Experimental Mathematics 2020

Alex Ghitza

Version of Tue 9th Jun, 2020 at 10:21

1 Gauss's agM

From Gauss's diary, entry dated 30 May 1799:

> Terminum medium arithmetico–geometricum inter 1 et $\sqrt{2}$ esse = $\frac{\pi}{\varpi}$ usque ad figuram undecimam comprobavimus, qua re demonstrata prorsus novus campus in analysis certo aperietur.

For those as Latin-challenged as I am, this translates to

> We have established that the arithmetic–geometric mean between 1 and $\sqrt{2}$ is = $\frac{\pi}{\varpi}$ to the eleventh decimal place; the demonstration of this fact will surely open an entirely new field of analysis.

This probably raises more questions than it answers. So let's look at it in more detail.

1.1 Two sequences, one limit

Consider the two sequences of real numbers (a_n) and (b_n) defined by the recursions

$$a_0 = \sqrt{2}, \qquad a_{n+1} = \frac{a_n + b_n}{2}$$
$$b_0 = 1, \qquad b_{n+1} = \sqrt{a_n b_n}$$

This looks a little strange, as the two definitions are intertwined (and what's up with the initial values $\sqrt{2}$ and 1?), but we can recognise some familiar features, at least locally. The recursive rule defining a_{n+1} takes the arithmetic mean of the numbers a_n and b_n. And the recursive rule defining b_{n+1} takes the geometric mean of the same two numbers.

I wonder how the two sequences behave as n increases. If I were Gauss, I would compute the first few terms by hand (to 11 decimal places, indeed). Having the luxury of living in the era of funny cat YouTube videos (and incidental infrastructure), I'll instead call upon SageMath on my computer:

```
sage: def a(n):
....:     if n == 0:
....:         return RR(sqrt(2))
....:     else:
....:         return RR((a(n-1)+b(n-1))/2)
sage: def b(n):
....:     if n == 0:
....:         return RR(1)
....:     else:
....:         return RR(sqrt(a(n-1)*b(n-1)))
```

Having defined the two sequences, I can now ask for values:

```
sage: a(1)
1.20710678118655
sage: a(2)
1.19815694809463
sage: a(3)
1.19814023479388
sage: a(4)
1.19814023473559
sage: a(5)
1.19814023473559
```

One might guess now that (a_n) converges to a number close to 1.19814023473559. How about (b_n)?

```
sage: b(1)
1.18920711500272
sage: b(2)
1.19812352149312
sage: b(3)
1.19814023467731
sage: b(4)
1.19814023473559
sage: b(5)
1.19814023473559
```

This also seems to converge, and to the same limit? Let's be bold and proclaim:

Proposition 1.1. *Given any starting values $a_0 = a \geqslant b_0 = b \in \mathbb{R}_{>0}$, the sequences (a_n) and (b_n) both converge to the same limit $M(a, b)$.*

Proof. More of a sketch, really.

There are two things to prove: that the sequences converge, and that they have the same limit[1].

Here is one approach to this, courtesy of Wikipedia:

- $b_n \leqslant a_n$ for all n. This is clear for $n = 0$ and true by simple algebraic manipulation in general (the arithmetic mean of two numbers is never smaller than their geometric mean).
- Use the previous part to note that $b_{n+1} \geqslant b_n$, so the sequence (b_n) is non-decreasing.
- Note that $b_n \leqslant a$ for all $n \geqslant 0$, so the sequence (b_n) is bounded above. Hence it has a limit L.
- Note that $b_n \geqslant b > 0$ for all n so $L \geqslant b > 0$, in particular $L \neq 0$.
- Note that
$$a_n = \frac{b_{n+1}^2}{b_n},$$
so using a theorem about the limit of the quotient of two convergent sequences we conclude that (a_n) converges to $\frac{L^2}{L} = L$.

[1]It would not be sufficient to only show that the sequence of differences $(a_n - b_n)$ converges to 0

□

The real number $M(a, b)$ is called the *arithmetic-geometric mean (agM) of a and b*. The above numerical experiment leads us to believe that

$$M(\sqrt{2}, 1) = 1.19814023473559\ldots$$

But Gauss's diary entry said more about this value. What's that about?

1.2 An elliptic integral

Consider the *lemniscate (of Bernoulli)*, a plane curve given in polar coordinates (r, θ) by the equation

$$r^2 = \cos(2\theta).$$

It would be nice to graph it, wouldn't it? As far as I know, neither Sage nor Mathematica have a built-in command for implicit polar plots. After a bit of algebraic manipulation with $x = r\cos\theta$, $y = r\sin\theta$, I get the implicit Cartesian equation

$$(x^2 + y^2)^2 = x^2 - y^2.$$

This is more amenable to plotting:

```
sage: var('x, y')
(x, y)
sage: f = (x^2 + y^2)^2 - (x^2 - y^2)
sage: p = implicit_plot(f, (x, -1, 1), (y, -1, 1))
sage: p.show()
None
```

which produces the pretty infinity sign in the picture.

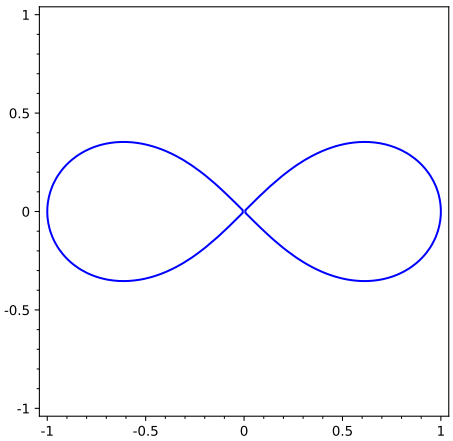


Figure 1.1: The lemniscate of Bernoulli, from Sage

Or you can follow a Mathematica lead from StackExchange:

https://mathematica.stackexchange.com/a/549

```
ContourPlot[
 Evaluate@With[
   {r = Sqrt[x^2 + y^2],
    theta = ArcTan[x, y]},
   r^2 - Cos[2*theta] == 0
   ],
{x, -1, 1}, {y, -1, 1}
]
```

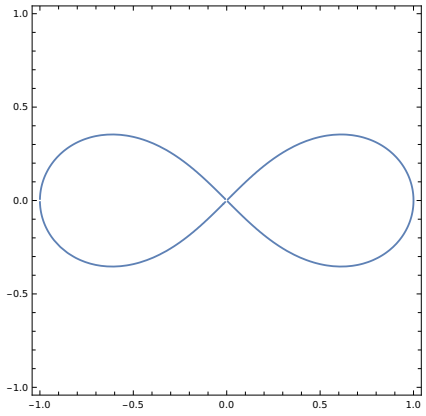


Figure 1.2: The lemniscate of Bernoulli, Mathematica style

The question is: what is the arclength of this curve?
The standard arclength formula in polar coordinates gives

$$4\int_0^{\pi/4}\left(r^2+\left(\frac{dr}{d\theta}\right)^2\right)^{1/2}d\theta = 4\int_0^{\pi/4}(\cos(2\theta))^{-1/2}\,d\theta.$$

If we introduce a new variable α with the property that $\cos(2\theta) = \cos^2\alpha$, the integral becomes

$$4\int_0^{\pi/2}\left(1+\cos^2\alpha\right)^{-1/2}d\alpha = 4\int_0^{\pi/2}\left(2\cos^2\alpha+\sin^2\alpha\right)^{-1/2}d\alpha.$$

One-half of this value is what Gauss denoted by ϖ in his diary entry.
We shouldn't just believe him like this. Let's check his calculations by estimating the integral numerically.
In Sage:

```
sage: f = 2/sqrt(2*cos(x)^2 + sin(x)^2)                                    37
sage: varpi = numerical_integral(f, 0, pi/2)[0]                            38
sage: RR(pi/varpi)                                                         39
1.19814023473559                                                           40
```

Or in Mathematica:

```
In[1]:= Pi/NIntegrate[2/Sqrt[2*Cos[t]^2+Sin[t]^2],{t,0,Pi/2},
                      {WorkingPrecision->15}]

Out[1]= 1.19814023473559
```

Fine, Gauss seems to have gotten his decimals right. More generally, he proved

Theorem 1.2. *If $a \geqslant b > 0$ then*

$$\int_0^{\pi/2} \left(a^2 \cos^2\alpha + b^2 \sin^2\alpha\right)^{-1/2} \, d\alpha = \frac{\pi}{2M(a,b)}$$

(These are examples of so-called *elliptic integrals*, and the arithmetic-geometric mean is a very efficient way of approximating them.)

Proof. Write

$$I(a,b) = \int_0^{\pi/2} \left(a^2 \cos^2\alpha + b^2 \sin^2\alpha\right)^{-1/2} \, d\alpha$$

We introduce a variable φ with the property that

$$\sin\alpha = \frac{2a \sin\varphi}{a + b + (a-b)\sin^2\varphi} \tag{1.1}$$

Then some secret magic sauce (in Gauss's words *"after the development has been done correctly, it will be seen"*[2]):

$$\left(a^2 \cos^2\alpha + b^2 \sin^2\alpha\right)^{-1/2} \, d\alpha = \left(a_1^2 \cos^2\varphi + b^2 \sin^2\varphi\right)^{-1/2} \, d\varphi$$

which leads us to the conclusion

$$I(a,b) = I(a_1, b_1).$$

A moment's further thought brings us to continue this as

$$I(a,b) = I(a_1,b_1) = I(a_2,b_2) = \cdots = I(a_n,b_n) = \ldots$$

and some judicious (real) analysis allows us to pass to the limit and get

$$I(a,b) = I(M(a,b), M(a,b)) = \frac{\pi}{2M(a,b)}$$

as the integral $I(c,c)$ is a piece of cake. □

Want to know more? There is plenty more where this came from, namely David Cox's articles [Cox85] (the shorter version) and [Cox84] (the longer one).

Ah, I can't resist mentioning one more thing, which I learned from Cox's papers. In 1973, Salamin discovered the formula

$$\pi = \frac{2M(\sqrt{2},1)^2}{1 - \sum_{n=1}^{\infty} 2^{n+1}(a_n^2 - b_n^2)}$$

which, as you can see, uses an infinite sum involving the terms a_n and b_n of the sequence we started with, as well as the common limit $M(\sqrt{2},1)$ of these two sequences. The existence of such a formula for π is weird enough, but it turns out [Sal76] that it's actually a pretty efficient way to compute lots of digits of π (in that the number of significant digits doubles at each step).

[2] Jacobi must have been as unimpressed as us by Gauss weaseling out of this, because he wrote down a couple of intermediate steps: Given the relation (1.1) between α and φ, show that

$$\cos\alpha = \frac{(2\cos\varphi)\left(a_1^2\cos^2\varphi + b_1^2\sin^2\varphi\right)^{1/2}}{a + b + (a-b)\sin^2\varphi}$$

and then that

$$\left(a^2\cos^2\alpha + b^2\sin^2\alpha\right)^{1/2} = a\,\frac{a + b - (a-b)\sin^2\varphi}{a + b + (a-b)\sin^2\varphi}$$

After this and implicitly differentiating (1.1), it's all smooth sailing to Gauss's equality of differentials.

1.3 Built-in agM commands

You might wonder if the arithmetic-geometric mean is already implemented in the software we're playing with.

But of course! In Mathematica, it is as simple as `ArithmeticGeometricMean`.

In Sage, it is a method (i.e. function attached to an object) of elements of `RR` (or other `RealField`s, and variants):

```
sage: a = RR(sqrt(2))
sage: a.agm(1)
1.19814023473559
```

Exercise 1.3. Implement your own function `myagm(a,b)` that returns the agM of two positive real numbers a and b. If you are so inclined, think about numerical issues.

2 Introduction to Sage and Mathematica

We look at a couple of innocent questions as an excuse to familiarise ourselves with the basics of the software. I would recommend reading through this in parallel with Sage's guided tour:

https://doc.sagemath.org/html/en/tutorial/tour.html

and bits of Mathematica's fast introduction for math students:

https://www.wolfram.com/language/fast-introduction-for-math-students/en/

2.1 Lines on a plane

What is the maximal number of regions $R(n)$ that you can obtain by drawing n lines in the plane $\mathbb{R}^2$?

We'll explore this by hand first. Clearly

$$R(0) = 1, \qquad R(1) = 2, \qquad R(2) = 4.$$

This is a good point to try guessing (a) the next number $R(3)$ in the sequence and (b) a formula for the general term of the sequence. Popular choices are $R(3) = 8$ and $R(n) = 2(n+1)$ or $R(n) = 2^n$. This is actually incorrect, as paper-and-pen will convince you that

$$R(3) = 7$$

As you continue exploring this by hand, you may notice that, given an existing configuration of lines, a new line creates k new regions if and only if it crosses $k-1$ old lines. So if we want to maximise k then we need to maximise the number of old lines we cross. This leads to the recursion

$$\begin{aligned} R(0) &= 1 \\ R(n) &= R(n-1) + n \qquad \text{if } n > 1 \end{aligned}$$

We can implement this in Sage as

```
sage: def R(n):
....:     if n == 0:
....:         return 1
....:     return R(n-1) + n
```

Let's check the first few terms:

```
sage: R(1)
2
sage: [R(n) for n in range(4)]
[1, 2, 4, 7]
```

This matches what we already knew. And now, into the great unknown:

```
sage: [R(n) for n in range(20)]
[1, 2, 4, 7, 11, 16, 22, 29, 37, 46, 56, 67, 79, 92, 106, 121, 137, 154, 172, 191]
```

In Mathematica, we could do something like

```
In[1]:= Clear[R]
In[2]:= R[0] = 1;
In[3]:= R[n_] := R[n-1] + n
In[4]:= Table[R[n], {n, 0, 19}]

Out[4]= {1, 2, 4, 7, 11, 16, 22, 29, 37, 46, 56, 67, 79, 92, 106, 121, 137,
>    154, 172, 191}
```

Very satisfying. How about a formula for $R(n)$ though? We could think about it and figure it out fairly quickly in this example, but let's instead search for the first few terms in the Online Encyclopedia of Integer Sequences (OEIS):

oeis.org

Feeding it $1, 2, 4, 7, 11, 16, 22, 29$, it gives a number of suggestions, the foremost of which is

> A000124 Central polygonal numbers (the Lazy Caterer's sequence): n(n+1)/2 + 1; or, maximal number of pieces formed when slicing a pancake with n cuts.

In fact, this is such a common workflow (work out the first few terms, consult OEIS), that you can do it entirely from Sage[1]:

```
sage: lst = [R(n) for n in range(8)]
sage: lst
[1, 2, 4, 7, 11, 16, 22, 29]
sage: oeis(lst)
0: A000124: Central polygonal numbers (the Lazy Caterer's sequence): n(n+1)/2 + 1;
    or, maximal number of pieces formed when slicing a pancake with n cuts.
1: A152947: a(n) = 1 + (n-2)*(n-1)/2.
2: A098574: a(n) = Sum_{k=0..floor(n/7)} C(n-5*k,2*k).
```

OEIS helpfully gives us a closed form for the number of regions:

$$R(n) = \frac{n(n+1)}{2} + 1$$

Of course!

$$\begin{aligned}
R(0) &= 1 \\
R(1) &= R(0) + 1 = 1 + 1 \\
R(2) &= R(1) + 2 = 1 + 1 + 2 \\
R(3) &= R(2) + 3 = 1 + 1 + 2 + 3 \\
R(n) &= R(n-1) + n = 1 + \sum_{k=1}^{n} k = 1 + \frac{n(n+1)}{2}
\end{aligned}$$

By the way, in case you forgot the sum of the first n positive integers, both Sage and Mathematica can help:

[1] Check out `oeis?` for more information.

```
sage: k, n = var("k, n")
sage: sum(k, k, 1, n)
1/2*n^2 + 1/2*n
sage: sum(k, k, 1, n).factor()
1/2*(n + 1)*n
```

```
In[1]:= Sum[k, {k, 1, n}]

        n (1 + n)
Out[1]= ---------
            2
```

Exercise 2.1. Go one dimension up and play with it: what is the maximal number of regions that can be obtained from n planes in $\mathbb{R}^3$?

As frivolous as the topic may seem (slicing a pancake with n cuts, indeed), it is an active area of research. Look up hyperplane arrangements on the web, or [Sta12, Section 3.11].

2.2 Pascal mod 2

What is the distribution of even and odd numbers in Pascal's triangle?

Let's start by generating part of Pascal's triangle. The glorious way to approach this is to code the recursive construction of the triangle, but we're after answers rather than glory so we'll just use the built-in `binomial` to get the binomial coefficients:

```
sage: lst = [binomial(m, n) for m in range(2^3) for n in range(2^3)]
sage: lst
[1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 2, 1, 0, 0, 0, 0, 0, 1, 3, 3,
   1, 0, 0, 0, 0, 1, 4, 6, 4, 1, 0, 0, 0, 1, 5, 10, 10, 5, 1, 0, 0, 1, 6, 15, 20,
   15, 6, 1, 0, 1, 7, 21, 35, 35, 21, 7, 1]
sage: matrix(2^3, lst)
[ 1  0  0  0  0  0  0  0]
[ 1  1  0  0  0  0  0  0]
[ 1  2  1  0  0  0  0  0]
[ 1  3  3  1  0  0  0  0]
[ 1  4  6  4  1  0  0  0]
[ 1  5 10 10  5  1  0  0]
[ 1  6 15 20 15  6  1  0]
[ 1  7 21 35 35 21  7  1]
```

Looks promising. Let's reduce modulo 2:

```
sage: lst = [binomial(m, n) % 2 for m in range(2^3) for n in range(2^3)]
sage: matrix(2^3, lst)
[1 0 0 0 0 0 0 0]
[1 1 0 0 0 0 0 0]
[1 0 1 0 0 0 0 0]
[1 1 1 1 0 0 0 0]
[1 0 0 0 1 0 0 0]
[1 1 0 0 1 1 0 0]
[1 0 1 0 1 0 1 0]
[1 1 1 1 1 1 1 1]
```

We can maybe see a pattern, but a bigger version might help. And dots are more easily visualised than 0s and 1s. So let's plot a dot whenever the binomial coefficient is odd:

```
sage: lst = [(m, n) for m in range(2^8) for n in range(2^8) \
....:        if (binomial(m, n) % 2) == 1]
sage: p = list_plot(lst)
sage: p.show()
None
```

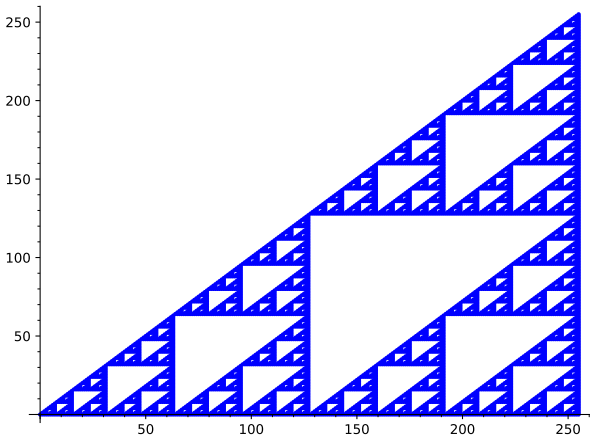


Figure 2.1: Pascal's triangle modulo 2

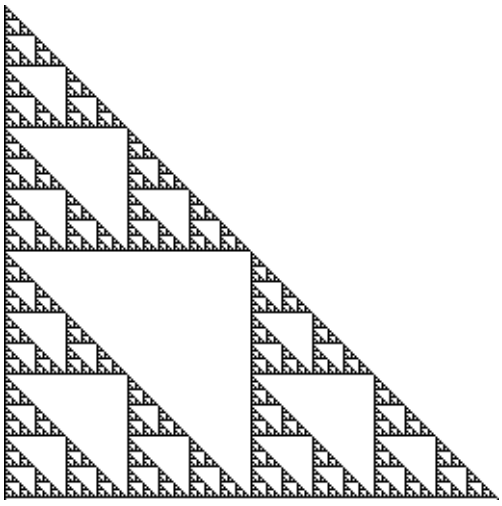
For a proof of what you are observing visually (i.e. that the picture looks a lot like Sierpiński's gasket), see [Ste95, Encounter 2].

The Mathematica version:

```
In[1]:= Table[Binomial[n, m], {n, 0, 2^3}, {m, 0, 2^3}] // MatrixForm

Out[1]//MatrixForm= 1    0    0    0    0    0    0    0    0

                    1    1    0    0    0    0    0    0    0

                    1    2    1    0    0    0    0    0    0

                    1    3    3    1    0    0    0    0    0

                    1    4    6    4    1    0    0    0    0

                    1    5    10   10   5    1    0    0    0

                    1    6    15   20   15   6    1    0    0

                    1    7    21   35   35   21   7    1    0

                    1    8    28   56   70   56   28   8    1
```

```
In[2]:= Table[Mod[Binomial[n, m], 2], {n, 0, 2^3}, {m, 0, 2^3}] // MatrixForm

Out[2]//MatrixForm= 1   0   0   0   0   0   0   0   0

                    1   1   0   0   0   0   0   0   0

                    1   0   1   0   0   0   0   0   0

                    1   1   1   1   0   0   0   0   0

                    1   0   0   0   1   0   0   0   0

                    1   1   0   0   1   1   0   0   0

                    1   0   1   0   1   0   1   0   0

                    1   1   1   1   1   1   1   1   0

                    1   0   0   0   0   0   0   0   1
```

```
In[3]:= Table[Mod[Binomial[n, m], 2], {n, 0, 2^8}, {m, 0, 2^8}] // Image // ColorNegate
```



Exercise 2.2. Look up how to plot a matrix in Sage. Use this to produce a visualisation of Pascal's triangle mod 2.

Exercise 2.3. Visualise the reduction of Pascal's triangle modulo other integers. Start with 3, 4, 5, ... Maybe give different colours to the different remainders.

3 Constant recognition

A common occurrence in computer-assisted mathematics is obtaining a numerical approximation to a number we are interested in. How can we recognise whether this number is of a special type (e.g. rational, or algebraic, or a simple combination of other numbers we like such as π, e, $M(\sqrt{2},1)$)? There are surprisingly robust ways of approaching such an ill-defined question, as we'll see now.

3.1 Finding the fraction

The Riemann zeta-function is a function of a complex variable s defined, for $\mathrm{Re}(s) > 1$, by the infinite series

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

Despite having Riemann's (1826–1866) name attached to it, particular aspects had been considered two hundred years before Riemann. As one such example, Mengoli asked in 1650 for the value

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} = \dots$$

What Mengoli meant by "value" was not an approximation such as

```
sage: RR(zeta(2))                                                    93
1.64493406684823                                                     94
```

which he could compute himself (maybe to slightly fewer decimals). It was the exact value, in closed form, a notion that is more metaphysical than mathematical. He was basically saying "I want a simple and pretty answer involving terms that I know already, and a proof that the answer is correct." This became known as the Basel problem and Euler knocked it out of the ballpark in 1741.

We're going to leave the historical trail and imagine for a moment that, by some stroke of genius, we thought it would be a good idea to divide $\zeta(2)$ by π^2 (why not?)

```
sage: RR(zeta(2)/pi^2)                                               95
0.166666666666667                                                    96
```

That is very compelling. Is it an artifact of the low precision?

```
sage: Rbig = RealField(1000)                                         97
sage: Rbig(zeta(2)/pi^2) == Rbig(1/6)                                98
True                                                                 99
```

Indeed, as Euler proved, it is the case that

$$\frac{\zeta(2)}{\pi^2} = \frac{1}{6}$$

Let's try another one. What could this be:

```
sage: RR(zeta(4))                                                    100
```

```
1.08232323371114
```

Maybe more divine revelation can help:

```
sage: RR(zeta(4)/pi^4)
0.0111111111111111
```

Yes,

$$\frac{\zeta(4)}{\pi^4} = \frac{1}{90}$$

This is easy! I can't believe there's a whole chapter devoted to this. Okay, one more:

```
sage: RR(zeta(12)/pi^12)
1.08220214040320e-6
```

Huh? Surely we need more decimals:

```
sage: Rbig = RealField(250)
sage: a = Rbig(zeta(12)/pi^12)
sage: a
1.0822021404031986042568053150063732074314084896095478106060116642127224138e-6
```

There's a certain gadget called a *continued fraction*, which is tailor-made for our problem[1]:

```
sage: c = continued_fraction(a)
sage: c
[0; 924041, 1, 3, 1, 2, 2, 1, 14]
```

What this really means is

$$\frac{\zeta(12)}{\pi^{12}} = 0 + \cfrac{1}{924041 + \cfrac{1}{1 + \cfrac{1}{3 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{1 + \frac{1}{14}}}}}}}} = \frac{691}{638512875}$$

More generally, a continued fraction expansion of a positive real number β is an expression of the form

$$\beta = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \frac{1}{\cdots}}}}$$

where $a_0 \in \mathbb{Z}_{\geqslant 0}, a_1, a_2, \cdots \in \mathbb{Z}_{>0}$.

Given β, the continued fraction expansion is easily computed by the recursion

$$\begin{aligned} \beta_0 &= \beta \\ a_n &= \lfloor \beta_n \rfloor \qquad \text{for } n \geqslant 0 \\ \beta_n &= \frac{1}{\beta_{n-1} - a_{n-1}} \qquad \text{for } n \geqslant 1 \end{aligned}$$

[1] And it's been around forever. According to Wikipedia, the first documented use of continued fractions is in Sanskrit and goes back to 499.

The truncation at a_n of a continued fraction is called its *n-th convergent*

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\cdots + \cfrac{1}{a_n}}}} = \frac{p_n}{q_n}$$

Exercise 3.1. If you harbour any nostalgia for the good ol' Real Analysis days, try your hand at this: prove that the sequence (p_n/q_n) converges as $n \to \infty$. One possible approach is to first show that

$$p_{n-1}q_n - p_n q_{n-1} = (-1)^n$$

then conclude that the sequence (p_n/q_n) is Cauchy.

For our constant-recognition purposes, some relevant facts are

- β is rational if and only if its continued fraction expansion is finite. This is closely related to the Euclidean algorithm for computing the greatest common denominator of two integers.
- β is quadratic irrational if and only if its continued fraction expansion is (eventually) periodic. (A quadratic irrational is a number of the form $a + b\sqrt{d}$ with $a, b \in \mathbb{Q}$ and d a squarefree positive integer.)
- Given a denominator upper bound Q, the convergent p_n/q_n with the largest n such that $q_n \leqslant Q$ is the *best rational approximation to β with denominator at most Q*, in other words it is the fraction p/q that minimises the quantity

$$|q\beta - p| \qquad \text{subject to } 1 \leqslant q \leqslant Q.$$

Some well-known constants have regular-looking continued fractions:

$$\frac{1+\sqrt{5}}{2} = \phi = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{\cdots}}}}}}}}}$$

$$e = 2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{4 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{6 + \cfrac{1}{1 + \cfrac{1}{\cdots}}}}}}}}}$$

while others do not:

$$\pi = 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{292 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{\ddots}}}}}}}}}}$$

3.2 Finding the integer relation

Let's take some of the discussion in the previous section and twist it around a little bit. Suppose I consider the real numbers $\zeta(2)$ and π^2 and I ask: are there integers a and b such that

$$a\,\zeta(2) + b\,\pi^2 = 0? \tag{3.1}$$

With the 20/20 hindsight afforded by the previous section, we can confidently answer: Yes, take $a = 6$ and $b = -1$. But the point is that in Equation (3.1) we organised this in the form of a linear relation with integer coefficients between the two numbers $\zeta(2)$ and π^2. And one can consider linear relations involving more than two numbers.

More precisely, given a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, we define an *integer relation for* $\mathbf{x}$ to be a nonzero vector $\mathbf{m} = (m_1, m_2, \ldots, m_n) \in \mathbb{Z}^n$ with integer entries such that

$$\mathbf{m} \cdot \mathbf{x} = m_1 x_1 + m_2 x_2 + \cdots + m_n x_n = 0.$$

Of course, there is no guarantee that such a magical $\mathbf{m}$ exists. (Take, for instance, $\mathbf{x} = (1, \sqrt{2})$.) This leads us to state the

Integer relation problem: Given $\mathbf{x} \in \mathbb{R}^n$, either find a "small" integer relation $\mathbf{m}$ for $\mathbf{x}$ or prove that no such "small" integer relation exists.

You may find the rather imprecise use of the adjective "small" distasteful. In that case, there is a more assertive version of the problem that fixes a bound 2^k and asks for $\mathbf{m} \in \mathbb{Z}^n$ with $\|\mathbf{m}\| \leqslant 2^{n+k}$ or a proof that there is no $\mathbf{m} \in \mathbb{Z}^n$ with $\|\mathbf{m}\| < 2^k$.

There is yet another version that is most attractive in practice, where we take the input vector $\mathbf{x} \in \mathbb{Z}^n$ as well. The following example shows how this might come about.

Example 3.2. Let

$$\begin{aligned} x_1 &= \arctan(1) = 0.785398\ldots \\ x_2 &= \arctan(1/5) = 0.197395\ldots \\ x_3 &= \arctan(1/239) = 0.004184\ldots \end{aligned}$$

Can we find an integer relation $\mathbf{m}$ for $\mathbf{x} = (x_1, x_2, x_3)$:

$$m_1 x_1 + m_2 x_2 + m_3 x_3 = 0?$$

We turn this into a question about integers by fixing a multiplier A, say $A = 10^6$, and considering

$$m_1 \lfloor A x_1 \rfloor + m_2 \lfloor A x_2 \rfloor + m_3 \lfloor A x_3 \rfloor \approx 0,$$

that is

$$785398m_1 + 197395m_2 + 4184m_3 \approx 0.$$

There are several approaches to finding m_1, m_2, m_3 and we will be looking in more detail at one of them, the LLL algorithm. For now I will just say that this algorithm takes the matrix

$$M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 785398 & 197395 & 4184 \end{bmatrix}$$

and returns the matrix

$$\begin{bmatrix} 1 & -13 & -52 \\ -4 & 58 & 203 \\ 1 & -296 & 184 \\ 2 & 272 & 345 \end{bmatrix}$$

The first column of this matrix will be telling us that

$$785398 \cdot 1 + 197395 \cdot (-4) + 4184 \cdot 1 = 2,$$

or, going back to our original real numbers, that

$$0.785398 \cdot 1 + 0.197395 \cdot (-4) + 0.004184 \cdot 1 = \frac{2}{10^6} \approx 0.$$

In other words, $\mathbf{m} = (1, -4, 1)$ seems to be an integer relation.

Indeed, Machin found this formula in 1706:

$$\arctan(1) = 4\arctan(1/5) - \arctan(1/239).$$

It works remarkably well as a rapidly convergent approximation to $\pi = 4\arctan(1)$.

Obvious questions remain. What is this LLL algorithm? What exactly is it doing to that matrix? And how come the first column of the resulting matrix gives us the integer relation we have been looking for?

A quick answer is that LLL is a *lattice reduction algorithm*, and that the integer relation problem can be solved via lattice reduction. To make sense of this, we need to know something about lattices.

3.3 Lattice reduction

Fix a natural number n and let V be an n-dimensional real vector space endowed with an inner product $\mathbf{u} \cdot \mathbf{v}$. (You may think of V as being $\mathbb{R}^n$ with the usual dot product.)

A *lattice* in V is a subset $L \subset V$ such that there exists a basis $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ of V with

$$L = \mathrm{Span}_{\mathbb{Z}}\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\} = \{r_1\mathbf{b}_1 + r_2\mathbf{b}_2 + \cdots + r_n\mathbf{b}_n \mid r_1, r_2, \ldots, r_n \in \mathbb{Z}\}.$$

We say that $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ is a $\mathbb{Z}$-*basis* for the lattice L, and we call n the *rank* of L. The *determinant* $d(L)$ of L is the determinant of the matrix with columns $\mathbf{b}_1, \ldots, \mathbf{b}_n$. This turns out to be independent of the choice of basis.

It would be good to recall the setup of Gram–Schmidt orthogonalisation. Let $\mathbf{b}_1, \ldots, \mathbf{b}_n$ be a basis for V. For $i = 1, \ldots, n$ let $V_i = \mathrm{Span}_{\mathbb{R}}\{\mathbf{b}_1, \ldots, \mathbf{b}_i\}$.

The Gram–Schmidt process returns vectors $\mathbf{b}_1^*, \dots, \mathbf{b}_n^* \in V$ and scalars $\mu_{ij} \in \mathbb{R}$ for $1 \leqslant j < i \leqslant n$, defined inductively by

$$\mathbf{b}_i^* = \operatorname{proj}_{V_{i-1}^\perp}(\mathbf{b}_i) = \mathbf{b}_i - \operatorname{proj}_{V_{i-1}}(\mathbf{b}_i) = \mathbf{b}_i - \sum_{j=1}^{i-1} \mu_{ij} \mathbf{b}_j^*$$

$$\mu_{ij} = \frac{\mathbf{b}_i \cdot \mathbf{b}_j^*}{\mathbf{b}_j^* \cdot \mathbf{b}_j^*}$$

with $\mathbf{b}_1^* = \mathbf{b}_1$.

Note that, for all $i = 2, \dots, n$,

$$V_{i-1} = \operatorname{Span}_{\mathbb{R}}\{\mathbf{b}_1, \dots, \mathbf{b}_{i-1}\} = \operatorname{Span}_{\mathbb{R}}\{\mathbf{b}_1^*, \dots, \mathbf{b}_{i-1}^*\}$$

and $\mathbf{b}_1^*, \dots, \mathbf{b}_n^*$ is an orthogonal basis of V.

Let's go back to the setup of a lattice $L \subset V$. We say that a basis $\mathbf{c}_1, \dots, \mathbf{c}_n$ for L is *(LLL-)reduced* if

$$|\mu_{ij}| \leqslant \frac{1}{2} \qquad \text{for all } 1 \leqslant j < i \leqslant n$$

$$\|\mathbf{c}_i^* + \mu_{i,i-1}\mathbf{c}_{i-1}^*\|^2 \geqslant \frac{3}{4}\|\mathbf{c}_{i-1}^*\|^2 \qquad \text{for all } 1 < i \leqslant n.$$

Here $3/4$ can be replaced by anything in the open interval $(1/4, 1)$.

The advantage of a reduced basis is that its vectors are in some sense small:

Proposition 3.3. *If $\{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ is a reduced basis of a lattice L, then*

$$\|\mathbf{c}_j\|^2 \leqslant 2^{i-1}\|\mathbf{c}_i^*\|^2 \qquad \textit{for all } 1 \leqslant j \leqslant i \leqslant n$$

$$d(L) \leqslant \prod_i \|\mathbf{c}_i\| \leqslant 2^{n(n-1)/4} d(L)$$

$$\|\mathbf{c}_1\| \leqslant 2^{(n-1)/4} d(L)^{1/n}$$

$$\|\mathbf{c}_1\|^2 \leqslant 2^{n-1}\|\mathbf{x}\|^2 \qquad \textit{for all } \mathbf{x} \in L - \{0\}$$

It is worth dwelling a little on this last inequality. If the constant (once L is fixed) multiplier 2^{n-1} were not there, we would have that $\mathbf{c}_1$ is a shortest nonzero vector in L. This may seem like a desirable outcome (and indeed many problems rely on finding a shortest vector), but it has the great disadvantage that its time complexity is exponential in the dimension n. A reduced basis gives up some control on the size of $\mathbf{c}_1$, to the extent shown in the last inequality of the Proposition. What is gained however is that this can be computed in time polynomial in the dimension n.

The latter is achieved by the LLL reduction algorithm. Its name is derived from its authors, Arjen Lenstra, Hendrik Lenstra, and László Lovász. The algorithm starts with some given basis of L and iteratively modifies it to achieve a reduced one. It is not particularly complicated, and the exposition in the original paper [LLL82] is quite good.

Example 3.4. Going back to the situation of Example 3.2, the lattice is $\mathbb{Z}$-spanned by the columns of the matrix

$$M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 785398 & 197395 & 4184 \end{bmatrix}$$

The norms of the three basis vectors are roughly

$$785398.0, 197395.0, 4184.0$$

The LLL algorithm returns the matrix

$$\begin{bmatrix} 1 & -13 & -52 \\ -4 & 58 & 203 \\ 1 & -296 & 184 \\ 2 & 272 & 345 \end{bmatrix}$$

whose columns are the vectors in the reduced basis, with norms roughly

$$4.7, 406.4, 443.6$$

Remember that Proposition 3.3 only guarantees that the first basis vector is within a factor of $\sqrt{2^{3-1}} = 2$ of the shortest nonzero vector. In this example however, it can be checked that the first basis vector is actually a shortest vector. This does happen sometimes, an instance of the fact that the worst-case performance and the average-case performance of an algorithm can be quite different.

How does lattice reduction help with finding integer relations? (It may be good to follow along with Example 3.2.) Given $x_1, \dots, x_n \in \mathbb{R}$, let $x'_1, \dots, x'_n \in \mathbb{R}$ be close approximations (such as, for instance, truncating after a certain number of digits). Let A be a multiplier. Consider the $\mathbb{Z}$-linear map $\mathbb{Z}^n \to \mathbb{R}^{n+1}$ given by

$$\begin{bmatrix} m_1 \\ \vdots \\ m_n \end{bmatrix} \mapsto \begin{bmatrix} m_1 \\ \vdots \\ m_n \\ A\sum_i m_i x'_i \end{bmatrix}$$

Let L denote the image of this map; it is a lattice (of rank n) inside an n-dimensional subspace V of $\mathbb{R}^{n+1}$.

In other words, we consider the matrix representation of the $\mathbb{Z}$-linear map above

$$M = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ Ax'_1 & Ax'_2 & \dots & Ax'_n \end{bmatrix}$$

Letting $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ denote the columns of this matrix, we have

$$L = \mathrm{Span}_{\mathbb{Z}}\{\mathbf{b}_1, \dots, \mathbf{b}_n\} \subset V = \mathrm{Span}_{\mathbb{R}}\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$$

The crucial observation is that

$$\begin{bmatrix} m_1 \\ \vdots \\ m_n \\ \varepsilon \end{bmatrix} \in L \qquad \text{if and only if} \qquad m_1 x'_1 + \dots + m_n x'_n = \frac{\varepsilon}{A}.$$

If ε is relatively small then (since $x'_i \approx x_i$)

$$m_1 x_1 + \dots + m_n x_n \approx 0.$$

So we are interested in finding elements of L that have small components. But, as we have seen above, the first vector in a reduced basis for L will be relatively small. So we apply LLL to the basis $\mathbf{b}_1, \dots, \mathbf{b}_n$ of L, obtain a reduced basis $\mathbf{c}_1, \dots, \mathbf{c}_n$, and take the relation determined by the first vector $\mathbf{c}_1$.

4 Polynomial systems and Gröbner bases

From linear algebra, we know how to solve a system of linear equations by doing Gaussian elimination (aka row reduction). But what can we do if we are faced with a system of nonlinear polynomial equations such as

$$\begin{aligned} x^2 + y^2 - 1 &= 0 \\ x^2 - y &= 0 \end{aligned}$$

Okay, that's not a very good example since we recognise the first equation as describing the unit circle and the second as a parabola and it is easy to do a substitution and solve for the two intersection points.

Exercise 4.1. Do this. (Find the real solutions (x, y) of this system.)
Does anything change if I ask for complex solutions?
Optional: how about solutions in a finite field $\mathbb{F}_p$?

Such adhoc methods will not always be available, for instance given

$$\begin{aligned} 6x^2y - x + 4y^3 - 1 &= 0 \\ 2xy + y^3 &= 0 \end{aligned}$$

There is a systematic approach to simplifying such systems with a view towards understanding their solutions. This approach is based on the notion of Gröbner basis, for which we need to describe/review a bit of algebra.

4.1 Polynomial rings and ideals

Just as vector spaces are the natural environment for reasoning about systems of linear equations, (multivariate) polynomial rings are the natural environment for systems of polynomial equations.

The first ingredient is a field of coefficients k. For the purposes of our discussion you can imagine this to be whatever field you prefer, e.g. $\mathbb{Q}$ or $\mathbb{R}$ or $\mathbb{C}$. Then we form *polynomial rings* like

$$k[x_1, \ldots, x_n] = \left\{ \sum c_{a_1,\ldots,a_n} x_1^{a_1} \ldots x_n^{a_n} \mid a_1, \ldots, a_n \in \mathbb{Z}_{\geqslant 0}, c_{a_1,\ldots,a_n} \in k \right\},$$

where the sums have only finitely many terms.

We will refer to the elements of $k[x_1, \ldots, x_n]$ as *polynomials* and perform the arithmetic operations of addition and multiplication in the usual way. We might sometimes call them *formal* polynomials as opposed to the polynomial *functions* that you encounter in calculus (broadly construed). The link between the two is given by the operation of *evaluation*. For any $b = (b_1, \ldots, b_n) \in k^n$ we have $\mathrm{ev}_b \colon k[x_1, \ldots, x_n] \to k$ given by

$$\mathrm{ev}_b(f) = f(b_1, \ldots, b_n)$$

where we substitute the field element b_1 for the formal variable x_1, b_2 for x_2, etc. This allows us to think of $f \in k[x_1, \ldots, x_n]$ as (giving rise to) a function $F\colon k^n \to k$ defined by

$$F(b_1, \ldots, b_n) = \mathrm{ev}_b(f).$$

While it is important to have an awareness of the distinction between polynomial and polynomial function, we'll now join the rest of the planet in abusing notation and writing $f(b_1, \ldots, b_n)$ when we mean $F(b_1, \ldots, b_n)$.

We can think of the solutions of the system of equations given above as

$$V(f_1, f_2) = \left\{(b_1, b_2) \in \mathbb{R}^2 \mid f_1(b_1, b_2) = 0, f_2(b_1, b_2) = 0\right\}$$

where $f_1(x, y) = 6x^2y - x + 4y^3 - 1 \in \mathbb{R}[x, y]$, $f_2(x, y) = 2xy + y^3 = 0 \in \mathbb{R}[x, y]$, and V stands for vanishing set.

We can define similarly the vanishing set of any subset of polynomials $A \subset k[x_1, \ldots, x_n]$:

$$V(A) = \{(b_1, \ldots, b_n) \in k^n \mid f(b_1, \ldots, b_n) = 0 \text{ for all } f \in A\}.$$

And we can turn this idea on its head and define, for any subset $B \subset k^n$, the set of polynomials that vanish identically on B:

$$I(B) = \{f \in k[x_1, \ldots, x_n] \mid f(b_1, \ldots, b_n) = 0 \text{ for all } (b_1, \ldots, b_n) \in B\}.$$

The interplay between these two functions V and I is the starting point of algebraic geometry and a beautiful piece of mathematics.

But I digress. The subset $I(B) \subset k[x_1, \ldots, x_n]$ exhibits some interesting extra structure:

- if f_1 and f_2 are in $I(B)$, then $f_1 + f_2 \in I(B)$
- if f is in $I(B)$ and r is in $k[x_1, \ldots, x_n]$, then $rf \in I(B)$

A nonempty subset of $k[x_1, \ldots, x_n]$ with these properties is called an *ideal*.

A great way of getting lots of ideals is choosing some elements $f_1, \ldots, f_m \in k[x_1, \ldots, x_n]$ and combining them as follows:

$$\langle f_1, \ldots, f_m \rangle = \{r_1 f_1 + \ldots r_m f_m \mid r_1, \ldots, r_m \in k[x_1, \ldots, x_n]\}.$$

(Compare this to the span of vectors $f_1, \ldots, f_m$. It is similar in shape but definitely not the same thing.)

Exercise 4.2. Prove that $\langle f_1, \ldots, f_m \rangle$ is indeed an ideal of $k[x_1, \ldots, x_n]$. It is called the *ideal generated by* $f_1, \ldots, f_m$.

Going back to a system such as

$$\begin{aligned} 6x^2y - x + 4y^3 - 1 &= 0 \\ 2xy + y^3 &= 0 \end{aligned}$$

our strategy will be to form the ideal $I = \langle f_1, f_2 \rangle$ of $\mathbb{R}[x, y]$ with $f_1 = 6x^2y - x + 4y^3 - 1$ and $f_2 = 2xy + y^3$ and try to gain a better understanding of this ideal. More precisely, we will aim to find a simpler set of generators for this ideal, one that will illuminate the solutions of the system.

To get a feel for this in more familiar settings, we consider two very special cases of the problem.

4.2 Linear systems

We go back to the linear case momentarily, because we know how to deal with it. Consider the ideal $I = \langle x + y - z, 2x + 3y + 2z \rangle$ in $\mathbb{R}[x, y, z]$. Its vanishing set is the set of solutions of the linear system

$$\begin{aligned} x + y - z &= 0 \\ 2x + 3y + 2z &= 0 \end{aligned}$$

In order to solve this system, we apply Gaussian elimination to the associated matrix

$$\begin{bmatrix} 1 & 1 & -1 \\ 2 & 3 & 2 \end{bmatrix} \to \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & 4 \end{bmatrix}$$

which tells us that $I = \langle x + y - z, y + 4z \rangle$, an arguably simpler set of generators than the one we started with. We could even go all in with Gauss–Jordan elimination to get the reduced row-echelon form

$$\begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & 4 \end{bmatrix}$$

telling us that $I = \langle x - 5z, y + 4z \rangle$.

From these generators we see that we can express both x and y in terms of z, so the solution set is one-dimensional and we can write down an explicit parametrisation in terms of z.

So in this special case we used elementary row operations in order to simplify the generators of I. Note that we implicitly ordered our variables so that x comes before y comes before z.

Had we decided to order them differently, say z before y before x, the same method applies but we would get the differently-looking answer $I = \langle -z + y + x, 2z + 3y + 2x \rangle = \langle z - x/5, y + 4x/5 \rangle$. This is a feature of the whole multivariate polyonomial business, and we'll have to think about it more carefully soon.

4.3 Single variable higher degree case

The other special situation we consider is that of higher-degree polynomials in a single variable x. In other words, we now work with arbitrary elements of $k[x]$.

Let's take the ideal $I = \langle x^3 - 2x^2 + 2x + 8, 2x^2 + 3x + 1 \rangle$. It's not clear what the vanishing set of I might be, so we try to find a simpler set of generators.

We can do this by polynomial long division:

$$x^3 - 2x^2 + 2x + 8 = \left(\frac{1}{2}x - \frac{7}{4}\right)\left(2x^2 + 3x + 1\right) + \left(\frac{27}{4}x + \frac{39}{4}\right),$$

from which we conclude that $I = \left\langle 2x^2 + 3x + 1, \frac{27}{4}x + \frac{39}{4} \right\rangle$. This is simpler than before (the generators have degrees 2 and 1 instead of the original 3 and 2), but still not as simple as could be[1].

We repeat the process:

$$2x^2 + 3x + 1 = \left(\frac{8}{27}x + \frac{4}{243}\right)\left(\frac{27}{4}x + \frac{39}{4}\right) + \frac{68}{81},$$

so $I = \left\langle \frac{27}{4}x + \frac{39}{4}, \frac{68}{81} \right\rangle$. We could do another long division of the first generator by the second, but that's overkill since we already have the condition $\frac{68}{81} = 0$, which indicates that the system has no solutions.

[1] You can however use the degree 1 generator to solve for x and then plug the solution into the other generator to reach the same conclusion as in the next paragraph.

Note that there is no ambiguity of ordering of the variables here, as there is a single variable.

The Euclidean algorithm for $k[x]$ (which is what this process of repeated long division is called) shows that every ideal in $k[x]$ can be generated by a single element. This ceases to be the case for the general situation $k[x_1, \ldots, x_n]$ (can you think of an ideal in there that cannot be generated by a single element), but something useful is still true:

Theorem 4.3 (Hilbert Basis Theorem)**.** *Every ideal in $k[x_1, \ldots, x_n]$ can be generated by a finite set of elements.*

Having considered the two special cases of this section and the previous one, we turn our attention to the general case of an ideal (or system of equations) in $k[x_1, \ldots, x_n]$. Before we investigate the common generalisation of Gaussian elimination and of single-variable polynomial long division, let's note that many questions about the solution set of a system can be expressed in terms of the associated ideal.

For instance, if k is an algebraically closed field (e.g. $\mathbb{C}$) and I is the ideal associated with a system of equations in $k[x_1, \ldots, x_n]$, then the system has no solutions if and only if I contains the constant 1. (As we observed in the example of this section, one direction of this claim is very easy and holds without any conditions on the field k.)

For the remainder of this chapter, we will focus on the ideals themselves rather than the systems of equations that they may have originated from.

4.4 Monomial orders

Polynomial long division in one variable x makes crucial use that people tend to order monomials $1, x, x^2, x^3, \ldots$ by degreeso that the leading term of $3x^4 + x^3 - x^2 + 1$ is $3x^4$ and that of $x^2 - x + 2$ is x^2. So one step of the division algorithm simply divides one leading term $3x^4$ by the other x^2.

How should we proceed, however, in order to divide $x^2y^2z^2 + xy^4z - 1$ by $xy^2z - 2z + 3$? There does not seem to be any "natural" reason to prefer $x^2y^2z^2$ over xy^4z as the leading term of the first polynomial. So it is a matter of choice.

What we are choosing here is a so-called *monomial order*, that is an order relation $>$ on the set of monomials

$$\{x^a = x_1^{a_1} \ldots x_n^{a_n} \mid a_1, \ldots, a_n \in \mathbb{Z}_{\geqslant 0}\}$$

with the following nice properties:

(a) it is a total order relation: given any two monomials x^a and x^b with $a \neq b$, either $x^a > x^b$ or $x^b > x^a$;

(b) it is compatible with multiplication in $k[x_1, \ldots, x_n]$: if $x^a > x^b$ and x^c is any monomial, then $x^{a+c} > x^{b+c}$;

(c) it is a well-ordering: every nonempty set of monomials has a smallest element under $>$.

The ring of polynomials in one variable x has a unique monomial order, namely

$$\cdots > x^n > x^{n-1} > \cdots > x^2 > x > 1$$

that we already alluded to.

If there is more than one variable, then there are many monomial orders, even after we choose an order on the variables themselves, which we do now once and for all:

$$x_1 > x_2 > \cdots > x_n.$$

Here are some popular monomial orders, together with their effect on the monomials $\{x_1x_2^2x_3, x_3^2, x_1^3, x_1^2x_3^2\}$:

- *Lexicographic order*: we say $x^a >_{\text{lex}} x^b$ if the vector $a - b \in \mathbb{Z}^n$ has positive leftmost nonzero entry.

  This gives $x_1^3 > x_1^2x_3^2 > x_1x_2^2x_3 > x_3^2$.

- *Graded (or degree) lexicographic order*: we say $x^a >_{\text{grlex}} x^b$ if

$$\begin{aligned}&\text{either } \sum a_i > \sum b_i\\ &\text{or } \sum a_i = \sum b_i \text{ and } x^a >_{\text{lex}} x^b.\end{aligned}$$

  This gives $x_1^2x_3^2 > x_1x_2^2x_3 > x_1^3 > x_3^2$.

- *Graded reverse lexicographic order*: we say $x^a >_{\text{grevlex}} x^b$ if

$$\begin{aligned}&\text{either } \sum a_i > \sum b_i\\ &\text{or } \sum a_i = \sum b_i \text{ and the vector } a-b \text{ has negative rightmost nonzero entry.}\end{aligned}$$

  This gives $x_1x_2^2x_3 > x_1^2x_3^2 > x_1^3 > x_3^2$.

While lexicographic order is the most familiar (it's used in dictionaries, for instance), it is almost never the best choice in terms of working with ideals. Which of the other orders is best to use is a subtle question whose answer depends on the application one has in mind. We'll ignore such details and move on.

4.5 Multivariate polynomial division

Assume now that we have fixed a monomial order on $k[x_1, \dots, x_n]$.

Given a polynomial $g \in k[x_1, \dots, x_n]$, we write $LT(g)$ for the *leading term* of g, i.e. the largest term of g with respect to the monomial order $>$. For instance, with lex order we have

$$LT(3x_1^3x_2^2 + x_1^2x_2x_3^3) = 3x_1^3x_2^2,$$

while with grevlex order:

$$LT(3x_1^3x_2^2 + x_1^2x_2x_3^3) = x_1^2x_2x_3^3.$$

Let $F = (f_1, \dots, f_s)$ be an s-tuple of polynomials in $k[x_1, \dots, x_n]$. Given $f \in k[x_1, \dots, x_n]$, there is an expression of the form

$$f = q_1f_1 + \cdots + q_sf_s + r, \qquad \text{with } q_i, r \in k[x_1, \dots, x_n]$$

where

- for each i, $q_if_i = 0$ or $LT(f) \geqslant LT(q_if_i)$;
- $r = 0$ or r is made of monomials that are not divisible by any of $LT(f_1), \dots, LT(f_s)$.

We refer to this as a *division of f by F* and to r as the *remainder* of this division.

Example 4.4. Consider $\mathbb{Q}[x_1, x_2, x_3]$ with lex order. Let $f = 3x_1^3x_2^2 + x_1^2x_2x_3^3$ and $F = (f_1, f_2)$ with

$$\begin{aligned}f_1 &= x_1^2x_2 + 2x_3\\ f_2 &= x_3^2\end{aligned}$$

Then the result of a division of f by F looks like

$$f = (3x_1x_2 + x_3^3)f_1 - (2x_3^2)f_2 - 6x_1x_2x_3.$$

4.6 Gröbner bases

In the one variable case, the ideal membership problem is solved by the division algorithm as follows: starting with an ideal $I = \langle f_1, \ldots, f_s \rangle$, apply the division algorithm repeatedly to find the gcd g of the f_i's, so that $I = \langle g \rangle$. Then given any f, divide f by g with remainder r. We have that $f \in I$ if and only if $r = 0$.

We would like to adapt this to the multivariate case, using the division algorithm discussed in the previous section. But it turns out that we cannot simply use division by $F = (f_1, \ldots, f_s)$, where $f_1, \ldots, f_s$ are arbitrary generators of the ideal I.

The extra property that we need to make this work is encompassed by the following definition: Fix a monomial order on $k[x_1, \ldots, x_n]$ and let I be an ideal. A *Gröbner basis for* I is a finite set $G = \{g_1, \ldots, g_t\} \subset I$ such that for every $f \in I$, $LT(f)$ is divisible by $LT(g_i)$ for some i.

Here are some useful (not necessarily obvious) facts about Gröbner bases:

(a) A Gröbner basis for I is a set of generators of I.

(b) For every ideal I and every choice of monomial order, there is a Gröbner basis. (For the zero ideal $\langle 0 \rangle$, by convention $\varnothing$ is a Gröbner basis.)

(c) There are algorithms (the first one of which was due to Buchberger) that start with a generating set $f_1, \ldots, f_s$ for the ideal I and produce a Gröbner basis for I.

(d) Even with a fixed choice of monomial order, there is not a *unique* Gröbner basis for each ideal I. It is possible to impose some extra conditions on the generators to arrive at the notion of a *monic Gröbner basis*. Then it is true that, a monomial order having been fixed, every ideal in $k[x_1, \ldots, x_n]$ has a unique monic Gröbner basis.

(e) If G is a Gröbner basis for I and $f \in I$, then the remainder of the division of f by G is zero. (This of course gives us a solution for the ideal membership problem in $k[x_1, \ldots, x_n]$.)

Example 4.5. Going back to Example 4.4, $\{f_1, f_2\}$ is a Gröbner basis for the ideal $I = \langle f_1, f_2 \rangle \subset \mathbb{Q}[x_1, x_2, x_3]$, and we can conclude that $f \notin I$ since the remainder $-6x_1x_2x_3 \neq 0$.

5 Hypergeometric summation machine

Did you know that for any $n \in \mathbb{N}$

$$\sum_k (-1)^k \binom{2n}{k}^3 = (-1)^n \frac{(3n)!}{(n!)^3} \quad ?$$

(The exclamation marks are of course factorials; the question mark is honest.)

In this chapter we'll not only see that this is true, but how to get your computer to prove this and many other identities that are too tedious to do by hand.

5.1 Hypergeometric stuff

We will consider *definite sums* of the form

$$F = \sum_{k=-\infty}^{\infty} t_k$$

especially when the general term t_k has *finite support*, i.e. it is zero for all but finitely many values of k. As a prototype for this, fix $n \in \mathbb{N}$ and think of

$$\sum_{k=-\infty}^{\infty} \binom{n}{k}$$

A *hypergeometric series* is a sum as above in which the ratio $\frac{t_{k+1}}{t_k}$ is a rational function of k, that is

$$\frac{t_{k+1}}{t_k} = \frac{P(k)}{Q(k)}$$

with P and Q both polynomials in k. We then refer to t_k as a *hypergeometric term*.

Example 5.1. If $t_k = \binom{n}{k}$ then

$$\frac{t_{k+1}}{t_k} = \frac{n-k}{k+1}$$

So the sum

$$\sum_{k=-\infty}^{\infty} \binom{n}{k}$$

is hypergeometric.

Every hypergeometric series has a representation in terms of the *generalised hypergeometric function*

$${}_pF_q\left[\begin{matrix}\alpha_1, & \ldots, & \alpha_p \\ \beta_1, & \ldots, & \beta_q\end{matrix}; x\right] = \sum_{k=0}^{\infty} A_k x^k = \sum_{k=0}^{\infty} \frac{(\alpha_1)_k \ldots (\alpha_p)_k}{(\beta_1)_k \ldots (\beta_q)_k} \frac{x^k}{k!},$$

where $(y)_k$ denotes the *Pochhammer symbol, aka the rising factorial*

$$(y)_k = y(y+1)\ldots(y+k-1)$$

We note that none of the lower parameters β_i are allowed to be negative integer or zero (otherwise we get a zero appearing in the denominator).

In the cases we'll consider the convergence of ${}_pF_q$ will not be an issue. We think of ${}_pF_q$ as a function of x, and most of the time it will be a polynomial.

Noting that

$$\frac{(y)_{k+1}}{(y)_k} = y + k,$$

we can investigate whether ${}_pF_q$ really is hypergeometric

$$\frac{A_{k+1}x^{k+1}}{A_k x^k} = \frac{(k+\alpha_1)\dots(k+\alpha_p)}{(k+\beta_1)\dots(k+\beta_q)}\,\frac{x}{(k+1)}$$

This is indeed a rational function of k (with some parameters thrown in).

Example 5.2 (Geometric series)**.** For $|x| < 1$ we have

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k = {}_1F_0\left[\begin{matrix}1\\-\end{matrix}; x\right]$$

Example 5.3 (Exponential)**.**

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = {}_0F_0\left[\begin{matrix}-\\-\end{matrix}; x\right]$$

Example 5.4. Life is not always so accommodating. Consider

$$\sum_{k=0}^{n} k\binom{n}{k}$$

which is certainly hypergeometric as the ratio of consecutive terms is

$$\frac{t_{k+1}}{t_k} = \frac{n-k}{k}$$

a rational function of k. But direct pattern matching would give us a lower parameter of 0, which is not allowed.

Here we need a shift: let $u_k = t_{k+1}$, then

$$\frac{u_{k+1}}{u_k} = \frac{n-k-1}{k+1} = \frac{k+1-n}{1}\,\frac{-1}{k+1},$$

so that, since $u_0 = t_1 = n$,

$$\sum_{k=0}^{n} t_k = \sum_{k+1=1}^{n+1} u_k = u_1 + u_2 + \dots + u_{n+1} = u_0\left(\frac{u_1}{u_0} + \frac{u_2}{u_0} + \dots + \frac{u_{n+1}}{u_0}\right) = n \cdot {}_1F_0\left[\begin{matrix}1-n\\-\end{matrix}; -1\right]$$

5.2 Celine Fasenmyer's algorithm for finding recurrence relations

Let's consider, for $n \in \mathbb{N}$,

$$s_n = \sum_{k=-\infty}^{\infty} f(n,k)$$

with

$$f(n,k) = k\binom{n}{k}$$

Our aim is to find a recurrence relation for s_n (as a function of n). The approach is to first find a k-free recurrence relation for the summand $f(n,k)$, that is a relation of the form

$$\sum_{i=0}^{I}\sum_{j=0}^{J} a_{ij}(n)\, f(n+j,k+i) = 0,$$

for suitable I and J and coefficients $a_{ij}(n)$ that are polynomials in n (but do not depend on k).

In our example, we'll try with $I = J = 1$ and we are looking for

$$a_{00}f(n,k) + a_{01}f(n+1,k) + a_{10}f(n,k+1) + a_{11}f(n+1,k+1) = 0,$$

or more precisely

$$a_{00}k\binom{n}{k} + a_{01}k\binom{n+1}{k} + a_{10}(k+1)\binom{n}{k+1} + a_{11}(k+1)\binom{n+1}{k+1} = 0.$$

Divide through by $k\binom{n}{k}$ to get something involving rational functions of k and n:

$$a_{00} + a_{01}\,\frac{n+1}{n+1-k} + a_{10}\,\frac{n-k}{k} + a_{11}\,\frac{n+1}{k} = 0$$

Get rid of the denominators by multiplying the whole thing by $k(n+1-k)$:

$$a_{00}k(n+1-k) + a_{01}k(n+1) + a_{10}(n-k)(n+1-k) + a_{11}(n+1)(n+1-k) = 0$$

Next rewrite the left hand side as a polynomial in k:

$$(-a_{00}+a_{10})k^2 + (a_{00}(n+1)+a_{01}(n+1)-a_{10}(2n+1)-a_{11}(n+1))k + (a_{10}n(n+1)+a_{11}(n+1)^2) = 0.$$

If we equate coefficients we get the system of linear equations

$$\begin{bmatrix} -1 & 0 & 1 & 0 \\ n+1 & n+1 & -2n-1 & -n-1 \\ 0 & 0 & n & n+1 \end{bmatrix} \begin{bmatrix} a_{00} \\ a_{01} \\ a_{10} \\ a_{11} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The space of solutions is one-dimensional with free parameter a_{11}. Taking $a_{11} = 1$ we have the solution

$$a_{00} = -\frac{n+1}{n} \qquad a_{01} = 0 \qquad a_{10} = -\frac{n+1}{n} \qquad a_{11} = 1,$$

which gives the k-free recurrence relation

$$-(n+1)f(n,k) - (n+1)f(n,k+1) + nf(n+1,k+1) = 0$$

It remains to turn this into a relation for s_n. Note that

$$s_n = \sum_{k=-\infty}^{\infty} f(n,k) = \sum_{k=-\infty}^{\infty} f(n,k+1)$$

so summing the k-free relation over k we get

$$-2(n+1)s_n + ns_{n+1} = 0$$

This was the original objective, but we can push this a little further to get a simple expression for s_n. We have $s_1 = 1$ and

$$s_{n+1} = 2\,\frac{n+1}{n}\,s_n$$

and by unrolling this recursion we conclude that

$$s_n = 2^{n-1}\, n$$

So

$$\sum_k k\binom{n}{k} = 2^{n-1}\, n.$$

In the example we looked at, we were able to find a k-free recurrence relation with $I = J = 1$. In general this may not be possible, i.e. the resulting homogeneous linear system may have no nonzero solutions. In that case, we increase I and/or J and try again.

But how do we know that suitably large I and J will work? The answer is that it is possible to write down explicit bounds for I and J that guarantee success, if $f(n,k)$ is a *proper hypergeometric term*:

(a) f has finite support

(b) f can be written in the form

$$f(n,k) = P(n,k)\,\frac{Q(n,k)}{R(n,k)}\,w^n\,z^k,$$

where P is a polynomial in n and k and Q and R are finite products of Γ-factors of the form $\Gamma(\alpha n + \beta k + \gamma)$.

5.3 Indefinite hypergeometric summation

Let's move temporarily from the discrete to the continuous. A basic task in calculus is, given a function $f(x)$, finding an antiderivative $F(x)$ such that

$$f(x) = F'(x)$$

If we can solve this *indefinite integral* problem, then we can trivially solve the *definite integral* problem using the fundamental theorem of calculus:

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

Our interest is in the discrete realm. Here the continuous (real) variable x is replaced by the discrete (integer) variable k, the function $f(x)$ is replaced by the sequence (a_k), the differentiation operation

$$F'(x) = \lim_{h\to 0} \frac{F(x+h) - f(x)}{h}$$

is replaced by the difference operation

$$\Delta A_k = A_{k+1} - A_k,$$

and the integration

$$\int_{-\infty}^{\infty} f(x)\,dx$$

is replaced by the summation

$$\sum_{k=-\infty}^{\infty} a_k$$

What is the equivalent of the fundamental theorem of calculus? Suppose we are given a sequence (a_k) and we can find an *antidifference* (A_k) for it, so that

$$a_k = \Delta A_k = A_{k+1} - A_k$$

Then the definite summation problem is a simple case of telescoping:

$$\sum_{k=b}^{c} a_k = (A_{c+1} - A_c) + (A_c - A_{c-1}) + \cdots + (A_{b+1} - A_b) = A_{c+1} - A_b$$

So how do we find an antidifference (A_k) of (a_k)? This is the problem of indefinite summation, and we will discuss an algorithmic solution for it in the hypergeometric case.

More precisely, we say that (a_k) is *Gosper-summable* if there exists an antidifference (A_k) that is a hypergeometric term. *Gosper's algorithm* is a procedure that decides whether (a_k) is Gosper-summable, and if yes, produces an antidifference hypergeometric term.

Here is an outline of this algorithm:

(a) Identify the ratio a_{k+1}/a_k as a rational function in k:

$$\frac{a_{k+1}}{a_k} = \frac{n(k)}{d(k)},$$

where n and d are polynomials in k.

(b) Using Lemma 5.5, find polynomials p, q, r in k with the property that

$$\frac{a_{k+1}}{a_k} = \frac{p(k+1)}{p(k)} \frac{q(k+1)}{r(k+1)}$$

and $\gcd(q(k), r(k+j)) = 1$ for all $j \in \mathbb{Z}_{\geqslant 0}$.

(c) Compute a degree bound N for the auxiliary polynomial f, using Lemma 5.6. If $N < 0$, stop because no antidifference hypergeometric term exists.

(d) Take a general polynomial of degree at most N in k

$$f(k) = c_0 + c_1 k + \cdots + c_N k^N,$$

plug it into the equation

$$p(k) = q(k+1)f(k) - r(k)f(k-1),$$

and solve the resulting system of linear equations in the unknowns c_j.

If the system has no solution, stop because no antidifference hypergeometric term exists.

(e) Calculate

$$A_k = \frac{r(k)}{p(k)} f(k-1) a_k, \tag{5.1}$$

the solution to the antidifference problem.

Lemma 5.5. *The functions $p(k)$, $q(k)$, and $r(k)$ can be chosen in such a way that*

$$\gcd(q(k), r(k+j)) = 1 \qquad \text{for all } j \in \mathbb{Z}_{\geqslant 0}.$$

We're not going to prove this, but here is the idea of how it works: we have already from step (a) an expression

$$\frac{a_{k+1}}{a_k} = \frac{n(k)}{d(k)}$$

So as a first approximation we can take $p(k) = 1$, $q(k) = n(k-1)$, and $r(k) = d(k-1)$. If the gcd condition is satisfied, we're done!

If not, there exists j such that

$$\gcd(q(k), r(k+j)) = g(k) \qquad \text{with } \deg g(k) > 0.$$

For such a j, we can fix our original choice by putting

$$p'(k) = p(k)g(k)g(k-1)\dots g(k-j+1), \quad q'(k) = \frac{q(k)}{g(k)}, \quad r'(k) = \frac{r(k)}{g(k-j)}$$

and we can easily check that

$$\frac{p'(k+1)}{p'(k)}\,\frac{q'(k+1)}{r'(k+1)} = \frac{p(k+1)}{p(k)}\,\frac{q(k+1)}{r(k+1)} = \frac{a_{k+1}}{a_k}$$

and the gcd condition holds for q' and r'.

Lemma 5.6. *The polynomial $f(k)$ has degree $\leqslant N$, where N is determined as follows.*

Define polynomials σ, δ by $\sigma(k) = q(k+1) + r(k)$ and $\delta(k) = q(k+1) - r(k)$. Let s be the degree of σ and d the degree of δ. If $s \leqslant d$, then

$$N = \deg p(k) - d.$$

Otherwise, let a be the coefficient of k^s in $\sigma(k)$ and let b be the coefficient of k^{s-1} in $\delta(k)$. If $(-2b/a) \notin \mathbb{Z}_{\geqslant 0}$ then

$$N = \deg p(k) - s + 1.$$

If $(-2b/a) \in \mathbb{Z}_{\geqslant 0}$ then

$$N = \max\left\{-\frac{2b}{a}, \deg p(k) - s + 1\right\}.$$

Note: In applying Lemma 5.6 we hay find ourselves in the situation where one of the polynomials is the constant zero. Here the degree convention is that

$$\deg(\text{nonzero constant}) = 0 \qquad \text{but} \qquad \deg(0) = -\infty.$$

Looking at Equation (5.1), we see that when (a_k) is Gosper-summable, then the hypergeometric term antidifference A_k is a rational function multiple of a_k:

$$A_k = R(k)\,a_k \qquad \text{where} \qquad R(k) = \frac{r(k)}{p(k)}\,f(k-1)$$

We call the rational function $R(k)$ the *rational certificate* of a_k, as it allows us to certify the result A_k without going through all the work of determining A_k in the first place:

$$A_k = R(k)\,a_k \qquad \text{if and only if} \qquad \frac{R(k)+1}{R(k+1)} = \frac{a_{k+1}}{a_k}$$

5.4 Definite hypergeometric summation via the Wilf–Zeilberger method

Gosper's algorithm for indefinite summation can be used in certain circumstances to evaluate definite summations.

Proposition 5.7. *Suppose $a(n,k)$ is a hypergeometric term with respect to both variables $n \in \mathbb{Z}_{\geqslant 0}$ and $k \in \mathbb{Z}$, Gosper-summable with respect to k, and with finite support (i.e. for any $n \in \mathbb{Z}_{\geqslant 0}$, $a(n,k) \neq 0$ for finitely many $k \in \mathbb{Z}$). Then*

$$\sum_k a(n,k) = 0$$

for all but finitely many $n \in \mathbb{Z}_{\geqslant 0}$.

More precisely, if $A(n,k) = R(n,k)\,a(n,k)$ is an antidifference of $a(n,k)$ with respect to k, then the above definite sum is zero for all $n \in \mathbb{Z}_{\geqslant 0}$ for which the denominator of $R(n,k)$ is not identically zero.

Proof. This is quite straightforward. If $n \in \mathbb{Z}_{\geqslant 0}$ is not a singular point of $R(n,k)$ then

$$a(n,k) = A(n,k+1) - A(n,k)$$

so after summing over k

$$\sum_k a(n,k) = \sum_k \left(A(n,k+1) - A(n,k)\right)$$

and the sum has only finitely many nonzero terms as $a(n,k)$ has finite support. But the right hand side is telescoping so we get zero. □

Example 5.8. The term

$$a(n,k) = (-1)^k \binom{n}{k}$$

is Gosper-summable with antidifference

$$A(n,k) = -\frac{k}{n}\,a(n,k)$$

So for $n > 0$ the proposition gives

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} = \sum_k (-1)^k \binom{n}{k} = 0.$$

If $n = 0$ the proposition does not apply and we can check manually that

$$\sum_{k=0}^{0} (-1)^k \binom{0}{k} = (-1)^0 \binom{0}{0} = 1.$$

Example 5.9. The term

$$a(n,k) = \frac{\binom{n}{k}\binom{n+1}{k}}{\binom{2n}{2k}}$$

is Gosper-summable with

$$A(n,k) = \frac{(2n-2k+1)k}{n+1}\,a(n,k)$$

As it stands though, $a(n,k)$ is undefined for $k<0$ or $k>n$. The alternative representation

$$b(n,k) = \frac{(-n-1)_k\,(2k)!}{(-n+1/2)_k\,4^k\,(k!)^2}$$

can be seen to be equal to $a(n,k)$ in the cases where the latter is defined and zero outside these cases. It also has finite support $\{0,1,\dots,n+1\}$.

Gosper's algorithm gives us the antidifference

$$B(n,k) = \frac{(2n-2k+1)k}{n+1}\,b(n,k)$$

Since the denominator in the rational certificate does not vanish for any $n \in \mathbb{Z}_{\geqslant 0}$, the proposition says that for all $n \geqslant 0$ we have

$$\sum_{k=0}^{n+1} \frac{\binom{n}{k}\binom{n+1}{k}}{\binom{2n}{2k}} = \sum_k b(k,n) = 0.$$

In addition to this direct application of Gosper's algorithm to definite summation, we have the Wilf–Zeilberger (WZ for those in the know) method for proving identities of the form

$$s_n = \sum_k a(n,k) = 1, \tag{5.2}$$

where $a(n,k)$ is hypergeometric in both n and k and has finite support.

The idea is to apply Gosper's algorithm not to $a(n,k)$ directly, but to the difference

$$d_k = a(n+1,k) - a(n,k)$$

with respect to k. If (d_k) is Gosper-summable, we get an antidifference D_k so that

$$a(n+1,k) - a(n,k) = d_k = D(n,k+1) - D(n,k)$$

and we have effectively moved the difference operator from the variable n to the variable k. It remains to sum over k and get

$$s_{n+1} - s_n = \sum_k \big(a(n+1,k) - a(n,k)\big) = \sum_k \big(D(n,k+1) - D(n,k)\big) = 0$$

This means that s_n is constant to respect to n, so $s_n = s_0$, and it only remains to show that $s_0 = 1$.

Example 5.10. Suppose we want to prove the well-known identity

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

Of course there's a very quick proof using the binomial expansion theorem. But let's instead try the WZ method. To do so we have to rewrite the identity in the form

$$\sum_{k=0}^{n} \frac{1}{2^n}\binom{n}{k} = 1.$$

We let

$$a(n,k) = \frac{1}{2^n}\binom{n}{k}$$

$$d_k = a(n+1,k) - a(n,k) = \frac{1}{2^{n+1}}\binom{n+1}{k} - \frac{1}{2^n}\binom{n}{k}$$

We apply Gosper's algorithm to (d_k) and get the antidifference

$$D_k = \frac{k}{n+1-2k}\, d_k$$

We encounter a potential problem as the denominator $n+1-2k$ is zero at $k=(n+1)/2$, so when n is odd this is in the set $\{0,1,\ldots,n\}$ over which we are summing. We are in luck however as

$$d_k = \frac{1}{2^{n+1}}\left(\frac{(n+1)!}{k!(n+1-k)!} - \frac{2n!}{k!(n-k)!}\right) = \frac{n!}{2^{n+1}k!(n+1-k)!}\,(2k-n-1)\,,$$

which is zero at $k=(n+1)/2$ and cancels out the zero in the denominator of D_k, leaving D_k well-defined for all k.

It remains to check out the case $n=0$, where the sum is trivially 1.

As the WZ method is derived from Gosper's algorithm, it also has the advantage of a certificate. We know that Gosper's result is of the form

$$D_k = R(k)d_k$$

where $R(k)$ is a rational function of k.

The *WZ certificate of* $a(n,k)$ is defined to be

$$\tilde{R}(n,k) = \frac{D_k}{a(n,k)} = R(k)\left(\frac{a(n+1,k)}{a(n,k)} - 1\right)$$

Given $\tilde{R}(n,k)$, the WZ identity (5.2) is equivalent to

$$\frac{a(n+1,k)}{a(n,k)} - 1 + \tilde{R}(n,k) - \tilde{R}(n,k+1)\,\frac{a(n,k+1)}{a(n,k)} = 0,$$

which is a simple exercise in arithmetic with rational functions.

Example 5.11. For the binomial identity in Example 5.10, the WZ certificate is

$$\tilde{R}(n,k) = -\frac{k}{2(n+1-k)},$$

so to certify the identity we need to check that

$$\frac{\frac{1}{2^{n+1}}\binom{n+1}{k}}{\frac{1}{2^n}\binom{n}{k}} - 1 - \frac{k}{2(n+1-k)} + \frac{k+1}{2(n-k)}\,\frac{\frac{1}{2^n}\binom{n}{k+1}}{\frac{1}{2^n}\binom{n}{k}} = 0,$$

which may look scary but really is not.

5.5 Zeilberger's method

Given a sum of the form

$$s_n = \sum_{k=-\infty}^{\infty} F(n,k)$$

where $F(n,k)$ is hypergeometric with respect to both n and k, and has finite support, we wish to find a recurrence relation of the form

$$\sum_{j=0}^{J} P_j(n)s_{n+j} = 0,$$

where the coefficients P_j are polynomials in $\mathbb{Q}[n]$. (Such a recurrence relation is called *holonomic*.)

Zeilberger's method consists of applying Gosper's algorithm to the sequence (a_k) defined by

$$a_k = F(n,k) + \sum_{j=1}^{J} \sigma_j(n) F(n+j,k)$$

where $\sigma_j \in \mathbb{Q}(n)$ are indeterminates that we wish to compute. Like in Fasenmyer's algorithm, we would start with $J = 1$ and increase its value until we find a solution or we decide that it's not worth the trouble.

One can check easily that (a_k) is indeed a hypergeometric term, so we can pass it to Gosper's algorithm.

Here are the steps of Zeilberger's algorithm:

(a) Set $J = 1$.

(b) Set

$$a_k = F(n,k) + \sum_{j=1}^{J} \sigma_j(n) F(n+j,k)$$

(c) Apply Gosper to (a_k), but in the step where we solve for the coefficients of the auxiliary polynomial f, solve also for the unknown σ_j's.

If Gosper succeeds, look at the denominator of the rational certificate to determine any singular values of n.

If Gosper fails, increment J and try again.

(d) Once we have achieved success, we have a recurrence

$$s_n + \sum_{j=1}^{J} \sigma_j(n) s_{n+j} = 0$$

with $\sigma_j \in \mathbb{Q}(n)$.

Clear the denominators to get a holonomic recurrence relation.

Example 5.12. Consider

$$s_n = \sum_k \binom{n}{k}$$

We try $J = 1$, setting

$$a_k = F(n,k) + \sigma_1 F(n+1,k) = \binom{n}{k} + \sigma_1 \binom{n+1}{k}$$

After simplifications

$$\frac{a_{k+1}}{a_k} = \frac{(n+1-k)(n-k+\sigma_1 n+\sigma_1)}{(k+1)(n+1-k+\sigma_1 n+\sigma_1)}$$

The polynomial representation in Gosper algorithm is then

$$p(k) = n+1-k+\sigma_1 n+\sigma_1, \qquad q(k) = n+2-k, \qquad r(k) = k$$

The upper bound on the degree of the auxiliary polynomial f is $N = 0$, so we put $f(k) = b_0$.

It has to satisfy

$$p(k) = q(k+1)f(k) - r(k)f(k-1),$$

which in this example becomes

$$n + 1 - k + \sigma_1 n + \sigma_1 = (n + 1 - k)b_0 - kb_0$$

Equating coefficients gives

$$-1 + 2b_0 = 0, \qquad (n + 1)(1 + \sigma_1 - b_0) = 0,$$

so $b_0 = 1/2$, $\sigma_1 = -1/2$.

The resulting recurrence relation is

$$s_n - \frac{1}{2} s_{n+1} = 0,$$

which gives $s_{n+1} = 2s_n$, and since $s_0 = 1$ we conclude that $s_n = 2^n$.

To recap, we proved the identity

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

We had already proved this using Wilf–Zeilberger in the last section, but with Zeilberger we didn't need to know a priori what the right hand side is, as it was found in the process!

5.6 More hypergeometric goodness

There is much more that can be said about hypergeometric identities. I recommend looking at [PWZ96] and [Koe14] for improvements and variants on the algorithms we discussed. Another reference that is related but also addresses other aspects is [GKP94].

Some answers/solutions/hints/more questions

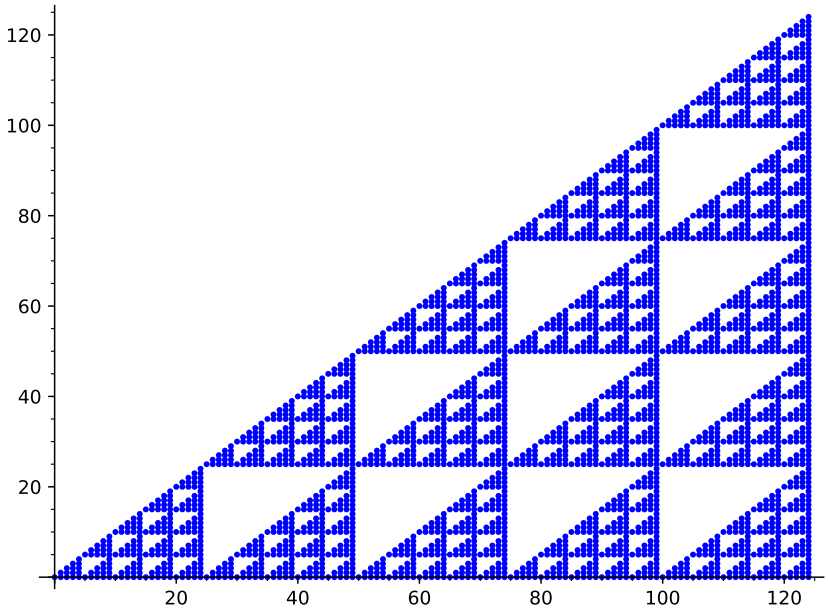
Exercise (2.2). For example:

```
sage: lst = [binomial(n, m) % 2 for n in range(2^8) for m in range(2^8)]
sage: mat = matrix(2^8, lst)
sage: p = mat.plot()
```

Exercise (2.3). Here's one approach for modulo k.

```
sage: def pascalmod(k, size=16, multicolour=True):
....:     p = list_plot([])
....:     for r in range(1, k):
....:         lstr = [(m, n) for m in range(size) for n in range(size) \
....:                 if binomial(m, n) % k == r]
....:         if multicolour:
....:             colour = hue(r/k)
....:         else:
....:             colour = 'blue'
....:         pr = list_plot(lstr, color=colour)
....:         p = p + pr
....:     return p
```

```
sage: p5 = pascalmod(5, 5^3, False)
sage: p5.show()
None
```



Exercise (3.1)**.** See [Sil13, Theorem 47.2].

Bibliography

[Cox84] David A. Cox. The arithmetic-geometric mean of Gauss. *Enseign. Math. (2)*, 30(3-4):275–330, 1984.

[Cox85] David A. Cox. Gauss and the arithmetic-geometric mean. *Notices Amer. Math. Soc.*, 32(2):147–151, 1985.

[GKP94] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete mathematics*. Addison-Wesley Publishing Company, Reading, MA, second edition, 1994. A foundation for computer science.

[Koe14] Wolfram Koepf. *Hypergeometric summation*. Universitext. Springer, London, second edition, 2014. An algorithmic approach to summation and special function identities.

[LLL82] A. K. Lenstra, H. W. Lenstra, Jr., and L. Lovász. Factoring polynomials with rational coefficients. *Math. Ann.*, 261(4):515–534, 1982.

[PWZ96] Marko Petkovšek, Herbert S. Wilf, and Doron Zeilberger. $A = B$. A K Peters, Ltd., Wellesley, MA, 1996. With a foreword by Donald E. Knuth, With a separately available computer disk.

[Sal76] Eugene Salamin. Computation of π using arithmetic-geometric mean. *Math. Comp.*, 30(135):565–570, 1976.

[Sil13] Joseph H. Silverman. *A friendly introduction to number theory*. Pearson, fourth edition, 2013.

[Sta12] Richard P. Stanley. *Enumerative combinatorics. Volume 1*, volume 49 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2012.

[Ste95] Ian Stewart. Four encounters with Sierpiński's gasket. *Math. Intelligencer*, 17(1):52–64, 1995.