Dodgson condensation, alternating signs and square ice

Andrew N.W Hone

References	**This article cites 14 articles** http://rsta.royalsocietypublishing.org/content/364/1849/3183.full.html#ref-list-1
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *Phil. Trans. R. Soc. A* go to:
http://rsta.royalsocietypublishing.org/subscriptions

Dodgson condensation, alternating signs and square ice

By Andrew N. W. Hone*

Institute of Mathematics & Statistics, University of Kent at Canterbury, Kent CT2 7NF, UK

Starting with the numbers $1, 2, 7, 42, 429, 7436$, what is the next term in the sequence? This question arose in the area of mathematics called algebraic combinatorics, which deals with the precise counting of sets of objects, but it goes back to Lewis Carroll's work on determinants. The resolution of the problem was only achieved at the end of the last century, and with two completely different approaches: the first involved extensive verification by computer algebra and a huge posse of referees, while the second relied on an unexpected connection with the theory of 'square ice' in statistical physics. This paper, aimed at a general scientific audience, explains the background to this problem and how subsequent developments are leading to a fruitful interplay between algebraic combinatorics, mathematical physics and number theory.

Keywords: alternating sign matrices; six-vertex model; Laurent property

1. Introduction

The nineteenth-century mathematician Charles Dodgson is best known by the name Lewis Carroll, the pseudonym under which he wrote the *Alice* books. However, he also wrote a treatise on the theory of determinants (Dodgson 1867), and worked on a method for calculating these, which is sometimes referred to as Dodgson condensation. There is an apocryphal story that Queen Victoria was so delighted with *Alice in Wonderland* and *Through the Looking Glass* that she requested a copy of his next book, and was distinctly unamused to receive a book about mathematics, but Dodgson himself denied the veracity of this tale.

Although it is a useful method, Dodgson condensation was largely forgotten for more than a century, until it was investigated by Mills *et al.* (1983). In the course of their study, they were led to the problem of enumerating so-called alternating sign matrices (or ASMs for short), which are square arrays consisting only of the numbers 0, 1 and -1, where for each row and column the entries sum to 1, the non-zero entries begin and end with $+1$, and the signs alternate in between. The unique ASM of order 1 is just a single number (1), while for 2×2 arrays there are only two possibilities, *viz.*

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{1.1}$$

*anwh@kent.ac.uk

and at order 3, there are seven different 3×3 matrices of this sort, namely

$$
\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},
\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},
\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},
\begin{pmatrix} 0 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix},
\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},
$$

$$
\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},
\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}.
\tag{1.2}
$$

As the size of the matrices increases, the number of different possibilities grows quite rapidly: if A_n denotes the number of ASMs of order n, then the sequence A_n goes

$$
1, 2, 7, 42, 429, 7436, 218\,348, 10\,850\,216, \ldots.
\tag{1.3}
$$

Mills *et al.* did calculations with a computer to work out the first few terms of the sequence A_n, and they were able to conjecture an exact formula which would hold for any n. However, a complete proof of this ASM conjecture remained elusive for over 10 years, until Zeilberger (1996*a*) found a remarkably intricate approach that required a computer to verify a vast number of identities, as well as 88 human referees who collectively checked the different steps that linked these identities. Almost immediately afterwards, Kuperberg (1996) found a much shorter proof that exploited a surprising connection between ASMs and configurations of 'square ice' (also known as the six-vertex model), which is a simplified model from statistical physics that describes the packing of water molecules in ice.

In what follows, we aim to give the general reader a flavour of some of the ingredients in the above story: determinants and condensation; plane partitions; and solvable models of statistical mechanics. Moreover, we shall see that this story is part of an ongoing saga that connects seemingly distant areas of mathematics and physics, ranging from recurrence sequences in number theory to the propagation of waves on shallow water.

2. Determinants and condensation

Matrices are ubiquitous in mathematics and its applications, and finding the determinant of a matrix means associating a single number to it. Probably, the simplest interpretation of a matrix is a geometrical one: a square matrix corresponds to a transformation of Euclidean space that leaves the origin unchanged. For example, a 2×2 matrix

$$
\boldsymbol{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix},
$$

corresponds to a transformation of the plane,

$$
\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \boldsymbol{M}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix},
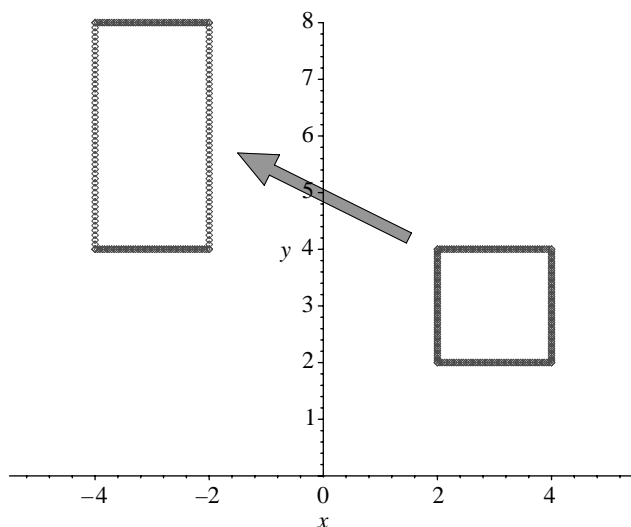$$

Figure 1. A shape is transformed to another shape with twice the original area by a matrix M with determinant 2.

and the magnitude of its determinant,

$$\det M = ad - bc,$$

gives the proportional change in the area of a shape when it is moved according to this transformation. Figure 1 shows a square with area 4 that is transformed by the matrix

$$M = \begin{pmatrix} 0 & -1 \\ 2 & 0 \end{pmatrix},$$

which has the determinant

$$\det M = 0 \times 0 - (-1) \times 2 = 2,$$

so the area of the image rectangle is 8, i.e. twice that of the original square.

Square matrices of the same size can be multiplied together, which corresponds to the composition of transformations, but an important feature of this multiplication is that it is not commutative, so the order matters, i.e. given a generic pair of matrices L and M, we find that $LM \neq ML$. For example, if M is as before and

$$L = \begin{pmatrix} 3 & 3 \\ -3 & 3 \end{pmatrix} \quad \text{then} \quad LM = \begin{pmatrix} 6 & -3 \\ 6 & 3 \end{pmatrix} \neq \begin{pmatrix} 3 & -3 \\ 6 & 6 \end{pmatrix} = ML,$$

and the fact that the order of the composition changes the result can be seen geometrically in figure 2. However, in both the cases, the area of the final image is scaled by the same factor, since $\det LM = \det ML = \det L \times \det M = 2 \times 18 = 36$.
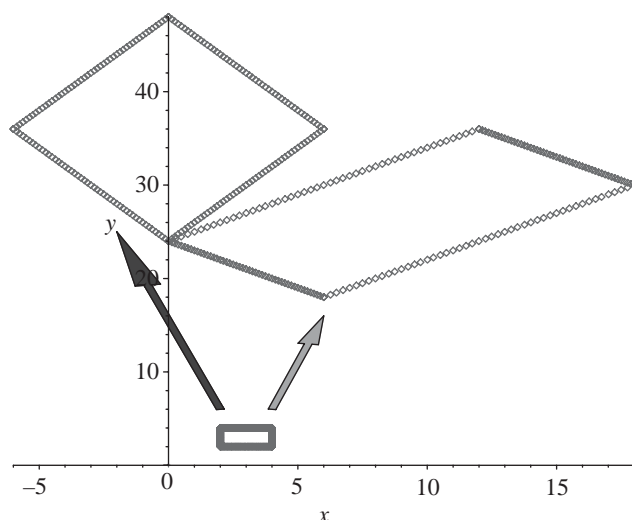
Figure 2. Matrix composition depends on the order: either $\boldsymbol{M}$ followed by $\boldsymbol{L}$, which is the product $\boldsymbol{LM}$ (short arrow), or $\boldsymbol{L}$ followed by $\boldsymbol{M}$, i.e. the product $\boldsymbol{ML}$ (long arrow).

For a $3{\times}3$ matrix,

$$\boldsymbol{M} = \begin{pmatrix} p & q & r \\ s & t & u \\ v & w & x \end{pmatrix}, \tag{2.1}$$

the determinant,

$$\det \boldsymbol{M} = ptx - puw - sqx + srw + vqu - vrt, \tag{2.2}$$

corresponds to the volume scale factor for a transformation, and there is an analogous interpretation for the determinant of an $n{\times}n$ matrix. However, as we shall see, determinants acquire a multitude of meanings in different applications. For an $n{\times}n$ matrix, the general formula defining the determinant involves products of the entries which are summed over all the different permutations of the n rows (or, equivalently, the n columns), with a plus/minus sign in the front depending on the permutation. For small matrices, this formula is easy to calculate with, but as the size grows it rapidly becomes impractical, because the number of permutations of n things is given by the factorial,

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1,$$

which grows very fast with n; its growth is given by $n! \sim \sqrt{2\pi n}(n/e)^n$, which is Stirling's formula. Many other ways have been developed for evaluating determinants more efficiently, and Dodgson condensation is one such way.

Basically, the method considered by Dodgson boils down to calculating a sequence of smaller and smaller matrices, so for a particular $3{\times}3$ example, one gets

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 2 \\ 1 & 6 \end{pmatrix} \rightarrow 2,$$

and hence the determinant of the 3×3 matrix on the left-hand side is 2. But how does the method actually work? Well, at the first stage, one works out the determinant of each 2×2 block, and then forms the next matrix out of these numbers. Hence, starting from a 3×3 matrix of the general form (2.1), one gets the 2×2 matrix

$$\begin{pmatrix} pt-qs & qu-rt \\ sw-tv & tx-uw \end{pmatrix}. \tag{2.3}$$

At each subsequent stage, one must again work out the determinant of each 2×2 block, but one must also divide each of these numbers by the corresponding central entry of the matrix from the previous stage before forming the next matrix. Therefore, for instance, one works out the determinant of the 2×2 matrix (2.3), and then one must divide it by the central entry t from the previous matrix (2.1) to get

$$\{(pt-qs)(tx-uw)-(qu-rt)(sw-tv)\}/t,$$

which (after multiplying out and collecting all terms) gives

$$(1)ptx + (-1)puw + (-1)qsx + (0)qsuwt^{-1} + (1)quv + (1)rsw + (-1)rtv. \tag{2.4}$$

The final expression (2.4) for the determinant is equal to the formula (2.2), but the inclusion of the term with a zero seems strange. In fact, Mills *et al.* (1983) found that in addition to the usual terms coming from permutations, Dodgson condensation produces extra terms that lead to final cancellations (like the zero above). Moreover, they showed that each of the terms is associated with an ASM, which is the matrix formed by placing a $+1$ in the position corresponding to one of the entries (like p, q, ..., etc.), a -1 in the position corresponding to the inverse of an entry (like t^{-1}) and 0 elsewhere. For example, the term $qsuwt^{-1}$ in (2.4) gives a 1 in the q, s, u and w positions and a -1 in the t position, with zeros everywhere else, giving the ASM

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

There are seven terms in expression (2.4), corresponding to the seven ASMs of order 3 given in (1.2).

As the matrix size n gets bigger, it gets much harder to enumerate all the ASMs. From computer calculations, Mills *et al.* found the first few terms in the sequence (1.3) of numbers of ASMs of order n, and by looking at their prime factors they were able to guess that the exact formula for these numbers should be given as a ratio of products of factorials, namely

$$A_n = \frac{1!4!7!\ldots(3n-2)!}{n!(n+1)!(n+2)!\ldots(2n-1)!}. \tag{2.5}$$

Using Stirling's formula, one gets an idea of how fast these numbers grow,

$$\log A_n \sim Cn^2, \qquad C = \log\sqrt{\frac{27}{16}} \approx 0.262. \tag{2.6}$$

Despite all the evidence for formula (2.5), it took more than a decade of work and insights from many different people before a proof was forthcoming. Furthermore, although the apparently simple problem of counting ASMs had

its origins in Dodgson condensation, it turned out to be related to a completely different problem about partitions.

3. Plane partitions

A partition of an integer is a way of breaking it up into a sum of smaller integers. For example, $(6, 6)$, $(4, 4, 2, 2)$ and $(5, 3, 2, 1, 1)$ are three different partitions of the number 12, because $6 + 6 = 4 + 4 + 2 + 2 = 5 + 3 + 2 + 1 + 1 = 12$. Partitions were first studied by the mathematician Leonhard Euler in the eighteenth century, and as well as being part of the additive theory of numbers, nowadays they have many important applications to such diverse areas as the theory of group representations in algebra and statistical problems in physics. In the nineteenth century, a generalization called plane partitions was developed by an English-man, Percy MacMahon, who was a major in the army as well as being a distinguished mathematician. MacMahon's plane partitions are given by rows of numbers represented in the plane, where the numbers are weakly decreasing along each row (so each number is less than or equal to the preceding one), and also weakly decreasing in each column. An example of this is

$$
\begin{array}{cccccc}
6 & 6 & 6 & 5 & 5 & 3 \\
6 & 5 & 5 & 3 & 3 & 1 \\
6 & 5 & 5 & 3 & 3 & 1 \\
5 & 3 & 3 & 1 & 1 & \\
3 & 1 & 1 & & &
\end{array}
\tag{3.1}
$$

which is a particular plane partition of the number 95. MacMahon made a lot of conjectures about plane partitions, but these difficult problems did not really begin to be tackled until the latter half of the twentieth century, in the work of Macdonald, Andrews, Stanley and others.

So how is this related to ASMs? Well, Andrews considered another type of plane partitions, called descending plane partitions of order n, and for small values he had worked out that the number of these increased with n according to 1,2,7,42,429, …—exactly the same as sequence (1.3)! When Mills *et al.* heard about this, they were able to prove one of MacMahon's conjectures, but they still could not prove their original ASM conjecture (2.5). However, they went on to consider plane partitions with special symmetries, known by the acronym totally symmetric self-complementary plane partitions (TSSCPPs), and Robbins noted that the sequence of these also began as 1,2,7,42,429, …—a third instance of the same sequence of numbers, which could not be a mere coincidence. Andrews then showed that the right-hand side of equation (2.5) *does* give the right formula for counting these TSSCPPs, and this eventually led Zeilberger to his proof of the ASM conjecture: he was able to show that the number of ASMs of size n is the same as the corresponding number of TSSCPPs.

The proof of Zeilberger (1996a) is incredibly intricate and not at all transparent, but the way it is structured is masterful. The main result is built upon a tower of smaller results: lemmas; sublemmas; subsublemmas; and so on. At each floor of this tower, one of Zeilberger's referees was required to check that the structure was solidly built, in the sense that the result on that floor did

indeed follow correctly from the results on the lower floors. At each level, there is a particular identity for constant terms in the expansion of certain rational functions, and the electronic version of the paper is accompanied by a software package called ROBBINS, which allows these identities to be checked. Zeilberger's paper was over 80 pages long, and although he finished it at the end of 1992, it could not be declared formally correct and watertight until all the checkers had finished doing their job, which took until 1995; it was finally published the next year.

In the introduction to his paper, Zeilberger says

> In this paper I present the *first* proof of this fact. I do hope it is not the *last* proof, and that a shorter, more direct, elegant and combinatorial proof will be found one day.
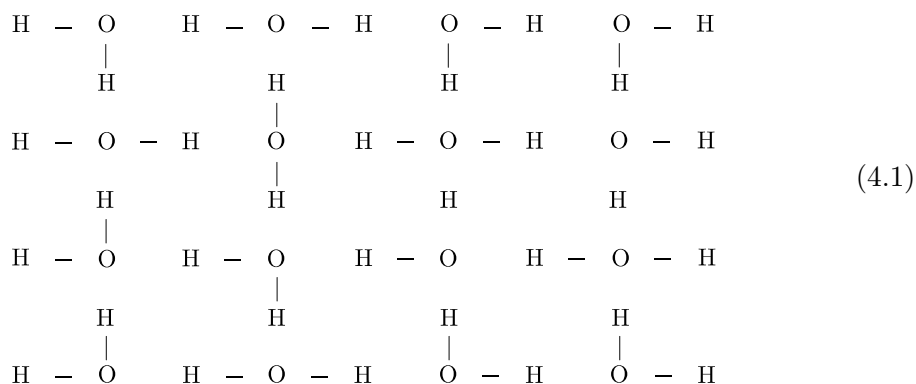
<div align="right">(Zeilberger 1996*a*)</div>

In fact, he did not have to wait very long, because almost immediately Kuperberg (1996) presented a very short alternative proof, making use of ideas and techniques from statistical mechanics (described in §4). However, this did not render Zeilberger's work obsolete, because the method he used is itself a masterpiece of technique that might be applied to harder problems in the future; furthermore, he simultaneously proved some stronger results which were not covered by Kuperberg. Zeilberger's paper also makes very interesting reading by virtue of an appendix, which contains potted biographies of all his referees.

4. Square ice

Although Zeilberger's proof resolved the ASM conjecture by showing that the number of ASMs of order n was the same as the number of plane partitions of a certain sort (TSSCPPs), it did not actually establish a direct correspondence between these two sets of things. Kuperberg's approach made use of a rather different observation, namely that there is a direct correspondence (bijection) between ASMs and the states of a lattice model for water molecules known as 'square ice'. Real ice crystals are three-dimensional, and square ice is a two-dimensional simplification, i.e. imagine a crystal that looks the same when it is cleaved open along any horizontal plane. The advantage of considering this simpler model in two dimensions is that it turns out to be exactly solvable.

Here is a small chunk of square ice.

$$
\begin{array}{cccccccccc}
\text{H} & - & \text{O} & & \text{H} & - & \text{O} & - & \text{H} & & & \text{O} & - & \text{H} & & & \text{O} & - & \text{H} \\
& & | & & & & & & & & & | & & & & & | & & \\
& & \text{H} & & & & \text{H} & & & & & \text{H} & & & & & \text{H} & & \\
& & & & & & | & & & & & & & & & & & & \\
\text{H} & - & \text{O} & - & \text{H} & & \text{O} & & \text{H} & - & \text{O} & - & \text{H} & & & \text{O} & - & \text{H} \\
& & & & & & | & & & & & & & & & & & & \\
& & \text{H} & & & & \text{H} & & & & \text{H} & & & & & \text{H} & & & \\
& & | & & & & & & & & & & & & & & & & \\
\text{H} & - & \text{O} & & \text{H} & - & \text{O} & & \text{H} & - & \text{O} & & \text{H} & - & \text{O} & - & \text{H} \\
& & & & & & | & & & & & & & & | & & & & \\
& & \text{H} & & & & \text{H} & & & & \text{H} & & & & \text{H} & & & & \\
& & | & & & & | & & & & & & & & | & & & & \\
\text{H} & - & \text{O} & & \text{H} & - & \text{O} & - & \text{H} & & \text{O} & - & \text{H} & & \text{O} & - & \text{H} \\
\end{array}
\tag{4.1}
$$

The numbers 1, -1 and 0 can be assigned to the water molecules, so that horizontal molecules correspond to $+1$, vertical molecules to -1 and molecules with bonds at right angles to 0. In this way, suitable configurations of the model are seen to be in bijection with ASMs. The block of square ice shown above corresponds to the 4×4 ASM

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

More often than not, the configurations of square ice are drawn as a directed graph on a square lattice, so that the picture (4.1) becomes

$$
\begin{array}{ccccccccc}
& \uparrow & & \uparrow & & \uparrow & & \uparrow & \\
\rightarrow & \bullet & \rightarrow & \bullet & \leftarrow & \bullet & \leftarrow & \bullet & \leftarrow \\
& \uparrow & & \downarrow & & \uparrow & & \uparrow & \\
\rightarrow & \bullet & \leftarrow & \bullet & \rightarrow & \bullet & \leftarrow & \bullet & \leftarrow \\
& \downarrow & & \uparrow & & \downarrow & & \uparrow & \\
\rightarrow & \bullet & \rightarrow & \bullet & \rightarrow & \bullet & \rightarrow & \bullet & \leftarrow \\
& \downarrow & & \uparrow & & \downarrow & & \downarrow & \\
\rightarrow & \bullet & \rightarrow & \bullet & \leftarrow & \bullet & \leftarrow & \bullet & \leftarrow \\
& \downarrow & & \downarrow & & \downarrow & & \downarrow &
\end{array}
\qquad (4.2)
$$

The directed graph above is built out of six different types of vertices, corresponding to six different possible orientations of water molecules, namely the $+1$ and -1 configurations, which are

$$
\begin{array}{ccc}
\uparrow & & \downarrow \\
\rightarrow \; \bullet \; \leftarrow & \quad \text{and} \quad & \leftarrow \; \bullet \; \rightarrow , \\
\downarrow & & \uparrow
\end{array}
$$

respectively, together with the four states associated with the number 0,

$$
\begin{array}{cccc}
\uparrow & \downarrow & \uparrow & \downarrow \\
\rightarrow \; \bullet \; \rightarrow & \leftarrow \; \bullet \; \leftarrow & \leftarrow \; \bullet \; \leftarrow & \rightarrow \; \bullet \; \rightarrow . \\
\uparrow & \downarrow & \uparrow & \downarrow
\end{array}
$$

Owing to these six types of vertices, square ice is commonly referred to as the six-vertex model.

Note that the arrows in (4.2) all point inwards along the left- and right-hand edges, while they all point outwards along the top and bottom edges. This requirement is called the domain wall boundary condition for the six-vertex model, and it is precisely the states satisfying this boundary condition that are in

direct correspondence with ASMs. Moreover, these boundary conditions were studied by Korepin, a Russian theoretical physicist, from the beginning of the 1980s (Korepin *et al.* 1993).

In statistical mechanics, each configuration or state S of a system is assigned an energy $E[S]$, and the probability that the state occurs at temperature T is proportional to $\exp(-E[S]/(kT))$, where $k \approx 1.38 \times 10^{-23}$ J K^{-1} is Boltzmann's constant (Mandl 1971). This probability increases with decreasing energy, which means that states of low energy are more probable than high-energy ones. Of central importance is the partition function of the model, which is given by a weighted sum over all possible states,

$$Z = \sum_S \exp\left(-\frac{E[S]}{kT}\right).$$

Building on Korepin's work, Izergin showed that the partition function for the six-vertex model could be given explicitly in terms of a determinant. Thus, here we see a completely different situation where determinants appear!

The full determinant formula for Z in the six-vertex model actually contains several parameters. Therefore, in order to tackle the ASM conjecture, Kuperberg had to do some very elegant manipulations. In this way, he was able to massage the original partition function into the form of an enumeration of ASMs, by doing a suitable limiting procedure. Thus, Kuperberg's alternative proof appeared shortly after Zeilberger's had been verified. Zeilberger wasted no time in absorbing the finer points of Kuperberg's technique, to the extent that he was subsequently able to extend it and thence prove a more refined result about ASMs (Zeilberger 1996*b*). For the interested reader, I can recommend two excellent survey papers (Bressoud & Propp 1999; Propp 2001) and a marvellous book (Bressoud 1999).

It would appear that all the main problems concerning Dodgson condensation and ASMs were already dealt with by the mid-1990s, but, in fact, the outlook is more interesting. On the one hand, Kuperberg and others have continued to obtain further insights into the symmetries of ASMs, while on the other hand, his work is having an impact on the deeper understanding of statistical and quantum mechanical models. Thus, while Kuperberg applied various techniques coming from solvable models in physics to solve a problem in combinatorics, it seems that the flow of technology is now going in the opposite direction, so that these combinatorial insights are yielding new results in physics (e.g. Batchelor *et al.* 2001). In a completely different direction, the work of Mills *et al.* on Dodgson condensation has stimulated another set of developments that links a diverse set of ideas ranging from number theory, via commutative algebra, to the theory of nonlinear waves. This next chapter of the saga is where we now turn.

5. Recurrence sequences and solitons

The Fibonacci sequence $1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots$ is the most famous example of a recurrence sequence. It is generated by the linear recurrence

$$F_{n+2} = F_{n+1} + F_n, \qquad F_0 = 0, \quad F_1 = 1.$$

When Robins and Rumsey were performing their analysis of the combinatorics of Dodgson condensation, they made a new discovery concerning a recurrence given by

$$D_{\ell,m,n+1} = (D_{\ell+1,m,n}D_{\ell-1,m,n} + \lambda D_{\ell,m+1,n}D_{\ell,m-1,n})/D_{\ell,m,n-1}, \qquad (5.1)$$

where the indices ℓ, m and n take integer values. For $\lambda = -1$, this recurrence encodes the relationship between the different determinants (called minors) that appear at each stage of the condensation process. What they observed was that the terms obtained by iterating this recurrence are polynomials in the initial values and their inverses, called Laurent polynomials, and these polynomials have integer coefficients.

In order to illustrate this *Laurent phenomenon*, and see why it is remarkable, we will take a simpler recurrence with only one index n, namely the Somos-4 recurrence

$$\tau_{n+4}\tau_n = \tau_{n+3}\tau_{n+1} + (\tau_{n+2})^2, \qquad (5.2)$$

with initial values $\tau_0 = \tau_1 = \tau_2 = \tau_3 = 1$, which was found by Michael Somos when he was investigating the combinatorics of theta functions. Propp noted that this can be obtained as a reduction of (5.1), e.g. by taking $\lambda = 1$ and $D_{\ell,m,n} = \tau_{2n+m}$ (see also Speyer 2004). With all four initial values equal to 1, the fourth-order recurrence (5.2) yields a sequence of integers, that is

$$1, 1, 1, 1, 2, 3, 7, 23, 59, 314, 1529, 8209, 83\,313, \ldots. \qquad (5.3)$$

Now why is this surprising? Well, in order to obtain each new term τ_{n+4} in the sequence from the preceding four, it is necessary to divide by τ_n at each stage, *viz.*

$$\tau_{n+4} = \frac{(\tau_{n+3}\tau_{n+1} + (\tau_{n+2})^2)}{\tau_n}. \qquad (5.4)$$

Miraculously, all the factors in the denominator always cancel with some factors from the numerator, and only integers (not fractions) come out!

Various proofs that the terms of the sequence (5.3) are all integers were found in around 1990 (Gale 1991), but a deeper understanding was lacking. It turned out that the deeper reason behind it was the Laurent property alluded to above: if the initial values $\tau_0 = a$, $\tau_1 = b$, $\tau_2 = c$, $\tau_3 = d$ are regarded as variables, then the terms generated by (5.2) turn out to be polynomials in these variables and their inverses a^{-1}, b^{-1}, etc. Polynomials of this kind are called Laurent polynomials, and in this case their coefficients are all integers. To see what this means explicitly, observe that if we iterate (5.2) starting from a,b,c,d, then we get

$$\tau_4 = bda^{-1} + c^2a^{-1}, \qquad \tau_5 = d^2b^{-1} + cda^{-1} + c^3a^{-1}b^{-1}$$
$$\tau_6 = d^3b^{-1}c^{-1} + d^2a^{-1} + b^2d^2a^{-2}c^{-1} + c^2da^{-1}b^{-1} + 2bcda^{-2} + c^3a^{-2}. \qquad (5.5)$$

Each of the expressions above is a sum of monomials of the form $a^pb^qc^rd^s$ for integers p, q, r, s, and the coefficients in front of each monomial are all integers. So far, there are no surprises, and clearly for τ_7 one also gets a Laurent polynomial (in fact, this one has 12 terms), because the first iterations only involve successive divisions by a, b, c, d. However, to find τ_8, one has to divide by τ_4, and the polynomial factor $(bd + c^2)$ that appears in the denominator on the right-hand side of equation (5.4) when $n = 4$ happens to cancel out from the numerator, so that τ_8 is a Laurent polynomial given by a sum of 17 monomial terms.

Furthermore, this remarkable cancellation keeps on happening at every step, so that τ_n is a Laurent polynomial with integer coefficients for all n. If we take the particular values $a = b = c = d = 1$, then it is easy to see that the Somos-4 sequence of integers given by (5.3) arises as a consequence of the Laurent property for the recurrence (5.2).

A complete understanding of the Laurent property has only begun to emerge very recently as an offshoot of the theory of cluster algebras due to Fomin and Zelevinsky. We have seen that matrix transformations can be combined, and that this composition is not commutative. Sets of invertible matrix transformations form groups, and if they can be specified by algebraic equations then they are called algebraic groups. In that case, they can also be viewed geometrically as spaces of some dimension (i.e. as algebraic varieties). Although the groups themselves are not commutative, the coordinates that define these spaces do commute, and it was precisely to understand the structure of such coordinates that Fomin and Zelevinsky introduced a new class of commutative algebras called cluster algebras. It transpired that sets of coordinates in these algebras (called clusters) satisfy a certain type of recurrence relation, of which (5.1) and (5.2) are particular cases. Therefore, in the general setting of cluster algebras, it was possible to give proofs of the Laurent property for a wide variety of these recurrences (Fomin & Zelevinsky 2002).

Meanwhile, various people (Swart 2003; van der Poorten 2005) had begun to explain the connection between sequences of points on elliptic curves and the iterates of the general Somos-4 recurrence

$$\tau_{n+4}\tau_n = \alpha\tau_{n+3}\tau_{n+1} + \beta(\tau_{n+2})^2, \qquad (5.6)$$

with arbitrary constant coefficients α and β. Earlier, such a connection was understood for fifth-order (Somos-5) recurrences in unpublished results of Zagier[1] and Elkies (quoted in Buchholz & Rathbun 1997). Independently, I was able to show (Hone 2005) that, for any sequence τ_n generated by the recurrence (5.6), the nth term is given by the explicit formula

$$\tau_n = AB^n \frac{\sigma(z_0 + nh)}{\sigma(h)^{n^2}}, \qquad (5.7)$$

for suitable parameters A, B, z_0, h, g_2, g_3, with $\sigma(z) = \sigma(z; g_2, g_3)$ being the Weierstrass sigma function, with invariants g_2 and g_3 associated with the elliptic curve

$$y^2 = 4x^3 - g_2 x - g_3. \qquad (5.8)$$

Using this formula, one can show that the original Somos-4 sequence (5.3) corresponds to a sequence of points on the particular curve with $g_2 = 4$ and $g_2 = -1$, and the leading-order asymptotics of this sequence of integers has the form

$$\log\tau_n \sim \hat{C}n^2, \qquad \hat{C} = \frac{\zeta(\omega_1)h^2}{2\omega_1} - \log|\sigma(h)| \approx 0.102, \qquad (5.9)$$

where the constant $\hat{C}$ has been given in terms of the sigma and zeta functions of the curve; the parameter $h \approx -1.134$; and the half-period $\omega_1 \approx 1.497$. This asymptotic behaviour should be compared with equation (2.6).

Actually, the Somos-4 recurrence has an ancestor from the 1940s, in Morgan Ward's work on elliptic divisibility sequences. Ward considered the recurrence (5.6) with the initial data given by the four integer values τ_1, τ_2, τ_3, τ_4, where

[1] See http://www-groups.dcs.st-and.ac.uk/john/Zagier/Problems.html.

these values and the parameters are constrained by $\tau_1 = 1$, $\tau_2 | \tau_4$, $\alpha = (\tau_2)^2$, $\beta = -\tau_1 \tau_3$. He showed that these special sequences correspond to multiples nP of a single point P on an elliptic curve (5.8), and all of the terms are integers satisfying the divisibility property $\tau_n | \tau_m$ whenever $n | m$. (These sequences also extend to negative n, with $\tau_0 = 0$ and $\tau_{-n} = -\tau_n$.) For example, taking the recurrence (5.2) with $\tau_1 = 1$, $\tau_2 = -1$, $\tau_3 = -1$, $\tau_4 = -1$ gives the sequence $1, -1, -1, 2, 1, -3, 5, 7, 4, -23, -29, 59, -129, -314, \ldots$, whose odd index terms are (up to a sign) the same as the sequence (5.3); indeed, this is associated with the multiples of the point $P = (x, y) = (0, 1)$ on the same elliptic curve, with $g_2 = 4$ and $g_3 = -1$.

Somos sequences are currently of interest to number theorists owing to the way that new prime divisors appear therein (see Everest *et al.* 2003, ch. 10 & 11). In a certain sense, they are a natural generalization of linear recurrence sequences, as they have similar properties. For example, the even index terms F_{2n} of the Fibonacci sequence form a divisibility sequence, with $F_{2n} | F_{2m}$ whenever $n | m$, and $\tau_n = F_{2n}$ satisfies a Somos-4 recurrence, that is (5.6) with $\alpha = 9$ and $\beta = -8$. This corresponds to the degenerate case when the curve (5.8) has vanishing discriminant, hence $g_2^3 - 27 g_3^2 = 0$ and the sigma function in (5.7) becomes a hyperbolic sine.

I got interested in these problems when I heard Graham Everest give a talk about the arithmetical properties of Somos-4 sequences, and I immediately guessed that an analytic expression like (5.7) should exist. However, the inspiration behind my proof of this formula did not come from number theory, but rather from the theory of nonlinear waves called solitons. The archetypal partial differential equation (PDE) in soliton theory is the KdV equation,

$$u_t + u_{xxx} + 6uu_x = 0, \tag{5.10}$$

named after Korteweg and deVries, who derived it in 1895 as a model for the propagation of long waves on shallow water. The KdV equation is completely integrable, in the sense that the initial-value problem can be solved exactly by the inverse scattering transform (Ablowitz & Clarkson 1991). Moreover, it has solutions corresponding to the superposition of an arbitrary number of waves (solitons), which preserve their shape and speed even after colliding with each other. Hirota developed a very neat (and purely algebraic) way of finding these multi-soliton solutions (Hirota 1972). He substituted the second logarithmic derivative

$$u = 2(\log \tau)_{xx},$$

of a new quantity τ, known as the tau-function, into (5.10), thus obtaining the fourth-order PDE

$$\tau_{xt}\tau - \tau_t\tau_x + \tau_{xxxx}\tau - 4\tau_{xxx}\tau_x + 3(\tau_{xx})^2 + f\tau^2 = 0 \tag{5.11}$$

(where $f = f(t)$ is an arbitrary function of t). With this equation, which is called the Hirota bilinear form of KdV, it is much easier to find the solitons than to work with the original nonlinear PDE (5.10), since then τ is just a polynomial in exponential functions of x and t. For a single soliton, one can just take $\tau = 1 + \exp \gamma(x - ct)$, $c = \gamma^2$ with $f = 0$, while for multi-solitons τ can be written as a determinant—thus, we see yet another alternative use for determinants!
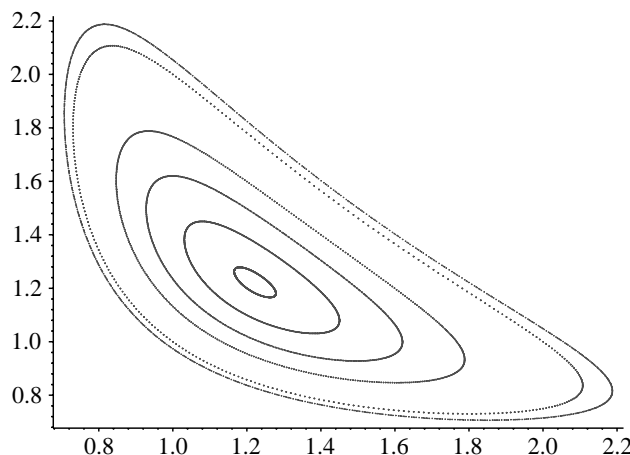
Figure 3. A family of orbits for the nonlinear recurrence (5.12) associated with Somos-4.

Making the reduction of (5.11) to travelling waves moving with speed c, by setting $\tau = \tau(x - ct)$, gives a fourth-order bilinear ordinary differential equation (ODE), which is solved in terms of the sigma function; the single soliton is a degenerate (hyperbolic function) case of the sigma function. The fourth-order recurrence (5.6) is a discrete analogue of this ODE. In the discrete setting, the associated nonlinear equation is obtained by taking the finite difference version of the second logarithmic derivative, namely with $\tau_n = \tau(z)$ and $z = z_0 + nh$ we have

$$u_n = \frac{\tau_{n+1}\tau_{n-1}}{(\tau_n)^2} \equiv \frac{\tau(z+h)\tau(z-h)}{\tau(z)^2} \approx 1 + h^2 \frac{d^2}{dz^2}(\log \tau),$$

as $h \to 0$, giving the second-order nonlinear recurrence

$$u_{n+2}u_n = (\alpha u_{n+1} + \beta)/(u_{n+1})^2. \tag{5.12}$$

This second-order equation for u_n also has a physical interpretation: it is a discrete analogue of a conservative system in classical mechanics, with total 'energy'

$$J = u_{n+1}u_n + \alpha\left(\frac{1}{u_{n+1}} + \frac{1}{u_n}\right) + \frac{\beta}{u_{n+1}u_n}, \tag{5.13}$$

and the symplectic form $\omega = (du_n \wedge du_{n+1})/u_n u_{n+1}$, which are conserved with n. This leads to an orbit structure for the solutions (figure 3), and the quartic curve defined by (5.13) can be parameterized by elliptic functions, which ultimately explains formula (5.7).

In the future, there are many more connections to be explored between all these different ideas. The recurrence (5.1) is equivalent to a partial difference soliton equation known as the discrete Hirota equation, and this describes the properties of quantum integrable systems (Krichever *et al.* 1997). The Laurent property should also be shared by non-autonomous bilinear equations like those for discrete Painlevé transcendents, and tau-functions and determinantal calculus are very important in this setting too (Noumi 2004). From another angle, if the sequence (1.3) counts the number of ASMs, then what do sequences

A. N. W. Hone

like (5.3) count? This question has already begun to be answered by James Propp and colleagues (Carroll & Speyer 2004; Speyer 2004), several of whom were undergraduates taking part in his REACH project (Research Experiences in Algebraic Combinatorics at Harvard). It was found by the REACH team, and independently by Bousquet-Mélou and West, that for a certain family of graphs,[2] the numbers (5.3) enumerate the so-called perfect matchings (also known as dimer coverings in the language of statistical mechanics). In the future, I envisage that there will be a deeper understanding of the interplay between combinatorics, algebra, number theory and integrable physical models, which will be beneficial to all concerned.

References

Ablowitz, M. J. & Clarkson, P. A. 1991 *Solitons, nonlinear evolution equations and inverse scattering*. Cambridge, UK: Cambridge University Press.

Batchelor, M. T., de Gier, J. & Nienhuis, B. 2001 The quantum symmetric XXZ chain at $\Delta=-1/2$, alternating-sign matrices and plane partitions. *J. Phys. A: Math. Gen.* **34**, L265–L270. (doi:10.1088/0305-4470/34/19/101)

Bressoud, D. M. 1999 *Proofs and confirmations: the story of the alternating sign matrix conjecture*. Cambridge, UK: Cambridge University Press.

Bressoud, D. M. & Propp, J. 1999 How the alternating sign matrix conjecture was solved. *Notices Am. Math. Soc.* **46**, 637–646.

Buchholz, R. H. & Rathbun, R. L. 1997 An infinite set of heron triangles with two rational medians. *Am. Math. Mon.* **104**, 107–115. (doi:10.2307/2974977)

Carroll, G. & Speyer, D. 2004 The cube recurrence. *Elec. J. Comb.* **11**, #R73.

Dodgson, C. L. 1867 *An elementary treatise on determinants, with their application to simultaneous linear equations and algebraical geometry*. London, UK: Macmillan.

Everest, G., van der Poorten, A., Shparlinski, I. & Ward, T. 2003 *Recurrence sequences AMS mathematical surveys and monographs*, vol. 104. Providence, RI: American Mathematical Society.

Fomin, S. & Zelevinsky, A. 2002 The Laurent phenomenon. *Adv. Appl. Math.* **28**, 119–144. (doi:10.1006/aama.2001.0770)

Gale, D. 1991 The strange and surprising saga of the Somos sequences. *Math. Intell.* **13**, 40–42.

Hirota, R. 1972 Exact solution of the Korteweg-deVries equation for multiple collision of solitons. *Phys. Rev. Lett.* **27**, 1192–1194. (doi:10.1103/PhysRevLett.27.1192)

Hone, A. N. W. 2005 Elliptic curves and quadratic recurrence sequences. *Bull. Lond. Math. Soc.* **37**, 161–171. Corrigendum **38**, 741–742. (doi:10.1112/S0024609306018844)

Korepin, V. E., Bogoliubov, N. M. & Izergin, A. G. 1993 *Quantum inverse scattering method and correlation functions*. Cambridge, UK: Cambridge University Press.

Krichever, I., Lipan, O., Wiegmann, P. & Zabrodin, A. 1997 Quantum integrable systems and elliptic solutions of classical discrete nonlinear equations. *Commun. Math. Phys.* **188**, 267–304. (doi:10.1007/s002200050165)

Kuperberg, G. 1996 Another proof of the alternating sign matrix conjecture. *Int. Math. Res. Notices* **3**, 139–150. (doi:10.1155/S1073792896000128)

Noumi, M. 2004 *Painlevé equations through symmetry AMS translations of mathematical monographs*, vol. 223. Providence, RI: American Mathematical Society.

Mandl, F. 1971 *Statistical physics*. Chichester, UK: Wiley.

Mills, W. H., Robbins, D. P. & Rumsey, H. 1983 Alternating-sign matrices and descending plane partitions. *J. Comb. Theory Ser. A* **34**, 340–359. (doi:10.1016/0097-3165(83)90068-7)

[2] See http://www.math.wisc.edu/propp/somos.html.

van der Poorten, A. J. 2005 Elliptic curves and continued fractions. *J. Integer Seq.* **8**, 5. Article 05.2.5.

Propp, J. 2001 The many faces of alternating-sign matrices. *Disc. Math. Theoret. Comp. Sci. Proc.* **AA (DM-CCG)**, 43–58.

Speyer, D. 2004 Perfect matchings and the octahedron recurrence. math. CO/0402452.

Swart, C. S. 2003 Elliptic curves and related sequences. Ph.D. thesis, Royal Holloway, University of London.

Zeilberger, D. 1996*a* Proof of the alternating sign matrix conjecture. *Electron. J. Comb.* **3**, #R13.

Zeilberger, D. 1996*b* Proof of the refined alternating sign matrix conjecture. *NY J. Math.* **2**, 59–68.

AUTHOR PROFILE

Andrew N.W. Hone

Andrew Hone was born in London in 1972. He was an undergraduate at Balliol College, University of Oxford, where he was awarded a first class honours degree and the Prosser Prize in Mathematics in 1993. After obtaining a Ph.D. in Mathematics from the University of Edinburgh in 1997, he did his postdoctoral research in the Physics Department of Rome 3 University (funded by the Leverhulme Trust) and in the Pure Mathematics Department of the University of Adelaide (funded by the Australian Research Council). He is currently a senior lecturer in applied mathematics at the University of Kent, where he was appointed to a lectureship in 2000, and is a member of the EPSRC Peer Review College. His scientific interests include nonlinear systems and their applications in physics, biology, number theory and computing. At the time of writing he holds an EPSRC Springboard Fellowship in the Mathematical Sciences. He is married with three children, and enjoys practicing yoga, cooking and singing (but not simultaneously).