Na^2 - 2Na + N + N^2a^2 - 2N^2a \\ &\leq 1 + Na^2 - N - N^2a^2 \\ &= (1-N)\left(1 + Na^2\right) \ .\end{aligned}$$

Hence, (because $1+N$ and $1+Na^2$ are both positive), $\frac{1+Na^2-2Na}{1+Na^2} \leq \frac{1-N}{1+N}$, and so the point $\left(\frac{2Na^2}{1+Na^2}, \frac{2}{1+Na^2}, \frac{1+Na^2-2Na}{1+Na^2}\sqrt{\frac{2}{N}}\right)$ satisfies the second of the three conditions of C_N (namely '$z \leq \frac{1-N}{1+N}\sqrt{\frac{2}{N}}$'), as $\sqrt{\frac{2}{N}}$ is positive.

Further, since $(a, b, 0) \in G_N$, whence $\frac{1}{ab} = N$ and (since $G_N \subseteq R$) $a \leq b$,

$$Na^2 \leq Nab = 1 \ .$$

So, adding 1,

$$1 + Na^2 \leq 2 = 2\left(1 + Na^2\right) - 2Na^2 \ .$$

Dividing by $1 + Na^2$ (which is positive),

$$1 \leq \frac{2\left(1 + Na^2\right) - 2Na^2}{1 + Na^2} = 2 - \frac{2Na^2}{1 + Na^2} \ .$$

Hence, the point $\left(\frac{2Na^2}{1+Na^2}, \frac{2}{1+Na^2}, \frac{1+Na^2-2Na}{1+Na^2}\sqrt{\frac{2}{N}}\right)$ satisfies the third condition of C_N (namely '$2 - x \geq 1$'), and so is on C_N.

Conversely, let $(a, 2-a, c)$ be an arbitrary point on C_N. By the first condition of C_N, $c = \pm\sqrt{2 - 2(a-1)^2} + \sqrt{\frac{2}{N}}$, and by the second, $c = -\sqrt{2 - 2(a-1)^2} + \sqrt{\frac{2}{N}}$, so $(a, 2-a, c) = \left(a, 2-a, \sqrt{\frac{2}{N}} - \sqrt{2 - 2(a-1)^2}\right)$. The line that passes through both this point and P_N is given by

$$\left\{ \begin{array}{ll} & \left(a, 2-a, \sqrt{\frac{2}{N}} - \sqrt{2 - 2(a-1)^2}\right) \\ + & \gamma\left(a, 2-a, -\sqrt{2 - 2(a-1)^2}\right) \end{array} : \gamma \in \mathbb{R} \right\} \ ;$$

that is, by

$$\left\{ \left((\gamma+1)a, (\gamma+1)(2-a), \sqrt{\frac{2}{N}} - (\gamma+1)\sqrt{2 - 2(a-1)^2} \right) : \gamma \in \mathbb{R} \right\} \ .$$

This meets the plane $z = 0$ when $\gamma + 1 = \sqrt{\frac{1}{Na(2-a)}}$, which corresponds to the point $A := \left(\sqrt{\frac{a}{N(2-a)}}, \sqrt{\frac{2-a}{Na}}, 0\right)$, which we wish to show to be in G_N. Note that, by the second condition of C_N ('$z \leq \frac{1-N}{1+N}\sqrt{\frac{2}{N}}$'), $\sqrt{\frac{2}{N}} - \sqrt{2-2\left(a-1\right)^2} \leq \frac{1-N}{1+N}\sqrt{\frac{2}{N}}$, whence $\sqrt{2-2\left(a-1\right)^2} \geq \sqrt{\frac{2}{N}}\left(1 - \frac{1-N}{1+N}\right) = \sqrt{\frac{2}{N}} \cdot \frac{2N}{1+N}$, so $2 - 2\left(a-1\right)^2 \geq \frac{2}{N} \cdot \frac{4N^2}{(1+N)^2} = \frac{8N}{(1+N)^2}$, whence $a\left(2-a\right) \geq \frac{4N}{(1+N)^2}$, and so $-a^2 + 2a - \frac{4N}{(1+N)^2} \geq 0$; from this, we have that $\frac{2}{N+1} \leq a \leq \frac{2N}{N+1}$. Further, by the third condition of C_N ('$2 - x \geq 1$'), $a \leq 1$, so

$$\frac{2}{N+1} \leq a \leq 1 \ .$$

So,

1. since a, N and $2-a$ are positive, $0 \leq \sqrt{\frac{a}{N(2-a)}}$;
2. since $a \leq 1$, $a^2 \leq 4 - 4a + a^2 = \left(2-a\right)^2$, whence $\frac{a}{N(2-a)} \leq \frac{2-a}{Na}$, and so $\sqrt{\frac{a}{N(2-a)}} \leq \sqrt{\frac{2-a}{Na}}$; and
3. since $\frac{2}{N+1} \leq a$, $2 \leq a\left(N+1\right)$, so $2 - a \leq Na$ and $\frac{2-a}{Na} \leq 1$, whence $\sqrt{\frac{2-a}{Na}} \leq 1$.

That is, $0 \leq \sqrt{\frac{a}{N(2-a)}} \leq \sqrt{\frac{2-a}{Na}} \leq 1$, and so $A \in R$.

Further, we have that the product of the first and second coordinates of A is $\sqrt{\frac{a}{N(2-a)} \cdot \frac{2-a}{Na}} = \sqrt{\frac{1}{N^2}} = \frac{1}{N}$, and so the point is, as claimed, in G_N. □

Remark 5. By Remark 4 and Proposition 7, the radiation from P_N arriving at C_N passes through the curve in R corresponding to the curve $y = \frac{N}{x}$ (that is, through G_N). Such a ray passes through a point corresponding to an integer solution on this curve if and only if the point displays the interference pattern of S at maximum amplitude; that this is the case is then evident at C_N.

3 Interpreting Results

Remark 6. Recall from Remark 5 that the radiation arriving from P_N at a point on C_N will display high-amplitude interference (because of the standing wave from S) if and only if the point $(x, y, 0)$ of R through which it passes offers a factorization of N (in that $\frac{1}{x} \cdot \frac{1}{y} = N$, where $\frac{1}{x}$ and $\frac{1}{y}$ are integers). Thus, the interpretation of results consists mainly of converting the coordinates of a point (that which displays high-amplitude interference) on C_N into those of a point in R (that through which the ray passes). Proposition 8 describes this conversion.

Proposition 8. *Radiation from P_N incident on a point $(a, 2-a, c)$ on C_N has passed through $\left(\sqrt{\frac{a}{N(2-a)}}, \sqrt{\frac{2-a}{Na}}, 0\right)$.*

Proof. This claim is justified in the proof of Proposition 7. □

Corollary 1. *If the radiation from P_N at $(a, 2-a, c)$ on C_N displays high-amplitude interference, then $\sqrt{\frac{Na}{2-a}}$ and $\sqrt{\frac{N(2-a)}{a}}$ are factors of N; conversely, all factors of N have an analogous point on C_N.*

Proof. This follows from Remark 4, Remark 5 and Proposition 8. □

Remark 7. Having set up the apparatus as described in Definitions 1 and 2, the factors of N are found as in Corollary 1. Since all factors are represented by points on C_N displaying high-amplitude interference (and since there are no other such points), a value of N produces

1. no such points if and only if N is not an integer,
2. a single such point (corresponding to the factorization $N = 1.N$) if and only if N is prime (or one), and
3. two or more such points if and only if N is composite.

In particular, by sweeping continuously through a range of values of N (by continuously altering the wavelength of S, for example with a variable resistor, and the height—that is, z-coordinate—of P_N and C_N), primes can be quickly identified.

Remark 8 (Aside). As mentioned in the abstract, the security of the RSA cryptographic system relies on the intractability of factorization. However, the ability of the proposed method to factorize virtually instantaneously does not compromise this security: if technology were sufficient to allow the method to reliably factorize, say, n-digit numbers (and, hence, decrypt information encoded with RSA using an n-digit key), then, by Remark 7, n-digit primes can be found, and, by multiplying two such, a $(2n-1)$- or $2n$-digit RSA key can be formed.

4 Generalization

The task of factorization has a geometric formulation as the extraction of integer solutions of an equation of which the graph is part of a conic section; this formulation is exploited by the proposed method of factorization.

It is clear that, by repositioning P_N and C_N so as to implement a different cone, an identical method allows computation of integer solutions of different conic section graphs (parabolae, hyperbolae, circles and ellipses) or parts thereof. So, while factorization is chosen for discussion because of its wide range of applications and its notoriety as a difficult problem, the task is merely an illustration of a larger class of problems that the general method presented here can be used to solve.

5 Summary

In Sect. 1, the task of finding factors of a given integer is restated as a geometric problem, wherein points both on a curve and in a grid are sought. Noting in particular that this curve is a conic section, the physical implementation of the geometric formulation is detailed in Sect. 2.

The method whereby a number is input and a factorization found using the apparatus described is given in Sect. 3.

Sect. 4 notes that factorization is just one use of the general method proposed, and describes the way in which the process can be altered in order to find integer solutions to other equations.

The method presented here addresses the computational difficulty encountered when using traditional algorithms to find integer solutions to certain equations (e.g. when factorizing integers). The proposed system of factorization is qualitatively different from existing processes because it uses a direct physical implementation of the problem in preference to the standard model of computation; this allows for much-improved calculation times.

That the system enjoys both time and space complexities that are constant in the size of the input value, however, serves to highlight not the power of the method but the incompleteness of traditional complexity theory. As N increases, the system does in fact require more resource (though neither specifically time nor space) to function; namely, the *precision* with which N must be input (by setting the wavelength of the grid source and the height of the cone) and its factors read (by measuring the positions of points on C_N) increases with N. This suggests that, for some analogue computers, traditional 'algorithmic' complexity theory is inadequate; more suitable notions of complexity are currently under consideration by the author and others.

28.vi.2007