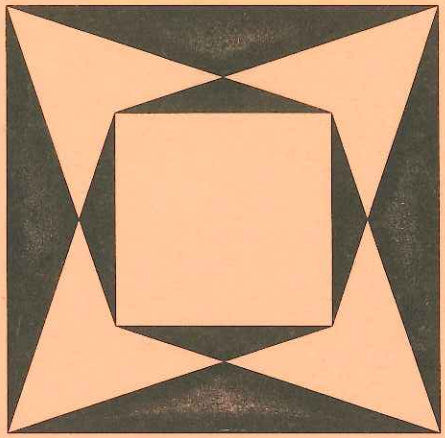
# IRISH MATHEMATICAL SOCIETY

# cumann matamaitice na héireann



BULLETIN

NUMBER 33 DECEMBER 1994

**ISSN 0791-5578**

**IRISH MATHEMATICAL SOCIETY**
**BULLETIN**

**EDITOR: Dr R. Gow**
**BOOK REVIEW EDITOR: Dr Michael Tuite**
**PROBLEM PAGE EDITOR: Dr Phil Rippon**
**PRODUCTION MANAGER: Dr Mícheál Ó Searcóid**

The aim of the Bulletin is to inform Society members about the activities of the Society and about items of general mathematical interest. It appears twice each year, in March and December. The Bulletin is supplied free of charge to members; it is sent abroad by surface mail. Libraries may subscribe to the Bulletin for IR£20.00 per annum.

The Bulletin seeks articles of mathematical interest written in an expository style. All areas of mathematics are welcome, pure and applied, old and new. The Bulletin is typeset using TEX. Authors are invited to submit their articles in the form of TEX input files if possible, in order to ensure speedier processing.

Correspondence concerning the Bulletin should be addressed to:

**Irish Mathematical Society Bulletin**
**Department of Mathematics**
**University College**
**Dublin**
**Ireland**

Correspondence concerning the Problem Page should be sent directly to the Problem Page Editor at the following address:

**Faculty of Mathematics**
**Open University**
**Milton Keynes, MK7 6AA**
**UK**

---

Printed in the University of Limerick



## CONTENTS

# cumann matamaiticе na héireann
# THE IRISH MATHEMATICAL SOCIETY

## Officers and Committee Members

| | | |
|---|---|---|
| **President** | Dr B. Goldsmith | President,<br>Dublin Institute of Technology<br>Kevin Street, Dublin |
| **Vice-President** | Dr D. Hurley | Department of Mathematics<br>University College, Cork |
| **Secretary** | Dr P. Mellon | Department of Mathematics<br>University College, Dublin |
| **Treasurer** | Dr M. Vandyck | Department of Physics<br>Regional Technical College, Cork<br>*also* University College, Cork |

Dr R. Timoney, Dr E. Gath, Dr J. Pulé, Dr G. Lessells, Dr R. Gow, Dr C. Nash, Dr B. McCann, Dr M. Ó Searcóid.

## Local Representatives

| | | |
|---|---|---|
| **Cork** | RTC | Mr D. Flannery |
| | UCC | Dr M. Stynes |
| **Dublin** | DIAS | Prof. J. Lewis |
| | Kevin St | Dr B. Goldsmith |
| | DCU | Dr M. Clancy |
| | St Patrick's | Dr J. Cosgrave |
| | TCD | Dr R. Timoney |
| | UCD | Dr F. Gaines |
| | Tallaght | Dr E. O'Riordan |
| **Dundalk** | RTC | Dr J. Harte |
| **Galway** | UCG | Dr R. Ryan |
| **Limerick** | MICE | Dr G. Enright |
| | UL | Dr E. Gath |
| | Thomond | Mr J. Leahy |
| **Maynooth** | | Prof. A. G. O'Farrell |
| **Waterford** | RTC | Mr T. Power |
| **Belfast** | QUB | Dr D. W. Armitage |



# NOTES ON APPLYING FOR I.M.S. MEMBERSHIP

1. The Irish Mathematical Society has reciprocity agreements with the American Mathematical Society and the Irish Mathematics Teachers Association.

2. The current subscription fees are given below.

| | |
|---|---|
| Institutional member | IR£50.00 |
| Ordinary member | IR£10.00 |
| Student member | IR£4.00 |
| I.M.T.A. reciprocity member | IR£5.00 |

   The subscription fees listed above should be paid in Irish pounds (puint) by means of a cheque drawn on a bank in the Irish Republic, a Eurocheque, or an international money-order.

3. The subscription fee for ordinary membership can also be paid in a currency other than Irish pounds using a cheque drawn on a foreign bank according to the following schedule:

   If paid in United States currency then the subscription fee is US$18.00.
   If paid in sterling then the subscription fee is £10.00 stg.
   If paid in any other currency then the subscription fee is the amount in that currency equivalent to US$18.00.

   The amounts given in the table above have been set for the current year to allow for bank charges and possible changes in exchange rates.

4. Any member with a bank account in the Irish Republic may pay his or her subscription by a bank standing order using the form supplied by the Society.

5. The subscription fee for reciprocity membership by members of the American Mathematical Society is US$10.00.

6. Subscriptions normally fall due on 1 February each year.

7. Cheques should be made payable to the Irish Mathematical Society. If a Eurocheque is used then the card number should be written on the back of the cheque.

8. Any application for membership must be presented to the Committee of the I.M.S. before it can be accepted. This Committee meets twice each year.

9. Please send the completed application form with one year's subscription fee to

   The Treasurer, I.M.S.
   Department of Physics
   Regional Technical College, Cork
   *also* University College, Cork
   Ireland



## Minutes of the Meeting of the Irish Mathematical Society

### Ordinary Meeting
31st March 1994

The Irish Mathematical Society held an ordinary meeting on Thursday 31st March 1994 in the Dublin Institute for Advanced Studies, 10 Burlington Road. Fourteen members were present. The president, B. Goldsmith, was in the chair.

1. The **minutes** of the meeting of 21st December 1993 were approved and signed.

2. **Matters arising**

It was reported that M. Tuite of UCG had been co-opted on to the committee.

3. There was no **correspondence**.

4. **Bulletin**

It was reported that the committee appointed the following members to form a bulletin editorial board: R. Gow, G. Lessells, M. Ó Searcóid and M. Tuite. The editor reported that the bulletin is fairly well on schedule but that more material is required for the 1995 issues. The editor encouraged members to submit articles which might appeal to a general mathematical audience.

5. **Subcommittee to manage interim affairs**

It was reported that the following committee members: D. Hurley, E. Gath, P. Mellon and M. Vandyck, were appointed to form a subcommittee which will manage the affairs of the society between committee meetings.

6. **Treasurer's business**

The treasurer's report was circulated. A proposal to accept this report was seconded and unanimously accepted.

As future bulletin publication costs will be considerable it was noted that collection of subscriptions should be a priority.

The treasurer reported that many members were in arrears with their subscriptions and announced that the subcommittee would be taking steps to correct this.

The possibility of corporate sponsorship and an increase in subscriptions were briefly discussed

### 7. September Meeting

Preparations for the September meeting are well under way. Provisional dates for the 1995 September meeting are 7th and 8th September.

### Other business

It was suggested that the subcommittee on behalf of the society make a formal submission to the Science, Technology and Innovation Advisory Council on matters to be considered in the production of the forthcoming white paper on science and technology policy. A discussion took place on this matter. The following points were made:

- In the development of science and technology it is counterproductive to take a short term economic perspective. In the long run, any "knowledge based" economy must have a solid research base in the mathematical sciences as this underpins most of today's scientific and technological research.
- The mathematical sciences community should be represented at every level of the "new" science council.

Members were asked to send their suggestions for the submission to the secretary via e-mail before 15th April. The subcommittee will consider such suggestions and formulate the submission by the end of April.

The meeting closed at 1.15pm.

Pauline Mellon
University College Dublin.

## Conference Announcement

*The Legacy of*
**GEORGE BOOLE**
UNIVERSITY COLLEGE, CORK, IRELAND
28TH-30TH JUNE, 1995

In 1995, University College, Cork will celebrate the 150th anniversary of its foundation. As part of this celebration, the University will hold a Conference honouring the genius of George Boole, who was its first Professor of Mathematics.

The last major Conference on Boole was held in 1954. Since then, significant advances have been made in many areas influenced by him, so the time is right for a re-assessment of his contribution to learning.

Speakers who have accepted the invitation to participate include:

G. K. Batchelor (University of Cambridge)
Robert L. Devaney (Boston University)
Keith Devlin (St Mary's College of California)
I. Grattan-Guinness (Middlesex University)
Theodore Hailperin (Lehigh University)
Desmond MacHale (University College, Cork)
John McCarthy (Stanford University)
Roger Penrose (University of Oxford)

ORGANIZING COMMITTEE

James Bowen (Computer Science, UCC)
Donal Hurley (Mathematics, UCC)
Desmond MacHale (Mathematics, UCC)
Lucette Murray (UCC 150)

For further enquiries: UCC 150 Office, UCC.
Tel: 021-276871 Ext: 2090 Fax: 021-276647 e-mail ucc150@iruccvax.ucc.ie

# REMARKS ON A PROBLEM OF FINBARR HOLLAND CONCERNING TRIGONOMETRIC POLYNOMIALS

David H. Armitage

Let $P_n$ denote the set of all non-negative trigonometric polynomials of degree at most $n$, normalized to have constant term equal to 1. Thus a typical element of $P_n$ has the form

$$p(t) = 1 + \sum_{j=1}^{n} (a_j \cos jt + b_j \sin jt) \geq 0 \quad \text{for all real } t.$$

A problem posed by Holland [1, Problem 4.26] essentially asks for the value of

$$\Lambda_n = \sup_{p \in P_n} \frac{1}{2\pi} \int_0^{2\pi} (p(t))^2 \, dt.$$

A much simpler problem is the determination of

$$M_n = \sup_{p \in P_n} \max p(t).$$

This was solved by Fejér [4] (or see [7; pp. 78-79]). It will be helpful to discuss this first, for it leads easily to rough bounds for $\Lambda_n$. Fejér showed that $M_n = n+1$; for a short proof see [2; §3.2]. He also showed that $M_n$ is an attained supremum: in fact if

$$q_n(t) = 1 + \frac{2}{n+1}(n \cos t + (n-1) \cos 2t + \cdots + \cos nt), \quad (1)$$

then $q_n \in P_n$ (for an easy calculation shows that

$$q_n(t) = \frac{1}{n+1} \left( \sin\{(n+1)\frac{t}{2}\} / \sin \frac{t}{2} \right)^2 \geq 0 \quad (0 < t < 2\pi))$$





and

$$\max q_n(t) = q_n(0) = 1 + \frac{2}{n+1}(n + (n-1) + \cdots + 1) = n+1.$$

Goldstein and McDonald [6] observed that Fejér's result leads to bounds on $\Lambda_n$ as follows. If $p \in P_n$, then

$$\frac{1}{2\pi} \int_0^{2\pi} (p(t))^2 dt \leq \frac{1}{2\pi} \max p(t) \int_0^{2\pi} p(t) dt = \max p(t) \leq n+1.$$

On the other hand,

$$\begin{aligned} \frac{1}{2\pi} \int_0^{2\pi} (q_n(t))^2 dt &= 1 + \frac{2}{(n+1)^2}(n^2 + (n-1)^2 + \cdots + 1^2) \\ &= 1 + \frac{n(2n+1)}{3(n+1)} \\ &> \frac{2}{3}(n+1). \end{aligned}$$

Hence $2/3 < \Lambda_n/(n+1) \leq 1$. In [6] there is further evidence favouring the conjecture that $(\Lambda_n/(n+1))$ converges to a limit $C \in [2/3, 1]$. In fact, a proof of this conjecture, yielding the value $C = 0.68698\ldots$, is implicit in earlier work of Garsia, Rodemich and Rumsey [5]. In work based partly on [5], Brown, Goldstein and McDonald [2, Theorem 2] showed further that $(n+1)C \leq \Lambda_n < 1 + (n+1)C$ for all $n \geq 1$. Quite intricate arguments are used in both [5] and [2], and it seems worthwhile to give an elementary, self-contained, and comparatively short proof of the existence of the limit $C$.

**Theorem.** *The sequence $(\Lambda_n/(n+1))$ converges to a limit $C$ in $[2/3, 1]$ and*

$$C = \inf_{n \geq 1} \Lambda_n/(n+1). \quad (2)$$

The main step in our proof is to establish the inequality

$$\frac{\Lambda_{nk+k-1}}{nk+k} \leq \frac{\Lambda_n}{n+1} \qquad (k \geq 2, n \geq 1). \quad (3)$$

Suppose for the moment that (3) is true and let $C$ be defined by (2). Fix $\epsilon > 0$ and let $N$ be such that $\Lambda_N/(N+1) < C + \epsilon$. If $n > N+1$ and $k(n)$ is the least integer such that $(N+1)k(n) > n$, then $(N+1)k(n) \le n+N+1$, and hence using (3) and the obvious fact that $(\Lambda_n)$ is non–decreasing, we obtain

$$\begin{aligned}\frac{\Lambda_n}{n+1} &\le \frac{\Lambda_{Nk(n)+k(n)-1}}{(N+1)k(n)} \cdot \frac{(N+1)k(n)}{n+1} \\ &\le \frac{\Lambda_N}{N+1}(1+\frac{N}{n+1}),\end{aligned}$$

so that $\limsup \Lambda_n/(n+1) < C+\epsilon$ and hence $\Lambda_n/(n+1) \to C$.

We write

$$J(p) = \frac{1}{2\pi}\int_0^{2\pi} (p(t))^2\,dt.$$

To prove (3), it suffices to show that if $p \in P_{nk+k-1}$, then

$$J(p) \le k\Lambda_n. \tag{4}$$

Let such a function $p$ be given by

$$p(t) = 1 + \sum_{j=1}^{nk+k-1} (a_j \cos jt + b_j \sin jt).$$

Since $p \ge 0$,

$$\begin{aligned}0 &\le \frac{1}{2\pi}\sum_{m=1}^{k-1}\int_0^{2\pi} p(t)p(t+2m\pi/k)dt \\ &= k - 1 + \frac{1}{2}\sum_{j=1}^{nk+k-1}\left\{(a_j^2+b_j^2)\sum_{m=1}^{k-1}\cos(2mj\pi/k)\right\} \\ &= k - 1 - \frac{1}{2}\sum_{j=1}^{nk+k-1}(a_j^2+b_j^2) + \frac{1}{2}k\sum_{\ell=1}^{n}(a_{\ell k}^2+b_{\ell k}^2);\end{aligned}$$



the last-written equation follows from the fact that

$$\sum_{m=1}^{k-1}\cos(2mj\pi/k) = \begin{cases} k-1 & \text{if } k|j \\ -1 & \text{if } k \not| j.\end{cases}$$

Hence

$$J(p) = 1 + \frac{1}{2}\sum_{j=1}^{nk+k-1}(a_j^2+b_j^2) \le k\left(1 + \frac{1}{2}\sum_{\ell=1}^{n}(a_{\ell k}^2+b_{\ell k}^2)\right). \tag{5}$$

Note that

$$1 + \frac{1}{2}\sum_{\ell=1}^{n}(a_{\ell k}^2+b_{\ell k}^2) = J(q), \tag{6}$$

where

$$q(t) = 1 + \sum_{\ell=1}^{n}(a_{\ell k}\cos \ell t - b_{\ell k}\sin \ell t). \tag{7}$$

If we can show that $q$ is non-negative, then we shall have $q \in P_n$ and hence $J(q) \le \Lambda_n$. From (5) and (6) it will then follow that $J(p) \le kJ(q) \le k\Lambda_n$, and (4) and hence (3) will be established.

To show that $q$ is non-negative, we first associate to $p$ the harmonic polynomial $h$ defined by

$$h(re^{it}) = 1 + \sum_{j=1}^{nk+k-1} r^j(a_j\cos jt + b_j \sin jt).$$

Let $\Delta$ denote the unit disc. Since $h(e^{it}) = p(t) \ge 0$ for all $t \in [0, 2\pi]$, we have $h \ge 0$ on $\partial\Delta$ and hence, by the minimum principle, $h \ge 0$ on $\Delta$. Also define $K$ on $\Delta$ by

$$K(re^{it}) = 1 + 2\sum_{\ell=1}^{\infty} r^\ell \cos \ell t. \tag{8}$$

It is easy to verify that

$$K(re^{it}) = \frac{1-r^2}{1-2r\cos t + r^2} > 0 \qquad (re^{it} \in \Delta).$$

(In fact, $K$ is the Poisson kernel of $\Delta$ with pole 1.) Since the series in (8) is locally uniformly convergent on $\Delta$, we have for all $r \in (0,1)$ and all real $\theta$,

$$0 \le \frac{1}{2\pi}\int_0^{2\pi} h(re^{it})K(re^{i(kt+\theta)})dt$$

$$= 1 + \frac{1}{\pi}\sum_{\ell=1}^{\infty} r^{\ell}\left(\sum_{j=1}^{nk+k-1} r^j \int_0^{2\pi} (a_j \cos jt + b_j \sin jt)\cos(\ell kt + \ell\theta)dt\right)$$

$$= 1 + \sum_{\ell=1}^{n} r^{\ell+\ell k}(a_{\ell k}\cos \ell\theta - b_{\ell k}\sin \ell\theta).$$

Letting $r \to 1-$, we find that the function $q$ defined by (7) is indeed non-negative and, as explained earlier, (3) now follows and therefore $(\Lambda_n/(n+1))$ converges to the limit $C$ given by (2). The bounds on $\Lambda_n$ obtained from Fejér's work show that $2/3 \le C \le 1$.

Calculations using Mathematica and based on a representation of $\Lambda_n$ obtained by Goldstein and McDonald [6, Corollary 2] suggest the values given in the table below. I am grateful to Tony Wickstead for his help with these calculations. Our values for $\Lambda_2, \ldots, \Lambda_5$ confirm those obtained in [6, p.87], except for a small discrepancy in the value of $\Lambda_3$.

| $n$ | $\Lambda_n$ | $\Lambda_n/(n+1)$ |
|---|---|---|
| | | .75 |
| 1 | 1.5 | |
| 2 | 2.142857142... | .714285714... |
| 3 | 2.808840165... | .702210041... |
| 4 | 3.4834502..... | .6966900..... |
| 5 | 4.1622565..... | .6937094..... |
| 6 | 4.8434275..... | .6919182..... |
| 7 | 5.5260645..... | .6907580..... |
| 8 | 6.2096738..... | .6899637..... |
| 9 | 6.8939613..... | .6893961..... |

To the best of my knowledge, the conjecture that $(\Lambda_n/(n+1))$ is decreasing remains open.

One obvious generalization of Holland's question concerns

$$\Lambda_{n,\alpha} = \sup_{p \in P_n} \frac{1}{2\pi}\int_0^{2\pi} (p(t))^{\alpha} dt \qquad (\alpha > 0).$$



If $0 < \alpha < 1$, then Hölder's inequality shows that $\Lambda_{n,\alpha} \le \Lambda_{n,1} = 1$, and since we can always take $p(t) \equiv 1$, it follows that $\Lambda_{n,\alpha} = 1$ for all $n$. If $\alpha > 1$, then there exists a positive constant $c_\alpha$ such that

$$c_\alpha (n+1)^{\alpha-1} \le \Lambda_{n,\alpha} \le (n+1)^{\alpha-1}.$$

Here the upper bound is obtained from Fejér's result $M_n = n+1$ and the lower bound is obtained by estimating

$$\int_0^{2\pi} (q_n(t))^{\alpha} dt,$$

where $q_n$ is given by (1). It seems plausible that $\lim_{n\to\infty}(n+1)^{1-\alpha}\Lambda_{n,\alpha}$ exists when $\alpha > 1$, but this appears to be an open question, except for $\alpha = 2$.

**References**


[1] J. M. Anderson, K. F. Barth and D. A. Brannan, *Research problems in complex analysis*, Bull. London Math. Soc. **9** (1977), 129-162.

[2] D. H. Armitage, *The Poisson kernel as an extremal function*, Irish Math. Soc. Bulletin **32** (1994), 19-31. .

[3] J. Brown, M. Goldstein and J. McDonald, *A sequence of extremal problems for trigonometric polynomials*, J. Math. Anal. Appl. **130** (1988), 545-551.

[4] L. Fejér, *Über trigonometrische Polynome*, J. Reine Angew. Math. **146** (1916), 53-82.

[5] A. Garsia, E. Rodemich and H. Rumsey, *On some extremal positive definite functions*, J. Math. Mech. **18** (1969), 805-834.

[6] M. Goldstein and J. N. McDonald, *An extremal problem for non-negative trigonometric polynomials*, J. London Math. Soc. (2) **29** (1984), 81-88.

[7] G. Pólya and G. Szegö, Problems and theorems in analysis, vol. II. Springer: 1976.



D. H. Armitage,
Department of Pure Mathematics,
The Queen's University of Belfast,
Belfast BT7 1NN,
Northern Ireland.

# ANALYSIS AND TOPOLOGY IN MATHEMATICAL ECONOMICS

Alan F. Beardon

**Abstract** This article is an expanded set of notes of a lecture given at University College, Cork in April 1993. Mathematical economics provides a fertile source of applications of general topology, and we illustrate this here, as well as discussing some topics which warrant further study.

## 1. The consumer

We wish to model the behaviour of a consumer faced with the task of buying quantities $x_1, \ldots, x_n$ of $n$ *goods* $G_1, \ldots, G_n$. The *bundle* of goods is $x = (x_1, \ldots, x_n)$ and, to avoid boundary conditions, we usually assume that $x$ lies in the open set

$$\Omega = \{\, x \in \mathbf{R}^n : x_j > 0, j = 1, \ldots, n \,\}.$$

The good $G_j$ has a *unit price* $p_j$, the *price vector* $p$ is $(p_1, \ldots, p_n)$, and the *cost* of the bundle $x$ is the scalar product $p.x$. Given two bundles $x$ and $y$, the consumer is assumed to have a (weak) preference for one of them and, formally, this is described as follows.

**Definition** A *weak preference relation* is a binary relation $\succeq$ on $\Omega$ which is

(1) *complete* (either $x \succeq y$ or $y \succeq x$; in particular, $x \succeq x$), and
(2) *transitive* ($x \succeq y$ and $y \succeq z$ implies $x \succeq z$).

We make the natural definitions (i) $x \sim y$ if and only if $x \succeq y$ and $y \succeq x$ (the consumer is then said to be *indifferent* between $x$ and $y$), and (ii) $x \succ y$ if and only if $x \succeq y$ but not $x \sim y$ (the consumer then has a *strict preference* for $x$). Of course, we assume that each good is desirable, so that if the bundle $x$ contains as much of each good as bundle $y$, and more of some good, then





$x \succ y$. Finally, we allow ourselves to use $x \prec y$ for $y \succ x$ and similarly for $\preceq$.

It is easy to see that $\sim$ is an equivalence relation. The $\sim$equivalence class of $x$ is denoted by $I(x)$ and is called the *indifference class* of $x$. In general, each $I(x)$ is an $(n-1)$-dimensional manifold (for example, a curve if $n = 2$) and these sets play a basic role in any discussion of the standard problems in economics. We may ask, for example, what is the most preferred bundle that can be purchased with a fixed sum of money; do we always buy less of $G_j$ when $p_j$ increases (the answer is 'no'), and so on. For a general account of these ideas and those in the next section, see [1], [2], [8] and [10].

## 2. Utility functions

The simplest way to construct a preference relation $\succeq$ on $\Omega$ is to take a real-valued function $u(x_1, \ldots, x_n)$ that is strictly increasing in each variable $x_j$, and then define $x \succ y$ if and only if $u(x) > u(y)$. With this, the indifference classes are the level sets of $u$, and we say that $u$ is a *utility function representing* $\succeq$. As a (popular) example, we mention the Cobb-Douglas utility function given by

$$u(x_1, \ldots, x_n) = x_1^{a_1} \ldots x_n^{a_n}, \quad a_j > 0.$$

It should be noted that we do not attach any significance to the numerical value of $u(x)$, but only to the relative values of $u(x)$ and $u(y)$. It follows that if $u$ is a utility function representing $\succeq$, then so is $h(u(x))$ for any strictly increasing real function $h$.

A significant part of the theory is devoted to the problem of when (or how) can we represent a given preference relation $\succeq$ by a utility function. This is not a trivial question for, as the next example shows, *such a representation is not always possible.*

**Example: the lexicographic ordering** Take $n = 2$ and define $(x_1, x_2) \succeq (y_1, y_2)$ if and only if either $x_1 > y_1$, or both $x_1 = y_1$ and $x_2 \geq y_2$. Note that in this case $I(x) = \{\, x \,\}$, a single point. To see that this relation cannot be represented by a utility function, we simply observe that if this were possible, each vertical line in $\Omega$ would map by $u$ into an open interval, and these intervals

would constitute an uncountable number of disjoint non-empty open intervals in $\mathbf{R}$. As this cannot be done, the function $u$ cannot exist.

It is natural to attempt to represent a preference relation by a *continuous* utility function $u$, and if $\succeq$ can be so represented then, for each $y$ in $\Omega$, the set

$$A(y) = \{\, x \in \Omega : x \succ y \,\} = \{\, x \in \Omega : u(x) > u(y) \,\}$$
$$= u^{-1}(u(y), +\infty)$$

is open, as is $B(y) = \{\, x \in \Omega : y \succ x \,\}$. In fact, these conditions are sufficient.

**Theorem 2.1.** *A preference relation $\succeq$ on $\Omega$ can be represented by a continuous utility function if and only if for all $y$ in $\Omega$, the sets $A(y)$ and $B(y)$ are open.*

*Proof:* We have to construct a continuous utility function given that all the sets $A(y)$ and $B(y)$ are open. Let $D$ be the 'diagonal' in $\Omega$ (given by $x_1 = x_2 = \cdots = x_n$) and consider the bundle $y$. It is not hard to see that $I(y)$ meets $D$ (otherwise $A(y)$ and $B(y)$ disconnect $D$) and, as the goods are desirable, $I(y) \cap D$ must be a single point, say $y_D$. We now define $u : \Omega \to \mathbf{R}$ by $u(y) = \|y_D\|$, and, because the sets $A(y)$ and $B(y)$ are open, it is easy to show that $u$ is continuous. This result is a simplified version of the more general result in the fundamental paper [21].

The construction of the map $y \mapsto y_D$ applies not just to $D$ but to any ray from the origin, and this really means that, in the circumstances described in Theorem 2.1, each indifference class is radially homeomorphic to that part of the unit sphere lying in $\Omega$, and so is an $(n-1)$-dimensional manifold.

### 3. Abstract preference relations

We now turn to discuss a preference relation $\preceq$ defined on an arbitrary non-empty set $X$ (such circumstances are of interest to some economists). The definition remains valid, but we lose the concept of having 'more' of a good, and the topology on $X$ (if there is one) may be quite bizarre. However, the induced indifference relation $\sim$ is still an equivalence relation, and the problem



of representing $\preceq$ by a utility function remains. We say that the pair $(X, \preceq)$ is a *commodity space*.

The question of representation of $\preceq$ by a *continuous* utility function implies the existence of a topology on $X$, and even if $X$ comes equipped with a topology (for example, the Euclidean topology), there is no reason at all to suppose that this topology should be related in any intrinsic way to $\preceq$. On the other hand, there *is* a natural topology on $X$, namely the *order topology* $\mathcal{T}_O$ generated by the intervals $\{\, x : x \succ y \,\}$ and $\{\, x : y \succ x \,\}$, and this topology is obviously equivalent to the preference relation used to define it. Note that the condition given in Theorem 2.1 can now be rephrased as the set-theoretic inclusion $\mathcal{T}_O \subset \mathcal{E}$, where $\mathcal{E}$ is the Euclidean topology on $\Omega$. This type of inclusion arises frequently, and for good reason. The natural question concerning continuity is continuity with respect to the order topology $\mathcal{T}_O$, and if $u$ is continuous with respect to $\mathcal{T}_O$, and if $\mathcal{T}_O \subset \mathcal{T}$, then $u$ is also continuous with respect to $\mathcal{T}$. For more details, see [2], [9], [10] and [14].

We mention, in passing, that if $\mathcal{S}$ is the closed unit square in $\mathbf{R}^2$ ordered lexicographically, then the topological space $(\mathcal{S}, \mathcal{T}_O)$ is an example (different from the usual $\sin(1/x)$ curve) of a space that is connected but not arcwise connected.

### 4. Existence of utility functions

If we now consider the commodity space $(X, \preceq)$ with the order topology, the quotient space $X/\sim$ is linearly ordered in a natural way and the order topology on this is indeed the quotient topology. Abstractly, then, the existence of a utility function is equivalent to the quotient space $X/\sim$ being order-isomorphic to a subset of $\mathbf{R}$, and there are various results of this type available. For example, we have (see [12])

**Theorem 4.1.** *An ordered set is order-isomorphic to a subset of $\mathbf{R}$ if and only if it has a countable dense subset (in the order topology), and has only countably many pairs $x$ and $y$ such that $x \prec y$ and $(x, y) \cap X = \emptyset$.*

We call such a pair $\{\, x, y \,\}$ a *jump*. If a linearly ordered set $X$

has a jump $\{x, y\}$, then the disjoint open intervals $(-\infty, y)$ and $(x, +\infty)$ disconnect $X$, so we have (see [11] and [12])

**Corollary 4.2.** *If a commodity space $(X, \preceq)$ is connected in its order topology, or in any larger topology, and if it has a countable dense subset, then $\preceq$ can be represented by a utility function.*

Another existence result ([2], [9], [10] and [19]) is

**Theorem 4.3.** *Let $(X, \preceq)$ be a commodity space. Then $\preceq$ can be represented by a utility function if and only if the order topology $\mathcal{T}_O$ is second countable.*

*Proof:* It is clear that if a utility function exists, then $\mathcal{T}_O$ is second countable (because $\mathbf{R}$ is). Now let $O_1, O_2, \ldots$ be a countable basis for $\mathcal{T}_O$. Given $x$, define

$$N(x) = \{\, n : O_n \subset (-\infty, x) \,\}, \quad v(x) = \sum_{n \in N(x)} \frac{1}{2^n}.$$

It is clear that $v(x)$ is weakly increasing with preferences. If $x \prec y$, then $(-\infty, x)$ is a *proper* subset of $(-\infty, y)$ (because the latter set contains $x$) so that $N(x)$, and hence $v(x)$, is strictly increasing with preferences and $v$ is the required utility function. This completes the proof.

Yet another approach is to mimic the idea in the proof of Theorem 2.1. Suppose for the moment, that a topological space $(X, \mathcal{T})$ is arcwise connected, and that it supports a preference relation $\preceq$ with respect to which there is a maximally preferable point $z$ and a minimally preferable point $w$. Suppose also that $\mathcal{T}_O \subset \mathcal{T}$. Join $z$ and $w$ by a curve $\gamma$ in $X$. As in the proof of Theorem 2.1, every indifference class meets $\gamma$ and provided that we can construct a utility function on $\gamma$, we can extend this to a utility function on $X$. With a little more work these ideas lead to (see [17])

**Theorem 4.4.** *Let $(X, \mathcal{T})$ be an arcwise connected topological space, and let $\preceq$ be a preference relation on $X$ such that $\mathcal{T}_O \subset \mathcal{T}$. If $X$ contains a countable subset $X_0$ such that for all $x$ in $X$ there*



*are points $a$ and $b$ in $X_0$ such that $a \preceq x \preceq b$, then $\preceq$ can be represented by a utility function.*

Finally, we observe that each preference relation $\succeq$ on $X$ is a subset of $X \times X$, and that this space carries the product topology. As the class of all subsets of a topological space also carries a natural (product) topology, it follows that *there is a natural topology on the space of all preference relations.* With this, we can begin to discuss more sophisticated results (and problems): for example, under certain circumstances, we can find utility functions which are *jointly continuous* with respect to both $x$ and $\preceq$.

## 5. The non-existence of utility functions

Preference relations which cannot be represented by a utility function are hard to find; indeed, the lexicographic order (or some variation of it) seems to be the only explicit example that is known. A more sophisticated (and implicit) example is that of the 'long-line'.

**Example: the long-line** A linearly ordered set $X$ is *well-ordered* by $<$ if every non-empty subset $X_0$ of $X$ has a smallest element, and (assuming the Axiom of Choice) every set can be well-ordered. Let $A$ be any well-ordered, uncountable set ordered, say, by $<$. Now construct another well-ordered set $B$ as follows:
*Case 1*: if, for each $a$ in $A$, $(-\infty, a)$ is countable, we put $B = A$;
*Case 2*: if there exists some $a$ for which $(-\infty, a)$ is uncountable, let $b$ be the smallest such $a$ (which exists as $A$ is well-ordered) and put $B = (-\infty, b)$.
It is immediate that (in both cases)
(1) $B$ is well-ordered (it is a subset of $A$);
(2) for all $x$ in $B$, $(-\infty, x)$ is countable;
(3) $B$ is uncountable;
(4) for each $x$ in $B$, there is a unique $x'$ in $B$ such that $x' < x$ and $(x, x') = \emptyset$.
Note that (4) holds because given $x$ in $B$, (2) and (3) imply that there is some $t$ in $B$ with $x < t$, and the well-ordering implies that the set of all such $t$ has a smallest member $x'$. The key properties of $B$ are expressed in the next result.

**Theorem 5.1.** *The ordered space $(B, <)$ has the properties*
*(a) for each $x$ in $B$, the order $<$ on $(-\infty, x)$ can be represented by a real-valued function;*
*(b) the order $<$ cannot be represented by a real-valued function on $B$.*

*Proof:* It is well-known (and easy to prove) that any countable set can be mapped in an order-preserving way into the set of rational numbers; thus (a) follows. On the other hand, $<$ cannot be represented on $B$ by a real-valued function $u$, for if it could, then the intervals $(u(x), u(x'))$ as $x$ varies over $B$ would constitute an uncountable set of pairwise disjoint subintervals of $\mathbf{R}$ and we know that this cannot happen. Note that (b) is also a consequence of Theorem 4.1 as, by (3) and (4), $B$ has uncountably many jumps.

There are results which suggest that, under fairly weak conditions, any preference relation which cannot be represented by a utility function must look (roughly speaking) rather like either the lexicographic order or the long-line ([17]). This seems to me to be an area worthy of much more study; there is a need to understand and illustrate the reasons why a preference relation cannot be represented by a utility function.

## 6. The existence of continuous utility functions

It is a rather surprising fact that the continuity of utility functions with respect to the order topology is not an issue at all. We have

**Theorem 6.1.** *If $\preceq$ can be represented by a utility function on $X$, then it can also be represented by a utility function that is continuous with respect to the order topology on $X$.*

Clearly, if $\preceq$ can be represented by a continuous utility function $u$ on $X$, then $h(u(x))$ is also a continuous utility function for every continuous map $h : u(X) \to \mathbf{R}$. It is natural to identify (that is, not distinguish between) the functions $u$ and $hu$ when $h$ is a homeomorphism, and with this we have the following uniqueness result.

**Theorem 6.2.** *Any two continuous utility functions representing $\preceq$ differ by composition with a homeomorphism.*



The original proof of Theorem 6.1 was given by the economist Debreu (1952) who, for this purpose, proved

**Theorem 6.3: the Gap Theorem.** *Let $E$ be a subset of $\mathbf{R}$. Then there is a strictly increasing map $\phi : E \to \mathbf{R}$ such that the complement of $\phi(E)$ has no bounded components of the form $[a, b)$ or $(a, b]$.*

The significance of the Gap Theorem is that if we have a utility function $u$ on $X$, and if we take $E = u(X)$, then we find that $\phi u$ is a continuous utility function. The original 'proof' of the Gap Theorem was simply to 'collapse' all of the offending intervals $[a, b)$ and $(a, b]$ to a single point: however, it was soon realized that this argument is not valid because in some cases (in which $E$ has measure zero) this would also collapse $E$ to a single point. A valid argument seems to require (in one form or another) a process which *at the same time* 'expands' $E$ (possibly from zero measure to positive measure) and collapses the intervals in its complement. There are now a variety of proofs of the Gap Theorem available (see [3], [5], [7], [9], [13], [16] and [20]), and the next result [5] (based on the fundamental paper [21] by the economist Wold in 1943) contains, as a Corollary, the Gap Theorem and several other important results in this area.

**Theorem 6.4.** *Let $\sim$ be an equivalence relation on the closed interval $[0, 1]$ with the property that each equivalence class is a closed interval. Then there is an increasing continuous function $u : [0, 1] \to [0, 1]$ such that $u(x) = u(y)$ if and only if $x \sim y$.*

To relate the Gap Theorem to a more concrete example, consider for the moment a strictly increasing map $f$ of $[0, 2]$ to $[0, 1] \cup (2, 3]$. The map is not continuous (because the complement of its range has a component that is a half-open half-closed interval) and, equally importantly, the set $[0, 1]$ is open in the subspace topology of $f(E)$ but *not* in the intrinsic order topology (namely, the order topology derived from the induced order on $f(E)$). More generally, given a subset $K$ of a linearly ordered space $X$, $K$ carries both the subspace topology and its own intrinsic order topology (found by first restricting the order to $K$ and then creating

the corresponding order topology), and these two topologies need not be the same. The 'gap condition' on the range $\phi(E)$ in the Gap Theorem is precisely the condition that these two topologies coincide.

These ideas lead to a study of general questions about continuity of increasing functions between linearly ordered spaces, and we can establish the following general result (see [4]).

**Theorem 6.5.** *Let $X$ and $Y$ be linearly ordered spaces, and suppose that the subspace topology and the intrinsic order topology on a subset $E$ of $X$ are the same. Then for every strictly increasing map $f : E \to Y$, the map $f^{-1} : f(E) \to E$ is continuous (with respect to the subspace topologies).*

Observe first that the result generalizes the well-known elementary result that if $f : [a, b] \to \mathbf{R}$ is strictly increasing, then $f^{-1} : f([a, b]) \to [a, b]$ is continuous (most elementary texts assume, unnecessarily, that $f$ is continuous). Indeed, this follows from Theorem 6.5 because on any compact interval the subspace and intrinsic order topologies coincide.

Next, if we consider an increasing map $f : E \to f(E)$ which has the property that for both $E$ and $f(E)$ the subspace and intrinsic order topologies coincide, then, by Theorem 6.5, $f$ is a homeomorphism from $E$ to $f(E)$. If this is taking place in the context of the real line, the statement about the two topologies can be replaced by a statement about half-open half-closed gaps and this leads, ultimately, to the uniqueness expressed in Theorem 6.2.

## 7. Utility functions with values in a linearly ordered group

A major unresolved question is *what can be said if a given preference relation cannot be represented by a utility function?* From a purely mathematical point of view, the insistence that utility functions be real-valued is bound to lead to difficulties for, after all, if $X$ is 'very' large, then the range of any utility function on $X$ must necessarily be correspondingly large to cope with this, and in many case it will be significantly larger than $\mathbf{R}$. Indeed, it can be



argued that to reject utility functions whose range is 'larger' than $\mathbf{R}$ is simply refusing to tackle the real issue. It is natural, then, to try to develop a theory of utility functions whose range can be any suitably large ordered set. Without further restrictions, however, this is an empty problem (for each ordered set $X$ can be represented by the identity map into itself, or by the quotient map onto $X/\sim$). We need, then, to find a class of ordered sets with some additional structure, and then allow a utility function to map into one of these sets.

One possibility is to seek utility functions with values in a suitable *linearly ordered group* (that is a group which is ordered in a way that is compatible with the group operation) and to help the theory along, there is a large theory of linearly ordered groups available, [15]. Moreover, the theory of linearly ordered groups contains much on lexicographically ordered products of groups, so there does seem to be a strong link here with the earlier discussion. It is perhaps worth noting that with its order topology, a linearly ordered group becomes a topological group (that is, the operations $(g, h) \mapsto gh$ and $g \mapsto g^{-1}$ are continuous).

In fact, it is known that any preference relation $\preceq$ on any space $X$ *can* be represented by a utility function with values in an abelian linearly ordered group. To see this, we take the group of integer-valued functions on $X$ which are zero except at a finite set of points of $X$. The group operation is addition in the usual way, and we write $f < g$ if $f(x) < g(x)$, where $x$ is the least preferable point of $X$ at which $f$ and $g$ disagree

We have not really solved our problem, however, for as a linearly ordered group is an ordered space, it supports its own order topology and knowing this, we must surely seek the existence of a *continuous* utility function. In this respect, there is an amusing, and tantalizing, observation to be made. The very example that is universally quoted as a preference relation which cannot be represented by a utility function, namely the lexicographic order on the first quadrant, is itself a linearly ordered group with the group operation

$$(x, y) \oplus (u, v) = (xu, yv).$$

This means, of course, that the lexicographic preference relation

*can be represented by a continuous utility function with values in an abelian linearly ordered group*, namely the identity function mapping $\Omega$ onto itself!

We end with an avenue for further study. The Gap Theorem has been proved for maps from $\mathbf{R}$ to $\mathbf{R}$, and, in the light of the remarks just made, we now ask for which linearly ordered groups is there a corresponding 'Gap Theorem' available? If we could show that $\mathbf{R}$ is the only linearly ordered group for which such a result is true, then this would provide a *mathematical* justification for restricting our attention to real-valued utility functions. If, on the other hand, there were other groups for which such a result existed, it might lead to a theory of more general utility functions which would, on mathematical grounds at least, have an equal claim to our attention.

**References**


[1] K. J. Arrow and F. Hahn, General Competitive Analysis. Oliver and Boyd: 1988.

[2] K. J. Arrow and M. D. Intriligator, Handbook of Mathematical Economics, Volume II. North Holland: 1982.

[3] A. F. Beardon, *Debreu's Gap Theorem*, Economic Theory **2** (1992a), 150-152.

[4] A. F. Beardon, *Utility theory and continuous monotonic functions*, Economic Theory, to appear.

[5] A. F. Beardon and G. B. Mehta, *The utility theory of Wold, Debreu and Arrow-Hahn*, Econometrica, to appear.

[6] G. Birkhoff, Lattice Theory. Amer. Math. Soc. Pub. 25: New York, 1948.

[7] R. Bowen, *A new proof of a theorem in utility theory*, International Economic Review **9** (1968), 374.

[8] J. W. S. Cassels, Economics for Mathematicians. London Math. Soc. Lecture Notes, No. 62. Cambridge University Press: 1981.

[9] G. Debreu, *Representation of a preference ordering by a numerical function*, *in* Decision Processes, ed. by R. Thrall, C. Coombs and R. Davis, 159-165. Wiley: 1954.

[10] G. Debreu, Mathematical Economics, Chapters 6 and 12. Cambridge University Press: 1983.





[11] S. Eilenberg, *Ordered topological spaces*, Amer. J. Math. **63** (1941), 39-45.

[12] I. Fleischer, *Numerical representations of utility*, Journ. S.I.A.M **9** (1961), 48-50.

[13] J. Jaffray, *Existence of a continuous utility function: an elementary proof*, Econometrica **43** (1975), 81-83.

[14] E. Klein and A. C. Thompson, Theory of Correspondences. Wiley Interscience: 1984.

[15] A. I. Kokorin and V. M. Kopytov, Fully Ordered Groups. Wiley: 1974.

[16] G. B. Mehta, *Recent developments in utility theory*, Indian Economic Journal **30** (1983), 103-124.

[17] P. K. Monteiro, *Some results on the existence of utility functions on path connected spaces*, J. Mathematical Economics **16** (1987), 147-156.

[18] L. Nachbin, Topology and Order. Van Nostrand: 1965.

[19] T. Rader, *The existence of a utility function to represent preferences*, Review of Economic Studies **30** (1963), 229-232.

[20] M. Richter, *Continuous and semi-continuous utility*, International Economic Review **21** (1980), 293-299.

[21] H. Wold, *A synthesis of pure demand analysis, I, II and III*, Scandinavisk Aktuarietidskrift **26** (1943-44), 85-118, 220-263, 69-120.



A. F. Beardon,
Department of Pure Mathematics and Mathematical Statistics,
University of Cambridge,
16 Mill Lane,
Cambridge CB2 1SB,
England.

# USING APPLIED MATHEMATICS IN INDUSTRIAL PROBLEMS

Stephen B. G. O'Brien

## 1. Introduction

Traditional mathematics degree courses in Ireland have often placed the emphasis on the pure mathematics and even the applied courses offered have paid little attention to the possibilities for applying mathematics to real world problems. In the U.S., in Britain and in Europe, applied mathematicians have retained far stronger links with industry and the courses taught in the universities in these countries tend to be more closely attuned to the needs of the modern applied mathematician. This also means that mathematicians in these countries tend to be quite successful in their attempts at solving industrial problems and this, in turn, has led to mathematics obtaining more research funding from industrial sources than is the case in Ireland. I believe that we can redress this situation by a change in our approach to teaching applied mathematics in this country.

The applied mathematician in industry must, to some extent, be a jack-of-all-trades. The emphasis on teaching applied mathematics in Ireland has been placed on solving already formulated problems (for example, the student is presented with a differential equation and suitable boundary conditions), while the final "solution" will usually entail some complicated mathematical expression. The problem may be garnished with vague references to incompressible liquids/ cantilever beams, etc, but the emphasis is on solving the mathematical problem. The applied mathematician in industry realizes that solution of the formulated problem is only a small part of his overall task. In the industrial setting, he will typically be shown some experimental process and asked why it





does not work! This is a far cry from starting with a neat, properly formulated problem.

In my experience, *formulation* of the problem is possibly the most important step during the complex process of solving an industrial problem. A certain amount of basic physics is required, and at the very least the ability to communicate with physicists and engineers and ask them the right sorts of questions. Once the basic physics has been captured, for example by writing down a system of differential equations, physical insight is required to make reasonable simplifications without losing the essence of the problem. Then a full non-dimensionalization should be carried out leading (usually) to a reduction in the number of parameters in the problem (via the Buckingham Pi theorem) and the possibility for further simplification via asymptotic means by exploiting the occurrence of small parameters. At this point, the system can hopefully be analysed using asymptotic techniques or, if it is still too complicated, it can be solved numerically. Having attained solutions, the applied mathematician is certainly not yet finished. The (non-dimensional) solutions must now be *interpreted* to see what physical insight can be gained. Non-dimensional solutions often contain a wealth of physical information, but this has to be translated back into physics and suggestions made for improving the industrial process under consideration.

To summarize, the applied mathematician's approach to industrial problems can be divided into four steps:
(i) formulation (physics to mathematics);
(ii) simplification, non-dimensionalization of mathematical problem;
(iii) analytical/numerical solution of mathematical problem;
(iv) interpretation (mathematics to physics).

Traditional applied mathematics courses have concentrated on step (iii). Obviously a certain amount of physical intuition is required, and one extremely useful way of developing students' feel for physics is by including a *physical* fluid mechanics course at undergraduate level. This has the advantage that it is a subject rich in physical mechanisms (viscous effects, inertia terms, diffus-

ive and convective phenomena) and approximations (slow creeping flow, thin film flow, boundary layer flow). It should certainly not be treated in a purely theoretical manner. Applied mathematics students should also be encouraged to take a certain number of physics courses and should be exposed to the elements of modelling and non-dimensionalization as early as possible. Finally, the basic *ideas* behind asymptotic methods (regular and singular perturbation theory) should be introduced as early as possible at an introductory level.

The rest of this article will consider a physical problem which I worked on at the Philips Research Laboratories in Eindhoven (in Holland), and illustrates the diversity of difficulties facing the industrial applied mathematician. In particular, while the formulation of the actual physical problem is carried out in Section 4 (followed approximately by steps (ii) to (iv) above), it should be noted that in order to reach this point a number of sub-problems must first be solved. Though this is not pursued in any detail here, each of these sub-problems also requires its own formulation, non-dimensionalization, simplification, etc.

## 2. Problem description

A dirt particle adhering to an integrated circuit (IC) decreases the stability and reliability of this IC. For ICs of current interest, particles of the order of $0.1\mu$m are critical. Existing cleansing methods (for example, scrubbing, jet cleaning) generally exert removal forces proportional to the surface area or volume of a dirt particle. Their success has been rather limited for particles of radius less than $1\mu$m. The existence of such a lower limit may at first sight seem puzzling but can be explained by the fact that adhesion is caused primarily by:

$$F_A = \frac{AR}{6H^2}; \qquad H << R, \tag{1}$$

where $F_A$ is the London-van der Waals force, A is the Hamaker constant, $R$ is the particle radius and $H$ is the gap between the particle and the substrate. During cleansing a force must be exerted on the particle which opposes the adhesion force. The methods



mentioned above exert a force proportional to the second or third power of the particle radius R. All other factors remaining the same, if we reduce the particle size then the forces of adhesion as in equation (1) will eventually dominate the removal forces. A cleansing technique which just succeeds for $R = 1\mu$m will fail when $R = 0.1\mu$m. A new cleansing method is illustrated in fig.1. The substrate to be cleaned is immersed in water and as the dirt particle passes through the liquid/air phase boundary, the surface tension forces which originate at the contact line around the sphere can oppose the adhesion forces, given favourable wetting conditions (contact angles) i.e. conditions which result in a favourable removal force as denoted by $F_\gamma$. The crucial factor is that the capillary forces can be shown to be linear in $R$, so the method is essentially independent of particle size because the adhesion forces (1) which cause the particle to stick to the substrate are also linear in R. In the next section we summarize some of the experimental work done.

### 2.1. Experimental work

Precise details of the experiments are to be found in [2] and [4]. In summary, a number of silicon substrates were contaminated with $TiO_2$ (rutile), $\alpha$-$Fe_2O_3$ (haematite) and $SiO_2$ (amorphous silica). Where necessary, the contaminated substrates were silylated to change the contact angles favourably. The substrates were then passed slowly through an air/water interface. Before and after immersion, the particles were sized and counted using an electron microscope. In general $70\% - 97\%$ of the particles were removed, given favourable wetting conditions. The method was equally successful for particles as small as $0.1\mu$m provided the immersion velocity was in the range $1\mu$m/sec to several cm/sec. Almost no particles were removed for velocities of the order of 10 cm/sec or higher.

## 3. Developing a mathematical model

Experiment suggests the existence of a critical velocity above which the removing process no longer works. We begin by ruling out gravity forces and hydrostatic pressure as being of any importance, since they will both be $O(R^2)$ or smaller and can

easily be shown to be negligible for particles of radius less than about 10 microns. The surface tension forces can be shown to be:

$$F_\gamma = 2\pi R\gamma \sin\phi \sin(\theta - \phi), \tag{2}$$

$\gamma$ being the surface tension of a liquid-gas interface where we consider the situation depicted in fig.2 where the water level is rising. $R$ is the radius of a dirt particle (assumed spherical), $\theta$ is the contact angle between water and the dirt particle and $\phi = \phi(t)$ indicates the position of the water/air interface at any time t. In fig.2 the net surface tension force $F_\gamma$ clearly opposes the adhesion force $F_A$. In order to use (2) to write down the magnitude of the the surface tension forces *as a function of the time*, it is necessary to solve a sub-problem which concerns the shape of the liquid free surface and involves finding an expression for $\phi$ as a function of time t. This is done in Section 3.1.

In what follows we assume constant contact angle. A full analysis of the problem would involve a Stokes' flow problem with its concomitant difficulties. Instead, we proceed in an adhoc fashion. It is probable that the solution to the problem lies in the occurrence of viscous effects which tend to oppose relative motion between the dirt particle and the substrate. In Section 3.2, we show how an expression for these forces can be found. We consider the situation represented in fig.2 as the fluid level rises. If at a particular point the resultant surface tension force (which in this case is wholly opposing the adhesion force) is greater than the adhesion force, the particle will tend to move away from the wall. Then, as in lubrication theory, we expect the generation of large viscous forces which offer great resistance to the further separation of sphere and substrate. It is not sufficient for the surface tension forces merely to be greater than the adhesion forces: this no longer guarantees removal. The viscous forces, which will presumably vary monotonically with the velocity of the *particle*, also retard its removal. Thus we require the particle to attain sufficient separation before the surface tension forces die away. As the particle starts from rest, we have

$$\int_0^t (F_\gamma - F_A - F_\mu)\, dt = mv(t), \tag{4}$$



where $F_\mu$ represents the viscous forces, and it becomes clear that the time for which the surface tension forces operate, and *hence the velocity of fluid immersion*, is crucial to ensure that the separation of the particle from the substrate is large enough, i.e. that the area under the curve (representing the impulse delivered to the particle) in fig.3 is as large as possible. We now propose a one dimensional model for a particle moving away from a substrate. Before we can proceed with formulating an equation of motion for the particle, we must first obtain expressions for the different forces acting.

### 3.1. Surface tension forces as a function of the time

The basic form of the surface tension forces is given by eqn (2) with $\phi = \phi(t)$, the form of $\phi$ being as yet unknown. We consider the situation occurring in fig.2 as the fluid level rises. The velocity at which the undisturbed meniscus at $\infty$ rises is constant but will not be the same as the rate of rise of the circle of contact on the sphere. The former is given by: $dh_\infty/dt = V$, where $h_\infty$ is the height of the water meniscus far from the particle (at $\infty$), i.e. the globally observed height. We thus view the dynamic formation of the meniscus as a series of quasi-steady state problems and seek a relationship between the undisturbed fluid meniscus height and the position of the contact line on the sphere as indicated by the angle $\phi$ or $\phi_0$ (see fig.2).

We will not go into the details of solving this sub-problem here (see for example [3]), but the shape of the fluid meniscus is described by a non-linear ordinary differential equation and on non-dimensionalizing and exploiting the occurrence of a small parameter, $\epsilon$ ($\sim 10^{-3}$) which is the ratio between the particle radius $R$ and the capillary length $a = (\gamma/\rho g)^{1/2}$, where $\rho$ is liquid density and $g$ is gravitational acceleration, we obtain a solution using matched asymptotic expansions and deduce the following relationship between the physical height, $h$ (see fig.2), and the angle $\phi_0 = (\pi/2 - \phi)$, defining the position of the circle of contact as:

$$h = a\epsilon C(\ln\epsilon + \ln\left(\cos\phi_0 + \sqrt{\cos^2\phi_0 - C^2}\right) - \ln 4 + \gamma_c), \tag{5}$$

where $C = \cos(\theta + \phi_0)\cos\phi_0$, $\gamma_c$ being Euler's constant. As $h$ is known as a function of $t$, so too is $\phi$ or $\phi_0$.

### 3.2. The viscous forces

As a particle starts to move away from the substrate (see fig.2), we expect the occurrence of strong viscous forces opposing the separation. In order to approximate this flow, we note that this so-called 'squeeze flow' between the sphere and substrate may be modelled using the lubrication approximation, as the layer of fluid separating the two bodies is so thin. We again neglect the details of this sub-problem. For the case in question where $H << R$, it was shown in [3] that the lubrication forces opposing separation are given by:

$$F = \frac{6\pi\mu V R^2}{H}, \tag{6}$$

where $V$ is the relative velocity between sphere and substrate. Thus the force required to effect separation is inversely proportional to the separation of sphere and substrate and directly proportional to the velocity of removal. Beer drinkers will have experienced similar 'squeeze flow' phenomena while trying to lift their glasses from a smooth wet table top!

## 4. Formulation, simplification, solution and interpretation

### 4.1. Formulation

Referring to fig.2 we consider a force balance on a sphere moving away from the substrate, the sphere being considered a particle. Equilibrium occurs as result of a balancing of the inertial, surface tension, viscous and adhesion (van der Waals) forces on the particle. Then, considering eqns (1), (2) (incorporating (5)) and (6), we can formulate an initial value problem non-dimensionalized using the following scales:

$$x^* = \frac{x}{H}, \quad t^* = \frac{tV}{R}, \tag{7}$$

where $x(t)$ is the *separation* between sphere and substrate. We



thus formulate:

$$K\ddot{x}^* = G(t*) - \frac{\delta}{(x^*)^2} - \lambda\frac{1}{x^*}\frac{dx^*}{dt^*}, \tag{8}$$

where

$$G(t*) = \sin\left[\phi(Rt^*/V)\right]\sin\left[\theta - \phi(Rt^*/V)\right],$$

$$K = \frac{2\rho_p V^2 H}{3\gamma}, \quad \delta = \frac{A}{H^2\gamma 12\pi}, \quad \lambda = \frac{3\mu V}{\gamma} = 3Ca$$

and $\rho_p$ is the density of the particle and $Ca$ is the capillary number, while differentiation with respect to the time is denoted by a dot. As boundary conditions in dimensionless form we have:

$$x(0) = 1, \quad \dot{x}(0) = 0, \tag{9}$$

indicating that the particle starts form rest at a known distance from the substrate.

### 4.2. Simplification

Equations (8) and (9) can be considered a singular perturbation problem, as $K << 1$ (see also [3]). However it turns out that the outer solution obtained by solving eqn (8) ignoring the inertia terms and using only the first of the boundary conditions gives rise to a solution which also satisfies the second condition. Thus the problem does not display boundary layer behaviour at lowest order, and can be solved in closed form.

### 4.3. Solution

The solution of (8) satisfying the first boundary condition of (9) and neglecting the $O(K)$ terms (this gives a Bernoulli equation) is easily shown to be:

$$x^{*2}(t^*) = \exp\left(\frac{2}{\lambda}[I(t^*) - I(0)]\right) - \frac{2\delta}{\lambda}\exp\left(\frac{2}{\lambda}I(t^*)\right)\int_0^{t^*}\exp\left(-\frac{2}{\lambda}I(\theta)\right)\,d\theta, \tag{10}$$

where

$$I(t^*) = \int^t G^*(\tau)\,d\tau.$$

### 4.4. Interpretation

The important factor in (10) is the $2/\lambda$ which identifies the dimensionless group $\gamma/\mu V$ (as $\lambda = 1/(3Ca)$), the reciprocal of the capillary number $Ca = \mu V/\gamma$, which has arisen naturally in the course of the analysis. The process of non-dimensionalization has indicated that the particular values of the surface tension $\gamma$, viscosity $\mu$ and velocity of immersion $V$ are not in themselves important. Rather, the way in which they combine to form the capillary number $Ca$ is what determines the dynamics of the motion. To optimize the process it is necessary to keep $Ca$ as small as possible. (10) would traditionally have signified the end of the applied mathematician's work, but as pointed out in the introduction, the present phase is just as important.

In dimensional form the solution (10) becomes:

$$x^2(t) = H^2\exp\left[\frac{2\gamma}{3\mu R}\int_0^t F_\phi(\tau)d\tau\right] - \frac{2A}{36\pi\mu R}\exp\left[\frac{2\gamma}{3\mu R}\int^t F_\phi(\tau)d\tau\right]\int_0^t\exp\left[-\frac{2\gamma}{3\mu R}\int^\tau F_\phi(\upsilon)\,d\upsilon\right]\,d\tau$$

where $F_\phi(t) = \sin\phi(t)\sin(\theta - \phi(t))$. In this instance the relevance of fig.3 can easily be seen. Decreasing the velocity of immersion increases the area under the surface tension-time graph and maximizes the separation between particle and substrate. Application of the results attained here to the physical problem (fig.1), which



makes the rather crude assumption that the meniscus is axisymmetric, in order to attain an estimate for the length of time for which the fluid meniscus remains in contact with the sphere (fig. 1; $\theta = \alpha = 70°$, $R = 0.3\mu$m, $\mu = 10^{-3}$kgm$^{-1}$s$^{-1}$, $A = 1.5\ 10^{-19}$J, $\rho = 10^3$kgm$^{-3}$), predicts a cut-off velocity of the order of 25cm/s. Above this value the process should no longer work, although experiment suggests that it is somewhat lower in the 10-15cm/s range.

## 5. Discussion of the physical model

There are a number of factors in the physical process which produce uncertainties in the results, e.g. the Hamaker constant $A$ and initial separation $H$ (see (1)). The latter is taken here to be 1nm but Kim and Lawrence, [1], suggest that a more realistic value would be 0.6nm. A further possible source of error is estimating when a particle is actually free from the substrate. As the nature of the van der Waals forces is known to change for a separation of 10nm, we arbitrarily assumed a particle to be free when it reached this distance, which is ten times the initial separation. Nevertheless the approximate model derived here captures the essential features of the experimental process, i.e. the significance of the viscous forces and the velocity dependent nature of the mechanism. One dynamic factor missing from the model is the variation of contact angle with substrate velocity for the simple reason that no values of the dynamic contact angle for the case considered in this paper are known, be it theoretically or experimentally. Leenaars [2] assumes that the contact angle remains constant.

The basic analysis identifies the capillary number as the most significant dimensionless parameter and indicates that the critical velocity of immersion can be increased by decreasing the ratio of $\mu/\gamma$ for the cleaning fluid. A further rather obvious improvement is also indicated: the experimental process as shown in fig.1 has the disadvantage that part of the removal power of the surface tension forces is lost due to the effect of the inclination $\alpha$. The theoretical set-up in fig.2 removes this problem completely and results in a much higher theoretical critical velocity. In practice this arrangement would be difficult to obtain when dealing with

submicron particles. However, a more favourable arrangement than fig.1 could be obtained by submerging the substrates at an angle, thus striving for a set-up between the extremes of figs.1 and 2. This also has the advantage of speeding up the process and of removing the uncertainty about the contact angle from the analysis.

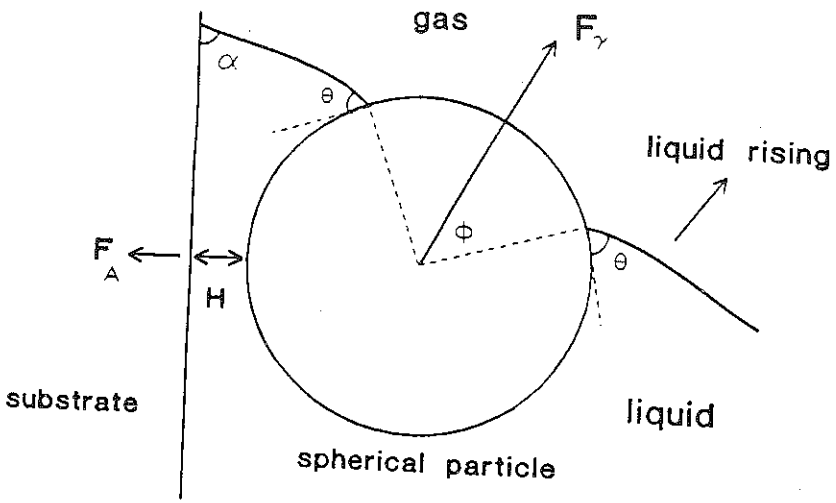



Fig. 1. Spherical particle adhering to substrate during passage of phase boundary.



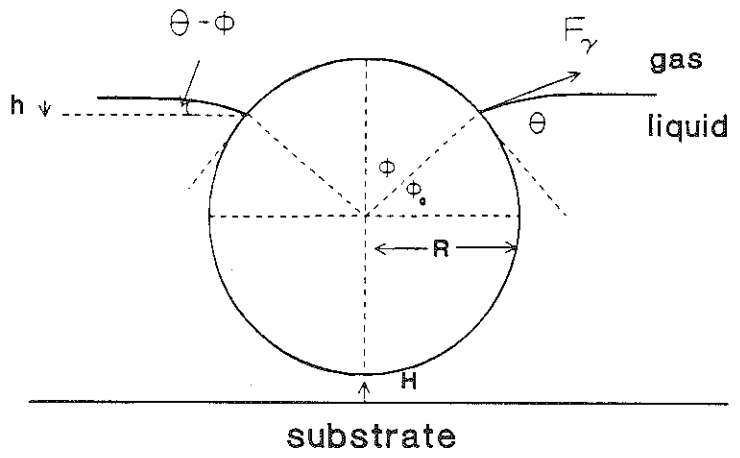



Fig. 2. Removal of particle from horizontal substrate.

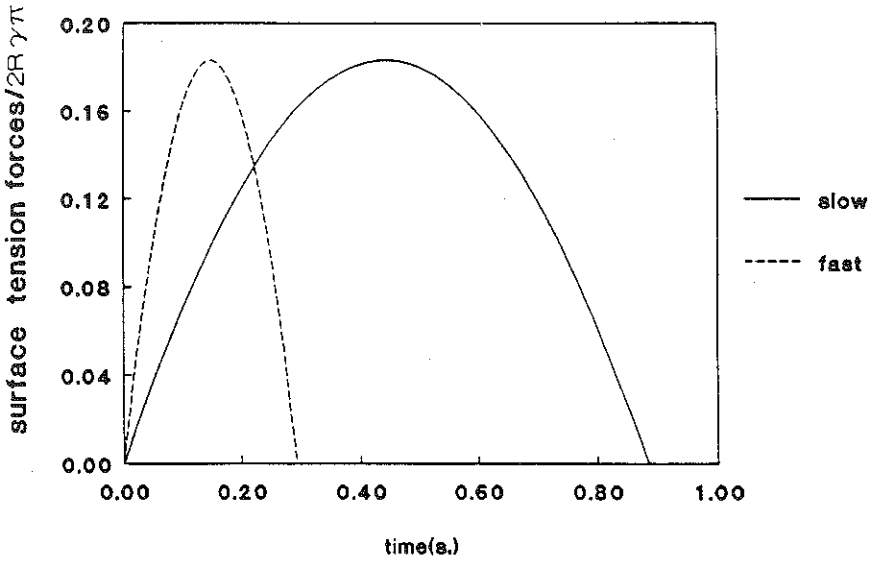



Fig. 3. Comparison of impulse delivered by $F_\gamma$ for slow and fast immersion.

**References**

[1] S. Kim and C. J. Lawrence, Chem. Eng. Sci. **43** (1988), 991-1004.

[2] A. F. M. Leenaars, *in* Particles on Surfaces: Detection, Adhesion and Removal, ed. K. L. Mittal. Plenum: New York, 1988.

[3] S. B. G. O'Brien and B. H. A. A. van den Brule , J. Colloid Interface Sci. **144** (1991), 210-222.

[4] S.B.G. O'Brien and A. F. M. Leenaars , Philips J. Res. **44** (1992), 183-209.

S. B. G. O'Brien,
Centre for Industrial and Applied Mathematics,
University of Limerick,
Limerick.

# THE 35TH INTERNATIONAL MATHEMATICAL OLYMPIAD

Fergus Gaines

The International Mathematical Olympiad (IMO) is the most prestigious mathematical competition in the world for pre-university students. It is held annually and the 1994 contest took place in Hong Kong in July. The number of countries and regions officially participating was 68. Each participating country sent a team of up to six members. The competition consisted of two four and a half-hour examinations, each exam made up of three problems. Each student competed as an individual and medals were awarded to the top performers.

IMO problems are celebrated for their extreme level of difficulty and some of them can even defeat professional mathematicians. It is no surprise, therefore, to find that a young student stands little chance of success in the competition, without a considerable amount of training. Some countries have a whole series of mathematics competitions–one for each year of the school programme–and in this way they can identify and encourage talented students from an early age. The first task in the process of choosing a team to represent Ireland in Hong Kong was to identify suitable candidates for training. Because a certain basis of mathematical knowledge is required in order to benefit from the training programme, generally only students who have completed the Junior Certificate are eligible. In November 1993 most secondary schools were invited to send up to three of their most mathematically talented pupils to attend training sessions in one of UCC, UCD, UCG and the University of Limerick. From information supplied by the Department of Education the top two hundred performers in the 1993 Junior Certificate mathematics examination were also personally invited to attend. The training sessions

took place once a week in each of the centres. As well as giving basic instruction in the areas of geometry, inequalities, combinatorics, number theory and algebra, the training sessions also concentrated on problem-solving techniques in these areas. Because the numbers attending in UCD were quite large (about 200, initially) an elimination test was held after four training sessions and the best 30 students were kept on in the training programme there. It was not necessary to have an elimination test in the other centres.

The Seventh Irish Mathematical Olympiad was held on Saturday, 7 May 1994 and consisted of two three-hour examinations. The top six performers in this contest were:

1. John Sullivan, Coláiste an Spioraid Naoimh, Cork.
2. Eoghan Flanagan, St Nessan's Community School, Limerick.
3. Mark Flanagan, St Benildus College, Dublin 14.
4. Deirdre O'Brien, Mount Mercy College, Cork.
5. Richard Murphy, St Munchen's College, Limerick.
6. Mark Dukes, Newpark Comprehensive School, Blackrock, Co Dublin.

and these were invited to form the Irish team for the IMO in Hong Kong. They all accepted the invitation.

A final three-day training camp for the team was held in the University of Limerick from 29 June to 1 July. The training camp is very important for the students–as well as concentrating on problem-solving strategies it gives the students the opportunity to get to know each other well, it helps to generate a team spirit and it helps to increase the motivation of the students to do their best in the competition.

I was the leader of the team and the deputy leader was Donal Hurley of University College Cork. I flew to Hong Kong on Thursday, 7 July and was taken to the Panda Hotel in Kowloon, where the leaders of all the teams were accommodated. The team leaders formed the jury for the final selection of the six problems that were to form the competition. Each participating country had already been invited by the organizers to submit up to six original problems for consideration by the jury. A local committee in Hong Kong had formed a short list of 24 problems from all those submitted. After three days of very long, and sometimes acrimonious, meetings, the six problems for the competition were selected. Although problems submitted are supposed to be original, a number of problems on the list of 24 were rejected because they, or problems very like them, had already appeared in other competitions. There was some dissatisfaction expressed at the quality of the shortlisted problems. Versions of the final six problems were prepared in the four official languages, English, French, Russian and Spanish, and the official text agreed. Finally, translations were made into all the languages required by the students. All the jury meetings took place in the Chinese University of Hong Kong.



The team, accompanied by Donal Hurley, arrived in Hong Kong on Monday, 11 July and were taken to their accommodation in summer camp-style residences in a rural part of Kowloon. There was no contact of any kind between them and the team leader until after the competition. The accommodation was adequate, if a little spartan, and the students found the lack of air-conditioning a bit trying in the humid, summer heat of Hong Kong. They found it difficult to adjust to the Chinese food, but they were able to buy food more to their taste in the local shops and the local McDonalds! The opening ceremony took place on Tuesday, 12 July, performed by the governor of Hong Kong, Mr Christopher Patten. The first exam was held at the Chinese University on Wednesday, 13 July when the first three problems were examined in a four and a half-hour exam. I received the exams of the Irish students that evening and began the work of reading their answers. The second exam was held on the morning of Thursday, 14 July and, that afternoon, all the deputy team leaders moved from their accommodation with the students to join the team leaders in the Panda Hotel.

Donal Hurley and I spent many hours reading the students' work and working out for each student the marks he would be expected to get, based on the marking scheme prepared by the coordinators. We also spent a considerable amount of time pursuing some of the students' lines of thought to see if they would

lead to solutions. This needs to be done if a case is to be made for extra marks. To give an idea of the amount of work involved, it happened twice that more than three hours were spent on the work of one student on one problem to get a complete understanding of that piece of work. On Friday and Saturday, 15 and 16 July, we went seven times to the coordinators to agree the marks to be awarded for each question. The seventh trip was needed because John Sullivan had a particularly complicated (and essentially correct) solution to problem no. 6 and a large amount of time was needed to understand his work before it could be presented for coordination. The coordinators for each problem consisted of a group of three local mathematicians. Donal Hurley and I explained, in great detail, what each of the students had done on that problem and agreed, in some cases after much argument, the mark to be awarded to each student.

This year's IMO exam was considered by most observers to be somewhat easier than usual and this was reflected in the high scores of many of the students. The rules of the IMO state that medals can be awarded to at most half of the contestants. It is also stipulated that at most 1/6th of those eligible can receive gold medals and that the corresponding fractions for silver and bronze are 1/3 and 1/2, respectively. A perfect answer to a question gains 7 points and, thus, the maximum number of points that a student can score is 42. Partial credit for a question is awarded, but a student has to do some significant work before any marks at all are given. The marks gained by the Irish students were:

| | |
|---|---|
| Mark Dukes | 11 |
| Eoghan Flanagan | 16 |
| Mark Flanagan | 10 |
| Richard Murphy | 15 |
| Deirdre O'Brien | 2 |
| John Sullivan | 14 |

Thus the team score was 68, which meant that Ireland got 49th place out of 69 competing countries. In order to win a bronze medal a contestant had to score at least 19 points, so the Irish won no medals this year. However, since any student who does



not get a medal and who scores 7 points on at least one question gets an "honourable mention", three of the students, Eoghan Flanagan, Mark Flanagan and Richard Murphy received this honour. Question 6 was by far the most difficult on the exam and this fact was underlined by the large number of students who scored zero on it.

There were some surprises in the results of some other countries. In recent years China has become the strongest team in the IMO, so it was quite a surprise when they were beaten into second place by the United States. The U.S. team created a record because every one of their students scored full points and this has never happened previously in the IMO. The Chinese team caused quite a stir when three of their students scored zero on question 6! This makes John Sullivan's mark of 4 on that question all the more meritorious.

The team scores, out of a maximum of 252, for the leading ten countries were:

| | |
|---|---|
| United States | 252 |
| China | 229 |
| Russia | 224 |
| Bulgaria | 223 |
| Hungary | 221 |
| Vietnam | 207 |
| United Kingdom | 206 |
| Iran | 203 |
| Romania | 198 |
| Japan | 180 |

It would not be possible for Ireland to participate in the IMO without considerable support from many people and organizations. The organizers are extremely grateful to the sponsors for financial and other assistance. The sponsors of the Irish participation in the 1994 IMO were

An Roinn Oideachais
Forbairt
University of Limerick
Arts Faculty, University College Dublin

Royal Irish Academy
Irish National Mathematics Contest.

I give here the six problems of the 35th IMO. Solutions are given below on pages 74 to 76.

**1.** Let $m$ and $n$ be positive integers. Let $a_1, a_2 \ldots, a_m$ be distinct elements of $\{1, 2, \ldots, n\}$ such that whenever $a_i + a_j \leq n$ for some $i$ and $j$, $1 \leq i \leq j \leq m$, there exists $k$, where $1 \leq k \leq m$, with $a_i + a_j = a_k$. Prove that

$$\frac{a_1 + a_2 + \cdots + a_m}{m} \geq \frac{n+1}{2}.$$

**2.** $ABC$ is an isosceles triangle with $AB = AC$. Suppose that
(*i*) $M$ is the midpoint of $BC$ and $O$ is the point on the line $AM$ such that $OB$ is perpendicular to $AB$;
(*ii*) $Q$ is an arbitrary point on the segment $BC$ different from $B$ and $C$;
(*iii*) $E$ lies on the line $AB$ and $F$ lies on the line $AC$ such that $E$, $Q$ and $F$ are distinct and collinear.
Prove that $OQ$ is perpendicular to $EF$ if and only if $QE = QF$.

**3.** For any positive integer $k$, let $f(k)$ be the number of elements in the set $\{k+1, k+2, \ldots, 2k\}$ whose base 2 representation has exactly three 1's.
(*a*) Prove that, for each positive integer $m$, there exists at least one positive integer $k$ such that $f(k) = m$.
(*b*) Determine all positive integers $m$ for which there exists exactly one $k$ with $f(k) = m$.

**4.** Determine all ordered pairs $(m, n)$ of positive integers such that

$$\frac{n^3 + 1}{mn - 1}$$

is an integer.

**5.** Let $S$ be the set of real numbers strictly greater than $-1$. Find all functions $f : S \to S$ satisfying the two conditions:
(*i*) $f(x + f(y) + xf(y)) = y + f(x) + yf(x)$ for all $x$ and $y$ in $S$;



(*ii*) $\frac{f(x)}{x}$ is strictly increasing on each of the intervals $-1 < x < 0$ and $x > 0$.

**6.** Show that there exists a set $A$ of positive integers with the following property: for any infinite set $S$ of primes there exist positive integers $m \in A$ and $n \notin A$ each of which is a product of $k$ distinct elements of $S$ for some $k \geq 2$.

Fergus Gaines,
Department of Mathematics,
University College,
Belfield,
Dublin 4.

# DAVID HILBERT AND THE THEORY OF ALGEBRAIC INVARIANTS

David W. Lewis

## 1. Introduction

The theory of algebraic invariants was at the forefront of mathematics in the latter half of the 19-th century. It attracted the interest of many top-class mathematicians. For example Cayley and Sylvester in England were known as the "Invariant Twins", and when Salmon in Dublin made useful contributions to the subject the trio were christened by Hermite as the "Invariant Trinity". Another Irish link with invariant theory is provided by George Boole who spent much of his working life in Cork. In 1841 Boole wrote a paper [1] which is often regarded as the beginnings of invariant theory, and in 1845 he wrote another paper on the subject but seemed to do nothing further on invariants. (Admittedly Boole was still in England when he wrote these papers. He moved to Ireland to become Professor of Mathematics at Queen's College, Cork in 1849). See [10] for an excellent account of the life and work of Boole. The Italian mathematician Faà di Bruno wrote a book on invariant theory which was highly regarded by Hilbert. In Germany the first mathematician to draw attention to the theory of invariants was Aronhold. He was followed by Clebsch and Gordan who worked extensively on the subject and developed symbolic calculation in invariant theory. Indeed Gordan was known as the "King of Invariants" and apparently would talk interminably about invariant theory to anyone who was willing to listen. (The names of Clebsch and Gordan will be familiar to students of quantum mechanics via the Clebsch-Gordan series and Clebsch-Gordan coefficients. The Clebsch-Gordan series played an important role in their theory of invariants of binary forms. See Weyl





[19].) Their work involved massive calculations. According to [3], there are papers of Gordan where twenty pages of formulae are not interrupted by a single text word, and it is alleged that Gordan often wrote only the formulae in his papers, the text being added later by friends.

Although invariant theory was a piece of pure mathematics, attempts were made to make use of invariant theory in other disciplines. For example, Sylvester in 1878, and later Gordan and Alexejeff, tried to apply invariant theory to chemistry, in connection with chemical valency. A brief account of this so-called "chemico-algebraic theory" appears in [5, pp.366-368]. In the period from 1885 to 1893 David Hilbert demolished the old-style invariant theory by solving, in a novel and unexpected way, the central finiteness problem of invariant theory. After Hilbert's work, many people thought that invariant theory was a dead subject. However it has refused to lie down and has resurrected itself on quite a few occasions in the 20th century. Indeed, to quote from the 1984 survey article by Kung and Rota [9], "the theory of invariants, pronounced dead at the turn of the century, is once again at the forefront of mathematics". Today, invariant theory is alive and well and the subjects of commutative algebra, algebraic geometry, representation theory , and combinatorics each owe an important debt to invariant theory.

## 2. David Hilbert

David Hilbert was born in 1862 in Königsberg, then part of East Prussia but renamed Kaliningrad after the Second World War and now a part of Russia. Königsberg has a long intellectual tradition, especially in mathematics and philosophy. (Mathematicians will all know of the famous "Königsberg bridge problem" solved by Euler in the 18th century. The philosopher Kant was one of the city's most famous sons. Clebsch was also born in Königsberg and attended the university there.) Hilbert went to university in Königsberg where he became a close friend of fellow student Hermann Minkowski, this friendship lasting until Minkowski's early death in 1909. After spending several very productive years lecturing at Königsberg, during which time he did all of his important

work in invariant theory, he was offered and accepted in 1895 a position at Göttingen. The mathematics department there, with Felix Klein as chairman, was possibly the most prestigious in Germany at that time. Hilbert spent the rest of his life in Göttingen and died there in 1943.

Hilbert has been described as "the last of the great universalists". Over his long career he made vital contributions to large and diverse areas of mathematics. One might say that he led mathematics out of the 19th century and into the 20th century. His famous list of unsolved problems at the International Congress of Mathematicians in Paris in 1900 pointed the way forward and profoundly influenced the direction of mathematical research in this century. His method of work led to great advances both in technical results and in the way in which mathematicians think about mathematics. Hilbert tended to concentrate almost exclusively on one particular area of mathematical research for a period of years and then move on to a different branch. His research work, according to [17], was roughly as follows:

1885-1893 - Invariant Theory
1893-1898 - Number Theory
1898-1902 - Foundations of Geometry and of Mathematics in general
1902-1912 - Integral Equations
1912-1922 - Mathematical Physics

The Japanese number theorist Takagi visited Hilbert at Göttingen in 1902 but Hilbert is reported to only have been being interested in talking about integral equations (the work of Takagi and of Hilbert forms the beginnings of class field theory). See [13, p.86]. There were exceptions to the list above. For example, in 1909 Hilbert successfully solved Waring's problem, a problem outstanding since 1770 about expressing a natural number as a sum of $n$-th powers. He produced this solution just at the time his friend Minkowski was dying of appendicitis and unfortunately Minkowski was unable to attend the seminar by Hilbert in which he described his solution to the problem. Also in 1899 Hilbert managed to resuscitate Dirichlet's Principle concerning the solution of boundary value problems, this being totally unrelated to



the main research work he was pursuing at this period. It was in the period on foundations of geometry that he made his famous pronouncement about the axiomatic method-"One must be able to say at all times-instead of points, straight lines and planes - tables, chairs, and beer mugs".

From 1912 on he worked on the idea of axiomatizing physics– this having been proposed as his 6th problem at the 1900 Paris congress. "Physics is much too hard for physicists", he said. He did not have much success however, the axiomatic method not seeming to be suitable for physics.

For a full account of the life of Hilbert the reader should refer to the book of Constance Reid, [13]. See also the book of Fang, [3].

### 3. Hilbert's 1897 lectures

In 1897 David Hilbert gave an introductory course of lectures on the theory of algebraic invariants at the University of Göttingen. These lectures, or rather a modern English translation by Reinhard C. Laubenbacher of the lecture notes, handwritten by Hilbert's student Sophus Marxsen, have recently been published by the Cambridge University Press, [6]. They provide a fascinating view of invariant theory and a glimpse of what it must have been like to have studied with Hilbert at Göttingen at that time. The course consisted of 51 lectures starting on 26 April 1897 and ending on 6 August 1897, 3 lectures per week for 17 weeks. Sophus Marxsen ended up with 527 pages of handwritten notes. As a lecturer Hilbert was inspiring but he sometimes ran into difficulty in a lecture because he had not prepared all the technical details. (This was in sharp contrast to his Göttingen colleague Felix Klein who is reported to always have prepared everything in meticulous detail. Klein was older and more famous than Hilbert at that time, although nowadays he is perhaps best remembered for his bottle, the "Klein bottle" being the famous one-sided surface loved by all topologists.) The year 1897 was an appropriate time for Hilbert to give an expository course on invariant theory because in two papers [7], [8], in 1890 and 1893 he had solved the major problems in invariant theory. Thus he was able in these lec-

tures to incorporate the work of his predecessors and his own new and revolutionary approach to the subject. For local interest we should remark that in his first lecture Hilbert refers to the book of Salmon [14], Modern Higher Algebra (fourth edition), Dublin 1885, as one of the best introductions to the subject of invariant theory.

An *m-ary n-form* $\phi$ is a homogeneous polynomial of degree $n$ in $m$ variables. (For $n = 2$, this is a quadratic form.) If we write $x_1, x_2, \ldots, x_m$ for the variables, then we may write

$$\phi = \sum a_{i_1 i_2 \ldots i_m} x_1^{i_1} x_2^{i_2} \ldots x_m^{i_m}$$

for some suitable constants $a_{i_1 i_2 \ldots i_m}$. In Hilbert's lectures these constants are allowed to be complex numbers.

For $m = 2$, the form is called a *binary* form, for $m = 3$ a *ternary* form etc. The *degree* of the term $a_{i_1 i_2 \ldots i_m} x_1^{i_1} x_2^{i_2} \ldots x_m^{i_m}$ is $i_1 + i_2 + \cdots + i_m$ (All terms of the form will have the same degree because the form is a homogeneous polynomial.)

Suppose we make a linear change of variables from $x_1, x_2, \ldots, x_m$ to $x_1', x_2', \ldots, x_m'$, i.e we write $x = Px'$, where $x = (x_i)$, $x' = (x_i')$ are column vectors and $P = (p_{ij})$ is an $m \times m$ matrix. Then the form may be written in terms of the new variables $x_i'$ with new coefficients $a'_{i_1 i_2 \ldots i_m}$. The determinant of the matrix $P$ is denoted $\delta$ and is called the *transformation determinant.*

Hilbert, in his lectures, limits himself to binary forms but says that generalizing to $m$-ary forms poses no difficulties in most cases. He writes a general binary form $\phi$ in the manner

$$\phi(x_1, x_2) = \sum_{i=0}^{n} \binom{n}{i} a_i x_1^i x_2^{n-i}.$$

(He always uses the word "coefficients" to mean the $a_i$ when the form is written in this way, i.e. not multiplied by the binomial coefficients!)

An *invariant* of the form $\phi$ above is a polynomial function $I(a_0, a_1, \ldots, a_n)$ of the coefficients of $\phi$ which changes only by a



factor equal to a power of the transformation determinant $\delta$ when one makes a linear transformation of the variables, i.e.

$$I(a_0', a_1', \ldots, a_n') = \delta^p I(a_0, a_1, \ldots, a_n)$$

for some natural number $p$. Here $a_0'$, $a_1', \ldots, a_n'$ are the coefficients of $\phi$ after a linear change of variables given by the matrix $P$. It can be shown by elementary considerations that $I$ must necessarily be homogeneous of degree $g$ where $ng = 2p$. We illustrate by a couple of examples.

**Example 1**

$\phi = a_0 x_1^2 + 2a_1 x_1 x_2 + a_2 x_2^2$ is a binary form of degree 2. $I_1 = a_0 a_2 - a_1^2$ is an invariant of $\phi$. (Those familiar with quadratic forms will recognize $\phi$ as a quadratic form of dimension 2 and $I$ as the discriminant of this form.)

**Example 2**

$\phi = a_0 x_1^4 + 4a_1 x_1^3 x_2 + 6a_2 x_1^2 x_2^2 + 4a_3 x_1 x_2^3 + a_4 x_2^4$ is a binary form of degree 4.

$I_2 = a_0 a_4 - 4a_1 a_3 + 3a_2^2$ is an invariant of $\phi$.

$I_3 = a_0 a_2 a_4 - a_0 a_3^2 - a_1^2 a_4 + 2a_1 a_2 a_3 - a_2^3$ is also an invariant of $\phi$.

Observe that $I_1$ in example 1 and $I_2$ in example 2 are each homogeneous of degree 2 and $I_3$ in example 2 is homogeneous of degree 3.

Hilbert proceeds in the first half of these lectures to characterize those polynomials which are invariants by utilizing the operator $D$ defined as follows:

$$D = a_0 \frac{\partial}{\partial a_1} + 2a_1 \frac{\partial}{\partial a_2} + 3a_2 \frac{\partial}{\partial a_3} + \cdots + n a_{n-1} \frac{\partial}{\partial a_n}.$$

An invariant $I$ is shown necessarily to be a homogeneous polynomial and it must satisfy $DI = 0$. Also it is shown that an invariant $I$ must be an isobaric function of $a_0, a_1, \ldots, a_n$. (A polynomial in $a_0, a_1, \ldots, a_n$ is said to be *isobaric* if each term has the same weight, where the weight of a term $a_0^{\nu_0} a_1^{\nu_1} a_2^{\nu_2} \ldots a_n^{\nu_n}$ is

$$\nu_1 + 2\nu_2 + 3\nu_3 + \cdots + n\nu_n.)$$

Hilbert shows that each isobaric homogeneous polynomial in $a_0$, $a_1, \ldots, a_n$ of degree $g$ and weight $p$, where $ng = 2p$, is an invariant whenever $DI = 0$. He also calculates the number of invariants of given degree $g$ for a form $\phi$. It turns out that $I_1$ in example 1 above is the only invariant of a binary quadratic form and that $I_2$ and $I_3$ of example 2 are the only invariants of a binary form of degree 4. He also discusses the notion of a covariant, of which invariants are a special case, but we omit discussion of these in this short article.

In lecture 24, Hilbert introduces the idea of simultaneous invariants. Here one begins with an arbitrary set of base forms, each in the same number of variables but not necessarily of the same degree, rather than a single form. One defines an invariant of the set to be a polynomial in the set of coefficients of all the base forms which changes only by a power of the transformation determinant when the same linear transformation is applied simultaneously to all of the base forms. For example, suppose we have binary forms

$$\phi_1(x_1, x_2) = \sum_{i=0}^{n} \binom{n}{i} a_i x_1^i x_2^{n-i}$$

$$\phi_2(x_1, x_2) = \sum_{i=0}^{m} \binom{m}{i} b_i x_1^i x_2^{m-i}.$$

Then a simultaneous invariant for the pair $\phi_1$, $\phi_2$ under a linear transformation changing the $a_i$, $b_i$ to $a_i'$, $b_i'$ is a polynomial $I$ in $n + m + 2$ variables such that

$$I(a_0', \ldots, a_n', b_0', \ldots, b_m') = \delta^p I(a_0, \ldots, a_n, b_0, \ldots, b_m)$$

for some natural number $p$, where $\delta$ is the transformation determinant. (Note that any invariant of $\phi_1$ alone yields a simultaneous invariant by viewing it as of degree zero in the $b_i$, and similarly for an invariant of $\phi_2$ alone.)



**Example 3**
Consider the two binary cubic forms

$$\phi_1(x_1, x_2) = a_0 x_1^3 + 3a_1 x_1^2 x_2 + 3a_2 x_1 x_2^2 + a_3 x_2^3$$
$$\phi_2(x_1, x_2) = b_0 x_1^3 + 3b_1 x_1^2 x_2 + 3b_2 x_1 x_2^2 + b_3 x_2^3.$$

One may check that $I = a_0 b_3 - 3a_1 b_2 + 3a_2 b_1 - a_3 b_0$ is a simultaneous invariant of these two cubic forms.

Starting with an arbitrary system of base forms, the simultaneous invariants of the system can in general be an infinite set. It was Cayley who first conjectured that any system of base forms has an invariant set which is finitely generated, i.e. there is a finite subset $I_1, I_2, \ldots, I_k$ of the invariant set such that each element of the invariant set is a polynomial in $I_1, I_2, \ldots, I_k$. However, Cayley soon began to doubt the validity of his conjecture and, in an 1856 memoir, he incorrectly claimed that the fundamental system of invariants is infinite for forms of degree more than six. His mistake arose from wrongly taking certain syzygies to be independent. (See below for more about syzygies.) Gordan, via cumbersome calculations using the symbolic method, succeeded in proving the finiteness theorem for an arbitrary system of binary base forms. This achievement in 1868 was what gained Gordan his title of "King of Invariants". However attempts by Gordan himself and others to prove finiteness for base forms of higher degree were unsuccessful. The finiteness problem, i.e. the proof of Cayley's conjecture for an arbitrary system of base forms of any degree, had become the main problem of invariant theory by the time Hilbert came on the scene. (The earlier stages of invariant theory had been concerned with finding the laws governing the structure of invariants and then with the enumeration and production of invariants systematically.) Hilbert solved the finiteness problem by realizing that one only needs to prove the *existence* of a finite basis (i.e. generating set) for the invariants. It was not necessary to construct a basis explicitly, which is what Gordan and others had tried to do. Hilbert's solution when it appeared in 1890, [7], caused consternation amongst mathematicians. His "existence theorem" was not accepted by some of them as being a solution

at all. Gordan commented about the proof, "Das ist nicht Mathematik. Das ist Theologie". Hilbert did indeed continue to work on invariant theory and in [8] he gave an essentially constructive and algorithmic method for obtaining a finite basis.

In the second part of his Göttingen lectures, (lecture 34 onwards), he begins by proving the finiteness theorem for an arbitrary system of *binary* forms. His proof uses a technique called representation by root differences which involves the elementary symmetric functions. It does not generalize to systems of forms of degree greater than two. A key lemma used in this proof asserts that a system of linear equations with coefficients in the natural numbers has a finite number of non-negative solutions which generate all the other non-negative solutions. This lemma is foundational nowadays in the theory of integer programming. See [15]. Hilbert proceeds (in lectures 34-36) to prove his general finiteness theorem as in his 1890 paper, using the key result known nowadays as the Hilbert Basis Theorem for polynomial ideals together with Cayley's $\Omega$-process. The $\Omega$-process is a differentiation process which behaves like a kind averaging and when applied repeatedly to a polynomial it yields an invariant. His Basis Theorem yields a finite set $I_1, I_2, \ldots, I_k$ such that any invariant $I$ is expressible in the form

$$I = F_1 I_1 + F_2 I_2 + \cdots + F_k I_k$$

for some forms $F_1, F_2, \ldots, F_k$. Applying $\Omega$ to each of the $F_i$ yields invariants $G_i$ such that we can write

$$I = G_1 I_1 + G_2 I_2 + \cdots + G_k I_k$$

and each $G_i$ clearly has degree less than the degree of $I$, since each $F_i$ has degree at least one. By expressing each $G_i$ in terms of the set $I_1, I_2, \ldots, I_k$ and repeating as necessary, we can eventually write $I$ as a polynomial in the set $I_1, I_2, \ldots, I_k$. The remainder of the lectures are based on Hilbert's 1893 paper, [8], where he gives his algorithmic method for constructing a finite basis. From a modern perspective there are two highly significant theorems



contained there, although their full importance and application was not apparent then. In lecture 39 he gives the theorem now known as the Hilbert Nullstellensatz, although he refers to [8] for a full proof. This theorem concerning the zero sets of families of polynomials is basic and fundamental for modern commutative algebra and algebraic geometry. Lecture 47 describes the result usually known now as Hilbert's Syzygy Theorem. The set $I_1, I_2, \ldots, I_k$ is not likely to be linearly independent, i.e. there will be a set of relations between them. This relation set also must have a finite basis $R_1, R_2, \ldots, R_h$ by the finiteness theorem. There may well be relations amongst the relations, i.e. expressions of the form

$$S_1 R_1 + S_2 R_2 + \cdots + S_h R_h = 0.$$

Such an expression is called a syzygy of the first order. These syzygies again form an ideal to which the finiteness theorem applies and a finite basis exists. Any relation for this basis is a syzygy of the second order. It may seem that this process can be repeated *ad infinitum*, but Hilbert's Syzygy Theorem says that the chain of syzygies breaks off after finitely many steps. In the last few lectures Hilbert outlines some applications of invariant theory to geometry and discusses possible generalizations of invariant theory.

## 4. The view from the end of the 20th century

In modern terms we may describe invariant theory as being concerned with the linear action of a group $G$ on a $K$-vector space $V$ for some field $K$. Writing $K[V]$ for the ring of all polynomial functions on $V$, the basic problem is to describe the subring $K[V]^G$ of all polynomials invariant under the action of the group $G$. In particular, we may ask whether $K[V]^G$ is finitely generated as a $K$-algebra and, if so, find an algorithm for determining a set of generators. In the classical case described above we have $K = \mathbf{C}$, the complex numbers, $G = \mathrm{GL}_m(\mathbf{C})$, the group of all invertible $m \times m$ matrices with complex entries, $V$ an $m$-dimensional vector space over $K$, and $K[V]$ the ring of all homogeneous polynomials in $m$ variables. Hilbert proved that $K[V]^G$ is finitely generated as a $K$-algebra in the classical case. Hilbert's 14th problem at the

1900 congress asked whether this finiteness theorem remains true if $G$ is an arbitrary subgroup of $\mathrm{GL}_m(\mathbf{C})$. It remained an open problem until 1959 when Nagata [12] answered it in the negative by producing an example of a group $G$ with $K[V]^G$ not finitely generated.

Three of Hilbert's results in the above lectures have turned out to have tremendous significance and importance and we will describe now how they fit into 20th century algebra. The Hilbert Basis Theorem is now usually stated in the form that the polynomial ring $K[x_1, x_2, \ldots, x_n]$ is a noetherian ring. A ring is said to be *noetherian* if every ascending chain of ideals terminates. It is not hard to prove that this ascending chain condition for the polynomial ring is equivalent to the ideals being finitely generated. The name noetherian is after Emmy Noether who, in the 1920's and 1930's, was the main influence in the development of modern abstract algebra. It is curious that Emmy Noether began her career as a student of Paul Gordan at Erlangen, writing a thesis in 1907 on invariant theory. Gordan was still doing very computational invariant theory. Noether later referred to invariant theory as a "jungle of formulae" (formelngestrüpp) and one may speculate that it was her distaste for this kind of mathematics which led her to develop the conceptual approach of modern abstract algebra.

Hilbert's Nullstellensatz is now usually regarded as the foundation of algebraic geometry, yielding the correspondence between geometric objects (varieties) and algebraic objects (co-ordinate rings), although we have seen that this was not the purpose for which Hilbert originally developed it.

Hilbert's Syzygy Theorem is now stated as a result in homological algebra, saying that the polynomial ring $\mathbf{C}[x_1, x_2, \ldots, x_n]$ has finite global dimension (in fact dimension $n$), i.e. every module over this polynomial ring admits a finite free resolution of length at most $n$.

We finish with a few words about how invariant theory has developed in the 20th century, although this author claims no great expertise in modern invariant theory. Weyl [18] developed invariant theory for all the classical Lie groups and linked it with



representation theory. Mumford [11] developed a geometric invariant theory. The survey by Kung and Rota, [9], describes invariant theory from the viewpoint of modern combinatorial theory. The books by Springer [14] and by Dieudonné and Carrell [2] are further modern references on the subject. As a final remark, we note that the "death" of invariant theory has even attracted the interest of a sociologist! See [4].

**References**


[1] G. Boole, *Exposition of a general theory of linear transformations, I and II* Cambridge Math. Journal **3** (1842), 1-20 and 106-119.

[2] J. Dieudonné and J. B. Carrell, Invariant Theory, old and new. Academic Press: New York and London, 1971.

[3] J. Fang, Hilbert: Towards a Philosophy of Modern Mathematics II. Paideia Press: New York, 1970.

[4] C. S. Fisher, *The death of a mathematical theory. A study in the sociology of knowledge*, Arch. Hist. of Exact Sciences **3** (1966), 137-159.

[5] J. H. Grace and A. Young, The Algebra of Invariants. Cambridge University Press: Cambridge, 1903.

[6] David Hilbert, Theory of Algebraic Invariants, (English translation). Cambridge University Press: Cambridge, 1993.

[7] David Hilbert, *Über die Theorie der algebraischen Formen*, Math. Ann. **36** (1890), 473-531.

[8] David Hilbert, *Über die vollen Invariantensysteme*, Math. Ann. **42** (1893), 313-370.

[9] J. P. S. Kung and G-C. Rota, *The invariant theory of binary forms*, Bull. Amer. Math. Soc. **10** (1984), 27-85.

[10] Desmond MacHale, George Boole. Boole Press: Dublin, 1985.

[11] D. Mumford, Geometric Invariant Theory. Springer-Verlag: Berlin-New York, 1965.

[12] M. Nagata, *On the 14th problem of Hilbert*, Amer. J. Math. **81** (1959), 766-772.

[13] Constance Reid, Hilbert. Springer-Verlag: Berlin-Heidelberg-New York, 1970.

[14] G. Salmon, Lessons Introductory to the Modern Higher Algebra (fourth edition). Hodges, Figgis: Dublin, 1885.

[15] A. Schrijver, Theory of Linear and Integer Programming. Wiley-Interscience: Chichester, 1986.

[16] T. A. Springer, Invariant Theory. Lecture Notes in Mathematics 585. Springer-Verlag: Berlin-Heidelberg-New York, 1977.

[17] Hermann Weyl, *David Hilbert and his mathematical work*, Bull. Amer. Math. Soc. **50** (1944), 612-654.

[18] Hermann Weyl, The Classical Groups, Their Invariants and Representations. Princeton University Press: New Jersey, 1939.

[19] Hermann Weyl, The Theory of Groups and Quantum Mechanics, (English translation). Methuen: London, 1931.

**Acknowledgement** I am indebted to Rod Gow for reading the first version of this article and supplying me with some additional historical information and references.

D. W. Lewis,
Department of Mathematics,
University College,
Belfield,
Dublin 4.

## SYLOW'S PROOF OF SYLOW'S THEOREM

Rod Gow

### 1. Introduction

While looking through some early volumes of *Mathematische Annalen*, we came across a paper with the following title:

Théorèmes sur les groupes de substitutions.

Par M. L. Sylow à Frederikshald en Norvege.

This was, of course, the paper containing Ludwig Sylow's fundamental contribution to group theory, [9]. We thought it might be interesting to see how Sylow actually proved his theorem and then to comment briefly on some later proofs and earlier work. It is likely that there have been prior discussions of the history of Sylow's theorem in the literature and we apologize for failing to acknowledge any relevant publications. (Our excuse is that the UCD library is badly stocked with periodicals on the history of science.)

Sylow's starting point is as follows: *On sait que si l'ordre d'un groupe de substitutions est divisible par le nombre premier $n$, le groupe contient toujours une substitution d'ordre $n$.* (The notation of Sylow is a little wayward to modern tastes. His prime is denoted by $n$, rather than the traditional $p$. Later in the paper, the expression $np+1$ appears as the number of Sylow subgroups, but $p$ denotes merely some non-negative integer. In virtually all later literature relating to the proof of Sylow's theorem and earlier literature on Cauchy's theorem that we have seen, the prime is represented by $p$. We shall follow standard practice and denote our prime by $p$ in this exposition, except when enunciating Sylow's

theorems in his own words.) We recognize the statement above as Cauchy's theorem, which we would normally state as: *if a prime $p$ divides the order of a finite group, then the group contains an element of order $p$.* The stipulation that we should have a group of permutations is irrelevant, although in Cauchy's day, abstract finite groups would not have been under consideration. Indeed in the fourth edition of Serret's book, there is some discussion of permutation groups and of groups of linear fractional transformations, but no discussion of abstract groups or of Cauchy's theorem. There is, however, a discussion of a construction, due to Cauchy, of a Sylow $p$-subgroup of the symmetric group $S_n$ in [8, p.302]. Cauchy's theorem underlies Sylow's proof. It is not proved. Later proofs sought to remove this reliance on Cauchy's theorem, whose original demonstration was quite complicated, although it contained germs of ideas vital to modern group theory. In particular, Frobenius was able to give a proof of the existence of Sylow subgroups which avoided Cauchy's theorem and became the standard proof of Sylow's theorem until the advent of Wielandt's proof in 1959, [11].

Sylow proves the existence of a Sylow $p$-subgroup $P$ in a finite group $G$, at the same time showing that if $N$ is the normalizer of $P$ in $G$, then $|G : N| \equiv 1 \bmod p$. Afterwards, he shows that any other Sylow $p$-subgroup $Q$ is conjugate to $P$ in $G$. A basic idea used by Sylow, the spirit of which really occurs in all proofs, is that of letting $P$ and $Q$ permute the cosets of $N$ by multiplication. Simple congruences modulo $p$ force out the desired conclusion. Needless to say, Sylow does not talk in terms of permuting cosets, but this is the way to interpret his procedures nowadays.

## 2. Sylow's proof

We now consider the details of Sylow's proof. We try to follow the spirit, as we see it, of Sylow's ideas but use more modern concepts to try to explain what is happening. We will make a few comments about Sylow's precise method later. Let $G$ be a non-trivial finite group and let $p^\alpha$ be the $p$-part of $|G|$, where $p$ is a prime and $\alpha \geq 1$. Let $P$ be a $p$-subgroup of $G$ of maximal order and let $N$ be its normalizer in $G$. Sylow first proves that



$|N : P|$ is not divisible by $p$. From our point of view, this is clear. If $p$ divides $|N : P|$, Cauchy's theorem guarantees the existence of a subgroup of $M/P$ of order $p$ in the quotient group $N/P$. $M$ is then a $p$-subgroup of order larger than $|P|$, which is impossible. Sylow next proves that $P$ contains all elements of $p$-power order in $N$. His proof is essentially the same as any modern one. Suppose $\phi$ is an element of $N$ not in $P$. The elements $\vartheta\phi^i$, where $\vartheta$ runs over $P$ and $i$ over the integers, form a subgroup of $N$ properly containing $P$. The order of this subgroup is $|P|m$, where $m$ is the smallest positive integer $j$ such that $\phi^j \in P$. But $m$ clearly divides the order of $\phi$. Since $P$ is a $p$-subgroup of maximal order, $\phi$ cannot have order a power of $p$, as required.

The crucial part of the proof is to show that $|G : N|$ is not divisible by $p$. Once this is known, we see that $|P| = p^a$, and the existence of Sylow $p$-subgroups is established. In fact, Sylow shows that $|G : N| \equiv 1 \bmod p$, which is another part of his basic theorem. The following would seem to be a modern version of his argument. $P$ permutes the left cosets of $N$ in $G$ by left multiplication. It fixes $N$, because it is contained in $N$. It fixes no other left coset. For if $P$ fixes the coset $\psi N$, we have $\psi^{-1}P\psi \leq N$. But the argument above shows that the $p$-subgroup $\psi^{-1}P\psi$ now contained in $N$ equals $P$, as it has the same order as $P$. Hence $\psi \in N$ and $\psi N = N$, as required. The left cosets of $N$ different from $N$ fall into $P$-orbits of size greater than 1, and the size of each orbit is a power of $p$, as it divides the order of $P$, by the orbit-stabilizer theorem. This proves what Sylow gives as his first theorem, where we return to Sylow's original notation:

*Si $n^\alpha$ désigne la plus grande puissance du nombre premier $n$ qui divise l'ordre du groupe $G$, ce groupe contient un autre $g$ de l'ordre $n^\alpha$; si de plus $n^\alpha\nu$ désigne l'ordre du plus grand groupe contenu dans $G$ dont les substitutions sont permutables à $g$, l'ordre de $G$ sera de la forme $n^\alpha\nu(np+1)$.*

Sylow's second theorem is the following:

*Tout étant posé comme au théorème précédent, le groupe $G$ contient précisement $np+1$ groupes distincts d'ordre $n^\alpha$; on les obtient tous en transformant l'un quelconque d'entre eux par les*

*substitutions de $G$, tout groupe étant donné par $n^\alpha\nu$ transformantes distinctes.*

His proof is the following (returning to our notation). Let $Q$ be a subgroup of order $|P|$. $Q$ permutes the left cosets of $N$ in $G$ into orbits, the size of each $Q$-orbit being a power of $p$. As the number of orbits is congruent to 1 modulo $p$, it fixes a coset. Thus $\psi^{-1}Q\psi \leq N$ for some $\psi$. But the argument of the previous paragraph shows that $\psi^{-1}Q\psi = P$, as $P$ contains all $p$-elements in $N$. Sylow notes that the same argument proves that any $p$-subgroup of $G$ is contained in a conjugate of $P$. Thus the standard results comprising Sylow's theorem are obtained.

Having proved his main theorems, Sylow continues his paper by considering the conjugating action of the $p$-group $P$ on itself. $P$ acts on $P$ according to the rule

$$\vartheta \to \phi^{-1}\vartheta\phi.$$

This is a permutation action, since

$$\vartheta_2^{-1}\vartheta_1^{-1}\phi\vartheta_1\vartheta_2 = (\vartheta_1\vartheta_2)^{-1}\phi(\vartheta_1\vartheta_2).$$

The orbits of $P$ acting in this way are its conjugacy classes (not so-called by Sylow) and their sizes are powers of $p$. Since the identity of $P$ forms a single orbit, we have an equation of the form

$$p^\alpha = 1 + p^a + p^b + \cdots$$

This implies that at least $p-1$ of the indices $a$, $b$, ..., are 0. Thus, in modern terminology, the centre of $P$ is non-trivial (the argument is unchanged to this day).

An element $\vartheta_0$ of order $p$ may then be found in the centre. If $\Theta_0$ denotes the subgroup of $P$ generated by $\vartheta_0$, Sylow essentially forms the quotient group $P/\Theta_0$ of order $p^{\alpha-1}$. This group has a non-trivial centre. Let $\vartheta_1$ project onto an element of order $p$ in the centre of $P/\Theta_0$. Then $\vartheta_1^p = \vartheta_0^a$. Furthermore

$$\vartheta^{-1}\vartheta_1\vartheta = \vartheta_0^b\vartheta_1$$



for all $\vartheta$ in $P$ (here $b$ depends on $\vartheta$). The elements of the form $\vartheta_0^i\vartheta_1^k$ form a (normal) subgroup of $P$ of order $p^2$. Continuing in this way, we then obtain $\vartheta_2$ so that

$$\vartheta_2^p = \vartheta_0^c\vartheta_1^d$$
$$\vartheta^{-1}\vartheta_2\vartheta = \vartheta_0^e\vartheta_1^t\vartheta_2$$

for all $\vartheta \in P$ (the exponents again depending on $\vartheta$). This leads to Sylow's third theorem:
*Si l'ordre d'un groupe est $n^\alpha$, $n$ étant premier, une substitution quelconque $\vartheta$ du groupe peut être exprimée par la formule*

$$\vartheta = \vartheta_0^i\vartheta_1^k\vartheta_2^l\ldots\vartheta_{\alpha-1}^r$$

*où*

$$\vartheta_0^n = 1$$
$$\vartheta_1^n = \vartheta_0^a$$
$$\vartheta_2^n = \vartheta_0^b\vartheta_1^c$$
$$\vartheta_3^n = \vartheta_0^d\vartheta_1^e\vartheta_2^f$$
$$\ldots\ldots\ldots\ldots$$

*et où l'on a*

$$\vartheta^{-1}\vartheta_0\vartheta = \vartheta_0$$
$$\vartheta^{-1}\vartheta_1\vartheta = \vartheta_0^\beta\vartheta_1$$
$$\vartheta^{-1}\vartheta_2\vartheta = \vartheta_0^\gamma\vartheta_1^\delta\vartheta_2$$
$$\vartheta^{-1}\vartheta_3\vartheta = \vartheta_0^\varepsilon\vartheta_1^\zeta\vartheta_2^\eta\vartheta_3$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$

Thus, Sylow obtains the beginnings of the structure theory for $p$-groups, showing in particular that such groups are solvable.

As we explained earlier, we have tried to render Sylow's proof into a modern formulation. To give some of the flavour of Sylow's version, we describe his proof of the fact that $p$ does not divide $\nu = |N:P|$. $P$ is a permutation group of degree $r$, say, and so may be thought to act on certain variables $x_1, \ldots, x_r$. Let $y_0$ be a rational function of the $x_i$ which is invariant under $P$ but fixed by no other

permutation. This function takes $\nu$ different values under the action of $N$, each of which is fixed by $P$. A homomorphic image $N'$ of $N$ acts faithfully and transitively on these functions (following the custom of the time, Sylow uses the word isomorphic rather than homomorphic). ($N'$ is just the quotient group $N/P$.) If $p$ divides $\nu$, then $N'$ contains a permutation of order $p$ by Cauchy's theorem and Sylow obtains a contradiction to this using the same line of reasoning that proved that $P$ contains all elements of $p$-power order in $N$.

### 3. Cauchy's theorem

We turn now to a brief look at Cauchy's theorem, which was vital to Sylow's proof. The paper, [2], in which Cauchy's proof appears is well worth studying. It is 102 pages long and its general spirit is quite close to modern algebra, unlike that of many ostensibly algebraic papers of the 19th century, which are often hopelessly vague. Much of the elementary theory of permutations may be found there. For example, the size of a conjugacy class of $S_n$ containing an element of a given cycle type is determined. Among other things, Cauchy gives an explicit construction of a Sylow $p$-subgroup of $S_n$ ([2, pp.195-196]). This is interesting in itself, as it requires the idea of a wreath product. Wreath products play a vital role in the study of permutation and linear groups.

The concept of a double coset decomposition of a group relative to two subgroups is implicit in §12 of [2]. To paraphrase Cauchy's argument, the following is proved. Let $G$ be a finite group containing subgroups $A$ and $B$. Suppose that no non-identity element of $A$ is conjugate to an element of $B$. Then the size of a double coset $AgB$ is $|A||B|$. Moreover, $G$ is the disjoint union of all the different double cosets. Consequently, with the hypothesis as above, $|A||B|$ divides $|G|$. Cauchy applies this when $G = S_n$, $A$ is a Sylow $p$-subgroup of $S_n$ (which he has already constructed) and $B$ is a subgroup whose order is divisible by the prime $p$. Since $|A||B|$ cannot divide $n!$, a non-trivial element of $A$ is conjugate to an element of $B$ and thus Cauchy's theorem is proved. With a little more care and determination, the existence part of Sylow's theorem might easily have been obtained by



Cauchy 30 or more years before Sylow's proof. Cauchy's proof of his theorem is reproduced in Jordan's famous treatise, [6, pp.26-29]. Cauchy's theorem applies to a subgroup of $S_n$, but Cayley's embedding theorem, that a finite group $G$ is isomorphic to a subgroup of $S_{|G|}$, shows that it applies to any abstract finite group.

### 4. Later proofs of Sylow's theorem

Fairly soon after the publication of Sylow's theorem in 1872, attempts were made to avoid the use of Cauchy's theorem in its proof. In 1877, Eugen Netto gave a proof in [7] which used only part of the proof of Cauchy's theorem. In Netto's situation, as in Sylow's, we have a subgroup $G$ of $S_n$ of order $k$, where $k$ is divisible by the prime $p$. (Note that Netto uses what has become standard notation in respect of $n$ and $p$). He assumes Cauchy's constructive result that $S_n$ contains a Sylow $p$-subgroup, $H$, say, of order $p^f$, and then proves that $G$ contains a Sylow $p$-subgroup. We found Netto's proof difficult to follow, but it seems clear that he is using a decomposition of $S_n$ into $(G, H)$-double cosets. He obtains the equation

$$\frac{n!}{k}\frac{n!}{p^f} = \frac{n!}{p^\alpha} + \frac{n!}{p^\beta} + \cdots = \frac{n!}{p^\alpha}s,$$

where $s$ is an integer, and $p^\alpha \geq p^\beta$, and so on. Multiplying each side above by $kp^f/n!$, we obtain the usual equation expressing the order of $S_n$ as the sum of the sizes of the different $(G, H)$-double cosets. The powers of $p$ that appear in the denominators are the orders of the intersections of $G$ with various conjugates of $H$. Netto's proof, in double coset form, has become a standard one. An alternative is to embed a finite group into a finite general linear group $\mathrm{GL}(n, p)$, where $p$ is a prime, and use the fact that the linear group contains an explicit Sylow subgroup, consisting of lower triangular matrices with all diagonal entries equal to 1. See, for example, the exercises on p.36 of [5].

The proof that was to become the standard proof of the existence of Sylow subgroups until 1959 is that of Frobenius, [3]. Although it appeared in 1887, it is dated Zürich, March 1884. Perhaps this is evidence that the publication backlogs of journals

are not a new phenomenon. Frobenius aims to remove all reference to Cauchy's work and keep the discussion as elementary as possible. The proof is by induction, the main tool to be used being the conjugacy class equation in a finite group. The concept of a quotient group is also required. Frobenius works with an abstract finite group $H$, noting that it may be considered as a group of permutations. He also notes that his abstract finite group is defined by three axioms, which he states. He then considers the centre, $G$, of $H$ and supposes that $p$ divides its order. He shows (without using Cauchy's theorem) that the centre contains an element $P$ of order $p$. He declares that two elements of $H$ are to be considered (relatively) equal if they differ only by a power of $P$. The relatively different elements form a group, whose order is $|H|/p$ (this is the quotient group of $H$ modulo the cyclic group generated by $P$). By induction, this group has a Sylow subgroup which lifts back to give a Sylow subgroup of $H$.

Finally, he supposes that $p$ does not divide $|G|$. The conjugacy class equation shows that there must be an element not in the centre, the size of whose conjugacy class is relatively prime to $p$. But then the $p$-part of the order of the centralizer of this element equals the $p$-part of $|H|$, and since the order of this centralizer is less than that of $H$, by induction, the centralizer contains a Sylow $p$-subgroup of $H$, as required. This proof may be found in such early textbooks as those of Burnside, [1], and Hilton, [4].

## 6. Life and work of Sylow

We close by making a few remarks about the life and career of Sylow. Sylow (1832-1918) taught from 1858 to 1898 at a school in Halden (Frederikshald) in Norway. A town of this name is located south of Oslo, near the Swedish border. A chair was created for him in 1898 at Christiana (Oslo) University. His other main paper is [10], devoted to complex multiplication of elliptic functions and associated singular moduli. He seems to have been drawn to this subject while editing a new edition of the collected work of Abel, his famous compatriot, who contributed important early work on elliptic functions. Sylow's 1872 paper showed that he had considerable talent in abstract algebra and it is a pity that he did not get more opportunity to put his talent into effect.



### References


[1] W. Burnside, Theory of Groups of Finite Order. Cambridge Univ. Press: Cambridge, 1897.

[2] A. L. Cauchy, *Mémoire sur les arrangements que l'on peut former avec des lettres données*, Exercices d'analyse et de physique mathématique, tome III. Bachelier: Paris, 1844.

[3] F. G. Frobenius, *Neuer Beweis des Sylowschen Satzes*, J. für die reine und angewandte Math. **100** (1887), 179-181.

[4] H. Hilton, An Introduction to the Theory of Finite Groups. Clarendon Press: Oxford, 1908.

[5] B. Huppert, Endliche Gruppen I. Springer-Verlag: Berlin-Heidelberg-New York, 1967.

[6] C. Jordan, Traité des substitutions et des équations algébriques. Gauthier-Villars: Paris, 1870.

[7] E. Netto, *Neuer Beweis eines Fundamentaltheorems aus der Theorie der Substitutionslehre*, Math. Ann. **13** (1878), 249-250.

[8] J.-A. Serret, Cours d'algèbre supérieure, tome II (quatrième édition). Gauthier-Villars: Paris, 1879.

[9] L. Sylow, *Théorèmes sur les groupes de substitutions*, Math. Ann. **5** (1872), 584-594.

[10] L. Sylow, *Sur la multiplication complexe des fonctions elliptiques*, Journal de mathématiques pures et appliquées (quatrième série) **3** (1887), 109-254.

[11] H. Wielandt, *Ein Beweis für die Existenz der Sylowgruppen*, Arch. Math. **10** (1959), 401-402.


Rod Gow,
Department of Mathematics,
University College,
Belfield,
Dublin 4.

## Book Review

### Algebraic Function Fields and Codes

H. Stichtenoth
Springer-Verlag 1993, x+260 pp.
ISBN 3-540-56489-6
Price $34.00.

**Reviewed by** Gary McGuire

The purpose of this book is two-fold. Firstly, to give an exposition of the basic theory of algebraic function fields using a purely algebraic approach. Secondly, to give the applications of this theory to the theory of error-correcting codes. We refer to the book under review as [S]. Before we begin the review, we shall briefly discuss the topics covered in the book.

### 1. Algebraic Function Fields

Let $K$ be any field. An algebraic function field $F/K$ of one variable over $K$ is an extension field $F$ of $K$ such that $F$ contains an element $x$ which is transcendental over $K$, and $F$ is an algebraic extension of finite degree over $K(x)$.

The algebraic approach to studying such extensions $F/K$ was first taken by Dedekind and Weber [4], with $K$ the complex numbers. Chevalley [3] treated arbitrary $K$ with this approach, and discussed geometry only with $K$ the complex numbers. Algebraic geometry enters the picture as soon as one considers the plane algebraic curve $C$ arising from $F/K$. This is defined by a polynomial equation $f(x,y) = 0$, where $F = K(x,y)$ and $f$ has coefficients in $K$. Conversely, given a curve $C$ defined by some irreducible polynomial $f \in K[x,y]$, the quotient field of the domain $K[x,y]/(f)$ is an algebraic function field of one variable, usually denoted $K(C)/K$. The geometric approach has been taken by many authors: Noether [13], Severi [14], Weil [19], and more recently one may consult Shafarevich [15], Hartshorne [7].





Two classics are Fulton [5] and Walker [18]. For a more sketchy presentation but with all the ideas, see Abhyankar [1] or Moreno [12].

In his review of [3], Weil [20] almost chastises Chevalley for the lack of geometry:

*Here is algebra with a vengeance; ... if it were not for a few hints ... one might never suspect him of ever having heard of algebraic curves or of taking any interest in them.*

He later concedes, it should be pointed out, that "this is a valuable and useful book". Not the least of the reasons for this is the strong analogy between the algebraic approach to algebraic functions (Chevalley) and the theory of algebraic numbers, *viz.* primes and irreducible polynomials, rational numbers and rational functions. For a simultaneous treatment of algebraic functions and algebraic numbers, see Artin [2].

Of course the "geometric" approach is through algebraic geometry, and involves a nontrivial amount of algebra itself. It would seem that this approach is the more popular. It is in reality a pleasant mixture of both algebra and geometry. For example, there is a one-to-one correspondence between points on a nonsingular curve $C$ and places (maximal ideals of valuation rings) of $K(C)/K$.

A cornerstone of either approach is the Riemann-Roch Theorem (see Chapter I). A divisor $A$ is a formal sum $\sum_P n_P P$ where the $n_P$ are integers and only finitely many are nonzero. The sum is over all points on a curve, or all places of a function field, depending on one's point of view. The degree of a divisor, $deg(A)$, is $\sum_P n_P$. Assuming the existence of something called a canonical divisor, $W$, and the divisor of any $f \in F$, denoted $(f)$, the Riemann-Roch theorem states that

$$\ell(A) - \ell(W - A) = deg(A) + 1 - g$$

where $g$ is the genus of the curve $C$, or the function field $F/K$, and $\ell(A)$ is the dimension of the $K$-vector space

$$\mathcal{L}(A) = \{f \in F : (f) \geq -A\} \cup \{0\}.$$

The Riemann-Roch Theorem has many important consequences. For example, if $f$ and $h$ are two curves of degrees $m$ and $n$ over the complex numbers (let us say), then by Bézout's Theorem [1] $f$ and $h$ intersect in $mn$ points (counted properly). Conversely, given $f$ and $mn$ points, does there exist a curve $h$ of degree $n$ which intersects $f$ in precisely those $mn$ points? Algebraically, we might ask: given specified poles and zeroes with multiplicities, does there exist $f \in K(x)$ with exactly those poles and zeroes? The Riemann-Roch Theorem provides answers.

## 2. Error-Correcting Codes

A code $C$ over an alphabet $Q$ is a subset of $Q^n$. Elements of $C$ are called codewords, and $n$ is called the length of the code. Usually $Q$ is taken to be $\mathbf{F}_q$, the finite field of $q$ elements. A linear code is a subspace of $\mathbf{F}_q^n$, and we assume linearity from now on. A code is called a code and not a subspace because of interest in a rather non-algebraic property, its minimum distance $d$. For $x, y \in Q^n$, the Hamming distance between $x$ and $y$, $d(x,y)$, is defined to be the number of coordinates where $x$ and $y$ differ. For example, the distance between 110101 and 111100 ($q = 2$) is 2. Then $d$ is defined by

$$d := min\{d(x,y) : x, y \in C, x \neq y\}.$$

If $C$ is a $k$-dimensional subspace of $\mathbf{F}_q^n$, we say that $C$ is a $q$-ary $[n,k,d]$ code. If $e = \lfloor \frac{d-1}{2} \rfloor$, $C$ is an $e$-error-correcting code. This is because in practice, codewords are transmitted over a channel to a receiver. Due to noise there may be errors introduced during transmission, but if there are not more than $e$ errors, the receiver can correct them and decode the received vector to the unique nearest codeword. Error-correcting codes are used every day in compact disc players, and have been used by NASA to receive data from space probes such as Mariner and Voyager. For an introduction to the theory of error-correcting codes, see [9] or [21].

For a fixed $n$ (and $q$), a central problem in coding theory is to find codes which maximize both $k$ and $d$. Unfortunately, the Singleton bound (trivial to prove) says

$$k + d \leq n + 1,$$



and so these aims are contradictory.

Much work has been done on bounds relating $n$, $k$, $d$ and $q$. For asymptotic bounds, applicable for large $n$, the simplest results are obtained when the rate $R = k/n$ is plotted as a function of $\delta = d/n$. Clearly

$$R + \delta \leq 1 + \frac{1}{n}.$$

In fact, in Shannon's Theorem, the desirable codes whose existence is proven have very large $n$. However, constructing such codes is another matter.

Following [9], a family of codes over $\mathbf{F}_q$ ($q$ fixed) is said to be *good* if it contains an infinite sequence of codes $C_i$, where $C_i$ is an $[n_i, k_i, d_i]$ code, such that both the rate $R_i = k_i/n_i$ and $\delta_i = d_i/n_i$ approach a nonzero limit as $i \to \infty$.

Examples of classical families of codes are Hamming codes, BCH codes, Reed-Solomon codes and Reed-Muller codes. These codes have nice algebraic constructions and properties. It turns out that all these families are bad. Construction of good families became a problem. J.L. Massey said

*... good codes just might be messy.*

Justesen (1972) constructed an infinite family of good binary codes, see [9].

That good codes exist was never in doubt: the Gilbert-Varshamov lower bound states that if $R$ is fixed, $0 \leq R \leq 1$, then there exist binary $[n,k,d]$ codes with $k/n \geq R$ and $d/n \geq H_2^{-1}(1-R) > 0$ where $H_2^{-1}(x)$ is the inverse of the entropy function $H_2(x) = -xlog_2(x) - (1-x)log_2(1-x)$.

There is an analogous statement of the Gilbert-Varshamov lower bound for any $q$, which can be translated into a lower bound for a function $\alpha_q(\delta)$ (which we leave undefined; see [S], Chapter VII).

## 3. Algebraic Function Fields and Codes

The main idea is that algebraic function fields can be used to construct codes which lead to an improved lower bound for $\alpha_q(\delta)$. It was thought for over thirty years that the Gilbert-Varshamov

lower bound would prove to be exact. Hence the improved lower bound caused a sensation in the field.

The improvement came in two stages. First came a construction from Goppa (1981) – after many years of trying to generalize his earlier work – of codes from algebraic function fields. We summarize this construction; it is fully described in [S], Chapter II. Let $F/\mathbf{F}_q$ be an algebraic function field of genus $g$ and let $P_1, \ldots, P_n$ be pairwise distinct places of $F/\mathbf{F}_q$ of degree one. Let $D$ be the divisor $P_1 + \cdots + P_n$ and let $G$ be a divisor of $F/\mathbf{F}_q$ with disjoint support from $D$. The geometric Goppa code $C(D, G)$ (also called an algebraic geometry code [16]) is the image of the linear map $\beta : \mathcal{L}(G) \to \mathbf{F}_q^n$ defined by

$$\beta(f) = (f(P_1), \ldots, f(P_n)).$$

The dimension and a bound on the minimum distance are found by using the Riemann-Roch Theorem.

Asymptotic values of the ratio of the number of places of degree one to the genus (as $g \to \infty$) are related to whether geometric Goppa codes are good. Hence bounds on these asymptotic values (from algebraic geometry) can be related to the Gilbert-Varshamov lower bound.

The definition above can be phrased in terms of nonsingular curves, which is how Goppa first described it.

Choosing $F = \mathbf{F}_q(t)$, these codes are the Reed-Solomon codes mentioned earlier. Hence geometric Goppa codes are a natural generalization of Reed-Solomon codes.

The second stage of the improvement came about by finding certain suitable function fields (curves) $F/\mathbf{F}_q$. This was done by Ihara [8], and independently by Tsfasman, Vladut and Zink [17], although work of Manin [10] (and presumably unpublished) paved the way. Manin and Vladut [11] gave a proof using Drinfeld modules. The improved lower bound is valid for $q \geq 49$ and square, so in particular we are still without any codes beating the Gilbert-Varshamov bound in the binary case. The suitable curves that were used are Shimura curves [8], or Drinfeld curves [16], first suggested in [10]. [S] does not describe this second stage.



## 4. Review

Chapter I contains Weil's proof of the Riemann-Roch theorem using adeles. The author defines a Weil differential as a linear map on the space of adeles, vanishing on some translate of $F$ (embedded). As the namesake of these differentials says,

*This rather abstract concept of differential is of course what makes possible such a brief proof of the Riemann-Roch Theorem.*

Later in Chapter IV these Weil differentials are identified with our "usual" notion of a differential. Chapter I also contains the Strong Approximation Theorem (crucial to many proofs in the book), Weierstrass gaps, and local components of Weil differentials (later to become residues of our usual differentials). This chapter is self-contained, requiring only basic graduate algebra. A useful appendix is provided with a summary of field theory.

In Chapter II the reader will find an introduction to coding theory and the definition of geometric Goppa codes. The dual code of $C(D, G)$ is also defined using local components of Weil differentials (later residues), and that it is the dual is proved using what is "really" the residue theorem (although not called such). Here we see perhaps the disadvantage to the algebraic approach. BCH and classical Goppa codes are constructed from geometric Goppa codes as subfield subcodes.

Chapter III (Extensions of Algebraic Function Fields) is the longest and most technical in the book. The presentation of many important ideas goes straight to the key theorems, and the proofs are concise but complete. As in Chapter I, however, the reader must make up his or her own examples in all sections except III.7. Topics covered include extensions and ramification, the different and the Hurwitz Genus formula, constant field extensions, Galois extensions (Kummer and Artin-Schreier), wild ramification, inseparability, and Castelnuovo's Inequality for the genus. A knowledge of algebraic number theory is useful (for familiarity purposes) but certainly not essential.

Chapter IV defines differentials via derivations and proves a one-to-one correspondence with Weil differentials from Chapter I.

We also find the $P$-adic completion of $F/K$ with respect to a place $P$, giving us $P$-adic power series, analogous to complex power series over $\mathbf{C}$. Here lies the residue theorem, another cornerstone of the theory.

The length of the code $C(D, G)$ is limited by how many places of degree one the extension $F/\mathbf{F}_q$ has. The Hasse-Weil Theorem (also known as the Riemann hypothesis for function fields over finite fields, proved by Weil) tells us approximately how many places of degree one we can expect. This theorem is famous and has many implications, both inside and outside this book. Chapter V defines the zeta function of $F/\mathbf{F}_q$ and presents Bombieri's short elementary proof of Hasse-Weil. It uses only the Riemann-Roch Theorem. As Manin [10] points out, the proof of the upper bound is "quite code-theoretic in spirit". Again the presentation is faultless. Improvements to the Hasse-Weil bound are given with proofs, including the asymptotic lower bound due to Drinfeld-Vladut. This bound was proved to be tight (for $q \geq 49$ and square) by Ihara and Tsfasman-Vladut-Zink (by constructing the suitable curves to give equality).

The reader may heave a sigh of relief upon seeing the title of Chapter VI – Examples of Algebraic Function Fields. The author does say (page 30) that "we defer such examples to Chapter VI at which point we will have better methods at hand for calculating the genus". Indeed, results from all previous chapters are drawn on to give a thorough treatment of elliptic function fields. It is a pleasure to see characteristic 2 not excluded. Next are hyperelliptic function fields, and more generally, function fields $F = K(x, y)$ defined by $y^n = f(x)$. Examples done are Fermat ($ax^n + by^n = c$) and Hermitian ($x^{q+1} + y^{q+1} = 1$) function fields.

Chapter VII concerns the Gilbert-Varshamov bound and the story recounted earlier. Automorphism groups of geometric Goppa codes are discussed and applied to Hermitian codes (from the Hermitian function field). A decoding algorithm for geometric Goppa codes due to Skorobogatov and Vladut (following Justesen) is presented. However these codes are still a long way from being used in practice.

The final chapter (VIII) discusses the trace code of a $q^m$-ary code $C$ of length $n$, which is defined as $Tr_q^{q^m}(C) \subseteq \mathbf{F}_q^n$, with trace taken componentwise. In certain special cases the minimum distance and perhaps all the weights in these codes can be found, or at least bounded. The bounds can be tight. Again results from previous chapters (especially chapter III) are used regularly.



A useful feature of this book is the two appendices, one containing a summary of field theory and the other explaining how to switch between the algebraic and geometric approaches, i.e. function fields and curves. The reader interested in curves can consult this second appendix and translate the results in the text to results about curves. Curves are not mentioned at all during the text, in keeping with the author's promise of an algebraic exposition. The reader may also find it helpful to glance at Chapter VI while reading the earlier chapters (especially I, III and IV), in order to see some examples.

The book is approximately 250 pages long and reasonably priced. It is typeset with some form of TEX, and one can have few complaints about that. A minor quibble concerns the letters of "Gal" and "Aut" (page 109), and "Der" (page 137), which are too close together. "Aut" has been corrected by page 209. The only typographical error this reviewer found (apart from a trivial one on page 144) is on page 243, where "monom" should be monomial. There are no exercises.

The exposition in this book is clean and tight. The quickest proofs of all the theorems are given, with no time wasted. The book is entirely self-contained. The author accomplishes what he set out to do with simplicity. It is recommended for anyone interested in algebraic function fields and their applications to codes.

### References

[S] H. Stichtenoth, Algebraic Function Fields and Codes. Springer-Verlag: Berlin, 1993.

[1] S.S. Abhyankar, Algebraic Geometry for Scientists and Engineers. Mathematical Surveys and Monographs, Vol. 35, American Mathematical Society: New York, 1990.

[2] E. Artin, Algebraic Numbers and Algebraic Functions. Gordon and Breach: New York, 1967.

[3] C. Chevalley, Introduction to the Theory of Algebraic Functions of One Variable. Mathematical Surveys and Monographs, Vol. 6, American Mathematical Society: New York, 1951.

[4] R. Dedekind and H. Weber, *Theorie der algebraischen functionen einer veränderlichen*, Crelle J. **92** (1882), 181-290.

[5] W. Fulton, Algebraic Curves. Benjamin: New York, 1969.

[6] V.D. Goppa, Geometry and Codes. Mathematics and Its Applications, Vol. 24, Kluwer Academic Publishers: Dordrecht, 1988.

[7] R. Hartshorne, Algebraic Geometry. Graduate Texts in Mathematics, Vol. 52, Springer-Verlag: New York, 1977.

[8] Y. Ihara, *Some remarks on the number of rational points of algebraic curves over finite fields*, J. Fac. Sci. Univ. Tokyo **28** (1981), 721-724.

[9] F. J. McWilliams and N. J. A. Sloane, The Theory of Error-Correcting Codes. North-Holland: Amsterdam, 1977.

[10] Yu. I. Manin, *What is the maximum number of points on a curve over* $\mathbf{F}_2$?, J. Fac. Sci. Univ. Tokyo **28** (1981), 715-720.

[11] Yu. I. Manin and S. G. Vladut, *Linear codes and modular curves*, J. Soviet Math. **30** (1985), 2611-2643.

[12] C. J. Moreno, Algebraic Curves over Finite Fields. Cambridge Tracts in Mathematics, Vol. 97, Cambridge University Press: Cambridge, 1991.

[13] M. Noether, *Über einen satz aus der theorie der algebraischen functionen*, Math. Ann. **6** (1873), 351-359.

[14] F. Severi, Vorlesungen über Algebraische Geometrie. Teubner: Leipzig, 1921.

[15] I. R. Shafarevich, Basic Algebraic Geometry. Grundlehren der Mathematischen Wissenschaften, Vol. 213, Springer-Verlag: Berlin, 1977.

[16] M. A. Tsfasman and S. G. Vladut, Algebraic-Geometric Codes. Kluwer: Dordrecht, 1991.

[17] M. A. Tsfasman, S. G. Vladut and T. Zink, *Modular curves, Shimura curves and Goppa codes, better than Varshamov-Gilbert bound*, Math. Nachr. **109** (1982) 21-28.

[18] R. J. Walker, Algebraic Curves. Princeton University Press: Princeton, 1950.

[19] A. Weil, Sur les courbes algébriques et les variétés qui s'en déduisent. Hermann: Paris, 1948.



[20] A. Weil, Review of "Introduction to the Theory of Algebraic Functions of One Variable", Bull. A.M.S. **57** (1951), 384-398.

[21] J. H. van Lint, Introduction to Coding Theory. Graduate Texts in Mathematics, Vol. 86, Springer-Verlag, Second Ed.: Berlin, 1992.

Gary McGuire,
Department of Mathematics,
California Institute of Technology,
Pasadena, CA 91125,
USA.

## Outline Solutions of the Problems for the 35th IMO

**1.** Without loss of generality $a_1 < a_2 < \ldots < a_m$. Suppose $a_i + a_{m+1-i} \leq n$, for some $i$ with $1 \leq i \leq m$. Then $a_j + a_{m+1-i} \leq n$, for $j = 1, 2, \ldots, i$. But then the $i$ distinct integers $a_j + a_{m+1-i}$, $j = 1, 2, \ldots, i$ must lie in the set $\{m, m-1, \ldots, m-i+2\}$, which contains only $i-1$ elements. Thus $a_i + a_{m+1-i} \geq n+1$, for $i = 1, 2, \ldots, m$. Add these inequalities to obtain the result.

**2.** Use coordinates. Without loss of generality, let $M = (0,0)$, $B = (-1,0)$ $C = (1,0)$. Let $A = (0,a)$ and $Q = (t,0)$. The rest of the solution is straightforward.

**3.** $(a)$ Let $A_k$ be the set of integers in $\{1, 2, \ldots, k\}$ whose base 2 representation contains exactly three 1's and let $g(k)$ be the number of elements in $A_k$. Then $f$ and $g$ are nondecreasing functions and $f(k) = g(2k) - g(k)$. Then

$$f(k+1) - f(k) = g(2k+2) - g(2k) - (g(k+1) - g(k)).$$

Now either both $2k+2 \in A_{2k+2}$ and $k+1 \in A_{k+1}$ or neither is true. Thus $f(k+1) - f(k) = 0$ or 1, depending on whether $2k+1 \in A_{2k+1}$ or not. Thus $f(k)$ does not skip any positive integer values. Since

$$g(2^n) = \binom{n}{3} = g(2^n - 1),$$

we get, after some calculation, $f(2^n) = \binom{n}{2}$. Thus $f$ is not bounded above and hence assumes every non-negative integer value.

$(b)$ Suppose $f(k) = m$ has a unique solution. Then

$$f(k+1) - f(k) = 1 = f(k) - f(k-1).$$

The former holds if and only if $2k+1 \in A_{2k+2}$, i.e. there are exactly two 1's in the base 2 digits of $k$. The same holds for $k-1$.





This is possible if and only if $k-1$ has exactly two 1's in its base 2 representation, where the last digit is 1 and the second last digit is 0, i.e. $k = 2^n + 2$ for some integer $n \geq 2$. A calculation gives

$$f(2^n + 2) = \binom{n}{2} + 1.$$

Thus the set of positive integers $m$ for which $f(k) = m$ has a unique solution is $\{\binom{n}{2} + 1 : n \geq 2\}$.

**4.** We note that

$$\frac{n^3+1}{mn-1} + 1 = \frac{n(n^2+m)}{mn-1}$$

and that

$$\frac{m(n^2+m)}{mn-1} - n = \frac{m^2+n}{mn-1}.$$

Thus $mn-1$ divides $n^3+1$ if and only if it divides $m^2+n$ and this holds if and only if $mn-1$ divides $m^3+1$.

If $m = n$ it is easy to see that $m = 2$.

If $m > n$, then $\frac{n^2+m}{mn-1} = k$, an integer, implies that $n^2 + k = m(kn-1) > kn^2 - n$ and thus $(k-1)n^2 - n - k < 0$. This implies that $n < \frac{k}{k-1}$, if $k > 1$.

If $k = 1$, then $n^2 + m = mn - 1$. Thus $m = n + 1 + \frac{2}{n-1}$. The fact that $n-1$ divides 2 proves that $n = 2$ or 3. If $n = 2$, then $m = 5$ and if $n = 3$ then $m = 5$.

If $k > 1$, then $n < \frac{k}{k-1} \leq 2$ implies that $n = 1$. Then $m = 2$ or 3.

Thus, if $\frac{n^3+1}{mn-1}$ is an integer, $(m,n)$ is one of the pairs:

$$(1,2), (1,3), (2,1), (3,1), (2,5), (3,5), (5,2), (5,3), (2,2).$$

It is clear that $\frac{n^3+1}{mn-1}$ is an integer if $(m,n)$ is one of these nine pairs.

**5.** It is clear that $\frac{f(x)}{x}$ can take the value 1 at most once in each of the intervals (-1,0) and $(0,\infty)$. Let $f(a) = a$, then property $(i)$

implies that $f(2a + a^2) = 2a + a^2$. If $-1 < a < 0$, then $-1 < 2a + a^2 < 0$ and thus $a = 2a + a^2$. This gives the contradiction $a = 0$ or $-1$. Similarly, the assumption that $a > 0$ leads to a contradiction. Thus $f(a) = a$ implies $a = 0$. Using this fact and letting $y = x$ in $(i)$ proves

$$x + f(x) + xf(x) = 0,$$

for all $x$ in $S$. Thus

$$f(x) = \frac{-x}{1+x}$$

for all $x$ in $S$. It is clear that this function satisfies $(i)$ and $(ii)$ and is the only function with these two properties.

**6.** First solution. Let $A$ be the set of all positive integers of the form $q_1 q_2 \dots q_{q_1}$, where $q_1 < q_2 < \dots < q_{q_1}$ are primes. For any infinite set $\{p_1, p_2, p_3, \dots\}$ of primes $p_1 < p_2 < p_3 < \dots$, we can satisfy the requirements of the problem, by taking

$$m = p_1 p_2 \dots p_{p_1} \text{ and } n = p_2 p_3 \dots p_{p_1+1}.$$

Second solution. Let $\Pi = \{p_1, p_2, p_3, \dots\}$ denote the set of all primes. Let

$$A_i = \{q_1 q_2 \dots q_i : q_1, q_2, \dots, q_i \in \Pi \text{ and } p_i \not| q_1 q_2 \dots q_i\}$$

and let $A = A_1 \cup A_2 \cup A_3 \cup \dots$. Let $S$ be any infinite subset of $\Pi$ and let $p_k$ be in $S$. Choose distinct primes $q_1, q_2, \dots, q_k$ in $S - \{p_k\}$. Then $m = q_1 q_2 \dots q_{k-1} q_k$ is in $A$, whereas $n = q_1 q_2 \dots q_{k-1} p_k$ is not in $A$.

## INSTRUCTIONS TO AUTHORS

The Bulletin is typeset with TEX. Authors should if possible submit articles to the Bulletin as TEX input files; if this is not possible typescripts will be accepted. Manuscripts are not acceptable.

### Articles prepared with TEX

Though authors may use other versions of TEX, It is preferred that they write plain TEX files using the standard IMS layout files. These files can be received by sending an e-mail message to `listserv@irlearn.ucd.ie`. The body of the message should contain the three lines:

```
get imsform tex
get mistress tex
get original syn
```

Instructions on the use of these is contained in the article on *Directions in Typesetting* in issue number 27, December 1991.

The TEX file should be accompanied by any non-standard style or input files which have been used. Private macros, reference input files and both METAFONT and TEX source files for diagrams should also accompany submissions.

The input files can be transmitted to the Editor either on an IBM or Macintosh diskette, or by electronic mail to the following Bitnet or EARN address:

RODGOW@IRLEARN.UCD.IE

Two printed copies of the article should also be sent to the Editor.

### Other Articles

Authors who prepare their articles with word processors can expedite the typesetting of their articles by submitting an ASCII input file as well as the printed copies of the article.

Typed manuscripts should be double-spaced, with wide margins, on numbered pages. Commencement of paragraphs should be clearly indicated. Hand-written symbols should be clear and unambiguous. Illustrations should be carefully prepared on separate sheets in black ink. Two copies of each illustration should be submitted: one with lettering added, the other without lettering. Two copies of the manuscript should be sent to the Editor.