Why Eisenstein proved the Eisenstein criterion and why Schönemann discovered it first

David A. Cox

Department of Mathematics
Amherst College
Amherst, MA 01002, USA
dac@cs.amherst.edu

The Eisenstein irreducibility critierion is part of the training of every mathematician. I first learned the criterion as an undergraduate and, like many before me, was struck by its power and simplicity. This article will describe the unexpectedly rich history of the discovery of the Eisenstein criterion and in particular the role played by Theodor Schönemann.

For a statement of the criterion, we turn to Dorwart's 1935 article "Irreducibility of polynomials" in the American Mathematical Monthly [9]. As you might expect, he begins with Eisenstein:

> The earliest and probably best known irreducibility criterion is the Schoenemann-Eisenstein theorem:
>
> *If, in the integral polynomial*
>
> $$a_0x^n + a_1x^{n-1} + \cdots + a_n,$$
>
> *all of the coefficients except a_0 are divisible by a prime p, but a_n is not divisible by p^2, then the polynomial is irreducible.*

Here's our first surprise—Dorwart adds Schönemann's name in front of Eisenstein's. He then gives a classic application:

> An important application of this theorem is the proof of the irreducibility of the so-called "cyclotomic polynomial"
>
> $$f(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \cdots + 1,$$
>
> where p is prime.

If, instead of $f(x)$, we consider $f(x+1)$, where

$$f(x+1) = \frac{(x+1)^p - 1}{(x+1) - 1} = x^{p-1} + \binom{p}{1} x^{p-2} + \cdots + p,$$

the theorem is seen to apply directly, and the irreducibility of $f(x+1)$ implies the irreducibility of $f(x)$.

The combination "Schönemann-Eisenstein" (often "Schoenemann-Eisenstein") was common in the early 20th century. An exception is Dorrie's delightful book *Triumph der Mathematik*, published in 1933 [8], where he states the "Satz von Schoenemann." Another exception is van der Waerden's *Moderne Algebra* from 1930 [28], where we find the "Eisensteinscher Satz."[1]

Given the influence of van der Waerden's book on succeeding generations of textbook writers, we can see how Schönemann's name got dropped. But how did it get added in the first place? Equally important, how did Eisenstein's get added? And why both names? To answer these questions, we need to explore some 19th century number theory. This is a rich subject, so by necessity my treatment will be far from complete. I will instead focus on specific highlights to trace the development of these ideas. There will be numerous quotes (with translations when necessary) to illustrate how mathematics was done at the time and what it looked like. We begin with Gauss.

Gauss

Disquisitiones Arithmeticae [13], published in 1801, contains an amazing amount of mathematics. In particular, Gauss proves that when p is prime, the cyclotomic polynomial $x^{p-1}+\cdots+x+1$ is irreducible. His proof uses an explicit representation of the roots and is not easy. However, he also uses the following general result that relates irreducibility over $\mathbb{Z}$ to irreducibility over $\mathbb{Q}$:

42.

Si coëfficientes $A, B, C \ldots . N$; $a, b, c \ldots . n$ duarum functionum formae

$$x^m + Ax^{m-1} + Bx^{m-2} + Cx^{m-3} \ldots . + N \ldots . . . . (P)$$
$$x^\mu + ax^{\mu-1} + bx^{\mu-2} + cx^{\mu-3} \ldots . + n \ldots . . . . (Q)$$

omnes sunt rationales, neque vero omnes integri, productumque ex (P) et (Q)

$$= x^{m+\mu} + \mathfrak{A}x^{m+\mu-1} + \mathfrak{B}x^{m+\mu-2} + \text{etc.} + \mathfrak{Z}$$

omnes coëfficientes $\mathfrak{A}, \mathfrak{B} \ldots . \mathfrak{Z}$ integri esse nequeunt.

---

[1] This edition included a reference to Schönemann that was dropped in the 1937 second edition.

42.

If the coefficients $A, B, C \ldots . N$; $a, b, c \ldots . n$ of two functions of the form

$$x^m + Ax^{m-1} + Bx^{m-2} + Cx^{m-3} \ldots . + N \ldots . . . . (P)$$
$$x^\mu + ax^{\mu-1} + bx^{\mu-2} + cx^{\mu-3} \ldots . + n \ldots . . . . (Q)$$

are all rational and not all integers, and if the product of (P) and (Q)

$$= x^{m+\mu} + \mathfrak{A}x^{m+\mu-1} + \mathfrak{B}x^{m+\mu-2} + \text{etc.} + \mathfrak{Z}$$

then not all the coefficients $\mathfrak{A}, \mathfrak{B} \ldots . \mathfrak{Z}$ can be integers.

This is what we now call *Gauss's Lemma*. His proof is essentially the same that appears in abstract algebra texts, though he states the result in the contrapositive form and never uses the term "polynomial." Gauss also doesn't use the three dots $\cdots$ that are standard today.

Another major result of *Disquisitiones* is Gauss's proof that $x^n - 1 = 0$ is solvable by radicals. The modern approach to solvability by radicals allows the introduction of arbitrary roots of unity, which implies that $x^n - 1 = 0$ is trivially solvable. Gauss instead followed the inductive strategy pioneered by Lagrange, where one constructs the roots recursively using polynomials of strictly smaller degree that are solvable by radicals. In modern terms, this gives an explicit description of the intermediate fields of the extension

$$\mathbb{Q} \subseteq \mathbb{Q}(e^{2\pi i/p})$$

when p is prime. This has degree $p-1$ by the irreducibility of $x^{p-1} + \cdots + x + 1$. From here, Gauss obtains his wonderful result about dividing the circle into n equal arcs by straightedge and compass.

The second paragraph of Section VII of *Disquisitiones* begins with a famous passage:

> Ceterum principia theoriae, quam exponere aggredimur, multo latius patent, quam hic extenduntur. Namque non solum ad functiones circulares, sed pari successu ad multas alias functiones transscendentes applicari possunt, e.g. ad eas, quae ad integrali $\int \frac{dx}{\sqrt{(1-x^4)}}$ pendent, praetereaque etiam ad varia congrueniarum genera: sed quoniam de illis functionibus transscendentibus amplum opus peculiare paramus, de congruentiis autem in continuatione disquitionum arithmeticarum copiose tractabitur, hoc loco solas functiones circulares considerare visum est.

The principles of the theory we are going to explain actually extend much farther than we will indicate. For they can be applied not only to circular functions but just as well to other transcendental functions, e.g. to those which depend on the integral $\int \frac{dx}{\sqrt{(1-x^4)}}$ and also to various types of congruences. Since, however, we are preparing a large work on those transcendental functions and since we will treat congruences at

> length in the continuation of these *Disquisitiones*, we have decided to consider only circular functions here.

In this quote, the reference to circular functions is clear. But what about transcendental functions that depend on the integral $\int \frac{dx}{\sqrt{(1-x^4)}}$? Here, any 19th century mathematician would immediately think of the lemniscate $r^2 = \cos 2\theta$, whose arc length is $4\int_0^1 \frac{dx}{\sqrt{(1-x^4)}}$. This integral and its relation to the lemniscate were discovered by the Bernoulli brothers in the late 17th century and played a key role in the development of elliptic integrals by Fagnano, Euler, and Legendre in the 18th century. Gauss's "large work" on these functions never appeared, though fragments found after Gauss's death contain some astonishing mathematics (see [3]).

The quote also mentions "various types of congruences" that will be discussed "in the continuation of these *Disquisitiones*." The published version of *Disquisitiones* had seven sections, but Gauss drafted an eighth section, *Disquisitiones generales de congruentiis*, that studied polynomial congruences $f(x) \equiv 0 \bmod p$, where $f \in \mathbb{Z}[x]$ and p is prime (see pp. 212–242 of [15, Vol. II] or pp. 602–629 of the German version of [13]). In modern terms, Gauss is studying the polynomial ring $\mathbb{F}_p[x]$. Here are some of his results:

- The existence and uniqueness of factorization of polynomials modulo p.
- A determination of the number (n) of monic irreducible polynomials modulo p. His result is

$$n(n) = p^n - \sum p^{\frac{n}{a}} + \sum p^{\frac{n}{ab}} - \sum p^{\frac{n}{abc}} \text{ etc.}$$

  where the sum $\sum p^{\frac{n}{a}}$ (resp. $\sum p^{\frac{n}{ab}}$) is over all distinct prime factors (resp. pairs of distinct prime factors) of n, and similarly for the remaining terms in the formula.

Gauss also had a theory of finite fields, though his approach is not easy for the modern reader because of his reluctance to introduce roots of polynomial congruences. Here is what Gauss says about the congruence $\xi \equiv 0 \bmod p$, where ξ is a polynomial with integer coefficients:

> ... ad hinc hihil obstat, quominus ξ in factores duram, trium pluriumve dimensionum resolvi possit, unde radices quasi *imaginariae* illi attribui possint. Revera, si simili licentia, quam recentiores mathematici usurparunt, uti talesque quantitates imaginarias introducere voluissemus, omnes nostras disquisitiones sequentes incomparabiter contrahere licuisset; ...

> ... but nothing prevents us from decomposing ξ, nevertheless, into factors of two, three or more dimensions [degrees], whereupon, in some sense, *imaginary* roots could be attributed to them. Indeed, we could have shortened incomparably all our following investigations, had we wanted to introduce such imaginary quantities by taking the same liberty some more recent mathematicians have taken; ...

Over the complex numbers, Gauss was the first to prove the existence of roots of polynomials. He was critical of those who simply assumed that roots exist, so he clearly wasn't going to assume that congruences of higher degree have solutions.

We refer the reader to [11] for a fuller account of Gauss's work on finite fields. Unfortunately, none of this was available until after Gauss's death in 1855. In particular, Schönemann was unaware of these developments when he rediscovered many of Gauss's results in the 1840s.

Abel

Gauss's cryptic comments about the lemniscate in *Disquisitiones* had a profound influence on Abel. He developed the theory of elliptic functions (as did Jacobi), based on the equation

$$y^2 = (1 - c^2x^2)(1 + e^2x^2), \tag{1}$$

and his elliptic functions were inverse functions to the elliptic integral

$$\int \frac{dx}{y} = \int \frac{dx}{\sqrt{(1 - c^2x^2)(1 + e^2x^2)}}.$$

When $e = c = 1$, we get the integral associated with the lemniscate.

The division problem for elliptic integrals goes back to Fagnano and Euler. Later in the article we will quote a letter from Eisenstein to Gauss, where he expressed the m-division problem in the lemniscatic case as the "algebraic integral of the equation

$$\int_0 dy/\sqrt{1 - y^4} = m \int_0 dx/\sqrt{1 - x^4}." \tag{2}$$

In modern language, we are talking about division points on elliptic curves, and the "algebraic integral" produces a polynomial P_m of degree m^2 whose solutions give (roughly speaking) the m-division points on the associated elliptic curve defined by (1). We will say more about the polynomial P_m and the equation (2) when we discuss Eisenstein.

For Abel and his contemporaries, a central question was whether polynomial equations such as $P_m(x) = 0$ were "solvable algebraically", which these days means solvable by radicals. Abel was uniquely qualified to pose this question, since just four years earlier he had proved that the general quintic was not solvable by radicals.

In his great paper *Recherches sur les functions elliptiques* [1, pp. 263–388], printed in volumes 2 and 3 of Crelle's journal[2] in 1827 and 1828, Abel considers the equation $P_{2n+1} = 0$ coming from $(2n + 1)$-division points on the elliptic curve (1). Here is what he has to say about this equation:

> Donc en dernier lieu la résolution d'équation $P_{2n+1} = 0$ est reduite á celle d'une seule équation du degré $2n + 2$; mais *en général* cette équation ne paraît pas être résoluble algébriquement. Néamoins on peut las résoudre complètement dans plusiers cas particuliers, par exemple, lorsque $e = c$, $e = c\sqrt{3}$, $e = c(2 \pm \sqrt{3})$ etc. Dans le course de ce mémoire

[2]The *Journal für die reine und angewandte Mathematik*, founded by August Leopold Crelle in 1826.

je m'occuperai de ces cas, dont le premier surtout est remarquable, tant pour la simplicité de la solution, que par sa belle application dans la géométrie.

En effet entre autres théorèmes je suis parvenu à celui-ci:

"On peut diviser la circonférence entière de la lemniscate en m parties égales *par la regle et le compas seuls*, si m est de la forme 2^n ou $2^n + 1$, ce dernier nombre étant en même temps premier; ou bien si m est un product de plusiers nombres de ces deux formes."

Ce théorème est, comme on le voit, précisément le même que celui de M. *Gauss*, relativement au cercle.

Thus finally the solution of the equation $P_{2n+1} = 0$ is reduced to a single equation of degree $2n + 2$; but *in general* this equation does not appear to be solvable algebraically. Nevertheless one can solve it completely in many particular cases, for example, when $e = c$, $e = c\sqrt{3}$, $e = c(2 \pm \sqrt{3})$ etc. In the course of this memoir I will concern myself with these cases, of which the first is especially remarkable, both for the simplicity of its solution, as well as by its beautiful application to geometry.

Indeed among other theorems I arrived at this one:

"One can divide the entire circumference of the lemniscate into m parts *by ruler and compass only*, if m is of the form 2^n or $2^n + 1$, the last number being at the same time prime, or if m is a product of several numbers of these two forms."

This theorem is, as one sees, precisely the same as that of M. *Gauss*, relative to the circle.

The reduction to an equation of degree $2n + 2$ was done by classical methods of Lagrange. Besides the mind-blowing result about the lemniscate ($e = c$), other aspects of this quote deserve comment:

- The cases $e = c$, $e = c\sqrt{3}$, $e = c(2 \pm \sqrt{3})$ etc. that Abel can solve by radicals correspond to elliptic curves with complex multiplication (see [5] for an introduction). Abel was the first to identify this important class of elliptic curves.

- From a modern standpoint, division points of elliptic curves with complex multiplication generate Abelian extensions and hence have Abelian Galois groups. Since Abelian groups are solvable, Galois theory implies that the extensions are solvable by radicals.

- When the curve doesn't have complex multiplication, Abel was more cautious: they do "not appear to be solvable algebraically." By deep work of Serre on Galois representations of elliptic curves [27], we now know that with at most finitely many exceptions, these equations aren't solvable by radicals.

Again we are in the presence of remarkably rich mathematics.

Abel thought deeply about why his equations $P_{2n+1} = 0$ were solvable by radicals when the curve has complex multiplication. He realized that the underlying reason was the structure of the roots and how they relate to each other. His general result appears in his *Mémoire sur une classe particulière d'équations résolubles algébriquement* [1, pp. 478–507], which was published in Crelle's journal in 1829. The article begins:

> Quoique la résolution algébrique des équations ne soit possible en général, il y a néamoins des équations particulières des tous les degrés qui admettant une telle résolution. Telles sont par example les équations de la forme $x^n - 1 = 0$. La résolution de ces équations est fondée sur certaines relations qui existent entre les racines.

> Although the algebraic solution of equations is not possible in general, there are nevertheless particular equations of all degrees which admit such a solution. Examples are the equations of the form $x^n - 1 = 0$. The solution of these equations is based on certain relations that exist among the roots.

The first sentence refers to Abel's result on the unsolvability of the general quintic and the solution of $x^n - 1 = 0$ described by Gauss in *Disquisitiones*. To give the reader a sense of what he means by "relations that exist among the roots," Abel takes a prime n and considers the cyclotomic equation $x^{n-1} + \cdots + x + 1 = 0$. Define the polynomial $\theta(x) = x^\alpha$, where α is a primitive root modulo n. Then the roots are given by

$$x,\ \theta(x) = x^\alpha,\ \theta^2(x) = x^{\alpha^2},\ \theta^3(x) = x^{\alpha^3}, \ldots,\ \theta^{n-2}(x) = x^{\alpha^{n-2}},\ \text{where } \theta^{n-1}(x) = x.$$

Abel goes on to say that the same property appears in a certain class of equations that he found in the theory of elliptic functions. He then states the main theorem of the paper:

> En général j'ai démontré le thèoréme suivant:
>
> „Si les racines d'une équation d'un degré quelquonque sont liées entre elles de telle sorte, que *toutes* ces racines puissent être exprimées rationnellement au moyen de l'une d'elles, que nous désignerons par x; si de plus, en désignant par θx, $\theta_1 x$ deux autres racines quelquonques, on a
>
> $$\theta\theta_1 x = \theta_1 \theta x,$$
>
> l'équation dont il s'agit sera toujours résoluble algébriquement. ... "

> In general I have proved the following theorem:
>
> „If the roots of an equation of arbitrary degree are related among themselves in such a way, that *all* of the roots can be rationally expressed

> in terms of one of them, which we designate by x; if in addition, designating by θx, $\theta_1 x$ two other arbitrary roots, one has
>
> $$\theta\theta_1 x = \theta_1\theta x,$$
>
> the equation in question is always solvable algebraicially. . . . "

Abel's "classe particulière" consist of all polynomials that satisfy the hypothesis of his theorem. To see what this says in modern terms, let $K \subseteq L$ be a Galois extension with primitive element α. For each element σ_i of the Galois group $\mathrm{Gal}(L/K)$, there is a polynomial $\theta_i(x) \in K[x]$ such that $\sigma_i(\alpha) = \theta_i(\alpha)$. Then one easily computes that

$$\sigma_i\sigma_j(\alpha) = \theta_j(\theta_i(\alpha)).$$

The switch of indices is correct—you should check why. Since σ_i is determined by its value on α,

$$\sigma_i\sigma_j = \sigma_j\sigma_i \iff \theta_j(\theta_i(\alpha)) = \theta_i(\theta_j(\alpha)).$$

Since the $\theta_i(\alpha)$ are the roots of the minimal polynomial $f(x)$ of α over K, we see that $f(x)$ is in the "classe particulière" if and only if $\mathrm{Gal}(L/K)$ is commutative. As noted earlier, this means that the Galois group is solvable, so that $f(x)$ is solvable by radicals by Galois theory.

Besides proving his general theorem, Abel intended to give two applications:

> Après avoir exposé cette théorie en général, je l'appliquerai aux fonctions circulares et elliptiques.
>
> ---
>
> After having developed this theory in general, I will apply it to circular and elliptic functions.

The version published in Crelle's journal has a section on circular functions, but ends with the following footnote by Crelle:

> *) L'auteur de ce mémoire donnera dans une autre occasion des applications aux fonctions elliptiques.
>
> ---
>
> *) The author of this memoir will give applications to elliptic functions on another occasion.

Alas, Abel died shortly after this article appeared.

After Abel

Abel's "classe particulière" had an important influence on Kronecker, Jordan, and Weber. Specifically:

- In 1853, Kronecker [18, Vol. IV, p. 11] defined $f(x) = 0$ to be "Abelian" provided it has roots $x, \theta(x), \ldots, \theta^{n-1}(x)$, $x = \theta^n(x)$. Here, as for Abel, θ is a rational function. This special case of Abel's "classe particulière" corresponds to polynomials with cyclic Galois groups.

- In 1870, Jordan [17, p. 287] defined $f(x) = 0$ to be "Abelian" in terms of its Galois group:

  > Nous appellerons donc *équations abéliennes* toutes celles dont le groupe ne contient que les substitutions échangeables entre elles.
  >
  > ---
  >
  > We thus call *Abelian equations* all of those whose group only contains substitutions that are exchangeable among each other.

  Here, "exchangeable" is Jordan's way of saying "commutative." He then proves [17, p. 288] that for irreducible equations, his definition is equivalent to Abel's "classe particulière."

- The first two volumes of Weber's monumental *Lehrbuch der Algebra* were published in 1894 and 1896. He gives the name "Abelian" to Abel's "classe particulière" [29, Vol. I, p. 576] and later defines a commutative group to be "Abelian" [29, Vol. II, p. 6]. As far as I know, this is the first appearence of the term "Abelian group" in the modern sense.[3]

The definition of "Abelian group" given in introductory algebra courses seems so simple. But in the background is a rich history involving Gauss, Abel, the leminiscate, elliptic functions, complex multiplication, and solvability by radicals.

Galois

One of the few papers published during Galois's lifetime was *Sur la théorie des nombres*, appearing in 1830 in the *Bulletin des sciences mathématiques de Ferussac* [12, pp. 113–127]. This paper develops the theory of finite fields. Galois begins with a congruence $F(x) \equiv 0 \bmod p$, or as he writes it, $Fx = 0$, where $F(x)$ is irreducible modulo p. Then he considers the roots:

> ... Il faut donc regarder les racines de cette congruence comme des espèces de symboles imaginaires ...
>
> ---
>
> ... One must regard the roots of this congruence as a kind of imaginary symbol ...

It is clear that Gauss would not approve. Galois used the symbol i to denote a root of $F(x) \equiv 0 \bmod p$, and he showed that the numbers

$$a + a_1 i + a_2 i^2 + \cdots + a_{\nu-1} i^{\nu-1},$$

where $\nu = \deg(F)$ and $a, a_1, \ldots, a_{\nu-1}$ are integers modulo p, form a finite field with p^ν elements. Galois went on to develop a complete theory of finite fields. The reason he needed finite fields is connected with his deep work on the structure of solvable primitive permutation groups (see the Historical Notes to [4, §14.3]).

We will not say more about Galois and finite fields, because Schönemann was not aware of Galois's 1830 paper when he began his own study of congruences and finite fields in the early 1840s.

[3] In 1870, Jordan used the term "groupe abélien" to refer to a group closely related to a symplectic group over a finite field [17, Livre II, §VIII].

Schönemann

Unlike the other people mentioned so far, Theodor Schönemann is not a familiar name. He has no biography at the MacTutor History of Mathematics archive [21]. According to the *Allgemeine Deustsche Biographie* [2, Vol. 32, pp. 293–294], Schönemann lived from 1812 to 1868 and was educated in Königsberg and Berlin under the guidance of Jacobi and Steiner. He got his doctorate in 1842 and became Oberlehrer and eventually Professor at a gymnasium in Brandenburg an der Havel. Lemmermeyer's book [19] includes several references to Schönemann's work in number theory, and some of his results are mentioned in Dickson's classic *History of the Theory of Numbers* [7], especially in the chapter on higher congruences in Volume I.

For us, Schönemann's most important work is a long paper printed in two parts in Crelle's journal in 1845 and 1846. The first part [24], consisting of §1–§50, appeared as *Grundzüge einer allgemeinen Theorie der höhern Congruenzen, deren Modul eine reelle Primzahl ist* (*Foundations of a general theory of higher congruences, whose modulus is a real prime number*). In the preface, Schönemann refers to Gauss:

> Der berühmte Verfasser der *Disquisitiones Arithmeticae* hatte für den achten abschnitt seines Werkes eine allgemeine Theorie der höhern Congruenzen bestimmt. Da indessen dieser achte Abschnitt nicht erscheinen, und auch, so viel ich weiss, über diesen Gegenstand sonst nichts von dem Herrn Verfasser bekannt gemacht oder nur bestimmt angedeutet worden ist ...

> The famous author of *Disquisitiones Arithmeticae* had intended a general theory of higher congruences for Section Eight of his work. Since, however, this Section Eight did not appear, and also, as far as I know, the author did not publish anything on this subject, nor indicate anything precisely ...

Schönemann suspects that he may have been scooped by Gauss, but is not worried:

> ... würde mich über die Einbusse der ersten Enteckung das Bewusstsein schadlos halten, auf selbständigem Wege mit dem Streben eines solchen Geistes zusammengetroffen zu sein.

> ... the loss of first discovery would be compensated by my knowing of having met in my own and independent way such a spirit.

Indeed, Schönemann had been scooped by both Gauss and Galois. Hence we should change "a spirit" to "spirits" in the quote, in which case the sentiment is even more apt.

Similar to what Gauss did, Schönemann gave a careful treatement of polynomials modulo p, including unique factorization. But then, in §14, he did something different. Let $f(x) \in \mathbb{Z}[x]$ be monic of degree n and irreducible modulo p, and let $\alpha \in \mathbb{C}$ be a root of $f(x)$ (Gauss would approve of this root). Then, given polynomials $\varphi, \psi \in \mathbb{Z}[x]$, Schönemann defined $\varphi(\alpha)$ and $\psi(\alpha)$ to be *congruent modulo* (p, α) if $\varphi(\alpha) = \psi(\alpha) + pR(\alpha)$ for some $R \in \mathbb{Z}[x]$. He also proved:

- The "allgemeine Form eines kleines Restes" ("general form of a smallest remainder") is $a_0\alpha^{n-1} + a_1\alpha^{n-2} + \cdots + a_{n-1}$, where $a_i \in \{0, \ldots, p-1\}$. This gives the finite field $\mathbb{F}_{p^n}$.
- The elements of $\mathbb{F}_{p^n}$ are the roots of $x^{p^n} - x \equiv 0 \bmod (p, \alpha)$.
- $f(x) \equiv (x-\alpha)(x-\alpha^p)\cdots(x-\alpha^{p^{n-1}}) \bmod (p, \alpha)$. Thus $\mathbb{F}_{p^n}$ is the splitting field of $f(x) \bmod (p, \alpha)$. Also, the Galois group (generated by Frobenius) is implicit in this factorization of f.

The first part of Schönemann's paper culminates in §50 with a lovely proof of the irreducibility of $\Phi_p(x) = x^{p-1} + \cdots + x + 1$. We will give the proof in modern notation. Pick a prime $\ell \neq p$ and consider the prime factorization

$$\Phi_p(x) \equiv f_1(x)\cdots f_r(x) \bmod \ell.$$

where the f_i are irreducible modulo ℓ. Standard properties of finite fields imply that

$$\begin{aligned} \deg(f_i) &= \text{the minimum } n \text{ such that } \mathbb{F}_{\ell^n}^* \text{ has an element of order } p \\ &= \text{the minimum } n \text{ such that } \ell^n \equiv 1 \bmod p \\ &= \text{the order of the congruence class of } \ell \text{ in } (\mathbb{Z}/p\mathbb{Z})^*. \end{aligned} \tag{3}$$

We leave this as a fun exercise for the reader. By Dirichlet's theorem on primes in arithmetic progressions (proved just a few years before Schönemann's paper), every congruence class modulo p contains a prime. In particular, the congruence class of a primitive root contains a prime ℓ. A primitive root modulo p gives a congruence class of order $p-1$ in $(\mathbb{Z}/p\mathbb{Z})^*$, so that $n = p-1$ in (3) for this choice of ℓ. This implies that $\Phi_p(x)$ is irreducible modulo ℓ and hence irreducible over $\mathbb{Z}$. Then $\Phi_p(x)$ is irreducible over $\mathbb{Q}$ by Gauss's Lemma.

This proof is simpler than Gauss's, though it does require knowledge of finite fields. The use of the auxiliary prime ℓ is especially elegant. When I studied Grothendieck-style algebraic geometry as a graduate student in the 1970s, I was always happy when a proof picked a prime different from the residue characteristic. This seemed so modern and cutting-edge. Little did I realize that Schönemann had used the same idea 120 years earlier.

The second part of Schönemann's paper [25], titled *Von denjenigen Moduln, welche Potenzen von Primzahlen sind* (*On those moduli, which are powers of prime numbers*), consists of §51–§66. In this paper, Schönemann considered the factorization of polynomials modulo p^m, and in particular, how the factorization changes as m varies. One of his major results, in §59, is what we now call *Hensel's Lemma*:

> Lehrsatz. Ist irgend ein einfacher Ausdruck von x nach dem Modul p in zwei einfache Factoren zerlegbar, die nach demselben Mdoul keinen gemeinsaftlichen Divisor haben: so ist dieser Ausdruck auch nach dem Modul p^m, ***aber nur auf einer Weise***, in zwei Factoren zerlegbar, welche jenen beiden ersten nach dem modul p congruent sind.

> Lemma. If any monic polynomial of x can be factored modulo p into two monic factors, which for this modulus have no common divisor: then this polynomial can be factored modulo p^m, ***in a unique manner***, into two factors, which are congruent to those first two factors modulo p.[4]

(Here, "einfacher Ausdruck von x" means a monic polynomial of x.) As a consequence, when an irreducible polynomial modulo p^m is reduced modulo p, the result must be a power of an irreducible polynomial modulo p. In §61, Schönemann asks about the converse:

> Aufgabe. Zu untersuchen, ob die Potenz eines nach dem Modul p irreductibeln Ausdrucks, nach dem Modul p^m irreducibel sei, oder nicht.

> Problem. To investigate, whether the power of irreducible polynomial modulo p is or is not irreducible modulo p^m.

An especially simple example is $(x-a)^n$, and for a polynomial congruent to $(x-a)^n$ modulo p, the first place to check for irreducibility is modulo p^2. Here is Schönemann's answer:

> ... man darf daher den Satz aussprechen: ***dass $(x-a)^n+pFx$ nach dem Modul p^2 irreductibel sein wurde, wenn Fx nach dem Modul p nicht den Factor $x-a$ in sich schliesst.*** ...

> ... hence one may state the theorem: ***that $(x-a)^n+pFx$ is irreducible modulo p^2, when the factor $x-a$ is not contained in Fx modulo p.*** ...

As stated, this is not quite correct—one needs to assume that $\deg(F) \le n$.[5] Since $x-a$ divides $F(x)$ modulo p if and only if $F(a) \equiv 0 \bmod p$, we can state Schönemann's result as follows.

> **Schönemann's Criterion.** Let $f(x) \in \mathbb{Z}[x]$ have degree $n > 0$ and assume that there is a prime p and an integer a such that
>
> $$f(x) = (x-a)^n + pF(x),$$
>
> If $F(a) \not\equiv 0 \bmod p$, then $f(x)$ is irreducible modulo p^2.

The proof is not difficult (assume $(x-a)^n+pF(x)$ factors modulo p^2, reduce modulo p and use unique factorization in $\mathbb{F}_p[x]$) and is left to the reader.

The pleasant surprise is that this result implies the Eisenstein criterion. To see why, suppose that $f(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n$ satisfies the hypothesis of

[4] The uniqueness assertion enables us to take the limit as $m \to \infty$, giving a factorization over the p-adic integers that reduces to the given factorization modulo p. This version of Hensel's Lemma is stated in [16, Thm. 3.4.6], and the discussion on [16, p. 72] explains how this relates to the more common version of Hensel's Lemma, which asserts that for $f(x) \in \mathbb{Z}_p[x]$, a solution of $f(x) \equiv 0 \bmod p$ of multiplicity one lifts to a solution of $f(x) = 0$ in $\mathbb{Z}_p$.

[5] For example, let $F(x) = x^3 - p^2x + 1$. Then $x^2 + pF(x) = (px+1)(x^2 - p^2x + p)$, yet x does not divide $F(x)$ modulo p.

the Eisentstein criterion. Multiplying by a suitable integer, we may assume $a_0 \equiv 1 \bmod p$. This allows us to write $f(x) = x^n + pF(x)$. Note also that $F(0) \not\equiv 0 \bmod p$ since p^2 does not divide a_n. Then $f(x)$ is irreducible modulo p^2 by Schönemann's criterion. This implies irreducibility over $\mathbb{Z}$ and hence (via Gauss's Lemma) over $\mathbb{Q}$.

As you might expect, Schönemann immediately applies his irreducibility criterion to a familiar polynomial:

> Wenden wir das erhaltene Resultat auf der Ausdruck $\dfrac{x^n-1}{x-1}$ an, wo n eine Primzahl bedeutet. Es ist für diesen Fall $x^n - 1 \equiv (x-1)^n$ (mod. n), und man erhält also
>
> $$\frac{x^n-1}{x-1} = x^{n-1} + x^{n-2} + \cdots + x + 1 = (x-1)^{n-1} + nFx.$$
>
> Für $x = 1$ erhält man $n = nF(1)$ und daher $F(1) = 1$, und nicht $\equiv 0$ (mod. n). ***Hieraus folgt, dass $\dfrac{x^n-1}{x-1}$ nach dem Modul n^2 stets irreductibel ist, wenn n eine Primzahl bedeutet; mithin muss dieser Ausdruck gewiss in algebraischer Beziehung irreductibel sein.***
>
> Die Leichtigkeit des Beweises dieses Satzes is auffallend, da derselbe in den „Disquisitiones" mit einem viel grössern Aufwande von Scharfsinn, und dennoch viel umständlicher geführt is. (Vergl. §. 50. Zus. 2.)

> Let us apply the result just obtained to the polynomial $\dfrac{x^n-1}{x-1}$, where n denotes a prime number. In this case $x^n - 1 \equiv (x-1)^n$ (mod. n), and one thus obtains
>
> $$\frac{x^n-1}{x-1} = x^{n-1} + x^{n-2} + \cdots + x + 1 = (x-1)^{n-1} + nFx.$$
>
> For $x = 1$ one obtains $n = nF(1)$ and thus $F(1) = 1$, and not $\equiv 0$ (mod. n). ***From this, it follows that $\dfrac{x^n-1}{x-1}$ is always irreducible modulo n^2, if n is a prime number; hence, this expression is certainly irreducible in the algebraic sense.***
>
> The ease of proof of this theorem is striking, because the proof in „Disquisitiones" requires much greater cleverness, and is much more elaborate. (See §. 50. Rem. 2.)

This proves the irreducibility of $x^{n-1} + \cdots + x + 1$ without the change of variable $x \leftrightarrow x + 1$ needed when one uses the Eisenstein criterion. Schönemann is clearly pleased that his proof is so much simpler than Gauss's. (The parenthetical comment at the end of the quote refers to Schönemann's earlier proof of irreducibility given in §50 of the first part of his article.)

Schönemann's criterion is lovely but is unknown to most mathematicians. So how did I learn about it? My book on Galois theory [4] gives Eisenstein's proof of Abel's theorem on the lemniscate. In trying to understand Eisenstein, I looked at Lemmermeyer's wonderful book *Reciprocity Laws*, where I found a reference to Schönemann. When I tried to read Schönemann's paper, I couldn't find the Eisenstein criterion, in part because the paper is long and my German isn't very good, and in part because I was looking for Eisenstein's version, not Schönemann's. I looked back at Lemmermeyer's book and noticed that Lemmermeyer thanked Michael Filaseta for the Schönemann reference. I wrote to Filaseta, who replied that Schönemann proved a criterion for a polynomial to be irreducible modulo p^2. This quickly led me to §61 of the article, which is where Schönemann states his result.

Back to Gauss

Besides discovering the Eisenstein criterion before Eisenstein, Schönemann also discovered Hensel's Lemma before Hensel. Unfortunately, Schönemann and Hensel were both scooped by Gauss. In his draft of the unpublished eighth section of *Disquisitiones* (p. 627 of the German version of [13] or p. 238 of [15, Vol. II]), Gauss takes a polynomial X with integer coefficients and studies its behavior modulo p and p^2:

> Problema. *Si functio X secundum modulum p in factores inter se primos ξ, ξ', ξ'' etc. sit resoluta, X secundum modulum pp in similes factores Ξ, Ξ', Ξ'' etc. resolvere ita, ut sit*
>
> $$\xi \equiv \Xi,\ \xi' \equiv \Xi',\ \xi'' \equiv \Xi'',\ \text{etc. } (\text{mod.}p)$$
>
> ---
>
> Problem. *If the polynomial X decomposes modulo p into mutually prime factors ξ, ξ', ξ'' etc., then similarly X decomposes modulo p^2 into factors Ξ, Ξ', Ξ'' etc. such that*
>
> $$\xi \equiv \Xi,\ \xi' \equiv \Xi',\ \xi'' \equiv \Xi'',\ \text{etc. } (\text{mod.}p)$$

Gauss proves this and then explains how the same principle applies modulo p^k for any k. His "Problema" is weaker than Schönemann's "Lemma" because it doesn't say that the lifted factorization is unique. So what Gauss really proved was a "proto-Hensel's Lemma." Nevertheless, Gauss was sufficiently pleased with this result that he recorded it in his famous mathematicial diary [14]. Here is entry 79, dated September 9, 1797:

> Principia detexi, ad quae congruentiarum secundum modulos multiplices resolutio ad congruentias secundum modulum linearem reducitur.
>
> ---
>
> Beginning to uncover principles, by which the resolution of congruences according to multiple moduli is reduced to congruences according to linear moduli.

Here, "resolution of congruences according to multiple moduli" means factoring polynomials modulo p^k, and similarly "congruences according to linear moduli" means working modulo p. This reading of Gauss's entry is carefully justified in [11].

Besides this elementary version of Hensel's Lemma, Gauss also considered the case when the factors modulo p are not distinct. For example, the congruence $X \equiv X'\,(x-a)^m \bmod p$ appears near the end of Gauss's draft of the eighth section. Had he pursued this, it is quite possible that he would have followed the same path as Schönemann and discovered the Eisenstein criterion. But instead, the draft ends abruptly in the middle of a congruence: the last thing Gauss wrote was

$$0 \equiv$$

As with many other projects, Gauss never returned to finish *Disquisitiones generales de congruentiis*. It came to light only after being published in 1863 in the second volume of his collected works, and today is still overshadowed by its more famous sibling, *Disquisitiones Arithmeticae*.

After Schönemann

Although Schönemann was scooped on finite fields by Gauss and Galois, he went beyond both of them in one significant way: he gave a rigorous description of the elements of a finite field. Gauss would have been very critical of the roots of congruences so blithely assumed by Galois. Schönemann, by starting with a complex root $\alpha \in \mathbb{C}$ of a monic polynomial $f(x)$ that is irreducible modulo p, constructed the field whose modern description is the quotient ring $\mathbb{Z}[\alpha]/\langle p\rangle$, where $\langle p\rangle$ is the ideal of $\mathbb{Z}[\alpha]$ generated by p.

Schönemann's construction, while rigorous, is not purely algebraic, since it depends on the root $\alpha \in \mathbb{C}$ of $f(x)$. This uses the Fundamental Theorem of Algebra, which in spite of its name is a theorem in analysis since it ultimately depends on the completeness of the real numbers. Of course, these days, we would express $\mathbb{Z}[\alpha]$ via the isomorphism

$$\mathbb{Z}[X]/\langle f(X)\rangle \simeq \mathbb{Z}[\alpha]$$

induced by $X \mapsto \alpha$, so that our finite field is

$$\mathbb{Z}[X]/\langle p, f(X)\rangle \simeq \mathbb{Z}[\alpha]/\langle p\rangle.$$

This algebraic version of finite fields was made explicit by Dedekind in his 1857 paper *Abriß einer Theorie der höheren Kongruenzen in bezug auf einen reellen Primzahl-Modulus* (*Outline of a theory of higher congruences for a real prime modulus*) [6]. Dedekind begins the paper by noting that the subject was initiated by Gauss and had been studied by Galois and Schönemann. Dedekind was at the time unaware of the full power of what Gauss had done, though later he became the editior in charge of publishing *Disquisitiones generales de congruentiis* in Volume II of Gauss's collected works in 1863.

Dedekind's construction is essentially what we did above with the quotient ring $\mathbb{Z}[X]/\langle p, f(X)\rangle$, $f(X)$ irreducible modulo p, though Dedekind was writing before the concept of quotient ring was fully established. Nevertheless, he shows that this

is a finite field with p^n elements, $n = \deg(f)$. For much of the 19th century, "finite field" meant this object. It has the advantage of being easy to compute with (even today, computers represent finite fields this way), but mathematically, it depends on the choice of $f(X)$ and hence is intrinsically non-canonical.

One of the first fully abstract definitions of finite field was given by E. H. Moore, whose paper [20] appeared in the proceedings of the 1893 international congress of mathematicians. Here is his definition:

> Suppose that we have a system of s distinct symbols or *marks**, $\mu_1, \ldots, \mu_s$ (s being some finite positive integer), and suppose that these marks may be combined by the four fundamental operations of algebra—addition, subtraction, multiplication, and division—these operations being subject to the ordinary abstract *operational identities* of algebra
>
> $$\mu_i + \mu_j = \mu_j + \mu_i;\ \mu_i\mu_j = \mu_j\mu_i;\ (\mu_i + \mu_j)\mu_k = \mu_i\mu_k + \mu_j\mu_k;\ \text{etc.}$$
>
> and that when the marks are so combined the results of these operations are in every case uniquely determined and belong to the system of marks. Such a system we shall call a *field of order* s, using the notation $F[s]$.
>
> We are led at once to seek *To determine all such fields of order* s, $F[s]$.

The words "system" and "marks" indicate that Moore was writing before the language of set theory was standardized. Moore went on to show that his definition was equivalent to the Dedekind-style representation of a finite field. So in 1893 we finally have a modern theory of finite fields.

The word "marks" in Moore's quote has an the asterisk that leads to the following footnote:

> * It is necessary that all *quantitative* ideas should be excluded from the concept *marks*. Note that the signs $>, <$ do not occur in the theory.

Moore was writing for a mathematically sophisticated audience, but he didn't assume that they had the apparatus of set theory in their heads—his footnote was intended to help them understand the abstract nature of what he was saying. This is something we should keep in mind when we teach abstract algebra to undergraduates.

Eisenstein

We finally get to Eisenstein, whose work on Abel's theorem on the lemniscate culminated in a long two-part paper in Crelle's journal in 1850 [10, pp. 536–619]. Eisenstein used Abel's notation φ for the lemniscatic function, so that

$$r = \varphi(s) \iff s = \int_0^r \frac{dr}{\sqrt{1 - r^4}}. \tag{4}$$

(We follow the 19th century practice of using the same letter for the variable and limit of integration.) In this equation, $0 \le r \le 1$ corresponds to $0 \le s \le \varpi =$

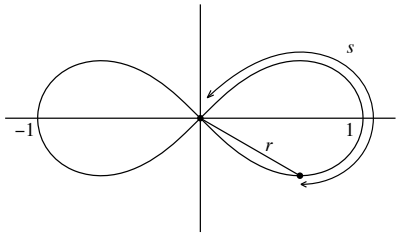


Figure 1: Arc length on the lemniscate

$\int_0^1 \frac{dr}{\sqrt{1-r^4}}$. Then define φ for $s \geq 0$ by considering the point on the lemniscate whose cumulative arc length is s when we start from the origin and follow the branch of the lemniscate in the first quadrant. See Figure 1. An arc length calculation shows that s and the radius r are related by the equation

$$r = \varphi(s)$$

(see [4, §15.2]). In particular, ϖ is one-fourth of the total arc length of the lemniscate, so that $\varphi(\varpi) = 1$ and $\varphi(2\varpi) = 0$. Hence, for any positive integer m, $r = \varphi(k \cdot 2\varpi/m)$, $k = 1, \ldots, m$, gives the radii of the points that divide the right half of the lemniscate into m equal pieces.

The change of variables $r = iu$ in (4) led Abel to define $\varphi(is) = i\varphi(s)$, and then Euler's addition law makes $\varphi(z) = \varphi(s + it)$ into a function of a complex variable $z \in \mathbb{C}$.[6] A key observation is that for any Gaussian integer $m \in \mathbb{Z}[i]$, $\varphi(mz)$ is a rational function of $\varphi(z)$ and its derivative $\varphi'(z)$. This is what complex multiplication means for the lemniscatic function φ.

When $m = a + ib$ is *odd* Gaussian integer, meaning that $a + b$ is odd, $\varphi(mz)$ is a rational function of $\varphi(z)$ alone. Here, one can find polynomials $U(x), V(x)$ with coefficients in $\mathbb{Z}[i]$ such that

$$y = \varphi(mz) \text{ is related to } x = \varphi(z)$$

via

$$y = \frac{U(x)}{V(x)} = \frac{A_0 x + A_1 x^5 + \cdots + A_{(N(m)-1)/4} x^{N(m)}}{1 + B_1 x^4 + \cdots + B_{(N(m)-1)/4} x^{N(m)-1}} \tag{5}$$

where $N(m) = a^2 + b^2$. See [4, Thm. 15.4.4] for a proof.

When m is an ordinary odd integer, we know that $r = \varphi(k \cdot 2\varpi/m)$ gives m-division points on the lemniscate. Setting

$$y = \varphi(m \cdot (k \cdot 2\varpi/m)) = \varphi(k \cdot 2\varpi) = 0 \text{ and } x = \varphi(k \cdot 2\varpi/m) = r$$

into (5), we see that

$$0 = \frac{U(r)}{V(r)}, \text{ hence } U(r) = 0.$$

[6] Gauss followed the same path in 1797, though he never published his findings. See [3] for more details.

This proves that the division radii r are roots of the polynomial equation $U(x) = 0$. When $m = 2n+1$, this is *precisely* the equation $P_{2n+1}(x) = 0$ considered by Abel.

To prove Abel's theorem, one can reduce to the case when $m = a + ib$ is an odd Gaussian prime. Since $U(x)$ has x as a factor. Eisenstein wrote $U(x) = xW(x)$, and the strategy of his proof was to show that $W(x)$ is irreducible. Once we know this, Abel's theorem follows—see [4, §15.5].[7]

But how do you prove that a polynomial such as $W(x)$ is irreducible? This is not easy. A key step for Eisenstein was when he noticed something about the coefficients of $W(x)$. He shared his thoughts with Gauss in a letter dated 18 August 1847 [10, p. 845]. Before quoting the letter, we need to observe that in terms of integrals, the equations $y = \varphi(mz)$ and $x = \varphi(z)$ imply that

$$\int_0^y \frac{dy}{\sqrt{1-y^4}} = m \int_0^x \frac{dx}{\sqrt{1-x^4}}.$$

In 19th century parlance, the relation between y and x given by (5) is an *algebraic integral* of this equality of integrals. Now the quote:

> Wenn $m = a + bi$ eine ungerade complexe Zahl, p deren Norm und $y = \dfrac{U}{V} = \dfrac{A_0x + A_1x^5 + \cdots + A_{(p-1)/4}x^p}{1 + B_1x^4 + \cdots + B_{(p-1)/4}x^{p-1}}$ das algebraische Integral der Gleichung
>
> $$\int_0 dy/\sqrt{1-y^4} = m \int_0 dx/\sqrt{1-x^4}$$
>
> ist, so hatte ich früher gezeigt, daß für eine *zweigliedrige* complexe Primzahl m die Coefficienten des Zählers bis auf den letzten, welcher eine complexe Einheit ist, und die Coefficienten des Nenners bis auf den Ersten, welcher $= 1$, alle durch m theilbar sind. Ich vermuthete, daß der Satz auch richtig sei, wenn m eine *eingliedrige* Primzahl ($\equiv 3 \pmod 4$ abgesehen vom Zeichen oder von einer complexen Einheit als Factor) ist;
>
> ---
>
> When $m = a + bi$ is an odd complex integer of norm p and $y = \dfrac{U}{V} = \dfrac{A_0x + A_1x^5 + \cdots + A_{(p-1)/4}x^p}{1 + B_1x^4 + \cdots + B_{(p-1)/4}x^{p-1}}$ is the algebraic integral of the equation
>
> $$\int_0 dy/\sqrt{1-y^4} = m \int_0 dx/\sqrt{1-x^4},$$
>
> so I had earlier shown that for a *two-term* complex prime number m the coefficients of the numerator up to the last, which is a complex unit, and the coefficients of the denominator except the first, which $= 1$, are all divisible by m. I conjectured that this proposition is also correct when m is a *one-term* prime number ($\equiv 3 \pmod 4$ apart from sign or a complex unit as factor);

[7] For a complete proof of Abel's theorem on the lemniscate, the reader should consult [4], [22] or [23]. The last reference gives a modern proof via class field theory.

(Note the use of four dots instead of three.) In the first part of the quote, Eisenstein sets up the situation, and after the displayed equation, describes the structure of the coefficients of the numerator and denominator. Odd Gaussian primes come in two flavors:

- *Two-term* primes of the form $m = a + ib$, where $p = a^2 + b^2$ is prime and $p \equiv 1 \bmod 4$.
- *One-term* primes of the form $m = \varepsilon q$, where ε is a unit in $\mathbb{Z}[i]$ and $q \equiv 3 \bmod 4$.

Now consider the polynomial

$$W(x) = \frac{1}{x}U(x) = A_0 + A_1x^4 + \cdots + A_{(p-1)/4}x^{p-1}.$$

For a two-term prime m, Eisenstein says that he earlier had shown that the last coefficient $A_{(p-1)/4}$ is a complex unit and the other coefficients $A_0, \ldots, A_{(p-1)/4-1}$ are divisible by m. He conjectures that the same is true for one-term primes.

This smells like the Eisenstein criterion, especially since Eisenstein notes in the letter that the constant term A_0 is m, which is not divisible by m^2. The difference is that m and the coefficients of W are Gaussian integers. A bit later in the letter, Eisenstein considers what happens if W is not irreducible over $\mathbb{Q}(i)$ [10, pp. 848–849]:

> ... wenn es möglich ist W das Produkt aus zwei rationalen ganzen Funktionen von x mit ganzen complexen Coefficienten, und deren Grade $< p-1$ sind. Es sei $W = PQ$; da das constante Glied von W, $= m$ ist, so kann, wenn m eine complexe Primzahl ist, das Constante Glied in einer der beiden ganzen Funktionen P, Q nur $= 1$, in der anderen $= m$ sein; denn die Coefficienten in P und Q müssen, wenn sie rational sind, nothwendig ganz sein, wie man durch dieselben Betrachtungen nachweisen kann, welche Ew. Hochwohlgeboren schon in der reellen Zahlentheorie (Disq. Sectio prima) angestellt haben.

> ... if it is possible that W is the product of two polynomials of x with Gaussian integer coefficients, and their degrees are $< p-1$. Let $W = PQ$; since the constant term of W is $= m$, so if m is a complex prime, the constant term in one of the two polynomials P, Q is $= 1$ and the other $= m$; then the coefficients of P and Q if rational, must necessarily be integral, as one can show by the same considerations which your Eminence[8] used in the real number theory (Disq. Section I).

[8]The literal translation of "Ew. Hochwohlgeboren" is "your High Well Born," which sounds silly in English. So I used "your Eminence" instead. The word "Hochwohlgeboren" originally applied to lesser German nobility and gentry. This flowery language is reflected in the letter's saluation, "Sr. Hochwohlgeboren, dem Geheimrath pp. Prof. Dr. Gauss", which translates "To his Eminence, the Distinguished, and so on, Professor Doctor Gauss." The word "Geheimrath," now spelled "Geheimrat," originated as the German equivalent of a "Privy councillor" in a governmental context and was an honorific for distinguished professors in German universities in the 19th century.

Here, "real number theory" means over $\mathbb{Z}$ rather than $\mathbb{Z}[i]$, and the reference to *Disquisitiones* is the first Gauss quote of this article. Thus Eisenstein is telling Gauss that Gauss's Lemma applies to the Gaussian integers. Mind-blowing. Then Eisenstein proceeds to prove that W is irreducible using one of the standard proofs of the Eisenstein criterion.[9] In other words, Eisenstein's first proof of his criterion

- was over the Gaussian integers;
- applied to a polynomial associated with the division problem on the lemniscate; and
- appeared in a letter to Gauss.

When Eisenstein wrote up his results for publication, he realized that his criterion was much more general. The first part of his long paper had the title *Über die Irreducibilität und einige andere Eigenschaften der Gleichung, von welcher die Theilung der ganzen Lemniscate abhängt* (*On the irreducibility and some other properties of equations that depend on the division of the lemniscate*) [10, pp. 536–555]. This paper contains Eisenstein's version of the Eisenstein criterion:

> „Wenn in einer ganzen Funktion $F(x)$ von x von beliebigem Grade „der Coëfficienten des höchstens Gleid $= 1$ ist, und alle folgenden Coëffi- „cienten ganze (reelle, complexe) Zahlen sind, in welchen eine gewisse „(reelle resp. complexe) Primzahl m aufgeht, wenn ferner der letzte „Coëfficient $= \varepsilon m$ ist, wo ε eine nicht durch m teilbare Zahl vorstellt: „so ist es unmöglich $F(x)$ auf die Form
>
> $$(x^\mu + a_1 x^{\mu-1} + \ldots + a_\mu)(x^\nu + b_1 x^{\nu-1} + \ldots + b_\nu)$$
>
> „zu bringen, wo μ und $\nu \geq 1$, $\mu + \nu =$ dem Grad von $F(x)$, und alle „a und b (reelle resp. complexe) ***ganze*** Zahl sind; und die Gleichung „$F(x) = 0$ is demnach irreductibel."

> If in a polynomial $F(x)$ of x of arbitrary degree the coefficient of the highest term is $= 1$, and all following coefficients are integers (real or complex), in which a certain (real resp. complex) prime number m appears, if further the last coefficient is $= \varepsilon m$, where ε represents a number not divisible by m: then it is impossible to bring $F(x)$ into the form
>
> $$(x^\mu + a_1 x^{\mu-1} + \ldots + a_\mu)(x^\nu + b_1 x^{\nu-1} + \ldots + b_\nu)$$
>
> where μ and $\nu \geq 1$, $\mu + \nu =$ the degree of $F(x)$, and all a and b are (real resp. complex) ***integers***; and the equation $F(x) = 0$ is accordingly irreducible.

[9] There are two standard proofs of the Eisenstein criterion. One proof (due to Eisenstein) works by studying which coefficients of the factors are divisible by the prime. The other proof (due to Schönemann) reduces modulo p and uses unique factorization in $\mathbb{F}_p[x]$.

After giving the proof (which works over any unique factorization domain), Eisenstein applies his criterion to the equation $W = 0$ that arises from division of the lemniscate and also to our friend $x^{p-1} + \cdots + 1$. Eisenstein's proof that the latter is irreducible is essentially identical to the one sketched on the first page of this article.

Eisenstein's paper is the first appearance of this classic proof of the irreducibility of $x^{p-1} + \cdots + 1$. Eisenstein is clearly pleased to have found such a splendid argument:

> ... Dies giebt also, wenn man will, einen neuen and höchst einfachen Beweis der Irreducibilität der Gleichung $x^{p-1} + x^{p-2} + \ldots + x + 1 = 0$; und zwar setzt dieser Beweis in Unterschiede mit früheren **) nicht die Kenntiss der Wurzeln und ihrer gegenseitigen Abhängigkeit voraus.
>
> **) Ausser dem Beweise von ***Gauss*** ist mir nur der von ***Kronecker*** im 29ten Bande dieses Journals Seite 280 bekannt.

> ... This thus gives, if you will, a new and most highly simple proof of the irreducibility of the equation $x^{p-1} + x^{p-2} + \ldots + x + 1 = 0$; and in constrast with earlier ones **), this proof does not presuppose knowledge of the roots and the relations among them.
>
> **) Besides the proof of ***Gauss***, only that of ***Kronecker*** in volume 29 of this journal, page 280, is known to me.

We know about Gauss's proof, and Kronecker's proof [18, Vol. I, pp. 1–4] from 1845 is simpler than Gauss's but still uses the explicit relations among the roots. But notice what footnote does *not* mention: Schönemann's two proofs of the irreducibility of $x^{p-1} + \cdots + 1$ given in his papers of 1845 and 1846. Yet Eisenstein's paper appears in the same journal in 1850!

Schönemann Complains

Eisenstein's paper, with the offending footnote, appeared in volume 39 of Crelle's journal. In volume 40, Schönemann published a *Notiz* [26], which began by describing two theorems from Eisenstein's paper:

- (ersten/first): The Eisenstein criterion for real primes (in $\mathbb{Z}$) and complex primes (in $\mathbb{Z}[i]$).
- (letzterem/last): The irreducibility of the cyclotomic polynomial $x^{p-1}+\cdots+1$, proved using the Eisenstein criterion.

Then Schönemann goes on to say:

> ... Da Herr ***Eisenstein*** ausdrücklich bemerkt, dass ihm von letzterem Satze nur der Beweis von ***Gauss*** und von ***Kronecker*** bekannt ist, so sehe ich mich veranlasst, daran zu erinnern, das ich bereits im Bande

> 31 dieses Journals §. 6, in meiner Abhandlung „Grundzüge einer allgemeinen Theorie der höhern Congruenzen etc.“ den ersten Satz für reelle Primzahlen beweisen und ach den folgenden aus demselben abgeleitet habe und das ferner die von Herrn etc. ***Eisenstein*** angewandete Methode nicht wesentlich von der meinigen verschieden is. Von dem letzteren Satze habe ich übrigens noch einen ganz verschiedenen Beweise im ersten Theile und §. 50 derselben Abhandlung gegeben.

> . . . Since ***Eisenstein*** expressly noted, that for the last theorem he only knew the proofs of ***Gauss*** and ***Kronecker***, I am led to recall that in §. 6 of my paper „Foundations of a general theory of higher congruences etc.” in volume 31 of this journal, I proved the first theorem for real primes and deduced the last from the first, and also the method used by ***Eisenstein*** is not significantly different from mine. For the last theorem, I in addition even gave an entirely different proof in §. 50 of the first part of the paper.

It seems clear that Eisenstein messed up by not citing Schönemann. However, there are some complications and confusions. First, Schönemann refers to §6 of his *Grundzüge . . .* paper in volume 31 of Crelle's journal, yet his irreducibility criterion and its application to $x^{p-1} + \cdots + 1$ are in §61 of the second part of his paper, which appeared in volume 32. The “§. 6” in his *Notiz* should have been “§. 61.” This explains part of the reason I had trouble finding Schönemann's criterion—I was looking in the wrong section!

But there was also confusion on Eisenstein's side as well. As already noted, Eisenstein's study of the division equations of the lemniscate was published in a two-part paper in Crelle's journal. The footnote quoted above appeared in the first part, in issue II of volume 39. The second part of the paper, *Über einige allgemeine Eigenschaften der Gleichung, von welcher die Theilung der ganzen Lemniscate abhängt, nebst Anwendungen derselben auf die Zahlentheorie* (*On some general properties of equations that depend on the division of the lemniscate, together with applications to number theory*) [10, pp. 555—619], appeared in issue III of the same volume. This paper included an *explicit reference* to Schönemann's first proof of the irreducibility of $x^{p-1}+\cdots+1$ (the one from §50 of Schönemann's paper in volume 31). Yet somehow this proof was unknown to Eisenstein when he wrote the first part of his paper. One can speculate on why this happened, but we will never know for sure.

Conclusion

We are now at the end of the amazing story of how Schönemann and Eisenstein independently discovered their criteria. Given that Schönemann discovered it first, the name “Schönemann-Eisenstein criterion” used by Dorwart is the most historically accurate. However, since most people use the Eisenstein's version, the name “Eisenstein-Schönemann criterion” is also reasonable. However, in my view, the name “Eisenstein criterion” does not do justice to Schönemann.

In the quote from Section VII of *Disquisitiones*, Gauss acknowledged two items of unfinished business: the extension from circular to transcendental functions such

as Abel's lemniscatic function φ, and the study of higher congruences. Both led to major areas of modern mathematics (elliptic curves and complex multiplication in the first case, p-adic numbers and local methods in number theory in the second), and both led to the Schönemann-Eisenstein criterion. Schönemann followed higher congruences to Hensel's Lemma to a question about irreducibility modulo p^2: his criterion appears in a completely natural way. Eisenstein followed Abel's work the lemniscate and considered the coefficients of the resulting division polynomials: his criterion appears in a completely natural way, completely different from the context considered by Schönemann. Yet both have their origin in the same paragraph in *Disquisitiones*. As I said, it is an amazing story.

Acknowledgements

The English translations of the first two Gauss quotes are from the English version of [13]. For the third Gauss quote and the first two Schönemann quotes, I used [11]. I would also like to thank Annemarie and Günter Frei for help in understanding the salutation in Eisenstein's letter to Gauss. Thanks also to Michael Filaseta for his help in pointing me to the right place in Schönemann's papers and to David Leep for bringing Dorrie's book [8] to my attention. I am also grateful to the referee for several useful suggestions.

I should also mention that the papers from Crelle's journal quoted in this article are available electronically through the Göttinger Digitalisierungzentrum at the web site

`http://gdz.sub.uni-goettingen.de/dms/load/toc/?IDDOC=238618`

References

[1] N. H. Abel, *Oeuvres complètes de Niels Henrik Abel*, Volume I, Edited by L. Sylow and S. Lie, Grøndahl & Søn, Christiana, 1881.

[2] *Allgemeine Deutsche Biographie*, Duncker & Humblot, Leipzig, 1875–1912. Available electronically at
`http://www.deutsche-biographie.de/~ndb/adb_index.html`
`http://de.wikisource.org/wiki/Allgemeine_Deutsche_Biographie`

[3] D. A. Cox, *The arithmetic-geometric mean of Gauss*, Enseign. Math. **30** (1984), 275–330. Reprinted in *Pi: A Source Book*, (L. Berggren, J. Borwein and P. Borwein, eds), third edition, Springer, 2003, 481–536.

[4] D. A. Cox, *Galois Theory*, Wiley, Hoboken, NJ, 2004.

[5] D. A. Cox, *Primes of the Form* $x^2 + ny^2$, Wiley, Hoboken, NJ, 1989.

[6] R. Dedekind, *Abriß einer Theorie der höheren Kongruenzen in bezug auf einen reellen Primzahl-Modulus*, J. Reine Angew. Math. **54** (1857), 269–325. Reprinted in *Gesammelte mathematische Werke*, Volume I, (E. Noether and O. Ore, eds), Vieweg, Braunschweig, 1930, 40–67.

[7] L. E. Dickson, *History of the Theory of Numbers*, Carnegie Institute, Washington, DC, 1919–1923. (Reprint by Chelsea, New York and AMS Chelsea, Providence, RI, 1969.)

[8] H. Dorrie, *Triumph der Mathematik: hundert berühmte Probleme aus zwei Jahrtausenden mathematischer Kultur*, Fredrich Hirt, Breslau, 1933. English translation of fifth edition published as *100 Great Problems of Elementary Mathematics*, Dover, Mineola, NY, 1965.

[9] H. L. Dorwart, *Irreducibility of polynomials*, Amer. Math. Monthly **42** (1935), 369–381.

[10] F. G. Eisenstein, *Mathematische Werke*, Volume II, (Reprint by Chelsea, New York and AMS Chelsea, Providence, RI, 1989).

[11] G. Frei, *The unpublished section eight: on the way to function fields over a finite field*, in *The Shaping of Arithmetic after C. F. Gauss's Disquisitiones arithmeticae*, (C. Goldstein, N. Schappacher and J. Schwermer, eds.), Springer, Berlin, 2007, 159–198.

[12] E. Galois, *Écrits et Mémoires Mathématiques D'Évariste Galois*, Edited by R. Bourgne and J.-P. Azra, Gauthier-Villars, Paris, 1962.

[13] C. F. Gauss *Disquisitiones Arithmeticae*, Leipzig, 1801. Republished in 1863 as Volume I of [15]. French translation, *Recherches Arithmétiques*, Paris, 1807. (Reprint by Hermann, Paris, 1910.) German translation, *Untersuchungen über Höhere Arithmetik*, Berlin, 1889. (Reprint by Chelsea, New York and AMS Chelsea, Providence, RI, 1965) English translation, Yale, New Haven, 1966. (Reprint by Springer, New York, 1986.) Catalan translation, Societat Catalana de Matemàtiques, Barcelona, 1996.

[14] C. F. Gauss, *Mathematical Diary*, Original manuscript in Latin: Handschriftenabteilung Niedersächsische Staats- und Universitätsbibliotek Göttingen, Cod. Ms. Gauß Math. 48 Cim. Ed. (Latin with German annotations). Reproduced in *Abrdruck des Tagebuchs (Notizenjournals)*, [15, Vol. X.1, pp. 483–575]. French translation by P. Eymard and J.-P. Lafon, *Le journal mathématique de Gauss*, Rev. Hist. Sci. Appl. **9** (1956), 21–51. English translation by J. Gray, *A commentary on Gauss's mathematical diary, 1796–1814*, Exposition. Math. **2** (1984), 97–130. German translation by E. Schumann, with a historical introduction by K.-R. Biermann, and annotations by H. Wußing and O. Neumann, *Mathematisches Tagebuch 1796–1814*, fourth edition, Ostwalds Klassiker der exakten Wissenschaften **256**, Akademische Verlagsgesellschaft Geest & Portig, Leipzig, 1985.

[15] C. F. Gauss, *Werke*, König. Gesell. Wissen., Göttingen, 1863–1927. Volumes I–IX are available electronically at
`http://www.wilbourall.org` (search for "Carl")

[16] F. Gouvêa, *p-adic Numbers: An Introduction*, Springer, 1993.

[17] C. Jordan, *Traité des substitutions et des équations algébriques*, Gauthier-Villars, Paris, 1870. Second edition, 1957.

[18] L. Kronecker, *Werke*, Leipzig, 1895–1931. (Reprint by Chelsea, New York and AMS Chelsea, Providence, RI, 1968.)

[19] F. Lemmermeyer, *Reciprocity Laws*, Springer, New York, 2000.

[20] E. H. Moore, *A doubly-infinite system of simple groups*, in *Mathematical Papers read at the International Mathematical Congress 1893*, Cambridge Univ. Press, Cambridge, 1896.

[21] J. J. O'Connor and E. F. Robertson, *Mactutor History of Mathematics archive*, available electronically at
`http://www-history.mcs.st-andrews.ac.uk/history/index.html`

[22] V. Prasolov and Y. Solovyev, *Elliptic Functions and Elliptic Integrals*, AMS, Providence, RI, 1997.

[23] M. Rosen, *Abel's theorem on the lemniscate*, Amer. Math. Monthly **88** (1981), 387–395.

[24] T. Schönemann, *Grundzüge einer allgemeinen Theorie der höhern Congruenzen, deren Modul eine reelle Primzahl ist*, J. Reine Angew. Math. **31** (1845), 269–325.

[25] T. Schönemann, *Von denjenigen Moduln, welche Potenzen von Primzahlen sind*, J. Reine Angew. Math. **32** (1846), 93–105.

[26] T. Schönemann, *Notiz*, J. Reine Angew. Math. **40** (1850), 188.

[27] J.-P. Serre, *Propriétés galoisiennes des points d'ordre fini des courbes elliptiques*, Invent. Math. **15** (1972), 259–331.

[28] B. L. van der Waerden, *Moderne algebra*, Springer, Berlin, 1930.

[29] H. Weber, *Lehrbuch der Algebra*, second edition, Vieweg, Braunschweig, 1898–1908. (Reprint by Chelsea, New York and AMS Chelsea, Providence, RI, 1961.)

Näringskedjor

L. Gårding

Matematik, Lunds universitet
larsg@maths.lth.se

Den naturliga utvecklingen av djurlivet på jorden har både givit nya arter och skapat strukturer baserade på relationen ätare-föda. En av dessa är näringskedjor av arter som antingen äter eller föder varandra. Dessa kedjor sägs vara stabila om de i längden behåller alla sina led, annars instabila. Det är klart att den naturliga utvecklingen favorerar stabila näringskedjor. För sådana kedjor finns en liten teori (se bibliografin) som använder lineär algebra för att förklara populationscykler och är ämnet för följande anspråkslösa betraktelse.

Stabila näringskedjor

En näringskedja består av populationer av ett antal arter som lever tillsammans och för sin föda beror av varandra. Det klassiska exemplet är populationer av betande djur som är föda åt en eller flera arter rovdjur. En näringskedja är stabil om den fungerar länge utan att förlora en population eller att en eller flera populationer växer okontrollerat. Detta detta hindrar inte att antalet individer i en population kan variera från år till år, ofta också regelbundet. Av detta kan man göra en enkel matematiska modell.

Låt $x = (x_1, ..., x_n)$ betyda årligen observerade individantal hos n populationer av arter som lever tillsammans i en eller flera näringskedjor och låt T vara en avbildning $x \to T(x)$ som beskriver den årliga förändringen av populationernas individantal. Vi antar också att att T är en bijektion av ett begränsat öppet område Ω i R^n som innehåller alla näringskedjans populationers individantal under upprepade årsövergångar.

Vi ska nu se att en reell lineär avbildning T alstrar en stabil kedja *om och endast om alla dess egenvärden ligger på enhetscirkeln.* Ty att kedjan är stabil betyder att för alla hela potenser T^k av T motsvarande områden $T^k\Omega$ tillhör en fix, kompakt del av R^n. Antag nu att T har ett egenvärde och egenvektor så att $T(u + iv) = \lambda e^{i\theta}(u + iv)$ med $\lambda \geq 0$. Vi kan anta att $u \neq 0$ och har då $T^k u = \lambda^n(\cos k\theta u - \sin k\theta v$. Med $w \in \Omega$ ligger också $w + cu$ i Ω då c är litet. Alltså ligger

$$T^k(w + cu) = T^k w + T^k u = T^k w + \lambda^k(\cos k\theta - \sin k\theta)$$

i en kompakt del av R^n. Här är första termen begränsad och alltså är $\lambda = 1$.

Cykler, perioder. Näringskedjans utveckling under den upprepade årliga förändringen sker i ett begränsat område och kan observeras under långa tider. Det som

då kan uppträda är cykler där ett svit av p antalsfördelningar, en cykel, ständigt upprepas. I vår modell betyder det att $T^p y = y$ för alla $y = T^k x_0$ med x_0 antalsfördelningen vid observationernas början. Talet p kallas cykelns period. För egenvärdena $e^{i\theta}$ till T betyder det att varje $p\theta$ är en multipel av 2π. Men det är naturligtvis mycket osannolikt att detta inträffar i praktiken och blir mer och mer osannolikt ju fler arter som är inblandade. Detta är också sant om man tillåter perioder som på ett eller annat sätt är approximativa. Men vi ska nu se ett par enkla fall där perioder finns och deras längder kan beräkna teoretiskt.

Tillämpningar

Vårt första exempel är kedjan ätare-föda. Välj enheter så att en enhet ätare i medeltal förbrukar en enhet föda årligen och låt P vara antalet enheter ätare och F antalet enheter föda vid ett års början. Om ätarnas genomsnittliga livslängd är A år dör i genomsnitt per år delen $a = 1/A$ av detta års ätare. I en stabil kedja kan vi också anse att årsantalet är nästan konstant, dvs att a också är ett genomsnittligt årligt födelsetal. Vi förutsätter att ungar blir mogna och kan räknas som vuxna vid årets slut.
Om $F = P$ kommer P att vara oförändrad vid årets slut. Annars leder brist på föda att P minskar och överflöd att P ökar genom extra överlevande avkomma. Detta kan åstadkommas enkelt med följande ierationsschema:

$$P \to P + a(F - P), F \to F - P + Q$$

där Q är ett årlig födotillskott. Schemat kan motiveras på följande sätt. Om $F > P$ finns det $F - P$ extra födoenheter som om de konsumeras av samma antal enheter ätare ger $a(F - P)$ ungar, ett nntal som vi antar har producerats av P ätare genom födoöverskottet. Om $F < P$ fattas föda för $P - F$ enheter ätare viLket ger en brist på motsvarande antal överlevande ungar.
Iterationer av schemat lämnr paret $F = P = Q$ invariant och betyder i allmänhet att paret P, F utför en cirkulär rörelse krin jämvikten Q, Q. Då $a > 4$ upphör kedjan att vara stabil och iterationerna divergerar mot oändligheten. För enheterna $P* = P - Q, F' = F - Q$ betyder schemat en lineär iteration med $Q = 0$ och en rörelse kring origo.
Det lineära schemat ovan innebär att $1 - a, a$ är den första raden i vår matris T och $-1, 1$ är den andra raden. Egenvärdena till T ges av

$$\lambda^2 - (2 - a)\lambda + 1 = 0$$

varav $\lambda = 1 - (a/2) \pm \sqrt{1 - (...)^2}$ så att $\lambda = e^{\pm i\theta}$ med $\cos\theta = 1 - a/2$. Denna formel ger egenvärden med absolutbelopp 1 precis då $a < 4$. Vi kan nu tillämpa den på lämmelcykeln i norr där lämlarna är ätare och deras föda är mossa och gräs. Lämmelns medellivslängd är omkring ett halvt år så att $a = 2$ varav $\cos\theta = 1 - a/2 = 0$ och $\theta = \pi/2$ så att lämmelcykeln är fyra år vilket stämmer med observationer. En andra tillämpning är den tioåriga cykeln för den kanadensiska snöskoharen i interaktion med sin ätare, lo. Medellivslängden för lo har uppskattas

till 2.6 år vilket enligt schemat ovan ger den observerade cykellängden 10.0 år.
Schemat kan utvidgas till tre aktörer, en predator Q som äter en art P som äter födan F. Det antas som tidigare att en enhet Q äter en enhet P som äter en enhet F i medeltal per år. Om Q, P har födotalen b, a så får man följande årsövergång:$Q \to (1-b)Q+bP, P \to P-Q+a(F-(P-Q)) = (1-a)(P-Q)+aF, F \to Q-P+F$ där mellanledet uttrycker att bara $P-Q$ av den andra predatorn äter föda när Q har fått sitt. Övergången ger en matris med raderna

$$(1-b, b, 0)/(a-1, 1-a, a)/(1, -1, 1)$$

skilda åt med snedsträck och ordnade uppifrån. Multiplicerad med en kolumn Q, P, F ger T årsövergången. Dess egenvärden visar sig vara 1 och ett par $e^{\pm i\theta}$ med $\cos\theta = 1-(b+a)/2$.
Här kan vi anta att P betyder lämmel. För Q finns flera val, arktiska rävar, rovfåglar eller vesslor som jagar lämmel i deras egna gångar. I de första fallen rör det sig om livslängder kring 2 år och alltså ett a kring $1/2$ som sänker lämmelperioden till 3.5. Vesslorna med en kortare livslängd på 1 år kortar den ytterligare till 3. I allt detta är lämlarnas precisa medellivslängd, här beräknad till ett halvår, litet osäker. Jag slutar med att låta detta vara ett exempel på mötet mellan en exakt teori och en mångfacetterad verklighet.
Alla tillämpningar nämnda här plus ett par till finns diskuterade i de två arbetena i referenslistan

En lång kedja

Denna sektion har ett exempel på en kedja med två arter predatorer A_1, A_2 och en art A_3 med vegetabilisk föda. Vi antar att alla lever i en stabil situation där den första arten äter den andra, den andra den tredje och den tredje äter av jordens gröda V. Enheterna är valda så att en enhet A_1 äter i genomsnitt en enhet A_2 årligen och analogt för A_2, A_3 och A_3, V. I fortsättningen låter vi F vara antalet årligen uppätna enheter gröda och låter P_1, P_2, P_3, F betyda antal enheter i motsvarande arter, allt under ett löpande år. Vi låter också a_1, a_2, a_3 betyda motsvarande predatorers födelsetal. Med mönster från fallet med två predatorer leder detta till årsövergångar för de tre populationerna givna av en matris T med raderna

$$(1-a_1, a_1, 0, 0)/(a_2-1, 1-a_2, a_2, 0)/(1-a_3, a_3-1, 1-a_3, a_3)/(-1, 1-1, 1)$$

tagna i ordning uppifrån och skilda åt med snedsträck. Koefficienterna i expansionen $\det(T+z) = z^4+c_1z^3+c_2z^2+c_3z+1$ är matrisens spår av ordningarna 1,2,3,4 och kan uträknas. Resultatet är

$$c_1 = 4-a_1-a_2-a_3, c_2 = 6-2(a_1+a_3)+a_1a_3, c_3 = c_1, c_4 = 1$$

Symmetrin visar att matrisens egenvärden är symmetriska under konjugation och invers vilket ännu inte betyder att den ger en stabil näringskedja. Två konjugerade egenvärden $re^{\pm i\theta}$ med $r<1$ innanför enhetscirkeln med inverserna $r^{-1}e^{\pm i\theta}$ utanför

är ännu tänkbart men ger inte stabilitet då $r < 1$. Detta fall ger alternativa spår för matrisen T,

$$c_1 = 2(r + 1/r)\cos\theta, c_2 = 2 + r^2 + r^{-2} + 2\cos 2\theta. c_3 = c_1, c_4 = 1$$

som inte kan uppkomma av T utan en koppling mellan födelsedata som beror på r och inte kan ge ett fall där dessa är lokalt fria. Vi går inte vidare med detta mycket speciella fall utan antar att att kedjan har alla sina egenvärden e î $\pm\,\theta_1, e^{\pm\theta_2}$ på enhetscirkeln. Spåren blir då

$$c_1 = 2\cos\theta_1 + 2\cos\theta_2; c_2 = 2 + 4\cos\theta_1\cos\theta_2; c_3 = c_1, c_4 = 1$$

så att

$$\cos\theta_1 + \cos\theta_2 = c_1/2; \cos\theta_1\cos\theta_2 = (c_2 - 2)/4$$

Våra två cosiner är alltså rötter absolut mindre än 1 till andragradsekvationen $t^2 - c_1 t/2 + (c_2 - 2)/4 = 0$. Detta ger villkor på koefficienterna, nämligen $0 < c_1/2 < 2$ och $-1 < (c_2 - 2)/4 < 1$, vilket för födelsetalen innebär villkoren

$$0 < a_1 + a_2 + a_3 < 4, -1 < 1 - (a_1 + a_3)/2 + a_1 a_3/4 < 1.$$

Alla är uppfyllda för små födelsetal. Om alla är lika med t betyder det att $1/4 < t < 3/4$, dvs att ingen art då kan ha ett högt födelsetal. Ett fall som påminner om situationen för lemmel är då de två första arterna i vår modell har ett litet födelsetal $a_1 = a_2 = t$ vilket betyder att

$$2t + a_3 < 4, 1 - (t + a_3)/2 + ta_3 < 1$$

där den andra olikheten inte begränsar a_3 och den första säger att ,$a_3 < 4 - 2t$. Det är alltså möjligt för stabilitet att A_3 är betesdjur som får mycket avkomma och sedan förser de två andra arterna med mat. I praktiken kan naturligtvis många andra fall uppträda men man skulle kanske ändå kunna dra ett par allmänna slutsatser i det stabila fallet.

Långa stabila näringskedjor av predatorer kan förekomma då alla deltagande arter (eller grupper av arter) har små födelsetal eller då ett led i kedjan äter vegetabilisk föda och har ett större födelsetal.

Speciella näringscykler har uppmärksammats mycket i den zoologiska litteraturen men en allmän teori för har hittills inte funnits. Vår anspråkslösa början på en sådan teori kan utsträckas till från predatorer och betesdjur till mera allmänna fall och kan kanske belysa de mäktiga näringskedjor som finns i havet och på den afrikanske savannen.

Referenser

[1] Gårding, L. A simple model for the interplay of predators, rodents and food. *J. Theoretical Biology* (2000) vol. 206, 73-80.

[2] Gårding, L. Applications of the eater-food model to predator-prey-food cycles. *Pol. J. Ecology 55.1* (2007) 187-189.

A minimum requiring angle trisection

Trond Steihaug and D. G. Rogers

Institutt for Informatikk, Universitetet i Bergen
PB7803, N5020, Bergen, Norge
trond.steihaug@ii.uib.no

For Audun Holme,
Editor, Normat*, 2005–2006,*
On his seventieth birthday, 1 December, 2008

1 Paul Nahin's envelope-folding problem

The solution to a problem can be *too* perfect, winning our admiration, but not our engagement. In contrast, a solution that is still rough and ready round the edges retains the power to draw us in, making us want to try our own hand. This was our response to reading Paul Nahin's account of what he calls the "envelope-folding problem" and to which he devotes a six-page section in his recent book *When Least is Best* [10, §3.3]. Perhaps attempting to share our enjoyment with readers of *Normat* is ill-advised, merely depriving them of similar pleasurable diversion. But the stimulus we should most like to transmit, in the hope that readers can resolve matters further, is that of discovering more than we can explain.

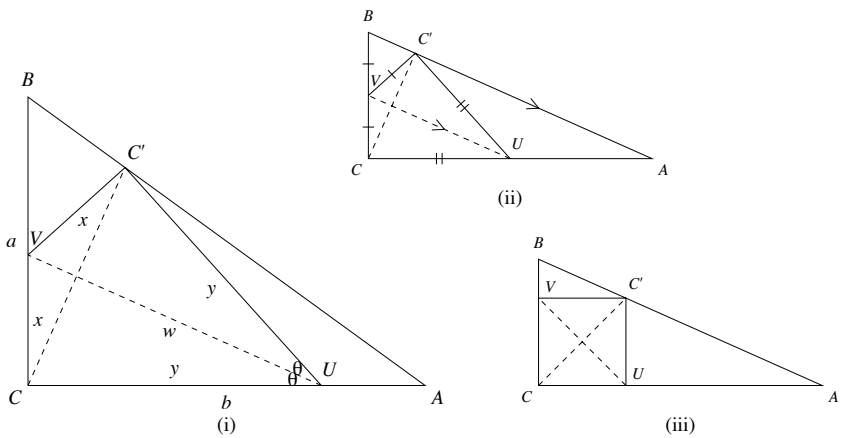


Figure 1: Folding a right triangle

Nahin's problem seems simple enough. We have a right triangle $\triangle ABC$ with right angle at C, with legs a and b and hypotenuse c, where $a^2+b^2=c^2$ and without loss of generality we may suppose that $a \le b$, as indicated in Figure 1(i). The vertex C is folded onto the hypotenuse AB at C' to create a crease UV. The envelope-folding problem asks: what is the least possible area of the folded right triangle $\triangle UVC$?

For instance, we can always fold parallel to the hypotenuse, as in Figure 1(ii). In this case, $\triangle UVC$ has an area one quarter that of $\triangle ABC$, that is, $ab/8$. Again, we can also fold along a diagonal on the inscribed square, as in Figure 1(iii). Since the inscribed square of $\triangle ABC$ is well-known to have side $s = ab/(a+b)$, this gives a folded triangle $\triangle UVC$ with area $s^2/2$. The inequality between the arithmetic and geometric means implies that this second area is always at least as large as the first, with equality only in the case of the isosceles right triangle. Symmetry considerations suggest that, in the isosceles case, the common fold does give the minimum area of the folded right triangle $\triangle UVC$.

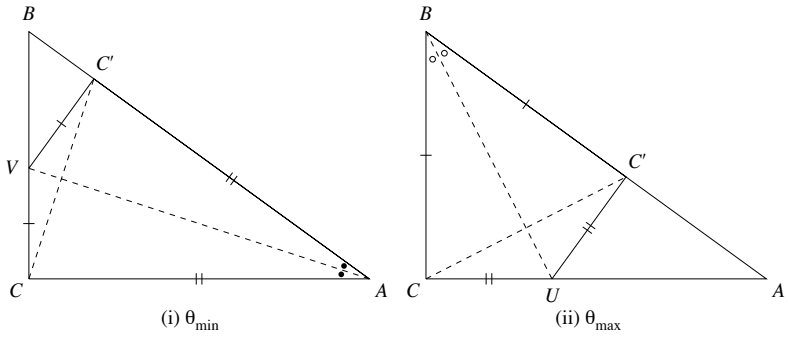


Figure 2: Angle bisectors as extreme folds

But otherwise there seems little in the way of intuition to guide us, so we must have recourse to a more analytical investigation. To this end, Nahin lets $\theta = \angle CUV$ be the angle that the crease UV makes with AC (see Figure 1(i)). If the crease is to cut off a triangle $\triangle UVC$, then it varies from the internal angle bisector of $\angle BAC$ to the internal angle bisector of $\angle ABC$ (compare Figure 2). Thus, writing as usual $\alpha = \angle BAC$,

$$\frac{\alpha}{2} \le \theta \le \frac{\pi}{4} + \frac{\alpha}{2}. \tag{1}$$

Nahin then builds expectation through a circuitous derivation of the area Δ of $\triangle UVC$ in terms of the parameter θ:

$$\Delta = \Delta(\theta) = \frac{(ab)^2}{8\cos\theta\sin\theta[a\sin\theta + b\cos\theta]^2}. \tag{2}$$

Now, it is possible to derive (2) much more directly, as we do in the next section. But that is not the point. Rather, the minimisation of (2) is easier than Nahin seems to suggest, amounting to the maximisation of the denominator. Thus, recalling only double and then treble angle formulae, we go on to locate the minimum of (2)

subject to (1) with startling simplicity by *angle trisection*:

$$\theta = \frac{1}{3}(\pi/2 + \alpha). \tag{3}$$

Now it is well-known that, in general, angle trisection involves the solution of a cubic equation, and is therefore not amenable to construction by straightedge and compasses. However, any angle can be trisected by paper-folding — an introduction [11] to paper-folding, including this result, was presented by Benedetto Scimemi in *Normat* in 1998. David Cox, who contributes an article [2, (c)] to the present issue, provides related results in his book [2, §10.3] (there is a connection between [11] and [2, (c)] *via* [2, (b)]). So, Nahin's envelope-folding problem has the happy outcome that the minimising solution can itself be obtained by folding. Readers who follow *Normat* closely may also be interested to note that the results in [11] also lead to foldings for the roots of the quartic equations arising in the ladder problems reviewed by Kent Holing in his series [8, esp. Pt. III] (compare [10, §3.4] and [5, Ex. 18]).

Unfortunately, Nahin misses these developments. For, his purpose in introducing the problem was to illustrate "how a computer can play a highly useful role in minimization analyses". So, having achieved (2) — and challenged his readers to try the "nasty business" of setting $d\Delta/d\theta = 0$ — he opts immediately to use a computer to study the behaviour of $\Delta(\theta)$ directly, plotting it in the isosceles case $a/b = 1$, where we already intuit the answer by symmetry, and in the case $a/b = 1/2$, where an answer is less obvious.

Is the neatness of (3) something special to Nahin's problem? To find out, we investigate some related minimisation problems. For example, in Section 3, we consider minimising variously the displacement CC' of C and the width UV of the resulting crease. The former is nice enough, as the displacement is minimised by folding parallel to the hypotenuse, as in Figure 1(ii), when the displacement is the altitude and the width is half the hypotenuse. But the latter is of the same order of difficulty as Nahin's problem, in that it, too, requires the solution of a cubic equation, but without an answer as neat as (3). Curiously enough, if we switch to folding A onto BC, then minimising either the area of folded triangle or the width of the crease require only the solution of quadratic equations, as we show in Section 4. In terms of the functions involved in these problems, some might have served Nahin's interests in computer-aided studies as well as, if not better than, minimising the area Δ in the envelope-folding problem. However, Nahin's choice seems inspired in the combination it offers of arresting challenge and attractive outcome.

So, we are left with the puzzle: what is the meaning of the angle trisection in (3) for minimisation? To point up this question, in Section 5, we leave the reader to consider what is in effect a limiting case of Nahin's problem where, for fixed a, the right triangle becomes a strip of width a — but now, following [5], in *doing without calculus*.

2 Minimising the area

Our first task is to provide a straightforward derivation of (2). Naturally, we should keep in mind the key property of folding, that the crease UV is the perpendicular bisector of the displacement CC'. Thus, $\angle C'UV = \angle CUV = \theta$, so that $\angle CUC' = 2\theta$ (see Figure 3).

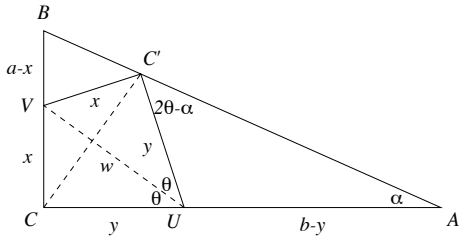


Figure 3: Setting for Sine rule

But $\angle CUC'$, as an exterior angle of $\triangle AC'U$, is the sum of the two opposite angles $\angle C'AU = \angle BAC = \alpha$ and $\angle AC'U$. Turning this equality around, we deduce that

$$\angle AC'U = \angle CUC' - \angle C'AU = 2\theta - \alpha.$$

This means that, with reference to Figure 3, application of the Sine rule yields

$$\frac{y}{\sin\alpha} = \frac{C'U}{\sin\angle C'AU} = \frac{AU}{\sin\angle AC'U} = \frac{b-y}{\sin(2\theta-\alpha)}.$$

Collecting terms in y gives

$$b\sin\alpha = y(\sin\alpha + \sin(2\theta-\alpha)) = y(\sin\alpha + \sin 2\theta\cos\alpha - \cos 2\theta\sin\alpha).$$

Hence, on recalling that $\cot\alpha = b/a$ and deploying the double angle formulae, we find that

$$y = \frac{ab}{2\sin\theta(a\sin\theta + b\cos\theta)}. \tag{4}$$

Turning now to the analogous triangle $\triangle BC'V$, we see that

$$\angle C'VB = \pi - \angle C'VC = \angle C'UC = 2\theta,$$

because $\angle UC'V = \angle UCV = \pi/2$, while

$$\angle BC'V = \pi - \angle UC'V - \angle AC'U = \frac{\pi}{2} + \alpha - 2\theta.$$

So, we are also in a position to apply the Sine rule in this second triangle, deriving thereby an expression for x matching (4):

$$x = \frac{ab}{2\cos\theta(a\sin\theta + b\cos\theta)}. \tag{5}$$

But, in view of Figure 1(i), $\Delta = xy/2$, so (4) and (5) delivers (2). Thus we are able to dispense with Nahin's much longer derivation, if at the price of knowing the double angle formulae. However, these are also suggestive on turning to the minimisation problem.

Now, as ab is a constant, (2) is really just a reciprocal. So, minimising $\Delta(\theta)$ is equivalent to maximising a multiple of the denominator, say,

$$f(\theta) = 2\sin\theta\cos\theta(a\sin\theta + b\cos\theta)^2,$$

taking care to avoid $f(\theta)$ vanishing. Differentiation will be simplified if we revert to double angles and write

$$f(\theta) = \sin 2\theta(a\sin\theta + b\cos\theta)^2.$$

It follows that

$$\begin{aligned} f'(\theta) &= 2\cos 2\theta(a\sin\theta + b\cos\theta)^2 + 2\sin 2\theta(a\sin\theta + b\cos\theta)(a\cos\theta - b\sin\theta) \\ &= 2[a(\sin\theta\cos 2\theta + \cos\theta\sin 2\theta) + b(\cos\theta\cos 2\theta - \sin\theta\sin 2\theta)](a\sin\theta + b\cos\theta) \\ &= 2(a\sin 3\theta + b\cos 3\theta)(a\sin\theta + b\cos\theta), \end{aligned} \tag{6}$$

making use of the treble angle formulae to achieve the final simplification in (6). Consequently, the only turning points of $f(\theta)$, with $f(\theta)$ non-zero, are given by

$$\tan 3\theta = -\frac{b}{a}. \tag{7}$$

But since $\tan\alpha = a/b$, we also know that

$$\tan(\pi/2 + \alpha) = -\frac{b}{a}.$$

Hence, the solution to (7) in the range (1) is given by (3); and this solution does give a maximun for $f(\theta)$, as expected.

What if we have (2), but do not (think to) make use of double and treble angles? Then there is some choice as to what might stand in place of the term in 3θ in the product in (6). But if we leave it as

$$a(3\cos^2\theta - \sin^2\theta)\sin\theta + b(\cos^2\theta - 3\sin^2\theta)\cos\theta,$$

then instead of (7), the location of the minimum of $\Delta(\theta)$ is given by a cubic in $t = \tan\theta$:

$$at^3 + 3bt^2 - 3at - b = 0. \tag{8}$$

Clearly, (8) disguises the angle trisection in (3). Nevertheless, it turns out to be tractable.

In order to recast (1) in terms of t, it is helpful to recall that the tangents of the half-angles of a right triangle can be expressed by means of the sides together with the radius r of the inscribed circle:

$$\tan\frac{\alpha}{2} = \frac{r}{b-r}, \quad \tan(\frac{\pi}{4} - \frac{\alpha}{2}) = \frac{r}{a-r},$$

where $r = (a + b - c)/2$. It follows that (1) translates to

$$\frac{a+b-c}{b+c-a} \leq t \leq \frac{a+c-b}{a+b-c}. \tag{9}$$

On the other hand, since $a \leq b$, it is straightforward to check that (8) has a root, $t = t_\Delta$, say, in the interval $[\frac{a}{b},1]$, so satisfying (9). The sign of the constant term in (8) implies there are either one or three positive real roots. But with three positive real roots, the cubic would have both its own turning points positive, which is not the case. Hence the root t_Δ in $[\frac{a}{b}, 1]$ is, in fact, the unique positive root of (8). Thus, comparing (8) with (6), at this turning point for $f(\theta)$, $f'(\theta)$ changes from positive to negative, ensuring that $f(\theta)$ assumes a maximum.

3 Minimising the crease

In obtaining (2) in the previous section, we made use of the fact that the area Δ of the folded triangle $\triangle UVC$ is given by $\Delta = xy/2$. Of course, this depends on the angle at C being right. But, whenever we fold, the displacement $l = CC'$ and the width $w = UV$ of the crease are orthogonal. Thus, with reference to Figure 1(i), we have an alternative expression for Δ:

$$\Delta = \frac{lw}{4}. \tag{10}$$

Now, since the angle at C is right, $w^2 = x^2 + y^2$. So, from (4) and (5), we find that

$$w = \frac{ab}{2\sin\theta\cos\theta(a\sin\theta + b\cos\theta)}, \tag{11}$$

and then comparison of (2), (10) and (11) gives

$$l = \frac{ab}{a\sin\theta + b\cos\theta}. \tag{12}$$

In (11) and (12), we have two further examples where, as with (2), minimisation can be effected through maximising a denominator. Of these, the minimisation of the displacement l in (12) is the simpler and neater. Indeed, differentiation of the denominator $a\sin\theta + b\cos\theta$ quickly reveals that l attains a minimum when $\theta = \alpha$. that is, when the crease is parallel to the hypotenuse. Consequently, in this case, C' is the foot of the altitude of $\triangle ABC$ from C, while U and V are the midpoints of AC and BC (compare Figure 1(ii)). As $\sin\alpha = a/c$ and $\cos\alpha = b/c$, with $c^2 = a^2 + b^2$, we infer from (11) and (12) that $w = ab/c$ and $l = c/2$.

On the other hand, minimisation of w itself proceeds along the lines of our discussion in the final paragraph of the previous section. The upshot is that the minimum is located by another cubic in $t = \tan\theta$:

$$at^3 + 2bt^2 - 2at - b = 0. \tag{13}$$

Without being alert to the quirk of angle trisection, there would seem little to tell (8) and (13) apart. In particular, arguing as for (8), (13) has a unique positive root $t = t_w$, say, which is located in the interval $[\frac{a}{b},1]$, so satisfying (9), and gives a minimum for w in (11). Moreover, these unique positive roots of (8) and (13) are related by

$$\frac{a}{b} \leq t_\Delta \leq t_w \leq 1.$$

Any one of these inequalities holds with equality only in the case $a = b$. that is, the right triangle $\triangle ABC$ is isosceles, and equality holds throughout. But Nahin's computer-aided study might come in more handy with (13), where we miss out on such a neat relation as (3).

4 Minimising another area

Envelopes make a handy source of right triangles ready for folding. But there is nothing in the nature of folding to restrict attention to right triangles and readers might be encouraged to experiment with folding vertices of triangles onto opposite sides for triangles and vertices of their choice. Entering into this spirit, we stay with Nahin's right triangle $\triangle ABC$, but now consider folding A onto BC at, say, A', by a crease UV with U on AC and V on AB.

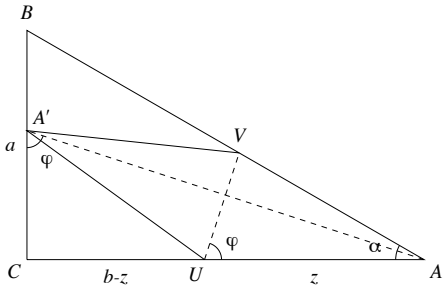


Figure 4: Another folding

If A' is on the line segment BC, rather than the line BC produced, the internal angle bisectors at B and C are inaccessible as folds for $a < b$, and the crease varies instead from the perpendicular bisector of AB, when A is folded onto B, to the perpendicular bisector of AC, when A is folded onto C. Hence, if we let $\phi = \angle AUV$ be the angle that the crease UV makes with CA, as shown in Figure 4, then

$$\frac{\pi}{2} - \alpha \leq \phi \leq \frac{\pi}{2}. \tag{14}$$

Further, let O be the point of intersection of the crease UV with the displacement AA'. Since the crease and the displacement intersect at right angles, the right triangles $\triangle AA'C$ and $\triangle AUO$ are similar, so that

$$\angle AA'C = \angle AUO = \angle AUV = \phi. \tag{15}$$

We convert our notation to this new setting, writing Δ for the area of the folded triangle $\triangle AUV$, l for the displacement AA', w for the width of the crease UV and $z = AU = A'U$. With this understanding, Δ is given by (10), as before. From Figure 4, we see that

$$l \sin\phi = b, \quad 2z \sin\phi = l.$$

Thus, z comes out as:

$$z = \frac{b}{2\sin^2\phi}. \tag{16}$$

Now, arguing on the lines of our derivation of (4) in Section 2, application of the Sine rule to $\triangle AUV$ yields, with reference to Figure 4,

$$\frac{w}{\sin\alpha} = \frac{UV}{\sin\angle UAV} = \frac{AU}{\sin\angle AVU} = \frac{z}{\sin(\pi-\alpha-\phi)},$$

that is,

$$z\sin\alpha = w\sin(\alpha+\phi) = w(\sin\alpha\cos\phi + \cos\alpha\sin\phi).$$

Noting once more that $\cot\alpha = b/a$ and eliminating z by means of (16) leads to

$$w = \frac{az}{a\cos\phi + b\sin\phi} = \frac{ab}{2\sin^2\phi(a\cos\phi + b\sin\phi)}. \tag{17}$$

Hence,

$$\Delta = \frac{lw}{4} = \frac{ab^2}{8\sin^3\phi(a\cos\phi + b\sin\phi)}. \tag{18}$$

It is immediate that l is minimised by folding A onto C. But the functions in (17) and (18) provide further candidates for computer-aided study of the sort favoured by Nahin. However, as with the counterpart expressions in the previous two sections, there is really no problem in locating minimums in (17) and (18) by maximising the denominators. Perhaps what is most interesting here is that, in contrast with the cubic equations (12) and (8), these new minimums are located as the positive roots of a *quadratic*, meaning that, in some sense, the problems in this section are simpler — for example, we can construct their solutions by means of straightedge and compasses, which is not possible for the solution of the cubics. Thus the minimum of w in (17) is located at the positive root of

$$at^2 - 4bt - 3a = 0,$$

while that of Δ in (18) is located at the positive root of

$$at^2 - 3bt - 2a = 0,$$

where now $t = \tan\phi$. It is easy the check that these positive roots are greater than b/a, in keeping with (14), although it is less easy to recognise the angles that have these tangents, even in the isosceles case.

There may be somewhat more transparency here in using as the parameter $h = A'C$. From (15), $ht = h\tan\phi = b$. Moreover, if L is the foot of the perpendicular from O onto AC, then L is the midpoint of AC while $OL = \frac{1}{2}h$. Thus, the discussion in this section can be reformulated in the language of coordinate geometry by means of h, if so wished. At all events, in terms of h, we find that

$$l = \sqrt{b^2 + h^2}, \quad w = \frac{a(b^2 + h^2)\sqrt{b^2 + h^2}}{2b(ah + b^2)}, \quad \Delta = \frac{a(b^2 + h^2)^2}{8b(ah + b^2)}.$$

Of course, the minimum of l is b. The minimum of w occurs for

$$h_w = b(\sqrt{8a^2 + 9b^2} - 3b)/4a;$$

that for Δ occurs for

$$h_\Delta = b(\sqrt{3a^2 + 4b^2} - 2b)/3a.$$

For what it is, worth $0 < h_\Delta < h_w < a$.

5 A stripped-down minimisation

Nahin has not been alone in combining folding with minimisation. For example, the questions of minimising the width of the crease when a corner C of a strip of width a is folded onto the opposite edge, as in Figure 5, has already achieved a certain veneration, perhaps because the answer is surprisingly neat: the minimum occurs when $x = 3a/4$, that is when $\tan\theta = \sqrt{2}/2$ (compare (13)). For the folded triangle in Figure 5, the minimum area occurs when $x = 2a/3$, that is, when $\theta = \pi/6$, the limiting case of (3) as α tends to zero. Clearly, Nahin's envelope-folding problem can be seen as extending this setting to allow the line through B onto which C is folded to be at an acute angle to BC, rather than a right angle.

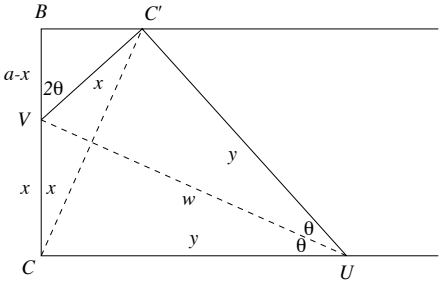


Figure 5: The folded strip

The well-known puzzlist Henry Ernest Dudeney (1857–1930) posed this minimum crease problem in the issue of *The Strand Magazine* for August, 1919, as part of his long-running series of *Perplexities* [3, (a), Prob. 469] that had appeared there since

May 1910 (and was to run until his death in 1930). He included it in *Modern Puzzles and How to Solve Them* [3, (b) Prob. 139, pp. 55, 144], a collection drawn mainly from this series that came out in 1926 (for a contemporary review, see [3, (c)]). However, in this instance no justification is given for the answer — the pages from Dudeney's book are reproduced in Figure 6 on p. 89. Similarly, Martin Gardner (1914–) gave the problem a further airing as an aside when discussing paper-folding in an early column on *Mathematical Games* [6] in the issue of *Scientific American* [6] for July, 1959. Although the problem has done long service as a staple in calculus textbooks, a fresh look has been taken recently in [4, 9].

But the questions of minimising the folded triangle and the width of the crease in Figure 5 are also among the examples in an article [5, Ex. 17 and Fig. 7] by Trevor James Fletcher (1921–) in 1971. Fletcher's intent is to show how to *dispense* with calculus in working a variety of maximisation and minimisation problems — an aim of which Dudeney and Gardner would no doubt approve. The "fundamental principle", as he puts it, running through his solutions is the inequality between the arithmetic and geometric means. But he concedes that the folding problems are "a little tricky" managed this way, and opts to leave them for "enthusiasts to solve for themselves". We are happy to pass on this encouragement to readers of *Normat*, now with Nahin's generalisation.

Indeed, there is a further puzzle we should like to include. Folding the corner of a paper strip over onto the opposite edge has been developed for classroom activity, not only as a computer-aided exercise in optimisation, for example in [1], but also as a construction for Pythagorean triples, for example in [7]. However, the triangle of interest in [1, 7] is not the folded triangle $\triangle UVC$, but rather the right triangle $\triangle C'VB$. Now, the maximum area of this triangle can be obtained by a simple direct argument. Let V' be the reflection of V in the line through B perpendicular to BC. Then $\triangle C'V'B$ is congruent to $\triangle C'VB$. Hence, $\triangle C'VV'$ is isosceles, with equal sides x and base $2(a-x)$, and so with constant perimeter $2a$. But among all (isosceles) triangles with given perimeter, the equilateral triangle has largest area. Thus, $\triangle C'VB$ attains its maximum area when $\triangle C'VV'$ is equilateral, that is, when $x = 2(a-x)$ or $x = 2a/3$. So, our further puzzle is to explain why this maximum area is achieved where the folded triangle $\triangle UVC$ attains its minimum area.

Moreover, this agreement is not peculiar to the folded strip, but holds more generally in Nahin's envelope-folding problem. For, let $\Gamma = \Gamma(\theta)$ denote the area of triangle $\triangle C'VB$ in Figure 3, so that

$$\Gamma = \Gamma(\theta) = \frac{1}{2}(a-x)x\sin 2\theta.$$

From (5), after a little algebra with trigonometric expansions, we find that

$$\Gamma(\theta) = \frac{a^2 b\sin\theta(a\sin 2\theta + b\cos 2\theta)}{4\cos\theta(a\sin\theta + b\cos\theta)^2}.$$

Now, if Nahin thought $\Delta(\theta)$ in (2) was messy to differentiate, then $\Gamma(\theta)$ looks even more formidable. Yet, the derivative comes out comparatively cleanly:

$$\Gamma'(\theta) = \frac{(ab)^2(a\sin 3\theta + b\cos 3\theta)}{4\cos^2\theta(a\sin\theta + b\cos\theta)^3}.$$

Hence, we see that turning points for $\Gamma(\theta)$ are also given by (7), as for $\Delta(\theta)$, and we can then verify that $\Gamma(\theta)$ attains a maximum at (3). But this working only quickens our interest in having some more transparent explanation.

References

[1] S. M. Arnold, Online mathematics resources, available at ⟨`http://www.compasstech.com.au/ARNOLD/maths.htm`⟩.

[2] D. A. Cox, (a) *Galois Theory* (Wiley Interscience, Hoboken, NJ, 2004). **MR**2119052; (b) with J. Shurman, Geometry and number theory on clovers, *Amer. Math. Monthly*, **112** (2005), 682–704. **MR**2167769; (c) Why Eisenstein proved the Eisenstein criterion and why Schönemann discovered it first, *Normat*, **57** (2009), *this issue*.

[3] H. E. Dudeney, (a) Perplexities, The Strand Magazine, **58** (August, 1919), 200; (b) *Modern Puzzles and How to Solve Them* (C. A. Pearson, London, 1926; 2nd ed., 1936) (c) review of [3, (b)] by W. Hope-Jones, *Math. Gaz.*, **13** (1927), 337–338.

[4] S. E. Ellermeyer, A closer look at the crease length problem, *Math. Mag.*, **81** (2008), 138–145.

[5] T. J. Fletcher, Doing without calculus, *Math. Gaz.*, **55** (1971), 4–17.

[6] M. Gardner, Mathematical Games: Origami, *Scientific American*, **201** (July 1959); reprinted as Chap. 16 *in* M. Gardner, *The Second Scientific American Book of Mathematical Puzzles and Diversions* (Simon and Schuster, New York, NY, 1961), esp. pp. 144–145; updated as *Origami, Eleusis, and the Soma Cube. Martin Gardner's Mathematical Diversions. New Martin Gardner Mathematical Library* (Cambridge University Press, Cambridge, UK; Math. Assoc. Amer., Washington, DC, 2008). **MR**2441570.

[7] G. Hatch, Note 80.40: Still more about the $(20, 21, 29)$ triangle, *Math. Gaz.*, **80** (1996), 548–550.

[8] K. Holing, På gjengrodde stiger I, *Normat*, **45** (1997), 62–78; II: tillegg og rettelser, *ibid*, **46** (1998), 45; III: geometriske løsninger, *ibid*, **48** (2000), 83–90; IV: epilog, *ibid*, **50** (2002), 92–95. **MR**1780823.

[9] R. B. Kirchner, The crease length problem revisited, Spring Meeting, MAA North Central Section, 25 April, 2009; demonstrations available at ⟨`http://public.me.com/rkirchne/CreaseLengthProblem`⟩; ⟨`http://demonstrations.wolfram.com/ExploringTheCreaseLengthProblem`⟩

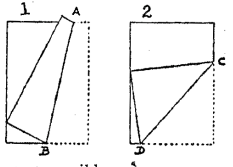
[10] P. J. Nahin, *When Least Is Best: How mathematicians discovered many clever ways to make things as small (or as large) as possible* (Princeton University Press, Princeton, NJ, 2004). **MR**2022170.

[11] B. Scimemi, Algebra og geometri ved hjelp av papirbretting, *Normat*, **46** (1998), 170–185, 188. **MR**1682447.

middle of a square, one on the edge of each side of the board and no man in a corner. It is a real puzzle, until you hit on the method of solution, and then to get the correct answer is absurdly easy.

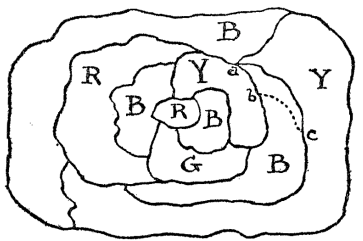
139.—A Crease Problem

Fold a page, so that the bottom outside corner touches the inside edge and the crease is the shortest possible. That is about as simple a question as we could put, but it will puzzle a good many readers to discover just where to make that fold. I give two examples of folding. It will be seen that the crease A B is considerably longer than C D, but the latter is not the shortest possible.



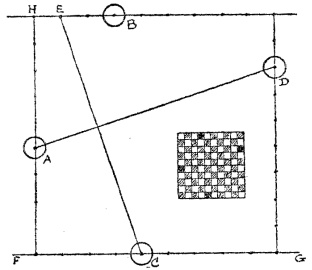
140.—The Four-colour Map Theorem

For just about fifty years various mathematicians, including De Morgan, Cayley, Kempe, Heawood, Heffter, Wernicke, Birkhoff, Franklin, and many others have attempted to prove the truth of this theorem, and in a long and learned article in the

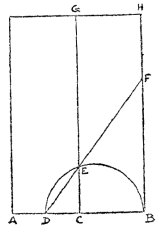


American Mathematical Monthly for July-August, 1923, Professor Brahana, of the University of Illinois, states that "the problem is still unsolved." It is simply this, that in colouring any map under the condition that no contiguous countries shall be coloured alike, not more than four colours can ever be neces-

sides. Also draw a line F G through C and parallel to E B and the lines through A and D perpendicular to E B and F G. Now, as H is the centre of a corner square, we can mark off the length H E all round the square and we find the board is 10×10.



If the size of the men were not given we might subdivide into more squares, but the men would be too large for the squares. As the distance between the centres of squares is the same as the width of the squares, we can now complete the board with ease, as shown in the diagram inset.



139.—A Crease Problem

Bisect A B in C and draw the line C G, parallel to B H. Then bisect A C in D and draw the semicircle D B, cutting the line C G in E. Now the line D E F gives the direction of the shortest possible crease under the conditions.

140.—The Four-colour Map Theorem

With two or more contiguous countries, two colours at least are obviously necessary (Fig. 1). If three countries are

Figure 6: Facscimiles from Dudeney's 1926 book *Modern Puzzles and How to Solve Them.*

Uppgifter

Uppgifterna 521, 522, 525 och 526 är hämtade från olympiadtävlingar i Ryssland, Vietnam och Japan.

521. Olikheten $\sin\alpha+\sin\beta+\sin\gamma\geq 2$ är uppfylld för vinklarna α, β och γ. Visa att $\cos\alpha+\cos\beta+\cos\gamma\leq\sqrt{5}$.

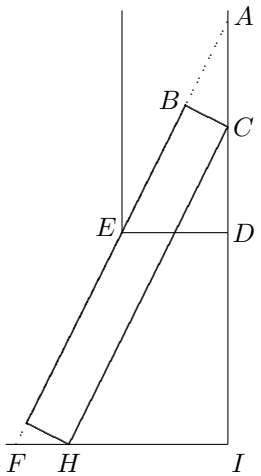
522. För tre icke-negativa reella tal a, b, c gäller att $ab+bc+ca+abc=4$. Visa att

$$a+b+c\geq ab+bc+ca.$$

När inträffar likhet?

523. Låt a, m, n vara positiva heltal med $a>1$. Antag att a^m+1 är en delare till a^n+1. Visa att m då måste vara en delare till n.

524. (Föreslaget av *Kent Holing, Trondheim, NO*) En del av en gammel prisoppgave i 1732 fra det kjente *Ladies' Diary* lyder som følger: Bestem den lengste sylindriske pålen med diameter $|BC|=1$ fot som akkurat får plass oppe i en pipe med en kvadratisk åpning med side $|DE|=4$ fot og som er $|DI|=8$ fot over golvet. Se vedlagt figur! Vi generaliserer: Karakteriser alle slike problem der $|DE|$, $|DI|$, $|AF|$, $|CH|$ og $|BC|$ er heltallig! (Hint: Se løsning til oppgave 404b) i hefte 1 av Normat 2003, side 36.)



525. Låt n vara ett givet positivt heltal. Betrakta en permutation $(a_1, a_2, \ldots, a_{2n})$ av heltalen $1, 2, \ldots, 2n$ sådan att talen $|a_{i+1} - a_i|$ för $i = 1, 2, \ldots, 2n-1$ alla är olika.
Visa att $a_1 - a_{2n} = n$ om och endast om $1 \leq a_{2k} \leq n$ för varje $k = 1, 2, \ldots, n$.

526. Låt $\mathbb{Q}$ vara mängden av rationella tal. Betrakta en funktion $f : \mathbb{Q} \to \mathbb{Q}$ som uppfyller följande tre villkor:
1) $f(0) = 2$, $f(1) = 3$;
2) För alla rationella tal x och alla heltal n gäller

$$f(x+n) - f(x) = n\big(f(x+1) - f(x)\big);$$

3) För alla rationella tal $x \neq 0$ är $f(x) = f(\frac{1}{x})$.
Bestäm alla rationella tal x för vilka $f(x) = 2009$.

Lösningar skickas senast 15 december 2009 till:
Dag Jonsson, dag@math.uu.se
Box 480
Uppsala universitet, Matematiska institutionen
SE-75106 Uppsala

Anm. Vi välkomnar givetvis lösningar även efter angivet datum så länge inga andra lösningar har presenterats i tidskriften.

Lösningar till tidigare uppgifter i Normat

512. (*Erik Hansen, Kalundborg, DK*) To snit (evt. ét) vil altid kunne dele perlekæden på den ønskede måde.
Vi tænker os perlekæden foldet sammen til en cirkel, hvor perlerne sidder ækvidistant. Cirklen deles af en diameter (med en pil), således at halvdelen af perlerne er i den højre halvdel af cirklen.
Antallet af røde perler i højre halvdel af cirklen er $r + h$. Antag at h ikke er nul. (Ellers snittes efter diameteren.) Hver gang diameteren drejes en vinkel på $\frac{\pi}{r+s}$ i positiv omløbsretning, glider en perle ud og en ny kommer ind i højre halvdel; h ændres derved med en af værdierne -1, 0 eller 1. Efter at have drejet diameteren en halv omgang er h-værdien det modsatte tal af det oprindelige. Undervejs har h altså været 0 mindst engang og i en sådan position snittes efter diameteren. Dermed er den ønskede deling opnået. (Också löst av *Hans Kaas Benner* och *Ebbe Thue Poulsen*)

513. (*Ebbe Thue Poulsen, Mårslet, DK*) Da udtrykket $a^2 + b^2 + c^2 + 4abc$ er symmetrisk i a, b og c, kan vi antage, at betegnelserne er valgt, så

$$a \le b \le c\,. \tag{1}$$

Da a, b og c er sider i en trekant, er

$$c < a + b\,, \tag{2}$$

og endvidere er det givet, at

$$a + b + c = 1\,. \tag{3}$$

Af (1), (2) og (3) følger dels, at

$$3c \ge a + b + c = 1$$

og

$$2c < a + b + c = 1\,,$$

der giver

$$\tfrac{1}{3} \le c < \tfrac{1}{2} \tag{4}$$

og dels, at

$$2a \le a + b = 1 - c$$

og

$$a = 1 - b - c \ge 1 - 2c\,,$$

der giver

$$1 - 2c \le a \le \tfrac{1}{2}(1 - c)\,. \tag{5}$$

Vi sætter $f(a,b,c) = a^2+b^2+c^2+4abc$, og omskriver med brug af (3):

$$\begin{aligned} f(a,b,c) &= (a+b)^2 + c^2 - 2ab(1-2c) \\ &= (1-c)^2 + c^2 - 2a(1-c-a)(1-2c)\,. \end{aligned} \tag{6}$$

For fastholdt $c < \frac{1}{2}$ er højre side af (6) et andengradspolynomium i a med minimum for $a = \frac{1}{2}(1-c)$. Heraf følger, at betragtet som funktion af a antager højre side af (6) sin største værdi på intervallet $[1-2c, \frac{1}{2}(1-c)]$ for $a = 1-2c$. Altså er, pga (4),

$$\begin{aligned} f(a,b,c) &\le (1-c)^2 + c^2 - 2(1-2c)^2 c \\ &= \tfrac{1}{2} + 2\left(c - \tfrac{1}{2}\right)^2 - 2(1-2c)^2 c \\ &= \tfrac{1}{2} + 2\left(c - \tfrac{1}{2}\right)^2 (1-4c) < \tfrac{1}{2}. \end{aligned}$$

(Också löst av *Erik Hansen*, *Con Amore Problemgruppe* och *Kåre Vedøy*)
Anm. Såväl *Erik Hansen* som *Con Amore Problemgruppe* påpekar (och visar) att under förutsättningarna i uppgiften gäller

$$\tfrac{13}{27} \le a^2+b^2+c^2+4abc < \tfrac{1}{2}.$$

Detta ger oss anledning att formulera följande *tilläggsuppgift:*
Visa att det för en triangel med sidlängderna a, b, c och omkretsen 1 gäller att

$$a^2+b^2+c^2+4abc \ge \tfrac{13}{27}.$$

När inträffar likhet?

514. (*Con Amore Problemgruppe, København, DK*) Vi skriver λ på formen $\lambda = re^{i\theta}$, hvor det er givet, at $r \ge 1$. Ved indsættelse i polynomiet fås så, idet vi sætter $\varepsilon = e^{-i\theta}$:

$$r^n\varepsilon^{-n} + a_{n-1}r^{n-1}\varepsilon^{-(n-1)} + \cdots + a_1 r\varepsilon^{-1} + a_0 = 0, \tag{1}$$

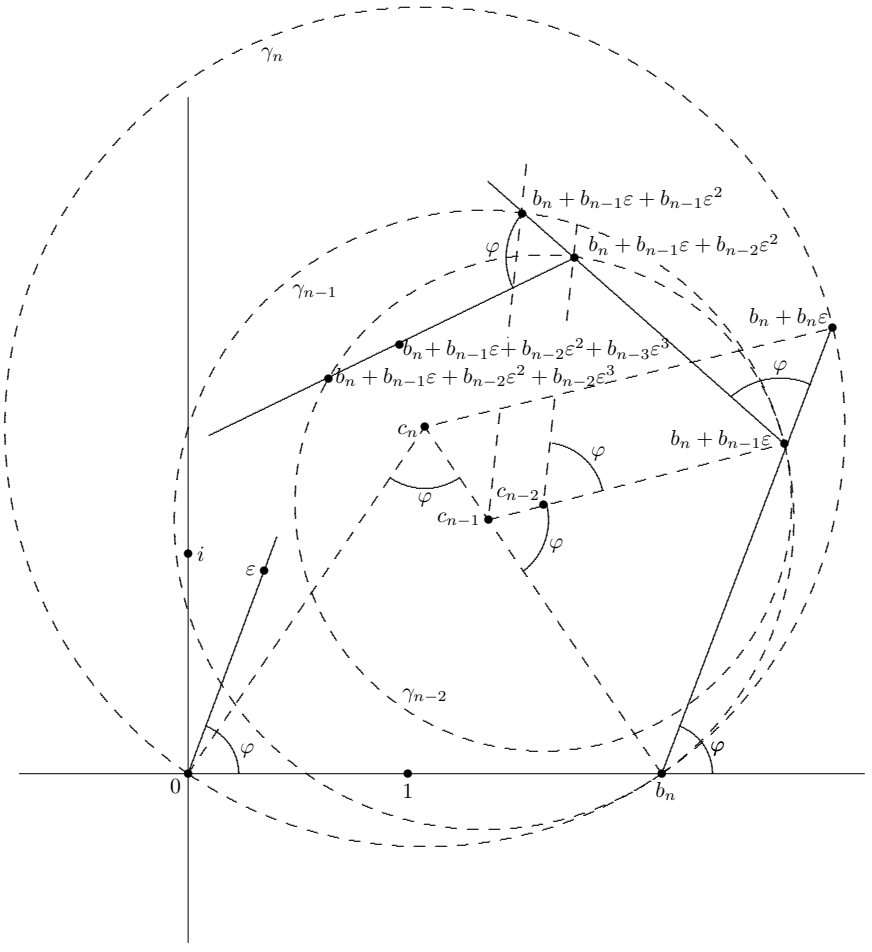
og videre, idet vi for kortheds skyld sætter $b_n = r^n, b_{n-1} = a_{n-1}r^{n-1}, \ldots, b_1 = a_1 r$ og $b_0 = a_0$ samt multiplicerer igennem med ε^n, at (1) er ensbetydende med

$$b_n + b_{n-1}\varepsilon + b_{n-2}\varepsilon^2 + \cdots + b_1\varepsilon^{n-1} + b_0\varepsilon^n = 0, \tag{2}$$

hvor

$$0 < b_0 \le b_1 \le \cdots \le b_{n-1} \le b_n. \tag{3}$$

Det fremgår let af (2), at ε ikke kan være 1, samt at hvis $\varepsilon = -1$, så må det gælde at $r = 1$ og n ulige, og dermed at $\lambda = -1$, og altså at $\lambda^{n+1} = 1$ som påstået.
Vi kan derfor nu forudsætte, at enhedsvektoren ε ligger överden reelle akse; thi med λ er jo også $\bar{\lambda}$ rod i polynomiet og kan derfor i givet fald betragtes i stedet for λ. Yderligere vil vi foreløbig forudsætte, at $\varepsilon^k \ne 1$ for $1 < k \le n$.
På nedenstående figur (hvor vi har benyttet φ som betegnelse for $-\theta$) har vi først fra punktet b_n afsat vektoren $b_{n-1}\varepsilon$.



Endepunktet for vektoren $b_n + b_{n-1}\varepsilon$ ligger ifølge (3) inden for eller på cirklen γ_n gennem punkterne 0, b_n og $b_n + b_n\varepsilon$. Fra punktet $b_n + b_{n-1}\varepsilon$ afsættes derefter vektoren $b_{n-2}\varepsilon^2$. Endepunktet for vektoren $b_n + b_{n-1}\varepsilon + b_{n-2}\varepsilon^2$ ligger tydeligvis inden for eller på cirklen γ_{n-1} gennem punkterne b_n, $b_n + b_{n-1}\varepsilon$ og $b_n + b_{n-1}\varepsilon + b_{n-1}\varepsilon^2$, som let ses att røre γ_n i b_n. Fra punktet $b_n + b_{n-1}\varepsilon + b_{n-2}\varepsilon^2$ afsættes derefter vektoren $b_{n-3}\varepsilon^3$. Endepunktet for vektoren $b_n + b_{n-1}\varepsilon + b_{n-2}\varepsilon^2 + b_{n-3}\varepsilon^3$ ligger tydeligvis inden for eller på cirklen γ_{n-2} gennem punkterne $b_n + b_{n-1}\varepsilon$, $b_n + b_{n-1}\varepsilon + b_{n-2}\varepsilon^2$ og $b_n + b_{n-1}\varepsilon + b_{n-2}\varepsilon^2 + b_{n-2}\varepsilon^3$, som let ses att røre γ_{n-1} i $b_n + b_{n-1}\varepsilon$. På den måde fortsættes, indtil vi har afsat vektoren

$$b_n + b_{n-1}\varepsilon + b_{n-2}\varepsilon^2 + b_{n-3}\varepsilon^3 + \cdots + b_1\varepsilon^{n-1} + b_0\varepsilon^n. \tag{4}$$

For cirklerne $\gamma_n, \gamma_{n-1}, \gamma_{n-2}, \ldots, \gamma_1, \gamma_0$ fremgår af ovenstående, at enhver af dem er indeholdt i den foregående, samt at γ_0 går gennem punkterne $b_n + b_{n-1}\varepsilon + b_{n-2}\varepsilon^2 + b_{n-3}\varepsilon^3 + \cdots + b_1\varepsilon^{n-1}$, $b_n + b_{n-1}\varepsilon + b_{n-2}\varepsilon^2 + b_{n-3}\varepsilon^3 + \cdots + b_1\varepsilon^{n-1} + b_0\varepsilon^n$ og $b_n + b_{n-1}\varepsilon + b_{n-2}\varepsilon^2 + b_{n-3}\varepsilon^3 + \cdots + b_1\varepsilon^{n-1} + b_0\varepsilon^n + b_0\varepsilon^{n+1}$, rører γ_1 i det første

af de nævnte punkter, og hvor det næste punkt netop er (4). Det følger heraf, at (4) er 0, hvis og kun hvis det gælder at

$$b_n = b_{n-1} = b_{n-2} = \cdots = b_1 = b_0. \tag{5}$$

Idet $r_n = b_n = b_{n-1} = a_{n-1}r^{n-1}$ medfører, at $r = a_{n-1}$, og da $r \geq 1$ og $a_{n-1} \leq 1$, må disse to tal begge være 1; og altså må det gælde, at $\lambda = \varepsilon^{-1}$. Ved bortforkortning af b_n i (2) samt multiplikation med λ^n følger derfor, at

$$0 = \lambda^n + \lambda^{n-1} + \cdots + \lambda^2 + \lambda + 1 = \frac{\lambda^{n+1} - 1}{\lambda - 1},$$

og dermed, at $\lambda^{n+1} = 1$ også gælder i den her betragtede situation.
Tilbage står det tilfælde, hvor det for et vist naturligt tal k med $1 < k \leq n$ gælder, at $\varepsilon^k = 1$. Lad da d være det mindste sådanne naturlige tal. Vi godtgør først ganske som ovenfor, at $r = 1$, og videre at $b_n = b_{n-1} = \cdots = b_{n-d+1}$; thi i modsat fald "foregår alt det følgende"inden for γ_n, muligvis bortset fra et fra 0 forskelligt punkt, dvs. (4) kan ikke være 0. Det gælder derfor at

$$b_n + b_{n-1}\varepsilon + \cdots + b_{n-d+1}\varepsilon^{d-1} = b_n \cdot (1 + \varepsilon + \cdots + \varepsilon^{d-1}) = b_n \cdot \frac{\varepsilon^d - 1}{\varepsilon - 1} = 0,$$

og følgelig ved benyttelse af dette i (2) samt bortforkortning af ε^d at

$$b_{n-d} + b_{n-d-1}\varepsilon + b_{n-d-2}\varepsilon^2 + \cdots + b_1\varepsilon^{n-d-1} + b_0\varepsilon^{n-d} = 0.$$

Ved gentagelse af dette argument et vist antal gange fås sluttelig, at det for et naturligt tal r med $1 \leq r \leq d$ gælder at

$$0 = b_{r-1} + b_{r-2}\varepsilon + \cdots + b_1\varepsilon^{r-2} + b_0\varepsilon^{r-1} = b_{r-1} \cdot (1 + \varepsilon + \cdots + \varepsilon^{r-1}) = b_{r-1} \cdot \frac{\varepsilon^r - 1}{\varepsilon - 1},$$

hvoraf det følger at $\varepsilon^r = 1$, og altså ifølge valget af d at $r = d$. Hermed har vi godtgjort, at der findes et naturligt tal m sådan at $md = n + 1$, og følgelig at

$$\lambda^{n+1} = \varepsilon^{-(n+1)} = (\varepsilon^d)^{-m} = 1.$$

Summary in English

David Cox, *Why Eisenstein proved the Eisenstein criterion and why Schönemann discovered it first.*
The Eisenstein criterion is well-known to all students of algebra, and it gives a very simple proof of the irreducibility of the cyclotomic polynomial $\Theta_p(x) = x^{p-1} + x^{p-2} + \cdots + 1$ for p prime. The irreducibility was (of course) known to Gauss but via a much more complicated proof. Simple ideas almost always arrive through tortuous detours, and the Eisenstein criterion being no exception, was originally stated for Gaussian integers in connection with polynomials associated to the division problem on the lemniscate! When it was published in Crelle in 1850, it provoked a complaint from a now forgotten mathematician Schönemann (1812-68) who had published a different proof of what was essentially the same criterion in the same journal only a few years earlier. His formulation is maybe even more elegant. Assume that $f(x) \in \mathbb{Z}[x]$ is of degree $n > 0$ and that there is some prime p and an integer a such that $f(x) = (x-a)^n + pF(x)$. If $F(a) \neq 0(p)$ then $f(x)$ irreducible mod p^2. This can be applied to the well known fact that $x^p - 1 = (x-1)^p(p)$ for p prime to get the irreducibility of $\Theta_p(x)$.

Schönemann not only deserves the priority for the Eisenstein criterion (and for some time his name was actually attached to Eisenstein's) but he also anticipated Hensel's lemma (which, however, in some weaker form was later found in the *Nachlaß* of Gauss), and did work out a theory for finite fields, although of course scooped by both Gauss and Galois. Yet the modern presentation of the latter subject did not appear until the very end of the 19th century when lectured on by E.H. Moore.

In addition to the main story a survey of the remarkable achievements of Gauss and Abel on elliptic functions is presented as providing the backdrop to the work of Schönemann and Eisenstein. Incidentally, the modern notion of an Abelian group, abstracted from the complicated equations on division points studied by Abel, did not appear until 1896 in Weber's classical *Lehrbuch der Algebra*.

Lars Gårding, *Näringskedjor* (Swedish).
A food-chain is an ordered collection of populations feeding on each other. A simple linear model for the variations of the populations is introduced. A condition for stability, i.e. that the variations of the populations are bounded, is that all the eigenvalues of the corresponding matrices lie on the unit-circle. The populations will then vary cyclically. The condition gives some restrictions on the fertility and life-expectancies of the various populations. Special cases of simple chains are studied and compared with real-life data, with some remarkable good fits.

Trond Steihaug and D.G. Rogers, *A minimum requiring angle trisection.*
Given a piece of foldable material (e.g. paper) in the shape of a right-angled triangle. Fold it by placing the right-angled corner on the hypothenuse. How should this be done in order to minimize the area of the folded triangle? The problem leads to a cubic equation whose relevant solution turns out to involve an angle trisection (and thus obtainable by a succession of foldings). Variations of the problem are considered, in particular letting one of the acute corners instead be placed on the opposite side.