A new approach to solving the cubic: Cardan's solution revealed [1]

RWD Nickalls [2]

The Mathematical Gazette (1993); 77 (November), 354–359 (JSTOR)
www.nickalls.org/dick/papers/maths/cubic1993.pdf

1 Introduction

The cubic holds a double fascination since not only is it interesting in its own right, but its solution is also the key to solving quartics. [3] This article describes five fundamental parameters of the cubic (δ, λ, h, x_N and y_N), and shows how they lead to a significant modification of the standard method of solving the cubic, generally known as *Cardan's solution.*

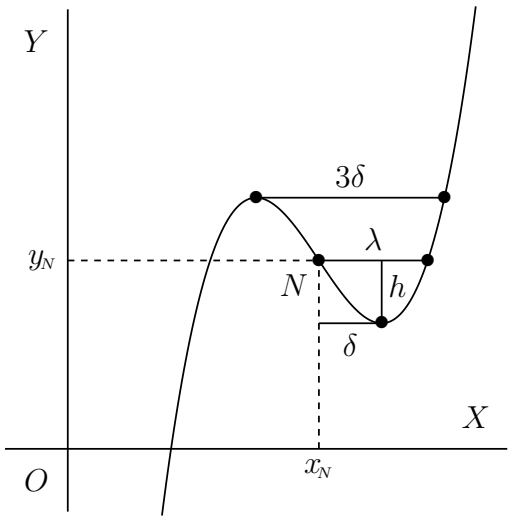

Figure 1:

[1] This minor revision of the original article corrects typographic errors and incorporates some explanatory footnotes, a special font for $\mathbb{G}$ and $\mathbb{H}$, a few minor text changes for clarification, as well as a minor improvement to Figure 2 . The original published version is available from www.jstor.org/stable/3619777.

[2] Department of Anaesthesia, Nottingham University Hospitals, City Hospital Campus, Nottingham, UK. email: dick@nickalls.org

[3] Nickalls RWD (2009). The quartic equation: invariants and Euler's solution revealed. *The Mathematical Gazette*; 93, 66–75. (www.nickalls.org/dick/papers/maths/quartic2009.pdf)

It is necessary to start with a definition. Let $N(x_N, y_N)$ be a point on a polynomial curve $f(x)$ of degree n such that moving the axes by putting $x = z + x_N$ makes the sum of the roots of the new polynomial $g(z)$ equal to zero (see Section 2.1). It is easy to show that for the polynomial equation

$$ax^n + bx^{n-1} + \ldots + k = 0$$

$x_N = -b/(na)$. If $f(x)$ is a cubic polynomial then $g(z)$ is known as the reduced cubic, and N is the point of inflection.

Now consider the general cubic

$$y = ax^3 + bx^2 + cx + d.$$

Here x_N is $-b/(3a)$, and N the point of symmetry of the cubic. Let the parameters δ, λ, h, be defined as the distances shown in Figure 1. It can be shown, and readers will easily do this, that λ and h are simple functions of δ namely [4, 5]

$$\lambda^2 = 3\delta^2 \quad \text{and} \quad h = 2a\delta^3,$$

where

$$\delta^2 = \frac{b^2 - 3ac}{9a^2}.$$

This result is found easily by locating the turning points. Thus the shape of the cubic is completely characterised by the parameter δ. Either the maxima and minima are distinct ($\delta^2 > 0$), or they coincide at N ($\delta^2 = 0$), or there are no real turning points ($\delta^2 < 0$). Furthermore, the quantity $a\delta\lambda^2/h$ is constant for any cubic, as follows

$$\frac{a\delta\lambda^2}{h} = \frac{3}{2}.$$

The relationship $\lambda^2 = 3\delta^2$ is a particular case of the general observation that

> If a polynomial curve passes through the origin, then the product of the roots $x_1, x_2, \ldots, x_{n-1}$ (excluding the solution $x = 0$) is related to the product of the x-coordinates of the turning points $t_1 t_2 \ldots t_{n-1}$ by
>
> $$x_1 x_2 \ldots x_{n-1} = n t_1 t_2 \ldots t_{n-1},$$

a result whose proof readers can profitably set to their classes, and which parallels a related but more difficult result about the y-coordinates of the turning points which we have discovered. [6]

[4] Unfortunately in the original printed version h was presented with a negative sign.

[5] Note that if m_N denotes the slope at the point of inflection N, then $m_N = -3a\delta^2$. See also Thomas Müller's interactive demonstration of how the parameters x_N, y_N, h, δ, m_N influence the shape and location of the cubic at demonstrations.wolfram.com/ParametersForPlottingACubicPolynomial/

[6] Nickalls RWD and Dye RH (1996). The geometry of the discriminant of a polynomial. *The Mathematical Gazette*, 80 (July), 279–285 (JSTOR). (www.nickalls.org/dick/papers/maths/discriminant1996.pdf)

2 Solution of the cubic

In addition to their value in curve tracing, I have found that the parameters δ, h, x_N and y_N, greatly clarify the standard method for solving the cubic since, unlike the Cardan approach (Burnside and Panton, 1886) [7] they reveal how the solution is related to the geometry of the cubic.

For example, the standard Cardan solution using the classical terminology, involves starting with an equation of the form [8]

$$ax^3 + 3b_1x^2 + 3c_1x + d = 0,$$

and then substituting $x = z - b_1/a$ to generate a reduced equation of the form [9]

$$az^3 + \frac{3\mathbb{H}}{a}z + \frac{\mathbb{G}}{a^2} = 0,$$

where

$$\mathbb{H} = ac_1 - b_1^2 \quad \text{and} \quad \mathbb{G} = a^2d - 3ab_1c_1 + 2b_1^3.$$

Subsequent development yields a discriminant of the form $\mathbb{G}^2 + 4\mathbb{H}^3$ where

$$\mathbb{G}^2 + 4\mathbb{H}^3 = a^2(a^2d^2 - 6ab_1c_1d + 4ac_1^3 + 4b_1^3d - 3b_1^2c_1^2).$$

The problem is that it is not clear geometrically what the quantities $\mathbb{G}$ and $\mathbb{H}$ represent. However, by using the parameters described earlier, not only is the solution just as simple but the geometry is revealed.

2.1 New approach

Start with the usual form of the cubic equation

$$f(x) \equiv ax^3 + bx^2 + cx + d = 0, \tag{1}$$

having roots α, β, γ, and obtain the reduced form by the substitution $x = z + x_N$ (see Figure 1). The equation will now have the form [10]

$$g(z) \equiv az^3 - 3a\delta^2z + y_N = 0, \tag{2}$$

and have roots $z_1 = \alpha - x_N$, $z_2 = \beta - x_N$, $z_3 = \gamma - x_N$; a form which allows the use of the usual identity

$$(p+q)^3 - 3pq(p+q) - (p^3 + q^3) = 0.$$

Thus $z = p + q$ is a solution of (2) where

$$pq = \delta^2 \quad \text{and} \quad p^3 + q^3 = -y_N/a.$$

[7] §§ 56–57 (pages 106–109) —see footnote 19 for URL.

[8] This is the (historical) binomial form as used by Burnside and Panton (1886).

[9] We use the maths font (`\mathbb`) here for $\mathbb{G}$ and $\mathbb{H}$ to indicate that these particular invariants are associated with the binomial equation format. A normal font can then be used for invariants associated with the normal equation format, as in Nickalls (2009) —see footnote 3 for URL.

[10] $y_N \equiv f(x_N) \equiv f(-b/3a) = \left\{2b^3/\left(27a^2\right)\right\} - \{bc/(3a)\} + d$

Solving these equations as usual by cubing the first, substituting for q in the second, and solving the resulting quadratic in p^3 gives

$$p^3 = \frac{1}{2a}\left\{-y_N \pm \sqrt{y_N^2 - 4a^2\delta^6}\ \right\},$$

and, since $h^2 = 4a^2\delta^6$, this becomes

$$p^3 = \frac{1}{2a}\left\{-y_N \pm \sqrt{y_N^2 - h^2}\ \right\}. \tag{3}$$

When this solution is viewed in the light of Figure 1, it is immediately clear that Equation 3 is particularly useful when there is a single real root, that is when

$$y_N^2 > h^2.$$

Contrast this with the standard Cardan approach which gives

$$p^3 = \frac{1}{2a^3}\left\{-\mathbb{G} \pm \sqrt{\mathbb{G}^2 + 4\mathbb{H}^3}\ \right\},$$

which completely obscures this fact. The values of $\mathbb{G}$, $\mathbb{H}$, and $\mathbb{G}^2 + 4\mathbb{H}^3$ are therefore found to be

$$\mathbb{G} = a^2 y_N, \quad \mathbb{H} = -a^2\delta^2 \quad \text{and} \quad \mathbb{G}^2 + 4\mathbb{H}^3 = a^4(y_N^2 - h^2).$$

However, Equation 3 can be rewritten as

$$p^3 = \frac{1}{2a}\left\{-y_N \pm \sqrt{(y_N + h)\,(y_N - h)}\ \right\}.$$

If the y-coordinate of a turning point is y_T then let

$$y_N + h = y_{T_1} \quad \text{and} \quad y_N - h = y_{T_2}.$$

Our solution (Equation 3) can therefore be written as

$$p^3 = \frac{1}{2a}\left\{-y_N \pm \sqrt{y_{T_1} y_{T_2}}\ \right\}.$$

Using the symbol Δ_3 for the (geometric) discriminant [11, 12] of the cubic, we have

$$\Delta_3 = y_{T_1} y_{T_2} = y_N^2 - h^2.$$

---

[11] The product $y_{T_1} y_{T_2}$ of the y-coordinates of the turning points is known as the *geometric* discriminant Δ_3 of the cubic; it is the geometric analogue of the algebraic (classical) discriminant (see Nickalls and Dye (1996)—for URL see footnote 6). The classical discriminant $\mathbb{G}^2 + 4\mathbb{H}^3$ has the same sign as the geometric discriminant since $\mathbb{G}^2 + 4\mathbb{H}^3 = a^4(y_N^2 - h^2) = a^4 y_{T_1} y_{T_2} = a^4\Delta_3$.

[12] The *algebraic* discriminant D_3 of the cubic is defined as the product of squared differences of the roots; $D_3 = (\alpha-\beta)^2(\alpha-\gamma)^2(\beta-\gamma)^2$ and hence $a^2 D_3 = -27(y_N^2 - h^2) = -27 y_{T_1} y_{T_2} = -27\Delta_3$. Note also that $a^4 D_3 = 18abcd - 4ac^3 - 27a^2d^2 + b^2c^2 - 4b^3d = -27a^2(y_N^2 - h^2)$.

Returning to the geometrical viewpoint, Figure 1 shows that the rest of the solution depends on the sign of the discriminant [13] as follows:

$$\begin{aligned} y_N^2 > h^2 \qquad & \text{1 real root,} \\ y_N^2 = h^2 \qquad & \text{3 real roots (two or three equal roots),} \\ y_N^2 < h^2 \qquad & \text{3 distinct real roots.} \end{aligned}$$

These are now dealt with in order.

2.2 $y_N^2 > h^2$ i.e. $y_{T_1} y_{T_2} > 0$, or Cardan's $\mathbb{G}^2 + 4\mathbb{H}^3 > 0$

Clearly, there can only be one real root of equation 1 under these circumstances (see Figure 1). As the discriminant is positive the value of the real root α is easily obtained as [14]

$$\alpha = x_N + \sqrt[3]{\frac{1}{2a}\left(-y_N + \sqrt{y_N^2 - h^2}\right)} + \sqrt[3]{\frac{1}{2a}\left(-y_N - \sqrt{y_N^2 - h^2}\right)}.$$

2.3 $y_N^2 = h^2$ i.e. $y_{T_1} y_{T_2} = 0$, or Cardan's $\mathbb{G}^2 + 4\mathbb{H}^3 = 0$

Providing $h \neq 0$ this condition yields two equal roots, the roots being $z = \delta$, δ and -2δ. The roots of (1) are then $x_N + \delta$, $x_N + \delta$ and $x_N - 2\delta$. Since there are two double root conditions the sign of δ is critical, and depends on the sign of y_N, and so in these circumstances δ has to be determined from

$$\delta = \sqrt[3]{\frac{y_N}{2a}}.$$

If $y_N = h = 0$ then $\delta = 0$, in which case there are three equal roots at $x = x_N$.

[13] Since the sign of the discriminant $(y_N^2 - h^2)$ reflects the relative magnitude of y_N^2 and h^2.

[14] The remaining two complex roots of equation 1 are given by

$$\beta,\, \gamma = x_N - \frac{z_1}{2} \pm j\frac{\sqrt{3}}{2}\sqrt{z_1^2 - 4\delta^2}$$

where $z_1 = \alpha - x_N$ (see equation 2) and $j^2 = -1$ (for derivation see: Nickalls RWD (2009). Feedback: 93.35: *The Mathematical Gazette*; 93 (Mar), 154–156. www.nickalls.org/dick/papers/maths/cubictables2009.pdf).

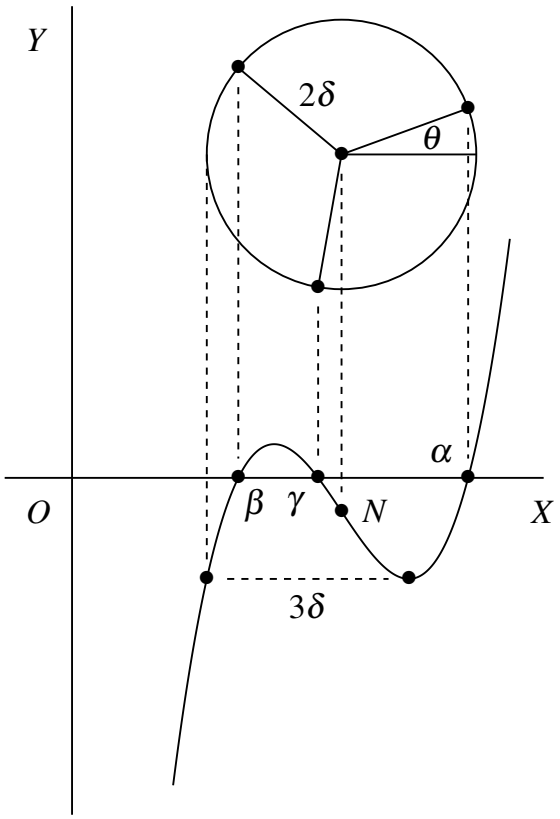

Figure 2:

2.4 $y_N^2 < h^2$ i.e. $y_{T_1} y_{T_2} < 0$, or Cardan's $\mathbb{G}^2 + 4\mathbb{H}^3 < 0$

From Figures 1 and 2 it is clear that there are three distinct real roots in this case. However, our solution requires that we find the cube root of a complex number, so it is easier to use trigonometry to solve the reduced form using the substitution [15] $z = 2\delta\cos\theta$ in Equation 2. This gives

$$2a\delta^3\left(4\cos^3\theta - 3\cos\theta\right) + y_N = 0,$$

and since $2a\delta^3 = h$, this becomes

$$\cos 3\theta = \frac{-y_N}{h}. \tag{4}$$

[15] To avoid the negative sign in (4) use instead $z = 2\delta\sin\phi$, as this leads to $\sin 3\phi = y_N/h$, for which the condition $y_N = 0$ is associated with $\phi = 0$. Use of this form is described in Nickalls, RWD (2006); Viète, Descartes and the cubic equation. *The Mathematical Gazette*; 90 (July), 203–208. (www.nickalls.org/dick/papers/maths/descartes2006.pdf). Alternatively, setting $h = -2a\delta^3$ yields $\cos 3\theta = y_N/h$ (see Müller T (2008); *Physical Review D*; 77, 124042. (dx.doi.org/10.1103/PhysRevD.77.124042))

The three roots α, β and γ are therefore given by [16]

$$\begin{cases} \alpha = x_N + 2\delta\cos\theta, \\ \beta = x_N + 2\delta\cos(\theta + 2\pi/3), \\ \gamma = x_N + 2\delta\cos(\theta + 4\pi/3). \end{cases} \tag{5}$$

These are shown in Figure 2 in relation to a circle, radius 2δ, centred above N. Note that the maximum between roots β and γ corresponds to the angle $2\pi/3$.

[16] To see this, note that it follows from Equation 4 that

$$\begin{cases} y_N = -h\cos 3\theta, \\ y_N^2 - h^2 = -h^2\sin^2 3\theta. \end{cases}$$

Substituting these into the Equation in Section 2.2 gives

$$\alpha = x_N + \sqrt[3]{\frac{1}{2a}\left(h\cos 3\theta + \sqrt{-h^2\sin^2 3\theta}\,\right)} + \sqrt[3]{\frac{1}{2a}\left(h\cos 3\theta - \sqrt{-h^2\sin^2 3\theta}\,\right)}.$$

Since $h = 2a\delta^3$ this is equivalent to

$$\alpha = x_N + \sqrt[3]{\delta^3(\cos 3\theta + j\sin 3\theta\,)} + \sqrt[3]{\delta^3(\cos 3\theta - j\sin 3\theta\,)},$$

i.e.,

$$\alpha = x_N + \delta(\cos\theta + j\sin\theta\,) + \delta(\cos\theta - j\sin\theta\,),$$

and hence

$$\alpha = x_N + 2\delta\cos\theta.$$

It is clear from Equation 4 that trigonometry [using real θ] [17] can only be used to solve the reduced cubic when

$$-1 \leq \frac{y_N}{h} \leq +1,$$

a point which is completely obscured by the corresponding classical form

$$\cos 3\theta = \frac{-\mathbb{G}}{2\,(-\mathbb{H})^{3/2}}.$$

[17] However, if we extend the range of θ to include complex angles ($\theta \in \mathbb{C}$), then *all* cubics can be solved using this approach (see Holmes GC (2002); The use of hyperbolic cosines in solving cubic polynomials. *The Mathematical Gazette*, 86, 473–477 (www.jstor.org/stable/3621149)).

For example, consider the reduced cubic $f(x) \equiv x^3 - 3x - 52 = 0$ (roots: 4, $-2 \pm 3i$), for which $x_N = 0$, $y_N = f(x_N) = -52$, $\delta^2 = 1$, $h = 2a\delta^3 = 2$, and $-y_N/h = 26$.

Using (5) let $x_0 = x_N + 2\delta\cos\theta$ be the principal root, where $\cos 3\theta = -y_N/h$. Since in this case $|y_N/h| > 1$ (i.e., $y_N^2 > h^2$) it follows that $\theta \in \mathbb{C}$. Consequently we start by setting $\theta = \alpha + i\beta$ (where α and β are real values), and hence the expression $\cos 3\theta = -y_N/h$ is equivalent to

$$\cos 3(\alpha + i\beta) = 26.$$

Expanding this using the identity $\cos(A + B) = \cos A\cos B - \sin A\sin B$, and then letting $\cos i3\beta \to \cosh 3\beta$ and $\sin i3\beta \to i\sinh 3\beta$, we obtain $\cos 3\alpha\cosh 3\beta - i\sin 3\alpha\sinh 3\beta = 26$. Separating the real ($Re$) and complex ($Im$) elements, we obtain a pair of simultaneous equations involving both circular and hyperbolic functions, as follows:

$$\begin{cases} Re: & \cos 3\alpha\,\cosh 3\beta = 26, \\ Im: & -\sin 3\alpha\,\sinh 3\beta = 0. \end{cases}$$

The solution for α and β is as follows:
We start by observing that either $\sin 3\alpha$ or $\sinh 3\beta$ must be equal to zero.
If $\sinh 3\beta = 0$ then $\cosh 3\beta = 1 \Longrightarrow \cos 3\alpha = 26$ which is impossible (since $-1 \leq \cos x \leq +1$). Consequently, it must be that $\sin 3\alpha = 0$ here, and hence either $3\alpha = 0$ (i.e., $\cos 3\alpha = +1$) or $3\alpha = \pm\pi$ (i.e., $\cos 3\alpha = -1$). However, $\cos 3\alpha = -1 \Longrightarrow \cosh 3\beta = -26$ which is impossible (since $\cosh x \geq +1$), and hence it must be that $\cos 3\alpha = +1$ (which implies $\cosh 3\beta = 26$). Let 3α denote the principal value of $\cos^{-1} 1$ ($= 0$). Let 3β denote the principal (+ve) value [**] of $\cosh^{-1} 26$ ($\approx +3{\cdot}95087$). Consequently we can write

$$\begin{cases} \alpha = (1/3)\cos^{-1} 1 = 0, \\ \beta = (1/3)\cosh^{-1} 26 \approx 1{\cdot}316957 \text{ radians}. \end{cases}$$

Thus the three roots x_k are given by

$$x_k = x_N + 2\delta\cos\left(\theta + \frac{2\pi k}{3}\right) \quad (k = 0, 1, 2),$$

where $\theta = \alpha + i\beta$ and $\cos 3\theta = -y_N/h$.

When $k = 0$ then $x_0 = 2\delta\cos(i\beta)$ (since in this particular case $x_N = 0$ and $\alpha = 0$). Since $\delta = 1$ (see above) we obtain the real root x_0, as follows:

$$x_0 = 0 + 2\delta\cos(0 + i\beta) = 2(1)\cos(i\beta) = 2\cosh\beta \approx 2\cosh(1{\cdot}316957) = 2 \times 2 = 4,$$

We leave it as an exercise for the reader to verify that $k = 1, 2$ do indeed generate the complex roots $-2 \pm 3i$.

For some insight as to why $\cosh(1{\cdot}316957)$ is (surprisingly) equal to the integer 2 see *Example 3* in: Nickalls (2021), On the structure of polynomial roots, *Mathematical Gazette*, 105, pp.253–262. www.nickalls.org/dick/papers/maths/rootstructure2021.pdf

[**] Observe that $\cosh^{-1} 26 \approx \pm 3{\cdot}95087$; we can use either sign since $\cosh\beta = \cosh(-\beta) \geq +1$.

3 Example

Solve the equation [18]

$$f(x) \equiv x^3 - 7x^2 + 14x - 8 = 0$$

The parameters are

$$x_N = 7/3, \quad y_N = f(x_N) \approx -0{\cdot}74074, \quad \delta^2 = 7/9 \quad \text{and} \quad h = 2a\delta^3 \approx 1{\cdot}37187.$$

Since $y_N^2 < h^2$, it follows (see Figures 1 and 2) that there are three distinct real roots, which are given by

$$x = x_N + 2\delta\cos\theta,$$

where

$$\cos 3\theta = \frac{-y_N}{h} \approx \frac{0{\cdot}74074}{1{\cdot}37187} \approx 0{\cdot}53995.$$

So $\theta \approx 19{\cdot}1066°$, and using (5) the three roots are

$$\begin{cases} \alpha = \dfrac{7}{3} + 2\sqrt{\dfrac{7}{9}}\cos 19{\cdot}1066° &= 4, \\ \beta = \dfrac{7}{3} + 2\sqrt{\dfrac{7}{9}}\cos 139{\cdot}1066° &= 1, \\ \gamma = \dfrac{7}{3} + 2\sqrt{\dfrac{7}{9}}\cos 259{\cdot}1066° &= 2. \end{cases}$$

4 Conclusion

Finally, I would like to suggest that the usual Cardan-type terminology for cubics and quartics, though it has been used for hundreds of years, be abandoned in favour of the parameters δ, h, x_N, y_N, which reveal to such advantage how the algebraic solution is related to the geometry of the cubic.

5 References

- Burnside W. S. and Panton A. W. (1886). [19] *The theory of equations: with an introduction to the theory of binary algebraic forms.* 2nd ed. Longmans, Green and Co., London [reprinted: Dover, 1960; 2 vols (7th ed., 1928)] [reprinted: Dover, 2005, as a single volume; ISBN 0-486-44140-7].

◇ ◇ ◇

[18] For another example see: Nickalls RWD (1996). A note on solving cubics. *The Mathematical Gazette*; 80 (November), 576–577 (JSTOR). (www.nickalls.org/dick/papers/maths/cubefink.pdf)

[19] The 1st ed (1881) is available on the web at http://quod.lib.umich.edu/cgi/t/text/text-idx?c=umhistmath;idno=ACA7397 See §§ 56–57 (pp. 104–107), and the excellent historical notes at back of the book, especially Note A *Algebraic solution of equations*, pp. 433–436, and Note B *Solution of numerical equations*, pp. 437–440.