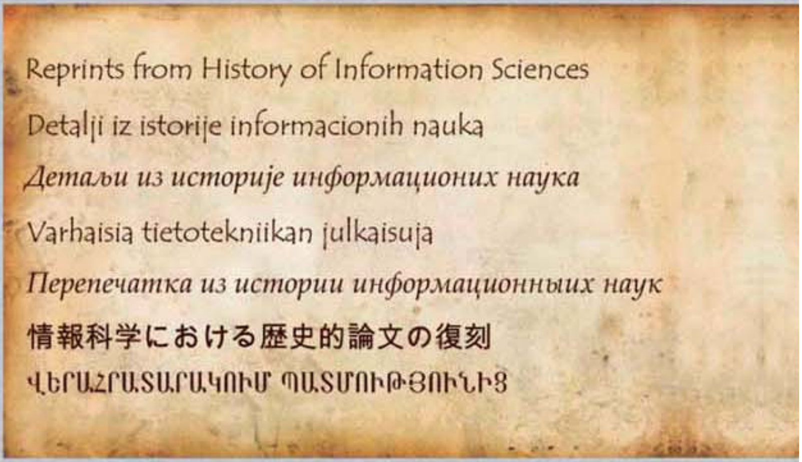
TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Tampere International Center for Signal Processing

Reprints from the Early Days of Information Sciences

Early Work of Aimo Tietäväinen in Number Theory and Coding Theory



Reprints from History of Information Sciences

Detalji iz istorije informacionih nauka

Детаљи из историје информационих наука

Varhaisia tietotekniikan julkaisuja

Перепечатка из истории информационныих наук

情報科学における歴史的論文の復刻

ՎԵՐԱՀՐԱՏԱՐԱԿՈՒՄ ՊԱՏՄՈՒԹՅՈՒՆԻՑ

Tampere 2012

Tampere International Center for Signal Processing. TICSP series # 59

Jaakko Astola & Radomir S. Stanković (eds.)

Reprints from the Early Days of Information Sciences
Early Work of Aimo Tietäväinen in Number Theory and Coding Theory

Tampere International Center for Signal Processing
Tampere 2012

ISBN 978-952-15-2817-0
ISSN 1456-2774

Reprints from the Early Days of Information Sciences

TICSP Series #59

Early Work of Aimo Tietäväinen

In Number Theory

and

Coding Theory

2012

Editors' Notice

This publication has been written and edited by
Jaakko T. Astola and Radomir S. Stanković.

Contents

Reprints from The Early Days of Information Sciences

Historical studies about a scientific discipline is a sign of its maturity. When properly understood and carried out, this kind of studies are more than enumeration of facts or giving credit to particular important researchers. It is more discovering and tracing the way of thinking that have lead to important discoveries. In this respect, it is interesting and also important to recall publications where for the first time some important concepts, theories, methods, and algorithms have been introduced.

In every branch of science there are some important results published in national or local journals or other publications that have not been widely distributed for different reasons, due to which they often remain unknown to the research community and therefore are rarely referenced. Sometimes the importance of such discoveries is overlooked or underestimated even by the inventors themselves. Such inventions are often re-discovered long after, but their initial sources may remain almost forgotten, and mostly remain sporadically recalled and mentioned within quite limited circles of experts. This is especially often the case with publications in other languages than the English language which is presently the most common language in the scientific world.

This series of publications is aimed at reprinting and, when appropriate, also translating some less known or almost forgotten, but important publications, where some concepts, methods or algorithms have been discussed for the first time or introduced independently on other related works.

Another aim of Reprints is to collect and present at the same place publications on certain particular subject of an important scholar whose scientific work is signified by contributions to different areas of sciences.

R.S. Stanković, J.T. Astola

Acknowledgments

Thanks are due to the Editors of journals from which the papers by Professor Aimo Tietäväinen are reprinted for the copyright permissions.

For the communication and help in clarification and receiving the copyright permissions we thank Mr. Mika Koskenoja (*Annales Academiae Scientiarum Fennicae*), Mr. Mikko Pennanen (*Annales Universitatist Turkuensis*), Ms. Bernadette Suarnaba (Copyright Clearance Center for the *Journal of Number Theory*), and Mr. Kelly Thomas (*SIAM Journal of Applied Mathematics*).

Special thanks are due to the artist Juhani Haaparinne, the art director of *Slave - Digital Boutique Agency*, Helsinki, Finland, for the permission to reprint the digital portrait of Professor Aimo Tietäväinen.

The Editors are grateful to Mrs. Marju Taavetti, the Librarian of the Library of Tempere University of Technology, and Mrs. Pirkko Ruotsalainen, the Development Manager of the Department of Signal Processing, Tampere University of Technology, for the help in collecting the related literature.

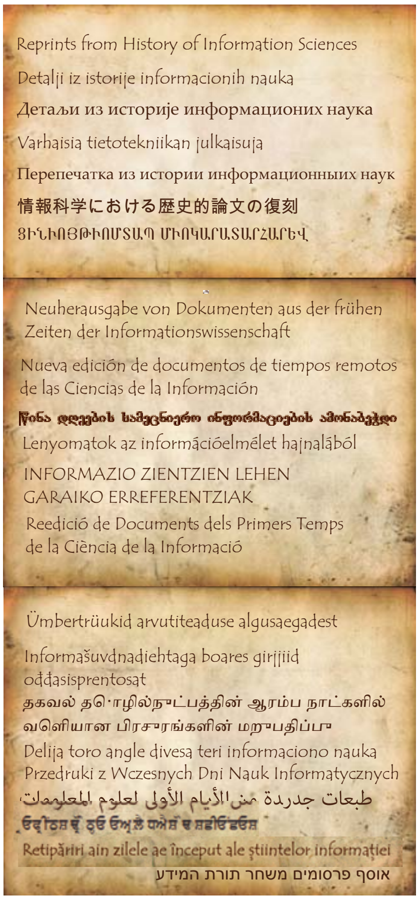
Early Work of Aimo Tietäväinen

Abstract

The present issue of Reprints from the Early Days of Information Sciences discusses research work of Aimo Tietäväinen. It presents 17 papers by Aimo Tietäväinen, and highlights the impact of this work to the research at the time in this area.

Notice

This book contains several reprints of pages from a book and a special issue of a journal edited in the honor of Professor Aimo Tietäväinen. They contain interesting historical information about Aimo Tietäväinen and the research community in general. We kindly ask for these reprints to not be considered simply as graphic illustrations from previous publications, but to be read as a part of the presentation in this book.



Reprints from History of Information Sciences

Detalji iz istorije informacionih nauka

Детаљи из историје информационих наука

Varhaisia tietotekniikan julkaisuja

Перепечатка из истории информационных наук

情報科学における歴史的論文の復刻

ՑԻՆԻՈՑԹԻՈՄՏԱՊ ՄԻՈԿԱՐԱՏԱՐՀԱՐԵՎ

Neuherausgabe von Dokumenten aus der frühen
Zeiten der Informationswissenschaft

Nueva edición de documentos de tiempos remotos
de las Ciencias de la Información

წინა დღეების სამეცნიერო ინფორმაციების ამონაბეჭდი

Lenyomatok az információelmélet hajnalából

INFORMAZIO ZIENTZIEN LEHEN
GARAIKO ERREFERENTZIAK

Reedició de Documents dels Primers Temps
de la Ciència de la Informació

Ümbertrüukid arvutiteaduse algusaegadest

Informašuvdnadiehtaga boares girjjiid
ođđasisprentosat

தகவல் தொழில்நுட்பத்தின் ஆரம்ப நாட்களில்
வெளியான பிரசுரங்களின் மறுபதிப்பு

Delija toro angle divesa teri informaciono nauka

Przedruki z Wczesnych Dni Nauk Informatycznych

طبعات جدردة من الأيام الأولى لعلوم المعلومات

ਓਦ੍ਰਿਠਸ ਦੇ ਠੁਓ ਓਅੁਲੇ ਧਐਸ ਦ ਸਫੀਓਡਓਸ

Retipăriri ain zilele ae început ale științelor informației

אוסף פרסומים משחר תורת המידע

English, Serbian Latin, Serbian Cyrillic, Finnish, Russian, Japanese, Armenian, German, Castilian, Georgian, Hungarian, Bask, Spanish, Estonian, Sami, Tamil (two lines), Balkan Romani, Polish, Arabic, Punjabi, Romanian, Hebrew.

ESIPUHE

Monet tieteelliset edistysaskeleet tai läpimurrot ovat syntyneet yhdistämällä kahden eri tutkimusalueen metodeja ja ehkäpä luonnollisin tapa tehdä tällainen tutkimusalueiden välinen harppaus on se, että tutkija siirtyy joko kokonaan tai osittain toimimaan uudella tutkimusalueella. Kaikki tutkijat tekevät tätä jossain määrin siirtyessään uusiin tutkimusongelmiin ja jotkut jopa toimivat useilla näennäisesti hyvin erilaisilla aloilla. Yleensä näitä aloja kuitenkin yhdistää vaikkapa se, että probleemien mallinnus perustuu samaan matematiikan alaan. Näin on myös Aimo Tietäväisen tutkimuksissa. Hän aloitti matematiikan opinnot Turun yliopistossa 1950 luvun puolivälissä ja väitteli yhtälöiden ratkeavuudesta äärellisissä kunnissa v. 1965.

Voi sanoa, että ensimmäiset kymmenen tutkimustyö vuotta kuluivat lukuteorian ja nimenomaan äärellisten kuntien teorian parissa. Samaan aikaan eli tietoliikenne ja informaatioteoriaan kuuluva virheitä korjaavien koodien teoria kiivasta kehitysvaihetta ja monilla koodausteorian ongelmilla oli yhteys ääraIlisten kuntien teoriaan. Näin oli myös kysymyksellä täydellisten koodien olemassaolosta. Sitä pidettiin merkittävänä ongelmana ja otaksuttiin, että tuntemattomia täydellisiä koodeja ei ole, mutta oli onnistuttu todistamaan vain muutamia erikoistapauksia. Vuonna 1971 Tietäväinen osoitti, että ei ole tuntemattomia binäärisiä täydellisiä koodeja ja vuonna 1973 lehdessä *SIAM J. Appl. Math.* ilmestyneessä artikkelissa hän todisti, että ei ole tuntemattomia täydellisiä koodeja yli minkään äärellisen kunnan. Tulos herätti suurta huomiota ja sitä pidetään yhtenä koodausteorian tutkimuksen kulmakivistä. Myöhemmin Tietäväinen on julkaissut monia muita syvällisiä tuloksia koodausteorian ja lukuteorian alalta.

Tieteellisten lehtien määrä ja levikki oli tuohon aikaan paljon vähäisempi ja tutkijat julkaisivat tuloksiaan pääosin kansallisissa ja yliopistojensa sarjoissa. Tulokset toki levisivät nopeastikin suoraan tutkijoiden keskinäisen kirjeenvaihdon kautta sekä referaattilehtien kautta, mutta vain saman alan tutkijoiden piirissä. Näin esimerkiksi Tietäväisen kansallisissa sarjoissa julkaistut varhaisemmat tulokset ovat vaikeasti saatavissa. Tähän reprint kokoelmaan onkin koottu hänen varhaisia julkaisujaan lukuteorian alalta sekä myös ensimmäiset koodausteorian työt.

Itse tutustuin Professori Tietäväiseen, kun aloitin matematiikan opinnot syksyllä 1968 ja seurasin matemaattisen analyysin peruskurssia, jota hän luennoi. Hän oli tavattoman pidetty ja kunnioitettu opettaja. Hänen luentonsa olivat erittäin hyvin suunniteltuja ja valmisteltuja, ja koko sali seurasi aina luentoa herkeämättä. Luennot oli myös höystetty pienillä annoksilla mainiota kuivaa huumoria sekä hyödyllisiä opiskelua koskevia neuvoja ja ohjeita. Opiskelijoiden kunnioitusta kuvastaa sekin, että kun hän oli ollut vuoden poissa Turusta, niin hänen tullessa luentosaliin ensimmäiselle luennolle opiskelijat nousivat spontaanisti seisomaan vaikka ajat olivat jo muuttuneet eikä tällainen enää ollut tapana.

Jatkoin opintoja hänen ohjauksessaan. Väitöskirjatyön ohjaajana hän oli vaativa, mutta kannusti itsenäiseen ajatteluun ja oli aina valmis keskusteluihin ja tarjoamaan ideoita kun opiskelijalla oli vaikeuksia päästä eteenpäin.

Tietäväinen perusti Turkuun koodausteorian koulukunnan ja hänet tunnetaan ehkä parhaiten täydellisten koodien probleeman ratkaisusta. Kysyin kerran häneltä mikä hänen toisen pääalueensa tulos on hänen omasta mieles tään mielenkiintoinen tai jännittävä. Hän sanoi, että esimerkiksi J. H. H. Chalkin Vinogradov-Mordell-Tietäväinen epäyhtälöiksi kutsumat tulokset ovat yllättäviä ja niiden todistukset ovat lyhyitä.

Jaakko Astola

Foreword

Many scientific advances or breakthroughs have been obtained by combining methods from two or more different research fields and perhaps the most natural way to make this kind of leap over the barriers between fields is that the researcher moves partially or completely to a new field. Actually, all researchers do this to some extent when they move to new problems, but some work simultaneously in seemingly very different fields. Typically, there is some underlying theme joining the fields, for instance, the field of mathematics that is used in modeling the problems. This is the case also in the research work of Professor Aimo Tietäväinen, where the underlying theme is number theory that is also called the *Queen of Mathematics.* He began his career by studying mathematics at University of Turku in middle 1950's and wrote his PhD thesis on the solvability equations over finite fields in 1965. His first ten years in research concentrated on number theory and especially on the theory of finite fields. At the same time the research in coding theory, a part of information theory, was rapidly expanding and many problems in coding theory are connected to finite fields. This is also the case with the famous conjecture of nonexistence of perfect codes over finite fields of which only isolated cases had been proved. In 1971, Tietäväinen proved that there are no unknown perfect codes over the binary field, which is the most important case. Soon after that, in an article that appeared in *SIAM J. Appl. Math.* in 1973, he proved the celebrated result that there are no unknown perfect codes over any finite fields, which is considered a cornerstone of the theory of error correcting codes.

At that time scientific and mathematical journals did not have as wide circulation as nowadays and many researchers published mainly in national publication series or in the publication series of their local university. The results spread quite quickly, nevertheless, via the correspondence between researchers, but mostly only between the researchers within the same field. Thus, for instance, the early works of Tietäväinen that were published in national series are not easy to get and, consequently, we have collected to this volume of reprints his early works on number theory as well as his key papers on perfect codes that reflect his movement to the new research field of coding theory.

I first met professor Tietäväinen when I began my studies of mathematics in fall 1968 by following the first course in mathematical analysis that he lectured. He was a highly liked and respected teacher. His lectures were superbly designed and prepared, and always the full room was keenly following the lecture.

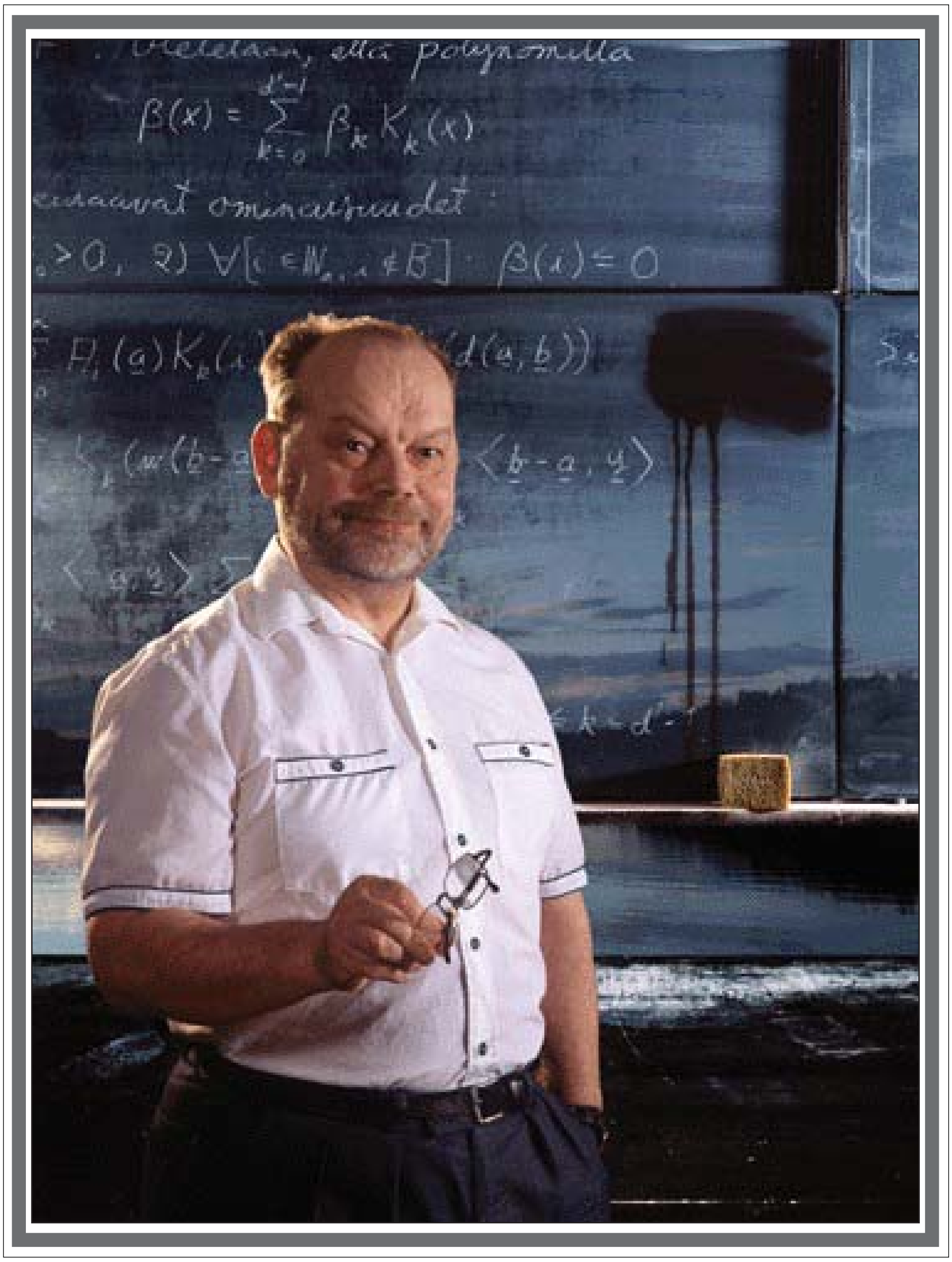
Mathematics Department had moved to a new building that had good lecture rooms with several pairs of blackboards that could be moved up and down. He started from the left upper board and usually at the end of the second hour of the lecture he put period in the right lower corner of the last board. He always managed to keep the theorems and key points of the proofs visible throughout the lecture so that a student could catch up with the reasoning of difficult parts. His lectures were supplemented with small doses of excellent dry humor as well as very useful hints and guidelines for studies.

I remember something that well describes the respect that the students held for him. When he returned to Turku after spending a year in another university and came to his first lecture, all the students spontaneously rose although this had not been the custom for a decade or so.

I continued my studies towards PhD under his guidance. As an advisor he was demanding but encouraged independent thinking, and he was also always ready to discuss and offer ideas when a student was stuck with a problem.

Tietäväinen founded Turku group of coding theory that is well known around the world. Among coding theorists he is best known for his proof of the nonexistence of perfect codes. I once asked him about his other field, number theory, what he considers to be the result he has found most pleasing or interesting. He answered that perhaps the results that J. H. H. Chalk calls Vinogradov-Mordell-Tietäväinen inequalities are such because the results are surprising while their proofs are short.

Jaakko Astola


Oletetaan, että polynomilla
$\beta(x) = \sum_{k=0}^{d'-1} \beta_k K_k(x)$
seuraavat ominaisuudet:
$>0,\ 2)\ \forall[i \in W_{a,i} \notin B] \cdot \beta(i) \leq 0$

A digital portrait of Aimo Tietäväinen by the artist Juhani Haaparinne reprinted by the courtesy of its author.

Biographical Data of Aimo Tietäväinen

Date and place of birth
July 6, 1937, Suistamo, Finland

Education
Doctor of Philosophy, University of Turku, 1965

Employment
Assistant in Mathematics, University of Turku, 1962-67
Associate Professor of Mathematics, Tampere University of Technology, 1967-68
University of Turku, 1968-71
Professor of Mathematics, Tampere University of Technology, 1971-72
University of Turku, 1972-

Scientific Activities
IEEE Transactions on Information Theory, Associate Editor 1990-93
Designs, Codes and Cryptography, Member of the Editorial Board 1990-
Journal of Discrete Mathematical Systems and Cryptography, Associate Editor 1998-
Institute of Combinatorics and Its Applications (ICA), Founding Fellow 1990-
Member of the Finnish Academy of Sciences 1978-
Supervisor of 10 Ph.D. Students

Referee of several selection committees for professorships and lecturerships in mathematics or computer science
(e.g., Technical University of Tampere (five times),
Universities of Oulu, York, Syracuse (USA), Illinois,
Tel-Aviv (twice), Puerto Rico and Bilkent)

Member of the organizing, program or advisory committees of several international symposiums; most recently
IEEE Information Theory Symposiums in Kobe (1988),
San Antonio (1993) and Ulm (1997)
International Conference on Sequences and Their Applications - SETA'98, Singapore (1998),
Developments in Language Theory in Turku (1993)

Occasional referee of several scientific journals
American Mathematical Monthly,
Discrete Mathematics,
IEEE Transactions on Information Theory,
Information and Control (new name *Information and Computation*),
Proceedings of the American Mathematical Society,
Archiv der Mathematik,
SIAM Journal on Discrete Mathematics,
Reviewer of several international and Academy of Finland research projects

Administrative Duties
Head of the Mathematics Department, University of Turku, 1973
Dean of the Faculty of Mathematics and Natural Sciences, University of Turku, 1975-77
Chairman of the Institute of Mathematical Sciences,
University of Turku, 1985-86
The responsible person of the Mathematics Department, University of Turku, until 1998.
For several years member of the senate, faculty councils,
institute councils and several committees,
Tampere University of Technology and/or University of Turku,
Member of the Natural Science Research Council of
the Academy of Finland, 1992-94

Honours
Knight 1st class of the Order the White Rose of Finland, 1993
The special issue of *Applicable Algebra in Engineering,*
Communication and Computing (1997) with articles dedicated to
Aimo Tietäväinen on the occasion of his 60th birthday, 347-435.
"The very knowledge of coding",
Studies in honor of Aimo Tietäväinen on the occasion of his
fiftieth birthday, 156 pages.

Conferences and scientific visits since 1990

"Applications of Algebraic Geometry", Puerto Rico, 1990, (invited lecture, session chairman)

IEEE Information Theory Symposium, San Diego, 1990 (lecture)

Royal Institute of Technology, Stockholm, 1990 (lecture)

"Codes and Designs", Oberwolfach, 1990 (invited lecture)

"The Fourth Nordic Discrete Mathematics Symposium",
Fredrikshavn, 1990 (lecture)

University of Linköping, 1990 (lecture)

"Algebraic and Combinatorial Coding Theory",
Leningrad, 1990 (invited lecture, session chairman),

IEEE Information Theory Symposium, Budapest, 1991 (the editors' meeting)

"Algebraic Coding Theory", Paris, 1991 (lecture, session chairman)

"Information Theory", Oberwolfach, 1992 (invited lecture, session chairman)

"Discrete Mathematics and Its Applications", Veldhoven, 1992 (invited lecture)

"Coding Theory", Bergen, 1992 (two lectures)

University of Tel-Aviv, 1992 (lecture)

Technion of Haifa, 1992 (lecture)

IEEE Information Theory Symposium, San Antonio 1993, (session chairman, the editors' meeting)

"Codes and Character Sums over Finite Fields", Puerto Rico, 1993 (two lectures)

"Discrete Metric Spaces", Bielefeld, 1994 (invited lecture)

Coding theory special session of the American Mathematical Society
Chicago meeting, 1995 (invited lecture)

"Mediterranean Workshop of Coding and Information Integrity", Mallorca, 1996 (lecture, session chairman)

"Number Theory and Its Applications", Bilkent, 1996, (two invited lectures)

University of Szeged, 1997 (lecture)

IEEE Information Theory Symposium, Ulm, 1997, (lecture given by P. Charpin)

"Designs and Codes", Oberwolfach, 1998

"Numbers, Information and Complexity", Bielefeld, 1998, (invited lecture; session chairman)

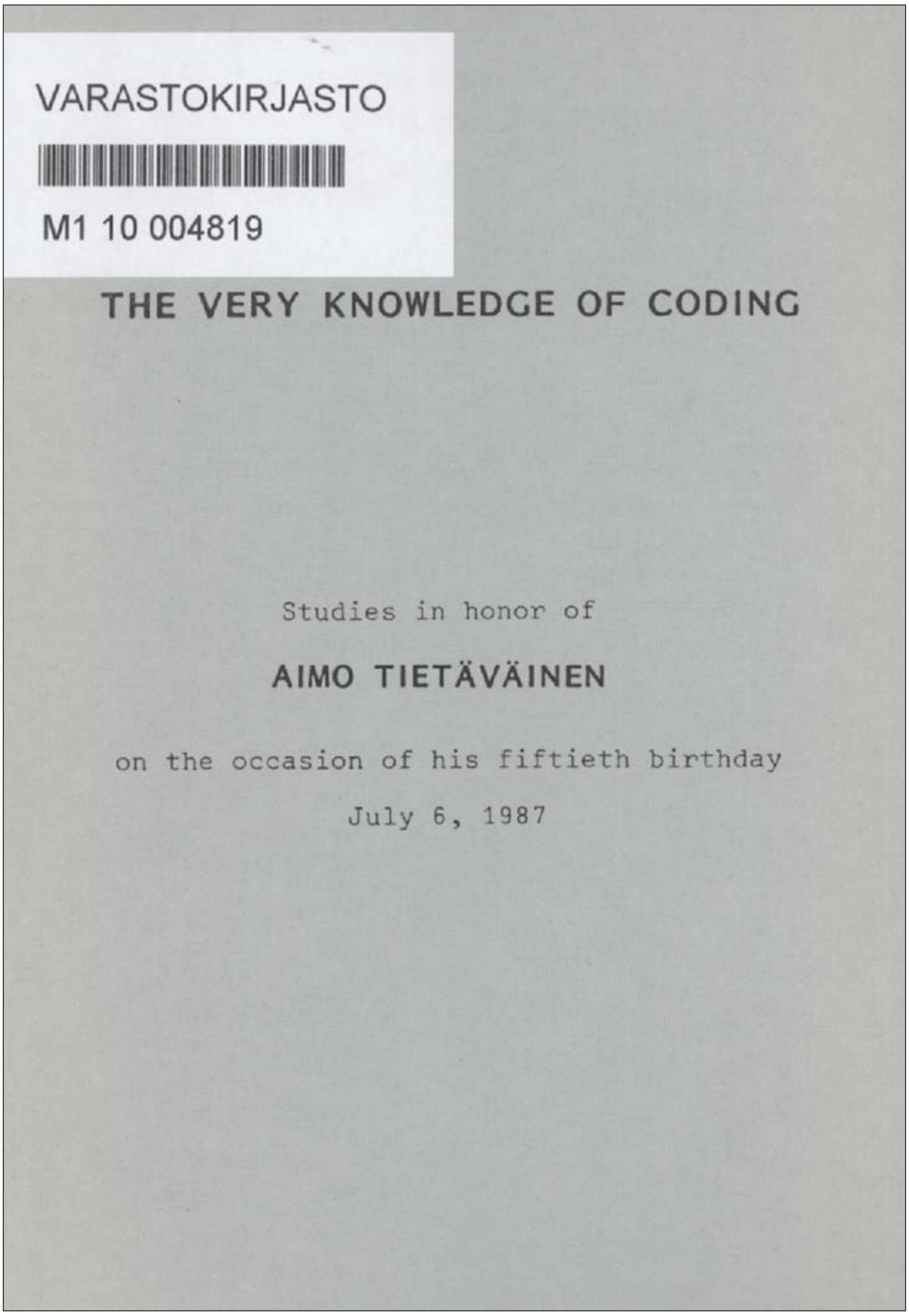
Table 1 lists the PhD students supervised by Professor Aimo Tietäväinen at the University of Turku, Turku, Finland, and the year of the defence.

Table 1: Former PhD students of Professor Aimo Tietäväinen.

Kauko Lindström	1975
Jaakko Astola	1978
Hannu Laakso	1979
Matti Aaltonen	1981
Antti Perttula	1982
Hannu Tarnanen	1982
Sirpa Ernvall	1983
Iiro Honkala	1989
Yrjö Kaipainen	1995
Tero Laihonen	1998

The cover, title-page, contents, and preface of the book
”The very knowledge of coding”,
Studies in honor of Aimo Tietäväinen on the occasion of his
fiftieth birthday.

The cover of The special issue of
Applicable Algebra in Engineering,
Communication and Computing (1997) with articles dedicated to
Aimo Tietäväinen on the occasion of his 60th birthday, 347-435.



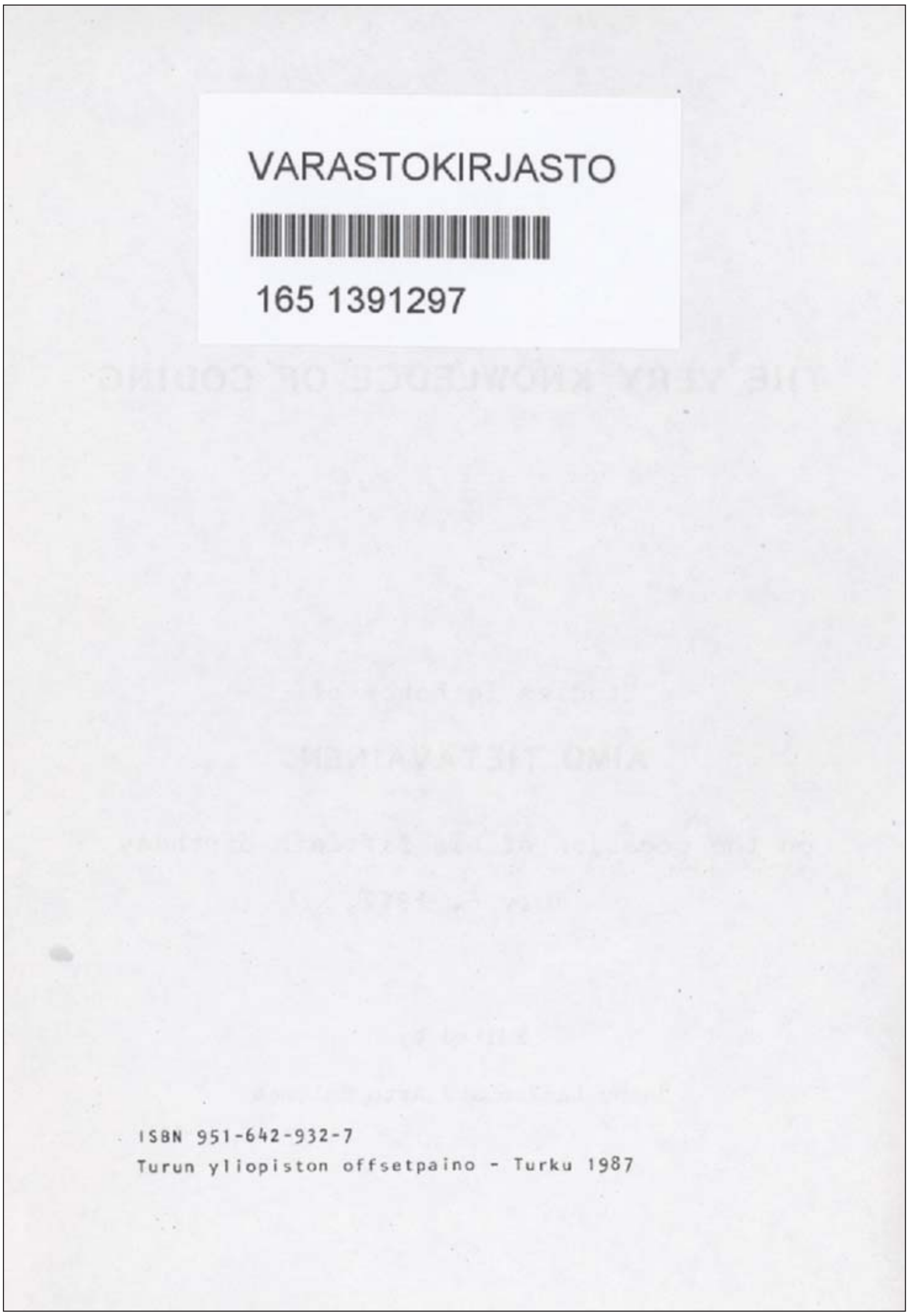
THE VERY KNOWLEDGE OF CODING

Studies in honor of

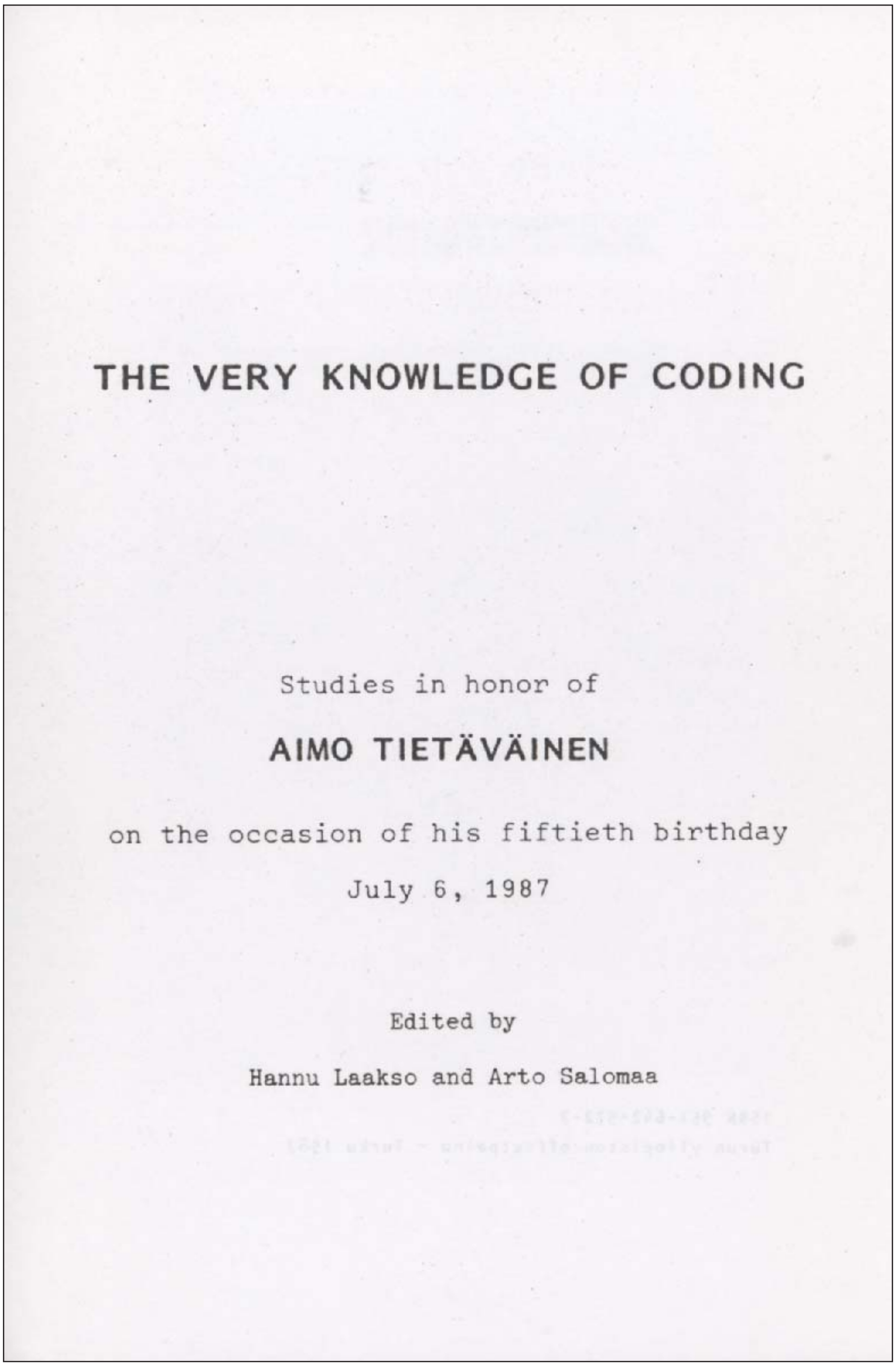
AIMO TIETÄVÄINEN

on the occasion of his fiftieth birthday

July 6, 1987



ISBN 951-642-932-7
Turun yliopiston offsetpaino - Turku 1987



THE VERY KNOWLEDGE OF CODING

Studies in honor of

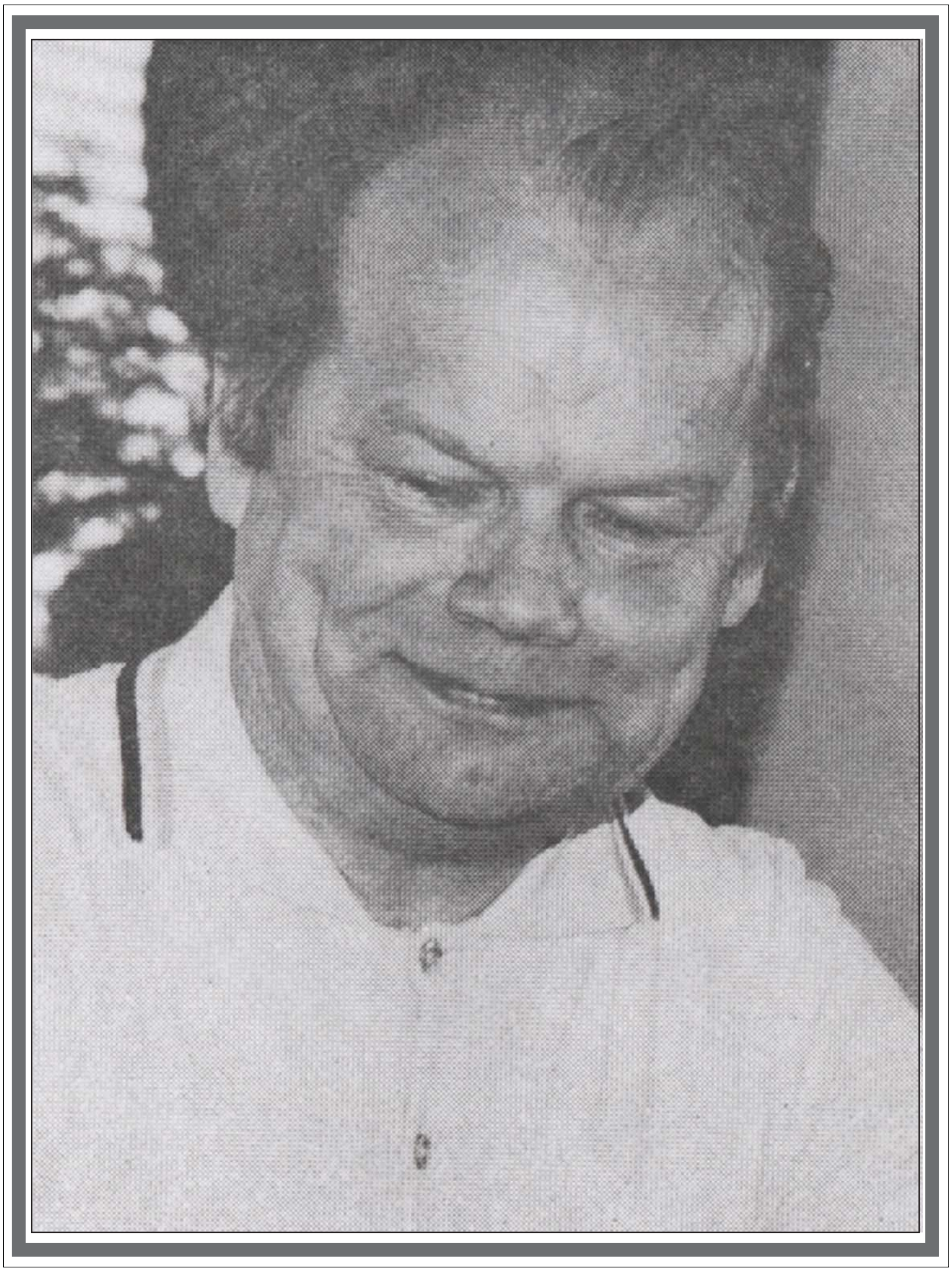
AIMO TIETÄVÄINEN

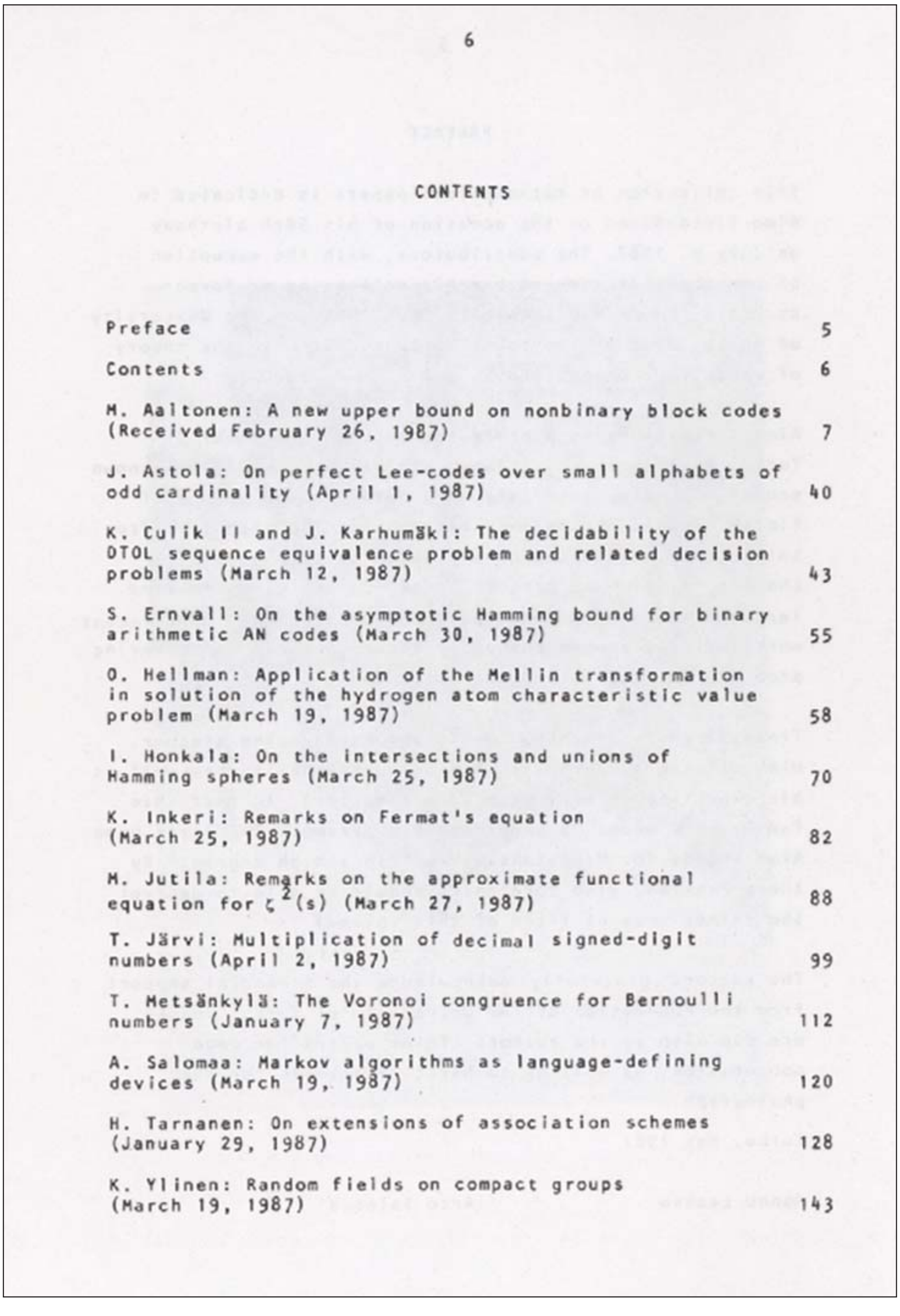
on the occasion of his fiftieth birthday

July 6, 1987

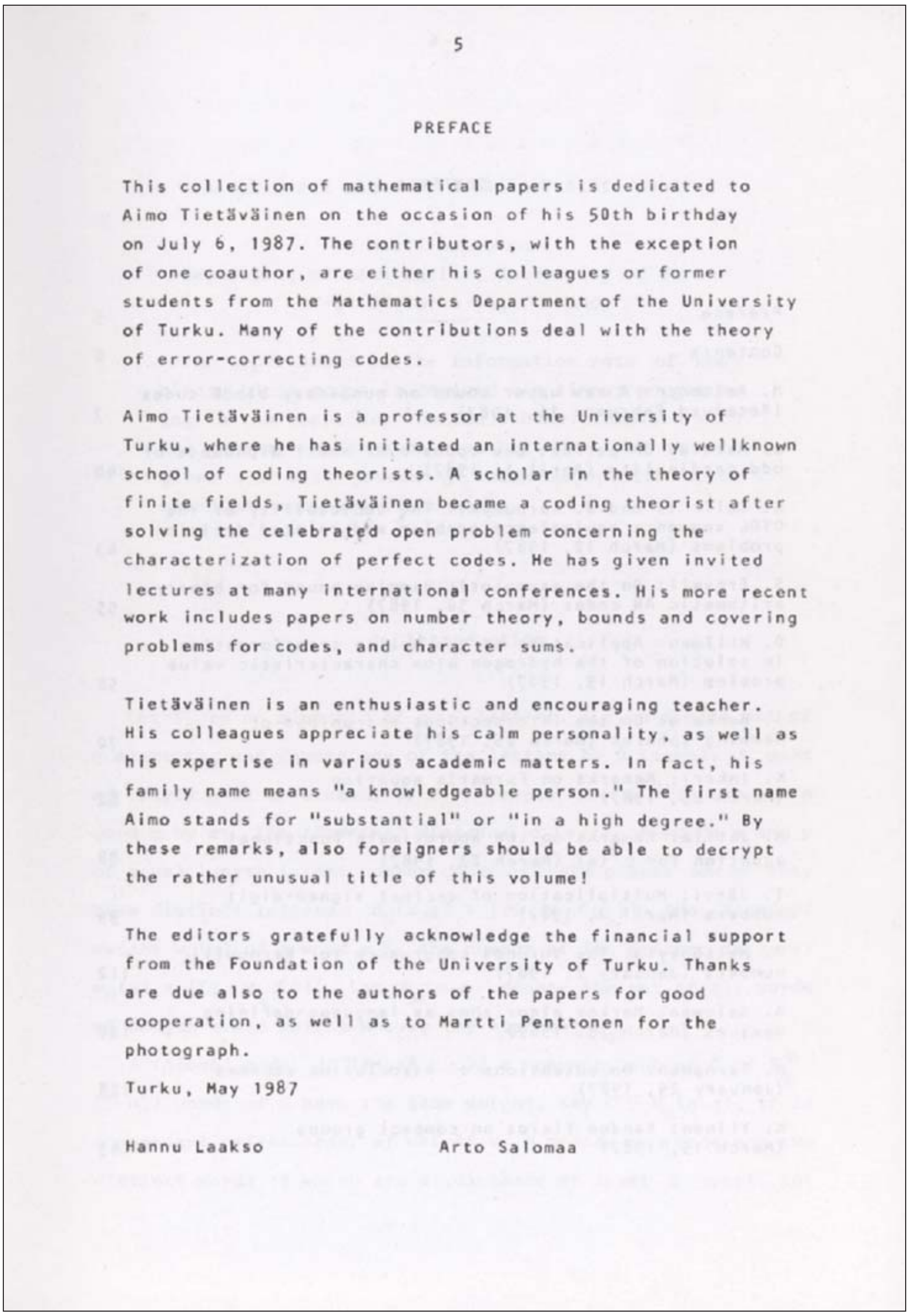
Edited by

Hannu Laakso and Arto Salomaa





CONTENTS



PREFACE

This collection of mathematical papers is dedicated to Aimo Tietäväinen on the occasion of his 50th birthday on July 6, 1987. The contributors, with the exception of one coauthor, are either his colleagues or former students from the Mathematics Department of the University of Turku. Many of the contributions deal with the theory of error-correcting codes.

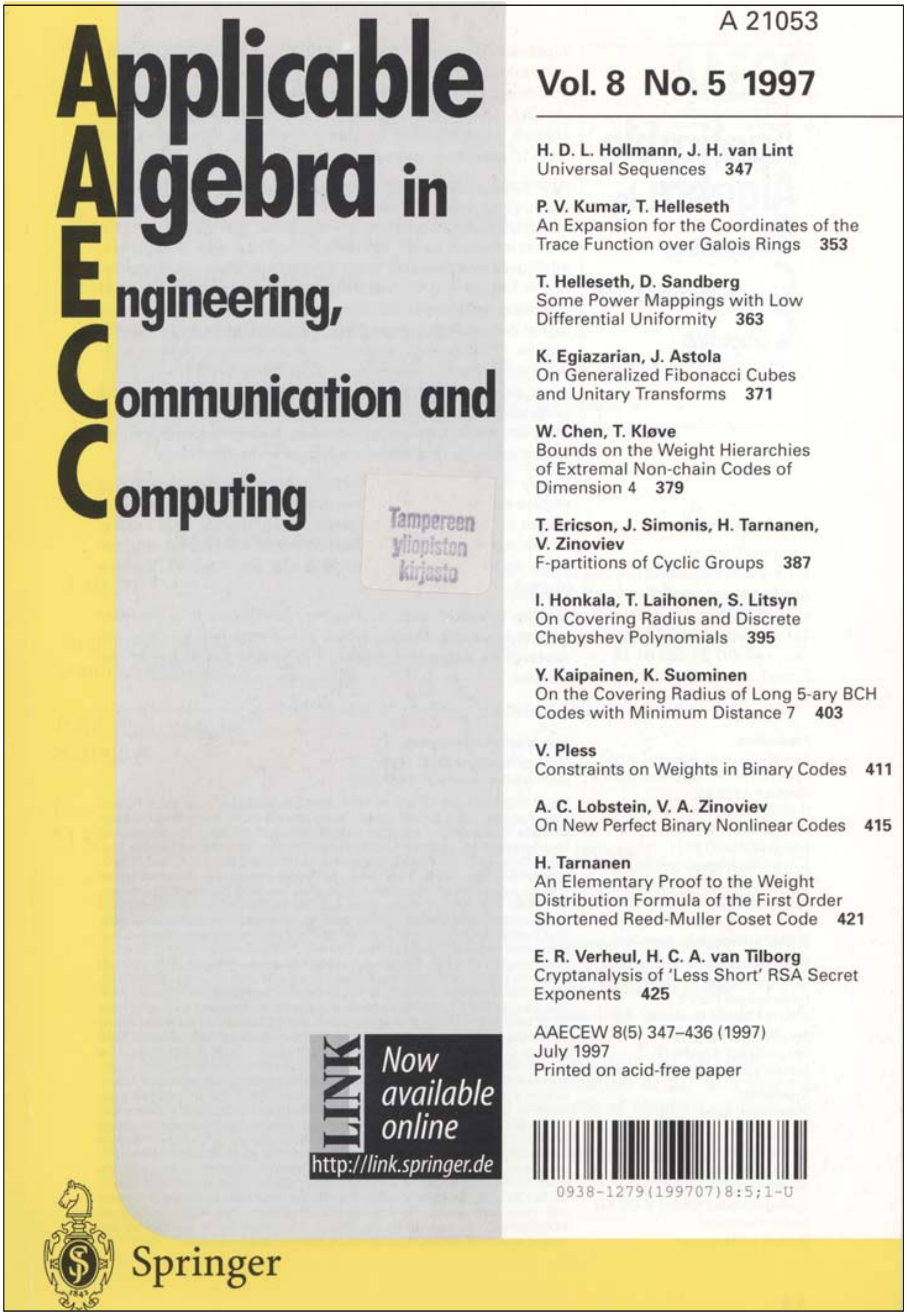
Aimo Tietäväinen is a professor at the University of Turku, where he has initiated an internationally wellknown school of coding theorists. A scholar in the theory of finite fields, Tietäväinen became a coding theorist after solving the celebrated open problem concerning the characterization of perfect codes. He has given invited lectures at many international conferences. His more recent work includes papers on number theory, bounds and covering problems for codes, and character sums.

Tietäväinen is an enthusiastic and encouraging teacher. His colleagues appreciate his calm personality, as well as his expertise in various academic matters. In fact, his family name means "a knowledgeable person." The first name Aimo stands for "substantial" or "in a high degree." By these remarks, also foreigners should be able to decrypt the rather unusual title of this volume!

The editors gratefully acknowledge the financial support from the Foundation of the University of Turku. Thanks are due also to the authors of the papers for good cooperation, as well as to Martti Penttonen for the photograph.

Turku, May 1987

Hannu Laakso Arto Salomaa



Applicable Algebra in Engineering, Communication and Computing

Vol. 8 No. 5 1997

AAECEW 8(5) 347–436 (1997)
July 1997
Printed on acid-free paper

0938-1279(199707)8:5;1-U

Springer

1 Reprints

The following table lists journals where Professor Aimo Tietäväinen used to publish in 1964 to 1974. The numbers in parentheses show the number of publications in a particular year.

Journal	Year of publishing
Ann. Univ. Turku.	1964, 1966 (2), 1967 (2), 1968 (2), 1969, 1971, 1974
Ann. Acad. Sci. Fenn.	1965 (2), 1966, 1970, 1973
J. Number Theory	1971, 1975
SIAM J. Appl. Math.	1973

Aimo Tietäväinen defended his PhD thesis in 1965 at the University of Turku, Turku, Finland, with the subject *On the Non-Trivial Solvability of Some Equations and Systems of Equations in Finite Fields* and under supervision of Professor Kustaa Inkeri. The thesis was published as the Paper 2 in the enumeration of papers reprinted in this book.

1. **Paper 1**
   "On the non-trivial solvability of some systems of equations in finite fields", *Ann. Univ. Turku.*, A I 71 (1964) 1-5.

2. **Paper 2**
   "On the non-trivial solvability of some equations and systems of equations in finite fields", *Ann. Acad. Sci. Fenn.*, A I 360 (1965) 1-38.

3. **Paper 3**
   "On systems of linear and quadratic equations in finite fields", *Ann. Acad. Sci. Fenn.*, A I 382 (1965) 1-5.

4. **Paper 4**
   "On systems of equations in finite fields", *Ann. Acad. Sci. Fenn.*, A I 386 (1966) 1-10.

5. **Paper 5**
   "On the trace of a polynomial over a finite field", *Ann. Univ. Turku.*, A I 87 (1966) 1-7.

6. **Paper 6**
   "On non-residues of a polynomial", *Ann. Univ. Turku.*, A I 94 (1966) 1-6.

7. **Paper 7**
   "On the solvability of equations in incomplete finite fields", *Ann. Univ. Turku.*, A I 102 (1967) 1-13.

8. **Paper 8**
   "On pairs of additive equations", *Ann. Univ. Turku.*, A I 112 (1967) 1-7.

9. **Paper 9**
   "On diagonal forms over finite fields", *Ann. Univ. Turku.*, A I 118 (1968) 1-10.

10. **Paper 10**
    "On the distribution of the residues of a polynomial", *Ann. Univ. Turku.*, A I 120 (1968) 1-4.

11. **Paper 11**
    "On a homogeneous congruence of odd degree", *Ann. Univ. Turku.*, A I 131 (1969) 1-6.

12. **Paper 12**
    "On the nonexistence of perfect 4-Hamming-error-correcting codes", *Ann. Acad. Sci. Fenn.*, A I 485 (1970) 1-6.

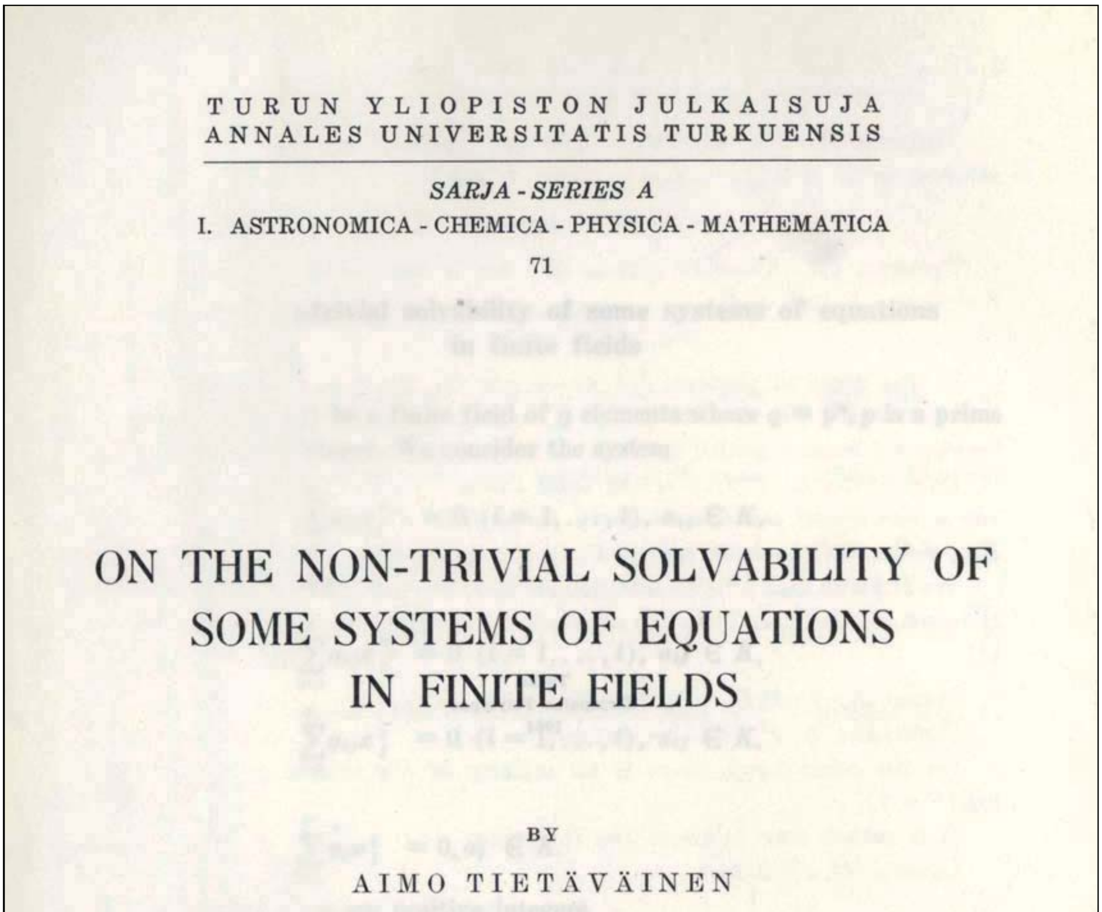
13. **Paper 13**
    "On a problem of Chowla and Shimura", *J. Number Theory*, 3 (1971) 247-252.

14. **Paper 14**
    (with A. Perko) "There are no unknown perfect binary codes", *Ann. Univ. Turku.*, A I 148 (1971) 1-10.

15. **Paper 15**
    "Note on Waring's problem (mod p)", *Ann. Acad. Sci. Fenn.*, A I 554 (1973) 1-7.

16. **Paper 16**
    "On the nonexistence of perfect codes over finite fields", *SIAM J. Appl. Math.*, 24 (1973) 88-96.

17. **Paper 17**
    "A short proof for the nonexistence of unknown perfect codes over $GF(q)$, $q > 2$", *Ann. Acad. Sci. Fenn.*, A I 580 (1974) 1-5.



TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SERIES A
I. ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA
71

ON THE NON-TRIVIAL SOLVABILITY OF SOME SYSTEMS OF EQUATIONS IN FINITE FIELDS

BY

AIMO TIETÄVÄINEN

Turku 1964
TURUN YLIOPISTO

Turku
Kirjapaino Polytypos
1964

On the non-trivial solvability of some systems of equations in finite fields

Let $K = GF(q)$ be a finite field of q elements where $q = p^n$, p is a prime and n a positive integer. We consider the system

$$(1) \qquad \sum_{j=1}^{s} a_{ij} x_j^{c_{ij}} = 0 \ (i = 1, \ldots, t), \ a_{ij} \in K,$$

and its special cases

$$(2) \qquad \sum_{j=1}^{s} a_{ij} x_j^{c_i} = 0 \ (i = 1, \ldots, t), \ a_{ij} \in K,$$

$$(3) \qquad \sum_{j=1}^{s} a_{ij} x_j^{c} = 0 \ (i = 1, \ldots, t), \ a_{ij} \in K,$$

and

$$(4) \qquad \sum_{j=1}^{s} a_j x_j^{c} = 0, \ a_j \in K.$$

Here the c-numbers are positive integers.

Definition 1. *The system (1) is correlated if the following condition is valid: If i_1 and i_2 are two of the integers $1, \ldots, t$, then either $c_{i_1 j} = c_{i_2 j}$, for every $j = 1, \ldots, s$, or $c_{i_1 j} \neq c_{i_2 j}$, for every $j = 1, \ldots, s$. The system (1) is strongly correlated if the a_{ij}'s are non-zero elements of K and $c_{i_1 j} \neq c_{i_2 j}$ when $i_1 \neq i_2$, for every $j = 1, \ldots, s$.*

Definition 2. *The system (1) is Chowla's system (cf. [3], [4]) if, for every $j = 1, \ldots, s$, there is an element h_j of K such that $h_j^{c_{ij}} = -1$, for every $i = 1, \ldots, t$.*

Clearly, the system (2) is correlated. Furthermore, the system (1) is Chowla's system at least in the case where the c_{ij}'s are odd.

By using the well-known result of Carlitz and Uchiyama [1] and an extension of a method of Chowla [4], we can prove the following theorems which are generalizations of some theorems of Chowla ([3], [4], cf. also [2]).

Theorem I. *The strongly correlated Chowla's system (1) has a non-trivial solution in K if $2 \leqq c_j = \max\limits_{(i)} c_{ij} \leqq p-1$, for every $j = 1, \ldots, s$, $\max c_{ij} \geqq 3$ and*

$$2^{s(s-2t)} \geqq \prod_{j=1}^{s} (c_j - 1)^{2t}.$$

THEOREM II. *The correlated Chowla's system* (1) *has a non-trivial solution in K if* $3 \leqq c = \max c_{ij} \leqq p-1$ *and*

$$s \geqq 2t(t + \log_2(c-1)).$$

THEOREM III. *Chowla's system* (3) *has a non-trivial solution in K if* $d = (c,q-1) \geqq 3$ *and*

$$s \geqq 2t(t + \log_2(d-1)).$$

In the proof of theorem III we do not use the deep result of CARLITZ and UCHIYAMA. The conditions $2 \leqq c_j$, $\max c_{ij} \geqq 3$ and $d \geqq 3$ are unessential because the cases excluded by them are easy to handle.

DEFINITION 3. $P(c,t)$ *is the least integer s such that the system* (3) *has a non-trivial solution in every finite field K, for every matrix* (a_{ij}). *Especially, we denote* $P(c,1) = P(c)$.

We suppose that C is an odd integer $\geqq 3$. Furthermore, we denote by $\{\alpha\}$ the least integer $\geqq \alpha$. Then the subsequent corollaries are implied by theorem III.

COROLLARY 1. $P(C,t) \leqq 2t^2 + \{2t \log_2(C-1)\}$.

COROLLARY 2. $P(C) \leqq 2 + \{2 \log_2(C-1)\}$.

On the other hand, there is an infinity of C's such that $P(C) \geqq 1 + \log_2(C+1)$.

It is rather easy to prove the following

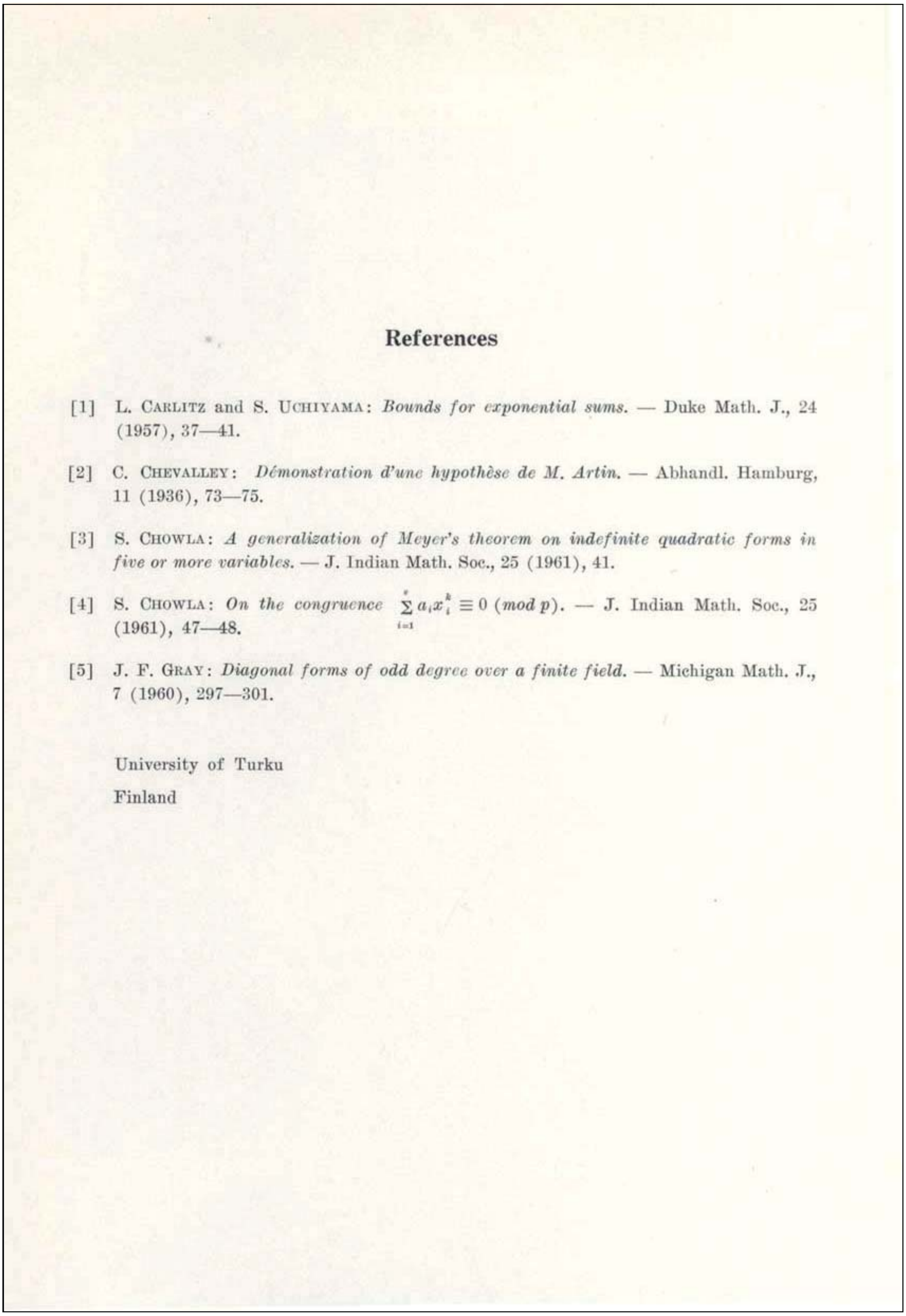
LEMMA. *If* $s \geqq 3$ *and*

$$q \geqq \frac{d(d-1)^{s/s-2}}{s},$$

where $d = (c,q-1)$, *then Chowla's equation* (4) *has a non-trivial solution in K.*

GRAY [5] has shown that $P(5) = 4$. By means of the lemma above, we can show by a very simple numerical calculation that $P(7) = 4, P(9) = 5$. Using the lemma, we obtain also an improvement of corollary 2.

The proofs of the results presented above will be published in the near future.



References

[1] L. Carlitz and S. Uchiyama: *Bounds for exponential sums.* — Duke Math. J., 24 (1957), 37—41.

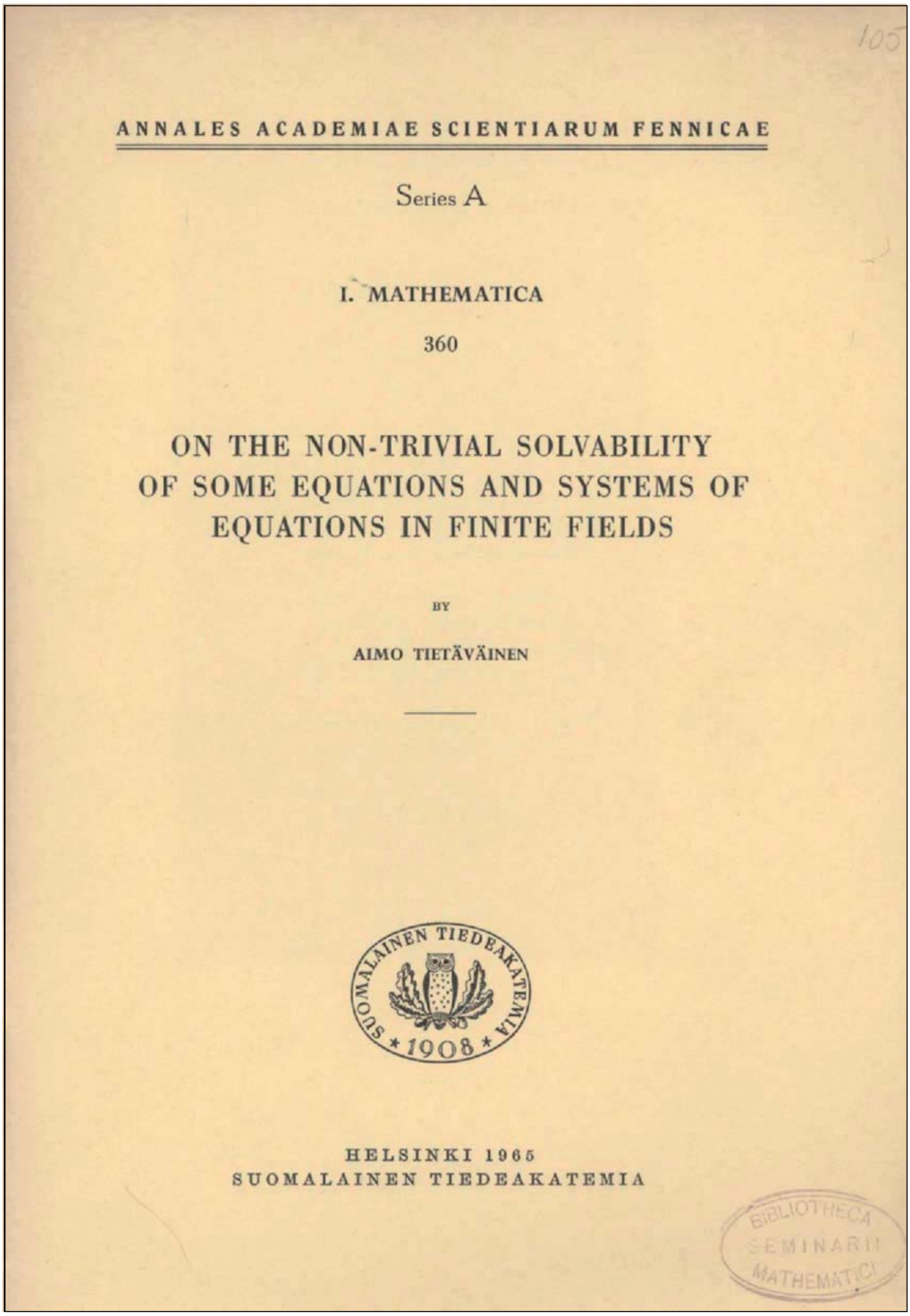
[2] C. Chevalley: *Démonstration d'une hypothèse de M. Artin.* — Abhandl. Hamburg, 11 (1936), 73—75.

[3] S. Chowla: *A generalization of Meyer's theorem on indefinite quadratic forms in five or more variables.* — J. Indian Math. Soc., 25 (1961), 41.

[4] S. Chowla: *On the congruence* $\sum_{i=1}^{s} a_i x_i^k \equiv 0 \ (mod\ p)$. — J. Indian Math. Soc., 25 (1961), 47—48.

[5] J. F. Gray: *Diagonal forms of odd degree over a finite field.* — Michigan Math. J., 7 (1960), 297—301.

University of Turku
Finland



ANNALES ACADEMIAE SCIENTIARUM FENNICAE

Series A

I. MATHEMATICA

360

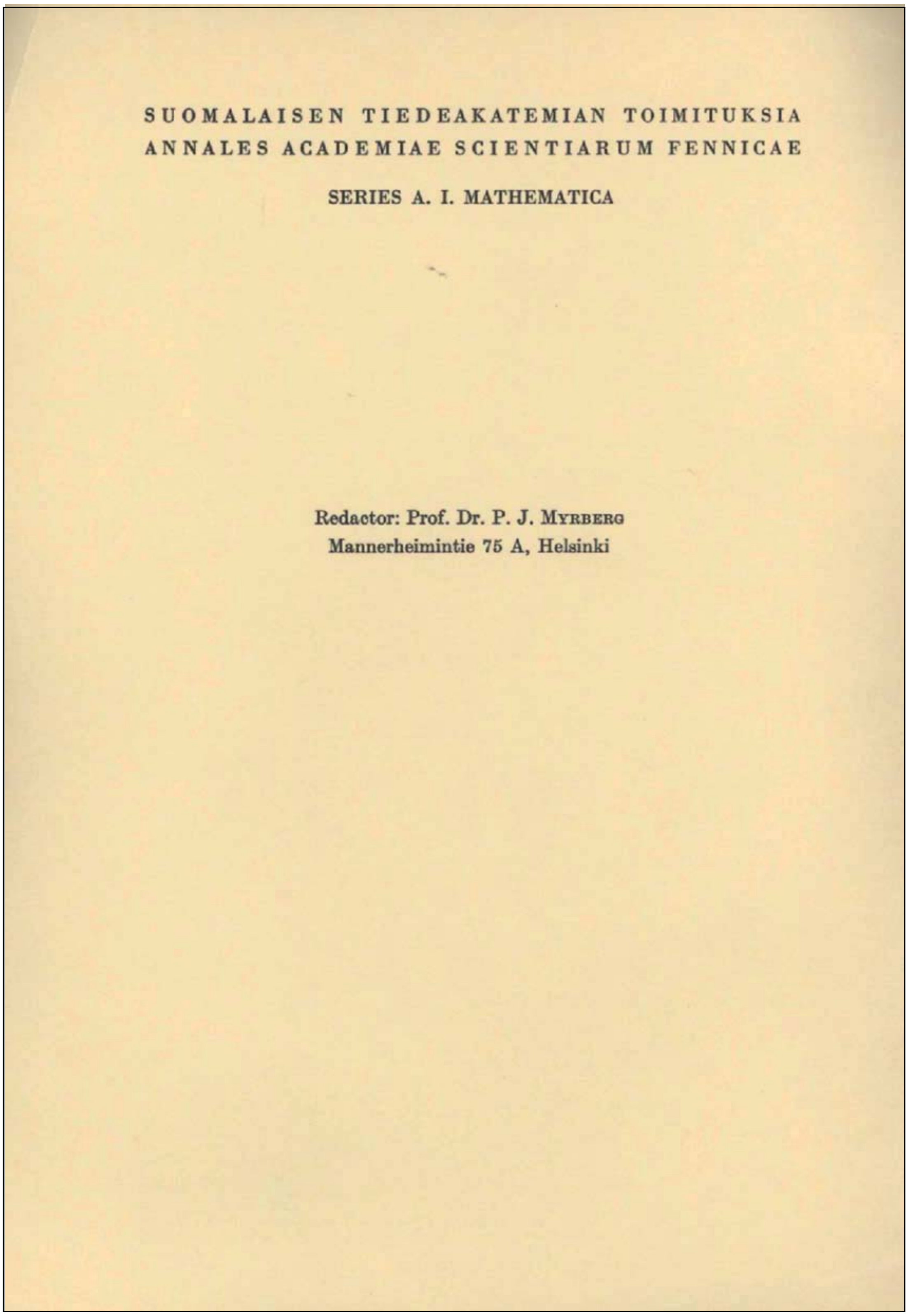
ON THE NON-TRIVIAL SOLVABILITY OF SOME EQUATIONS AND SYSTEMS OF EQUATIONS IN FINITE FIELDS

BY

AIMO TIETÄVÄINEN

SUOMALAINEN TIEDEAKATEMIA *1908*

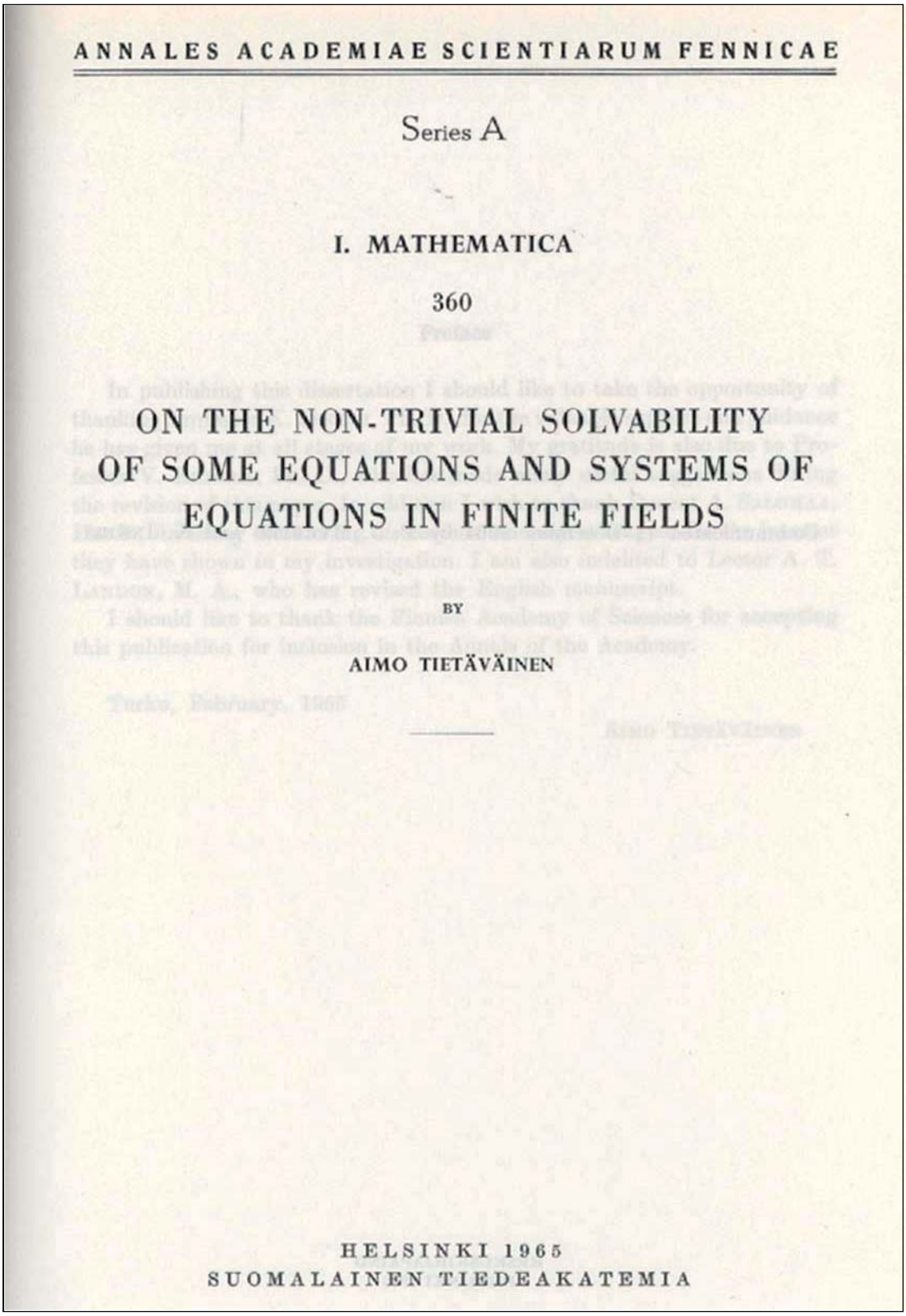
HELSINKI 1965
SUOMALAINEN TIEDEAKATEMIA



SUOMALAISEN TIEDEAKATEMIAN TOIMITUKSIA
ANNALES ACADEMIAE SCIENTIARUM FENNICAE

SERIES A. I. MATHEMATICA

Redactor: Prof. Dr. P. J. Myrberg
Mannerheimintie 75 A, Helsinki



ANNALES ACADEMIAE SCIENTIARUM FENNICAE

Series A

I. MATHEMATICA

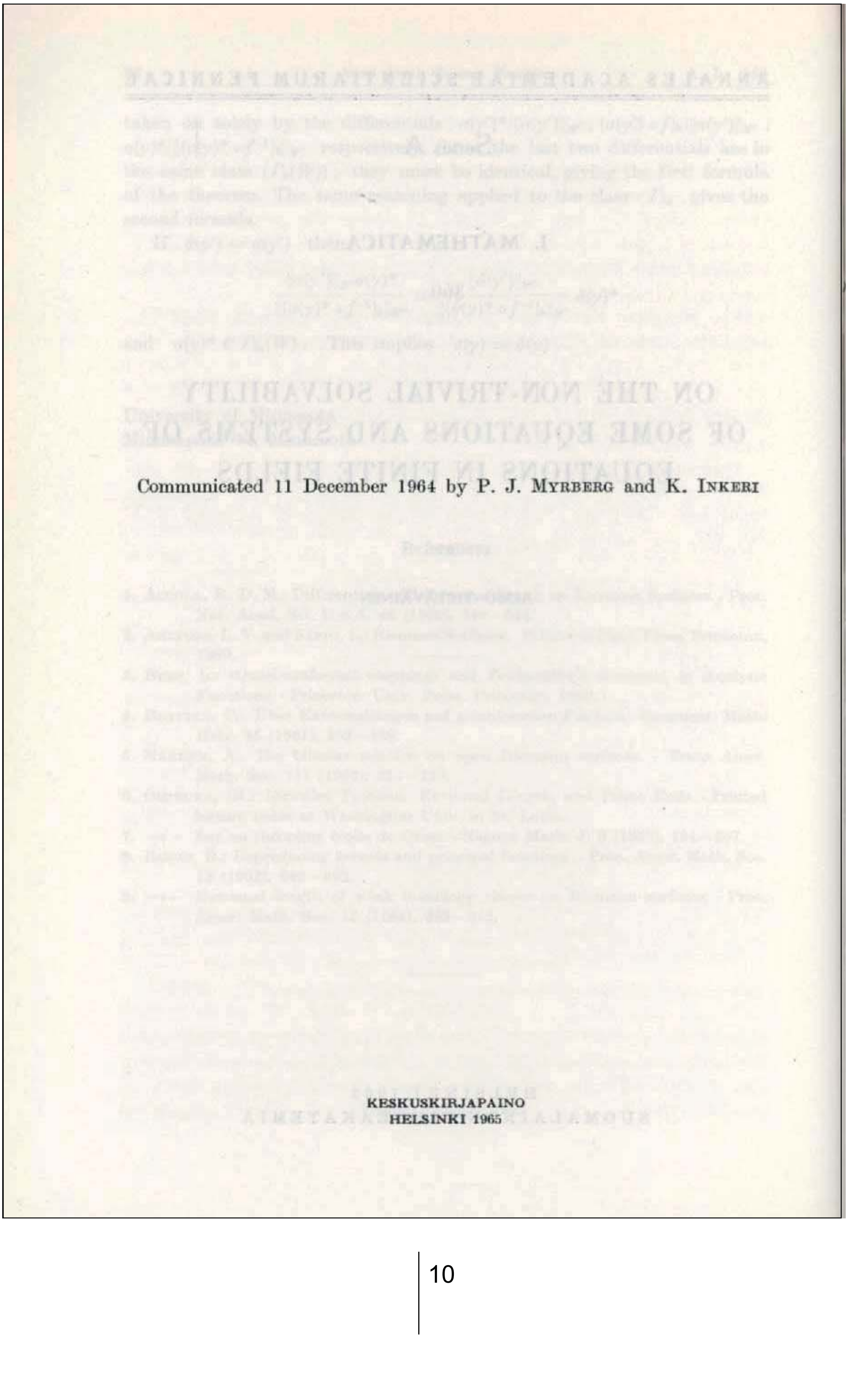
360

ON THE NON-TRIVIAL SOLVABILITY OF SOME EQUATIONS AND SYSTEMS OF EQUATIONS IN FINITE FIELDS

BY

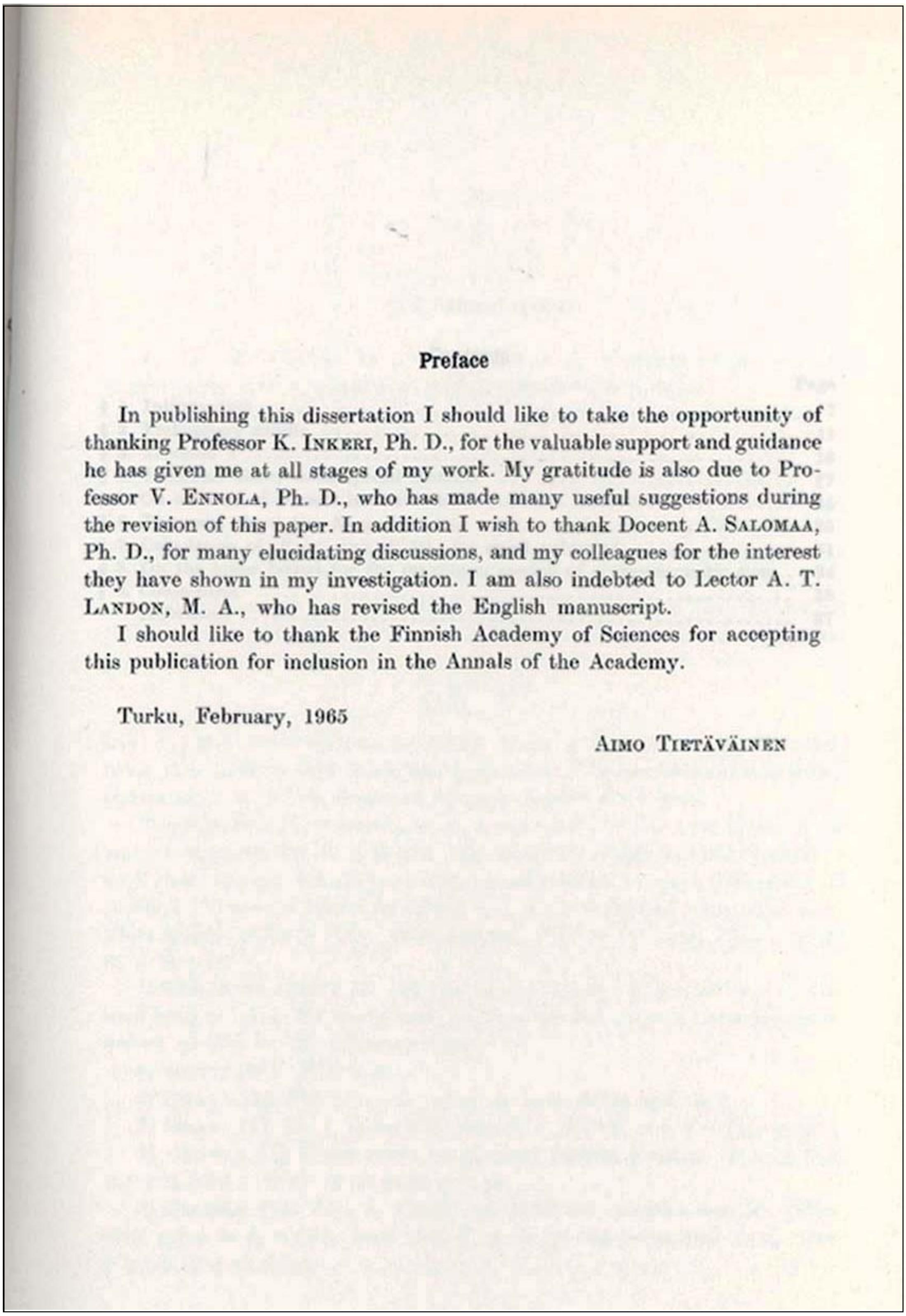
AIMO TIETÄVÄINEN

HELSINKI 1965
SUOMALAINEN TIEDEAKATEMIA



Communicated 11 December 1964 by P. J. MYRBERG and K. INKERI

KESKUSKIRJAPAINO
HELSINKI 1965



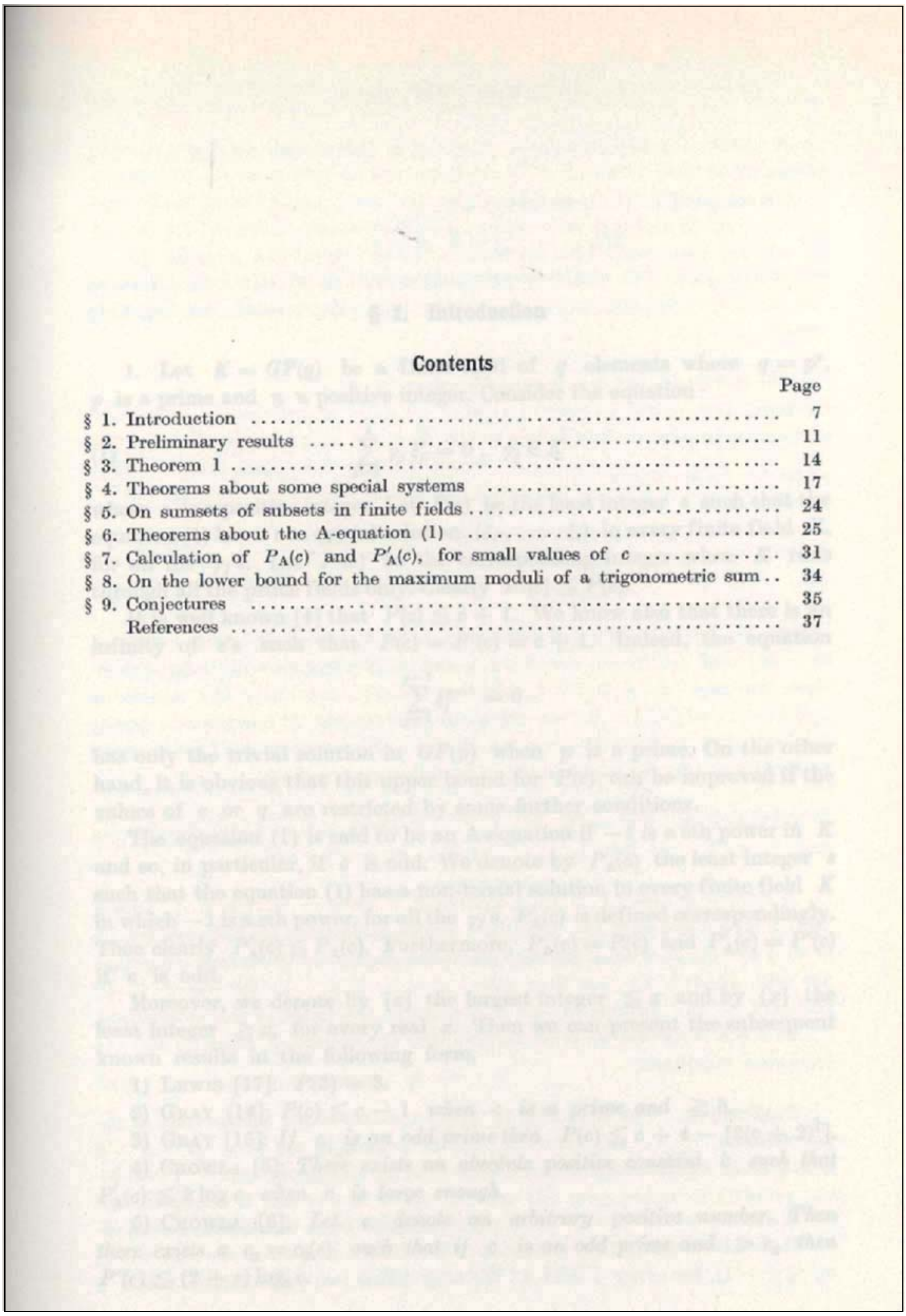
Preface

In publishing this dissertation I should like to take the opportunity of thanking Professor K. INKERI, Ph. D., for the valuable support and guidance he has given me at all stages of my work. My gratitude is also due to Professor V. ENNOLA, Ph. D., who has made many useful suggestions during the revision of this paper. In addition I wish to thank Docent A. SALOMAA, Ph. D., for many elucidating discussions, and my colleagues for the interest they have shown in my investigation. I am also indebted to Lector A. T. LANDON, M. A., who has revised the English manuscript.

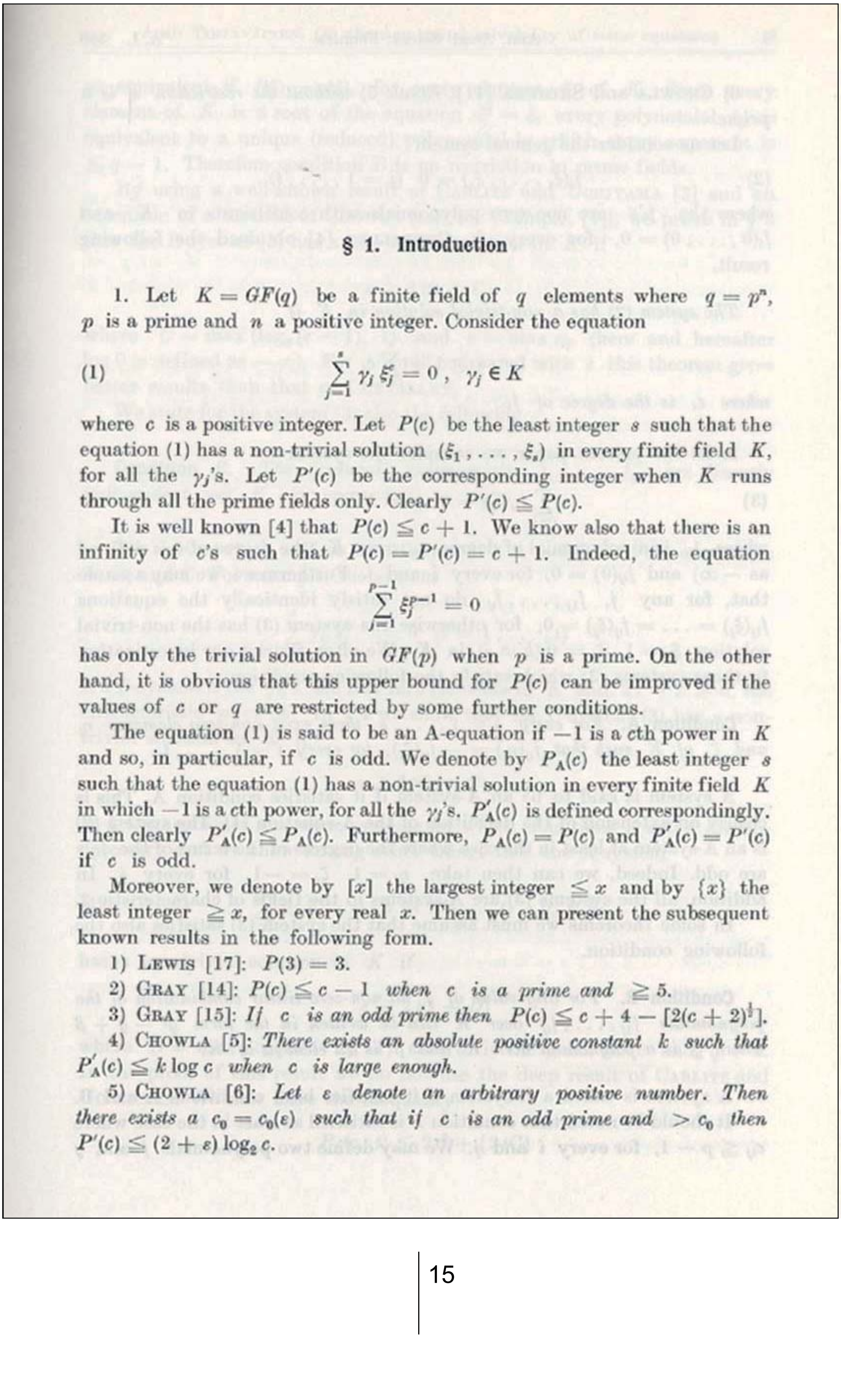
I should like to thank the Finnish Academy of Sciences for accepting this publication for inclusion in the Annals of the Academy.

Turku, February, 1965

AIMO TIETÄVÄINEN



Contents



§ 1. Introduction

1. Let $K = GF(q)$ be a finite field of q elements where $q = p^n$, p is a prime and n a positive integer. Consider the equation

$$(1) \qquad \sum_{j=1}^{s} \gamma_j \xi_j^c = 0 \,, \quad \gamma_j \in K$$

where c is a positive integer. Let $P(c)$ be the least integer s such that the equation (1) has a non-trivial solution $(\xi_1, \dots, \xi_s)$ in every finite field K, for all the γ_j's. Let $P'(c)$ be the corresponding integer when K runs through all the prime fields only. Clearly $P'(c) \leqq P(c)$.

It is well known [4] that $P(c) \leqq c + 1$. We know also that there is an infinity of c's such that $P(c) = P'(c) = c + 1$. Indeed, the equation

$$\sum_{j=1}^{p-1} \xi_j^{p-1} = 0$$

has only the trivial solution in $GF(p)$ when p is a prime. On the other hand, it is obvious that this upper bound for $P(c)$ can be improved if the values of c or q are restricted by some further conditions.

The equation (1) is said to be an A-equation if -1 is a cth power in K and so, in particular, if c is odd. We denote by $P_A(c)$ the least integer s such that the equation (1) has a non-trivial solution in every finite field K in which -1 is a cth power, for all the γ_j's. $P'_A(c)$ is defined correspondingly. Then clearly $P'_A(c) \leqq P_A(c)$. Furthermore, $P_A(c) = P(c)$ and $P'_A(c) = P'(c)$ if c is odd.

Moreover, we denote by $[x]$ the largest integer $\leqq x$ and by $\{x\}$ the least integer $\geqq x$, for every real x. Then we can present the subsequent known results in the following form.

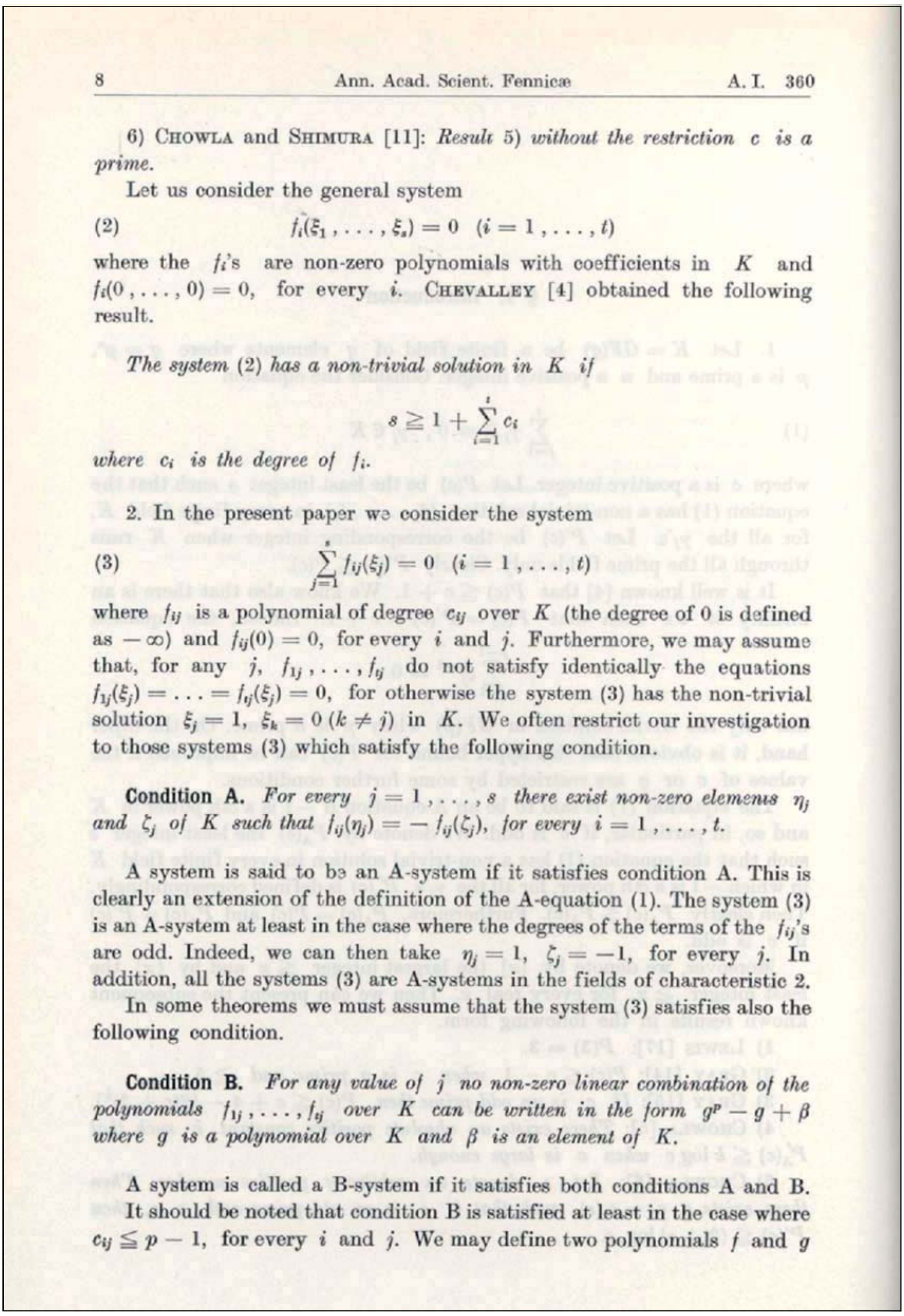
1) Lewis [17]: $P(3) = 3$.

2) Gray [14]: *$P(c) \leqq c - 1$ when c is a prime and $\geqq 5$.*

3) Gray [15]: *If c is an odd prime then $P(c) \leqq c + 4 - [2(c+2)^{\frac{1}{2}}]$.*

4) Chowla [5]: *There exists an absolute positive constant k such that $P'_A(c) \leqq k \log c$ when c is large enough.*

5) Chowla [6]: *Let ε denote an arbitrary positive number. Then there exists a $c_0 = c_0(\varepsilon)$ such that if c is an odd prime and $> c_0$ then $P'(c) \leqq (2 + \varepsilon) \log_2 c$.*



6) Chowla and Shimura [11]: *Result 5) without the restriction c is a prime.*

Let us consider the general system

$$f_i(\xi_1 , \ldots , \xi_s) = 0 \quad (i = 1 , \ldots , t) \tag{2}$$

where the f_i's are non-zero polynomials with coefficients in K and $f_i(0 , \ldots , 0) = 0$, for every i. Chevalley [4] obtained the following result.

The system (2) has a non-trivial solution in K if

$$s \geqq 1 + \sum_{i=1}^{t} c_i$$

where c_i is the degree of f_i.

2. In the present paper we consider the system

$$\sum_{j=1}^{s} f_{ij}(\xi_j) = 0 \quad (i = 1 , \ldots , t) \tag{3}$$

where f_{ij} is a polynomial of degree c_{ij} over K (the degree of 0 is defined as $-\infty$) and $f_{ij}(0) = 0$, for every i and j. Furthermore, we may assume that, for any j, $f_{1j} , \ldots , f_{tj}$ do not satisfy identically the equations $f_{1j}(\xi_j) = \ldots = f_{tj}(\xi_j) = 0$, for otherwise the system (3) has the non-trivial solution $\xi_j = 1$, $\xi_k = 0 \ (k \neq j)$ in K. We often restrict our investigation to those systems (3) which satisfy the following condition.

Condition A. *For every $j = 1 , \ldots , s$ there exist non-zero elemenis η_j and ζ_j of K such that $f_{ij}(\eta_j) = - f_{ij}(\zeta_j)$, for every $i = 1 , \ldots , t$.*

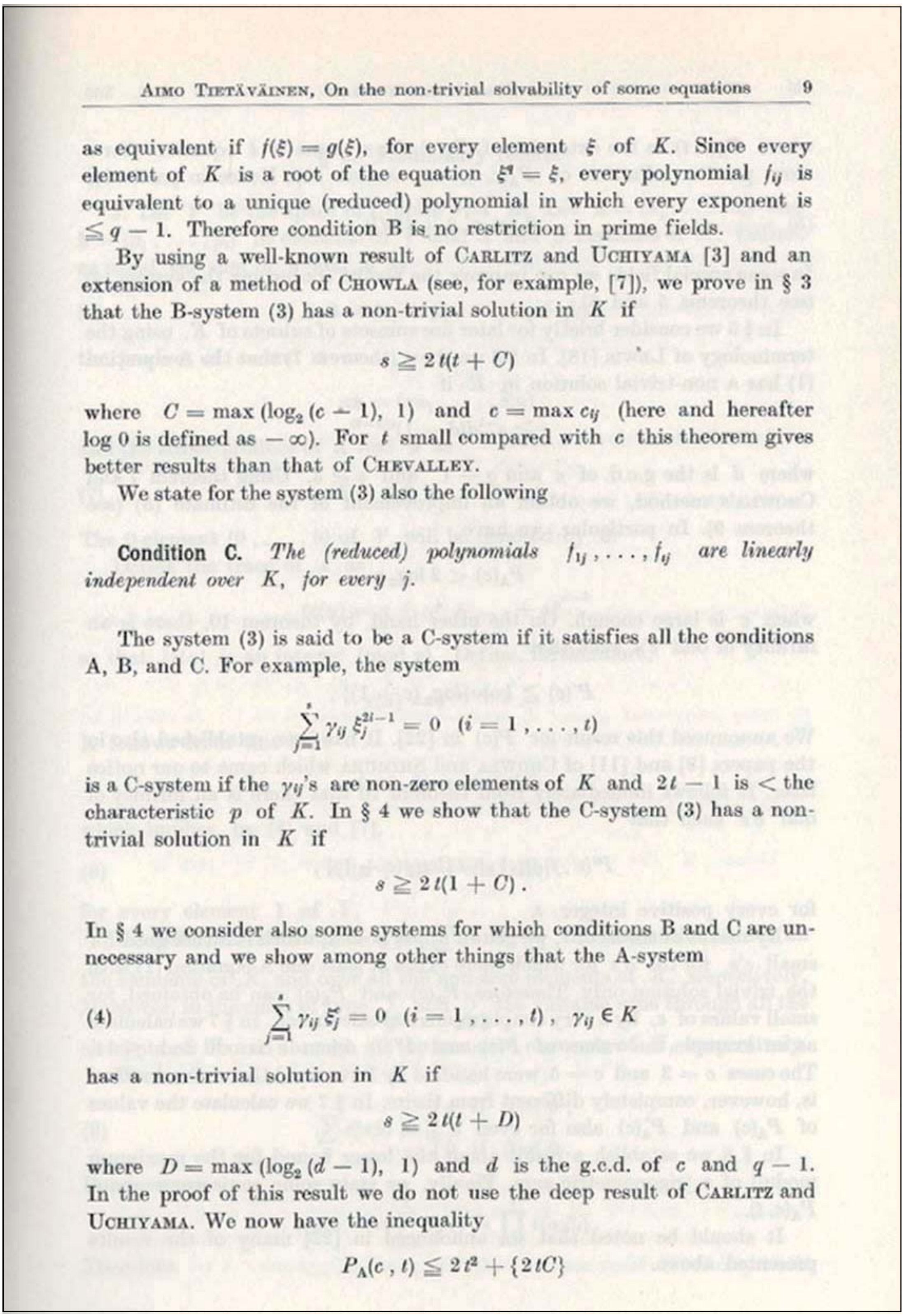
A system is said to be an A-system if it satisfies condition A. This is clearly an extension of the definition of the A-equation (1). The system (3) is an A-system at least in the case where the degrees of the terms of the f_{ij}'s are odd. Indeed, we can then take $\eta_j = 1$, $\zeta_j = -1$, for every j. In addition, all the systems (3) are A-systems in the fields of characteristic 2.

In some theorems we must assume that the system (3) satisfies also the following condition.

Condition B. *For any value of j no non-zero linear combination of the polynomials $f_{1j} , \ldots , f_{tj}$ over K can be written in the form $g^p - g + \beta$ where g is a polynomial over K and β is an element of K.*

A system is called a B-system if it satisfies both conditions A and B.

It should be noted that condition B is satisfied at least in the case where $c_{ij} \leqq p - 1$, for every i and j. We may define two polynomials f and g



as equivalent if $f(\xi) = g(\xi)$, for every element ξ of K. Since every element of K is a root of the equation $\xi^q = \xi$, every polynomial f_{ij} is equivalent to a unique (reduced) polynomial in which every exponent is $\leqq q - 1$. Therefore condition B is no restriction in prime fields.

By using a well-known result of Carlitz and Uchiyama [3] and an extension of a method of Chowla (see, for example, [7]), we prove in § 3 that the B-system (3) has a non-trivial solution in K if

$$s \geqq 2\,t(t + C)$$

where $C = \max(\log_2(c - 1),\ 1)$ and $c = \max c_{ij}$ (here and hereafter $\log 0$ is defined as $-\infty$). For t small compared with c this theorem gives better results than that of Chevalley.

We state for the system (3) also the following

Condition C. *The (reduced) polynomials* $f_{1j}, \ldots, f_{tj}$ *are linearly independent over* K, *for every* j.

The system (3) is said to be a C-system if it satisfies all the conditions A, B, and C. For example, the system

$$\sum_{j=1}^{s} \gamma_{ij}\,\xi_j^{2i-1} = 0 \quad (i = 1, \ldots, t)$$

is a C-system if the γ_{ij}'s are non-zero elements of K and $2t - 1$ is $<$ the characteristic p of K. In § 4 we show that the C-system (3) has a non-trivial solution in K if

$$s \geqq 2\,t(1 + C)\,.$$

In § 4 we consider also some systems for which conditions B and C are unnecessary and we show among other things that the A-system

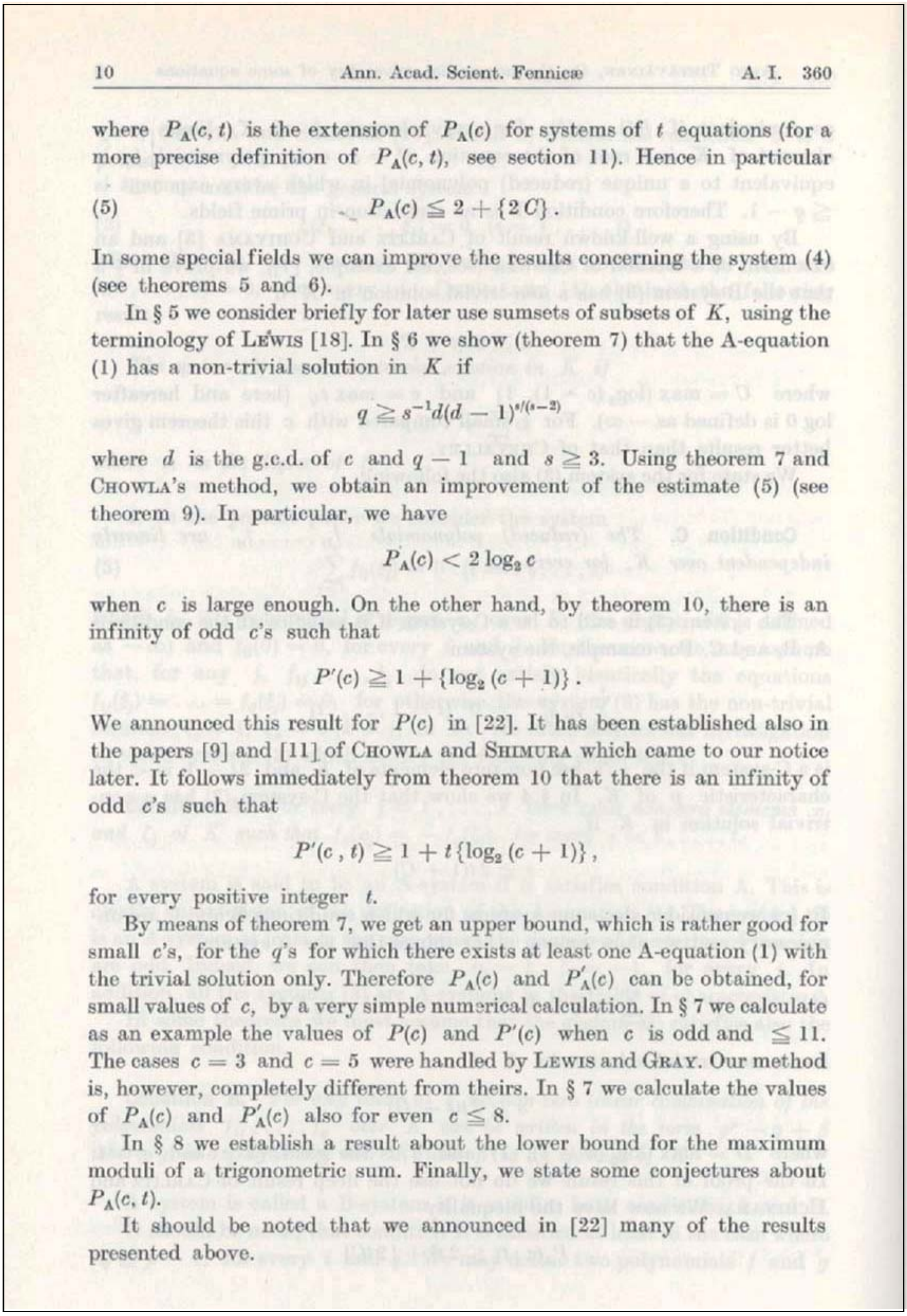
$$\text{(4)} \qquad \sum_{j=1}^{s} \gamma_{ij}\,\xi_j^{i} = 0 \quad (i = 1, \ldots, t)\,, \quad \gamma_{ij} \in K$$

has a non-trivial solution in K if

$$s \geqq 2\,t(t + D)$$

where $D = \max(\log_2(d - 1),\ 1)$ and d is the g.c.d. of c and $q - 1$. In the proof of this result we do not use the deep result of Carlitz and Uchiyama. We now have the inequality

$$P_A(c\,,t) \leqq 2\,t^2 + \{2\,tC\}$$



where $P_A(c, t)$ is the extension of $P_A(c)$ for systems of t equations (for a more precise definition of $P_A(c, t)$, see section 11). Hence in particular

$$(5) \qquad P_A(c) \leqq 2 + \{2C\} .$$

In some special fields we can improve the results concerning the system (4) (see theorems 5 and 6).

In § 5 we consider briefly for later use sumsets of subsets of K, using the terminology of Lewis [18]. In § 6 we show (theorem 7) that the A-equation (1) has a non-trivial solution in K if

$$q \geqq s^{-1}d(d-1)^{s/(s-2)}$$

where d is the g.c.d. of c and $q - 1$ and $s \geqq 3$. Using theorem 7 and Chowla's method, we obtain an improvement of the estimate (5) (see theorem 9). In particular, we have

$$P_A(c) < 2 \log_2 c$$

when c is large enough. On the other hand, by theorem 10, there is an infinity of odd c's such that

$$P'(c) \geqq 1 + \{\log_2 (c + 1)\} .$$

We announced this result for $P(c)$ in [22]. It has been established also in the papers [9] and [11] of Chowla and Shimura which came to our notice later. It follows immediately from theorem 10 that there is an infinity of odd c's such that

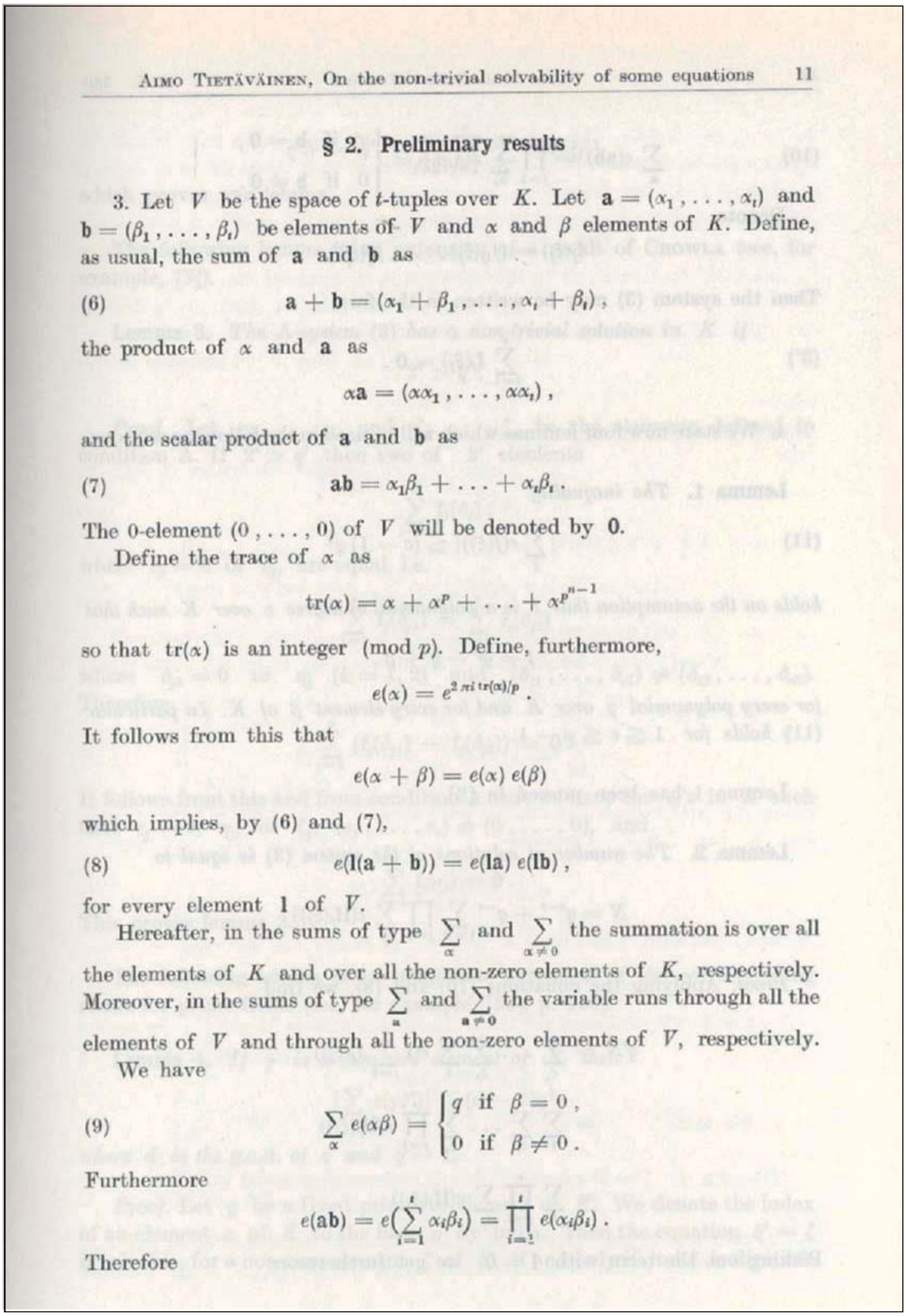
$$P'(c , t) \geqq 1 + t\{\log_2 (c + 1)\} ,$$

for every positive integer t.

By means of theorem 7, we get an upper bound, which is rather good for small c's, for the q's for which there exists at least one A-equation (1) with the trivial solution only. Therefore $P_A(c)$ and $P'_A(c)$ can be obtained, for small values of c, by a very simple numerical calculation. In § 7 we calculate as an example the values of $P(c)$ and $P'(c)$ when c is odd and $\leqq 11$. The cases $c = 3$ and $c = 5$ were handled by Lewis and Gray. Our method is, however, completely different from theirs. In § 7 we calculate the values of $P_A(c)$ and $P'_A(c)$ also for even $c \leqq 8$.

In § 8 we establish a result about the lower bound for the maximum moduli of a trigonometric sum. Finally, we state some conjectures about $P_A(c, t)$.

It should be noted that we announced in [22] many of the results presented above.



§ 2. Preliminary results

3. Let V be the space of t-tuples over K. Let $\mathbf{a} = (\alpha_1, \ldots, \alpha_t)$ and $\mathbf{b} = (\beta_1, \ldots, \beta_t)$ be elements of V and α and β elements of K. Define, as usual, the sum of $\mathbf{a}$ and $\mathbf{b}$ as

$$\mathbf{a} + \mathbf{b} = (\alpha_1 + \beta_1, \ldots, \alpha_t + \beta_t), \tag{6}$$

the product of α and $\mathbf{a}$ as

$$\alpha\mathbf{a} = (\alpha\alpha_1, \ldots, \alpha\alpha_t),$$

and the scalar product of $\mathbf{a}$ and $\mathbf{b}$ as

$$\mathbf{ab} = \alpha_1\beta_1 + \ldots + \alpha_t\beta_t. \tag{7}$$

The 0-element $(0, \ldots, 0)$ of V will be denoted by $\mathbf{0}$.

Define the trace of α as

$$\mathrm{tr}(\alpha) = \alpha + \alpha^p + \ldots + \alpha^{p^{n-1}}$$

so that $\mathrm{tr}(\alpha)$ is an integer (mod p). Define, furthermore,

$$e(\alpha) = e^{2\pi i\, \mathrm{tr}(\alpha)/p}.$$

It follows from this that

$$e(\alpha + \beta) = e(\alpha)\, e(\beta)$$

which implies, by (6) and (7),

$$e(\mathbf{l}(\mathbf{a} + \mathbf{b})) = e(\mathbf{la})\, e(\mathbf{lb}), \tag{8}$$

for every element $\mathbf{l}$ of V.

Hereafter, in the sums of type $\sum\limits_{\alpha}$ and $\sum\limits_{\alpha \neq 0}$ the summation is over all the elements of K and over all the non-zero elements of K, respectively. Moreover, in the sums of type $\sum\limits_{\mathbf{a}}$ and $\sum\limits_{\mathbf{a} \neq \mathbf{0}}$ the variable runs through all the elements of V and through all the non-zero elements of V, respectively.

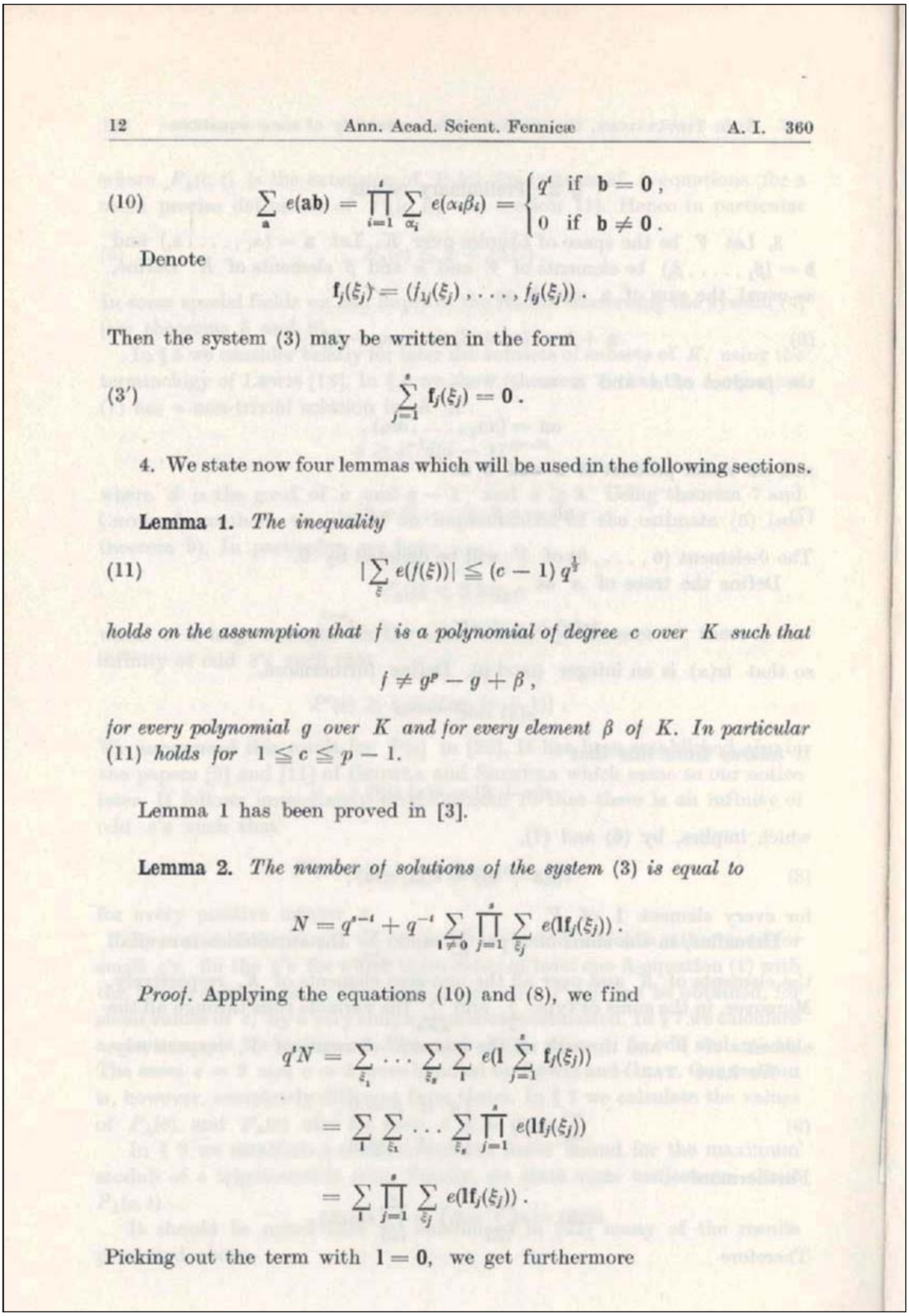
We have

$$\sum_{\alpha} e(\alpha\beta) = \begin{cases} q & \text{if } \beta = 0, \\ 0 & \text{if } \beta \neq 0. \end{cases} \tag{9}$$

Furthermore

$$e(\mathbf{ab}) = e\Big(\sum_{i=1}^{t} \alpha_i\beta_i\Big) = \prod_{i=1}^{t} e(\alpha_i\beta_i).$$

Therefore



$$\text{(10)} \qquad \sum_{\mathbf{a}} e(\mathbf{ab}) = \prod_{i=1}^{t} \sum_{\alpha_i} e(\alpha_i\beta_i) = \begin{cases} q^t & \text{if } \mathbf{b} = \mathbf{0}\,, \\ 0 & \text{if } \mathbf{b} \neq \mathbf{0}\,. \end{cases}$$

Denote

$$\mathbf{f}_j(\xi_j) = (f_{1j}(\xi_j)\,, \ldots, f_{tj}(\xi_j))\,.$$

Then the system (3) may be written in the form

$$\text{(3')} \qquad \sum_{j=1}^{s} \mathbf{f}_j(\xi_j) = \mathbf{0}\,.$$

4. We state now four lemmas which will be used in the following sections.

Lemma 1. *The inequality*

$$\text{(11)} \qquad |\sum_{\xi} e(f(\xi))| \leq (c-1)\, q^{\frac{1}{2}}$$

holds on the assumption that f *is a polynomial of degree* c *over* K *such that*

$$f \neq g^p - g + \beta\,,$$

for every polynomial g *over* K *and for every element* β *of* K. *In particular* (11) *holds for* $1 \leq c \leq p - 1$.

Lemma 1 has been proved in [3].

Lemma 2. *The number of solutions of the system* (3) *is equal to*

$$N = q^{s-t} + q^{-t} \sum_{\mathbf{l} \neq \mathbf{0}} \prod_{j=1}^{s} \sum_{\xi_j} e(\mathbf{l}\mathbf{f}_j(\xi_j))\,.$$

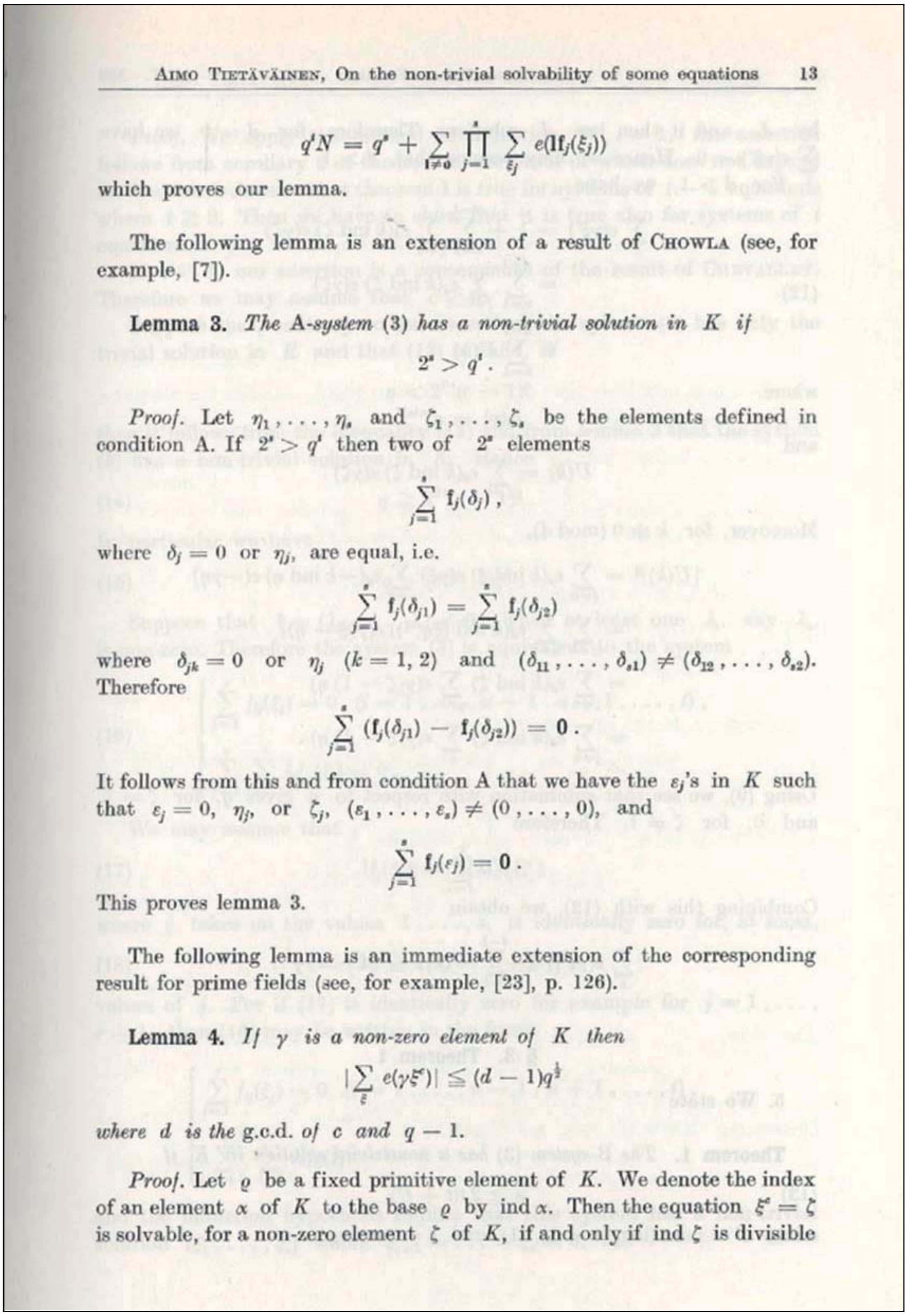
Proof. Applying the equations (10) and (8), we find

$$q^t N = \sum_{\xi_1} \cdots \sum_{\xi_s} \sum_{\mathbf{l}} e(\mathbf{l} \sum_{j=1}^{s} \mathbf{f}_j(\xi_j))$$

$$= \sum_{\mathbf{l}} \sum_{\xi_1} \cdots \sum_{\xi_s} \prod_{j=1}^{s} e(\mathbf{l}\mathbf{f}_j(\xi_j))$$

$$= \sum_{\mathbf{l}} \prod_{j=1}^{s} \sum_{\xi_j} e(\mathbf{l}\mathbf{f}_j(\xi_j))\,.$$

Picking out the term with $\mathbf{l} = \mathbf{0}$, we get furthermore



$$q^t N = q^s + \sum_{\mathfrak{l} \neq 0} \prod_{j=1}^{s} \sum_{\xi_j} e(\mathfrak{l}\mathfrak{f}_j(\xi_j))$$

which proves our lemma.

The following lemma is an extension of a result of Chowla (see, for example, [7]).

Lemma 3. *The* A-*system* (3) *has a non-trivial solution in* K *if*

$$2^s > q^t .$$

Proof. Let $\eta_1, \ldots, \eta_s$ and $\zeta_1, \ldots, \zeta_s$ be the elements defined in condition A. If $2^s > q^t$ then two of 2^s elements

$$\sum_{j=1}^{s} \mathfrak{f}_j(\delta_j) ,$$

where $\delta_j = 0$ or η_j, are equal, i.e.

$$\sum_{j=1}^{s} \mathfrak{f}_j(\delta_{j1}) = \sum_{j=1}^{s} \mathfrak{f}_j(\delta_{j2})$$

where $\delta_{jk} = 0$ or η_j $(k = 1, 2)$ and $(\delta_{11}, \ldots, \delta_{s1}) \neq (\delta_{12}, \ldots, \delta_{s2})$. Therefore

$$\sum_{j=1}^{s} (\mathfrak{f}_j(\delta_{j1}) - \mathfrak{f}_j(\delta_{j2})) = \mathbf{0} .$$

It follows from this and from condition A that we have the ε_j's in K such that $\varepsilon_j = 0$, η_j, or ζ_j, $(\varepsilon_1, \ldots, \varepsilon_s) \neq (0, \ldots, 0)$, and

$$\sum_{j=1}^{s} \mathfrak{f}_j(\varepsilon_j) = \mathbf{0} .$$

This proves lemma 3.

The following lemma is an immediate extension of the corresponding result for prime fields (see, for example, [23], p. 126).

Lemma 4. *If* γ *is a non-zero element of* K *then*

$$|\sum_{\xi} e(\gamma \xi^c)| \leqq (d-1) q^{\frac{1}{2}}$$

where d *is the* g.c.d. *of* c *and* $q - 1$.

Proof. Let ϱ be a fixed primitive element of K. We denote the index of an element α of K to the base ϱ by $\operatorname{ind} \alpha$. Then the equation $\xi^c = \zeta$ is solvable, for a non-zero element ζ of K, if and only if $\operatorname{ind} \zeta$ is divisible by d, and it then has d solutions. Therefore, for $d = 1$ we have $\sum_{\xi} e(\gamma\xi^c) = 0$. Hence we may assume that $d > 1$.

For $d > 1$, we have

$$\begin{aligned} \sum_{\xi} e(\gamma\xi^c) &= 1 + \sum_{k=0}^{d-1} \sum_{\zeta \neq 0} e_d(k \operatorname{ind} \zeta)\, e(\gamma\zeta) \\ &= \sum_{k=1}^{d-1} \sum_{\zeta \neq 0} e_d(k \operatorname{ind} \zeta)\, e(\gamma\zeta) \qquad (12) \\ &= \sum_{k=1}^{d-1} U(k) \end{aligned}$$

where

$$e_d(v) = e^{2\pi i v/d}$$

and

$$U(k) = \sum_{\zeta \neq 0} e_d(k \operatorname{ind} \zeta)\, e(\gamma\zeta)\,.$$

Moreover, for $k \not\equiv 0 \pmod{d}$,

$$\begin{aligned} |U(k)|^2 &= \sum_{\xi \neq 0} e_d(k \operatorname{ind} \xi)\, e(\gamma\xi) \sum_{\eta \neq 0} e_d(-k \operatorname{ind} \eta)\, e(-\gamma\eta) \\ &= \sum_{\xi \neq 0} \sum_{\eta \neq 0} e_d(k \operatorname{ind} (\xi\eta^{-1}))\, e(\gamma(\xi - \eta)) \\ &= \sum_{\zeta \neq 0} e_d(k \operatorname{ind} \zeta) \sum_{\eta \neq 0} e(\gamma(\zeta - 1)\,\eta) \\ &= \sum_{\zeta \neq 0} e_d(k \operatorname{ind} \zeta) \sum_{\eta} e(\gamma(\zeta - 1)\,\eta)\,. \end{aligned}$$

Using (9), we see that summation with respect to η gives q, for $\zeta = 1$, and 0, for $\zeta \neq 1$. Therefore

$$|U(k)|^2 = q\,.$$

Combining this with (12), we obtain

$$|\sum_{\xi} e(\gamma\xi^c)| \leqq \sum_{k=1}^{d-1} |U(k)| \leqq (d-1)\, q^{\frac{1}{2}}\,.$$

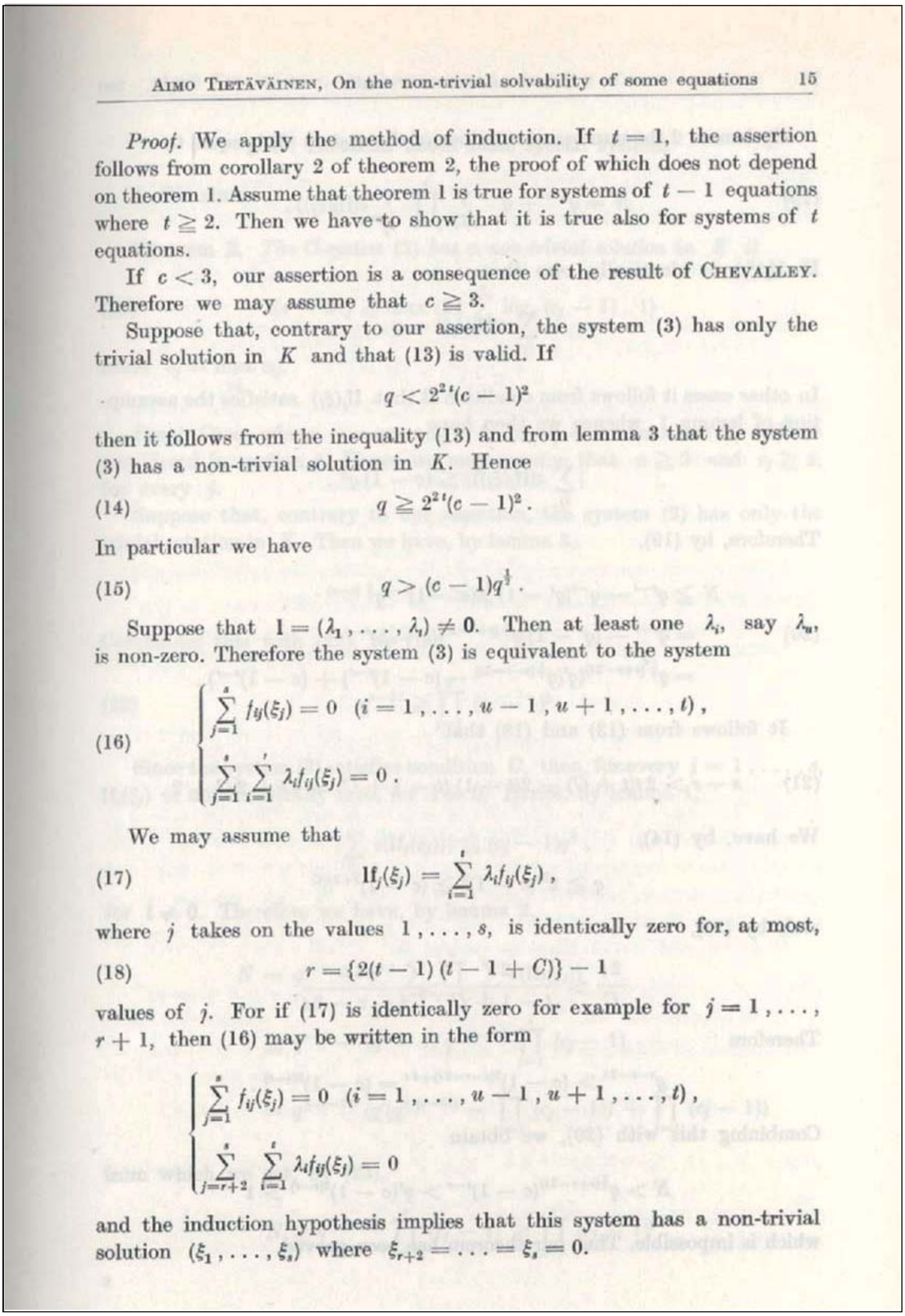
§ 3. Theorem 1

5. We state

Theorem 1. *The* B-*system* (3) *has a non-trivial solution in* K *if*

$$s \geqq 2\, t(t + C) \tag{13}$$

where $C = \max(\log_2(c-1),\ 1)$ *and* $c = \max c_{ij}$.



Proof. We apply the method of induction. If $t = 1$, the assertion follows from corollary 2 of theorem 2, the proof of which does not depend on theorem 1. Assume that theorem 1 is true for systems of $t - 1$ equations where $t \geqq 2$. Then we have to show that it is true also for systems of t equations.

If $c < 3$, our assertion is a consequence of the result of CHEVALLEY. Therefore we may assume that $c \geqq 3$.

Suppose that, contrary to our assertion, the system (3) has only the trivial solution in K and that (13) is valid. If

$$q < 2^{2t}(c-1)^2$$

then it follows from the inequality (13) and from lemma 3 that the system (3) has a non-trivial solution in K. Hence

$$q \geqq 2^{2t}(c-1)^2 . \tag{14}$$

In particular we have

$$q > (c-1)q^{\frac{1}{2}} . \tag{15}$$

Suppose that $\mathbf{l} = (\lambda_1, \ldots, \lambda_t) \neq \mathbf{0}$. Then at least one λ_i, say λ_u, is non-zero. Therefore the system (3) is equivalent to the system

$$\begin{cases} \sum\limits_{j=1}^{s} f_{ij}(\xi_j) = 0 \quad (i = 1, \ldots, u-1, u+1, \ldots, t), \\ \sum\limits_{j=1}^{s} \sum\limits_{i=1}^{t} \lambda_i f_{ij}(\xi_j) = 0 . \end{cases} \tag{16}$$

We may assume that

$$\mathbf{l}\mathbf{f}_j(\xi_j) = \sum_{i=1}^{t} \lambda_i f_{ij}(\xi_j) , \tag{17}$$

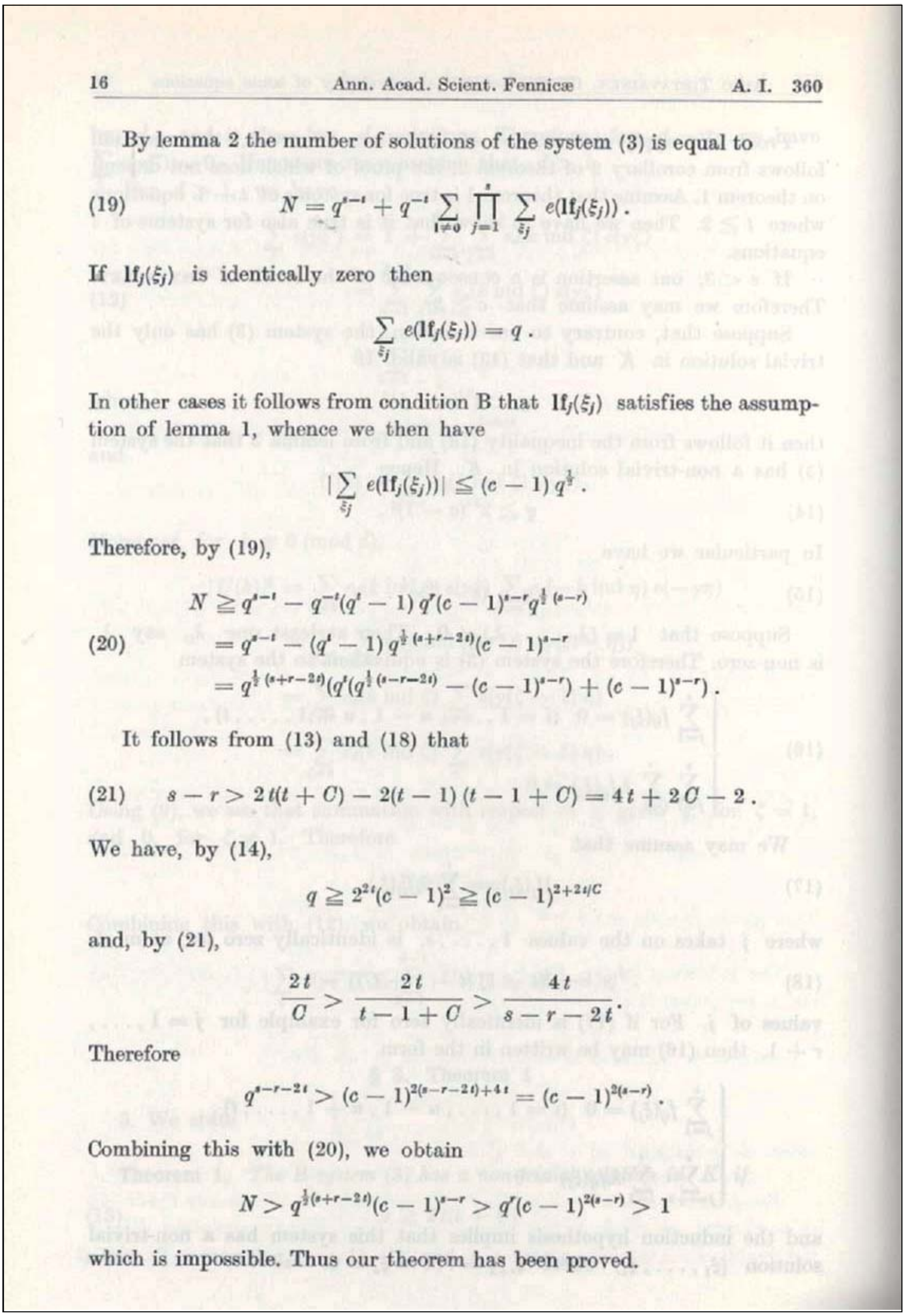
where j takes on the values $1, \ldots, s$, is identically zero for, at most,

$$r = \{2(t-1)(t-1+C)\} - 1 \tag{18}$$

values of j. For if (17) is identically zero for example for $j = 1, \ldots, r+1$, then (16) may be written in the form

$$\begin{cases} \sum\limits_{j=1}^{s} f_{ij}(\xi_j) = 0 \quad (i = 1, \ldots, u-1, u+1, \ldots, t), \\ \sum\limits_{j=r+2}^{s} \sum\limits_{i=1}^{t} \lambda_i f_{ij}(\xi_j) = 0 \end{cases}$$

and the induction hypothesis implies that this system has a non-trivial solution $(\xi_1, \ldots, \xi_s)$ where $\xi_{r+2} = \ldots = \xi_s = 0$.



By lemma 2 the number of solutions of the system (3) is equal to

$$N = q^{s-t} + q^{-t} \sum_{\mathbf{l} \neq 0} \prod_{j=1}^{s} \sum_{\xi_j} e(\mathbf{l}\mathbf{f}_j(\xi_j)) \,. \tag{19}$$

If $\mathbf{l}\mathbf{f}_j(\xi_j)$ is identically zero then

$$\sum_{\xi_j} e(\mathbf{l}\mathbf{f}_j(\xi_j)) = q \,.$$

In other cases it follows from condition B that $\mathbf{l}\mathbf{f}_j(\xi_j)$ satisfies the assumption of lemma 1, whence we then have

$$|\sum_{\xi_j} e(\mathbf{l}\mathbf{f}_j(\xi_j))| \leqq (c-1)\, q^{\frac{1}{2}} \,.$$

Therefore, by (19),

$$\begin{aligned} N &\geqq q^{s-t} - q^{-t}(q^t - 1)\, q^r (c-1)^{s-r} q^{\frac{1}{2}(s-r)} \\ &= q^{s-t} - (q^t - 1)\, q^{\frac{1}{2}(s+r-2t)}(c-1)^{s-r} \\ &= q^{\frac{1}{2}(s+r-2t)}(q^t(q^{\frac{1}{2}(s-r-2t)} - (c-1)^{s-r}) + (c-1)^{s-r}) \,. \end{aligned} \tag{20}$$

It follows from (13) and (18) that

$$s - r > 2t(t+C) - 2(t-1)(t-1+C) = 4t + 2C - 2 \,. \tag{21}$$

We have, by (14),

$$q \geqq 2^{2t}(c-1)^2 \geqq (c-1)^{2+2t/C}$$

and, by (21),

$$\frac{2t}{C} > \frac{2t}{t-1+C} > \frac{4t}{s-r-2t} .$$

Therefore

$$q^{s-r-2t} > (c-1)^{2(s-r-2t)+4t} = (c-1)^{2(s-r)} \,.$$

Combining this with (20), we obtain

$$N > q^{\frac{1}{2}(s+r-2t)}(c-1)^{s-r} > q^r (c-1)^{2(s-r)} > 1$$

which is impossible. Thus our theorem has been proved.

§ 4. Theorems about some special systems

6. We state

Theorem 2. *The* C-*system* (3) *has a non-trivial solution in* K *if*

$$s(s-2t) \geqq \max \left(2t \sum_{j=1}^{s} \log_2 (c_j - 1)\,,\ 1\right) \tag{22}$$

where $c_j = \max\limits_{(i)} c_{ij}$.

Proof. Cases where $c = \max c_{ij} < 3$ or $c_j < 2$, for some j, will be considered in section 8. Hence we may assume that $c \geqq 3$ and $c_j \geqq 2$, for every j.

Suppose that, contrary to our assertion, the system (3) has only the trivial solution in K. Then we have, by lemma 3,

$$2^s \leqq q^t\,.$$

Combining this with (22), we find

$$q^{s-2t} \geqq \prod_{j=1}^{s} (c_j - 1)^2\,. \tag{23}$$

Since the system (3) satisfies condition C, then, for every $j = 1\,, \ldots\,, s$, $\mathbf{l}f_j(\xi_j)$ is not identically zero, for $\mathbf{l} \neq \mathbf{0}$. Hence, by lemma 1,

$$\left|\sum_{\xi_j} e(\mathbf{l}f_j(\xi_j))\right| \leqq (c_j - 1) q^{\frac{1}{2}}\,,$$

for $\mathbf{l} \neq \mathbf{0}$. Therefore we have, by lemma 2,

$$\begin{aligned} N &= q^{s-t} + q^{-t} \sum_{\mathbf{l} \neq 0} \prod_{j=1}^{s} \sum_{\xi_j} e(\mathbf{l}f_j(\xi_j)) \\ &\geqq q^{s-t} - (q^t - 1)\, q^{\frac{1}{2}(s-2t)} \prod_{j=1}^{s} (c_j - 1) \\ &= q^{\frac{1}{2}(s-2t)} \left(q^t\left(q^{\frac{1}{2}(s-2t)} - \prod_{j=1}^{s} (c_j - 1)\right) + \prod_{j=1}^{s} (c_j - 1)\right) \end{aligned}$$

from which we get, by (23),

$$N \geqq q^{\frac{1}{2}(s-2t)} \prod_{j=1}^{s} (c_j - 1) \geqq \prod_{j=1}^{s} (c_j - 1)^2\,.$$

Consequently, by the assumptions $c \geqq 3$ and $c_j \geqq 2$, for every j, we have

$$N > 1$$

which is impossible. Hence theorem 2 is true.

7. Theorem 2 implies the following weaker but simpler result.

Corollary 1. *The* C-*system* (3) *has a non-trivial solution in* K *if*

$$s \geqq 2t(1 + C)$$

where C *is defined as in theorem* 1.

Let us consider the B-system (3) in the case $t = 1$, i.e. the B-equation

$$\sum_{j=1}^{s} f_j(\xi_j) = 0 \, . \tag{24}$$

Since the f_j's are non-zero polynomials, the equation (24) satisfies condition C and hence it is a C-equation. Consequently we have

Corollary 2. *The* B-*equation* (24) *has a non-trivial solution in* K *if*

$$s \geqq 2\,(1 + C)$$

where $C = \max\,(\log_2 (c - 1),\ 1)$, $c = \max c_j$, *and* c_j *is the degree of* f_j.

8. Let us consider in detail the cases of theorem 2 where $c < 3$ or $c_j < 2$, for some j. Now condition A is unnecessary. The case $c_j = 1$ is obvious. Indeed, in this case the system (3) which now consists of one equation has a non-trivial solution, for $s \geqq 2$. Moreover, if we have one quadratic equation, then the value 3 of s obtained by Chevalley is, by a well-known result of Dickson ([13], p. 46), the best possible.

If $t = 2$ and $c < 3$ then the system (3) is of the form

$$\sum_{j=1}^{s} (\alpha_{ij}\xi_j^2 + \beta_{ij}\xi_j) = 0 \quad (i = 1\,,\, 2)\,, \quad \alpha_{ij}\,, \beta_{ij} \in K \, . \tag{25}$$

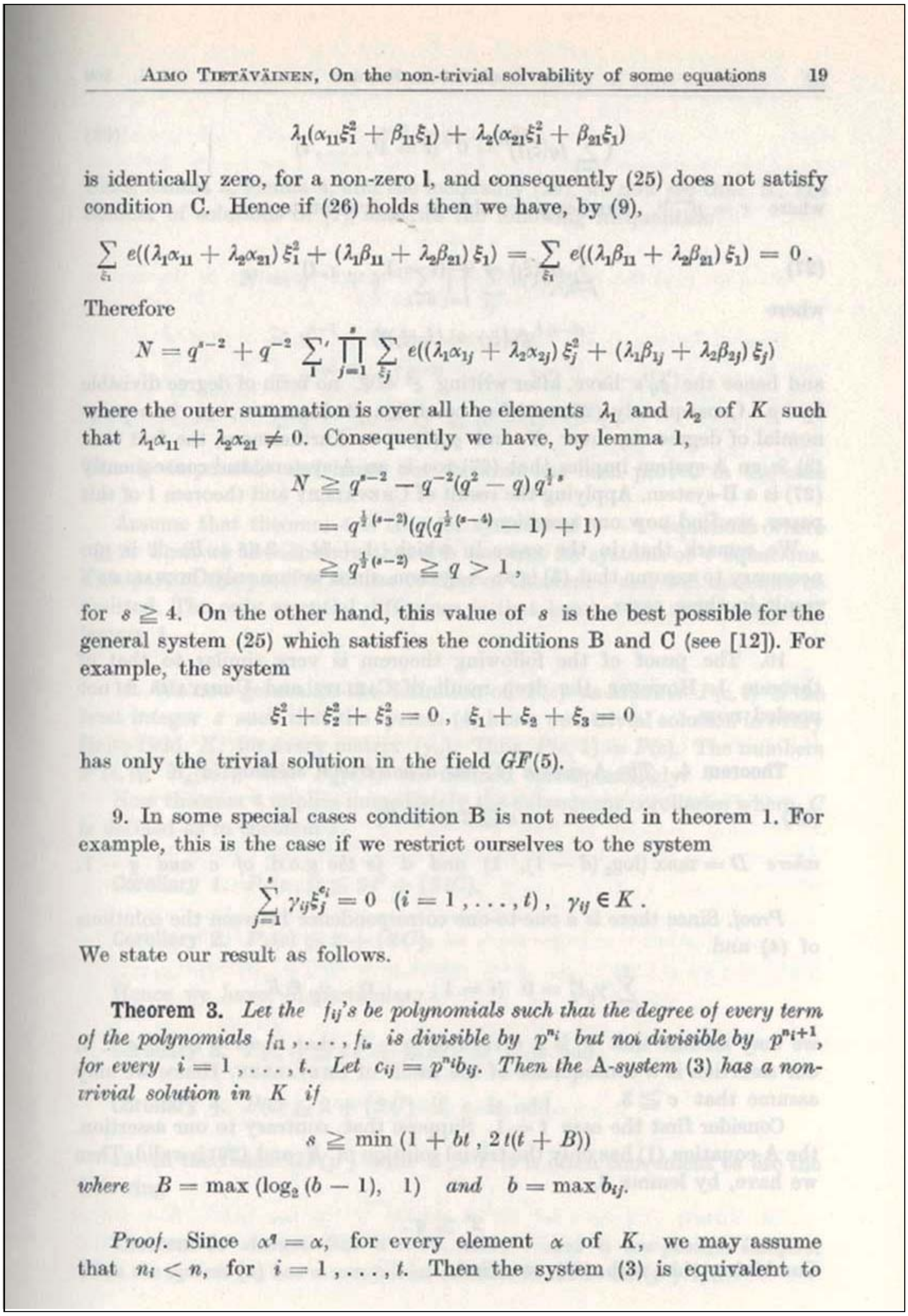
The number of solutions of (25) is equal to

$$N = q^{s-2} + q^{-2} \sum_{\mathbf{l} \neq \mathbf{0}} \prod_{j=1}^{s} \sum_{\xi_j} e((\lambda_1\alpha_{1j} + \lambda_2\alpha_{2j})\xi_j^2 + (\lambda_1\beta_{1j} + \lambda_2\beta_{2j})\xi_j)$$

where $\mathbf{l} = (\lambda_1, \lambda_2)$. Now $\lambda_1\beta_{11} + \lambda_2\beta_{21} \neq 0$ whenever

$$\lambda_1\alpha_{11} + \lambda_2\alpha_{21} = 0\,, \mathbf{l} \neq \mathbf{0}\,, \tag{26}$$

for otherwise



$$\lambda_1(\alpha_{11}\xi_1^2 + \beta_{11}\xi_1) + \lambda_2(\alpha_{21}\xi_1^2 + \beta_{21}\xi_1)$$

is identically zero, for a non-zero $\mathbf{l}$, and consequently (25) does not satisfy condition C. Hence if (26) holds then we have, by (9),

$$\sum_{\xi_1} e((\lambda_1\alpha_{11} + \lambda_2\alpha_{21})\,\xi_1^2 + (\lambda_1\beta_{11} + \lambda_2\beta_{21})\,\xi_1) = \sum_{\xi_1} e((\lambda_1\beta_{11} + \lambda_2\beta_{21})\,\xi_1) = 0\,.$$

Therefore

$$N = q^{s-2} + q^{-2} \sum_{\mathbf{l}}{}' \prod_{j=1}^{s} \sum_{\xi_j} e((\lambda_1\alpha_{1j} + \lambda_2\alpha_{2j})\,\xi_j^2 + (\lambda_1\beta_{1j} + \lambda_2\beta_{2j})\,\xi_j)$$

where the outer summation is over all the elements λ_1 and λ_2 of K such that $\lambda_1\alpha_{11} + \lambda_2\alpha_{21} \neq 0$. Consequently we have, by lemma 1,

$$\begin{aligned} N &\geqq q^{s-2} - q^{-2}(q^2 - q)\,q^{\frac{1}{2}s} \\ &= q^{\frac{1}{2}(s-2)}(q(q^{\frac{1}{2}(s-4)} - 1) + 1) \\ &\geqq q^{\frac{1}{2}(s-2)} \geqq q > 1\,, \end{aligned}$$

for $s \geqq 4$. On the other hand, this value of s is the best possible for the general system (25) which satisfies the conditions B and C (see [12]). For example, the system

$$\xi_1^2 + \xi_2^2 + \xi_3^2 = 0\,, \quad \xi_1 + \xi_2 + \xi_3 = 0$$

has only the trivial solution in the field $GF(5)$.

9. In some special cases condition B is not needed in theorem 1. For example, this is the case if we restrict ourselves to the system

$$\sum_{j=1}^{s} \gamma_{ij}\xi_j^{e_i} = 0 \quad (i = 1\,, \ldots, t)\,, \quad \gamma_{ij} \in K\,.$$

We state our result as follows.

Theorem 3. *Let the* f_{ij}*'s be polynomials such that the degree of every term of the polynomials* $f_{i1}\,, \ldots, f_{is}$ *is divisible by* p^{n_i} *but not divisible by* p^{n_i+1}, *for every* $i = 1\,, \ldots, t$. *Let* $c_{ij} = p^{n_i}b_{ij}$. *Then the* Δ*-system* (3) *has a non-trivial solution in* K *if*

$$s \geqq \min\,(1 + bt\,,\ 2\,t(t + B))$$

where $B = \max\,(\log_2(b-1),\ 1)$ *and* $b = \max b_{ij}$.

Proof. Since $\alpha^q = \alpha$, for every element α of K, we may assume that $n_i < n$, for $i = 1\,, \ldots, t$. Then the system (3) is equivalent to

$$\Big(\sum_{j=1}^{s} f_{ij}(\xi_j)\Big)^r = 0 \quad (i = 1, \ldots, t)$$

where $r = p^{n-n_i}$. Furthermore, this is equivalent to

$$\sum_{j=1}^{s} g_{ij}(\xi_j) = 0 \quad (i = 1, \ldots, t) \tag{27}$$

where

$$g_{ij}(\xi_j) = (f_{ij}(\xi_j))^r$$

and hence the g_{ij}'s have, after writing $\xi^q = \xi$, no term of degree divisible by p. Consequently (27) satisfies condition B. Moreover, g_{ij} is a polynomial of degree b_{ij} over K and $g_{ij}(0) = 0$. Furthermore, the fact that (3) is an A-system implies that (27) too is an A-system, and consequently (27) is a B-system. Applying the result of CHEVALLEY and theorem 1 of this paper, we find now our assertion.

We remark that in the cases in which $1 + bt \leqq 2\,t(t + B)$ it is unnecessary to assume that (3) is an A-system, since we use only CHEVALLEY's result in these cases.

10. The proof of the following theorem is very similar to that of theorem 1. However, the deep result of CARLITZ and UCHIYAMA is not needed now.

Theorem 4. *The* A*-system* (4) *has a non-trivial solution in* K *if*

$$s \geqq 2\,t(t + D) \tag{28}$$

where $D = \max\,(\log_2 (d - 1),\ 1)$ *and* d *is the* g.c.d. *of* c *and* $q - 1$.

Proof. Since there is a one-to-one correspondence between the solutions of (4) and

$$\sum_{j=1}^{s} \gamma_{ij}\xi_j^d = 0 \quad (i = 1, \ldots, t), \quad \gamma_{ij} \in K\,,$$

we may assume that c is a divisor of $q - 1$, that is, $d = c$. If $c < 3$, our assertion is a consequence of the result of CHEVALLEY. Hence we may assume that $c \geqq 3$.

Consider first the case $t = 1$. Suppose that, contrary to our assertion, the A-equation (1) has only the trivial solution in K and (28) is valid. Then we have, by lemma 3,

$$2^s \leqq q\,.$$

Combining this with (28), we find

(29) $$q^{s-2} \geqq (c-1)^{2s}.$$

Using lemma 2, lemma 4, and the inequality (29), we now see that N, the number of solutions of (1), satisfies the following inequalities:

$$\begin{aligned} N &= q^{s-1} + q^{-1} \sum_{\lambda \neq 0} \prod_{j=1}^{s} \sum_{\xi_j} e(\lambda \gamma_j \xi_j^{c}) \\ &\geqq q^{s-1} - (q-1)\,(c-1)^s q^{\frac{1}{2}(s-2)} \\ &= q^{\frac{1}{2}(s-2)}(q(q^{\frac{1}{2}(s-2)} - (c-1)^s) + (c-1)^s) \\ &\geqq q^{\frac{1}{2}(s-2)}(c-1)^s \geqq (c-1)^{2s} > 1\,. \end{aligned}$$

This is impossible and hence our theorem has been proved in the case $t = 1$.

Assume that theorem 4 is true for systems of $t-1$ equations where $t \geqq 2$. Then we have to show that it is also true for systems of t equations. This part of the proof is similar to that of theorem 1 and will therefore be omitted. The only essential difference is that lemma 4 is used instead of lemma 1.

11. We now generalize the definition of $P(c)$ as follows: $P(c, t)$ is the least integer s such that the system (4) has a non-trivial solution in every finite field K, for every matrix (γ_{ij}). Thus $P(c, 1) = P(c)$. The numbers $P'(c, t)$, $P_A(c, t)$, and $P'_A(c, t)$ are defined correspondingly.

Now theorem 4 implies immediately the subsequent corollaries where C is defined as in theorem 1.

Corollary 1. $P_A(c\,, t) \leqq 2t^2 + \{2tC\}$.

Corollary 2. $P_A(c) \leqq 2 + \{2C\}$.

Hence we have, in particular,

Corollary 3. $P(c, t) \leqq 2t^2 + \{2tC\}$ if c is odd.

Corollary 4. $P(c) \leqq 2 + \{2C\}$ if c is odd.

12. In the fields $GF(p^n)$ with $n > 1$ it is often convenient to use the following

Theorem 5. *Assume that $n = ru$ where r and u are positive integers. Then the system (4) has a non-trivial solution in $K = GF(p^n)$ if*

$$s \geqq 1 + a^{-1}\,dtu \tag{30}$$

where d *is the* g.c.d. *of* c *and* $q-1$, a *the* g.c.d. *of* d *and* $h = (p^n-1)/(p^r-1)$.

We note that the system (4) need not be an A-system in theorem 5.

Proof of theorem 5. We may again assume that c is a divisor of p^n-1, that is, $d = c$. Let ϱ be a fixed primitive element of K. Then ϱ^h is a primitive element of $GF(p^r)$. Moreover, ϱ^{hl} is a cth power in K if l is a multiple of $a^{-1}c$. Therefore the $(a^{-1}c)$th powers of the elements of $GF(p^r)$ are cth powers in K.

Let

$$\gamma_{ij} = \sum_{k=1}^{u} \gamma_{ijk}\vartheta^{k-1}$$

where ϑ is a fixed generator of K over $GF(p^r)$ and the γ_{ijk}'s are elements of $GF(p^r)$. Consider the system

$$\sum_{j=1}^{s} \gamma_{ijk}\eta_j^{c/a} = 0 \quad (i = 1, \ldots, t;\ k = 1, \ldots, u)\,. \tag{31}$$

If (30) holds, then, by the theorem of Chevalley, there exists a non-trivial solution $(\eta_1, \ldots, \eta_s)$ of (31) with the η_j's in $GF(p^r)$. Since the $(a^{-1}c)$th powers of the elements of $GF(p^r)$ are cth powers in K, there exist elements $\xi_1, \ldots, \xi_s$ in K which are not all zero and are such that $\xi_j^c = \eta_j^{c/a}$. Thus $(\xi_1, \ldots, \xi_s)$ is a non-trivial solution of (4) and so theorem 5 has been proved.

As an illustrative example we consider the system (4) in the field $GF(64)$. We find that (4) has a non-trivial solution in $GF(64)$ whenever

$$s \geqq \begin{cases} 1+t & \text{if } c = 1 \qquad\qquad (r = 6)\,, \\ 1+2t & \text{if } c = 3 \text{ or } 9 \quad (r = 3)\,, \\ 1+3t & \text{if } c = 7 \text{ or } 21 \quad (r = 2)\,, \\ 1+6t & \text{if } c = 63 \qquad\quad (r = 1)\,. \end{cases}$$

13. Finally, we consider the A-system (4) in the fields $K = GF(q)$ such that $q-1$ is divisible by $3c$. We shall first prove two lemmas (for lemma 5, cf. [21], theorem 1).

Lemma 5. *Suppose that there exist non-zero elements σ and τ of K such that $\sigma^e - \tau^e = 1$. Then the* Λ*-system* (4) *has a non-trivial solution in K if*

$$3^s > q^t . \tag{32}$$

Proof. Since $3^s > q^t$, two of 3^s vectors

$$\left(\sum_{j=1}^{s} \gamma_{1j}\delta_j^e , \ldots , \sum_{j=1}^{s} \gamma_{tj}\delta_j^e\right) ,$$

where $\delta_j = 0$, 1, or σ, are equal. Hence

$$\sum_{j=1}^{s} \gamma_{ij}\delta_{j1}^e = \sum_{j=1}^{s} \gamma_{ij}\delta_{j2}^e \quad (i = 1 , \ldots , t)$$

where $\delta_{jk} = 0$, 1, or σ, for $k = 1$, 2, and

$$(\delta_{11} , \ldots , \delta_{s1}) \neq (\delta_{12} , \ldots , \delta_{s2}) .$$

Therefore

$$\sum_{j=1}^{s} \gamma_{ij}\varepsilon_j = 0 \quad (i = 1 , \ldots , t)$$

where $\varepsilon_j = \delta_{j1}^e - \delta_{j2}^e = 0$, ± 1, $\pm \sigma^e$, or $\pm \tau^e$, and

$$(\varepsilon_1 , \ldots , \varepsilon_s) \neq (0 , \ldots , 0) .$$

Since (4) is an Λ-system, there exists an element η of K such that $\eta^e = -1$. Consequently $\varepsilon_j = \varkappa_j^e$ where $\varkappa_j = 0$, 1, η, σ, $\eta\sigma$, τ, or $\eta\tau$. The system (4) has therefore the non-trivial solution $(\varkappa_1 , \ldots , \varkappa_s)$ in K.

Lemma 6. *Let $q = 1 + 3\,kc$ where k is a positive integer. Then the* Λ*-system* (4) *has a non-trivial solution in K if* (32) *holds.*

Proof. Let ϱ be a primitive element of K. Then

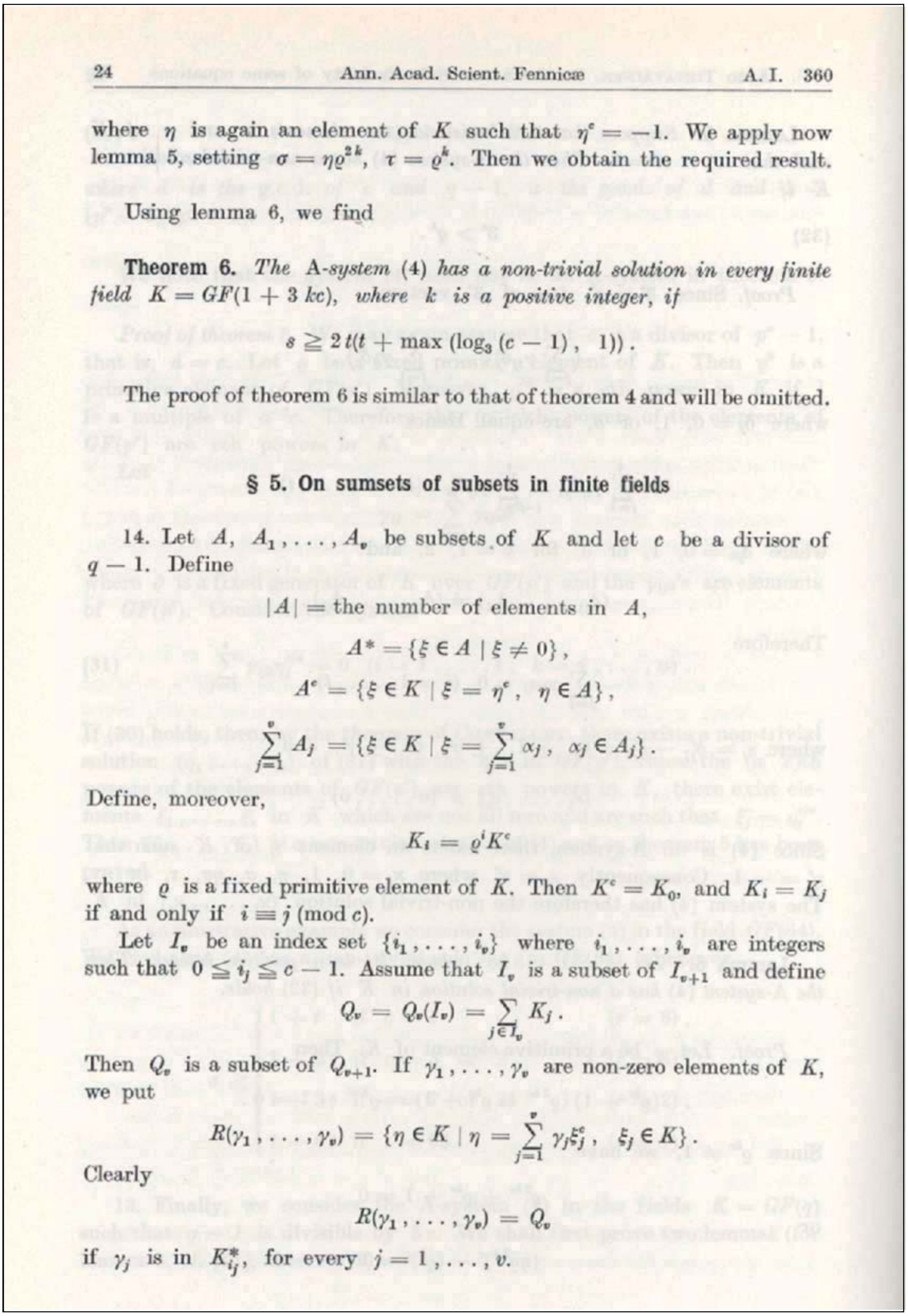
$$(\varrho^{kc} - 1)\,(\varrho^{2\,kc} + \varrho^{kc} + 1) = \varrho^{3\,kc} - 1 = 0 .$$

Since $\varrho^{kc} \neq 1$, we have

$$\varrho^{2\,kc} + \varrho^{kc} + 1 = 0$$

or

$$(\eta\varrho^{2\,k})^e - (\varrho^k)^e = 1$$



where η is again an element of K such that $\eta^c = -1$. We apply now lemma 5, setting $\sigma = \eta\varrho^{2k}$, $\tau = \varrho^k$. Then we obtain the required result.

Using lemma 6, we find

Theorem 6. *The* A-*system* (4) *has a non-trivial solution in every finite field* $K = GF(1 + 3\,kc)$, *where* k *is a positive integer, if*

$$s \geqq 2\,t(t + \max\,(\log_3 (c-1)\,,\ 1))\,.$$

The proof of theorem 6 is similar to that of theorem 4 and will be omitted.

§ 5. On sumsets of subsets in finite fields

14. Let A, $A_1, \ldots, A_v$ be subsets of K and let c be a divisor of $q-1$. Define

$$|A| = \text{the number of elements in } A,$$

$$A^* = \{\xi \in A \mid \xi \neq 0\}\,,$$

$$A^c = \{\xi \in K \mid \xi = \eta^c\,,\ \eta \in A\}\,,$$

$$\sum_{j=1}^{v} A_j = \{\xi \in K \mid \xi = \sum_{j=1}^{v} \alpha_j\,,\ \alpha_j \in A_j\}\,.$$

Define, moreover,

$$K_i = \varrho^i K^c$$

where ϱ is a fixed primitive element of K. Then $K^c = K_0$, and $K_i = K_j$ if and only if $i \equiv j \pmod{c}$.

Let I_v be an index set $\{i_1, \ldots, i_v\}$ where $i_1, \ldots, i_v$ are integers such that $0 \leqq i_j \leqq c-1$. Assume that I_v is a subset of I_{v+1} and define

$$Q_v = Q_v(I_v) = \sum_{j \in I_v} K_j\,.$$

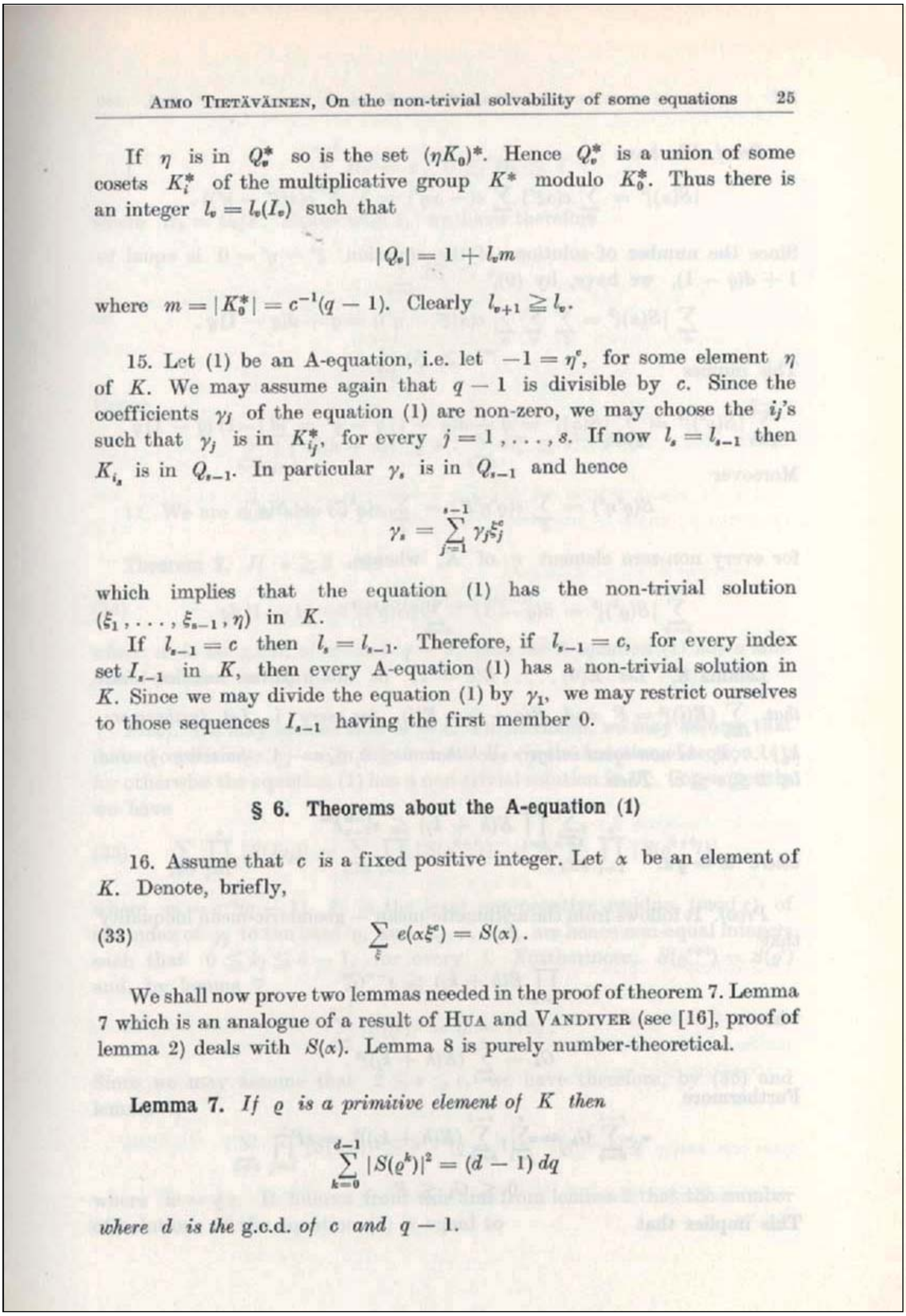
Then Q_v is a subset of Q_{v+1}. If $\gamma_1, \ldots, \gamma_v$ are non-zero elements of K, we put

$$R(\gamma_1, \ldots, \gamma_v) = \{\eta \in K \mid \eta = \sum_{j=1}^{v} \gamma_j \xi_j^c\,,\ \ \xi_j \in K\}\,.$$

Clearly

$$R(\gamma_1, \ldots, \gamma_v) = Q_v$$

if γ_j is in $K^*_{i_j}$, for every $j = 1, \ldots, v$.



If η is in Q_v^* so is the set $(\eta K_0)^*$. Hence Q_v^* is a union of some cosets K_i^* of the multiplicative group K^* modulo K_0^*. Thus there is an integer $l_v = l_v(I_v)$ such that

$$|Q_v| = 1 + l_v m$$

where $m = |K_0^*| = c^{-1}(q-1)$. Clearly $l_{v+1} \geqq l_v$.

15. Let (1) be an A-equation, i.e. let $-1 = \eta^c$, for some element η of K. We may assume again that $q - 1$ is divisible by c. Since the coefficients γ_j of the equation (1) are non-zero, we may choose the i_j's such that γ_j is in $K_{i_j}^*$, for every $j = 1, \ldots, s$. If now $l_s = l_{s-1}$ then K_{i_s} is in Q_{s-1}. In particular γ_s is in Q_{s-1} and hence

$$\gamma_s = \sum_{j=1}^{s-1} \gamma_j \xi_j^c$$

which implies that the equation (1) has the non-trivial solution $(\xi_1, \ldots, \xi_{s-1}, \eta)$ in K.

If $l_{s-1} = c$ then $l_s = l_{s-1}$. Therefore, if $l_{s-1} = c$, for every index set I_{s-1} in K, then every A-equation (1) has a non-trivial solution in K. Since we may divide the equation (1) by γ_1, we may restrict ourselves to those sequences I_{s-1} having the first member 0.

§ 6. Theorems about the A-equation (1)

16. Assume that c is a fixed positive integer. Let α be an element of K. Denote, briefly,

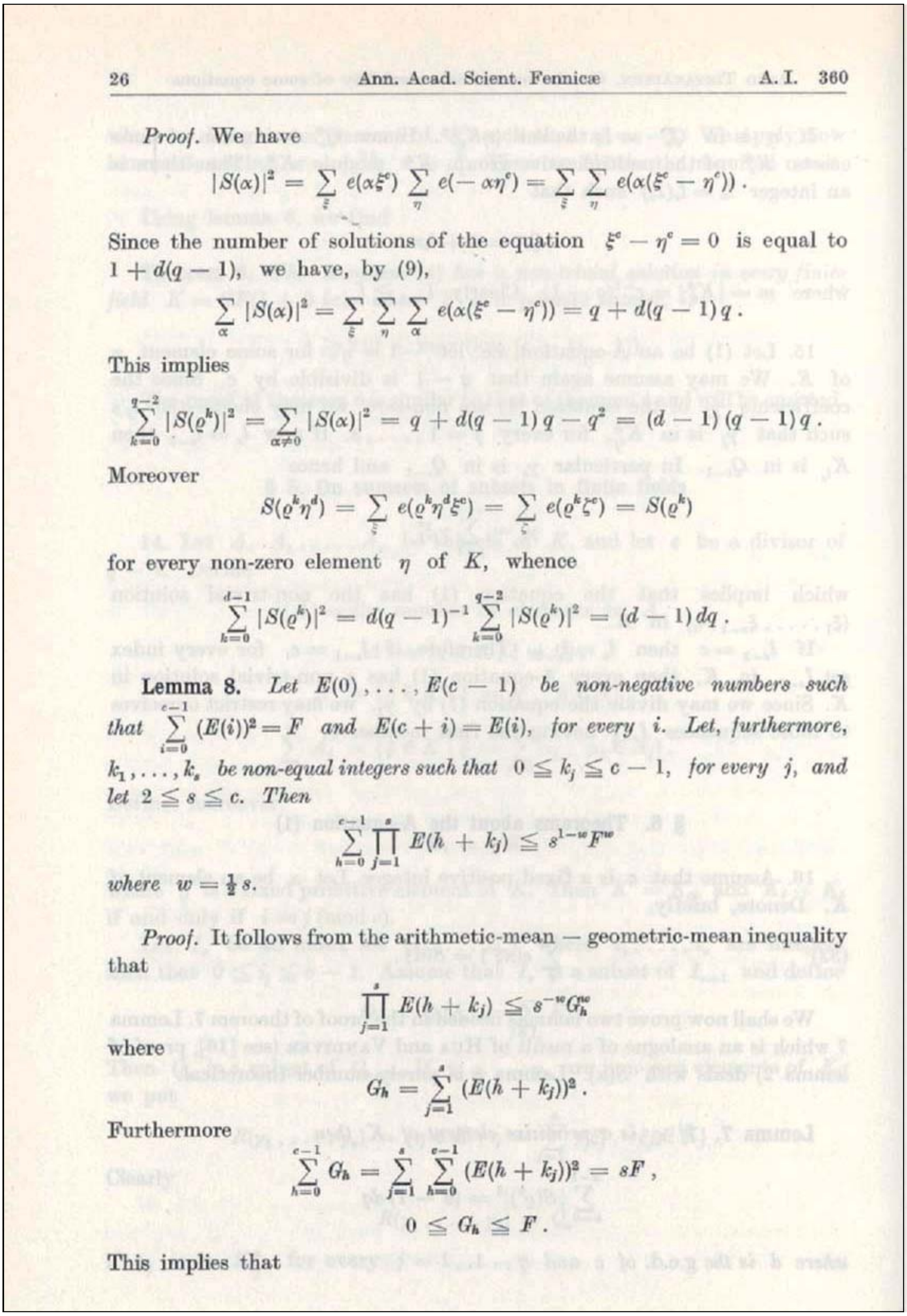
$$\sum_{\xi} e(\alpha \xi^c) = S(\alpha) . \tag{33}$$

We shall now prove two lemmas needed in the proof of theorem 7. Lemma 7 which is an analogue of a result of HUA and VANDIVER (see [16], proof of lemma 2) deals with $S(\alpha)$. Lemma 8 is purely number-theoretical.

Lemma 7. *If ϱ is a primitive element of K then*

$$\sum_{k=0}^{d-1} |S(\varrho^k)|^2 = (d-1)\, dq$$

where d is the g.c.d. *of c and $q-1$.*



Proof. We have

$$|S(\alpha)|^2 = \sum_{\xi} e(\alpha\xi^c) \sum_{\eta} e(-\alpha\eta^c) = \sum_{\xi} \sum_{\eta} e(\alpha(\xi^c - \eta^c)) .$$

Since the number of solutions of the equation $\xi^c - \eta^c = 0$ is equal to $1 + d(q-1)$, we have, by (9),

$$\sum_{\alpha} |S(\alpha)|^2 = \sum_{\xi} \sum_{\eta} \sum_{\alpha} e(\alpha(\xi^c - \eta^c)) = q + d(q-1)\,q .$$

This implies

$$\sum_{k=0}^{q-2} |S(\varrho^k)|^2 = \sum_{\alpha \neq 0} |S(\alpha)|^2 = q + d(q-1)\,q - q^2 = (d-1)\,(q-1)\,q .$$

Moreover

$$S(\varrho^k \eta^d) = \sum_{\xi} e(\varrho^k \eta^d \xi^c) = \sum_{\zeta} e(\varrho^k \zeta^c) = S(\varrho^k)$$

for every non-zero element η of K, whence

$$\sum_{k=0}^{d-1} |S(\varrho^k)|^2 = d(q-1)^{-1} \sum_{k=0}^{q-2} |S(\varrho^k)|^2 = (d-1)\,dq .$$

Lemma 8. *Let* $E(0)\,, \ldots, E(c-1)$ *be non-negative numbers such that* $\sum_{i=0}^{c-1} (E(i))^2 = F$ *and* $E(c+i) = E(i)$, *for every* i. *Let, furthermore,* $k_1\,, \ldots, k_s$ *be non-equal integers such that* $0 \leqq k_j \leqq c-1$, *for every* j, *and let* $2 \leqq s \leqq c$. *Then*

$$\sum_{h=0}^{c-1} \prod_{j=1}^{s} E(h + k_j) \leqq s^{1-w} F^w$$

where $w = \frac{1}{2} s$.

Proof. It follows from the arithmetic-mean — geometric-mean inequality that

$$\prod_{j=1}^{s} E(h + k_j) \leqq s^{-w} G_h^w$$

where

$$G_h = \sum_{j=1}^{s} (E(h + k_j))^2 .$$

Furthermore

$$\sum_{h=0}^{c-1} G_h = \sum_{j=1}^{s} \sum_{h=0}^{c-1} (E(h + k_j))^2 = sF ,$$

$$0 \leqq G_h \leqq F .$$

This implies that

$$\sum_{h=0}^{c-1} H_h = s\,, \quad 0 \leqq H_h \leqq 1$$

where $H_h = G_h/F$. Since $w \geqq 1$, we have therefore

$$\sum_{h=0}^{c-1} H_h^w \leqq s$$

or

$$\sum_{h=0}^{c-1} G_h^w \leqq sF^w\,.$$

Hence

$$\sum_{h=0}^{c-1} \prod_{j=1}^{s} E(h+k_j) \leqq s^{-w} \sum_{h=0}^{c-1} G_h^w \leqq s^{1-w} F^w\,.$$

17. We are now able to prove

Theorem 7. *If* $s \geqq 3$ *and*

$$q \geqq s^{-1} d(d-1)^{s/(s-2)}\,, \tag{34}$$

where d *is the* g.c.d. *of* c *and* $q-1$, *then the* A-*equation* (1) *has a non-trivial solution in* K.

Proof. We may assume that $d = c$. Furthermore, we may assume that no two coefficients γ_j are in the same set K_i (in the notation of section 14), for otherwise the equation (1) has a non-trivial solution in K. Consequently we have

$$\sum_{\lambda \neq 0} \prod_{j=1}^{s} |S(\lambda\gamma_j)| = \sum_{h=0}^{q-2} \prod_{j=1}^{s} |S(\varrho^{h+k_j})| = m \sum_{h=0}^{c-1} \prod_{j=1}^{s} |S(\varrho^{h+k_j})| \tag{35}$$

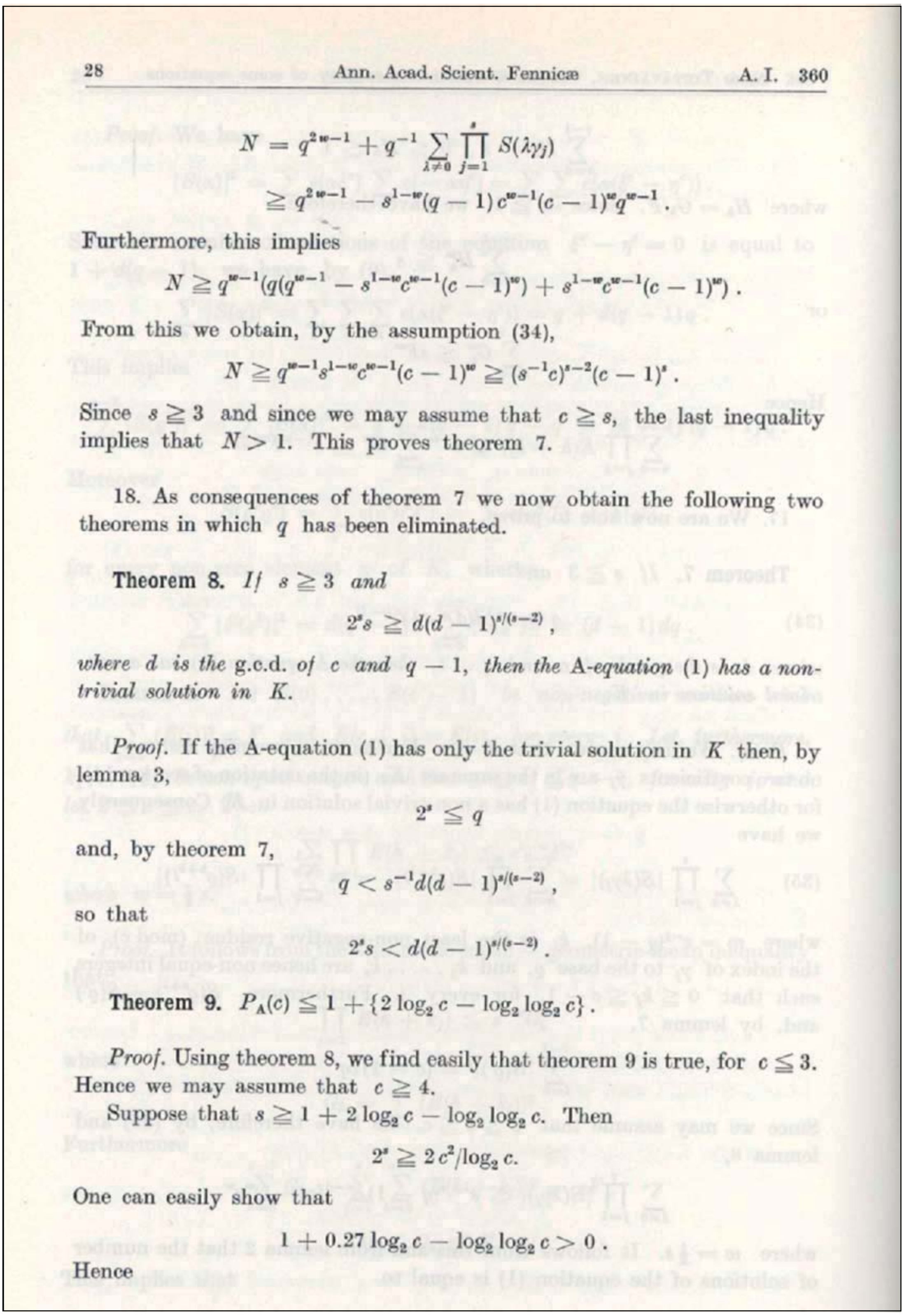
where $m = c^{-1}(q-1)$, k_j is the least non-negative residue $(\bmod\, c)$ of the index of γ_j to the base ϱ, and $k_1, \ldots, k_s$ are hence non-equal integers such that $0 \leqq k_j \leqq c-1$, for every j. Furthermore, $S(\varrho^{c+i}) = S(\varrho^i)$ and, by lemma 7,

$$\sum_{i=0}^{c-1} |S(\varrho^i)|^2 = (c-1)\,cq\,.$$

Since we may assume that $2 \leqq s \leqq c$, we have therefore, by (35) and lemma 8,

$$\sum_{\lambda \neq 0} \prod_{j=1}^{s} |S(\lambda\gamma_j)| \leqq s^{1-w}(q-1)\,c^{w-1}(c-1)^w q^w$$

where $w = \frac{1}{2}s$. It follows from this and from lemma 2 that the number of solutions of the equation (1) is equal to



$$N = q^{2w-1} + q^{-1} \sum_{\lambda \neq 0} \prod_{j=1}^{s} S(\lambda \gamma_j)$$
$$\geqq q^{2w-1} - s^{1-w}(q-1)\, c^{w-1}(c-1)^w q^{w-1} .$$

Furthermore, this implies

$$N \geqq q^{w-1}(q(q^{w-1} - s^{1-w}c^{w-1}(c-1)^w) + s^{1-w}c^{w-1}(c-1)^w) .$$

From this we obtain, by the assumption (34),

$$N \geqq q^{w-1}s^{1-w}c^{w-1}(c-1)^w \geqq (s^{-1}c)^{s-2}(c-1)^s .$$

Since $s \geqq 3$ and since we may assume that $c \geqq s$, the last inequality implies that $N > 1$. This proves theorem 7.

18. As consequences of theorem 7 we now obtain the following two theorems in which q has been eliminated.

Theorem 8. *If* $s \geqq 3$ *and*

$$2^s s \geqq d(d-1)^{s/(s-2)} ,$$

where d *is the* g.c.d. *of* c *and* $q-1$, *then the* A-*equation* (1) *has a non-trivial solution in* K.

Proof. If the A-equation (1) has only the trivial solution in K then, by lemma 3,

$$2^s \leqq q$$

and, by theorem 7,

$$q < s^{-1}d(d-1)^{s/(s-2)} ,$$

so that

$$2^s s < d(d-1)^{s/(s-2)} .$$

Theorem 9. $P_{\mathrm{A}}(c) \leqq 1 + \{2 \log_2 c - \log_2 \log_2 c\}$.

Proof. Using theorem 8, we find easily that theorem 9 is true, for $c \leqq 3$. Hence we may assume that $c \geqq 4$.

Suppose that $s \geqq 1 + 2 \log_2 c - \log_2 \log_2 c$. Then

$$2^s \geqq 2\, c^2/\log_2 c .$$

One can easily show that

$$1 + 0.27 \log_2 c - \log_2 \log_2 c > 0 .$$

Hence

$$s > 1.73 \log_2 c$$

and

$$2^s s > 3.46\, c^2 . \tag{36}$$

On the other hand

$$E \log_2 c - 1 - \log_2 \log_2 c \geqq 0$$

when

$$E = \begin{cases} 1\,, & \text{for } 4 \leqq c \leqq 7\,, \\ 0.87\,, & \text{for } c \geqq 8\,. \end{cases}$$

Hence

$$s - 2 \geqq (2 - E) \log_2 c$$

and

$$\frac{2}{s-2} \leqq \frac{2}{(2-E)\log_2 c} \leqq \begin{cases} \dfrac{2}{\log_2 c}\,, & \text{for } 4 \leqq c \leqq 7\,, \\ \dfrac{1.77}{\log_2 c}\,, & \text{for } c \geqq 8\,. \end{cases}$$

Consequently

$$(c-1)^{2/(s-2)} < c^{2/(s-2)} \leqq \begin{cases} 4\,, & \text{for } 4 \leqq c \leqq 7\,, \\ 3.42\,, & \text{for } c \geqq 8\,. \end{cases}$$

Therefore

$$(c-1)^{s/(s-2)} < \begin{cases} 4 \cdot 0.86\, c = 3.44\, c\,, & \text{for } 4 \leqq c \leqq 7\,, \\ 3.42\, c\,, & \text{for } c \geqq 8\,. \end{cases}$$

Combining this with (36), we find

$$2^s s > c(c-1)^{s/(s-2)}\,,$$

for every $c \geqq 4$. This proves our theorem.

Theorem 9 implies immediately

Corollary. *$P_A(c) < 2 \log_2 c$ when c is large enough.*

19. Theorem 9 gives an upper bound for $P_A(c)$. On the other hand, we have

Theorem 10. *There is an infinity of odd c's such that*

$$P'(c) \geqq 1 + \{\log_2 (c+1)\} .$$

Theorem 10 implies obviously the corresponding results for $P(c)$ (c odd), for $P'_\mathrm{A}(c)$, and for $P_\mathrm{A}(c)$. Since there is an infinity of primes of the form $4m+3$, theorem 10 is an immediate consequence of the following

Lemma 9. *If $2c+1$ is a prime then*

$$P'_\mathrm{A}(c) \geqq 1 + \{\log_2 (c+1)\} . \tag{37}$$

We remark that there exist c's such that (37) is not true. For example, $P'_\mathrm{A}(4) = 3$ (see section 22).

Proof of lemma 9. We use the notation of section 14. In the field $K = GF(2c+1)$ $K_0 = \{0, 1, -1\}$. Let

$$K_{i_j} = \{0 , 2^{j-1} , - 2^{j-1}\} \quad (j = 1 , 2 , \ldots) .$$

Then

$$Q_{v-1} = \{0 , \pm 1 , \ldots , \pm (2^{v-1} - 1)\} .$$

Hence

$$K_{i_v} \cap Q_{v-1} = \{0\}$$

when $2^v - 1 < 2c + 1$ i.e. $2^{v-1} < c + 1$. If $s < 1 + \log_2 (c+1)$ then the equation

$$\sum_{j=1}^{s} 2^{j-1} \xi_j^c = 0$$

has therefore only the trivial solution in $GF(2c+1)$. This proves lemma 9.

Denote $l = \{\log_2 (c+1)\}$. Let c be an odd integer and K a finite field such that there exists an equation

$$\sum_{j=1}^{l} \gamma_j \xi_j^c = 0 , \quad \gamma_j \in K$$

with the trivial solution only. Then the system

$$\sum_{j=1}^{l} \gamma_j \xi_{l(i-1)+j}^c = 0 \quad (i = 1 , \ldots , t)$$

has no non-trivial solution, for every positive integer t. Hence we have the following corollary of theorem 10.

Corollary. *There is an infinity of odd c's such that*

$$P'(c, t) \geqq 1 + t\{\log_2 (c+1)\},$$

for every t.

The corresponding results also hold for $P(c, t)$ (c odd), for $P'_{\mathrm{A}}(c, t)$, and for $P_{\mathrm{A}}(c, t)$.

§ 7. Calculation of $P_{\mathrm{A}}(c)$ and $P'_{\mathrm{A}}(c)$, for small values of c

20. We may again restrict ourselves to fields $K = GF(q)$ such that $q - 1$ is divisible by c. We require two lemmas (in which we use the same notation as earlier).

Lemma 10. *If q is a prime and $c < \frac{1}{2}(q-1)$, then*

$$l_{v+1} \geqq \min (l_v + 2, c).$$

This lemma has been proved in [10].

Lemma 11. *If q is a prime, $c < \frac{1}{2}(q-1)$, and*

$$l_2(\{0, i_2\}) \geqq c - 2(s-3) \tag{38}$$

whenever i_2 is one of the integers

$$1, 2, \ldots, [c/(s-1)], \tag{39}$$

then the A-*equation* (1) *has a non-trivial solution in K, for all the elements $\gamma_1, \ldots, \gamma_s$ of K.*

Proof. We showed in section 15 that if $l_{s-1} = l_{s-1}(I_{s-1}) = c$, for every index set

$$I_{s-1} = \{0, i_2, \ldots, i_{s-1}\} \tag{40}$$

in K, then the A-equation (1) has a non-trivial solution in K, for all the elements γ_j of K. Furthermore, it follows from lemma 10, by induction, that (38) implies the equation $l_{s-1}(I_{s-1}) = c$.

Obviously the following transformations of the index set (40) are allowable:

(i) permutation of the members of (40),

(ii) addition (mod c) of the fixed integer r to every member of (40) (corresponding to the multiplying of the equation (1) by ϱ^r).

Furthermore, we may assume, by the proof of theorem 7, that the i_j's are non-equal and each of them is $\neq 0$. Therefore the members of (40) are $s-1$ non-equal elements of the cycle

$$(0, 1, \ldots, c-1). \tag{41}$$

It follows from this that there are two members of (40) such that the distance between them in the cycle (41) is $\leqq [c/(s-1)]$. Consequently, applying the transformations (i) and (ii), we can carry every index set (40) to a form such that the first term is zero and the second term is one of the integers (39). This proves the lemma.

21. We now turn to the investigation of $P(c)$ and $P'(c)$, for $c \leqq 11$ and odd. We also consider cases where $c = 3$ or $c = 5$, because our method is completely different from that of Lewis and Gray.

I. $c = 3$. It follows from lemma 9 that $P'(3) \geqq 3$. By theorem 8 $P(3) \leqq 3$. Hence $P(3) = P'(3) = 3$.

II. $c = 5$. Lemma 9 implies that $P'(5) \geqq 4$. On the other hand, the equation

$$\sum_{j=1}^{4} \gamma_j \xi_j^5 = 0 \tag{42}$$

has a non-trivial solution in every finite field $GF(q)$, for $q < 16$ by lemma 3, for $q = 16$ by lemma 6, and for $q > 20$ by theorem 7. Hence $P(5) = P'(5) = 4$.

It should be mentioned that the existence of a non-trivial solution of the equation (42) in the field $GF(16)$ also follows from a deep theorem of Mitchell ([19], theorem 2), from a result of Segre ([20], p. 252), or from our theorem 5 (with $r = 2$).

III. $c = 7$. The equation

$$\sum_{j=1}^{4} \gamma_j \xi_j^7 = 0 \tag{43}$$

has a non-trivial solution in the field $GF(q)$, for $q < 16$ by lemma 3, for $q = 43$ by lemma 6, and for $q > 63$ by theorem 7. Therefore we are left to consider the equation (43) in the field $GF(29)$ only.

Using the element 2 as a primitive element of $K = GF(29)$, we get, by means of a simple numerical calculation,

$$K_0 + K_1 = K_0 \cup K_1 \cup K_2 \cup K_5 \cup K_6 ,$$
$$K_0 + K_2 = K_0 \cup K_1 \cup K_2 \cup K_3 \cup K_4 \cup K_5 .$$

Hence $l_2(\{0\,,\,i_2\}) \geqq 5$, for $i_2 = 1, 2$. Therefore, by lemma 11, (43) has a non-trivial solution in $GF(29)$. So we have $P(7) \leqq 4$. On the other hand, the equation

$$\xi_1^7 + 2\,\xi_2^7 + 8\,\xi_3^7 = 0$$

has only the trivial solution in $GF(29)$, since K_3 is not a subset of $K_0 + K_1$ in this field. Therefore $P'(7) \geqq 4$. Hence $P(7) = P'(7) = 4$.

IV. $c = 9$. It follows from lemma 9 that $P'(9) \geqq 5$. Therefore, by lemma 3 and theorem 7, $P(9) = P'(9) = 5$ if every equation

$$\sum_{j=1}^{5} \gamma_j \xi_j^9 = 0 \tag{44}$$

has a non-trivial solution in $GF(37)$. Using the element 2 as a primitive element, we find that in $GF(37)$

$$K_0 + K_1 = K_0 \cup K_1 \cup K_2 \cup K_3 \cup K_8\,,$$
$$K_0 + K_2 = K_0 \cup K_1 \cup K_2 \cup K_5 \cup K_6 \cup K_8\,.$$

Hence $l_2(\{0, i_2\}) \geqq 5$, for $i_2 = 1$, 2. Therefore, by lemma 11, every equation (44) has a non-trivial solution in $GF(37)$ and consequently $P(9) = P'(9) = 5$.

V. $c = 11$. It follows from lemma 9 that $P'(11) \geqq 5$. Therefore we have to consider the equation

$$\sum_{j=1}^{5} \gamma_j \xi_j^{11} = 0\,. \tag{45}$$

It follows from lemma 3, lemma 6, and theorem 7 that it is sufficient to consider the equation (45) in the field $GF(89)$. In this field

$$K_0 + K_1 = K_0 \cup K_1 \cup K_2 \cup K_5 \cup K_8 \cup K_9 \cup K_{10}\,,$$
$$K_0 + K_2 = K_0 \cup K_1 \cup K_2 \cup K_3 \cup K_4 \cup K_5 \cup K_7 \cup K_8 \cup K_9$$

when 3 is used as a primitive element. Hence $P(11) = P'(11) = 5$.

22. We now calculate the values of $P_A(c)$ and $P'_A(c)$ for even $c \leqq 8$.

I. $c = 2$. Theorem 9 and lemma 9 imply the well-known result $P_A(2) = P'_A(2) = 3$.

II. $c = 4$. Since $P'_A(2) = 3$, then $P'_A(4) \geqq 3$ On the other hand, the equation

$$\gamma_1\xi_1^4 + \gamma_2\xi_2^4 + \gamma_3\xi_3^4 = 0$$

has a non-trivial solution in $GF(q)$, for $q = 9$ by theorem 5 (with $r = 1$), for $q = 25$ by lemma 6, and for $q > 36$ by theorem 7. In addition, in the field $GF(17)$ $K_0 + K_1 = K_0 + K_2 = K$. This implies, by lemma 11, that $P_A(4) \leqq 3$. Hence $P_A(4) = P_A'(4) = 3$.

III. $c = 6$. It follows from lemma 9 that $P_A'(6) \geqq 4$. On the other hand $P_A(6) \leqq 4$ by lemma 3 (for $q < 16$), theorem 5 (for $q = 25$), lemma 6 (for $q = 37$), and theorem 7 (for $q > 37$). Hence $P_A(6) = P_A'(6) = 4$.

IV. $c = 8$. Lemma 9, lemma 3, and theorem 7 imply that $P_A(8) = P_A'(8) = 5$.

In cases where $c \geqq 10$ the calculations are more complicated.

§ 8. On the lower bound for the maximum moduli of a trigonometric sum

23. Let c be a fixed positive integer and $\mathbf{a} = (\alpha_1, \ldots, \alpha_c)$ a c-dimensional vector over $K = GF(q)$. Define

$$S(\mathbf{a}) = \sum_{\xi} e(\alpha_1\xi + \ldots + \alpha_c\xi^c) .$$

Anderson and Stiffler [1] have considered the function

$$M(c) = \max_{\mathbf{a} \neq 0} |S(\mathbf{a})| .$$

Carlitz and Uchiyama ([3], our lemma 1) have proved that

$$M(c) \leqq (c-1)\,q^{\frac{1}{2}} ,$$

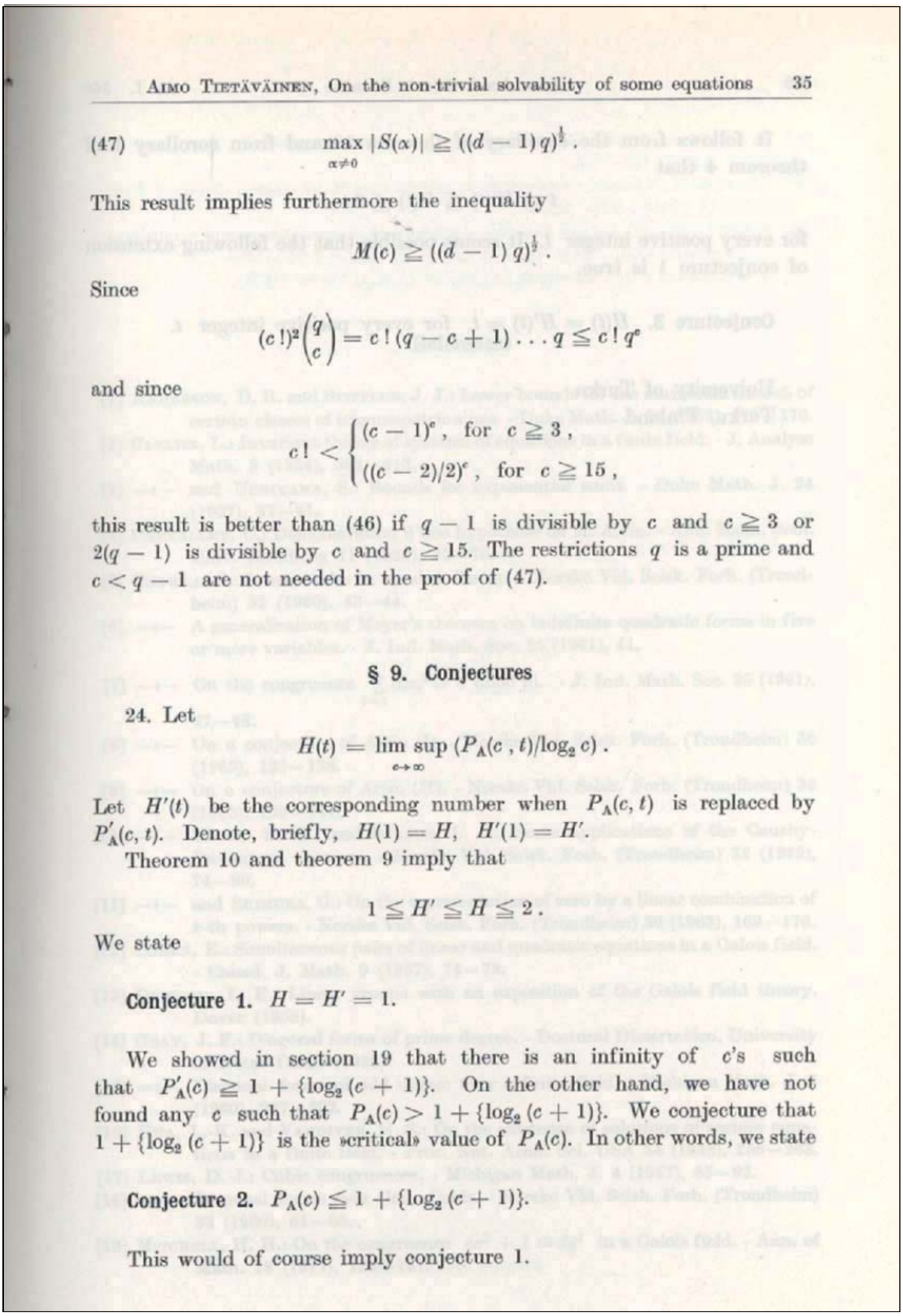
for $c \leqq p - 1$. On the other hand, it has been proved in [1] that if q is a prime and $c < q - 1$ then

$$M(c) > \left((c\,!)^2\binom{q}{c} - q^c\right)^{1/2c} . \tag{46}$$

We showed in the proof of lemma 7 that

$$\sum_{\alpha \neq 0} |S(\alpha)|^2 = (d-1)(q-1)q$$

where $S(\alpha)$ is defined by (33) and d is the g.c.d. of c and $q - 1$. It follows from this that



$$(47) \qquad \max_{\alpha \neq 0} |S(\alpha)| \geqq ((d-1)q)^{\frac{1}{2}} .$$

This result implies furthermore the inequality

$$\tilde{M}(c) \geqq ((d-1)\,q)^{\frac{1}{2}} .$$

Since

$$(c\,!)^2 \binom{q}{c} = c\,!\,(q-c+1)\ldots q \leqq c\,!\,q^c$$

and since

$$c\,! < \begin{cases} (c-1)^c , & \text{for } c \geqq 3 , \\ ((c-2)/2)^c , & \text{for } c \geqq 15 , \end{cases}$$

this result is better than (46) if $q-1$ is divisible by c and $c \geqq 3$ or $2(q-1)$ is divisible by c and $c \geqq 15$. The restrictions q is a prime and $c < q-1$ are not needed in the proof of (47).

§ 9. Conjectures

24. Let

$$H(t) = \limsup_{c \to \infty} \, (P_A(c\,,t)/\log_2 c)\,.$$

Let $H'(t)$ be the corresponding number when $P_A(c,t)$ is replaced by $P'_A(c,t)$. Denote, briefly, $H(1) = H$, $H'(1) = H'$.

Theorem 10 and theorem 9 imply that

$$1 \leqq H' \leqq H \leqq 2\,.$$

We state

Conjecture 1. $H = H' = 1$.

We showed in section 19 that there is an infinity of c's such that $P'_A(c) \geqq 1 + \{\log_2 (c+1)\}$. On the other hand, we have not found any c such that $P_A(c) > 1 + \{\log_2 (c+1)\}$. We conjecture that $1 + \{\log_2 (c+1)\}$ is the »critical» value of $P_A(c)$. In other words, we state

Conjecture 2. $P_A(c) \leqq 1 + \{\log_2 (c+1)\}$.

This would of course imply conjecture 1.

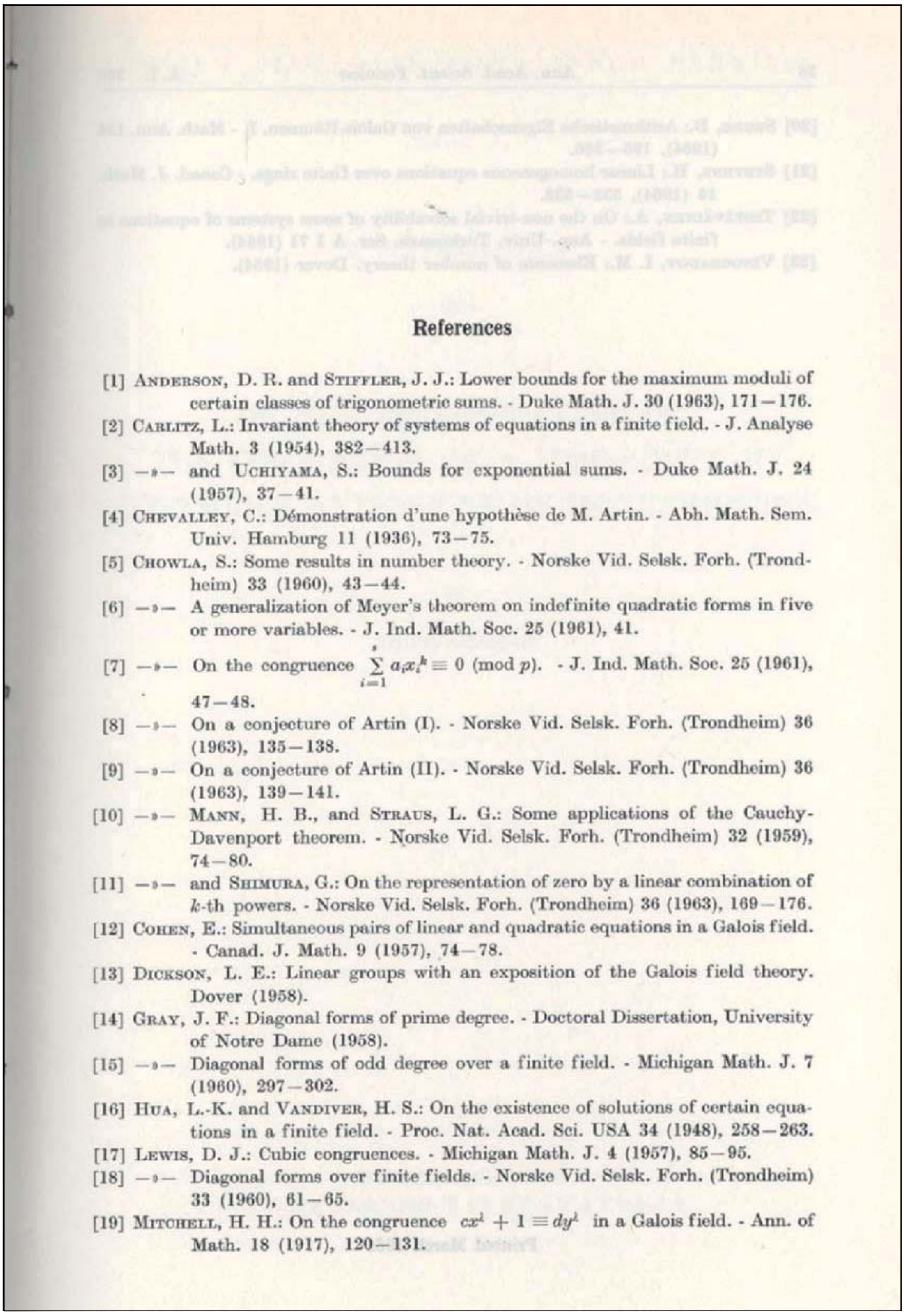
It follows from the corollary of theorem 10 and from corollary 1 of theorem 4 that

$$t \leqq H'(t) \leqq H(t) \leqq 2t\,,$$

for every positive integer t. It seems possible that the following extension of conjecture 1 is true.

Conjecture 3. $H(t) = H'(t) = t$, for every positive integer t.

University of Turku
Turku, Finland



References

[1] Anderson, D. R. and Stiffler, J. J.: Lower bounds for the maximum moduli of certain classes of trigonometric sums. - Duke Math. J. 30 (1963), 171—176.

[2] Carlitz, L.: Invariant theory of systems of equations in a finite field. - J. Analyse Math. 3 (1954), 382—413.

[3] —»— and Uchiyama, S.: Bounds for exponential sums. - Duke Math. J. 24 (1957), 37—41.

[4] Chevalley, C.: Démonstration d'une hypothèse de M. Artin. - Abh. Math. Sem. Univ. Hamburg 11 (1936), 73—75.

[5] Chowla, S.: Some results in number theory. - Norske Vid. Selsk. Forh. (Trondheim) 33 (1960), 43—44.

[6] —»— A generalization of Meyer's theorem on indefinite quadratic forms in five or more variables. - J. Ind. Math. Soc. 25 (1961), 41.

[7] —»— On the congruence $\sum_{i=1}^{s} a_i x_i^k \equiv 0 \pmod{p}$. - J. Ind. Math. Soc. 25 (1961), 47—48.

[8] —»— On a conjecture of Artin (I). - Norske Vid. Selsk. Forh. (Trondheim) 36 (1963), 135—138.

[9] —»— On a conjecture of Artin (II). - Norske Vid. Selsk. Forh. (Trondheim) 36 (1963), 139—141.

[10] —»— Mann, H. B., and Straus, L. G.: Some applications of the Cauchy-Davenport theorem. - Norske Vid. Selsk. Forh. (Trondheim) 32 (1959), 74—80.

[11] —»— and Shimura, G.: On the representation of zero by a linear combination of k-th powers. - Norske Vid. Selsk. Forh. (Trondheim) 36 (1963), 169—176.

[12] Cohen, E.: Simultaneous pairs of linear and quadratic equations in a Galois field. - Canad. J. Math. 9 (1957), 74—78.

[13] Dickson, L. E.: Linear groups with an exposition of the Galois field theory. Dover (1958).

[14] Gray, J. F.: Diagonal forms of prime degree. - Doctoral Dissertation, University of Notre Dame (1958).

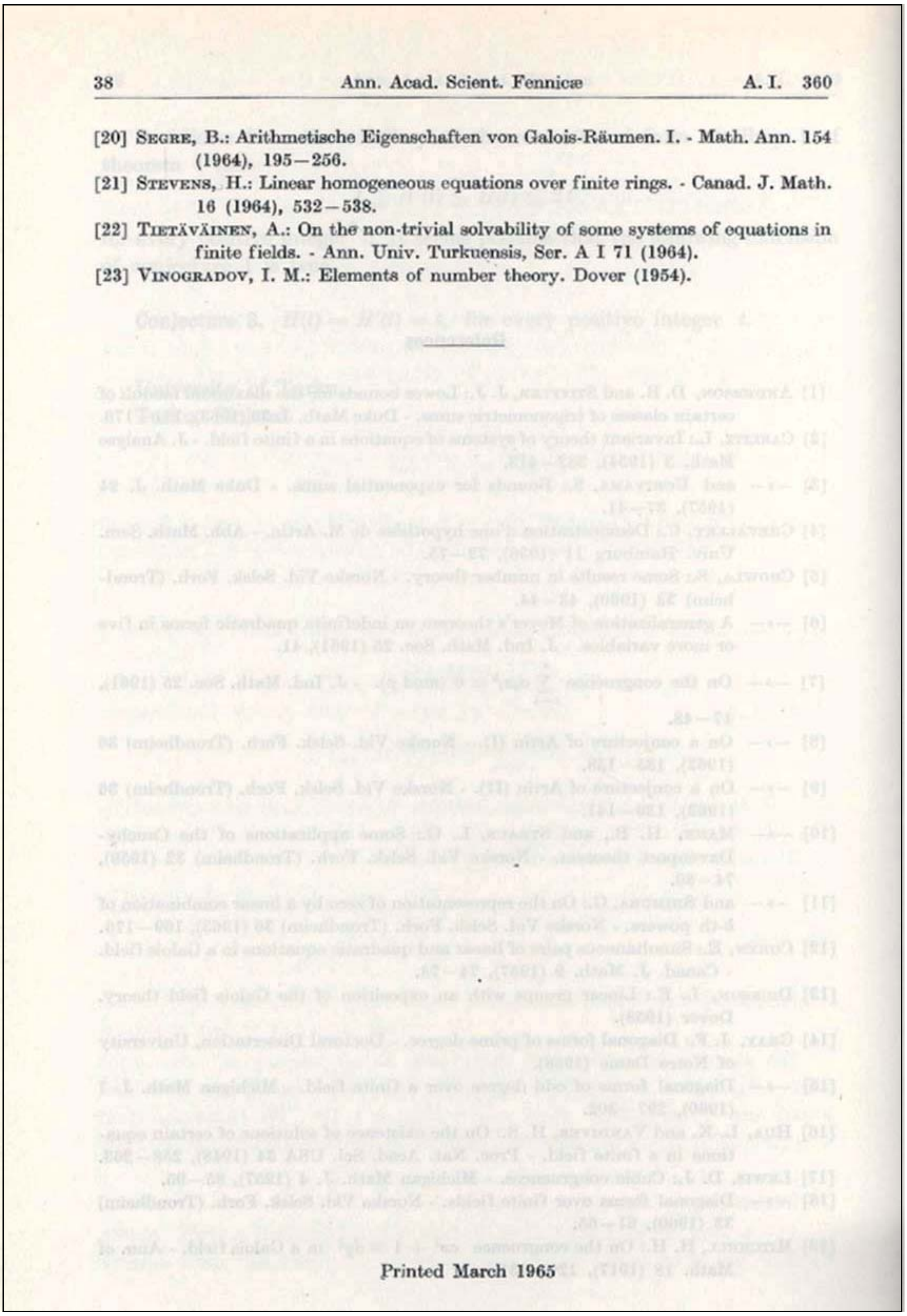
[15] —»— Diagonal forms of odd degree over a finite field. - Michigan Math. J. 7 (1960), 297—302.

[16] Hua, L.-K. and Vandiver, H. S.: On the existence of solutions of certain equations in a finite field. - Proc. Nat. Acad. Sci. USA 34 (1948), 258—263.

[17] Lewis, D. J.: Cubic congruences. - Michigan Math. J. 4 (1957), 85—95.

[18] —»— Diagonal forms over finite fields. - Norske Vid. Selsk. Forh. (Trondheim) 33 (1960), 61—65.

[19] Mitchell, H. H.: On the congruence $cx^l + 1 \equiv dy^l$ in a Galois field. - Ann. of Math. 18 (1917), 120—131.



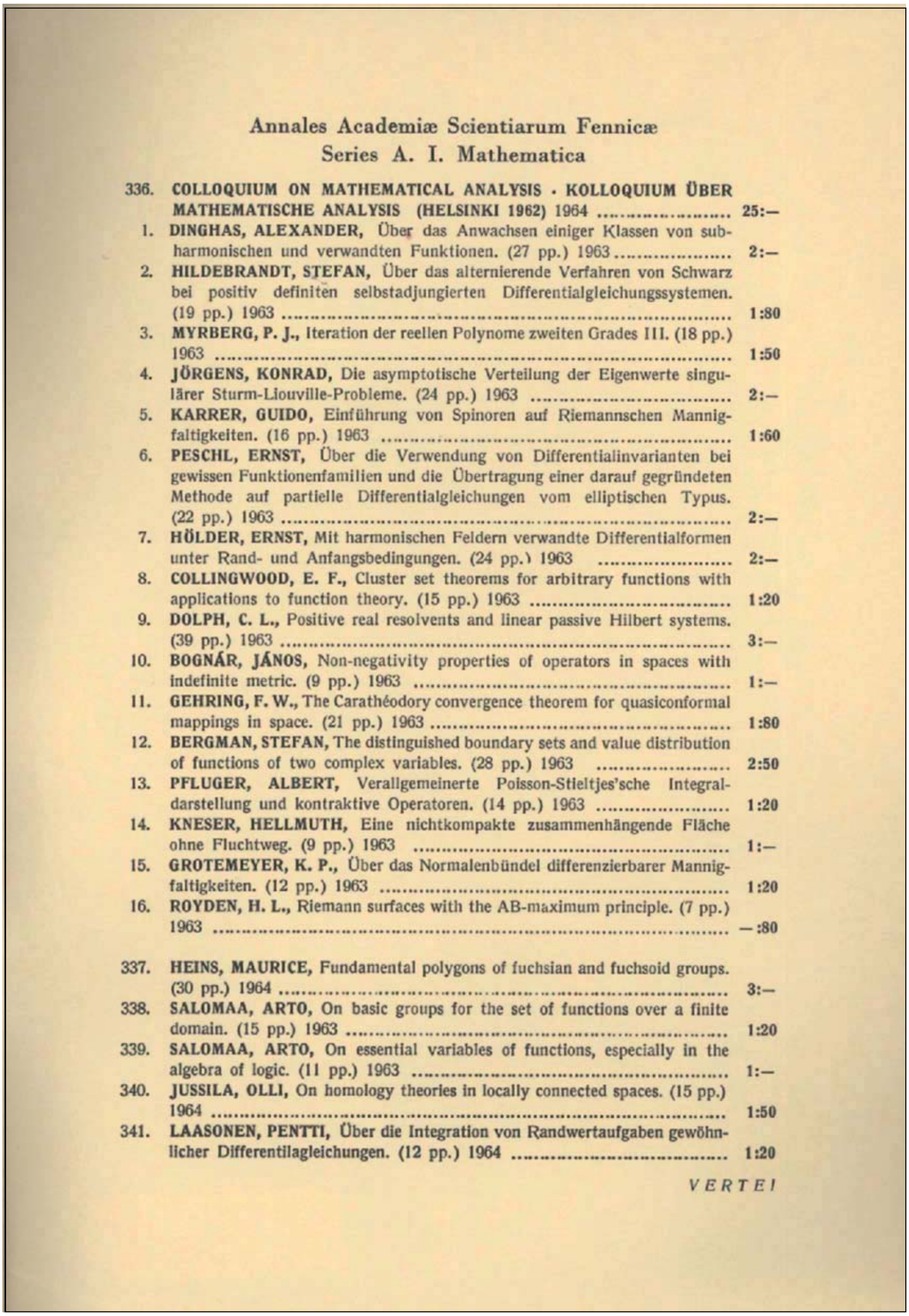
[20] Segre, B.: Arithmetische Eigenschaften von Galois-Räumen. I. - Math. Ann. 154 (1964), 195–256.

[21] Stevens, H.: Linear homogeneous equations over finite rings. - Canad. J. Math. 16 (1964), 532–538.

[22] Tietäväinen, A.: On the non-trivial solvability of some systems of equations in finite fields. - Ann. Univ. Turkuensis, Ser. A I 71 (1964).

[23] Vinogradov, I. M.: Elements of number theory. Dover (1954).

Printed March 1965



Annales Academiæ Scientiarum Fennicæ
Series A. I. Mathematica

336. **COLLOQUIUM ON MATHEMATICAL ANALYSIS · KOLLOQUIUM ÜBER MATHEMATISCHE ANALYSIS (HELSINKI 1962)** 1964 .... 25:—

1. **DINGHAS, ALEXANDER,** Über das Anwachsen einiger Klassen von subharmonischen und verwandten Funktionen. (27 pp.) 1963 .... 2:—
2. **HILDEBRANDT, STEFAN,** Über das alternierende Verfahren von Schwarz bei positiv definiten selbstadjungierten Differentialgleichungssystemen. (19 pp.) 1963 .... 1:80
3. **MYRBERG, P. J.,** Iteration der reellen Polynome zweiten Grades III. (18 pp.) 1963 .... 1:50
4. **JÖRGENS, KONRAD,** Die asymptotische Verteilung der Eigenwerte singulärer Sturm-Liouville-Probleme. (24 pp.) 1963 .... 2:—
5. **KARRER, GUIDO,** Einführung von Spinoren auf Riemannschen Mannigfaltigkeiten. (16 pp.) 1963 .... 1:60
6. **PESCHL, ERNST,** Über die Verwendung von Differentialinvarianten bei gewissen Funktionenfamilien und die Übertragung einer darauf gegründeten Methode auf partielle Differentialgleichungen vom elliptischen Typus. (22 pp.) 1963 .... 2:—
7. **HÖLDER, ERNST,** Mit harmonischen Feldern verwandte Differentialformen unter Rand- und Anfangsbedingungen. (24 pp.) 1963 .... 2:—
8. **COLLINGWOOD, E. F.,** Cluster set theorems for arbitrary functions with applications to function theory. (15 pp.) 1963 .... 1:20
9. **DOLPH, C. L.,** Positive real resolvents and linear passive Hilbert systems. (39 pp.) 1963 .... 3:—
10. **BOGNÁR, JÁNOS,** Non-negativity properties of operators in spaces with indefinite metric. (9 pp.) 1963 .... 1:—
11. **GEHRING, F. W.,** The Carathéodory convergence theorem for quasiconformal mappings in space. (21 pp.) 1963 .... 1:80
12. **BERGMAN, STEFAN,** The distinguished boundary sets and value distribution of functions of two complex variables. (28 pp.) 1963 .... 2:50
13. **PFLUGER, ALBERT,** Verallgemeinerte Poisson-Stieltjes'sche Integraldarstellung und kontraktive Operatoren. (14 pp.) 1963 .... 1:20
14. **KNESER, HELLMUTH,** Eine nichtkompakte zusammenhängende Fläche ohne Fluchtweg. (9 pp.) 1963 .... 1:—
15. **GROTEMEYER, K. P.,** Über das Normalenbündel differenzierbarer Mannigfaltigkeiten. (12 pp.) 1963 .... 1:20
16. **ROYDEN, H. L.,** Riemann surfaces with the AB-maximum principle. (7 pp.) 1963 .... —:80

337. **HEINS, MAURICE,** Fundamental polygons of fuchsian and fuchsoid groups. (30 pp.) 1964 .... 3:—

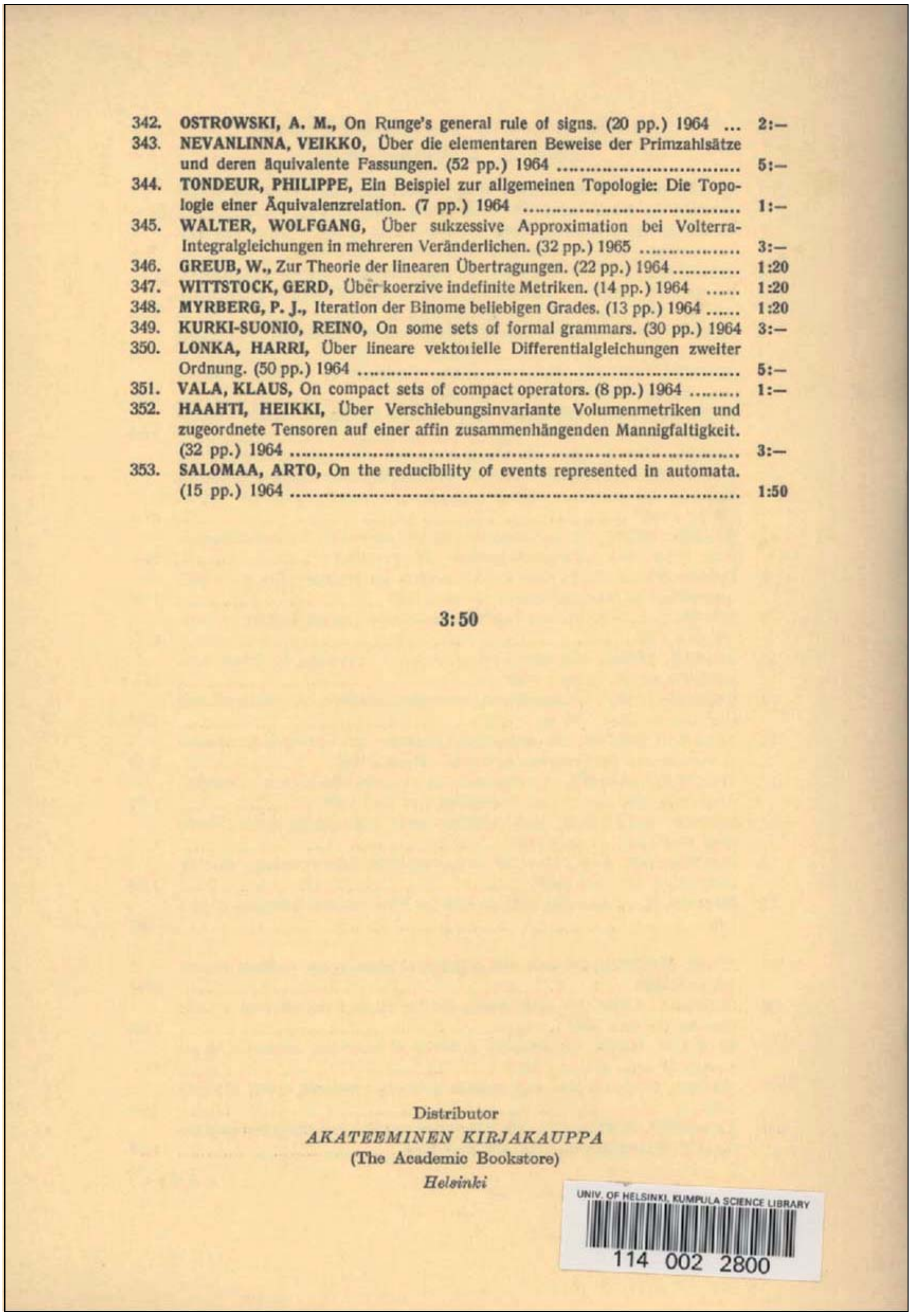
338. **SALOMAA, ARTO,** On basic groups for the set of functions over a finite domain. (15 pp.) 1963 .... 1:20

339. **SALOMAA, ARTO,** On essential variables of functions, especially in the algebra of logic. (11 pp.) 1963 .... 1:—

340. **JUSSILA, OLLI,** On homology theories in locally connected spaces. (15 pp.) 1964 .... 1:50

341. **LAASONEN, PENTTI,** Über die Integration von Randwertaufgaben gewöhnlicher Differentilagleichungen. (12 pp.) 1964 .... 1:20

VERTE!



342. **OSTROWSKI, A. M.,** On Runge's general rule of signs. (20 pp.) 1964 ... 2:—

343. **NEVANLINNA, VEIKKO,** Über die elementaren Beweise der Primzahlsätze und deren äquivalente Fassungen. (52 pp.) 1964 .... 5:—

344. **TONDEUR, PHILIPPE,** Ein Beispiel zur allgemeinen Topologie: Die Topologie einer Äquivalenzrelation. (7 pp.) 1964 .... 1:—

345. **WALTER, WOLFGANG,** Über sukzessive Approximation bei Volterra-Integralgleichungen in mehreren Veränderlichen. (32 pp.) 1965 .... 3:—

346. **GREUB, W.,** Zur Theorie der linearen Übertragungen. (22 pp.) 1964 .... 1:20

347. **WITTSTOCK, GERD,** Über koerzive indefinite Metriken. (14 pp.) 1964 ...... 1:20

348. **MYRBERG, P. J.,** Iteration der Binome beliebigen Grades. (13 pp.) 1964 ...... 1:20

349. **KURKI-SUONIO, REINO,** On some sets of formal grammars. (30 pp.) 1964 3:—

350. **LONKA, HARRI,** Über lineare vektoıielle Differentialgleichungen zweiter Ordnung. (50 pp.) 1964 .... 5:—

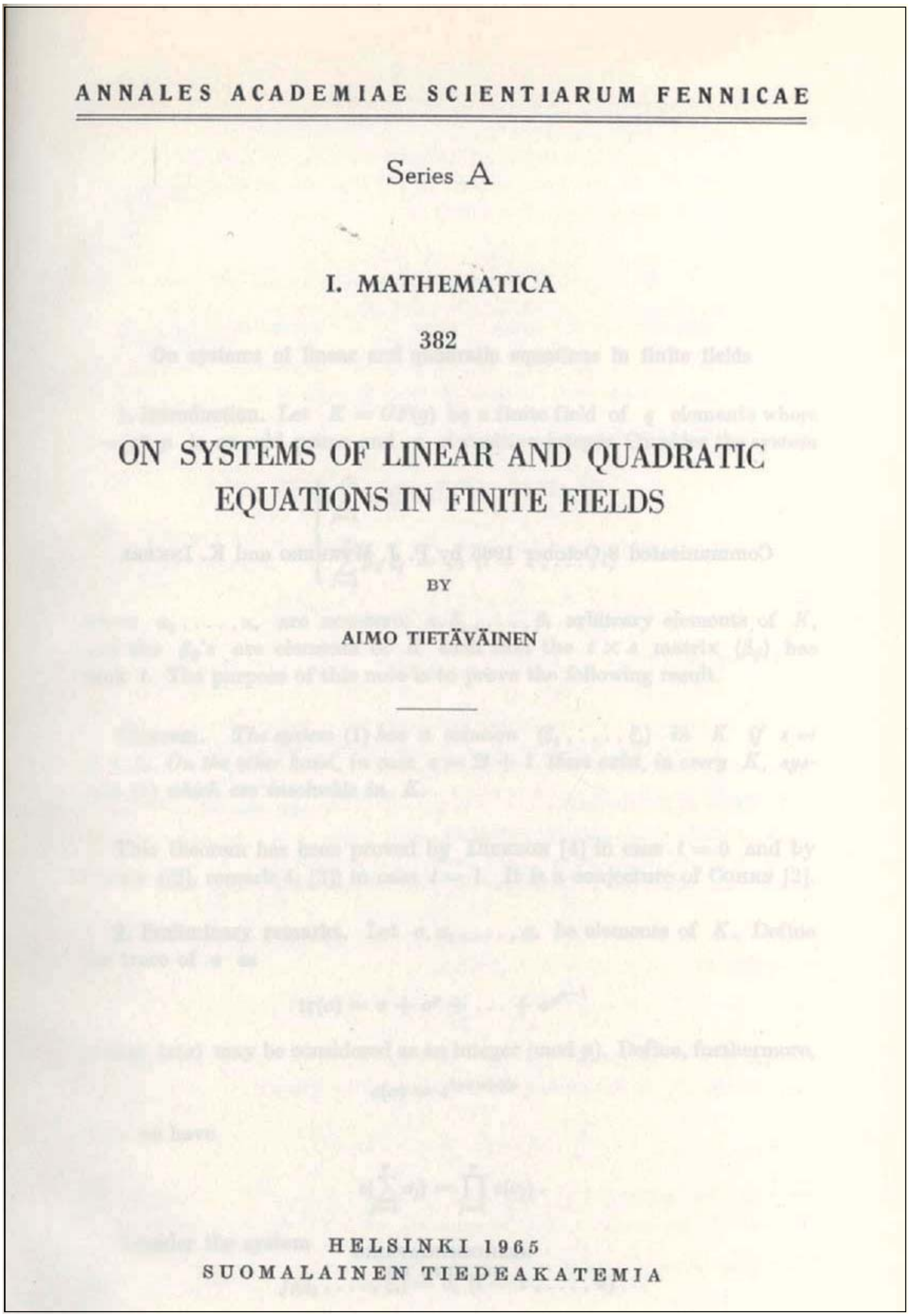
351. **VALA, KLAUS,** On compact sets of compact operators. (8 pp.) 1964 .... 1:—

352. **HAAHTI, HEIKKI,** Über Verschiebungsinvariante Volumenmetriken und zugeordnete Tensoren auf einer affin zusammenhängenden Mannigfaltigkeit. (32 pp.) 1964 .... 3:—

353. **SALOMAA, ARTO,** On the reducibility of events represented in automata. (15 pp.) 1964 .... 1:50

3:50

Distributor
AKATEEMINEN KIRJAKAUPPA
(The Academic Bookstore)
Helsinki



ANNALES ACADEMIAE SCIENTIARUM FENNICAE

Series A

I. MATHEMATICA

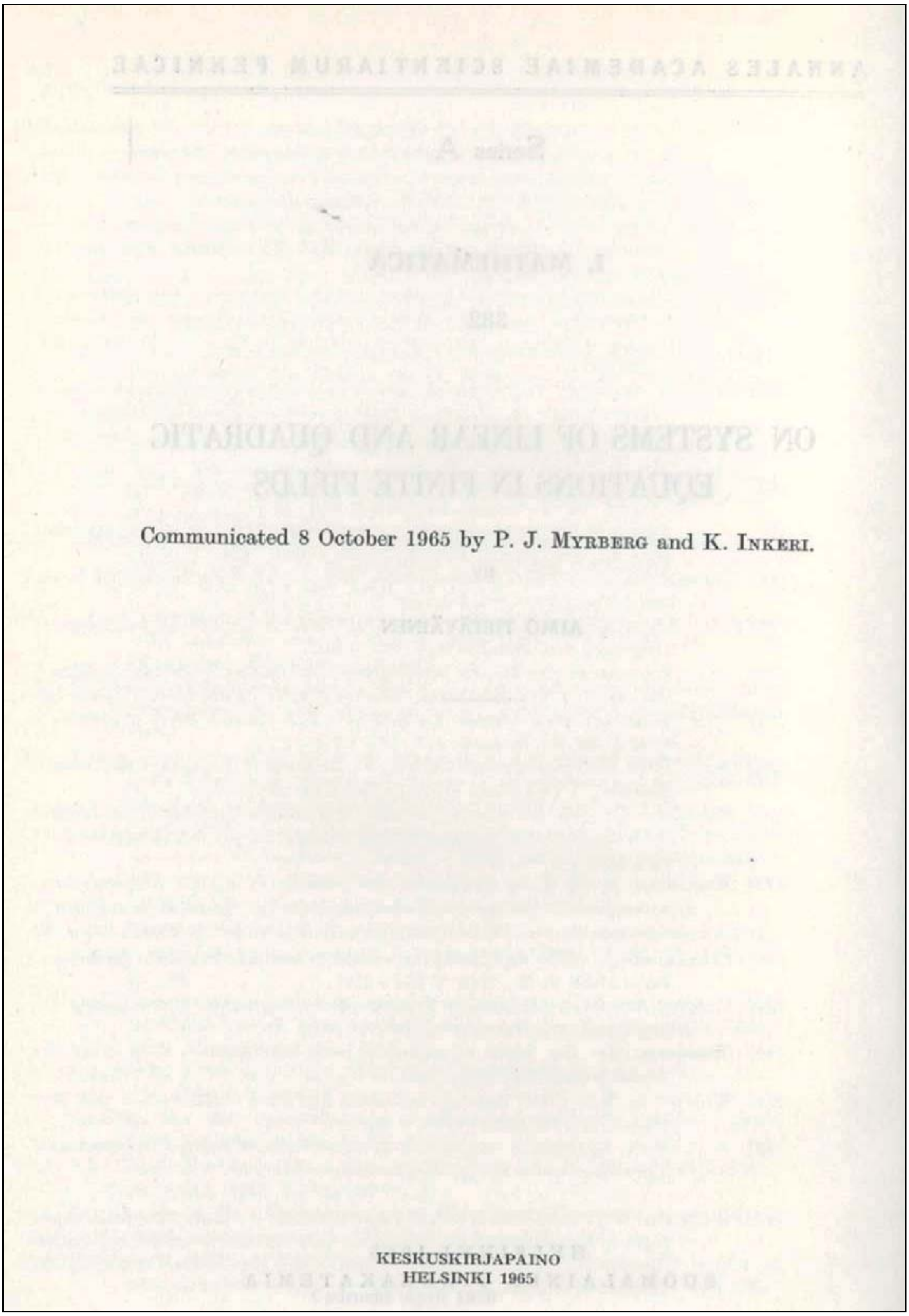
382

ON SYSTEMS OF LINEAR AND QUADRATIC EQUATIONS IN FINITE FIELDS

BY

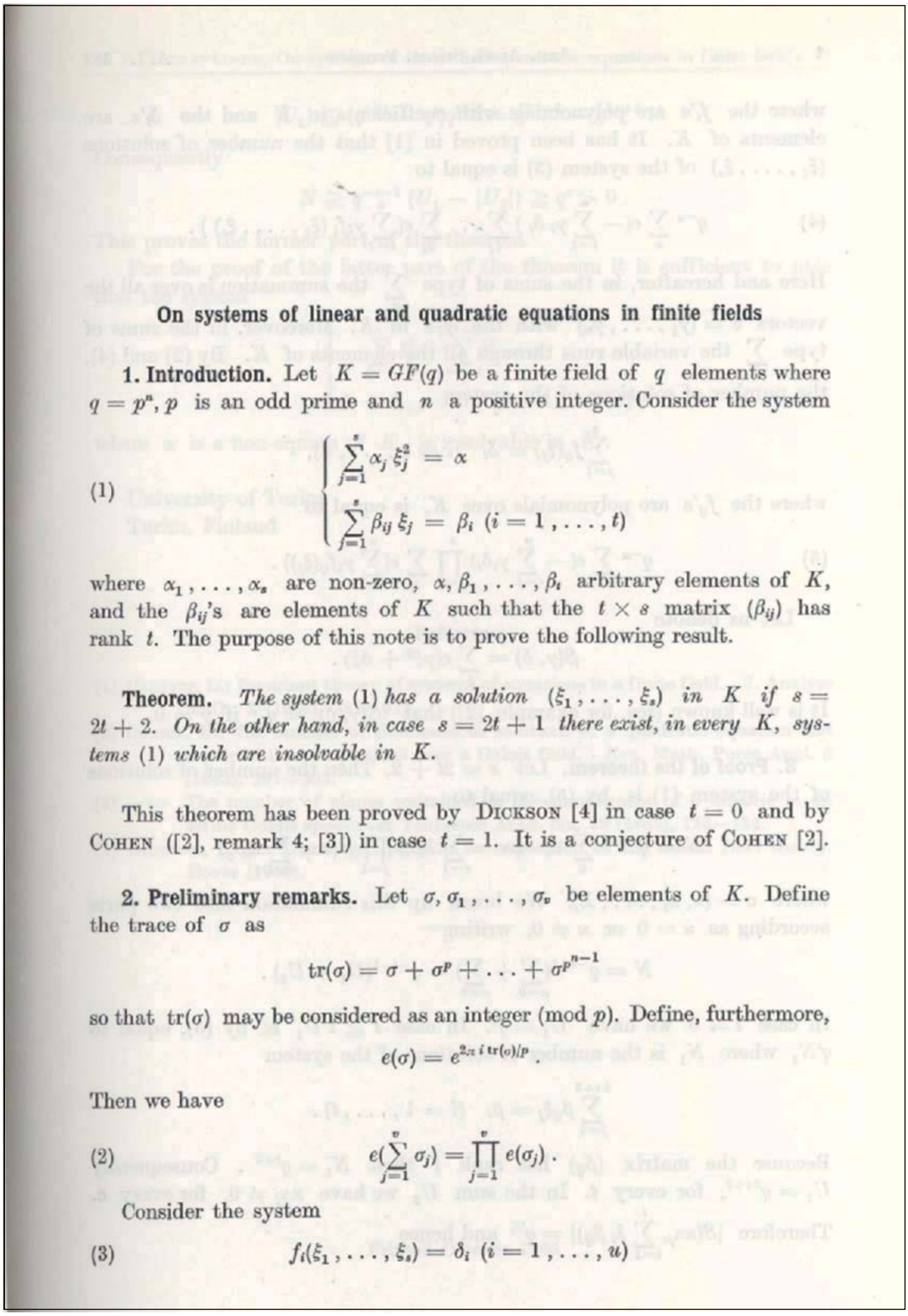
AIMO TIETÄVÄINEN

HELSINKI 1965
SUOMALAINEN TIEDEAKATEMIA



Communicated 8 October 1965 by P. J. Myrberg and K. Inkeri.

KESKUSKIRJAPAINO
HELSINKI 1965



On systems of linear and quadratic equations in finite fields

1. Introduction. Let $K = GF(q)$ be a finite field of q elements where $q = p^n$, p is an odd prime and n a positive integer. Consider the system

$$(1)\qquad \begin{cases} \sum_{j=1}^{s} \alpha_j \xi_j^2 = \alpha \\ \sum_{j=1}^{s} \beta_{ij} \xi_j = \beta_i \ (i = 1, \ldots, t) \end{cases}$$

where $\alpha_1, \ldots, \alpha_s$ are non-zero, $\alpha, \beta_1, \ldots, \beta_t$ arbitrary elements of K, and the β_{ij}'s are elements of K such that the $t \times s$ matrix (β_{ij}) has rank t. The purpose of this note is to prove the following result.

Theorem. *The system* (1) *has a solution* $(\xi_1, \ldots, \xi_s)$ *in* K *if* $s = 2t + 2$. *On the other hand, in case* $s = 2t + 1$ *there exist, in every* K, *systems* (1) *which are insolvable in* K.

This theorem has been proved by Dickson [4] in case $t = 0$ and by Cohen ([2], remark 4; [3]) in case $t = 1$. It is a conjecture of Cohen [2].

2. Preliminary remarks. Let $\sigma, \sigma_1, \ldots, \sigma_v$ be elements of K. Define the trace of σ as

$$\mathrm{tr}(\sigma) = \sigma + \sigma^p + \ldots + \sigma^{p^{n-1}}$$

so that $\mathrm{tr}(\sigma)$ may be considered as an integer (mod p). Define, furthermore,

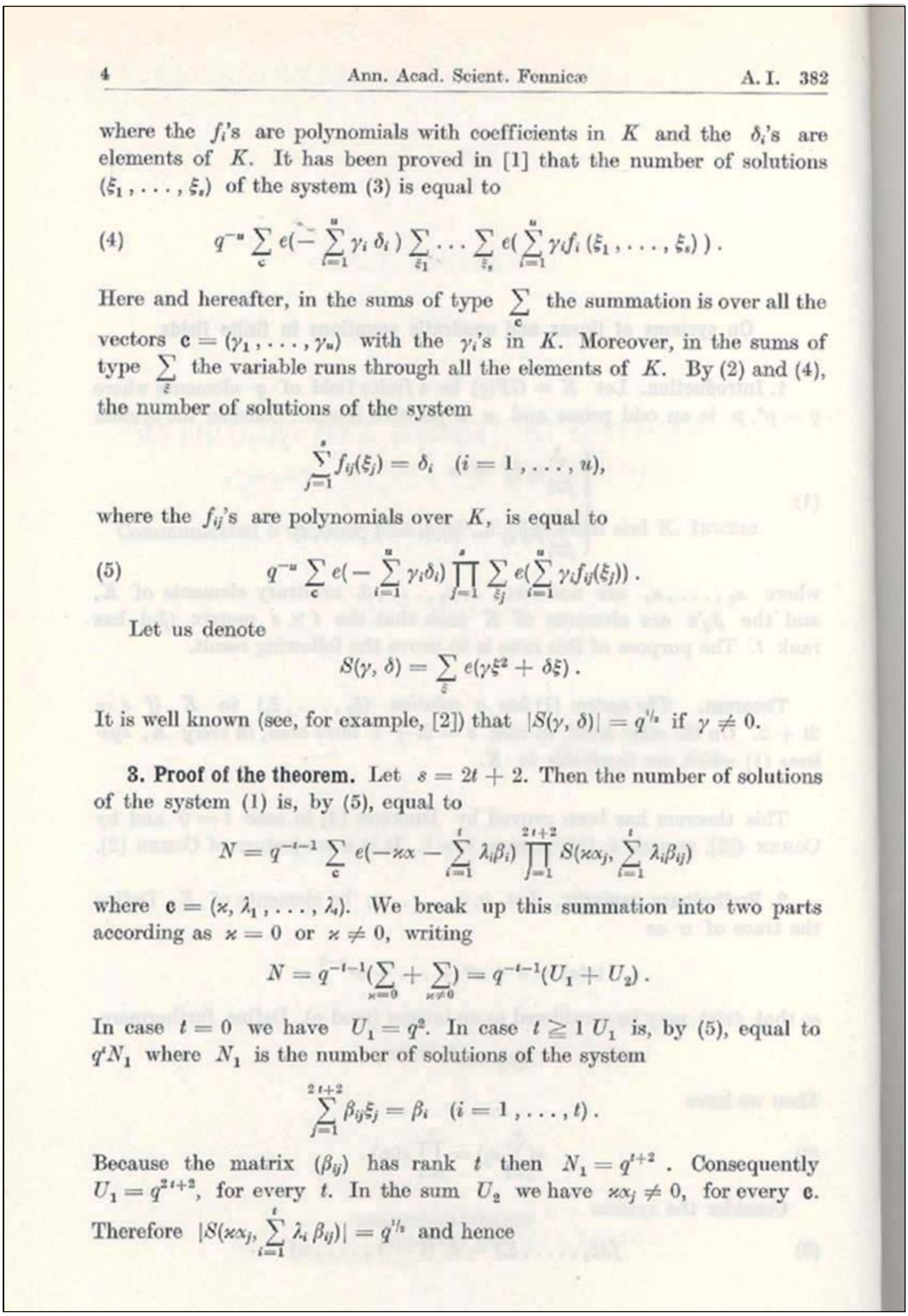
$$e(\sigma) = e^{2\pi i\, \mathrm{tr}(\sigma)/p} .$$

Then we have

$$(2)\qquad e(\sum_{j=1}^{v} \sigma_j) = \prod_{j=1}^{v} e(\sigma_j) .$$

Consider the system

$$(3)\qquad f_i(\xi_1, \ldots, \xi_s) = \delta_i \ (i = 1, \ldots, u)$$



where the f_i's are polynomials with coefficients in K and the δ_i's are elements of K. It has been proved in [1] that the number of solutions $(\xi_1, \ldots, \xi_s)$ of the system (3) is equal to

$$(4) \qquad q^{-u} \sum_{\mathfrak{c}} e(-\sum_{i=1}^{u} \gamma_i \delta_i) \sum_{\xi_1} \cdots \sum_{\xi_s} e(\sum_{i=1}^{u} \gamma_i f_i(\xi_1, \ldots, \xi_s)) .$$

Here and hereafter, in the sums of type $\sum_{\mathfrak{c}}$ the summation is over all the vectors $\mathfrak{c} = (\gamma_1, \ldots, \gamma_u)$ with the γ_i's in K. Moreover, in the sums of type $\sum_{\xi}$ the variable runs through all the elements of K. By (2) and (4), the number of solutions of the system

$$\sum_{j=1}^{s} f_{ij}(\xi_j) = \delta_i \quad (i = 1, \ldots, u),$$

where the f_{ij}'s are polynomials over K, is equal to

$$(5) \qquad q^{-u} \sum_{\mathfrak{c}} e(-\sum_{i=1}^{u} \gamma_i \delta_i) \prod_{j=1}^{s} \sum_{\xi_j} e(\sum_{i=1}^{u} \gamma_i f_{ij}(\xi_j)) .$$

Let us denote

$$S(\gamma, \delta) = \sum_{\xi} e(\gamma \xi^2 + \delta \xi) .$$

It is well known (see, for example, [2]) that $|S(\gamma, \delta)| = q^{1/2}$ if $\gamma \neq 0$.

3. Proof of the theorem. Let $s = 2t + 2$. Then the number of solutions of the system (1) is, by (5), equal to

$$N = q^{-t-1} \sum_{\mathfrak{c}} e(-\varkappa\alpha - \sum_{i=1}^{t} \lambda_i \beta_i) \prod_{j=1}^{2t+2} S(\varkappa\alpha_j, \sum_{i=1}^{t} \lambda_i \beta_{ij})$$

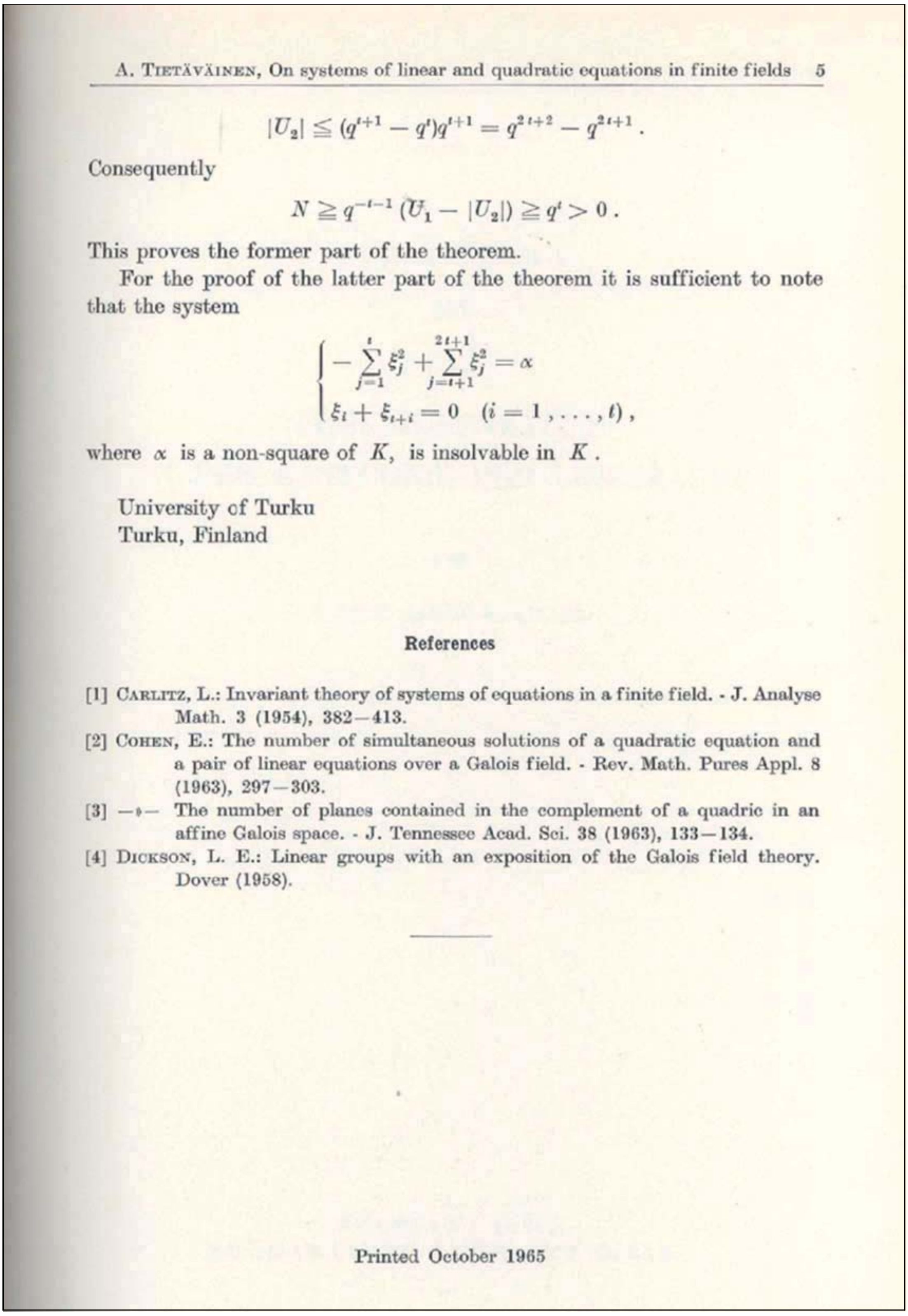
where $\mathfrak{c} = (\varkappa, \lambda_1, \ldots, \lambda_t)$. We break up this summation into two parts according as $\varkappa = 0$ or $\varkappa \neq 0$, writing

$$N = q^{-t-1}(\sum_{\varkappa=0} + \sum_{\varkappa\neq0}) = q^{-t-1}(U_1 + U_2) .$$

In case $t = 0$ we have $U_1 = q^2$. In case $t \geqq 1$ U_1 is, by (5), equal to $q^t N_1$ where N_1 is the number of solutions of the system

$$\sum_{j=1}^{2t+2} \beta_{ij} \xi_j = \beta_i \quad (i = 1, \ldots, t) .$$

Because the matrix (β_{ij}) has rank t then $N_1 = q^{t+2}$. Consequently $U_1 = q^{2t+2}$, for every t. In the sum U_2 we have $\varkappa\alpha_j \neq 0$, for every $\mathfrak{c}$. Therefore $|S(\varkappa\alpha_j, \sum_{i=1}^{t} \lambda_i \beta_{ij})| = q^{1/2}$ and hence



$$|U_2| \leqq (q^{t+1} - q^t)q^{t+1} = q^{2t+2} - q^{2t+1} .$$

Consequently

$$N \geqq q^{-t-1} (\tilde{U}_1 - |U_2|) \geqq q^t > 0 .$$

This proves the former part of the theorem.

For the proof of the latter part of the theorem it is sufficient to note that the system

$$\begin{cases} - \sum_{j=1}^{t} \xi_j^2 + \sum_{j=t+1}^{2t+1} \xi_j^2 = \alpha \\ \xi_i + \xi_{t+i} = 0 \quad (i = 1 , \ldots , t) , \end{cases}$$

where α is a non-square of K, is insolvable in K.

University of Turku
Turku, Finland

References

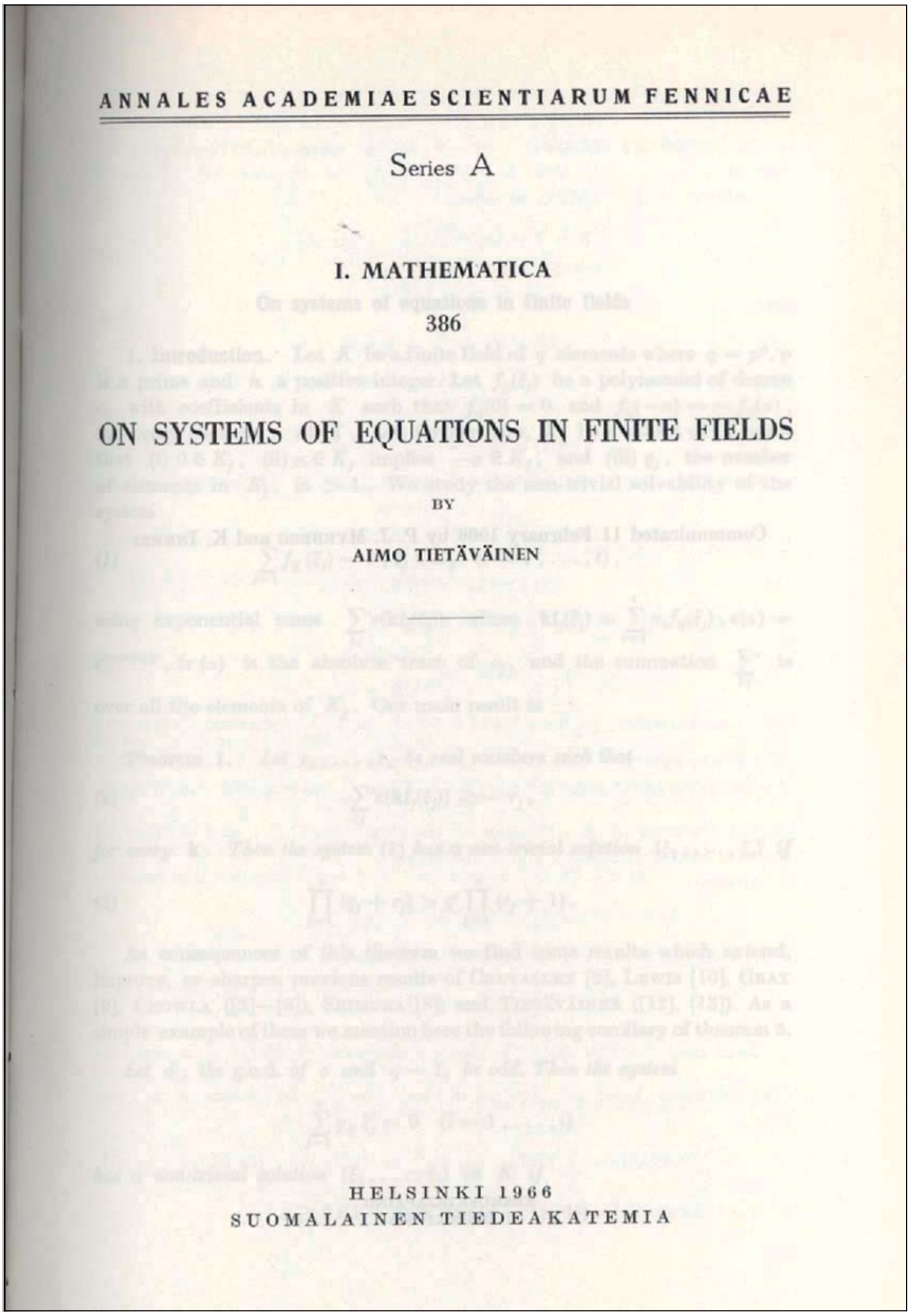
[1] CARLITZ, L.: Invariant theory of systems of equations in a finite field. - J. Analyse Math. 3 (1954), 382–413.

[2] COHEN, E.: The number of simultaneous solutions of a quadratic equation and a pair of linear equations over a Galois field. - Rev. Math. Pures Appl. 8 (1963), 297–303.

[3] —»— The number of planes contained in the complement of a quadric in an affine Galois space. - J. Tennessee Acad. Sci. 38 (1963), 133–134.

[4] DICKSON, L. E.: Linear groups with an exposition of the Galois field theory. Dover (1958).

Printed October 1965



ANNALES ACADEMIAE SCIENTIARUM FENNICAE

Series A

I. MATHEMATICA

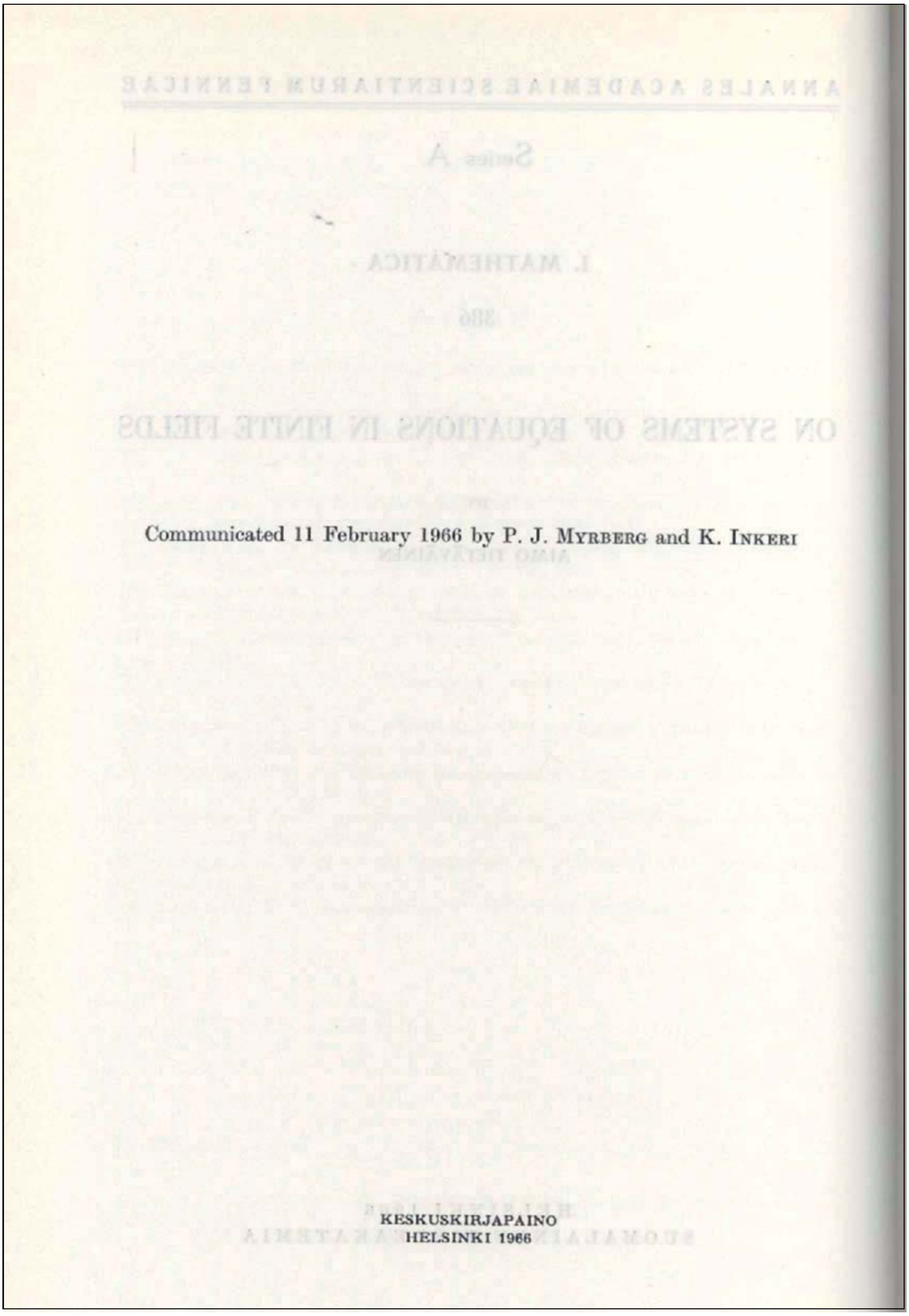
386

ON SYSTEMS OF EQUATIONS IN FINITE FIELDS

BY

AIMO TIETÄVÄINEN

HELSINKI 1966
SUOMALAINEN TIEDEAKATEMIA



Communicated 11 February 1966 by P. J. Myrberg and K. Inkeri

KESKUSKIRJAPAINO
HELSINKI 1966

On systems of equations in finite fields

1. Introduction. Let K be a finite field of q elements where $q = p^n$, p is a prime and n a positive integer. Let $f_{ij}(\xi_j)$ be a polynomial of degree c_{ij} with coefficients in K such that $f_{ij}(0) = 0$ and $f_{ij}(-\alpha) = -f_{ij}(\alpha)$, for every element α of K. Let, furthermore, K_j be a subset of K such that (i) $0 \in K_j$, (ii) $\alpha \in K_j$ implies $-\alpha \in K_j$, and (iii) q_j, the number of elements in K_j, is > 1. We study the non-trivial solvability of the system

$$\sum_{j=1}^{s} f_{ij}(\xi_j) = 0\,,\ \xi_j \in K_j \quad (i = 1, \ldots, t)\,, \tag{1}$$

using exponential sums $\sum_{\xi_j}{}' e(\mathbf{kf}_j(\xi_j))$ where $\mathbf{kf}_j(\xi_j) = \sum_{i=1}^{t} \varkappa_i f_{ij}(\xi_j)$, $e(\alpha) = e^{2\pi i \operatorname{tr}(\alpha)/p}$, $\operatorname{tr}(\alpha)$ is the absolute trace of α, and the summation $\sum_{\xi_j}{}'$ is over all the elements of K_j. Our main result is

Theorem 1. *Let $r_1, \ldots, r_s$ be real numbers such that*

$$\sum_{\xi_j}{}' e(\mathbf{kf}_j(\xi_j)) \geqq -r_j\,, \tag{2}$$

for every $\mathbf{k}$. *Then the system* (1) *has a non-trivial solution* $(\xi_1, \ldots, \xi_s)$ *if*

$$\prod_{j=1}^{s} (q_j + r_j) > q^t \prod_{j=1}^{s} (r_j + 1)\,. \tag{3}$$

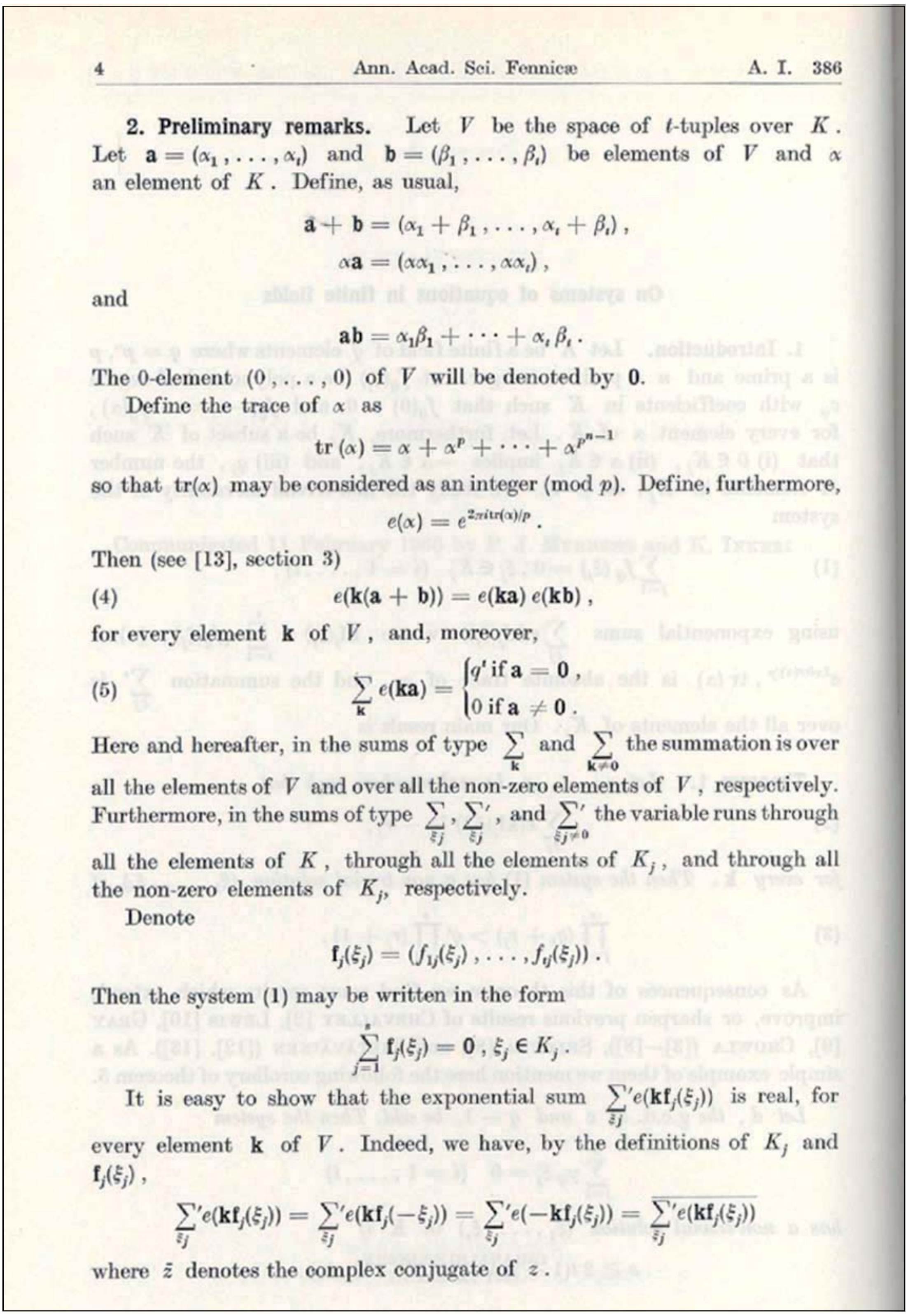
As consequences of this theorem we find some results which extend, improve, or sharpen previous results of Chevalley [2], Lewis [10], Gray [9], Chowla ([3]—[8]), Shimura [8], and Tietäväinen ([12], [13]). As a simple example of them we mention here the following corollary of theorem 5.

Let d, the g.c.d. *of c and $q - 1$, be odd. Then the system*

$$\sum_{j=1}^{s} \gamma_{ij}\, \xi_j^c = 0 \quad (i = 1, \ldots, t)$$

has a non-trivial solution $(\xi_1, \ldots, \xi_s)$ in K if

$$s \geqq 2\,t(1 + \max(\log_2(d-1), 1))\,.$$



2. Preliminary remarks. Let V be the space of t-tuples over K. Let $\mathbf{a} = (\alpha_1, \ldots, \alpha_t)$ and $\mathbf{b} = (\beta_1, \ldots, \beta_t)$ be elements of V and α an element of K. Define, as usual,

$$\mathbf{a} + \mathbf{b} = (\alpha_1 + \beta_1, \ldots, \alpha_t + \beta_t),$$

$$\alpha\mathbf{a} = (\alpha\alpha_1, \ldots, \alpha\alpha_t),$$

and

$$\mathbf{ab} = \alpha_1\beta_1 + \cdots + \alpha_t\beta_t .$$

The 0-element $(0, \ldots, 0)$ of V will be denoted by $\mathbf{0}$.

Define the trace of α as

$$\mathrm{tr}\,(\alpha) = \alpha + \alpha^p + \cdots + \alpha^{p^{n-1}}$$

so that $\mathrm{tr}(\alpha)$ may be considered as an integer (mod p). Define, furthermore,

$$e(\alpha) = e^{2\pi i \mathrm{tr}(\alpha)/p} .$$

Then (see [13], section 3)

$$e(\mathbf{k}(\mathbf{a} + \mathbf{b})) = e(\mathbf{ka})\, e(\mathbf{kb}), \tag{4}$$

for every element $\mathbf{k}$ of V, and, moreover,

$$\sum_{\mathbf{k}} e(\mathbf{ka}) = \begin{cases} q^t \text{ if } \mathbf{a} = \mathbf{0}, \\ 0 \text{ if } \mathbf{a} \neq \mathbf{0}. \end{cases} \tag{5}$$

Here and hereafter, in the sums of type $\sum\limits_{\mathbf{k}}$ and $\sum\limits_{\mathbf{k} \neq \mathbf{0}}$ the summation is over all the elements of V and over all the non-zero elements of V, respectively. Furthermore, in the sums of type $\sum\limits_{\xi_j}$, $\sum\limits_{\xi_j}{}'$, and $\sum\limits_{\xi_j \neq 0}{}'$ the variable runs through all the elements of K, through all the elements of K_j, and through all the non-zero elements of K_j, respectively.

Denote

$$\mathbf{f}_j(\xi_j) = (f_{1j}(\xi_j), \ldots, f_{tj}(\xi_j)) .$$

Then the system (1) may be written in the form

$$\sum_{j=1}^{s} \mathbf{f}_j(\xi_j) = \mathbf{0}, \ \xi_j \in K_j .$$

It is easy to show that the exponential sum $\sum\limits_{\xi_j}{}' e(\mathbf{kf}_j(\xi_j))$ is real, for every element $\mathbf{k}$ of V. Indeed, we have, by the definitions of K_j and $\mathbf{f}_j(\xi_j)$,

$$\sum_{\xi_j}{}' e(\mathbf{kf}_j(\xi_j)) = \sum_{\xi_j}{}' e(\mathbf{kf}_j(-\xi_j)) = \sum_{\xi_j}{}' e(-\mathbf{kf}_j(\xi_j)) = \overline{\sum_{\xi_j}{}' e(\mathbf{kf}_j(\xi_j))}$$

where $\bar{z}$ denotes the complex conjugate of z.

3. Proof of theorem 1. Let

$$J = J(\xi_1, \ldots, \xi_s) = \{j \in \{1, \ldots, s\} \mid \xi_j = 0\},$$

$$A(\xi_1, \ldots, \xi_s) = \begin{cases} 1 \text{ if } \xi_j \neq 0\,, \text{ for every } j\,, \\ \prod_{j \in J} (r_j + 1) \text{ otherwise,} \end{cases} \tag{6}$$

and

$$B(\xi_1, \ldots, \xi_s) = \begin{cases} A(\xi_1, \ldots, \xi_s) \text{ if } \sum_{j=1}^{s} \mathbf{f}_j(\xi_j) = \mathbf{0}\,, \\ 0 \text{ otherwise.} \end{cases} \tag{7}$$

Let, furthermore,

$$M = \sum_{\xi_1}{}' \cdots \sum_{\xi_s}{}' B(\xi_1, \ldots, \xi_s)\,. \tag{8}$$

Then (1) has a non-trivial solution if $M > \prod_{j=1}^{s} (r_j + 1)$.

We have, by (5), (7), (8), and (4),

$$\begin{aligned} q^t M &= \sum_{\xi_1}{}' \cdots \sum_{\xi_s}{}' A(\xi_1, \ldots, \xi_s) \sum_{\mathbf{k}} e(\mathbf{k} \sum_{j=1}^{s} \mathbf{f}_j(\xi_j)) \\ &= \sum_{\mathbf{k}} \sum_{\xi_1}{}' \cdots \sum_{\xi_s}{}' A(\xi_1, \ldots, \xi_s) \prod_{j=1}^{s} e(\mathbf{k}\mathbf{f}_j(\xi_j))\,. \end{aligned} \tag{9}$$

It can be shown, by induction, that

$$\sum_{\xi_1}{}' \cdots \sum_{\xi_s}{}' A(\xi_1, \ldots, \xi_s) \prod_{j=1}^{s} e(\mathbf{k}\mathbf{f}_j(\xi_j)) = \prod_{j=1}^{s} (r_j + \sum_{\xi_j}{}' e(\mathbf{k}\mathbf{f}_j(\xi_j)))\,. \tag{10}$$

Indeed, it is easy to see that the statement (10) is true for $s = 1$, and we assume it to be true for $s - 1$ variables ξ_j. Since, by (6),

$$A(\xi_1, \ldots, \xi_s) = \begin{cases} (r_s + 1)\, A(\xi_1, \ldots, \xi_{s-1}) \text{ if } \xi_s = 0\,, \\ A(\xi_1, \ldots, \xi_{s-1}) \text{ if } \xi_s \neq 0\,, \end{cases}$$

then the left side of (10) equals

$$(r_s + 1) \sum_{\xi_1}{}' \cdots \sum_{\xi_{s-1}}{}' A(\xi_1, \ldots, \xi_{s-1}) \prod_{j=1}^{s-1} e(\mathbf{k}\mathbf{f}_j(\xi_j)) +$$

$$\sum_{\xi_1}{}' \cdots \sum_{\xi_{s-1}}{}' A(\xi_1, \ldots, \xi_{s-1}) \prod_{j=1}^{s-1} e(\mathbf{k}\mathbf{f}_j(\xi_j)) \sum_{\xi_s \neq 0}{}' e(\mathbf{k}\mathbf{f}_s(\xi_s))\,.$$

Using the equation

$$\sum_{\xi_s \neq 0}{}' e(\mathbf{k}\mathbf{f}_s(\xi_s)) = \sum_{\xi_s}{}' e(\mathbf{k}\mathbf{f}_s(\xi_s)) - 1$$

and the induction hypothesis, we find that this is, moreover, equal to

$$(r_s + \sum_{\xi_s}{}' e(\mathbf{kf}_s(\xi_s)) \sum_{\xi_1}{}' \cdots \sum_{\xi_{s-1}}{}' A(\xi_1, \ldots, \xi_{s-1}) \prod_{j=1}^{s-1} e(\mathbf{kf}_j(\xi_j))$$

$$= \prod_{j=1}^{s} (r_j + \sum_{\xi_j}{}' e(\mathbf{kf}_j(\xi_j))) .$$

Thus we have proved the equation (10).

Using (9) and (10), we get

$$q^t M = \sum_{\mathbf{k}} \prod_{j=1}^{s} (r_j + \sum_{\xi_j}{}' e(\mathbf{kf}_j(\xi_j)))$$

$$= \prod_{j=1}^{s} (q_j + r_j) + \sum_{\mathbf{k} \neq 0} \prod_{j=1}^{s} (r_j + \sum_{\xi_j}{}' e(\mathbf{kf}_j(\xi_j))) .$$

We have hence, by (2) and (3).

$$M \geqq q^{-t} \prod_{j=1}^{s} (q_j + r_j) > \prod_{j=1}^{s} (r_j + 1)$$

which is the required inequality.

4. Consequences of theorem 1. Since $e(\mathbf{kf}_j(0)) = e(0) = 1$ then $\sum_{\xi_j}{}' e(\mathbf{kf}_j(\xi_j)) \geqq 2 - q_j$. Therefore we may take $r_j = q_j - 2$ in theorem 1. Then

$$\prod_{j=1}^{s} (q_j + r_j) = 2^s \prod_{j=1}^{s} (q_j - 1) = 2^s \prod_{j=1}^{s} (r_j + 1) .$$

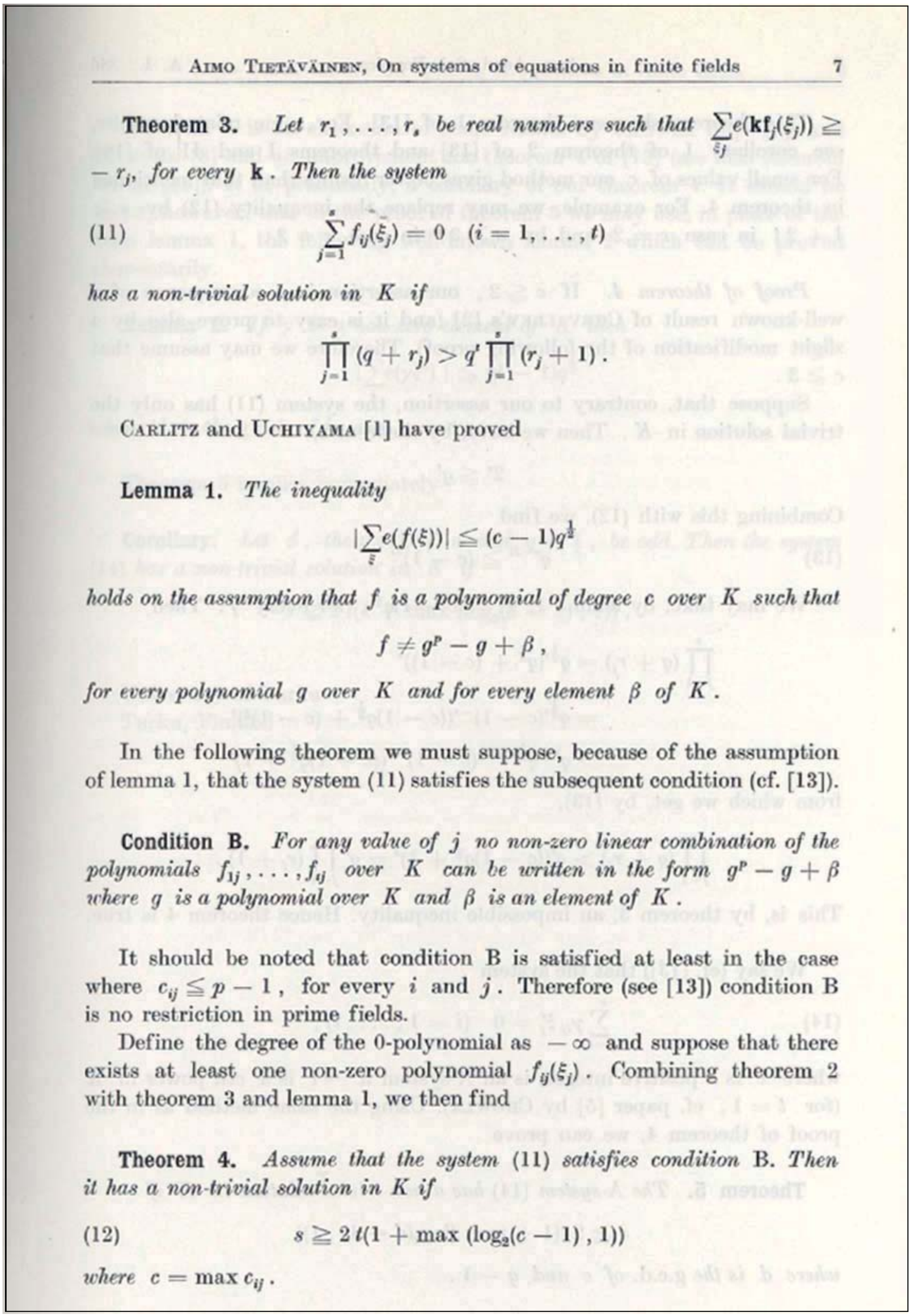
Consequently we have the following corollary of theorem 1.

Theorem 2. *The system* (1) *has a non-trivial solution if*

$$2^s > q^t .$$

This theorem is an extension of a result of CHOWLA's (see, for example, [5]) .For some related theorems, see [11], theorem 1, and [13], lemma 3. Theorem 2 can be proved also by using CHOWLA's method but it is interesting to see that all the theorems 1—5 can be proved by using exponential sum methods only.

If we put $K_1 = \cdots = K_s = K$, we get immediately, by theorem 1, the following result.



Theorem 3. *Let* $r_1, \ldots, r_s$ *be real numbers such that* $\sum_{\xi_j} e(\mathbf{k}\mathfrak{f}_j(\xi_j)) \geqq - r_j$, *for every* $\mathbf{k}$. *Then the system*

$$\sum_{j=1}^{s} f_{ij}(\xi_j) = 0 \quad (i = 1, \ldots, t) \tag{11}$$

has a non-trivial solution in K *if*

$$\prod_{j=1}^{s} (q + r_j) > q^t \prod_{j=1}^{s} (r_j + 1) .$$

CARLITZ and UCHIYAMA [1] have proved

Lemma 1. *The inequality*

$$|\sum_{\xi} e(f(\xi))| \leqq (c - 1)q^{\frac{1}{2}}$$

holds on the assumption that f *is a polynomial of degree* c *over* K *such that*

$$f \neq g^p - g + \beta ,$$

for every polynomial g *over* K *and for every element* β *of* K.

In the following theorem we must suppose, because of the assumption of lemma 1, that the system (11) satisfies the subsequent condition (cf. [13]).

Condition B. *For any value of* j *no non-zero linear combination of the polynomials* $f_{1j}, \ldots, f_{tj}$ *over* K *can be written in the form* $g^p - g + \beta$ *where* g *is a polynomial over* K *and* β *is an element of* K.

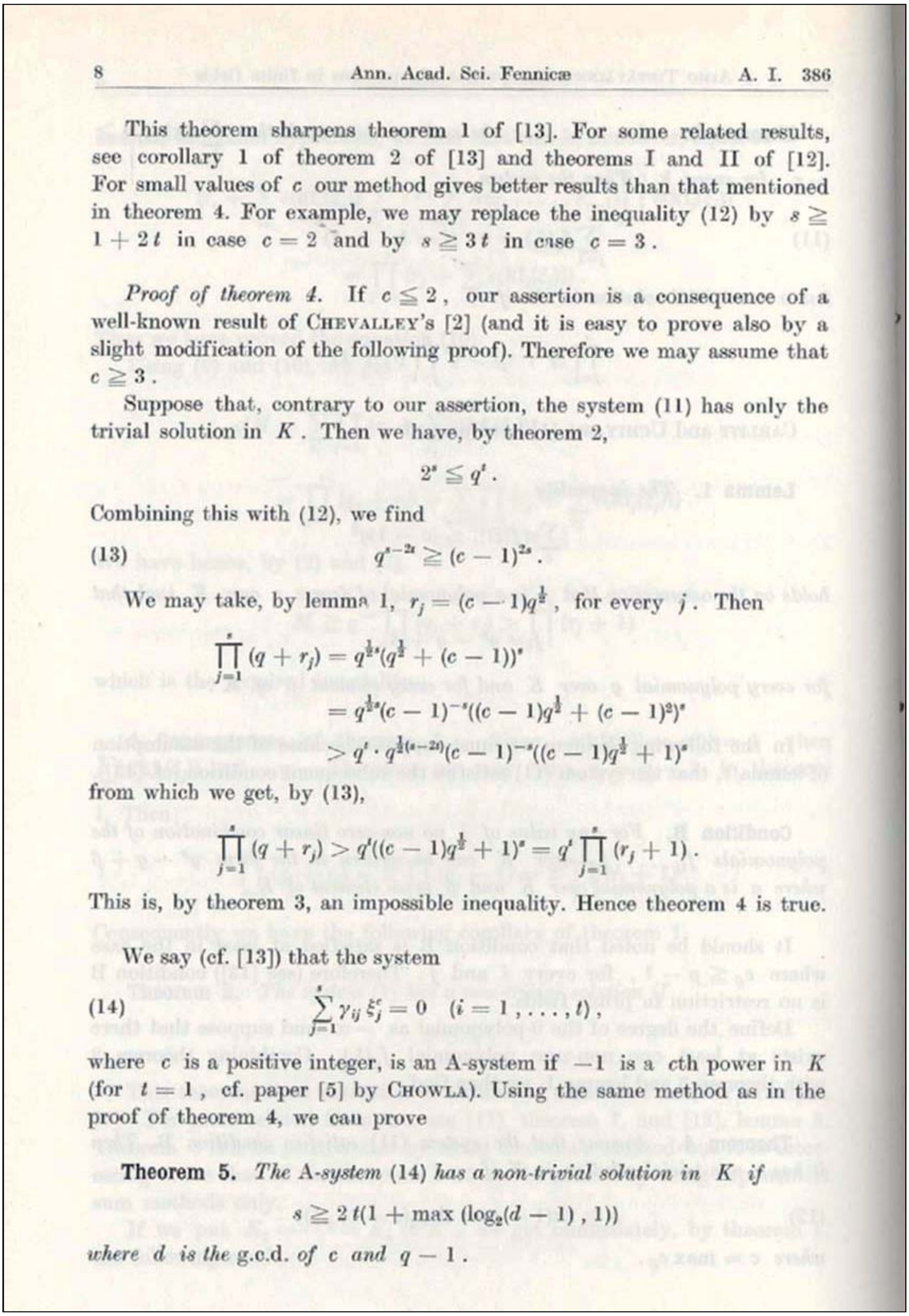
It should be noted that condition B is satisfied at least in the case where $c_{ij} \leqq p - 1$, for every i and j. Therefore (see [13]) condition B is no restriction in prime fields.

Define the degree of the 0-polynomial as $-\infty$ and suppose that there exists at least one non-zero polynomial $f_{ij}(\xi_j)$. Combining theorem 2 with theorem 3 and lemma 1, we then find

Theorem 4. *Assume that the system* (11) *satisfies condition* B. *Then it has a non-trivial solution in* K *if*

$$s \geqq 2\,t(1 + \max(\log_2(c - 1), 1)) \tag{12}$$

where $c = \max c_{ij}$.



This theorem sharpens theorem 1 of [13]. For some related results, see corollary 1 of theorem 2 of [13] and theorems I and II of [12]. For small values of c our method gives better results than that mentioned in theorem 4. For example, we may replace the inequality (12) by $s \geqq 1 + 2t$ in case $c = 2$ and by $s \geqq 3t$ in case $c = 3$.

Proof of theorem 4. If $c \leqq 2$, our assertion is a consequence of a well-known result of CHEVALLEY's [2] (and it is easy to prove also by a slight modification of the following proof). Therefore we may assume that $c \geqq 3$.

Suppose that, contrary to our assertion, the system (11) has only the trivial solution in K. Then we have, by theorem 2,

$$2^s \leqq q^t .$$

Combining this with (12), we find

$$q^{s-2t} \geqq (c-1)^{2s} . \tag{13}$$

We may take, by lemma 1, $r_j = (c-1)q^{\frac{1}{2}}$, for every j. Then

$$\begin{aligned}\prod_{j=1}^{s}(q+r_j) &= q^{\frac{1}{2}s}(q^{\frac{1}{2}}+(c-1))^s \\ &= q^{\frac{1}{2}s}(c-1)^{-s}((c-1)q^{\frac{1}{2}}+(c-1)^2)^s \\ &> q^t\cdot q^{\frac{1}{2}(s-2t)}(c-1)^{-s}((c-1)q^{\frac{1}{2}}+1)^s\end{aligned}$$

from which we get, by (13),

$$\prod_{j=1}^{s}(q+r_j) > q^t((c-1)q^{\frac{1}{2}}+1)^s = q^t\prod_{j=1}^{s}(r_j+1) .$$

This is, by theorem 3, an impossible inequality. Hence theorem 4 is true.

We say (cf. [13]) that the system

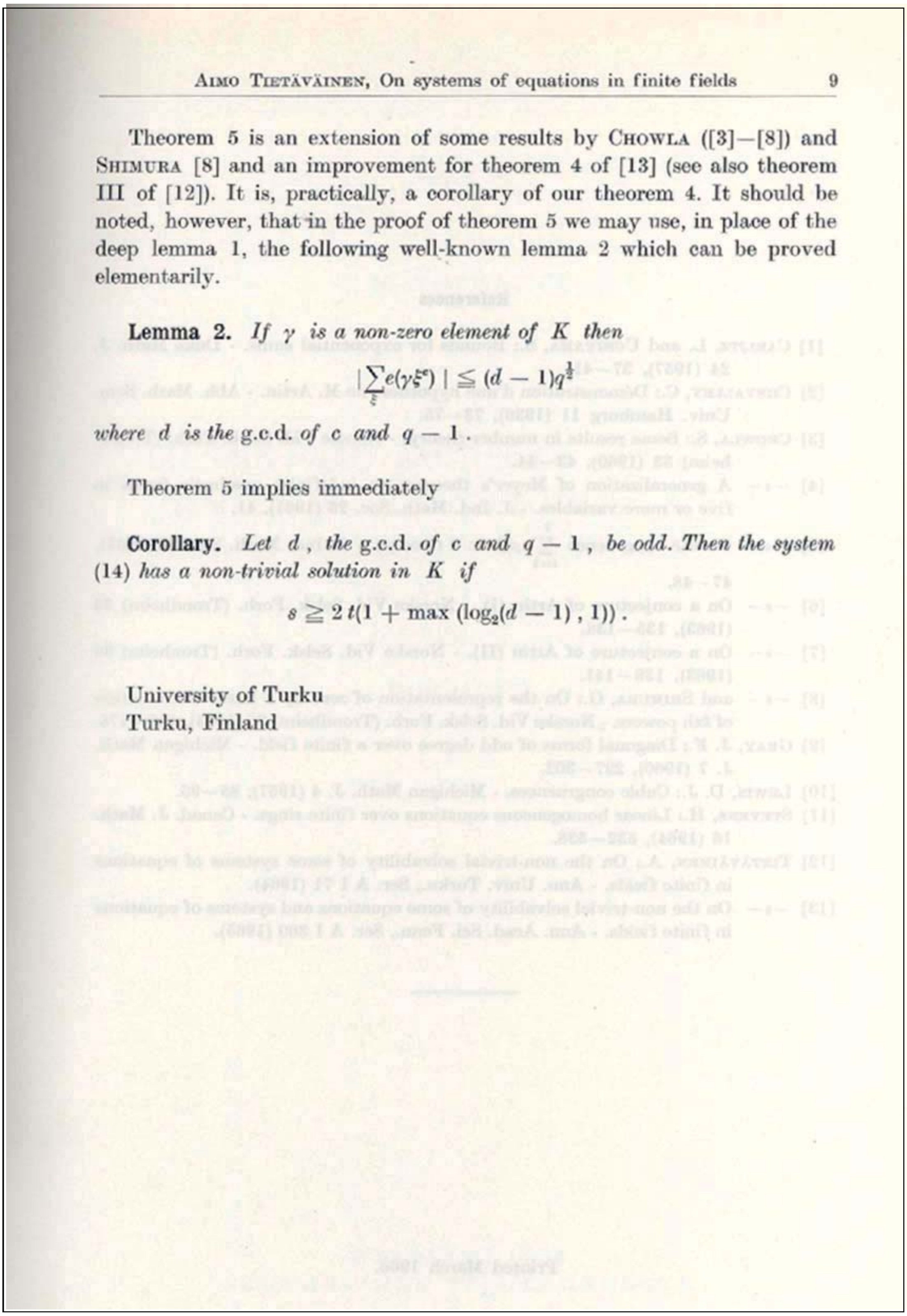
$$\sum_{j=1}^{s}\gamma_{ij}\,\xi_j^c = 0 \quad (i = 1, \ldots, t), \tag{14}$$

where c is a positive integer, is an A-system if -1 is a cth power in K (for $t = 1$, cf. paper [5] by CHOWLA). Using the same method as in the proof of theorem 4, we can prove

Theorem 5. *The* A-*system* (14) *has a non-trivial solution in* K *if*

$$s \geqq 2t(1 + \max(\log_2(d-1), 1))$$

where d *is the* g.c.d. *of* c *and* $q-1$.



Theorem 5 is an extension of some results by Chowla ([3]—[8]) and Shimura [8] and an improvement for theorem 4 of [13] (see also theorem III of [12]). It is, practically, a corollary of our theorem 4. It should be noted, however, that in the proof of theorem 5 we may use, in place of the deep lemma 1, the following well-known lemma 2 which can be proved elementarily.

Lemma 2. *If γ is a non-zero element of K then*

$$|\sum_{\xi} e(\gamma\xi^c)| \leqq (d-1)q^{\frac{1}{2}}$$

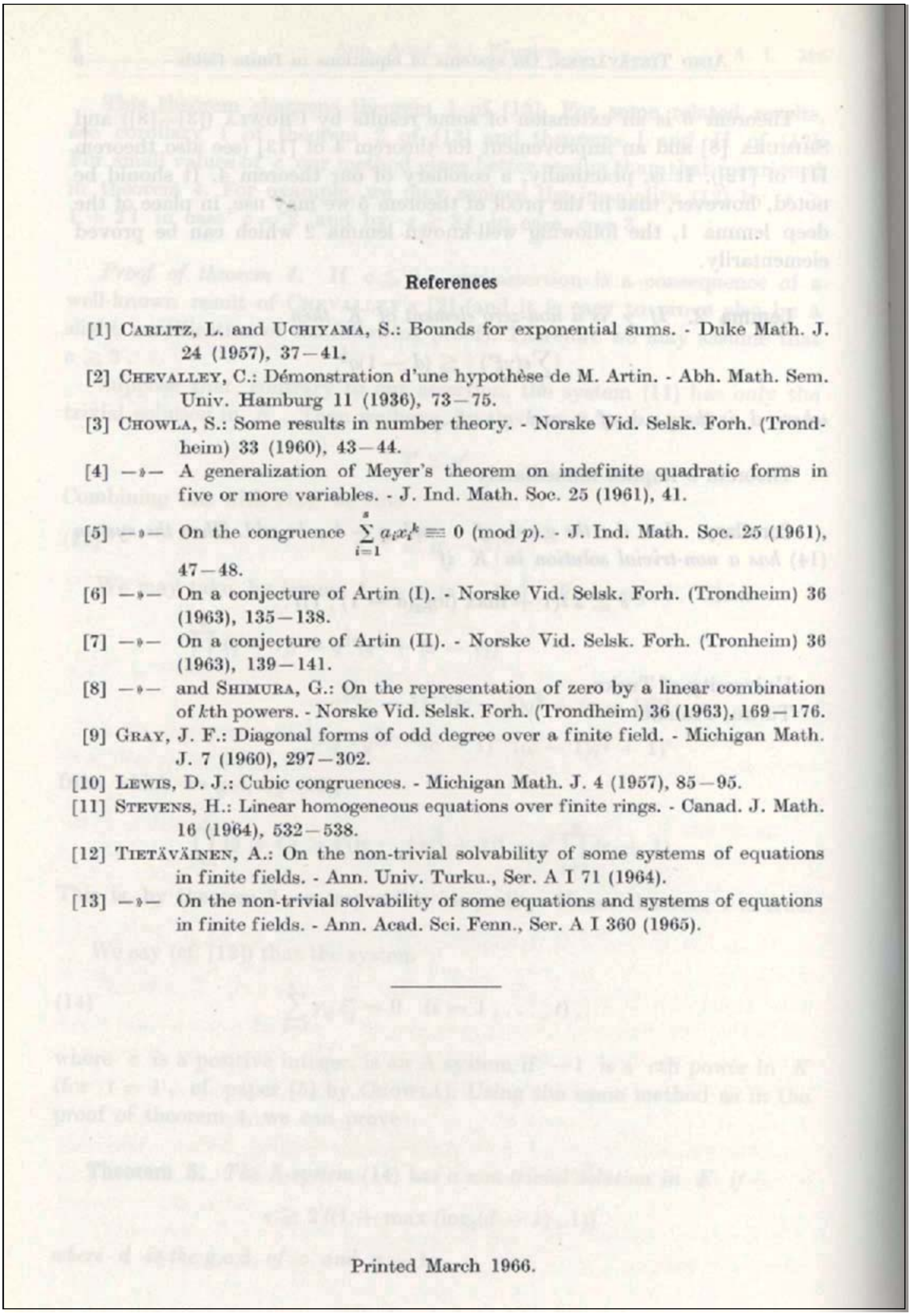
where d is the g.c.d. *of c and $q-1$.*

Theorem 5 implies immediately

Corollary. *Let d, the* g.c.d. *of c and $q-1$, be odd. Then the system* (14) *has a non-trivial solution in K if*

$$s \geqq 2\,t(1+\max(\log_2(d-1), 1)).$$

University of Turku
Turku, Finland



References

[1] Carlitz, L. and Uchiyama, S.: Bounds for exponential sums. - Duke Math. J. 24 (1957), 37–41.

[2] Chevalley, C.: Démonstration d'une hypothèse de M. Artin. - Abh. Math. Sem. Univ. Hamburg 11 (1936), 73–75.

[3] Chowla, S.: Some results in number theory. - Norske Vid. Selsk. Forh. (Trondheim) 33 (1960), 43–44.

[4] —»— A generalization of Meyer's theorem on indefinite quadratic forms in five or more variables. - J. Ind. Math. Soc. 25 (1961), 41.

[5] —»— On the congruence $\sum_{i=1}^{s} a_i x_i^k \equiv 0 \pmod{p}$. - J. Ind. Math. Soc. 25 (1961), 47–48.

[6] —»— On a conjecture of Artin (I). - Norske Vid. Selsk. Forh. (Trondheim) 36 (1963), 135–138.

[7] —»— On a conjecture of Artin (II). - Norske Vid. Selsk. Forh. (Tronheim) 36 (1963), 139–141.

[8] —»— and Shimura, G.: On the representation of zero by a linear combination of kth powers. - Norske Vid. Selsk. Forh. (Trondheim) 36 (1963), 169–176.

[9] Gray, J. F.: Diagonal forms of odd degree over a finite field. - Michigan Math. J. 7 (1960), 297–302.

[10] Lewis, D. J.: Cubic congruences. - Michigan Math. J. 4 (1957), 85–95.

[11] Stevens, H.: Linear homogeneous equations over finite rings. - Canad. J. Math. 16 (1964), 532–538.

[12] Tietäväinen, A.: On the non-trivial solvability of some systems of equations in finite fields. - Ann. Univ. Turku., Ser. A I 71 (1964).

[13] —»— On the non-trivial solvability of some equations and systems of equations in finite fields. - Ann. Acad. Sci. Fenn., Ser. A I 360 (1965).

Printed March 1966.

11.10.66

TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SERIES A

I. ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA

87

ON THE TRACE OF A POLYNOMIAL OVER A FINITE FIELD

BY

AIMO TIETÄVÄINEN

Turku 1966
TURUN YLIOPISTO

Turku
Kirjapaino Polytypos
1966

ON THE TRACE OF A POLYNOMIAL OVER A FINITE FIELD

1. Introduction. Let p be a prime and let $f(x)$ be a polynomial of degree n with integer coefficients. Let l denote the least non-negative residue of $f(x) \pmod p$. MORDELL [7] found the estimate

$$l \leqq np^{1/2} \log p. \tag{1}$$

Let $K = GF(q)$ denote the finite field of order q where $q = p^m$. Let $f(\xi)$ be a polynomial of degree n over K such that $f \not\equiv g^p - g + \beta$, for every polynomial g over K and for every element β of K. Let, moreover, l denote the least non-negative trace of $f(\xi)$ as ξ ranges over K. It follows from a result of CAVIOR's [3] that

$$l \leqq np^{1-m/2} \log p$$

which is a generalization of (1). We now show that, for $n > 1$,

$$l < 2(n-1)p^{1-m/2}. \tag{2}$$

In fact, we can prove this result in a more precise form which implies, furthermore, that $l = 0$ if

$$m > \frac{2\log((n-1)(p-1))}{\log p}. \tag{3}$$

The assumption ”f is not of the form $g^p - g + \beta$” is essential. For example, the field $GF(4)$ and the polynomial $\xi^2 - \xi + \rho$, where ρ is a primitive element of $GF(4)$, satisfy the condition (3), but, however, $\operatorname{tr}(\xi^2 - \xi + \rho) = 1$ for every element of $GF(4)$.

Suppose, moreover, that $f(0) = 0$ and $f(-\alpha) = -f(\alpha)$, for every element α of K. Let h denote the least non-negative trace of $f(\xi)$ as ξ runs through all the non-zero elements of K. Then, for $n > 2$,

$$h < (n-1)p^{1-m/2}. \tag{4}$$

It would be possible to prove a result like (4) in a manner similar to (2). However, we make use of another exponential sum method which is, probably, easier to generalize for systems of many polynomials.

It should be noted that in the special case

$$m = 1,\ f(\xi) = \gamma\xi^n \tag{5}$$

there exist better estimates than (4) (see [4], [5]). For example, it has been shown that there exists a positive $\eta = \eta(n)$ such that $h < p^{1/2-\eta}$, provided that (5) is satisfied and p is large enough. In particular, one may take $\eta(3) = 0.30$, $\eta(5) = 0.22$.

In proving the results (2) and (4), a deep estimate of CARLITZ and UCHIYAMA [2] for an exponential sum is required. A less precise result was found by MORDELL [6] more than thirty years ago. Other useful estimates are known, for example, in the special case $m = 1$, $f(\xi) = \alpha\xi^n + \beta\xi^k$, $\alpha \neq 0$, $\beta \neq 0$, $(n,k) = 1$ (see [1]).

2. Preliminary results. Let α be an element of K and let a and b be integers. Define the trace of α as

$$\operatorname{tr}(\alpha) = \alpha + \alpha^p + \ldots + \alpha^{p^{m-1}}$$

so that $\operatorname{tr}(\alpha)$ may be considered as an integer $(\operatorname{mod} p)$. Then

$$a\operatorname{tr}(\alpha) \equiv \operatorname{tr}(a\alpha) \pmod p. \tag{6}$$

Define, furthermore,

$$e_p(a) = e^{2\pi i a/p},\ e(\alpha) = e_p(\operatorname{tr}(\alpha)). \tag{7}$$

Then

$$e_p(a+b) = e_p(a)e_p(b), \tag{8}$$

$$\sum_{a=0}^{p-1} e_p(ab) = \begin{cases} p \text{ if } b \equiv 0 \pmod p, \\ 0 \text{ otherwise.} \end{cases} \tag{9}$$

Moreover (see [2])

$$|\sum_{\xi} e(f(\xi))| \leqq (n-1)p^{m/2} \tag{10}$$

provided that f cannot be written in the form $g^p - g + \beta$ where g is a polynomial over K and β is an element of K. Here and hereafter, in the sum $\sum_{\xi}$ the variable ξ runs through all the elements of K.

Let u be an integer such that $0 \leqq u \leqq (p-1)/2$. Denote $U = \{0,1,\ldots,u\}$ and

$$S(a) = \sum_{b=0}^{u} e_p(ab).$$

Then, by (8),

$$|S(a)|^2 = \sum_{b=0}^{u} e_p(ab) \sum_{c=0}^{u} e_p(-ac) = \sum_{b=0}^{u} \sum_{c=0}^{u} e_p(a(b-c)). \tag{11}$$

Since the number of solutions of the congruence

$$b - c \equiv 0 \pmod p, \quad b \in U, \quad c \in U$$

is equal to $u+1$, we have, by (11) and (9),

$$\sum_{a=0}^{p-1} |S(a)|^2 = \sum_{b=0}^{u} \sum_{c=0}^{u} \sum_{a=0}^{p-1} e_p(a(b-c)) = p(u+1).$$

This implies

$$\sum_{a=1}^{p-1} |S(a)|^2 = p(u+1) - (u+1)^2 = (u+1)(p-u-1). \tag{12}$$

3. Proof of the inequality (2).

It is clear that we may prove the assertion (2) in the following form: There exists an element ξ of K such that $-u \leqq \operatorname{tr}(f(\xi)) \leqq u$ where the integer u satisfies the inequalities

$$(n-1)p^{1-m/2} - 1 \leqq u < (n-1)p^{1-m/2}. \tag{13}$$

We may assume that $u \leqq (p-1)/2$ because (2) is trivial otherwise.

Consider the congruence

$$\operatorname{tr}(f(\xi)) - y + z \equiv 0 \pmod p \tag{14}$$

where $\xi \in K$, $y \in U$, $z \in U$ and hence $-u \leqq y - z \leqq u$. We see, by (9), (8), (6), (7), and (11), that N, the number of solutions (ξ, y, z) of (14), satisfies the equations

$$\begin{aligned} pN &= \sum_{\xi} \sum_{y=0}^{u} \sum_{z=0}^{u} \sum_{k=0}^{p-1} e_p(k(\operatorname{tr}(f(\xi)) - y + z)) \\ &= \sum_{k=0}^{p-1} \sum_{\xi} e_p(\operatorname{tr}(kf(\xi))) \sum_{y=0}^{u} \sum_{z=0}^{u} e_p(k(-y+z)) \\ &= q(u+1)^2 + \sum_{k=1}^{p-1} \sum_{\xi} e(kf(\xi))\,|S(k)|^2. \end{aligned}$$

Using (10) and (12), we find, furthermore,

$$\begin{aligned} pN &\geqq q(u+1)^2 - (n-1)p^{m/2} \sum_{k=1}^{p-1} |S(k)|^2 \\ &= q(u+1)^2 - (n-1)p^{m/2}(u+1)(p-u-1) \\ &= p^{m/2}(u+1)((u+1)p^{m/2} - (n-1)(p-u-1)). \end{aligned} \tag{15}$$

Thus, provided that u satisfies (13), we have

$$pN \geqq (u+1)^2(n-1)p^{m/2} > 0.$$

Consequently (2) has been proved.

If we put $u = 0$ in (15), we find

$$N \geqq p^{m/2-1}(p^{m/2}-(n-1)(p-1)).$$

Hence $l = 0$ if

$$p^{m/2} > (n-1)(p-1)$$

or if (3) is true.

4. Proof of the inequality (4). Denote, briefly, $(n-1)p^{m/2} = r$. Let

$$A(\xi) = \begin{cases} 1 \text{ if } \xi \neq 0, \\ r+1 \text{ if } \xi = 0, \end{cases} \tag{16}$$

$$B(\xi, y, z) = \begin{cases} A(\xi) \text{ if (14) is satisfied,} \\ 0 \text{ otherwise,} \end{cases} \tag{17}$$

and

$$M(u) = \sum_{\xi} \sum_{y=0}^{u} \sum_{z=0}^{u} B(\xi, y, z). \tag{18}$$

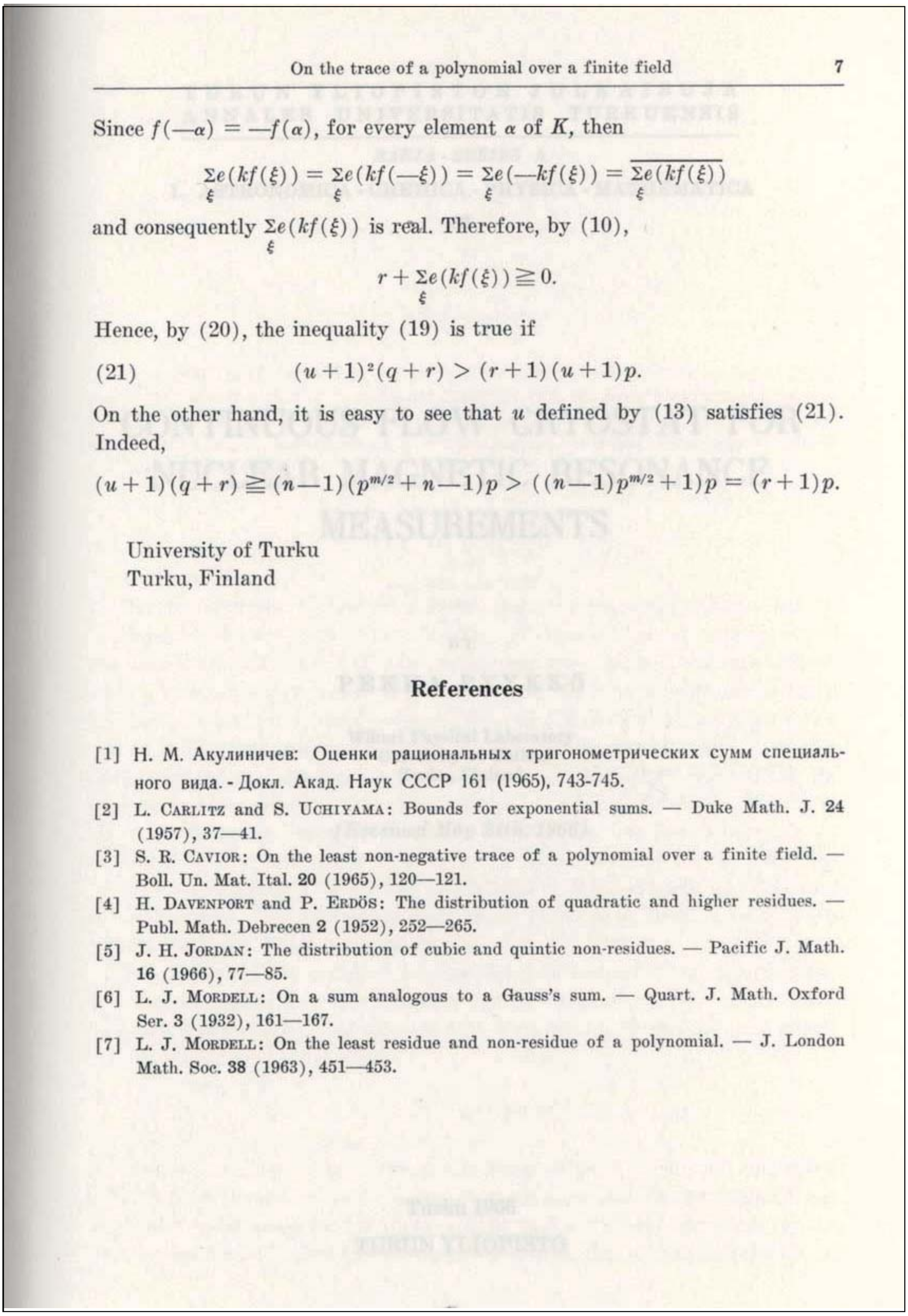
If

$$M(u) > (r+1)(u+1) \tag{19}$$

then (14) has a solution (ξ, y, z), where $\xi \in K - \{0\}$, $y \in U$, and $z \in U$, and consequently, by the condition $f(-\xi) = -f(\xi)$, there exists a non-zero element ξ of K such that $0 \leqq \operatorname{tr}(f(\xi)) \leqq u$. Therefore we have to prove that u defined by (13) satisfies the inequality (19).

Using the definitions (16), (17), and (18), and the same methods as in section 3, we find

$$\begin{aligned} pM(u) &= \sum_{\xi} \sum_{y=0}^{u} \sum_{z=0}^{u} A(\xi) \sum_{k=0}^{p-1} e_p(k(\operatorname{tr}(f(\xi)) - y + z)) \\ &= \sum_{k=0}^{p-1} |S(k)|^2 \sum_{\xi} A(\xi) e(kf(\xi)) \\ &= \sum_{k=0}^{p-1} |S(k)|^2 (r + \sum_{\xi} e(kf(\xi))) \\ &= (u+1)^2(q+r) + \sum_{k=1}^{p-1} |S(k)|^2 (r + \sum_{\xi} e(kf(\xi))). \end{aligned} \tag{20}$$



Since $f(-\alpha) = -f(\alpha)$, for every element α of K, then

$$\underset{\xi}{\Sigma} e(kf(\xi)) = \underset{\xi}{\Sigma} e(kf(-\xi)) = \underset{\xi}{\Sigma} e(-kf(\xi)) = \overline{\underset{\xi}{\Sigma} e(kf(\xi))}$$

and consequently $\underset{\xi}{\Sigma} e(kf(\xi))$ is real. Therefore, by (10),

$$r + \underset{\xi}{\Sigma} e(kf(\xi)) \geqq 0.$$

Hence, by (20), the inequality (19) is true if

$$(21) \qquad (u+1)^2(q+r) > (r+1)(u+1)p.$$

On the other hand, it is easy to see that u defined by (13) satisfies (21). Indeed,

$$(u+1)(q+r) \geqq (n-1)(p^{m/2}+n-1)p > ((n-1)p^{m/2}+1)p = (r+1)p.$$

University of Turku
Turku, Finland

References

[1] Н. М. Акулиничев: Оценки рациональных тригонометрических сумм специального вида. - Докл. Акад. Наук СССР 161 (1965), 743-745.

[2] L. Carlitz and S. Uchiyama: Bounds for exponential sums. — Duke Math. J. **24** (1957), 37—41.

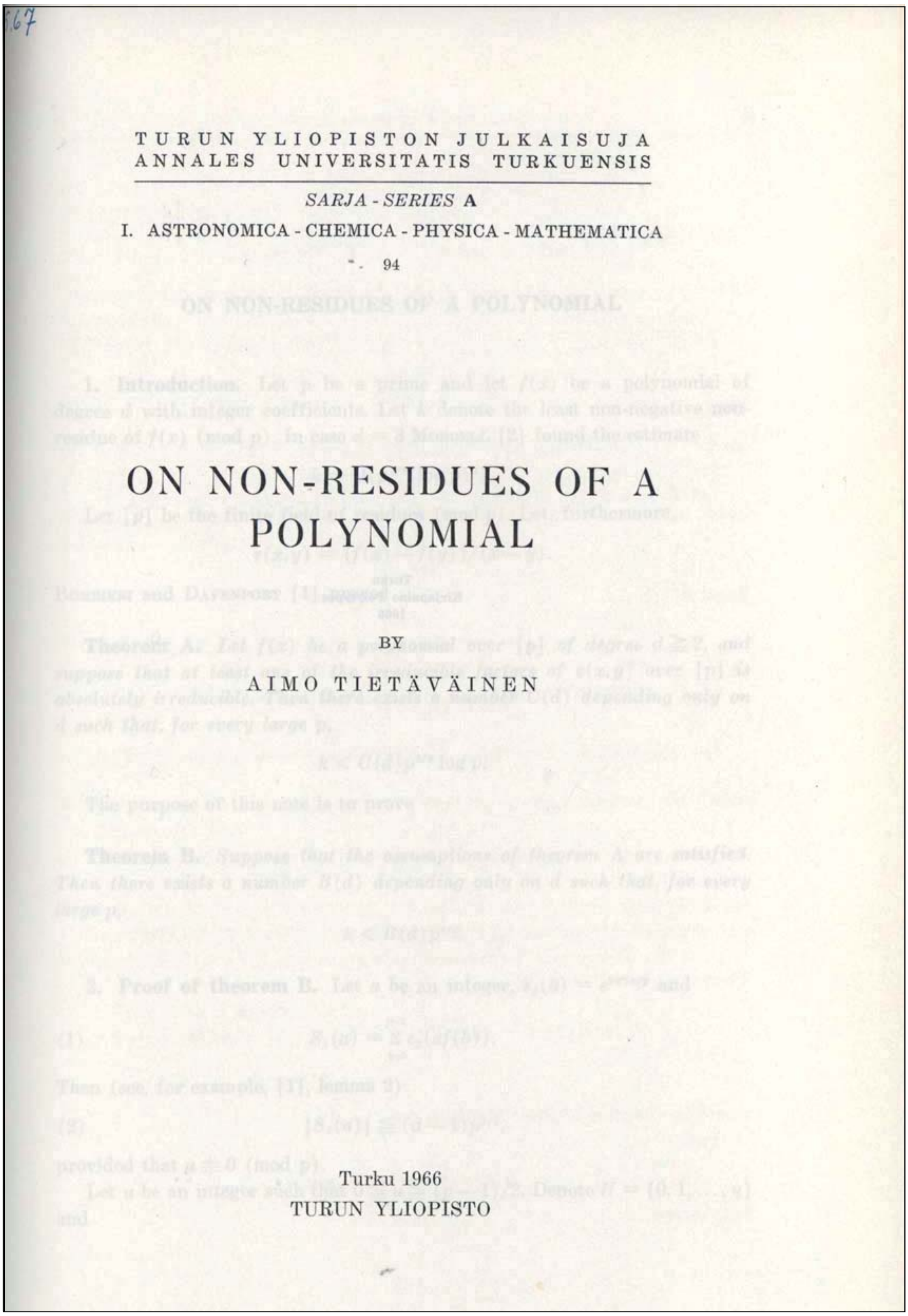
[3] S. R. Cavior: On the least non-negative trace of a polynomial over a finite field. — Boll. Un. Mat. Ital. **20** (1965), 120—121.

[4] H. Davenport and P. Erdös: The distribution of quadratic and higher residues. — Publ. Math. Debrecen **2** (1952), 252—265.

[5] J. H. Jordan: The distribution of cubic and quintic non-residues. — Pacific J. Math. **16** (1966), 77—85.

[6] L. J. Mordell: On a sum analogous to a Gauss's sum. — Quart. J. Math. Oxford Ser. **3** (1932), 161—167.

[7] L. J. Mordell: On the least residue and non-residue of a polynomial. — J. London Math. Soc. **38** (1963), 451—453.



67

TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SERIES A

I. ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA

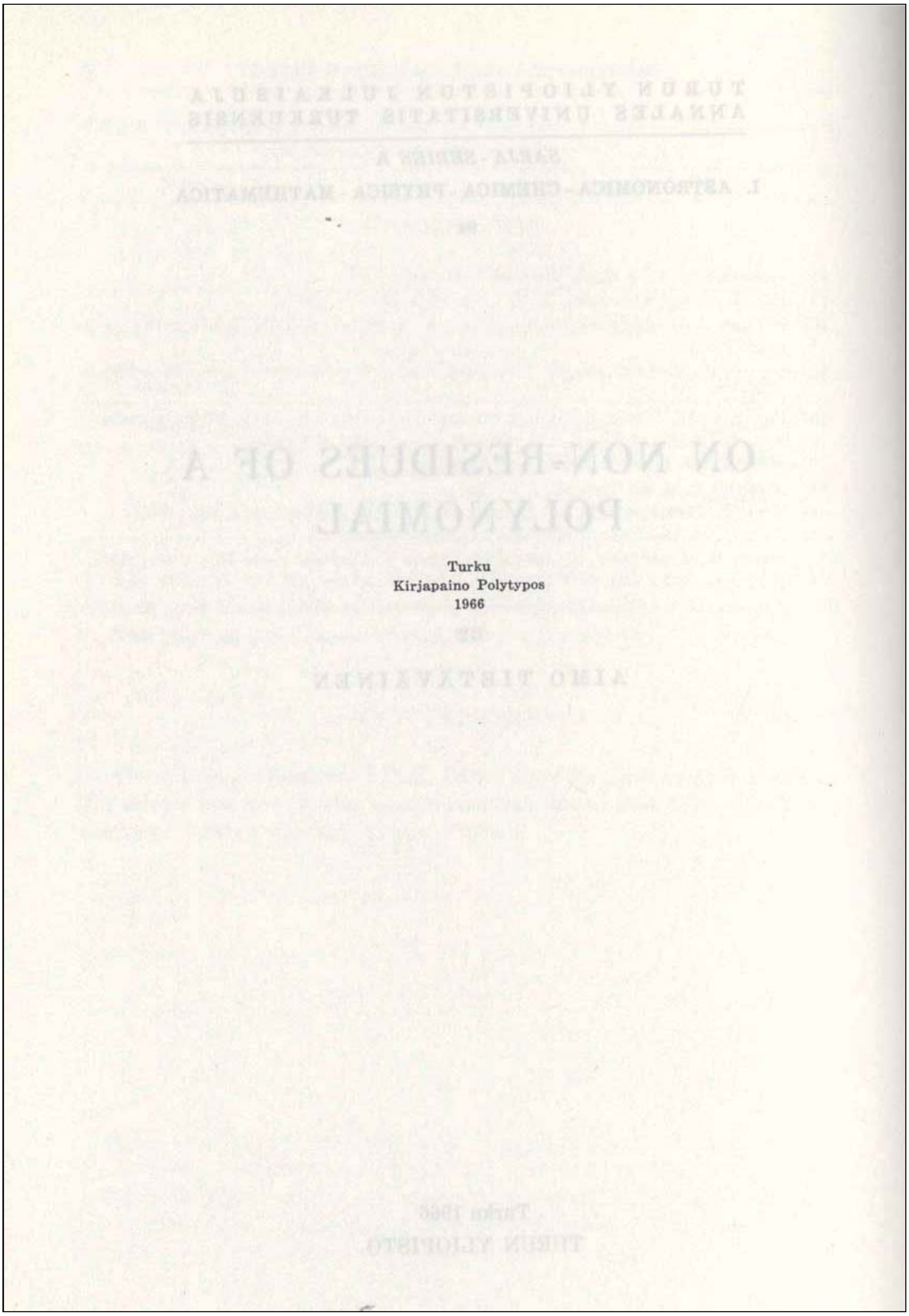
94

ON NON-RESIDUES OF A POLYNOMIAL

BY

AIMO TIETÄVÄINEN

Turku 1966
TURUN YLIOPISTO



Turku
Kirjapaino Polytypos
1966

ON NON-RESIDUES OF A POLYNOMIAL

1. Introduction. Let p be a prime and let $f(x)$ be a polynomial of degree d with integer coefficients. Let k denote the least non-negative non-residue of $f(x)$ (mod p). In case $d = 3$ Mordell [2] found the estimate

$$k = O(p^{1/2}(\log p)^2).$$

Let $[p]$ be the finite field of residues (mod p). Let, furthermore,

$$\varphi(x,y) = (f(x) - f(y))/(x - y).$$

Bombieri and Davenport [1] proved

Theorem A. *Let $f(x)$ be a polynomial over $[p]$ of degree $d \geqq 2$, and suppose that at least one of the irreducible factors of $\varphi(x,y)$ over $[p]$ is absolutely irreducible. Then there exists a number $C(d)$ depending only on d such that, for every large p,*

$$k < C(d)p^{1/2} \log p.$$

The purpose of this note is to prove

Theorem B. *Suppose that the assumptions of theorem* A *are satisfied. Then there exists a number $B(d)$ depending only on d such that, for every large p,*

$$k < B(d)p^{1/2}.$$

2. Proof of theorem B. Let a be an integer, $e_p(a) = e^{2\pi i a/p}$ and

$$S_1(a) = \sum_{b=0}^{p-1} e_p(af(b)). \tag{1}$$

Then (see, for example, [1], lemma 2)

$$|S_1(a)| \leqq (d-1)p^{1/2}, \tag{2}$$

provided that $a \not\equiv 0$ (mod p).

Let u be an integer such that $0 \leqq u \leqq (p-1)/2$. Denote $U = \{0, 1, \ldots, u\}$ and

$$S(a) = \sum_{b=0}^{u} e_p(-ab). \tag{3}$$

Then ([3], p. 5)

$$\sum_{a=1}^{p-1} |S(a)|^2 = (u+1)(p-u-1). \tag{4}$$

Assume that $k(m)$ is the number of solutions (x, y) of the congruence

$$x + y - m \equiv 0 \pmod{p},\ x \in U,\ y \in U.$$

Then

$$pk(m) = \sum_{t=0}^{p-1} \sum_{x=0}^{u} \sum_{y=0}^{u} e_p(t(m-x-y)). \tag{5}$$

Let T be a mapping of $[p]$ into the field of complex numbers. Put

$$V(t) = \sum_{m=0}^{p-1} T(m) e_p(tm).$$

Then, by (5),

$$\begin{aligned} p \sum_{m=0}^{p-1} T(m)k(m) &= \sum_{t=0}^{p-1} \sum_{m=0}^{p-1} T(m) e_p(tm) \sum_{x=0}^{u} e_p(-tx) \sum_{y=0}^{u} e_p(-ty) \\ &= \sum_{t=0}^{p-1} V(t)(S(t))^2. \end{aligned} \tag{6}$$

Let $r(m)$ denote the number of solutions of $f(x) \equiv m \pmod{p}$. Then

$$S_1(t) = \sum_{x=0}^{p-1} e_p(tf(x)) = \sum_{m=0}^{p-1} r(m) e_p(tm).$$

We now apply the equation (6), setting $T(m) = r(m)$. We then find, by (1) and (3),

$$p \sum_{m=0}^{p-1} r(m)k(m) = \sum_{t=0}^{p-1} S_1(t)(S(t))^2 = p(u+1)^2 + \sum_{t=1}^{p-1} S_1(t)(S(t))^2.$$

Therefore, by (2) and (4),

$$\begin{aligned} \left| \sum_{m=0}^{p-1} r(m)k(m) - (u+1)^2 \right| &\leqq p^{-1}(d-1)p^{1/2} \sum_{t=1}^{p-1} |S(t)|^2 \\ &= (u+1)O(p^{1/2}) \end{aligned} \tag{7}$$

where the implied constant depends only on d.

Put

$$R(t) = \sum_{m=0}^{p-1} (r(m))^2 e_p(tm).$$

Let, moreover,

$$R = \max |R(t)| \text{ for } t \not\equiv 0 \pmod{p}.$$

Then ([1], p. 66)

$$R = O(p^{1/2}), R(0) = (e+1)p + O(p^{1/2})$$

where e is the number of the irreducible factors of $\varphi(x,y)$ over $[p]$, which are absolutely irreducible, and consequently, by the assumption, $e \geqq 1$. We now apply the equation (6), setting $T(m) = (r(m))^2$. Then we find

$$\begin{aligned} \sum_{m=0}^{p-1} (r(m))^2 k(m) &= p^{-1} \sum_{t=0}^{p-1} R(t)(S(t))^2 \\ &= p^{-1}(u+1)^2 R(0) + (u+1)O(R) \qquad (8) \\ &= (e+1)(u+1)^2 + (u+1)O(p^{1/2}). \end{aligned}$$

Let $N_h = \Sigma k(m)$ where the summation is over all the m's for which $0 \leqq m \leqq p-1$ and $r(m) = h$. Then

$$\sum_{h=0}^{d} N_h = \sum_{m=0}^{p-1} k(m) = (u+1)^2.$$

Furthermore, by (7) and (8),

$$\sum_{h=1}^{d} hN_h = \sum_{m=0}^{p-1} r(m)k(m) = (u+1)^2 + (u+1)O(p^{1/2}),$$

$$\sum_{h=1}^{d} h^2 N_h = \sum_{m=0}^{p-1} (r(m))^2 k(m) = (e+1)(u+1)^2 + (u+1)O(p^{1/2}).$$

Multiplying by 1, $-1-d^{-1}$ and d^{-1} respectively and adding, we deduce (cf. [1], p. 66) that

$$N_0 \geqq ed^{-1}(u+1)^2 + (u+1)O(p^{1/2}).$$

Thus $N_0 > 0$ if p is large enough and $u \geqq A(d)p^{1/2}$, for a suitable $A(d)$. Consequently there exists, for every large p, an integer m such that $0 \leqq m < 2A(d)p^{1/2} + 2$ and $r(m) = 0$. This proves theorem B.

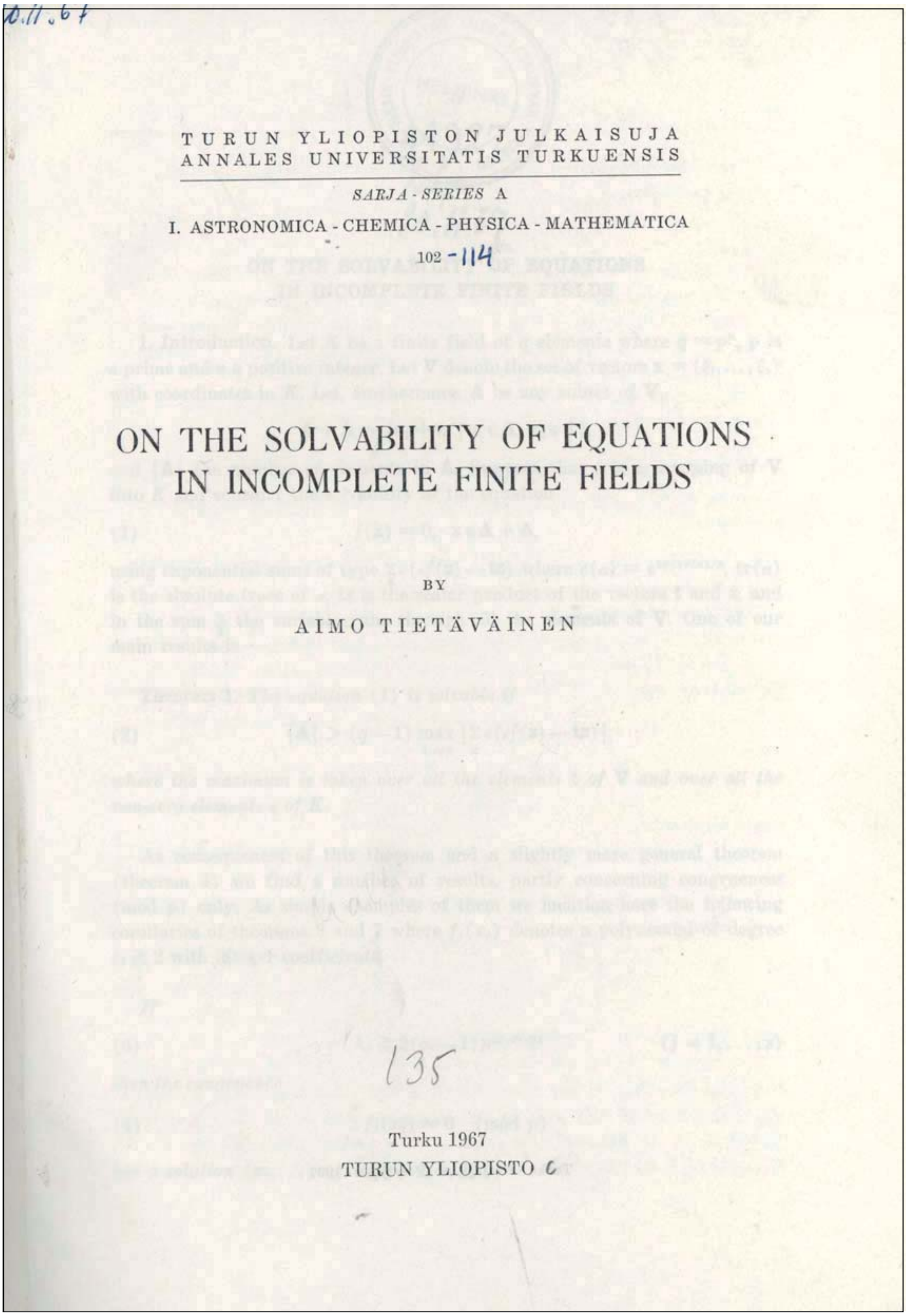
University of Turku
Turku, Finland

References

[1] E. Bombieri and H. Davenport: On two problems of Mordell. — Amer. J. Math. 88 (1966), 61—70.

[2] L. J. Mordell: On the least residue and non-residue of a polynomial. — J. London Math. Soc. 38 (1963), 451—453.

[3] A. Tietäväinen: On the trace of a polynomial over a finite field. — Ann. Univ. Turku., Ser. A I 87 (1966).



0.11.67

TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SERIES A

I. ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA

102 -114

ON THE SOLVABILITY OF EQUATIONS IN INCOMPLETE FINITE FIELDS

BY

AIMO TIETÄVÄINEN

135

Turku 1967
TURUN YLIOPISTO c

Turku - Kirjapaino Polytypos - 1967

ON THE SOLVABILITY OF EQUATIONS IN INCOMPLETE FINITE FIELDS

1. Introduction. Let K be a finite field of q elements where $q = p^n$, p is a prime and n a positive integer. Let $\mathbf{V}$ denote the set of vectors $\mathbf{x} = (\xi_1, \ldots, \xi_s)$ with coordinates in K. Let, furthermore, $\mathbf{A}$ be any subset of $\mathbf{V}$,

$$\mathbf{A} + \mathbf{A} = \{\mathbf{a}_1 + \mathbf{a}_2 \mid \mathbf{a}_1 \in \mathbf{A},\ \mathbf{a}_2 \in \mathbf{A}\},$$

and $|\mathbf{A}|$ the number of elements in $\mathbf{A}$. Suppose that f is a mapping of $\mathbf{V}$ into K and consider the solvability of the equation

$$f(\mathbf{x}) = 0, \quad \mathbf{x} \in \mathbf{A} + \mathbf{A}, \tag{1}$$

using exponential sums of type $\sum\limits_{\mathbf{z}} e(\kappa f(\mathbf{z}) - \mathbf{tz})$ where $e(\alpha) = e^{2\pi i\,\mathrm{tr}(\alpha)/p}$, $\mathrm{tr}(\alpha)$ is the absolute trace of α, $\mathbf{tz}$ is the scalar product of the vectors $\mathbf{t}$ and $\mathbf{z}$, and in the sum $\sum\limits_{\mathbf{z}}$ the variable runs through all the elements of $\mathbf{V}$. One of our main results is

Theorem 1. *The equation* (1) *is solvable if*

$$|\mathbf{A}| > (q-1) \max_{\mathbf{t}, \kappa \neq 0} \left|\sum_{\mathbf{z}} e(\kappa f(\mathbf{z}) - \mathbf{tz})\right| \tag{2}$$

where the maximum is taken over all the elements $\mathbf{t}$ *of* $\mathbf{V}$ *and over all the non-zero elements* κ *of* K.

As consequences of this theorem and a slightly more general theorem (theorem 4) we find a number of results, partly concerning congruences (mod p) only. As simple examples of them we mention here the following corollaries of theorems 3 and 7 where $f_j(x_j)$ denotes a polynomial of degree $c_j \geqq 2$ with integer coefficients.

If

$$h_j \geqq 2(c_j - 1)p^{(s+2)/2s} \qquad (j = 1, \ldots, s) \tag{3}$$

then the congruence

$$\sum_{j=1}^{s} f_j(x_j) \equiv 0 \pmod{p} \tag{4}$$

has a solution $(x_1, \ldots, x_s)$ *with* $0 \leqq x_j \leqq h_j$.

l, the least non-negative residue of the polynomial $\sum_{j=1}^{s} f_j(x_j)$ (mod p), *satisfies the inequality*

$$l < 2p^{(2-s)/2} \prod_{j=1}^{s} (c_j - 1). \tag{5}$$

The former of these corollaries improves the following result of CHALK's [4]: *the congruence* (4) *has a solution* $(x_1, \ldots, x_s)$ *with* $1 \leqq x_j \leqq h_j$ *if* $h_j \geqq Cp^{(s+2)/2s} \log p$ *where* C *is a suitably large constant depending only on* $c_1, \ldots, c_s$. It should be noted, however, that MORDELL ([7], [9]) has found better results in case $c_1 = \ldots = c_s = 2$.

The special case $s = 1$ of the latter corollary is an improvement of the result $l \leqq c_1 p^{1/2} \log p$ of MORDELL's [8] (see also [10]).

2. Preliminary remarks. Let $\mathbf{a} = (\alpha_1, \ldots, \alpha_s)$ and $\mathbf{b} = (\beta_1, \ldots, \beta_s)$ be elements of $\mathbf{V}$ and α and β elements of K. Define the scalar product of $\mathbf{a}$ and $\mathbf{b}$ as

$$\mathbf{ab} = \alpha_1\beta_1 + \ldots + \alpha_s\beta_s.$$

The 0-element $(0, \ldots, 0)$ of $\mathbf{V}$ will be denoted by $\mathbf{0}$.

Define the trace of α as

$$\operatorname{tr}(\alpha) = \alpha + \alpha^p + \ldots + \alpha^{p^{n-1}}$$

so that $\operatorname{tr}(\alpha)$ may be considered as an integer (mod p). Define, furthermore,

$$e(\alpha) = e^{2\pi i \operatorname{tr}(\alpha)/p}.$$

Then

$$e(\alpha + \beta) = e(\alpha) e(\beta) \tag{6}$$

which implies that

$$e(\mathbf{k}(\mathbf{a}+\mathbf{b})) = e(\mathbf{ka}) e(\mathbf{kb}) \tag{7}$$

for every element $\mathbf{k}$ of $\mathbf{V}$.

It is known that

$$\sum_{\alpha} e(\alpha\beta) = \begin{cases} q & \text{if } \beta = 0, \\ 0 & \text{if } \beta \neq 0, \end{cases} \tag{8}$$

and

$$\sum_{\mathbf{a}} e(\mathbf{ab}) = \begin{cases} q^s & \text{if } \mathbf{b} = \mathbf{0}, \\ 0 & \text{if } \mathbf{b} \neq \mathbf{0}. \end{cases} \tag{9}$$

Here and hereafter, in the sums of type $\sum_{\alpha}$ and $\sum_{\alpha \neq 0}$ the summation is over all the elements of K and over all the non-zero elements of K, respectively. Moreover, in the sums of type $\sum_{\mathbf{a}}$ and $\sum_{\mathbf{a}}'$ the variable runs through all the elements of $\mathbf{V}$ and through all the elements of $\mathbf{A}$, respectively.

Denote

$$S(\mathbf{a}) = \Sigma'_{\mathbf{b}}\, e(\mathbf{ab}). \tag{10}$$

Then

$$S(0) = |\mathbf{A}|. \tag{11}$$

Furthermore, by (7),

$$|S(\mathbf{a})|^2 = \Sigma'_{\mathbf{b}}\, e(\mathbf{ab})\Sigma'_{\mathbf{c}}\, e(-\mathbf{ac}) = \Sigma'_{\mathbf{b}}\Sigma'_{\mathbf{c}}\, e(\mathbf{a}(\mathbf{b}-\mathbf{c})).$$

Consequently, by (9),

$$\Sigma_{\mathbf{a}}|S(\mathbf{a})|^2 = \Sigma'_{\mathbf{b}}\Sigma'_{\mathbf{c}}\Sigma_{\mathbf{a}}\, e(\mathbf{a}(\mathbf{b}-\mathbf{c})) = q^s|\mathbf{A}|. \tag{12}$$

3. Lemmas. CARLITZ and UCHIYAMA [2] have proved the deep

Lemma 1. *The inequality*

$$|\Sigma_{\xi}\, e(f(\xi))| \leqq (c-1)q^{1/2}$$

holds on the assumption that f is a polynomial of degree c over K such that $f \neq g^p - g + \beta$, for every polynomial g over K and for every element β of K.

The subsequent lemmas can be proved elementarily.

Lemma 2. *If α is a non-zero element of K then*

$$|\Sigma_{\xi}\, e(\alpha\xi^c)| \leqq (d-1)q^{1/2}$$

where d is the g.c.d. *of c and $q-1$.*

Lemma 3. *If α and β are non-zero elements of K then*

$$|\Sigma_{\xi}\, e(\alpha\xi^c + \beta\xi)| \leqq d^{-1/2}q$$

where d is the g.c.d. *of c and $q-1$.*

The proof of lemma 2 is well-known (for the special case $q = p$, see [11], p. 47). AKULINIČEV [1] has proved lemma 3 in case $q = p$; we give now a slightly different proof for the general case.

Denote

$$T(\beta) = \Sigma_{\xi}\, e(\alpha\xi^c + \beta\xi).$$

Then

$$|T(\beta)|^2 = \sum_{\xi} e(\alpha\xi^c + \beta\xi) \sum_{\eta} e(-\alpha\eta^c - \beta\eta) = \sum_{\xi}\sum_{\eta} e(\alpha(\xi^c - \eta^c) + \beta(\xi - \eta))$$

and hence, by (8),

$$\sum_{\beta} |T(\beta)|^2 = \sum_{\xi}\sum_{\eta} e(\alpha(\xi^c - \eta^c)) \sum_{\beta} e(\beta(\xi - \eta)) = q^2. \tag{13}$$

Let $Y = \{\eta \in K \mid \eta^c = 1\}$. Then $|Y| = d$. Moreover

$$T(\beta\eta) = \sum_{\xi} e(\alpha\xi^c + \beta\eta\xi) = \sum_{\xi} e(\alpha(\eta\xi)^c + \beta\eta\xi) = \sum_{\zeta} e(\alpha\zeta^c + \beta\zeta) = T(\beta),$$

for every element η of Y. Therefore, by (13),

$$d|T(\beta)|^2 = \sum_{\eta \in Y} |T(\beta\eta)|^2 \leqq \sum_{\gamma} |T(\gamma)|^2 = q^2,$$

for every non-zero element β of K. This implies lemma 3.

4. Proof of theorem 1. The equation (1) is solvable if and only if the equation

$$f(\mathbf{x} + \mathbf{y}) = 0, \quad \mathbf{x} \in \mathbf{A}, \ \mathbf{y} \in \mathbf{A} \tag{14}$$

is solvable. Let M be the number of solutions of the equation (14). Using (8), we see that

$$M = q^{-1} \sum_{\mathbf{x}}{}' \sum_{\mathbf{y}}{}' \sum_{\kappa} e(\kappa f(\mathbf{x} + \mathbf{y})).$$

This implies, by (9) and (10),

$$\begin{aligned} M &= q^{-s-1} \sum_{\kappa}\sum_{\mathbf{z}} e(\kappa f(\mathbf{z})) \sum_{\mathbf{x}}{}' \sum_{\mathbf{y}}{}' \sum_{\mathbf{t}} e(\mathbf{t}(\mathbf{x} + \mathbf{y} - \mathbf{z})) \\ &= q^{-s-1} \sum_{\kappa}\sum_{\mathbf{t}} (S(\mathbf{t}))^2 \sum_{\mathbf{z}} e(\kappa f(\mathbf{z}) - \mathbf{tz}). \end{aligned} \tag{15}$$

Picking out the term with $\kappa = 0$, we get furthermore, by (9) and (11),

$$M = q^{-1}|\mathbf{A}|^2 + q^{-s-1} \sum_{\mathbf{t}} (S(\mathbf{t}))^2 \sum_{\kappa \neq 0} \sum_{\mathbf{z}} e(\kappa f(\mathbf{z}) - \mathbf{tz}). \tag{16}$$

Combining this with (12), we find

$$M \geqq q^{-1}|\mathbf{A}|^2 - q^{-1}(q-1)|\mathbf{A}|m = q^{-1}|\mathbf{A}|(|\mathbf{A}| - (q-1)m)$$

where

$$m = \max_{\mathbf{t}, \kappa \neq 0} \left|\sum_{\mathbf{z}} e(\kappa f(\mathbf{z}) - \mathbf{tz})\right|. \tag{17}$$

Consequently, by (2), $M > 0$. Thus theorem 1 has been proved.

5. Consequences of theorem 1. Let A_j be a subset of K and let $f_j(\xi_j)$ be a polynomial of degree c_j ($c_j \geqq 2$, c_j is not divisible by p) over K. By theorem 1, the equation

$$\sum_{j=1}^{s} f_j(\xi_j) = 0, \quad \xi_j \in A_j + A_j \tag{18}$$

is solvable if

$$\prod_{j=1}^{s} |A_j| > (q-1)m.$$

Now

$$m = \max_{\mathbf{t},\, \kappa \neq 0} \left| \sum_{\mathbf{z}} e\Big(\kappa \sum_{j=1}^{s} f_j(\zeta_j) - \sum_{j=1}^{s} \tau_j \zeta_j\Big) \right|$$

where $\mathbf{t} = (\tau_1, \ldots, \tau_s)$ and $\mathbf{z} = (\zeta_1, \ldots, \zeta_s)$. Furthermore, by (6) and lemma 1,

$$\begin{aligned} \left| \sum_{\mathbf{z}} e\Big(\kappa \sum_{j=1}^{s} f_j(\zeta_j) - \sum_{j=1}^{s} \tau_j \zeta_j\Big) \right| &= \prod_{j=1}^{s} \left| \sum_{\zeta_j} e(\kappa f_j(\zeta_j) - \tau_j \zeta_j) \right| \\ &\leqq q^{s/2} \prod_{j=1}^{s} (c_j - 1), \end{aligned}$$

for every $\mathbf{t}$ and for every non-zero κ. Consequently we have

Theorem 2. *The equation* (18) *is solvable if*

$$\prod_{j=1}^{s} |A_j| > (q-1) q^{s/2} \prod_{j=1}^{s} (c_j - 1).$$

Theorem 2 implies immediately

Corollary 1. *The equation* (18) *is solvable if*

$$|A_j| \geqq (c_j - 1) q^{(s+2)/2s} \qquad (j = 1, \ldots, s).$$

If we put $|A_j| = q$ in theorem 2, we get the following well-known result (for the special case $f_j(\xi_j) = \gamma_j \xi_j^c$, see, for example, [6]).

Corollary 2. *The equation*

$$\sum_{j=1}^{s} f_j(\xi_j) = 0$$

has a solution in K if

$$q^{s-2} \geqq \prod_{j=1}^{s} (c_j - 1)^2.$$

Consider now the congruence

$$(19) \qquad \sum_{j=1}^{s} f_j(x_j) \equiv 0 \pmod{p}, \quad 0 \leqq x_j \leqq h_j$$

where $f_j(x_j)$ is a polynomial of degree $c_j \geqq 2$ with integer coefficients and h_j is $< p$. Since $a^p \equiv a \pmod{p}$ for every integer a, we may suppose that $c_j < p$. If we choose $A_j = \{0, 1, \ldots, [h_j/2]\}$ then $2|A_j| > h_j$ and $A_j + A_j$ is a subset of $\{0, 1, \ldots, h_j\}$. Hence we have the following consequence of theorem 2.

Theorem 3. *The congruence* (19) *is solvable if*

$$\prod_{j=1}^{s} h_j \geqq 2^s (p-1) p^{s/2} \prod_{j=1}^{s} (c_j - 1).$$

Theorem 3 implies

Corollary. *The congruence* (19) *is solvable if the inequality* (3) *is true.*

6. Theorem 4 and its consequences. If $c_j = 1$ then it is possible that the sum

$$\sum_{\zeta_j} e(\kappa f_j(\zeta_j) - \tau_j \zeta_j)$$

doesn't satisfy the assumptions of lemma 1. Therefore the restriction $c_j \geqq 2$ is essential in the proof of theorem 2.

We now extend theorem 1 such that we consider the equation

$$(20) \qquad f(\mathbf{x}) - \xi = 0, \ \mathbf{x} \in \mathbf{A} + \mathbf{A}, \ \xi \in A + A$$

where A is a subset of K. This equation is solvable if and only if the equation

$$(21) \qquad f(\mathbf{x} + \mathbf{y}) - \xi - \eta = 0, \ \mathbf{x} \in \mathbf{A}, \ \mathbf{y} \in \mathbf{A}, \ \xi \in A, \ \eta \in A$$

is solvable. Using the same methods as in section 4, we observe that N, the number of solutions of the equation (21), satisfies the subsequent equations.

$$\begin{aligned} N &= q^{-1} \sum_{\mathbf{x}}{}' \sum_{\mathbf{y}}{}' \sum_{\xi}{}' \sum_{\eta}{}' \sum_{\kappa} e(\kappa(f(\mathbf{x}+\mathbf{y}) - \xi - \eta)) \\ &= q^{-s-1} \sum_{\mathbf{x}}{}' \sum_{\mathbf{y}}{}' \sum_{\mathbf{z}} \sum_{\mathbf{t}} e(\mathbf{t}(\mathbf{x}+\mathbf{y}-\mathbf{z})) \sum_{\xi}{}' \sum_{\eta}{}' \sum_{\kappa} e(\kappa(f(\mathbf{z}) - \xi - \eta)) \\ &= q^{-s-1} \sum_{\kappa} (S_0(\kappa))^2 \sum_{\mathbf{t}} (S(\mathbf{t}))^2 \sum_{\mathbf{z}} e(\kappa f(\mathbf{z}) - \mathbf{t}\mathbf{z}) \end{aligned}$$

where the summation in the sums of type $\sum_{\alpha}{}'$ is over all the elements of A,

$$S_0(\kappa) = \sum_{\alpha}{}' e(-\kappa\alpha), \tag{22}$$

and therefore (cf. the equations (11) and (12))

$$S_0(0) = |A|, \ \sum_{\kappa \neq 0} |S_0(\kappa)|^2 = |A|(q-|A|). \tag{23}$$

Hence

$$N \geqq q^{-1}|\mathbf{A}|^2|A|^2 - q^{-s-1}m \sum_{\mathbf{t}} |S(\mathbf{t})|^2 \sum_{\kappa \neq 0} |S_0(\kappa)|^2$$
$$= q^{-1}|\mathbf{A}|\,|A|(|\mathbf{A}|\,|A| - (q-|A|)m)$$

where m is defined by (17). Consequently we find

Theorem 4. *The equation* (20) *is solvable if*

$$|\mathbf{A}|\,|A| > (q-|A|) \max_{\mathbf{t},\, \kappa \neq 0} |\sum_{\mathbf{z}} e(\kappa f(\mathbf{z}) - \mathbf{t}\mathbf{z})|.$$

If we put $A = \{0\}$, we see that theorem 1 follows from theorem 4. Using the same methods as in the proof of theorem 2, we observe, furthermore, that theorem 4 implies

Theorem 5. *The equation*

$$\sum_{j=1}^{s} f_j(\xi_j) - \xi_{s+1} = 0, \ \xi_i \in A_i + A_i \ (i = 1, \ldots, s+1)$$

is solvable if

$$\prod_{i=1}^{s+1} |A_i| > (q-|A_{s+1}|)q^{s/2} \prod_{j=1}^{s} (c_j - 1).$$

If we take $A_1 = \ldots = A_s = K$, we find

Theorem 6. *The equation*

$$\sum_{j=1}^{s} f_j(\xi_j) - \xi_{s+1} = 0, \ \xi_j \in K, \ \xi_{s+1} \in A_{s+1} + A_{s+1}$$

is solvable if

$$|A_{s+1}| \geqq q^{(2-s)/2} \prod_{j=1}^{s} (c_j - 1).$$

In theorem 6 as well as in theorem 7 the condition "c_j is not divisible by p" may be replaced by the condition "f_j is not of the form $g^p - g + \beta$ where g is a polynomial over K and β is an element of K" (for a proof, cf. [10]). It should be noted, too, that the consideration of the special case $A_{s+1} = \{0\}$ shows that theorem 6 implies corollary 2 of theorem 2.

Let u be a non-negative integer. Denote

$$A_{s+1} = \{\xi \in K \mid 0 \leqq \mathrm{tr}(\xi) \leqq u\}.$$

Then $|A_{s+1}| = (u+1)p^{n-1}$ and-.

$$A_{s+1} + A_{s+1} = \{\xi \in K \mid 0 \leqq \mathrm{tr}(\xi) \leqq 2u\}.$$

Assume now that

$$\mathrm{p}^{(2-ns)/2} \prod_{j=1}^{s} (c_j - 1) - 1 \leqq u < p^{(2-ns)/2} \prod_{j=1}^{s} (c_j - 1).$$

Then, by theorem 6, there exist elements $\xi_1, \ldots, \xi_s$ of K such that

$$0 \leqq \mathrm{tr}(\sum_{j=1}^{s} f_j(\xi_j)) \leqq 2u.$$

We write this result in the following form.

Theorem 7. *The least non-negative trace of the polynomial* $\sum_{j=1}^{s} f_j(\xi_j)$ *is* $< 2p^{(2-ns)/2} \prod_{j=1}^{s} (c_j - 1)$.

Theorem 7 is an extension for the equation (2) of [10] ; this result was an improvement for a result of CAVIOR's [3]. The special case $q = p$ of theorem 7 can be written as

Corollary. *l, the least non-negative residue of the polynomial* $\sum_{j=1}^{s} f_j(x_j)$ (mod p), *satisfies the inequality* (5).

7. Some further results. In the special case

$$f(\mathbf{x}) = \sum_{j=1}^{s} \gamma_j \xi_j^c - \gamma$$

we can prove the following result.

Theorem 8. *Let d be the* g.c.d. *of c and $q-1$. Suppose that $d-1$ is equal to $aq^{(s-2)/2s}$ where $a < 1$. Then the equation*

$$\sum_{j=1}^{s} \gamma_j \xi_j^c = \gamma, \quad \xi_j \in A + A, \tag{24}$$

where $\gamma_1 \ldots \gamma_s \neq 0$, is solvable if

$$|A| \geqq (1-a)^{-1} d^{-1/2} q^{(s+1)/s}.$$

This theorem, in the proof of which we do not use the deep lemma 1, gives better results than corollary 1 of theorem 2 if

$$q < (1-a)^2 d(d-1)^2.$$

Proof of theorem 8. We have now, by (15),

$$M = q^{-s-1} \sum_{\kappa} \sum_{\mathbf{z}} e(\kappa(\sum_{j=1}^{s} \gamma_j \zeta_j^c - \gamma)) \sum_{\mathbf{x}}' \sum_{\mathbf{y}}' \sum_{\mathbf{t}} e(\mathbf{t}(\mathbf{x}+\mathbf{y}-\mathbf{z}))$$

$$= q^{-s-1} \sum_{\kappa} e(-\kappa\gamma) \prod_{j=1}^{s} \sum_{\xi_j}' \sum_{\eta_j}' \sum_{\zeta_j} \sum_{\tau_j} e(\tau_j(\xi_j + \eta_j - \zeta_j) + \kappa\gamma_j\zeta_j^c)$$

$$= q^{-s-1} \sum_{\kappa} e(-\kappa\gamma) \prod_{j=1}^{s} \sum_{\tau_j} (S_0(-\tau_j))^2 \sum_{\zeta_j} e(\kappa\gamma_j\zeta_j^c - \tau_j\zeta_j)$$

where S_0 is defined by (22). Picking out the term with $\kappa = 0$, we get furthermore, by (23) and lemma 2,

$$M = q^{-1}|A|^{2s} +$$

$$q^{-s-1} \sum_{\kappa\neq 0} e(-\kappa\gamma) \prod_{j=1}^{s} (\theta_j(d-1)q^{1/2}|A|^2 + \sum_{\tau_j\neq 0} (S_0(-\tau_j))^2 \sum_{\zeta_j} e(\kappa\gamma_j\zeta_j^c - \tau_j\zeta_j))$$

where θ_j is a complex number such that $|\theta_j| \leqq 1$. This implies, by (23) and lemma 3,

$$M \geqq q^{-1}|A|^{2s} - q^{-(s+2)/2}(q-1)|A|^s((d-1)|A| + d^{-1/2}q^{1/2}(q-|A|))^s$$

$$> q^{-1}|A|^s(|A|^s - q^{(2-s)/2}((d-1)|A| + d^{-1/2}q^{3/2})^s).$$

Consequently $M > 0$ if

$$|A| \geqq q^{(2-s)/2s}((d-1)|A| + d^{-1/2}q^{3/2})$$

or if

$$|A| \geqq \frac{d^{-1/2}q^{(s+1)/s}}{1-(d-1)q^{(2-s)/2s}} = (1-a)^{-1}d^{-1/2}q^{(s+1)/s}.$$

Thus theorem 8 has been proved.

Theorem 8 is an analogy of corollary 1 of theorem 2. As an analogy of the special case $s = 1$ of theorem 7 we get

Theorem 9. *Let α and β be non-zero elements of K. Then the least non-negative trace of the polynomial $\alpha\xi^c + \beta\xi$ is $< 2d^{-1/2}p$ where d is the* g.c.d. *of c and $q-1$.*

Proof. Define the integer u by the inequalities

$$d^{-1/2}p - 1 \leqq u < d^{-1/2}p. \tag{25}$$

By [10], the congruence

References

[1] Н. М. Акулиничев: Оценки рациональных тригонометрических сумм специального вида. — Докл. Акад. наук СССР 161 (1965), 743—745.

[2] L. Carlitz and S. Uchiyama: Bounds for exponential sums. — Duke Math. J. **24** (1957), 37—41.

[3] S. R. Cavior: On the least non-negative trace of a polynomial over a finite field. — Boll. Un. Mat. Ital. **20** (1965), 120—121.

[4] J. H. H. Chalk: The number of solutions of congruences in incomplete residue systems. — Canad. J. Math. **15** (1963), 291—296.

[5] J. H. H. Chalk and K. S. Williams: The distribution of solutions of congruences. — Mathematika **12** (1965), 176—192.

[6] L.-K. Hua and H. S. Vandiver: On the existence of solutions of certain equations in a finite field. — Proc. Nat. Acad. Sci. USA **34** (1948), 258—263.

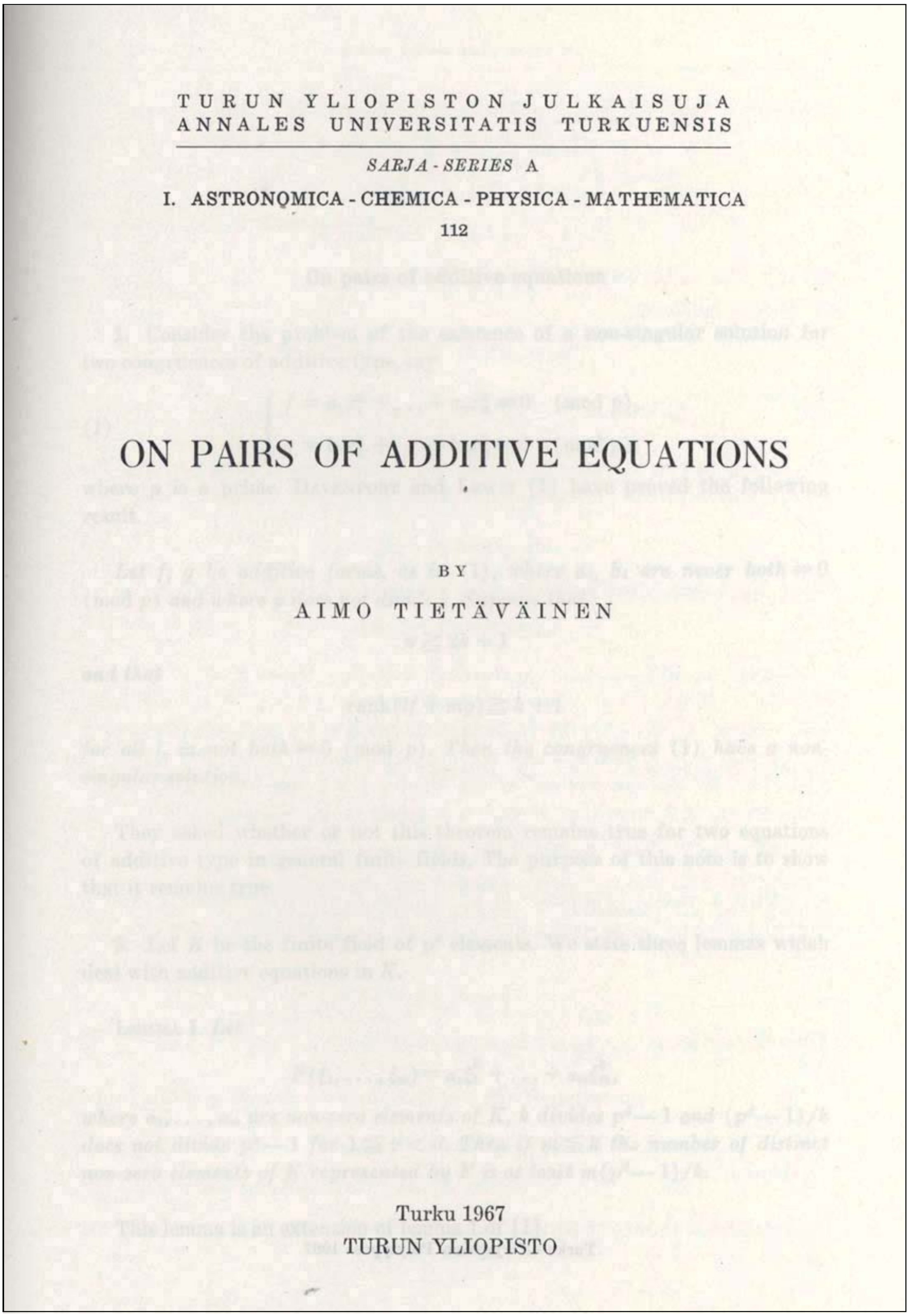
[7] L. J. Mordell: On the number of solutions in incomplete residue sets of quadratic congruences. — Arch. Math. **8** (1957), 153—157.

[8] L. J. Mordell: On the least residue and non-residue of a polynomial. — J. London Math. Soc. **38** (1963), 451—453.

[9] L. J. Mordell: Incomplete exponential sums and incomplete residue systems for congruences. — Czechoslovak Math. J. **14** (1964), 235—242.

[10] A. Tietäväinen: On the trace of a polynomial over a finite field. — Ann. Univ. Turku., Ser. A I **87** (1966).

[11] I. M. Vinogradov: The method of trigonometrical sums in the theory of numbers. London and New York (1954).



TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SERIES A

I. ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA

112

ON PAIRS OF ADDITIVE EQUATIONS

BY

AIMO TIETÄVÄINEN

Turku 1967
TURUN YLIOPISTO

Turku - Kirjapaino Polytypos - 1967

On pairs of additive equations

1. Consider the problem of the existence of a non-singular solution for two congruences of additive type, say

$$(1)\qquad \begin{cases} f = a_1x_1^k + \ldots + a_nx_n^k \equiv 0 \pmod{p}, \\ g = b_1x_1^k + \ldots + b_nx_n^k \equiv 0 \pmod{p}, \end{cases}$$

where p is a prime. Davenport and Lewis [1] have proved the following result.

Let f, g be additive forms, as in (1), *where a_i, b_i are never both $\equiv 0 \pmod{p}$ and where p does not divide k. Suppose that*

$$n \geqq 2k + 1$$

and that

$$\operatorname{rank}(lf + mg) \geqq k + 1$$

for all l, m not both $\equiv 0 \pmod{p}$. Then the congruences (1) *have a non-singular solution.*

They asked whether or not this theorem remains true for two equations of additive type in general finite fields. The purpose of this note is to show that it remains true.

2. Let K be the finite field of p^d elements. We state three lemmas which deal with additive equations in K.

Lemma 1. *Let*

$$F(\xi_1, \ldots, \xi_m) = \alpha_1\xi_1^k + \ldots + \alpha_m\xi_m^k,$$

where $\alpha_1, \ldots, \alpha_m$ are non-zero elements of K, k divides $p^d - 1$ and $(p^d - 1)/k$ does not divide $p^v - 1$ for $1 \leqq v < d$. Then if $m \leqq k$ the number of distinct non-zero elements of K represented by F is at least $m(p^d - 1)/k$.

This lemma is an extension of lemma 1 of [1].

Proof of lemma 1 (cf. [2]). Let A be a subset of K. Define

$$|A| = \text{the number of elements in } A,$$

$$A^* = \{\eta \in A \mid \eta \neq 0\},$$

$$H(A) = \{\eta \mid \alpha + \eta \in A \ \forall \ \alpha \in A\}.$$

Clearly $H(A)$ is an additive subgroup of K.

Define, moreover,

$$K_0 = \{\eta \mid \eta = \xi^k, \ \xi \in K\},$$

$$Q_w = Q_w(\alpha_1, \dots, \alpha_w) = \{\eta \mid \eta = \sum_{j=1}^{w} \alpha_j \xi_j^k, \ \xi_j \in K\}.$$

If η is in Q_w^* so ist the set ηK_0^*. Hence Q_w^* is the union of some cosets of the multiplicative group K^* modulo K_0^*. Thus there is an integer l_w such that

$$|Q_w^*| = l_w(p^d - 1)/k.$$

Since $l_1 = 1$, our assertion can be now written in the following form:

$$l_{w+1} \geqq \min(1 + l_w, k).$$

Clearly $l_{w+1} \geqq l_w$. If $l_{w+1} = l_w$ then $\alpha_{w+1} \in H(Q_w)$. Hence $H(Q_w) \neq \{0\}$ and therefore $|H(Q_w)| = p^v$ where $v \geqq 1$. If η is in $(H(Q_w))^*$ so is the set ηK_0^*. Hence $(H(Q_w))^*$ is the union of some cosets of the multiplicative group K^* modulo K_0^* and consequently $(p^d - 1)/k$ divides $|(H(Q_w))^*| = p^v - 1$. Therefore, by the assumptions of the lemma, $v = d$ and hence $H(Q_w) = K$. Since $0 \in Q_w$ it follows that $H(Q_w)$ is a subset of Q_w. Consequently $Q_w = K$ and hence $l_w = k$. Thus $l_{w+1} \geqq \min(1 + l_w, k)$.

LEMMA 2. *Let k be a factor of $p^d - 1$. Let G, H be any forms of degree k over K in $\eta_1, \dots, \eta_s$, where $s > k$. Let $\gamma_1, \dots, \gamma_u$ be distinct non-zero elements of K where*

$$u > (2k - s)(p^d - 1)/k.$$

Then there exist $\eta_1, \dots, \eta_s$, not all zero, such that

$$G(\eta_1, \dots, \eta_s) = 0$$

and

$$H(\eta_1, \dots, \eta_s) = 0 \text{ or } \gamma_i$$

for some i.

Lemma 2 is an immediate extension of lemma 2 of [1].

Lemma 3. *Let k be a factor of $p^d - 1$. Let v be a factor of d such that $p^v - 1$ is divisible by $(p^d - 1)/k$. Then the equations*

$$\gamma_1\eta_1^k + \ldots + \gamma_t\eta_t^k = 0, \ \delta_1\eta_1^k + \ldots + \delta_t\eta_t^k = 0$$

have a non-trivial solution in K if

$$t \geqq 1 + \frac{2dk(p^v - 1)}{v(p^d - 1)}.$$

This lemma follows from theorem 5 of [3].

3. We now state

Theorem. *Let*

$$\left\{ \begin{aligned} f &= \alpha_1\xi_1^k + \ldots + \alpha_n\xi_n^k = 0, \\ g &= \beta_1\xi_1^k + \ldots + \beta_n\xi_n^k = 0 \end{aligned} \right. \tag{2}$$

be additive equations over K. Suppose that p does not divide k and that α_i, β_i are never both $= 0$. Suppose, furthermore, that

$$n \geqq 2k + 1$$

and that

$$\operatorname{rank}(\lambda f + \mu g) \geqq k + 1$$

for all λ, μ not both $= 0$. Then the equations (2) have a non-singular solution in K.

Proof. We can suppose without loss of generality that k divides $p^d - 1$. Assume that the equations (2) have singular solutions only. As Davenport and Lewis ([1], p. 342) have shown, instead of considering (2) we then, by permuting the variables and by interchanging f, g if necessary, may consider the pair

$$\left\{ \begin{aligned} \alpha_1\xi_1^k + \ldots + \alpha_r\xi_r^k + \gamma_1\eta_1^k + \ldots + \gamma_s\eta_s^k &= 0, \\ \delta_1\eta_1^k + \ldots + \delta_s\eta_s^k &= 0, \end{aligned} \right. \tag{3}$$

where all the α_i and all the δ_i are non-zero and the equation

$$\alpha_1\xi_1^k + \ldots + \alpha_r\xi_r^k = 0 \tag{4}$$

has a non-trivial solution. Furthermore, we have

$$s \geqq k + 1. \tag{5}$$

Case 1. *$(p^d-1)/k$ does not divide p^v-1 for $1 \leqq v < d$.* The proof of this case is similar to that of the theorem of DAVENPORT and LEWIS. However, lemmas 1 and 2 of the present paper are used instead of their lemmas 1 and 2.

Case 2. *There exists an integer v such that $1 \leqq v < d$ and p^v-1 is divisible by $(p^d-1)/k$.* Now $d \geqq 2$ and therefore $p^d \geqq 4$. Furthermore, we can assume that v is a divisor of d and consequently $d \geqq 2v$, for $(p^d-1,\ p^v-1) = p^{(d,v)}-1$ and therefore $(p^d-1)/k$, which is a common divisor of p^d-1 and p^v-1, is a divisor of $p^{(d,v)}-1$.

It can be shown that the equations (3) have a non-singular solution if the equations

$$(6) \qquad \begin{cases} \alpha_r \xi_r^k + \gamma_1 \eta_1^k + \ldots + \gamma_s \eta_s^k = 0, \\ \qquad\quad \delta_1 \eta_1^k + \ldots + \delta_s \eta_s^k = 0 \end{cases}$$

have a non-trivial solution. Indeed, if $(\xi_r', \eta_1', \ldots, \eta_s')$ is a non-trivial solution of (6) and $(\xi_1'', \ldots, \xi_r'')$ is a non-trivial solution of (4), then the equations (3) have the non-singular solution $(0, \ldots, 0, \xi_r', \eta_1', \ldots, \eta_s')$ in case $\xi_r' \neq 0$ and the non-singular solution $(\xi_1'', \ldots, \xi_r'', \eta_1', \ldots, \eta_s')$ in case $\xi_r' = 0$.

Suppose firstly that $p^d = 4$. Case $k = 1$ is excluded, whence we may assume that $k = 3$. Since $s \geqq 4$, the equations (6) have, by lemma 3, a non-trivial solution.

Suppose now that $p^d > 4$. If $p^v = 2$ then

$$(p^d-1)/(p^v-1) = 2^d - 1 > 2d = 2d/v,$$

since $d \geqq 3$. If $p^v \geqq 3$ then

$$(p^d-1)/(p^v-1) = 1 + p^v + \ldots + p^{d-v} \geqq 1 + 3(d/v-1) \geqq 2d/v,$$

because $d \geqq 2v$. Hence, in every case,

$$v(p^d-1) \geqq 2d(p^v-1)$$

and consequently, by (5),

$$s \geqq 1 + \frac{2dk(p^v-1)}{v(p^d-1)}.$$

This inequality implies, by lemma 3, that the equations (6) have a non-trivial solution.

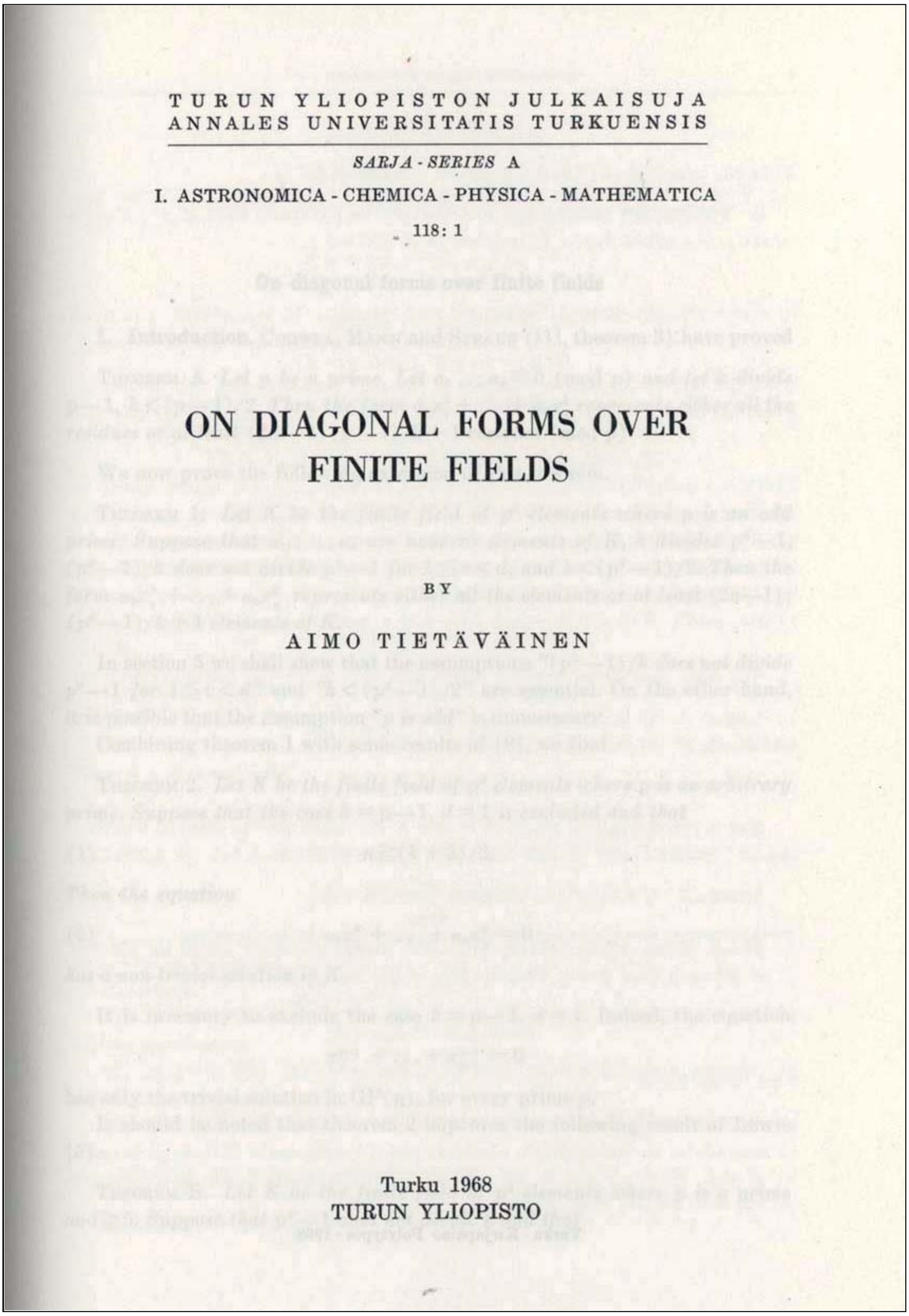
UNIVERSITY OF TURKU
TURKU, FINLAND

References

[1] H. Davenport and D. J. Lewis: Notes on congruences (III). — Quart. J. Math. Oxford (2) 17 (1966), 339—344.

[2] D. J. Lewis: Diagonal forms over finite fields. — Norske Vid. Selsk. Forh. (Trondheim) 33 (1960), 61—65.

[3] A. Tietäväinen: On the non-trivial solvability of some equations and systems of equations in finite fields. — Ann. Acad. Sci. Fenn., Ser. A I 360 (1965), 38 pp.



TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SERIES A

I. ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA

118: 1

ON DIAGONAL FORMS OVER FINITE FIELDS

BY

AIMO TIETÄVÄINEN

Turku 1968
TURUN YLIOPISTO

Turku · Kirjapaino Polytypos · 1968

On diagonal forms over finite fields

1. Introduction. Chowla, Mann and Straus ([1], theorem 3) have proved

Theorem A. *Let p be a prime. Let $a_1 \dots a_n \not\equiv 0$* (mod p) *and let k divide $p-1$, $k < (p-1)/2$. Then the form $a_1x_1^k + \dots + a_nx_n^k$ represents either all the residues or at least $(2n-1)(p-1)/k+1$ residues* (mod p).

We now prove the following extension of this theorem.

Theorem 1. *Let K be the finite field of p^d elements where p is an odd prime. Suppose that $\alpha_1, \dots, \alpha_n$ are nonzero elements of K, k divides p^d-1, $(p^d-1)/k$ does not divide p^v-1 for $1 \leqq v < d$, and $k < (p^d-1)/2$. Then the form $\alpha_1x_1^k + \dots + \alpha_nx_n^k$ represents either all the elements or at least $(2n-1) \cdot (p^d-1)/k+1$ elements of K.*

In section 5 we shall show that the assumptions "*$(p^d-1)/k$ does not divide p^v-1 for $1 \leqq v < d$*" and "$k < (p^d-1)/2$" are essential. On the other hand, it is possible that the assumption "*p is odd*" is unnecessary.

Combining theorem 1 with some results of [8], we find

Theorem 2. *Let K be the finite field of p^d elements where p is an arbitrary prime. Suppose that the case $k = p-1$, $d = 1$ is excluded and that*

$$n \geqq (k+3)/2. \tag{1}$$

Then the equation

$$\alpha_1x_1^k + \dots + \alpha_nx_n^k = 0 \tag{2}$$

has a non-trivial solution in K.

It is necessary to exclude the case $k = p-1$, $d = 1$. Indeed, the equation

$$x_1^{p-1} + \dots + x_{p-1}^{p-1} = 0$$

has only the trivial solution in GF(p), for every prime p.

It should be noted that theorem 2 improves the following result of Lewis [5].

Theorem B. *Let K be the finite field of p^d elements where p is a prime and $\geqq 5$. Suppose that p^d-1 does not divide k and that*

$$n \geqq (k+5)/(2-3/p).$$

Then the equation (2) *has a non-trivial solution in* K.

2. Preliminary results. Let $K = \mathrm{GF}(p^d)$ be the finite field of p^d elements where p is a prime. Let A be a subset of K. Define

$$|A| = \text{the cardinality of } A,$$

$$A^* = \{\eta \in A \mid \eta \neq 0\},$$

$$H(A) = \{\eta \in K \mid A + \eta = A\},$$

$$K_\circ = \{\eta \mid \eta = \xi^k, \xi \in K\},$$

$$Q_w = Q_w(\alpha_1, \dots, \alpha_w) = \{\eta \mid \eta = \sum_{j=1}^{w} \alpha_j \xi_j^k, \ \xi_j \in K\}$$

where k is a factor of $p^d - 1$. If $H(A) \neq \{0\}$, A is said to be periodic. If $H(A) = \{0\}$, A is aperiodic. Clearly $H(A)$ is an additive subgroup of K, and A is the union of some additive cosets of K modulo $H(A)$. Furthermore (see [5]), there is an integer l_w such that

$$|Q_w^*| = l_w(p^d - 1)/k.$$

We state now seven lemmas which will be used in the following sections.

LEMMA 1. *If* $\alpha_1, \dots, \alpha_{w+1}$ *are non-zero elements of* K *and* $(p^d-1)/k$ *does not divide* $p^v - 1$ *for* $1 \leqq v < d$ *then*

$$l_{w+1} \geqq \min(1 + l_w, k).$$

For a proof, see [9], proof of lemma 1. For some special cases of lemma 1, see [2], lemma 1, and [7].

LEMMA 2. *If* $k \leqq (p^d-1)/2$ *then* $\sum_{\alpha \in Q_w} \alpha = 0$.

Proof. Since $k \leqq (p^d-1)/2$, $|K_\circ^*| \geqq 2$. Therefore there exists an element β of $K_\circ^*$ such that $\beta \neq 1$. Clearly $\beta Q_w = Q_w$. Let $\sum_{\alpha \in Q_w} \alpha = \delta$. Then

$$\beta\delta = \sum_{\alpha \in Q_w} \beta\alpha = \sum_{\gamma \in Q_w} \gamma = \delta$$

and hence $\delta = 0$.

LEMMA 3. *If* A *is a periodic subset of* K *and* $p > 2$, *then* $\sum_{\alpha \in A} \alpha = 0$.

Proof. Since $H(A)$ is an additive subgroup of K, $H(A) \neq \{0\}$, and $p > 2$, then $2H(A) = H(A)$. Let $\sum_{\alpha \in H(A)} \alpha = \gamma$. Then

$$2\gamma = \sum_{\alpha \in H(A)} 2\alpha = \sum_{\beta \in H(A)} \beta = \gamma$$

and hence $\gamma = 0$. Let $B = \delta + H(A)$ be an additive coset of K modulo $H(A)$. Then

$$\sum_{\alpha \in B} \alpha = |H(A)|\,\delta + \sum_{\alpha \in H(A)} \alpha = 0,$$

because p divides $|H(A)|$. Since A is the union of some additive cosets of K modulo $H(A)$, this equation implies lemma 3.

Lemma 4. *Suppose that $(p^d - 1)/k$ does not divide $p^v - 1$ for $1 \leqq v < d$ and that A is a non-empty periodic subset of K such that $\alpha A = A$ for every element α of K_0^*. Then $A = K$.*

Proof. Since A is periodic, $H(A) \neq \{0\}$ and therefore $|H(A)| = p^v$ where $v \geqq 1$. If $\beta \in (H(A))^*$ then $\beta\alpha + A = \beta\alpha + \alpha A = \alpha(\beta + A) = \alpha A = A$ for every element α of K_0^* and therefore βK_0^* is a subset of $(H(A))^*$. Hence $(H(A))^*$ is the union of some cosets of the multiplicative group K^* modulo K_0^* and consequently $(p^d - 1)/k$ divides $|(H(A))^*| = p^v - 1$. Therefore, by the assumptions of the lemma, $v = d$ and hence $H(A) = K$. Consequently $A = K$.

Lemma 5. *Let A and B be subsets of K satisfying*

$$|A + B| \leqq |A| + |B| - 2.$$

Then $A + B$ is periodic.

Lemma 5 is due to Kneser [4].

Lemma 6. *Let A and B be subsets of K such that*

$$|A + B| = |A| + |B| - 1.$$

Let $\gamma_1, \dots, \gamma_t$ denote all the elements in $A + B$ having only one representation as $\gamma_j = \alpha_j + \beta_j$ $(\alpha_j \in A, \beta_j \in B)$.

Assertion:

(1) *If $t = 0$ then $A + B$ is either periodic or can be made periodic by adding one element.*

(2) *If $t = 1$ then $A + B$ is either periodic or can be made periodic by deleting one element.*

(3) *If $t \geqq 3$ then either $\alpha_1 = \cdots = \alpha_t$ or $\beta_1 = \cdots = \beta_t$. Moreover, every element in $A + B$, other than γ_i, has at least t representations as the sum of an element of A and an element of B.*

This lemma is a special case of theorem 6. 1. of [3].

LEMMA 7. *Let v be a factor of d such that p^v-1 is divisible by $(p^d-1)/k$. Then the equation* (2) *has a non-trivial solution in K if*

$$n \geqq 1+\frac{dk(p^v-1)}{v(p^d-1)}.$$

Lemma 7 follows from theorem 5 of [8].

3. Proof of theorem 1. Since $l_1=1$, the assertion can be written in the following form:

$$l_{s+1} \geqq \min(2+l_s, k) \qquad (s=1, \ldots, n-1). \tag{3}$$

By lemma 1, $l_{s+1} \geqq \min(1+l_s,\ k)$. Suppose that $l_{s+1}=1+l_s$. We shall say (see [5]) that an element γ of Q_{s+1} has a unique representation in Q_{s+1} if it has only one representation as $\gamma=\alpha+\beta$ with $\alpha \in Q_s,\ \beta \in \alpha_{s+1}K_o$. Let t be the number of the elements in Q_{s+1} which have unique representations in Q_{s+1}. If γ has a unique representation in Q_{s+1} then so does every element in the set γK^*_o. Hence $t \equiv 1 \pmod m$ if 0 has only the trivial representation in Q_{s+1}, and $t \equiv 0 \pmod m$ if 0 has at least one non-trivial representation in Q_{s+1}. Here and hereafter $m=(p^d-1)/k$.

Case 1. Q_{s+1} is periodic. Then $l_{s+1}=k$ by lemma 4. Hence the assertion (3) is true in this case.

Case 2. Q_{s+1} is aperiodic and $t=0$. Then, by lemma 6, Q_{s+1} can be made periodic by adding one element. Let β be this element. Then

$$\sum_{\alpha \in Q_{s+1}} \alpha = -\beta$$

by lemma 3. On the other hand,

$$\sum_{\alpha \in Q_{s+1}} \alpha = 0$$

by lemma 2. Hence $\beta=0$. This is impossible, since $0 \in Q_{s+1}$.

Case 3. Q_{s+1} is aperiodic and $t=1$. Then, by lemma 6, Q_{s+1} can be made periodic by deleting one element. By lemma 3 and lemma 2, this element is 0. Therefore Q^*_{s+1} is periodic. This is impossible by lemma 4.

Case 4. Q_{s+1} is aperiodic and $t=2$. This case is impossible, since $m>2$.

Case 5. Q_{s+1} is aperiodic and $t \geqq 3$. Let $\delta_1, \ldots, \delta_t$ be the uniquely represented elements in Q_{s+1}. Then, by lemma 6, either there exists some unique element β of Q_s such that

$$\delta_j=\beta+\alpha_{s+1}\gamma_j^k \ (j=1, \ldots, t), \tag{4}$$

or there exists some unique element γ^k of K_o such that

$$\delta_j = \beta_j + \alpha_{s+1}\gamma^k \quad (j = 1, \ldots, t) \tag{5}$$

where the β_j are elements of Q_s.

Suppose that (4) holds. Let ε^k be $\neq 0, 1$. Then $\varepsilon^k\delta_j$ has the unique representation

$$\varepsilon^k\delta_j = \varepsilon^k\beta + \alpha_{s+1}(\gamma_j\varepsilon)^k$$

in Q_{s+1}. Consequently $\varepsilon^k\delta_j$ is one of the elements $\delta_1, \ldots, \delta_t$ and therefore $\varepsilon^k\beta = \beta$. Hence $\beta = 0$. It follows from this that the set U of the uniquely represented elements in Q_{s+1} is $\alpha_{s+1}K_0$ or $\alpha_{s+1}K_0^*$. If $U = \alpha_{s+1}K_0$ then (see [5]) Q_s^* is periodic and therefore, by lemma 4, $Q_s^* = K$ which is an impossibility. If $U = \alpha_{s+1}K_0^*$ then

$$Q_s^* + \alpha_{s+1}K_0^* = Q_s$$

and hence, by lemma 5, Q_s is periodic. Consequently, by lemma 4, $Q_s = K$. Hence $Q_{s+1} = K$ which is an impossibility, because Q_{s+1} is aperiodic.

Suppose now that (5) holds. Then the product $\varepsilon^k\delta_j$, where $\varepsilon^k \neq 0, 1$, has the unique representation

$$\varepsilon^k\delta_j = \varepsilon^k\beta_j + \alpha_{s+1}\varepsilon^k\gamma^k$$

in Q_{s+1} and therefore $\alpha_{s+1}\varepsilon^k\gamma^k = \alpha_{s+1}\gamma^k$. Consequently $\gamma = 0$. If 0 has the trivial representation only, then, by [5], there exists a non-zero element α of K such that the difference set $Q_{s+1} \setminus \alpha K_0$ is periodic. This is impossible by lemma 4. Hence there exists a positive integer g such that $t = gm$, $g \leqq l_s$. Now, by lemma 6, those elements in Q_{s+1} not uniquely represented in Q_{s+1} have at least gm representations. As there are $(1 + l_s m)(1 + m)$ sums of an element of Q_s and an element of $\alpha_{s+1}K_0$, we get

$$(1 + (1 + l_s - g)m)gm + gm \leqq (1 + l_s m)(1 + m)$$

or

$$m^2(g - 1 - m^{-1})(g - l_s - m^{-1}) \geqq 0.$$

Since $g < l_s + m^{-1}$, we have the inequality $g \leqq 1 + m^{-1}$. Hence $g = 1$. Suppose that αK_0^* is the subset of Q_s which consists of exactly those elements which have a unique representation in Q_{s+1}. Then

$$Q_s + \alpha_{s+1}K_0^* = Q_{s+1} \setminus \alpha K_0^*$$

and consequently, by lemma 5, $Q_{s+1} \setminus \alpha K_0^*$ is periodic. This is impossible by lemma 4.

Now we have shown that the equation $l_{s+1} = 1 + l_s$ implies the equation $l_{s+1} = k$. Hence the inequality (3) is true.

4. Proof of theorem 2. We may assume that the coefficients $\alpha_1, \ldots, \alpha_n$ are non-zero, for if $\alpha_j = 0$, then the equation (2) has the non-trivial solution

$x_j = 1$, $x_k = 0$ $(k \neq j)$. Furthermore, we can assume that k divides $p^d - 1$, for the equation $x^k = \alpha$ is soluble if and only if the equation $y^{(k, p^d - 1)} = \alpha$ is soluble.

Case 1. Suppose that all the assumptions of theorem 1 are satisfied. Then, by theorem 1,

$$F(x_1, \ldots, x_{n-1}) = \alpha_1 x_1^k + \ldots + \alpha_{n-1} x_{n-1}^k$$

represents all the elements of K. In particular, there exist elements $\xi_1, \ldots, \xi_{n-1}$ in K such that $F(\xi_1, \ldots, \xi_{n-1}) = -\alpha_n$. Hence the equation (2) has the non-trivial solution $(\xi_1, \ldots, \xi_{n-1}, 1)$.

Case 2. Suppose that there exists an integer v such that $1 \leqq v < d$ and $p^v - 1$ is divisible by $(p^d - 1)/k$. Now $d \geqq 2$ and therefore $p^d \geqq 4$. Furthermore (see [9]), we can assume that v is a divisor of d and consequently $d \geqq 2v$.

Suppose firstly that $p^d = 4$. Case $k = 1$ is excluded, whence we may assume that $k = 3$. Since $n \geqq 3$, the equation (2) has, by lemma 7, a non-trivial solution in K.

Suppose now that $p^d > 4$. If $p^v = 2$ then

$$(p^d - 1)/(p^v - 1) = 2^d - 1 > 2d = 2d/v$$

since $d \geqq 3$. If $p^v \geqq 3$ then

$$(p^d - 1)/(p^v - 1) = 1 + p^v + \ldots + p^{d-v} \geqq 1 + 3(d/v - 1) \geqq 2d/v$$

because $d \geqq 2v$. Hence, in every case,

$$v(p^d - 1) \geqq 2d(p^v - 1)$$

and consequently, by (1),

$$n \geqq \frac{3}{2} + \frac{dk(p^v - 1)}{v(p^d - 1)}.$$

This inequality implies, by lemma 7, that the equation (2) has a non-trivial solution in K.

Case 3. Suppose $p = 2$. Then -1 is a kth power in K and therefore (2) is a so-called A-equation (see [8]). Using the results of the sections 18, 21 and 22 of [8], we can show, by means of some simple numerical calculations, that the equation (2) has a non-trivial solution in this case, too.

Case 4. Suppose $k \geqq (p^d - 1)/2$. If $k = (p^d - 1)/2$ then -1 is a kth power in K and the consideration is similar to that of case 3. If $k > (p^d - 1)/2$ then $k = p^d - 1$, $d > 1$, and hence $(p^d - 1)/k$ is a divisor of $p - 1$. Since $d > 1$, this case is therefore a special case of case 2.

Thus theorem 2 has been proved.

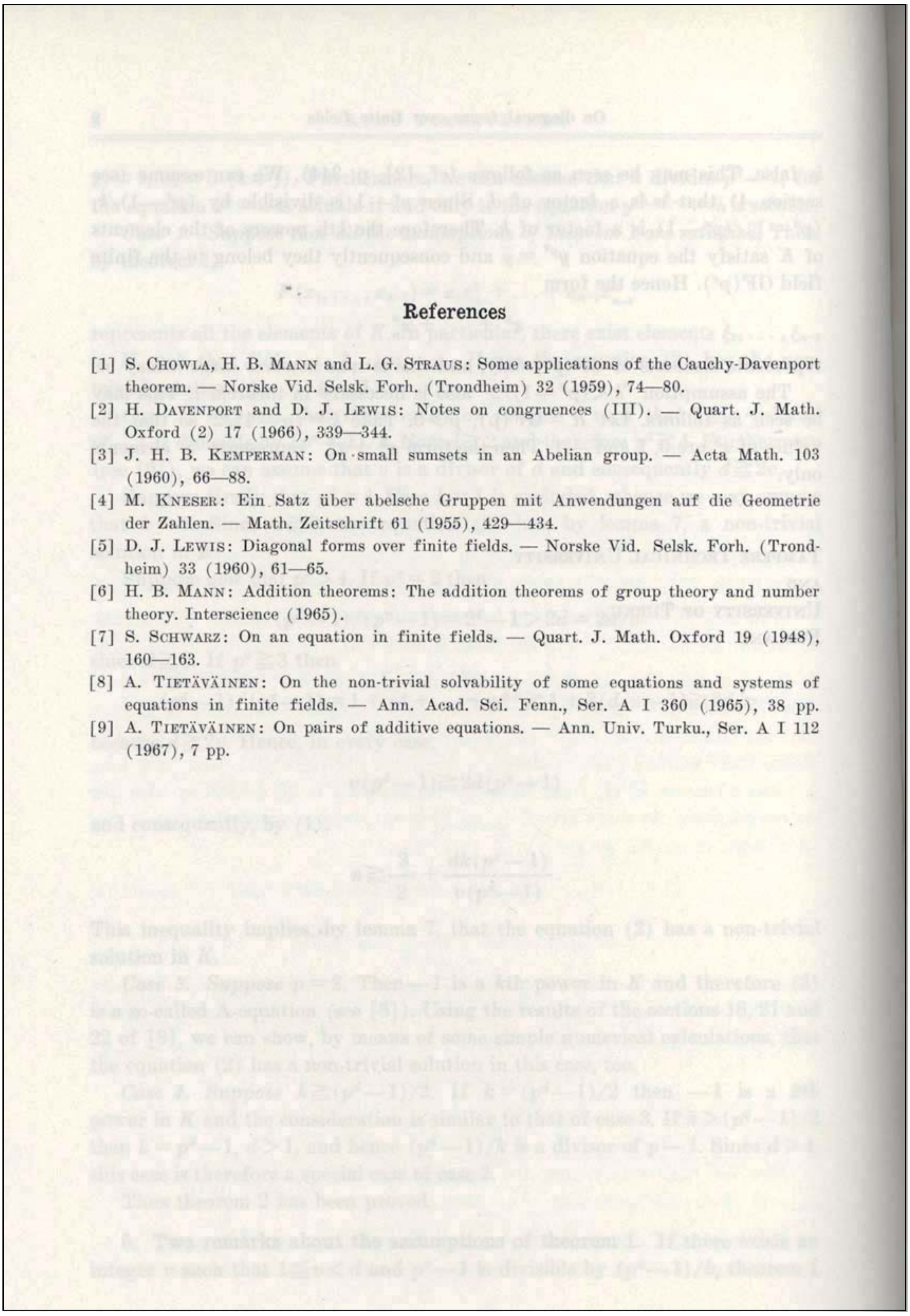
5. Two remarks about the assumptions of theorem 1. If there exists an integer v such that $1 \leqq v < d$ and $p^v - 1$ is divisible by $(p^d - 1)/k$, theorem 1 is false. This may be seen as follows (cf. [2], p. 344). We can assume (see section 4) that v is a factor of d. Since $p^v—1$ is divisible by $(p^d—1)/k$, $(p^d—1)/(p^v—1)$ is a factor of k. Therefore the kth powers of the elements of K satisfy the equation $y^{p^v} = y$ and consequently they belong to the finite field $\mathrm{GF}(p^v)$. Hence the form

$$x_1^k + \dots + x_n^k$$

represents elements of $\mathrm{GF}(p^v)$ only, and this is true however large n may be.

The assumption ”$k < (p^d—1)/2$” also is necessary in theorem 1. This may be seen as follows. Let $K = \mathrm{GF}(p)$, $p > 5$. Take $k = (p—1)/2$, so that the values of x^k are 0, 1 and —1. Then the form $x_1^k + x_2^k$ represents 5 elements only.

Tampere Technical University
and
University of Turku,
Finland



References

[1] S. Chowla, H. B. Mann and L. G. Straus: Some applications of the Cauchy-Davenport theorem. — Norske Vid. Selsk. Forh. (Trondheim) 32 (1959), 74—80.

[2] H. Davenport and D. J. Lewis: Notes on congruences (III). — Quart. J. Math. Oxford (2) 17 (1966), 339—344.

[3] J. H. B. Kemperman: On small sumsets in an Abelian group. — Acta Math. 103 (1960), 66—88.

[4] M. Kneser: Ein Satz über abelsche Gruppen mit Anwendungen auf die Geometrie der Zahlen. — Math. Zeitschrift 61 (1955), 429—434.

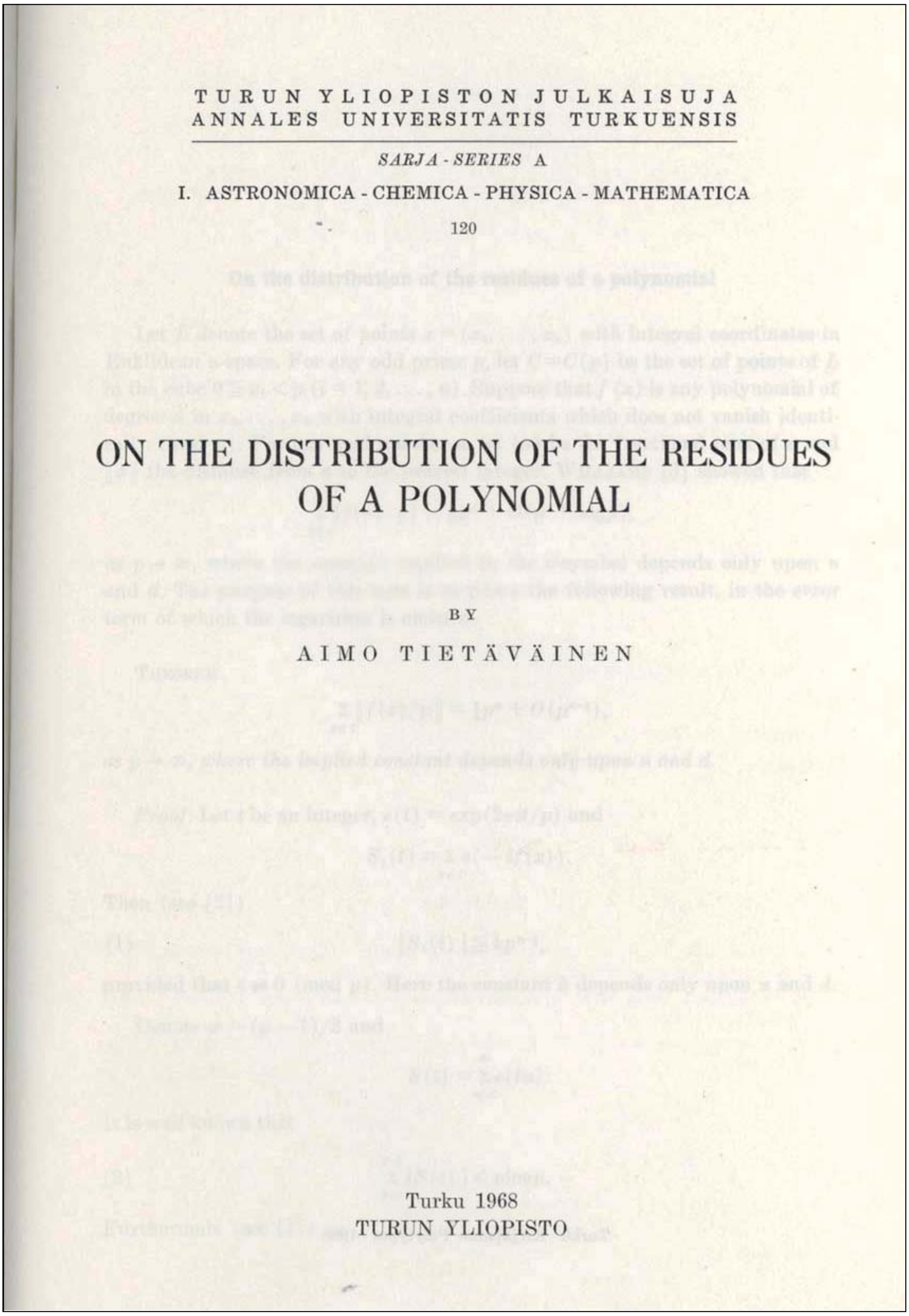
[5] D. J. Lewis: Diagonal forms over finite fields. — Norske Vid. Selsk. Forh. (Trondheim) 33 (1960), 61—65.

[6] H. B. Mann: Addition theorems: The addition theorems of group theory and number theory. Interscience (1965).

[7] S. Schwarz: On an equation in finite fields. — Quart. J. Math. Oxford 19 (1948), 160—163.

[8] A. Tietäväinen: On the non-trivial solvability of some equations and systems of equations in finite fields. — Ann. Acad. Sci. Fenn., Ser. A I 360 (1965), 38 pp.

[9] A. Tietäväinen: On pairs of additive equations. — Ann. Univ. Turku., Ser. A I 112 (1967), 7 pp.



TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SERIES A

I. ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA

120

ON THE DISTRIBUTION OF THE RESIDUES OF A POLYNOMIAL

BY

AIMO TIETÄVÄINEN

Turku 1968
TURUN YLIOPISTO

Turku - Kirjapaino Polytypos - 1968

On the distribution of the residues of a polynomial

Let L denote the set of points $x=(x_1, \dots, x_n)$ with integral coordinates in Euklidean n-space. For any odd prime p, let $C=C(p)$ be the set of points of L in the cube $0 \leqq x_i < p$ $(i=1,2,\dots,n)$. Suppose that $f(x)$ is any polynomial of degree d in $x_1, \dots, x_n$ with integral coefficients which does not vanish identically (mod p). For any real number a, let $\{a\}$ be the fractional part of a and $\| a \|$ the distance from a to the nearest integer. Williams [3] showed that

$$\sum_{x \in C} \{f(x)/p\} = \tfrac{1}{2}p^n + O(p^{n-\frac{1}{2}} \log p),$$

as $p \to \infty$, where the constant implied in the 0-symbol depends only upon n and d. The purpose of this note is to prove the following result, in the error term of which the logarithm is omitted.

Theorem.

$$\sum_{x \in C} \| f(x)/p \| = \tfrac{1}{4}p^n + O(p^{n-\frac{1}{2}}),$$

as $p \to \infty$, where the implied constant depends only upon n and d.

Proof. Let t be an integer, $e(t) = \exp(2\pi i t/p)$ and

$$S_1(t) = \sum_{x \in C} e(-tf(x)).$$

Then (see [2])

$$|S_1(t)| \leqq k p^{n-\frac{1}{2}}, \tag{1}$$

provided that $t \not\equiv 0$ (mod p). Here the constant k depends only upon n and d.

Denote $w = (p-1)/2$ and

$$S(t) = \sum_{u=0}^{w} e(tu).$$

It is well known that

$$\sum_{t=1}^{p-1} |S(t)| < p \log p. \tag{2}$$

Furthermore (see [1])

$$\sum_{t=1}^{p-1} |S(t)|^2 = \tfrac{1}{4}(p^2-1). \tag{3}$$

Assume that $l(m)$ is the number of solutions (u, v) of the congruence

$$u+v-m \equiv 0 \pmod p, \quad 0 \leqq u \leqq w, \ 1 \leqq v \leqq w.$$

Then

$$pl(m) = \sum_{t=0}^{p-1} \sum_{u=0}^{w} \sum_{v=1}^{w} e(t(u+v-m)) = \sum_{t=0}^{p-1} S(t)(S(t)-1)e(-tm)$$

and, on the other hand, $l(m) = p\,\|m/p\|$. Therefore

$$p^2 \sum_{x \in C} \|f(x)/p\| = p \sum_{x \in C} l(f(x)) = \sum_{t=0}^{p-1} S(t)(S(t)-1)S_1(t).$$

Since $S(0) = \frac{1}{2}(p+1)$, $S_1(0) = p^n$, we find, by (1),

$$\Bigl|\sum_{x \in C} \|f(x)/p\| - \tfrac{1}{4}(p^2-1)p^{n-2}\Bigr| \leqq kp^{n-5/2} \sum_{t=1}^{p-1} (|S(t)|^2 + |S(t)|),$$

from which the theorem follows in view of the results (2) and (3).

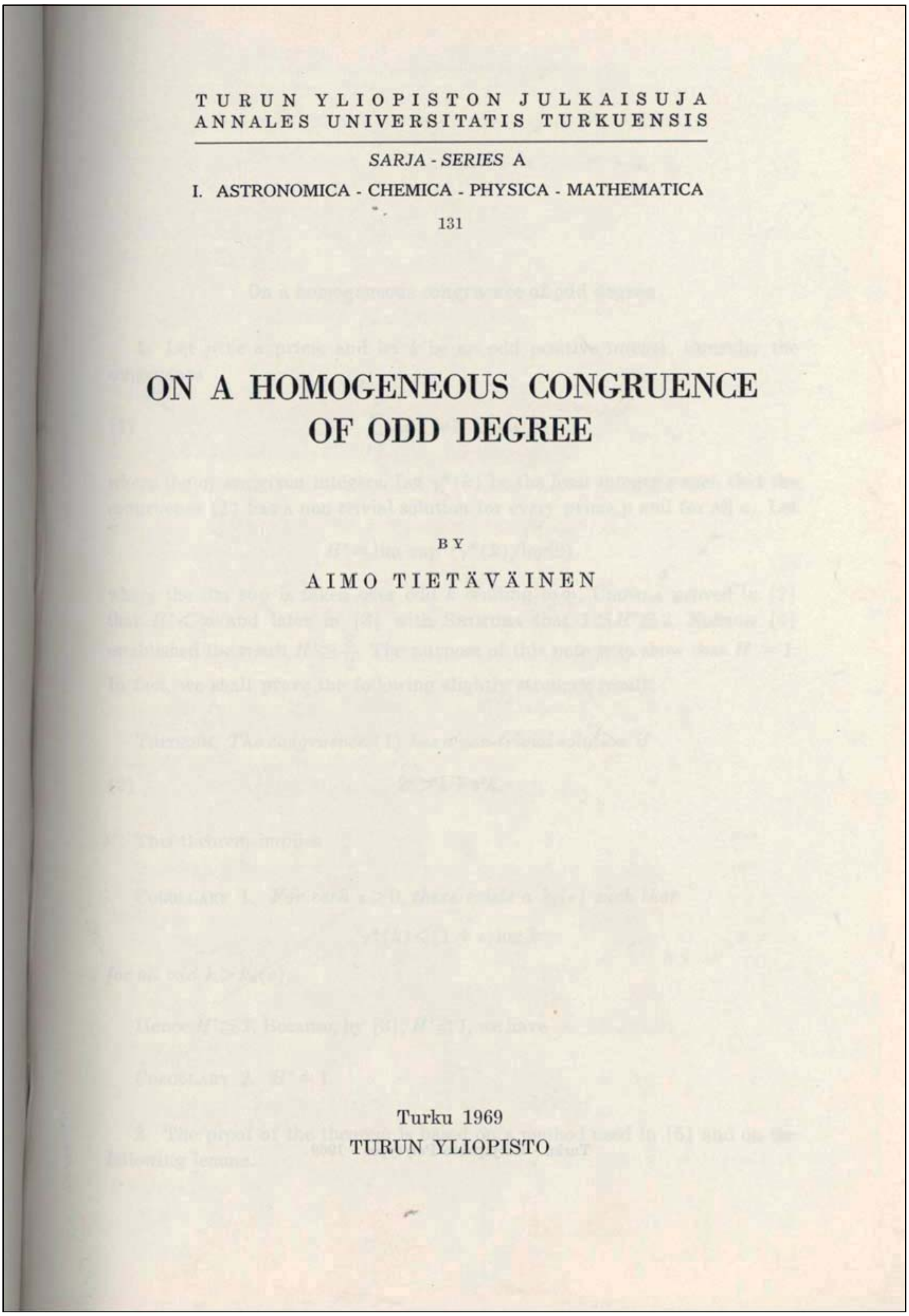
TAMPERE TECHNICAL UNIVERSITY
AND
UNIVERSITY OF TURKU,
FINLAND

References

[1] A. TIETÄVÄINEN: On the trace of a polynomial over a finite field. — Ann. Univ. Turku., Ser. A I 87 (1966), 7 pp.

[2] S. UCHIYAMA: On a multiple exponential sum. — Proc. Japan Acad. 32 (1956), 748—749.

[3] K. S. WILLIAMS: A sum of fractional parts. — Amer. Math. Monthly 74 (1967), 978—980.



TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SERIES A

I. ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA

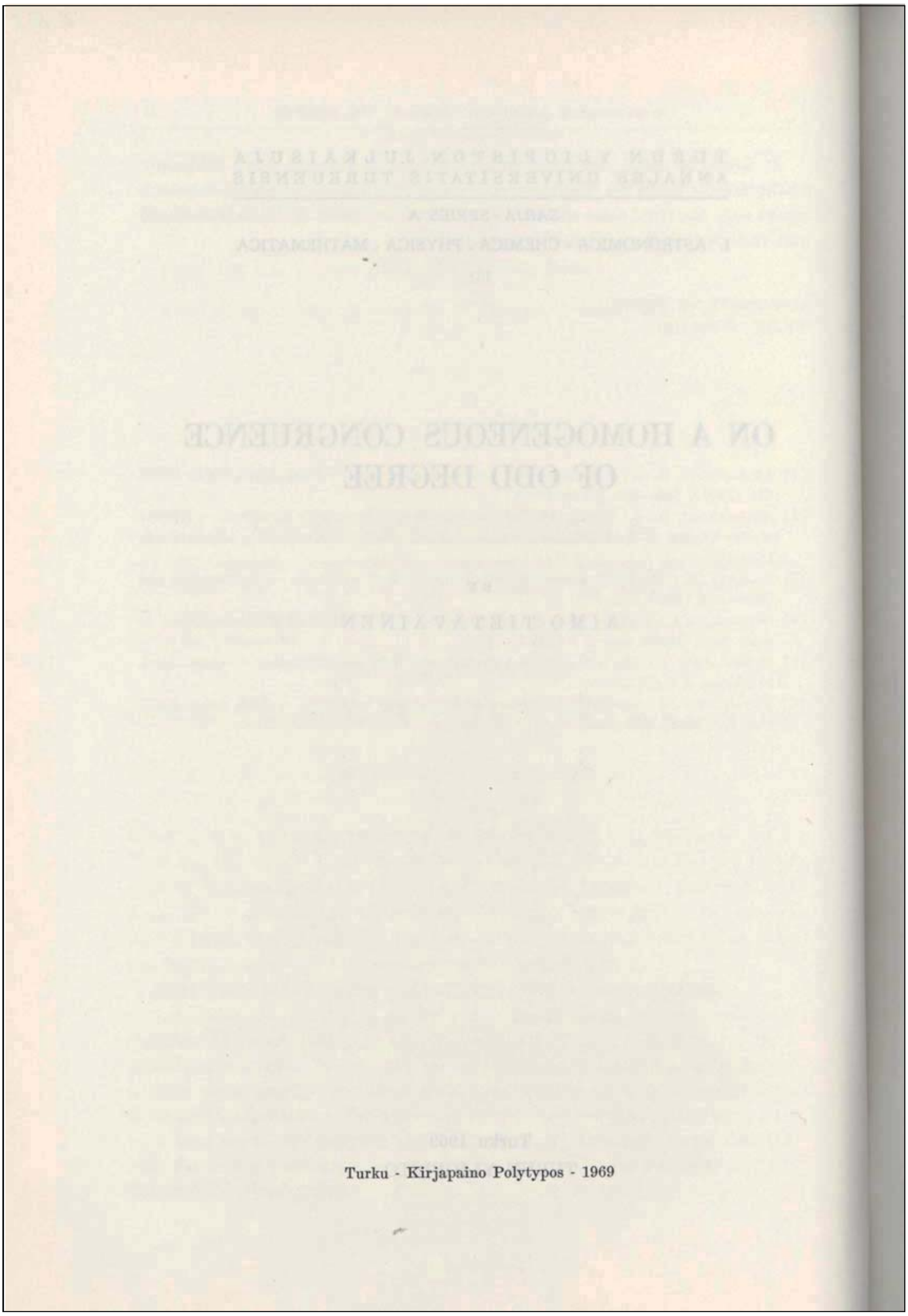
131

ON A HOMOGENEOUS CONGRUENCE OF ODD DEGREE

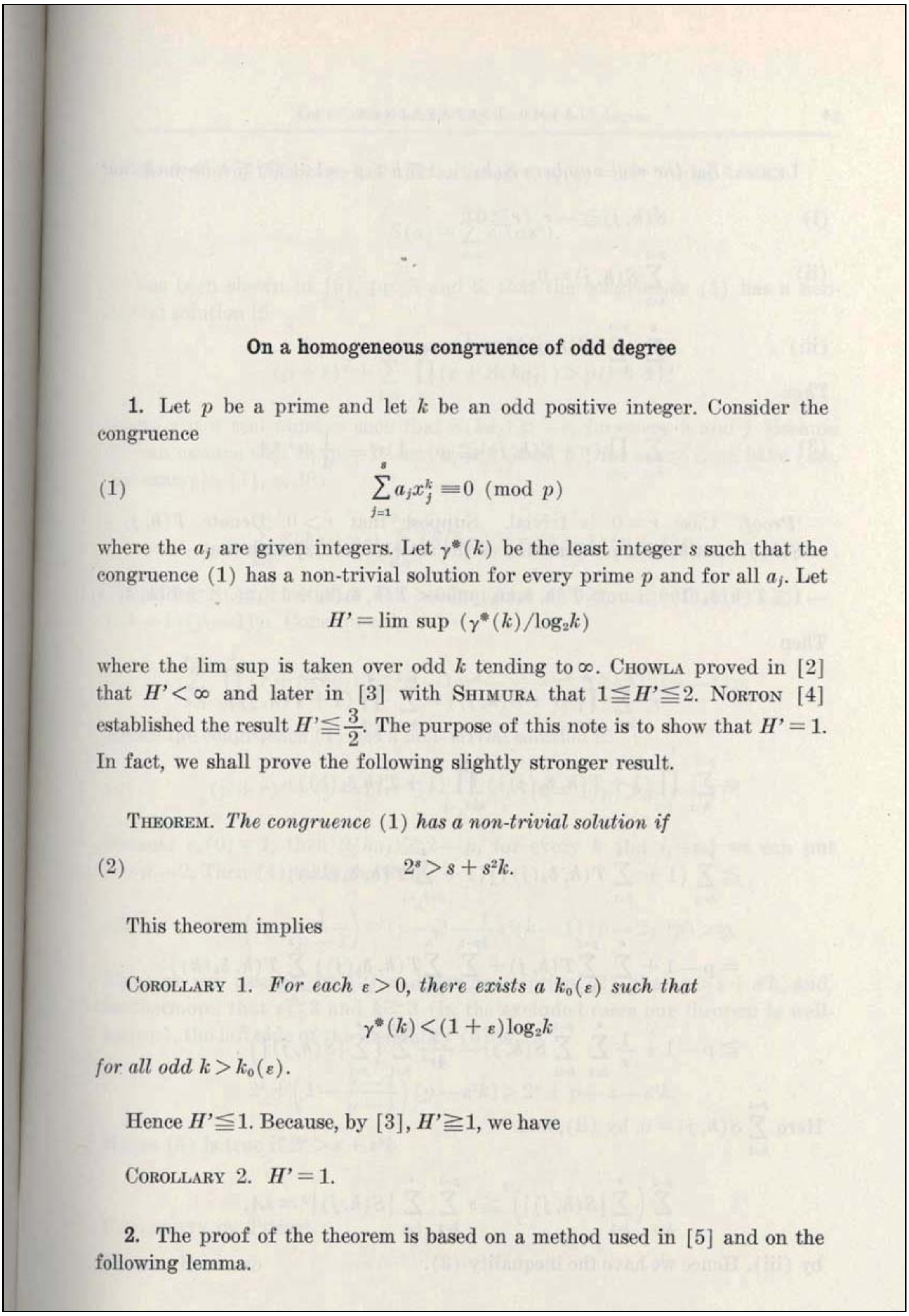
BY

AIMO TIETÄVÄINEN

Turku 1969
TURUN YLIOPISTO



Turku - Kirjapaino Polytypos - 1969



On a homogeneous congruence of odd degree

1. Let p be a prime and let k be an odd positive integer. Consider the congruence

$$\sum_{j=1}^{s} a_j x_j^k \equiv 0 \pmod{p} \tag{1}$$

where the a_j are given integers. Let $\gamma^*(k)$ be the least integer s such that the congruence (1) has a non-trivial solution for every prime p and for all a_j. Let

$$H' = \lim\sup\ (\gamma^*(k)/\log_2 k)$$

where the lim sup is taken over odd k tending to ∞. CHOWLA proved in [2] that $H' < \infty$ and later in [3] with SHIMURA that $1 \leqq H' \leqq 2$. NORTON [4] established the result $H' \leqq \frac{3}{2}$. The purpose of this note is to show that $H' = 1$. In fact, we shall prove the following slightly stronger result.

THEOREM. *The congruence* (1) *has a non-trivial solution if*

$$2^s > s + s^2 k. \tag{2}$$

This theorem implies

COROLLARY 1. *For each $\varepsilon > 0$, there exists a $k_0(\varepsilon)$ such that*

$$\gamma^*(k) < (1+\varepsilon)\log_2 k$$

for all odd $k > k_0(\varepsilon)$.

Hence $H' \leqq 1$. Because, by [3], $H' \geqq 1$, we have

COROLLARY 2. $H' = 1$.

2. The proof of the theorem is based on a method used in [5] and on the following lemma.

Lemma. *Let the real numbers* $S(h,j)$, $1 \leqq h \leqq p-1$, $1 \leqq j \leqq s$, *be such that*

$$\text{(i)} \qquad S(h,j) \geqq -r \quad (r \geqq 0),$$

$$\text{(ii)} \qquad \sum_{h=1}^{p-1} S(h,j) = 0,$$

$$\text{(iii)} \qquad \sum_{j=1}^{s} \sum_{h=1}^{p-1} (S(h,j))^2 = A.$$

Then

$$\text{(3)} \qquad \sum_{h=1}^{p-1} \prod_{j=1}^{s} (r + S(h,j)) \geqq (p-1) r^s - \frac{1}{4} s r^{s-2} A.$$

Proof. Case $r = 0$ is trivial. Suppose that $r > 0$. Denote $T(h,j) = r^{-1} S(h,j)$. Let δ_h be a permutation of the set $\{1, 2, \ldots, s\}$ such that

$$-1 \leqq T(h, \delta_h(1)) \leqq \ldots \leqq T(h, \delta_h(t_h)) \leqq 0 < T(h, \delta_h(t_h+1)) \leqq \ldots \leqq T(h, \delta_h(s)).$$

Then

$$r^{-s} \sum_{h=1}^{p-1} \prod_{j=1}^{s} (r + S(h,j)) = \sum_{h=1}^{p-1} \prod_{j=1}^{s} (1 + T(h,j))$$

$$= \sum_{h=1}^{p-1} \prod_{j=1}^{t_h} (1 + T(h, \delta_h(j))) \prod_{k=t_h+1}^{s} (1 + T(h, \delta_h(k)))$$

$$\geqq \sum_{h=1}^{p-1} (1 + \sum_{j=1}^{t_h} T(h, \delta_h(j)))(1 + \sum_{k=t_h+1}^{s} T(h, \delta_h(k)))$$

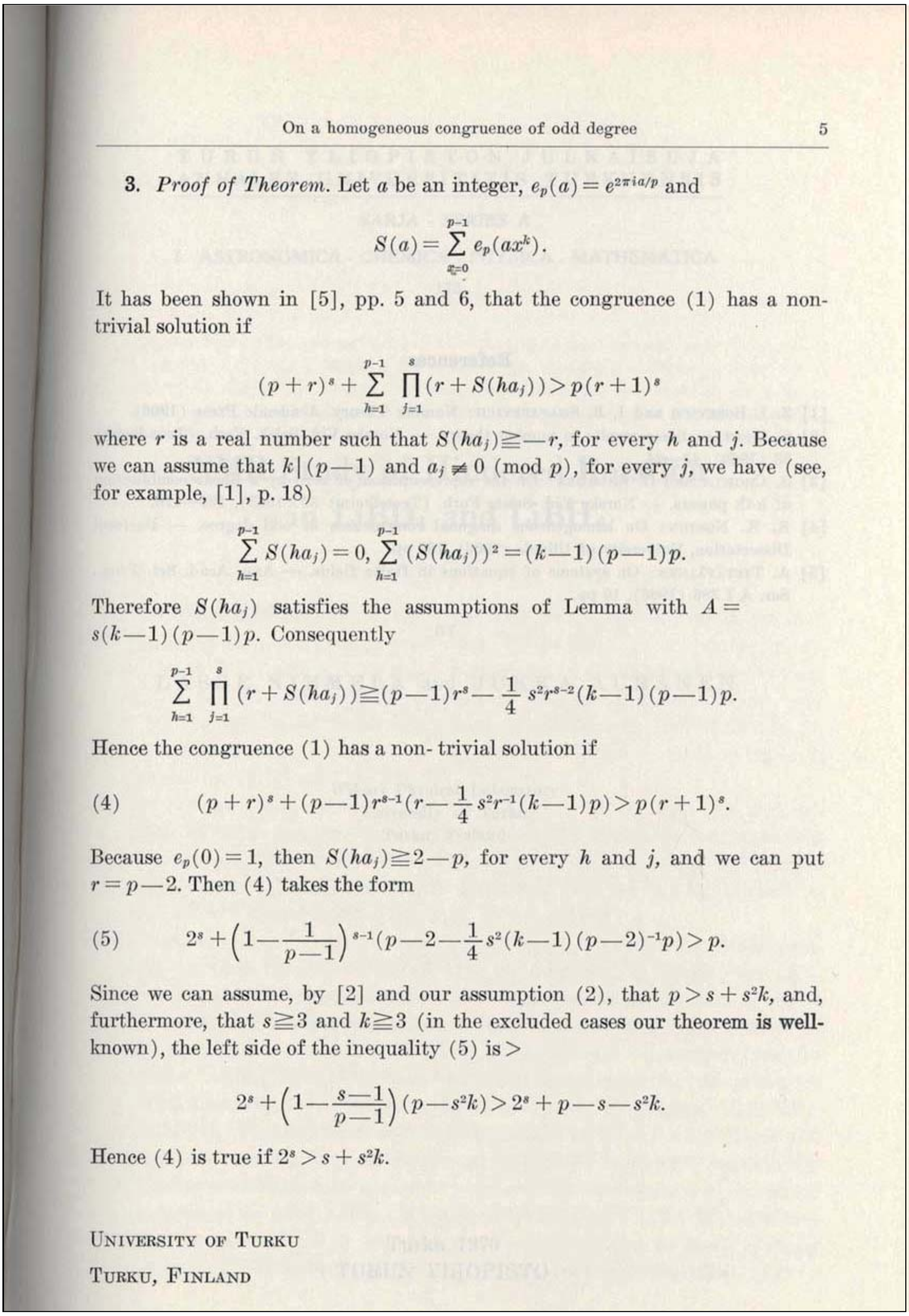
$$= p - 1 + \sum_{j=1}^{s} \sum_{h=1}^{p-1} T(h,j) + \sum_{h=1}^{p-1} \sum_{j=1}^{t_h} T(h, \delta_h(j)) \sum_{k=t_h+1}^{s} T(h, \delta_h(k))$$

$$\geqq p - 1 + \frac{1}{r} \sum_{j=1}^{s} \sum_{h=1}^{p-1} S(h,j) - \frac{1}{4r^2} \sum_{h=1}^{p-1} \Bigl(\sum_{j=1}^{s} |S(h,j)| \Bigr)^2.$$

Here $\sum_{h=1}^{p-1} S(h,j) = 0$, by (ii), and

$$\sum_{h=1}^{p-1} \Bigl(\sum_{j=1}^{s} |S(h,j)| \Bigr)^2 \leqq s \sum_{h=1}^{p-1} \sum_{j=1}^{s} |S(h,j)|^2 = sA,$$

by (iii). Hence we have the inequality (3).



3. *Proof of Theorem.* Let a be an integer, $e_p(a)=e^{2\pi ia/p}$ and

$$S(a)=\sum_{x=0}^{p-1} e_p(ax^k).$$

It has been shown in [5], pp. 5 and 6, that the congruence (1) has a non-trivial solution if

$$(p+r)^s+\sum_{h=1}^{p-1}\prod_{j=1}^{s}(r+S(ha_j))>p(r+1)^s$$

where r is a real number such that $S(ha_j)\geqq -r$, for every h and j. Because we can assume that $k\,|\,(p-1)$ and $a_j\not\equiv 0 \pmod p$, for every j, we have (see, for example, [1], p. 18)

$$\sum_{h=1}^{p-1} S(ha_j)=0,\ \sum_{h=1}^{p-1}(S(ha_j))^2=(k-1)(p-1)p.$$

Therefore $S(ha_j)$ satisfies the assumptions of Lemma with $A=s(k-1)(p-1)p$. Consequently

$$\sum_{h=1}^{p-1}\prod_{j=1}^{s}(r+S(ha_j))\geqq(p-1)r^s-\frac{1}{4}s^2r^{s-2}(k-1)(p-1)p.$$

Hence the congruence (1) has a non- trivial solution if

$$(4)\qquad (p+r)^s+(p-1)r^{s-1}(r-\frac{1}{4}s^2r^{-1}(k-1)p)>p(r+1)^s.$$

Because $e_p(0)=1$, then $S(ha_j)\geqq 2-p$, for every h and j, and we can put $r=p-2$. Then (4) takes the form

$$(5)\qquad 2^s+\left(1-\frac{1}{p-1}\right)^{s-1}(p-2-\frac{1}{4}s^2(k-1)(p-2)^{-1}p)>p.$$

Since we can assume, by [2] and our assumption (2), that $p>s+s^2k$, and, furthermore, that $s\geqq 3$ and $k\geqq 3$ (in the excluded cases our theorem is well-known), the left side of the inequality (5) is >

$$2^s+\left(1-\frac{s-1}{p-1}\right)(p-s^2k)>2^s+p-s-s^2k.$$

Hence (4) is true if $2^s>s+s^2k$.

University of Turku
Turku, Finland

References

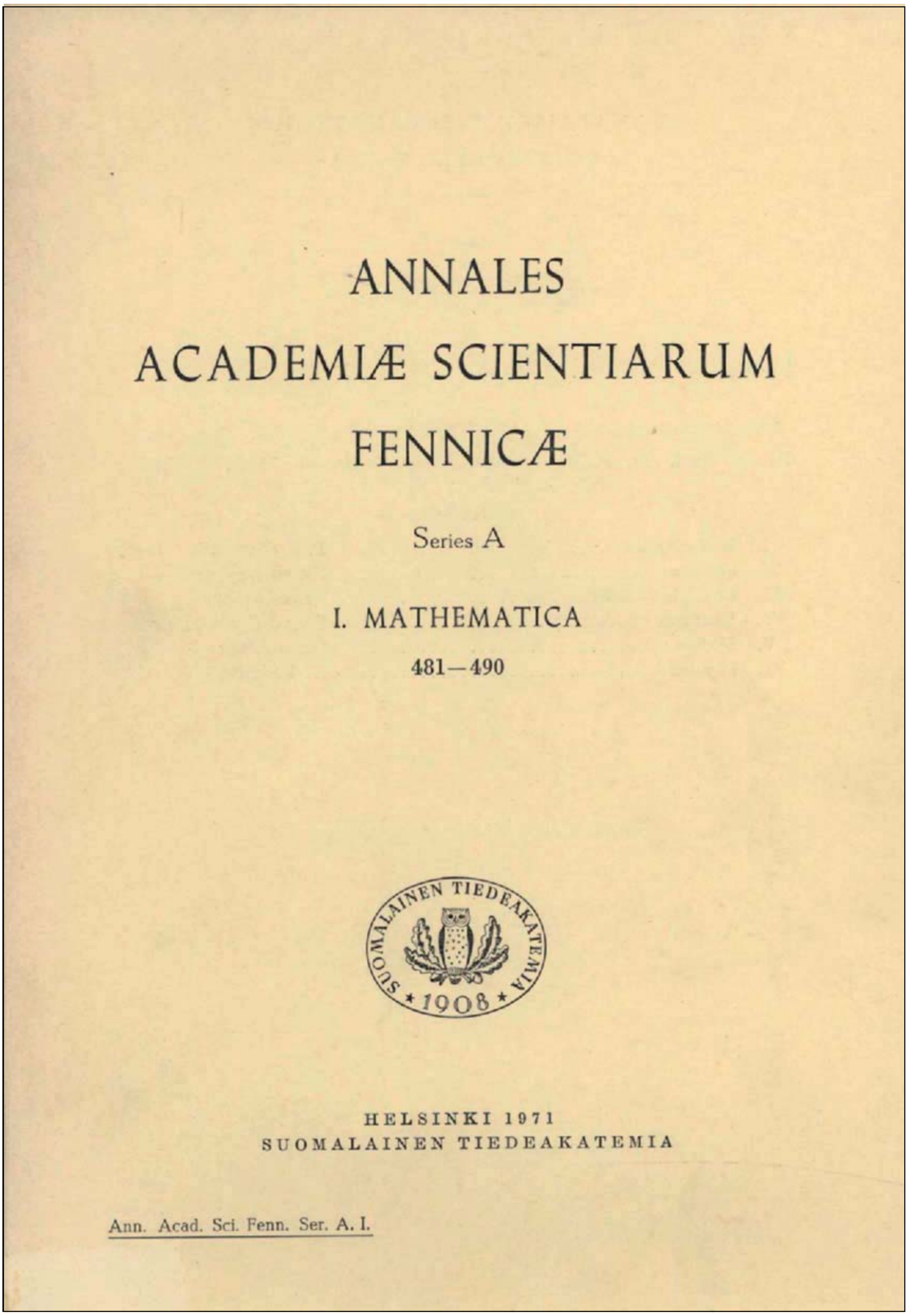
[1] Z. I. BOREVICH and I. R. SHAFAREVICH: Number Theory. Academic Press (1966).

[2] S. CHOWLA: Some results in number theory. — Norske Vid. Selsk. Forh. (Trondheim) 33 (1960), 43—44.

[3] S. CHOWLA and G. SHIMURA: On the representation of zero by a linear combination of k-th powers. — Norske Vid. Selsk. Forh. (Trondheim) 36 (1963), 169—176.

[4] K. K. NORTON: On homogeneous diagonal congruences of odd degree. — Doctoral Dissertation, University of Illinois (1966), 169 pp.

[5] A. TIETÄVÄINEN: On systems of equations in finite fields. — Ann. Acad. Sci. Fenn., Ser. A I 386 (1966), 10 pp.



ANNALES ACADEMIÆ SCIENTIARUM FENNICÆ

Series A

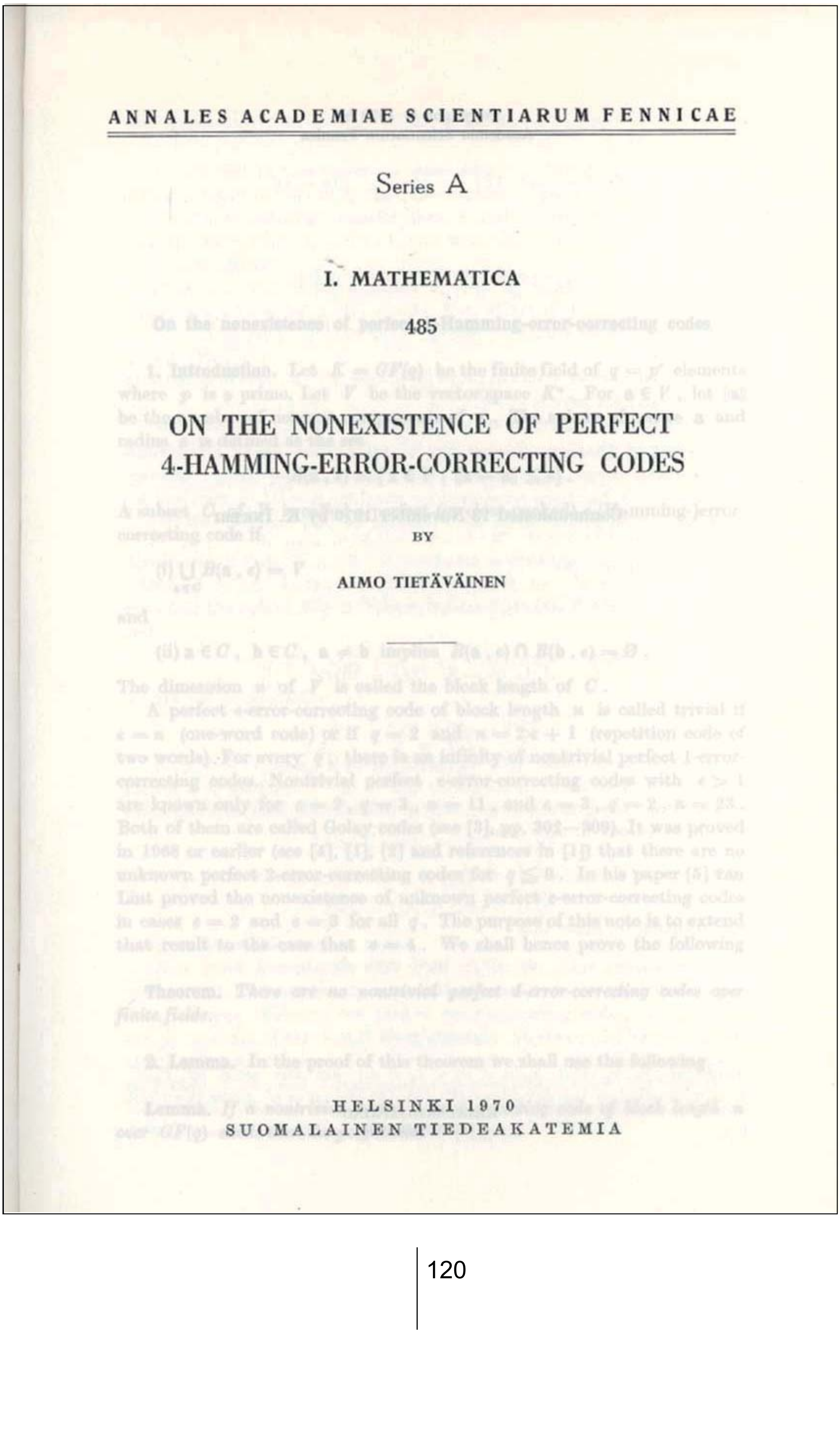
I. MATHEMATICA

481—490

SUOMALAINEN TIEDEAKATEMIA
1908

HELSINKI 1971
SUOMALAINEN TIEDEAKATEMIA

Ann. Acad. Sci. Fenn. Ser. A. I.



ANNALES ACADEMIAE SCIENTIARUM FENNICAE

Series A

I. MATHEMATICA

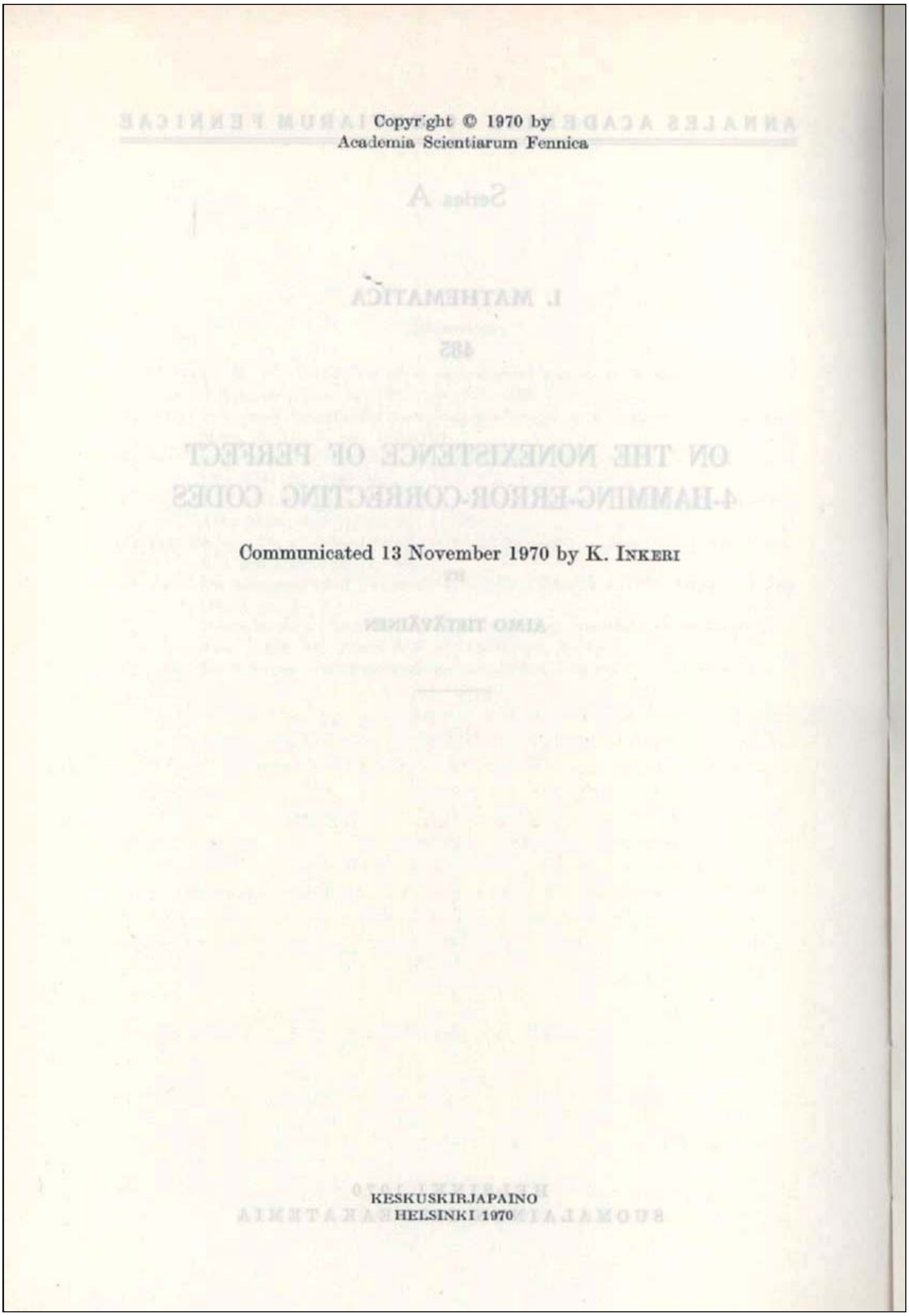
485

ON THE NONEXISTENCE OF PERFECT 4-HAMMING-ERROR-CORRECTING CODES

BY

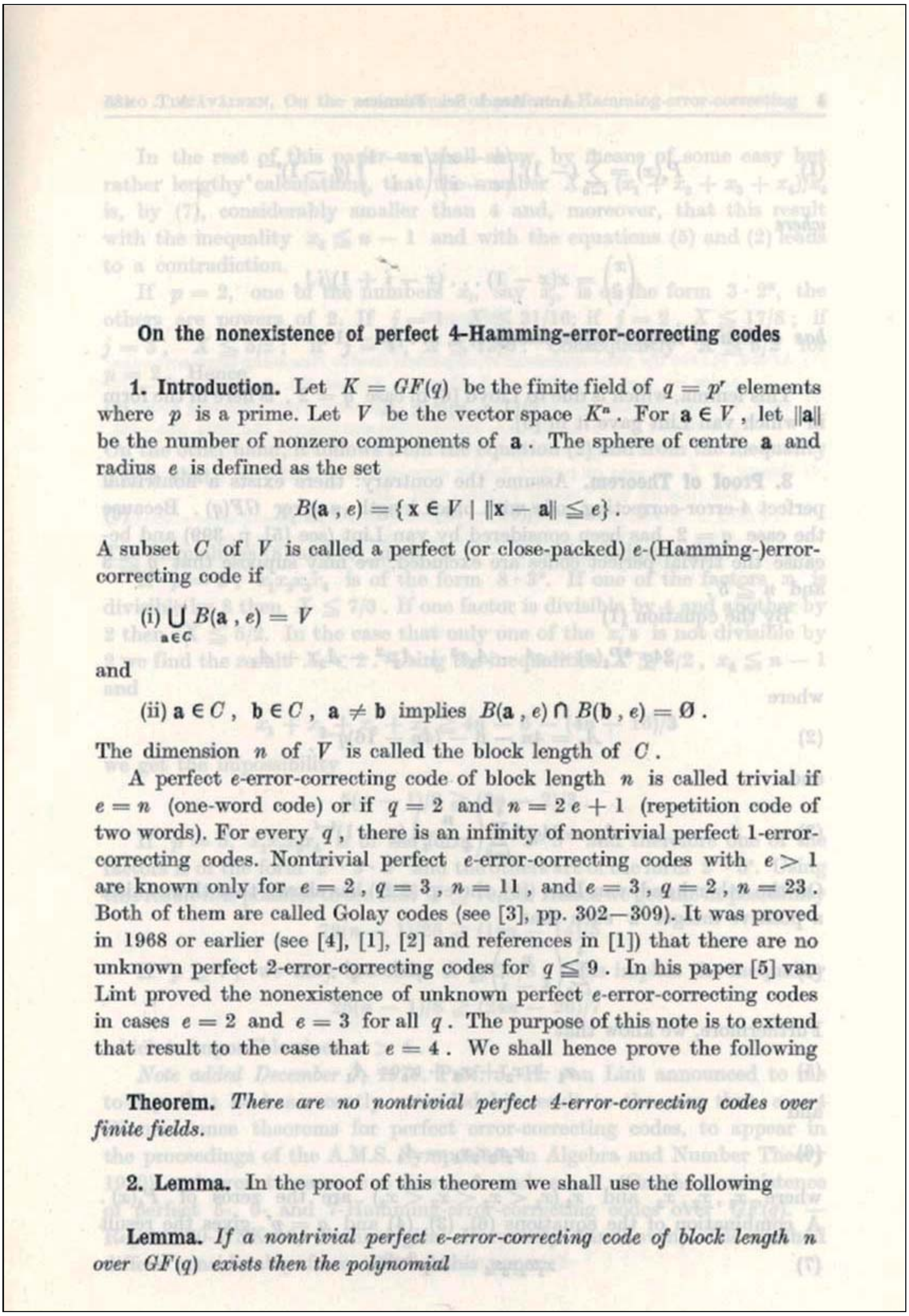
AIMO TIETÄVÄINEN

HELSINKI 1970
SUOMALAINEN TIEDEAKATEMIA



Communicated 13 November 1970 by K. Inkeri

KESKUSKIRJAPAINO
HELSINKI 1970



On the nonexistence of perfect 4-Hamming-error-correcting codes

1. Introduction. Let $K = GF(q)$ be the finite field of $q = p^r$ elements where p is a prime. Let V be the vector space K^n. For $\mathbf{a} \in V$, let $\|\mathbf{a}\|$ be the number of nonzero components of $\mathbf{a}$. The sphere of centre $\mathbf{a}$ and radius e is defined as the set

$$B(\mathbf{a}, e) = \{ \mathbf{x} \in V \mid \|\mathbf{x} - \mathbf{a}\| \leqq e \}.$$

A subset C of V is called a perfect (or close-packed) e-(Hamming-)error-correcting code if

(i) $\bigcup_{\mathbf{a} \in C} B(\mathbf{a}, e) = V$

and

(ii) $\mathbf{a} \in C$, $\mathbf{b} \in C$, $\mathbf{a} \neq \mathbf{b}$ implies $B(\mathbf{a}, e) \cap B(\mathbf{b}, e) = \emptyset$.

The dimension n of V is called the block length of C.

A perfect e-error-correcting code of block length n is called trivial if $e = n$ (one-word code) or if $q = 2$ and $n = 2e + 1$ (repetition code of two words). For every q, there is an infinity of nontrivial perfect 1-error-correcting codes. Nontrivial perfect e-error-correcting codes with $e > 1$ are known only for $e = 2$, $q = 3$, $n = 11$, and $e = 3$, $q = 2$, $n = 23$. Both of them are called Golay codes (see [3], pp. 302—309). It was proved in 1968 or earlier (see [4], [1], [2] and references in [1]) that there are no unknown perfect 2-error-correcting codes for $q \leqq 9$. In his paper [5] van Lint proved the nonexistence of unknown perfect e-error-correcting codes in cases $e = 2$ and $e = 3$ for all q. The purpose of this note is to extend that result to the case that $e = 4$. We shall hence prove the following

Theorem. *There are no nontrivial perfect 4-error-correcting codes over finite fields.*

2. Lemma. In the proof of this theorem we shall use the following

Lemma. *If a nontrivial perfect e-error-correcting code of block length n over $GF(q)$ exists then the polynomial*

$$(1) \qquad P_e(x) = \sum_{i=0}^{e} (-1)^i \binom{n-x}{e-i}\binom{x-1}{i}(q-1)^{e-i},$$

where

$$\binom{x}{i} = x(x-1)\ldots(x-i+1)/i\,!,$$

has e *distinct integral zeros among* $1, 2, \ldots, n-1$.

This lemma, which is due to Lloyd [6] in case $q = 2$, is here in the form in which van Lint gave it in [5].

3. Proof of Theorem. Assume the contrary: there exists a nontrivial perfect 4-error-correcting code with block length n over $GF(q)$. Because the case $q = 2$ has been considered by van Lint (see [5], p. 399) and because the trivial perfect codes are excluded, we may suppose that $q \geqq 3$ and $n \geqq 5$.

By the equation (1)

$$24q^{-4}P_4(x) = x^4 - A_1x^3 + A_2x^2 - A_3x + A_4$$

where

$$(2) \qquad A_1 = 4n - 6 - (4n - 16)q^{-1}$$

and

$$(3) \qquad A_4 = 24q^{-4}\sum_{i=0}^{4}\binom{n}{4-i}(q-1)^{4-i}.$$

On the other hand, van Lint ([5], the eq. (2.2)) has shown that there exists a positive integer k such that

$$(4) \qquad \sum_{i=0}^{4}\binom{n}{4-i}(q-1)^{4-i} = q^k.$$

Furthermore, we know that

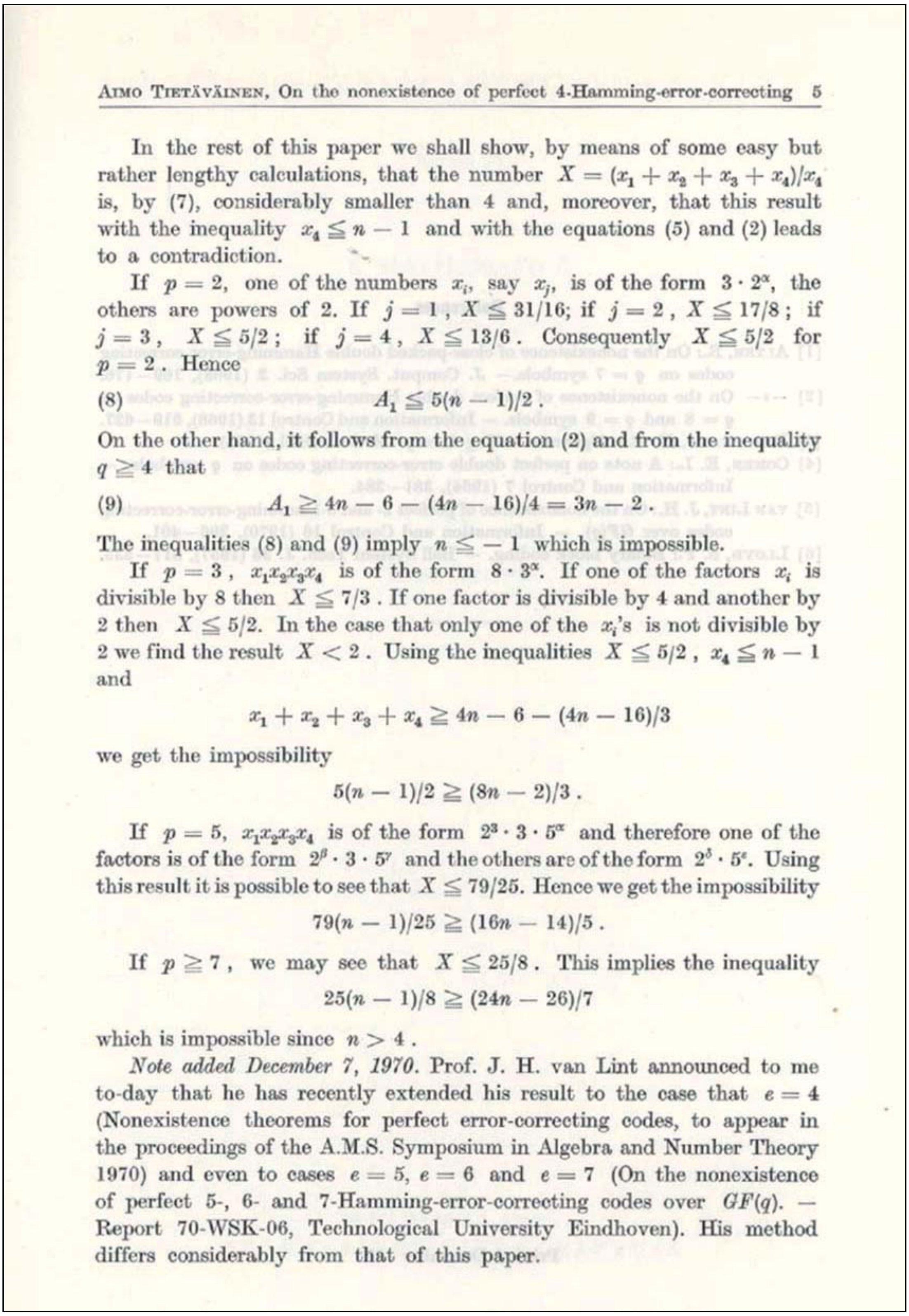
$$(5) \qquad x_1 + x_2 + x_3 + x_4 = A_1$$

and

$$(6) \qquad x_1x_2x_3x_4 = A_4$$

where x_1, x_2, x_3 and x_4 $(x_1 < x_2 < x_3 < x_4)$ are the zeros of $P_4(x)$. A combination of the equations (6), (3), (4) and $q = p^r$ gives the result

$$(7) \qquad x_1x_2x_3x_4 = 24p^{(k-4)r}.$$



In the rest of this paper we shall show, by means of some easy but rather lengthy calculations, that the number $X = (x_1 + x_2 + x_3 + x_4)/x_4$ is, by (7), considerably smaller than 4 and, moreover, that this result with the inequality $x_4 \leqq n - 1$ and with the equations (5) and (2) leads to a contradiction.

If $p = 2$, one of the numbers x_i, say x_j, is of the form $3 \cdot 2^\alpha$, the others are powers of 2. If $j = 1$, $X \leqq 31/16$; if $j = 2$, $X \leqq 17/8$; if $j = 3$, $X \leqq 5/2$; if $j = 4$, $X \leqq 13/6$. Consequently $X \leqq 5/2$ for $p = 2$. Hence

$$A_1 \leqq 5(n-1)/2 . \tag{8}$$

On the other hand, it follows from the equation (2) and from the inequality $q \geqq 4$ that

$$A_1 \geqq 4n - 6 - (4n - 16)/4 = 3n - 2 . \tag{9}$$

The inequalities (8) and (9) imply $n \leqq -1$ which is impossible.

If $p = 3$, $x_1 x_2 x_3 x_4$ is of the form $8 \cdot 3^\alpha$. If one of the factors x_i is divisible by 8 then $X \leqq 7/3$. If one factor is divisible by 4 and another by 2 then $X \leqq 5/2$. In the case that only one of the x_i's is not divisible by 2 we find the result $X < 2$. Using the inequalities $X \leqq 5/2$, $x_4 \leqq n - 1$ and

$$x_1 + x_2 + x_3 + x_4 \geqq 4n - 6 - (4n - 16)/3$$

we get the impossibility

$$5(n-1)/2 \geqq (8n - 2)/3 .$$

If $p = 5$, $x_1 x_2 x_3 x_4$ is of the form $2^3 \cdot 3 \cdot 5^\alpha$ and therefore one of the factors is of the form $2^\beta \cdot 3 \cdot 5^\gamma$ and the others are of the form $2^\delta \cdot 5^\varepsilon$. Using this result it is possible to see that $X \leqq 79/25$. Hence we get the impossibility

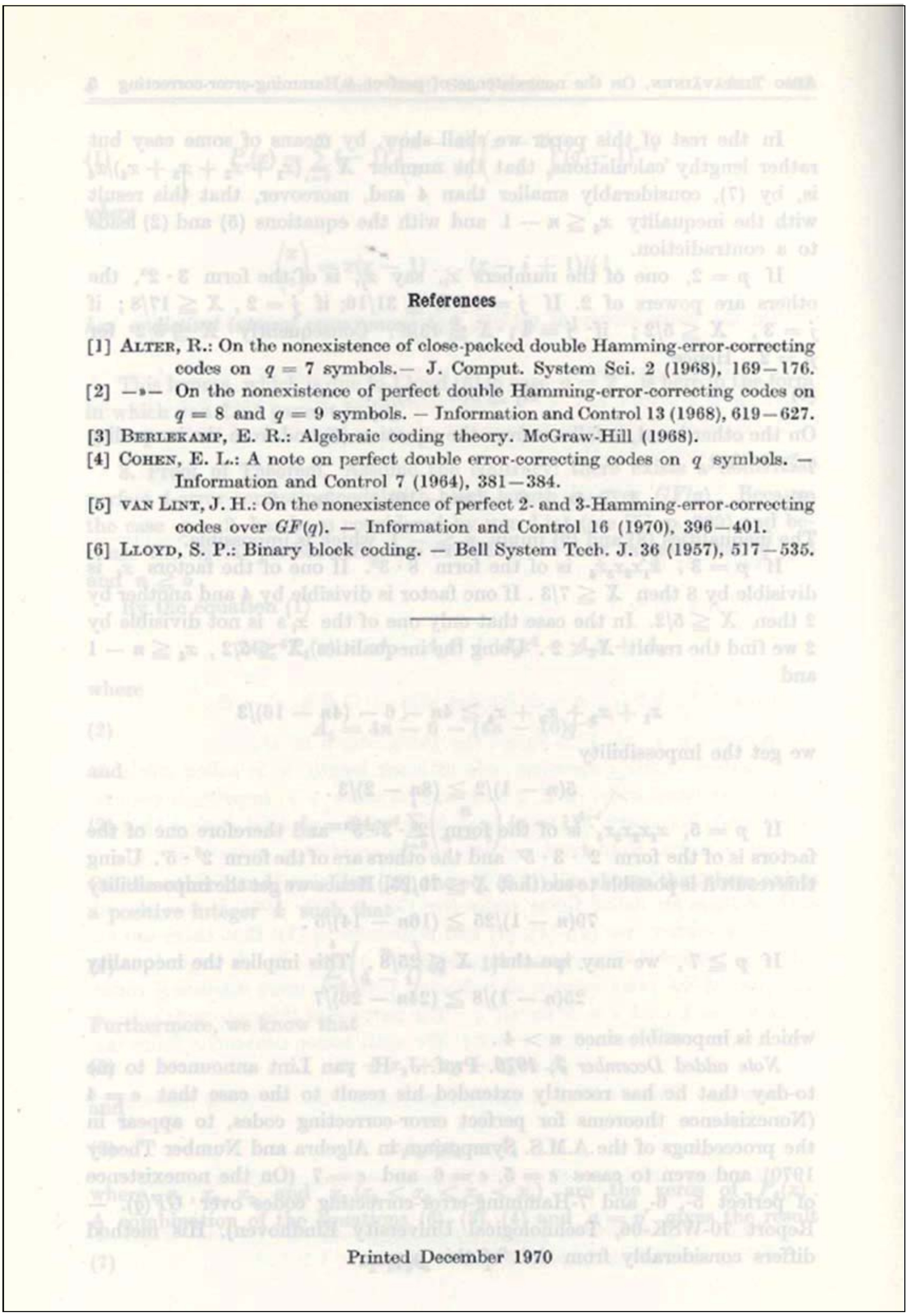
$$79(n-1)/25 \geqq (16n - 14)/5 .$$

If $p \geqq 7$, we may see that $X \leqq 25/8$. This implies the inequality

$$25(n-1)/8 \geqq (24n - 26)/7$$

which is impossible since $n > 4$.

Note added December 7, 1970. Prof. J. H. van Lint announced to me to-day that he has recently extended his result to the case that $e = 4$ (Nonexistence theorems for perfect error-correcting codes, to appear in the proceedings of the A.M.S. Symposium in Algebra and Number Theory 1970) and even to cases $e = 5$, $e = 6$ and $e = 7$ (On the nonexistence of perfect 5-, 6- and 7-Hamming-error-correcting codes over $GF(q)$. — Report 70-WSK-06, Technological University Eindhoven). His method differs considerably from that of this paper.



References

[1] Alter, R.: On the nonexistence of close-packed double Hamming-error-correcting codes on $q = 7$ symbols. — J. Comput. System Sci. 2 (1968), 169—176.

[2] —»— On the nonexistence of perfect double Hamming-error-correcting codes on $q = 8$ and $q = 9$ symbols. — Information and Control 13 (1968), 619—627.

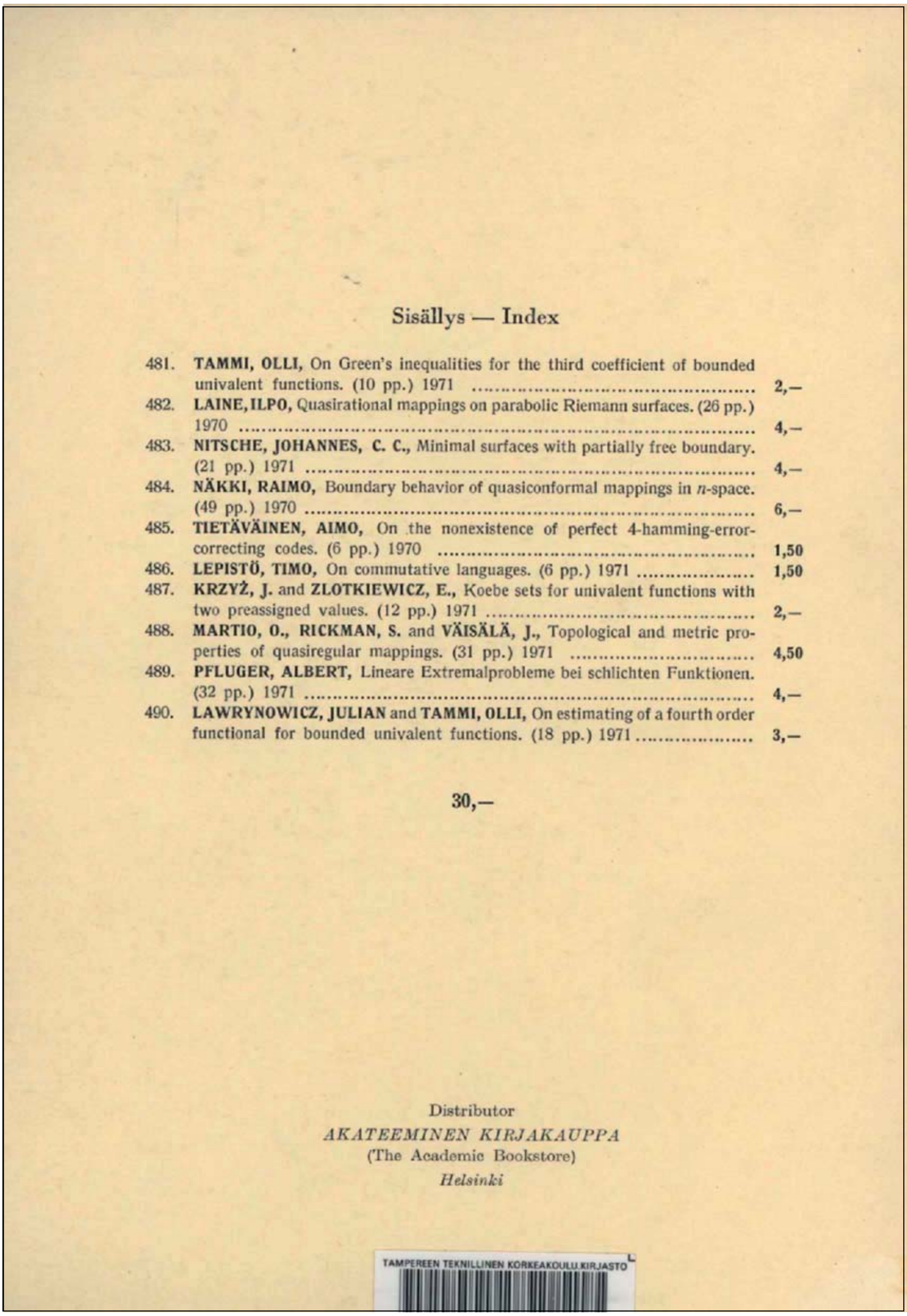
[3] Berlekamp, E. R.: Algebraic coding theory. McGraw-Hill (1968).

[4] Cohen, E. L.: A note on perfect double error-correcting codes on q symbols. — Information and Control 7 (1964), 381—384.

[5] van Lint, J. H.: On the nonexistence of perfect 2- and 3-Hamming-error-correcting codes over $GF(q)$. — Information and Control 16 (1970), 396—401.

[6] Lloyd, S. P.: Binary block coding. — Bell System Tech. J. 36 (1957), 517—535.

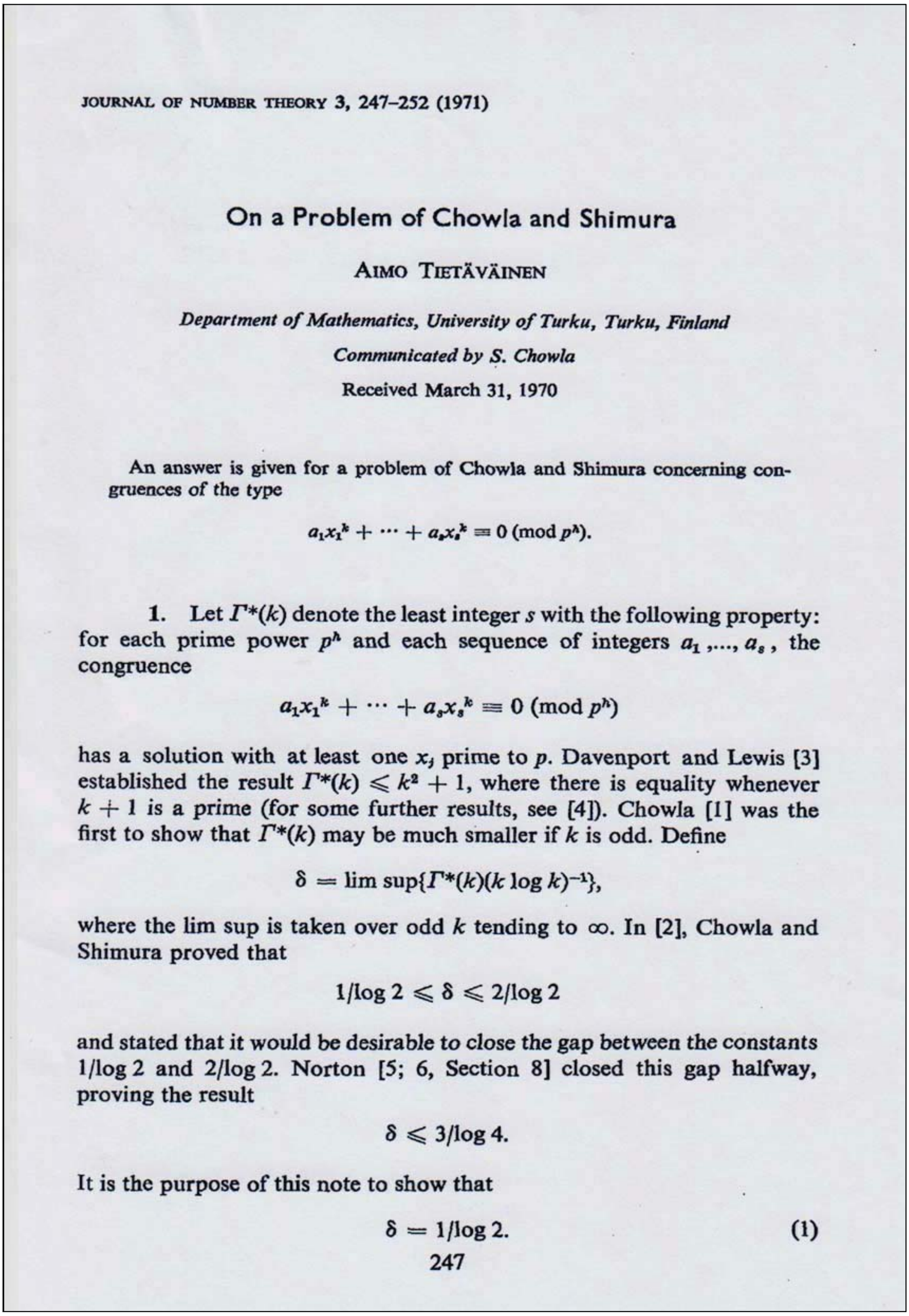
Printed December 1970



Sisällys — Index

30,—

Distributor
AKATEEMINEN KIRJAKAUPPA
(The Academic Bookstore)
Helsinki



JOURNAL OF NUMBER THEORY 3, 247–252 (1971)

On a Problem of Chowla and Shimura

AIMO TIETÄVÄINEN

Department of Mathematics, University of Turku, Turku, Finland

Communicated by S. Chowla

Received March 31, 1970

An answer is given for a problem of Chowla and Shimura concerning congruences of the type

$$a_1x_1^k + \cdots + a_sx_s^k \equiv 0 \pmod{p^h}.$$

1. Let $\Gamma^*(k)$ denote the least integer s with the following property: for each prime power p^h and each sequence of integers $a_1, \ldots, a_s$, the congruence

$$a_1x_1^k + \cdots + a_sx_s^k \equiv 0 \pmod{p^h}$$

has a solution with at least one x_j prime to p. Davenport and Lewis [3] established the result $\Gamma^*(k) \leqslant k^2 + 1$, where there is equality whenever $k + 1$ is a prime (for some further results, see [4]). Chowla [1] was the first to show that $\Gamma^*(k)$ may be much smaller if k is odd. Define

$$\delta = \lim\sup\{\Gamma^*(k)(k \log k)^{-1}\},$$

where the lim sup is taken over odd k tending to ∞. In [2], Chowla and Shimura proved that

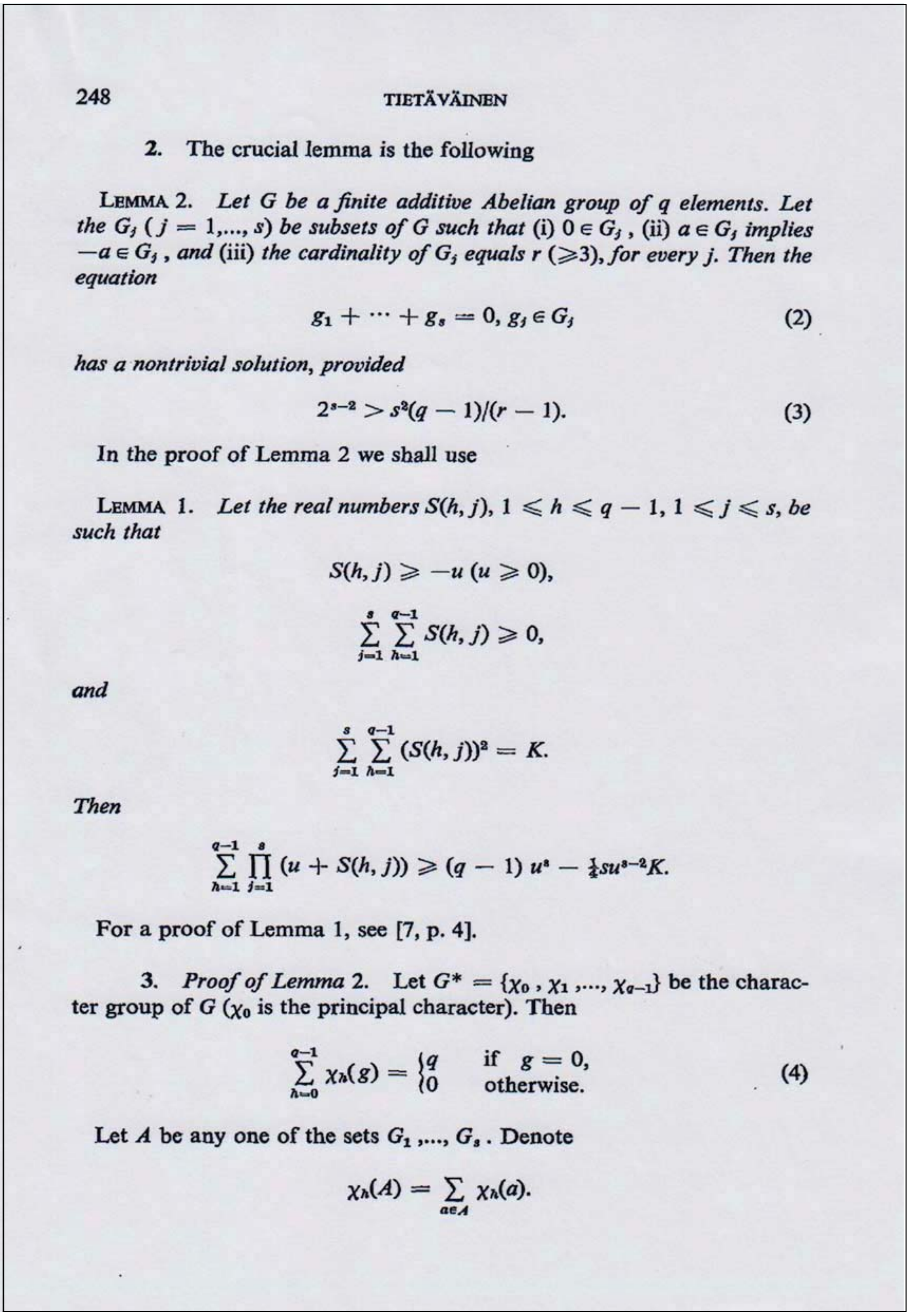
$$1/\log 2 \leqslant \delta \leqslant 2/\log 2$$

and stated that it would be desirable to close the gap between the constants 1/log 2 and 2/log 2. Norton [5; 6, Section 8] closed this gap halfway, proving the result

$$\delta \leqslant 3/\log 4.$$

It is the purpose of this note to show that

$$\delta = 1/\log 2. \tag{1}$$



2. The crucial lemma is the following

Lemma 2. *Let G be a finite additive Abelian group of q elements. Let the G_j ($j = 1,\ldots, s$) be subsets of G such that* (i) $0 \in G_j$, (ii) $a \in G_j$ *implies* $-a \in G_j$, *and* (iii) *the cardinality of G_j equals r ($\geqslant 3$), for every j. Then the equation*

$$g_1 + \cdots + g_s = 0,\ g_j \in G_j \tag{2}$$

has a nontrivial solution, provided

$$2^{s-2} > s^2(q-1)/(r-1). \tag{3}$$

In the proof of Lemma 2 we shall use

Lemma 1. *Let the real numbers $S(h, j)$, $1 \leqslant h \leqslant q-1$, $1 \leqslant j \leqslant s$, be such that*

$$S(h, j) \geqslant -u\ (u \geqslant 0),$$

$$\sum_{j=1}^{s} \sum_{h=1}^{q-1} S(h, j) \geqslant 0,$$

and

$$\sum_{j=1}^{s} \sum_{h=1}^{q-1} (S(h, j))^2 = K.$$

Then

$$\sum_{h=1}^{q-1} \prod_{j=1}^{s} (u + S(h, j)) \geqslant (q-1)\, u^s - \tfrac{1}{2} s u^{s-2} K.$$

For a proof of Lemma 1, see [7, p. 4].

3. *Proof of Lemma* 2. Let $G^* = \{\chi_0, \chi_1, \ldots, \chi_{q-1}\}$ be the character group of G (χ_0 is the principal character). Then

$$\sum_{h=0}^{q-1} \chi_h(g) = \begin{cases} q & \text{if } g = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Let A be any one of the sets $G_1, \ldots, G_s$. Denote

$$\chi_h(A) = \sum_{a \in A} \chi_h(a).$$

By assumption (ii), $\chi_h(A)$ is real, and, because $\chi_h(0) = 1$,

$$\chi_h(A) \geqslant -r + 2.$$

Furthermore, by Equation (4) and assumption (i),

$$\sum_{h=0}^{q-1} \chi_h(A) = \sum_{a \in A} \sum_{h=0}^{q-1} \chi_h(a) = q,$$

and hence, by (iii),

$$\sum_{h=1}^{q-1} \chi_h(A) = q - r \geqslant 0.$$

Moreover, it follows from (4), (ii) and (iii) that

$$\sum_{h=0}^{q-1} (\chi_h(A))^2 = \sum_{a \in A} \sum_{b \in A} \sum_{h=0}^{q-1} \chi_h(a + b) = rq.$$

Consequently, by (iii),

$$\sum_{h=1}^{q-1} (\chi_h(A))^2 = r(q - r).$$

Hence we can take in Lemma 1 $S(h, j) = \chi_h(G_j)$, $u = r - 2$ and $K = sr(q - r)$, and we get

$$\sum_{h=1}^{q-1} \prod_{j=1}^{s} (r - 2 + \chi_h(G_j)) \geqslant (q - 1)(r - 2)^s - \tfrac{1}{4}s^2(r - 2)^{s-2}\, r(q - r). \quad (5)$$

Suppose that, contrary to our assertion, Equation (2) has the trivial solution only. Let $1 \leqslant t \leqslant s$. Let $i_1, \ldots, i_t$ be different elements of the set $\{1, \ldots, s\}$. Because the equation

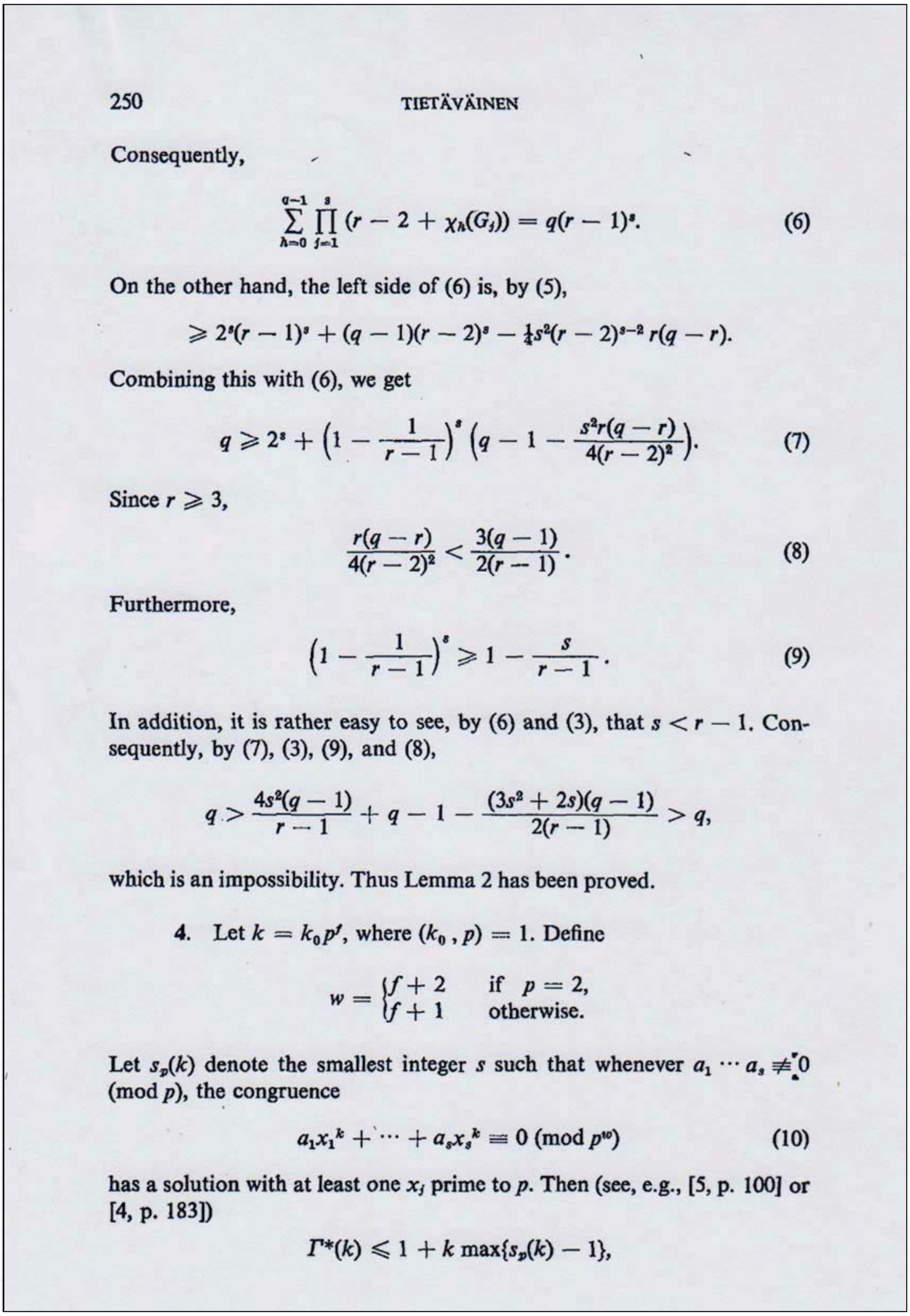
$$g_{i_1} + \cdots + g_{i_t} = 0, \qquad g_{i_k} \in G_{i_k}$$

has the trivial solution only, we have, by (4),

$$\sum_{g_{i_1} \in G_{i_1}} \cdots \sum_{g_{i_t} \in G_{i_t}} \sum_{h=0}^{q-1} \chi_h \left(\sum_{k=1}^{t} g_{i_k} \right) = q,$$

or

$$\sum_{h=0}^{q-1} \prod_{k=1}^{t} \chi_h(G_{i_k}) = q.$$



Consequently,

$$\sum_{h=0}^{q-1}\prod_{j=1}^{s}(r-2+\chi_h(G_j)) = q(r-1)^s. \tag{6}$$

On the other hand, the left side of (6) is, by (5),

$$\geqslant 2^s(r-1)^s + (q-1)(r-2)^s - \tfrac{1}{4}s^2(r-2)^{s-2}\, r(q-r).$$

Combining this with (6), we get

$$q \geqslant 2^s + \left(1-\frac{1}{r-1}\right)^s \left(q-1-\frac{s^2 r(q-r)}{4(r-2)^2}\right). \tag{7}$$

Since $r \geqslant 3$,

$$\frac{r(q-r)}{4(r-2)^2} < \frac{3(q-1)}{2(r-1)}. \tag{8}$$

Furthermore,

$$\left(1-\frac{1}{r-1}\right)^s \geqslant 1-\frac{s}{r-1}. \tag{9}$$

In addition, it is rather easy to see, by (6) and (3), that $s < r-1$. Consequently, by (7), (3), (9), and (8),

$$q > \frac{4s^2(q-1)}{r-1} + q - 1 - \frac{(3s^2+2s)(q-1)}{2(r-1)} > q,$$

which is an impossibility. Thus Lemma 2 has been proved.

4. Let $k = k_0 p^f$, where $(k_0, p) = 1$. Define

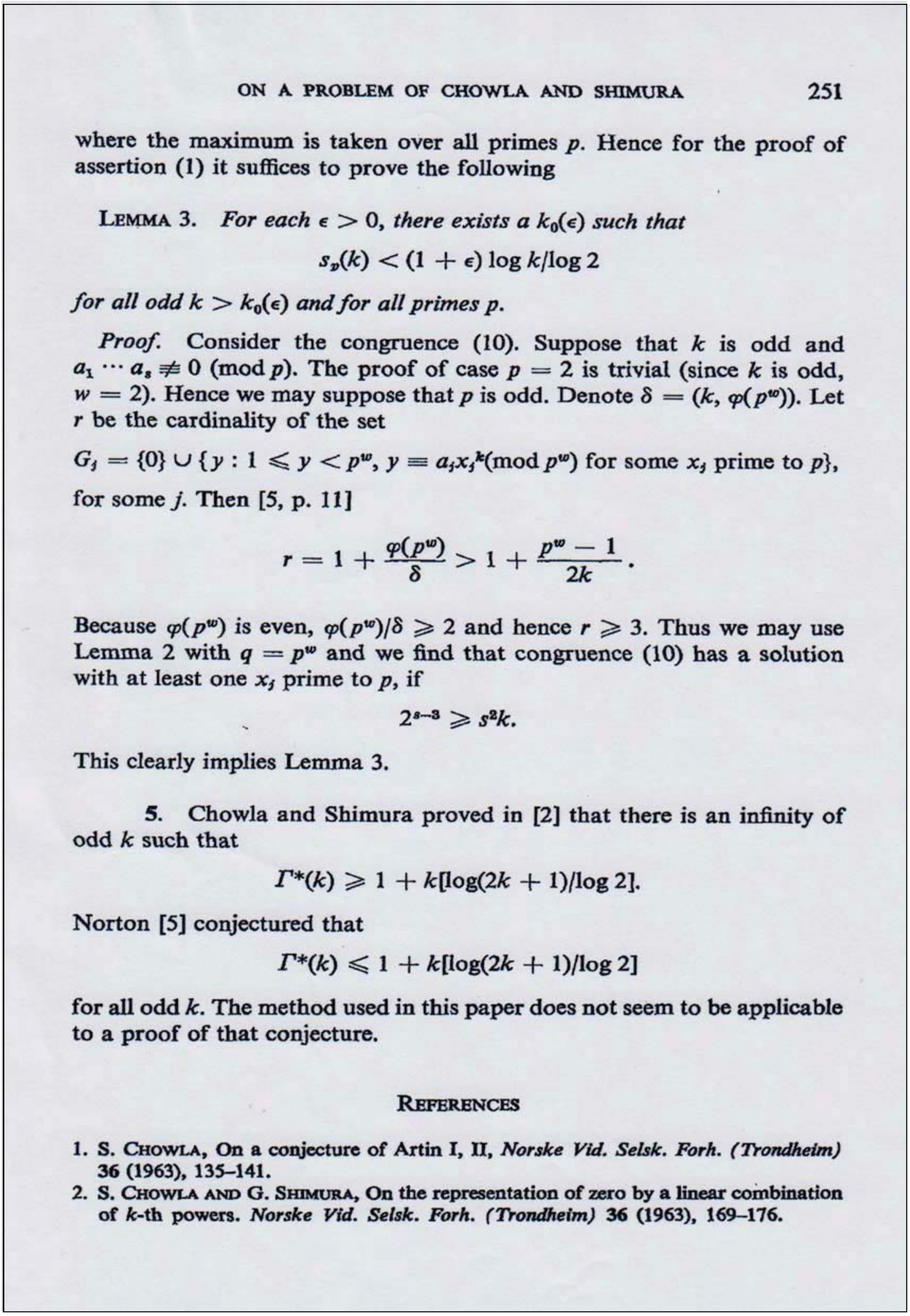
$$w = \begin{cases} f+2 & \text{if } p = 2, \\ f+1 & \text{otherwise.} \end{cases}$$

Let $s_p(k)$ denote the smallest integer s such that whenever $a_1 \cdots a_s \not\equiv 0$ (mod p), the congruence

$$a_1x_1^k + \cdots + a_sx_s^k \equiv 0 \pmod{p^w} \tag{10}$$

has a solution with at least one x_j prime to p. Then (see, e.g., [5, p. 100] or [4, p. 183])

$$\Gamma^*(k) \leqslant 1 + k\max\{s_p(k)-1\},$$



where the maximum is taken over all primes p. Hence for the proof of assertion (1) it suffices to prove the following

LEMMA 3. *For each $\epsilon > 0$, there exists a $k_0(\epsilon)$ such that*

$$s_p(k) < (1 + \epsilon) \log k/\log 2$$

for all odd $k > k_0(\epsilon)$ and for all primes p.

Proof. Consider the congruence (10). Suppose that k is odd and $a_1 \cdots a_s \not\equiv 0 \pmod p$. The proof of case $p = 2$ is trivial (since k is odd, $w = 2$). Hence we may suppose that p is odd. Denote $\delta = (k, \varphi(p^w))$. Let r be the cardinality of the set

$G_j = \{0\} \cup \{y : 1 \leqslant y < p^w,\ y \equiv a_j x_j{}^k (\mathrm{mod}\ p^w)$ for some x_j prime to $p\}$,

for some j. Then [5, p. 11]

$$r = 1 + \frac{\varphi(p^w)}{\delta} > 1 + \frac{p^w - 1}{2k}.$$

Because $\varphi(p^w)$ is even, $\varphi(p^w)/\delta \geqslant 2$ and hence $r \geqslant 3$. Thus we may use Lemma 2 with $q = p^w$ and we find that congruence (10) has a solution with at least one x_j prime to p, if

$$2^{s-3} \geqslant s^2 k.$$

This clearly implies Lemma 3.

5. Chowla and Shimura proved in [2] that there is an infinity of odd k such that

$$\Gamma^*(k) \geqslant 1 + k[\log(2k + 1)/\log 2].$$

Norton [5] conjectured that

$$\Gamma^*(k) \leqslant 1 + k[\log(2k + 1)/\log 2]$$

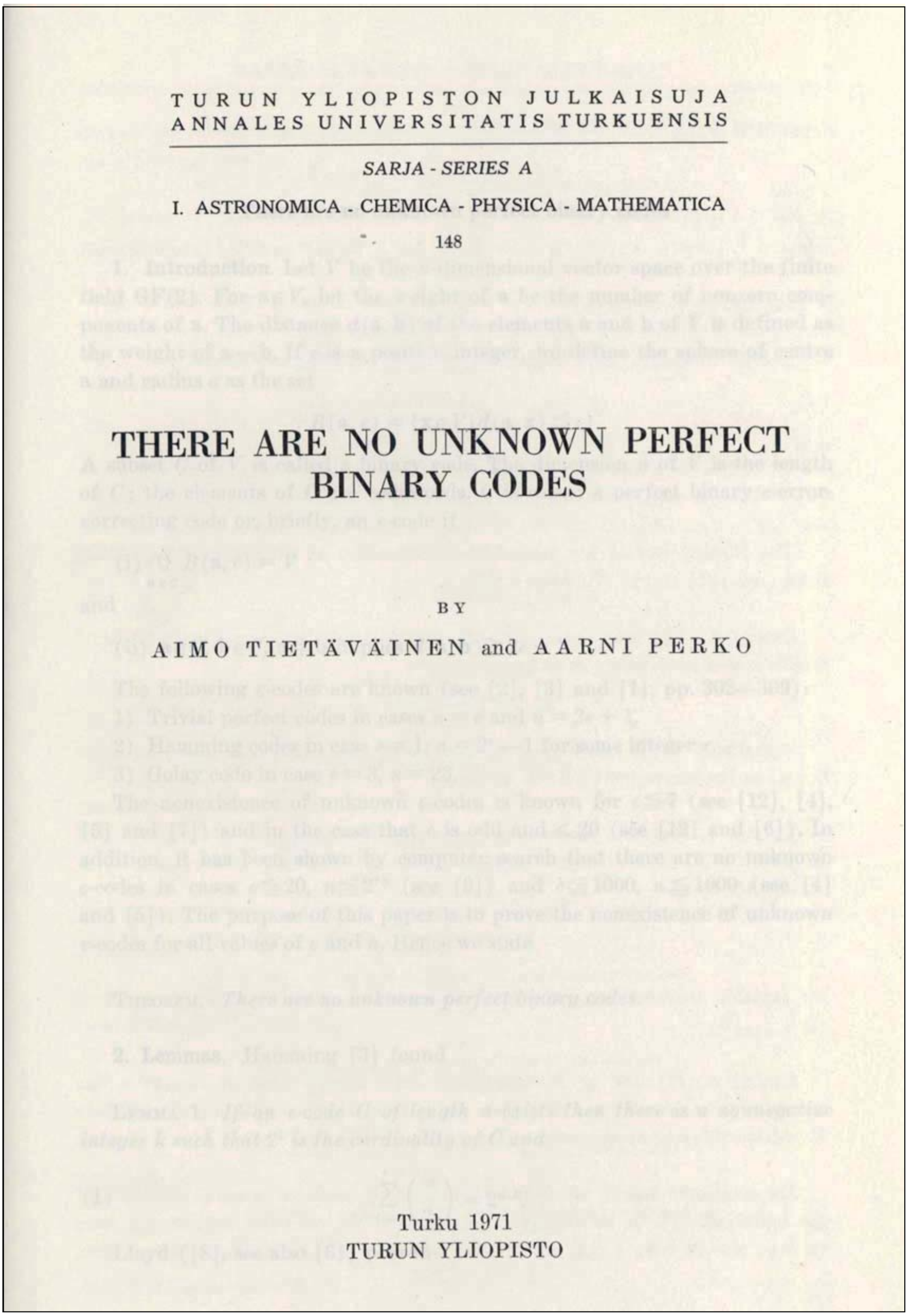
for all odd k. The method used in this paper does not seem to be applicable to a proof of that conjecture.

REFERENCES

1. S. CHOWLA, On a conjecture of Artin I, II, *Norske Vid. Selsk. Forh. (Trondheim)* **36** (1963), 135–141.
2. S. CHOWLA AND G. SHIMURA, On the representation of zero by a linear combination of k-th powers. *Norske Vid. Selsk. Forh. (Trondheim)* **36** (1963), 169–176.

3. H. Davenport and D. J. Lewis, Homogeneous additive equations, *Proc. Roy. Soc. Ser. A* **274** (1963), 443–460.
4. M. Dodson, Homogeneous additive congruences, *Philos. Trans. Roy. Soc. London Ser. A* **261** (1967), 163–210.
5. K. K. Norton, "On homogeneous diagonal congruences of odd degree," Ph.D. thesis, University of Illinois, Urbana, Ill. 1966.
6. K. K. Norton, Upper bounds for k-th power coset representatives modulo n. *Acta Arith.* **15** (1969), 161–179.
7. A. Tietäväinen, On a homogeneous congruence of odd degree, *Ann. Univ. Turku. Ser. A. I* **131** (1969), 3–6.

Printed in Belgium



TURUN YLIOPISTON JULKAISUJA
ANNALES UNIVERSITATIS TURKUENSIS

SARJA - SERIES A

I. ASTRONOMICA - CHEMICA - PHYSICA - MATHEMATICA

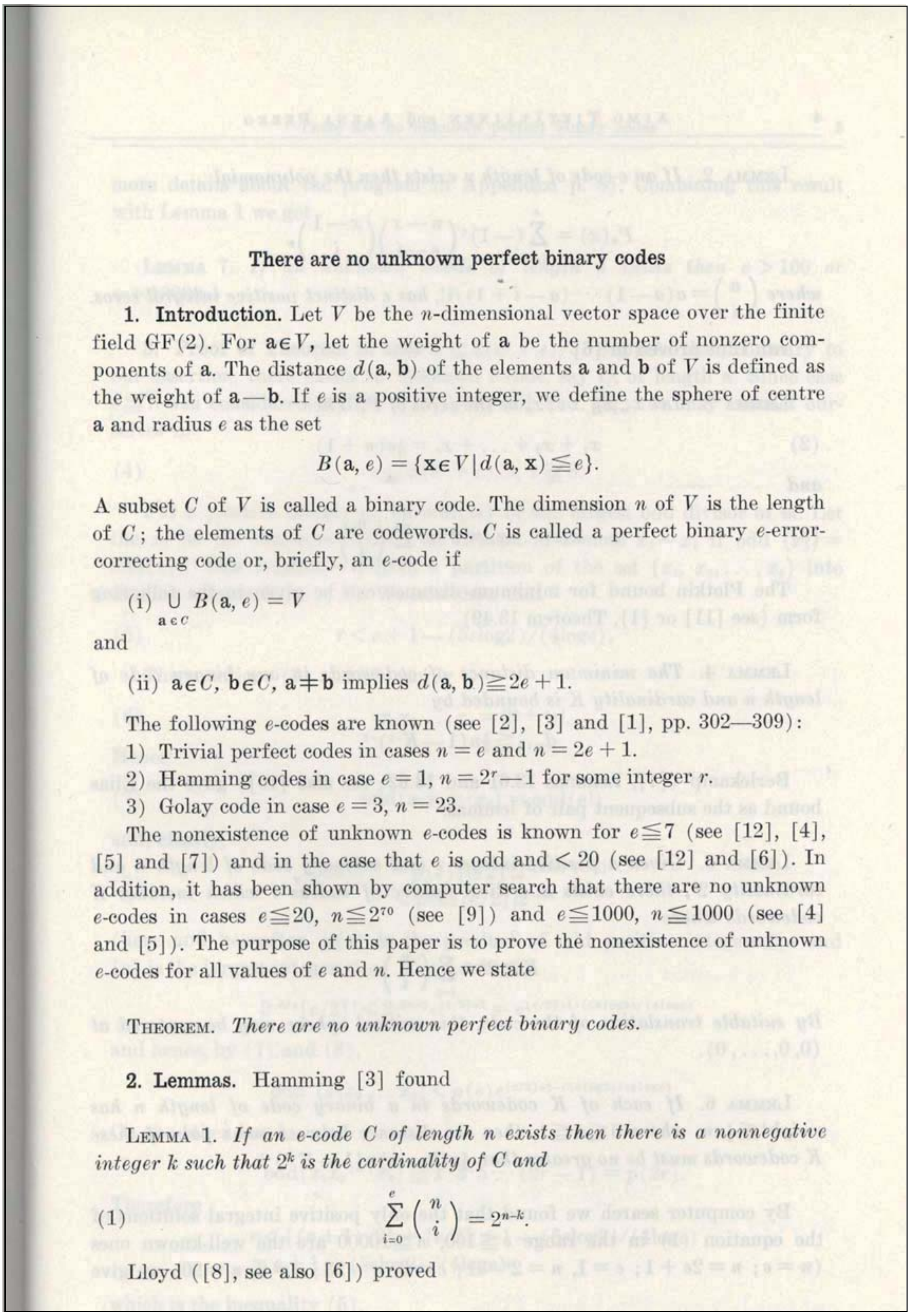
148

THERE ARE NO UNKNOWN PERFECT BINARY CODES

BY

AIMO TIETÄVÄINEN and AARNI PERKO

Turku 1971
TURUN YLIOPISTO



There are no unknown perfect binary codes

1. Introduction. Let V be the n-dimensional vector space over the finite field GF(2). For $\mathbf{a} \in V$, let the weight of $\mathbf{a}$ be the number of nonzero components of $\mathbf{a}$. The distance $d(\mathbf{a}, \mathbf{b})$ of the elements $\mathbf{a}$ and $\mathbf{b}$ of V is defined as the weight of $\mathbf{a}-\mathbf{b}$. If e is a positive integer, we define the sphere of centre $\mathbf{a}$ and radius e as the set

$$B(\mathbf{a}, e) = \{\mathbf{x} \in V \mid d(\mathbf{a}, \mathbf{x}) \leqq e\}.$$

A subset C of V is called a binary code. The dimension n of V is the length of C; the elements of C are codewords. C is called a perfect binary e-error-correcting code or, briefly, an e-code if

(i) $\bigcup_{\mathbf{a} \in C} B(\mathbf{a}, e) = V$

and

(ii) $\mathbf{a} \in C$, $\mathbf{b} \in C$, $\mathbf{a} \neq \mathbf{b}$ implies $d(\mathbf{a}, \mathbf{b}) \geqq 2e+1$.

The following e-codes are known (see [2], [3] and [1], pp. 302—309):

1) Trivial perfect codes in cases $n = e$ and $n = 2e+1$.

2) Hamming codes in case $e = 1$, $n = 2^r - 1$ for some integer r.

3) Golay code in case $e = 3$, $n = 23$.

The nonexistence of unknown e-codes is known for $e \leqq 7$ (see [12], [4], [5] and [7]) and in the case that e is odd and < 20 (see [12] and [6]). In addition, it has been shown by computer search that there are no unknown e-codes in cases $e \leqq 20$, $n \leqq 2^{70}$ (see [9]) and $e \leqq 1000$, $n \leqq 1000$ (see [4] and [5]). The purpose of this paper is to prove the nonexistence of unknown e-codes for all values of e and n. Hence we state

THEOREM. *There are no unknown perfect binary codes.*

2. Lemmas. Hamming [3] found

LEMMA 1. *If an e-code C of length n exists then there is a nonnegative integer k such that 2^k is the cardinality of C and*

$$(1) \qquad \sum_{i=0}^{e} \binom{n}{i} = 2^{n-k}.$$

Lloyd ([8], see also [6]) proved

Lemma 2. *If an e-code of length n exists then the polynomial*

$$P_e(x) = \sum_{i=0}^{e} (-1)^i \binom{n-x}{e-i}\binom{x-1}{i},$$

where $\binom{a}{i} = a(a-1)\cdots(a-i+1)/i!$, *has e distinct positive integral zeros.*

van Lint proved in [6]

Lemma 3. *Let* $x_1, x_2, \ldots, x_e$ *be the zeros of* P_e. *Then*

$$x_1 + x_2 + \ldots + x_e = \tfrac{1}{2}e(n+1) \tag{2}$$

and

$$x_1x_2\ldots x_e = 2^{-e}e!\sum_{i=0}^{e}\binom{n}{i}. \tag{3}$$

The Plotkin bound for minimum distance can be given in the following form (see [11] or [1], Theorem 13.49).

Lemma 4. *The minimum distance of codewords in any binary code of length n and cardinality K is bounded by*

$$d_{\min} \leqq \tfrac{1}{2}n(1-K^{-1})^{-1}.$$

Berlekamp ([1], Lemmas 13.61 and 13.62; see also [10]) gave the Elias bound as the subsequent pair of lemmas.

Lemma 5. *Given a positive integer t and a binary code of length n and cardinality* 2^k, *there exists a critical sphere of radius t which includes K codewords where*

$$K \geqq 2^{k-n}\sum_{i=0}^{t}\binom{n}{i}.$$

By suitable translation of the code, this critical sphere may be centered at $(0, 0, \ldots, 0)$.

Lemma 6. *If each of K codewords in a binary code of length n has weight* $\leqq \frac{1}{2}xn$, *where* $0 \leqq x \leqq 1$, *then the distance between some pair of these K codewords must be no greater than* $\frac{1}{2}x(2-x)n(1-K^{-1})^{-1}$.

By computer search we found that the only positive integral solutions of the equation (1) in the range $e \leqq 100$, $n \leqq 10000$ are the well-known ones ($n = e$; $n = 2e+1$; $e = 1$, $n = 2^r-1$; $e = 3$, $n = 23$; $e = 2$, $n = 90$; we give more details about the program in Appendix, p. 8). Combining this result with Lemma 1 we get

LEMMA 7. *If an unknown e-code of length n exists then* $e>100$ *or* $n>10000$.

3. Proof of Theorem in case $n \geqq 2(e^2+e)/3$. Suppose that, contrary to our assertion, there exists an unknown e-code, say C, of length n. Since case $e \leqq 7$ was considered earlier ([12], [4], [5] and [7]), we may restrict ourselves to

$$n \geqq 2(e^2+e)/3, \quad e \geqq 8. \tag{4}$$

For a positive integer m, let $\operatorname{odd}(m)$ be the largest odd divisor of m. Let the x_j be the numbers defined in Lemma 3. Denote $x_j \sim x_l$ if $\operatorname{odd}(x_j)=\operatorname{odd}(x_l)$. This relation $\sim$ defines a partition of the set $\{x_1, x_2, \dots, x_e\}$ into disjoint subsets $X_1, \dots, X_r$. We now show that

$$r < e+1-(5e\log 2)/(4\log e). \tag{5}$$

The equations (3) and (1) imply

$$x_1x_2 \dots x_e = 2^{n-k-e}e!. \tag{6}$$

Hence

$$\operatorname{odd}(x_1x_2\cdots x_e) = \operatorname{odd}(e!) \tag{7}$$

and, clearly,

$$\begin{aligned} \operatorname{odd}(e!) &= p(e)[e/2]!2^{-[e/4]-[e/8]-\dots} \\ &< p(e)[e/2]!2^{-e/4} \end{aligned} \tag{8}$$

(here and hereafter, $p(a)$ is the product of odd positive integers $\leqq a$ and $[a]$ is the largest integer $\leqq a$). Furthermore,

$$2^{-e/4}[e/2]! < 2^{-5e/4}e^{[e/2]+1} = e^{[e/2]+1-(5e\log 2)/(4\log e)}$$

and hence, by (7) and (8),

$$\operatorname{odd}\ (x_1x_2\dots x_e) < p(e)e^{[e/2]+1-(5e\log 2)/(4\log e)}.$$

On the other hand,

$$\operatorname{odd}(x_1x_2\cdots x_e) \geqq 1\cdot 3\cdot 5\cdots(2r-1) = p(2r).$$

Therefore

$$\begin{aligned} r &< [(e+1)/2] + [e/2] + 1 - (5e\log 2)/(4\log e) \\ &= e+1-(5e\log 2)/(4\log e) \end{aligned}$$

which is the inequality (5).

Let X_i be any one of the sets $X_1, \dots, X_r$. Let $s(i)$ be the cardinality of X_i. Denote

$$R_i = (\prod_{x \in X_i} x) / (\sum_{x \in X_i} x/s(i))^{s(i)}.$$

Because $y \in X_i$, $z \in X_i$, $y > z$ implies $y/z \geqq 2$, we see that in case $s(i) = 2$

$$R_i \leqq (z \cdot 2z)/((z+2z)/2)^2 = 8/9,$$

and it is rather easy to prove, by induction, that generally

$$R_i \leqq (8/9)^{s(i)-1}.$$

Therefore

$$R_1 \cdots R_r \leqq \prod_{i=1}^{r} (8/9)^{s(i)-1} = (8/9)^{e-r}.$$

Consequently

$$x_1 x_2 \cdots x_e \leqq (8/9)^{e-r} \prod_{i=1}^{r} (\sum_{x \in X_i} x/s(i))^{s(i)}, \tag{9}$$

and the arithmetic-mean — geometric-mean inequality implies that the right side of (9) is $\leqq$

$$(8/9)^{e-r}((x_1 + x_2 + \cdots + x_e)/e)^e.$$

Hence, by Lemma 3,

$$e! \sum_{i=0}^{e} \binom{n}{i} \leqq (8/9)^{e-r}(n+1)^e,$$

and, consequently,

$$(8/9)^{e-r} > (n+1)^{-e} e! \binom{n}{e} = \prod_{j=1}^{e} \left(1 - \frac{j}{n+1}\right) > 1 - \frac{e^2+e}{2(n+1)}. \tag{10}$$

The inequalities (10), (5) and (4) imply

$$(8/9)^{(5e\log 2)/(4\log e)-1} > \frac{1}{4},$$

and therefore

$$((5e\log 2)/(4\log e) - 1)\log(9/8) < \log 4$$

or

$$\frac{e\log 2}{\log e} < \frac{4}{5}\left(\frac{\log 4}{\log(9/8)} + 1\right) < \frac{41}{4}.$$

Hence we have

$$e < 64; \tag{11}$$

and then it follows from Lemma 7 that

$$(12) \qquad n > 10000.$$

The inequalities (11) and (12) imply

$$(13) \qquad 1 - (e^2 + e)/(2n+2) > 3/4 > (8/9)^3.$$

Since $e \geqq 8$, we have, by (5), the inequality

$$(14) \qquad e - r \geqq 3.$$

Substituting the estimates (14) and (13) in (10) we get the impossible inequality $(8/9)^3 > (8/9)^3$.

4. Proof of Theorem in case $n < 2(e^2 + e)/3$. Assume the contrary: there is an unknown e-code C of cardinality 2^k and length n where

$$(15) \qquad n < 2(e^2 + e)/3.$$

Denote, as in Lemma 4, $d_{\min} = \min\{d(\mathbf{a}, \mathbf{b}) \mid \mathbf{a} \in C,\ \mathbf{b} \in C,\ \mathbf{a} \neq \mathbf{b}\}$. Since the trivial codes are excluded, we know that $k \geqq 2$ and consequently, by Lemma 4,

$$(16) \qquad d_{\min} \leqq \tfrac{1}{2} n \left(1 - \frac{1}{4}\right)^{-1} = 2n/3.$$

On the other hand, by the definition of e-codes,

$$(17) \qquad d_{\min} \geqq 2e + 1.$$

The inequalities (16) and (17) imply

$$(18) \qquad n \geqq 3e + 2.$$

Put $t = e + 2$ in Lemma 5. Then, by (1),

$$\begin{aligned}
\mathrm{K} &\geqq 2^{k-n}\left(\sum_{i=0}^{e}\binom{n}{i} + \binom{n}{e+1} + \binom{n}{e+2}\right) \\
&= 1 + \left(\binom{n}{e+1} + \binom{n}{e+2}\right) \Big/ \sum_{i=0}^{e}\binom{n}{i} \\
&> 1 + \binom{n+1}{e+2} \Big/ \left(\binom{n}{e}\left(1 + \frac{e}{n-e+1} + \left(\frac{e}{n-e+1}\right)^2 + \cdots\right)\right) \\
&= 1 + \frac{(n+1)(n-e)(n-2e+1)}{(e+1)(e+2)(n-e+1)} \\
&> 1 + \frac{n(n-2e)}{(e+1)(e+2)}
\end{aligned}$$

and hence

$$(1-K^{-1})^{-1} = 1 + (K-1)^{-1} < 1 + \frac{(e+1)(e+2)}{n(n-2e)}.$$

Choosing $x = 2(e+2)/n$ in Lemma 6 we get therefore

$$d_{\min} < \frac{2(e+2)(n-e-2)}{n}\left(1 + \frac{(e+1)(e+2)}{n(n-2e)}\right).$$

Combining this with (17) we obtain

$$2(e+2)(n-e-2)(n^2-2en+e^2+3e+2) > n^2(n-2e)(2e+1)$$

or

$$\text{(19)} \qquad 3n^3-(2e^2+14e+8)n^2+(6e^3+26e^2+32e+8)n \\ -(2e^4+14e^3+36e^2+40e+16) > 0.$$

If $e \leqq 100$ then, by (15), $n < 10000$, and we have the case considered by Lemma 7. Therefore we may suppose that $e > 100$. Hence

$$\text{(20)} \qquad (2e+8)n^2+(2e^3/3-49e^2/3-32e-8)n \\ +(12e^3+36e^2+40e+16) > 0.$$

Combining the inequalities (19) and (20) we find

$$F(n) = 3n^3-(2e^2+12e)n^2+(20e^3/3+29e^2/3)n-(2e^4+2e^3) > 0.$$

Because the zeros of F are $e/3$, $3e$ and $2(e^2+e)/3$, and because, by (18), $n > 3e$, n must be $> 2(e^2+e)/3$. This contradicts (15).

APPENDIX

The computer search for the solutions of the equation (1) was carried out as follows:

Denote the sums

$$\sum_{i=0}^{e}\binom{n}{i} = S(n,e).$$

Then it is well known that

$$\text{(21)} \qquad S(n+1,e) = S(n,e) + S(n,e-1) \qquad (e = 1,\ldots,n).$$

The sums $S(n,e)$ $(e = 1,\ldots,m)$ were determined for certain initial value $n = n_a$ directly from binomial coefficients, which were calculated recursively from

$$\text{(22)} \qquad \binom{n}{0} = 1, \binom{n}{j} = \left((n-j+1)\binom{n}{j-1}\right)/j \qquad (j = 1,\ldots,n).$$

Then the successive rows of sums $S(n, 1), \ldots, S(n, m)$ $(n = n_a + 1, n_a + 2, \ldots)$ were obtained by using (21). When the row was formed in order $e = m, m-1, \ldots, 1$, then all extra movements of data were avoided. The only operations for each $S(n,e)$ were one addition and the inspection whether the sum is of form 2^{n-k}.

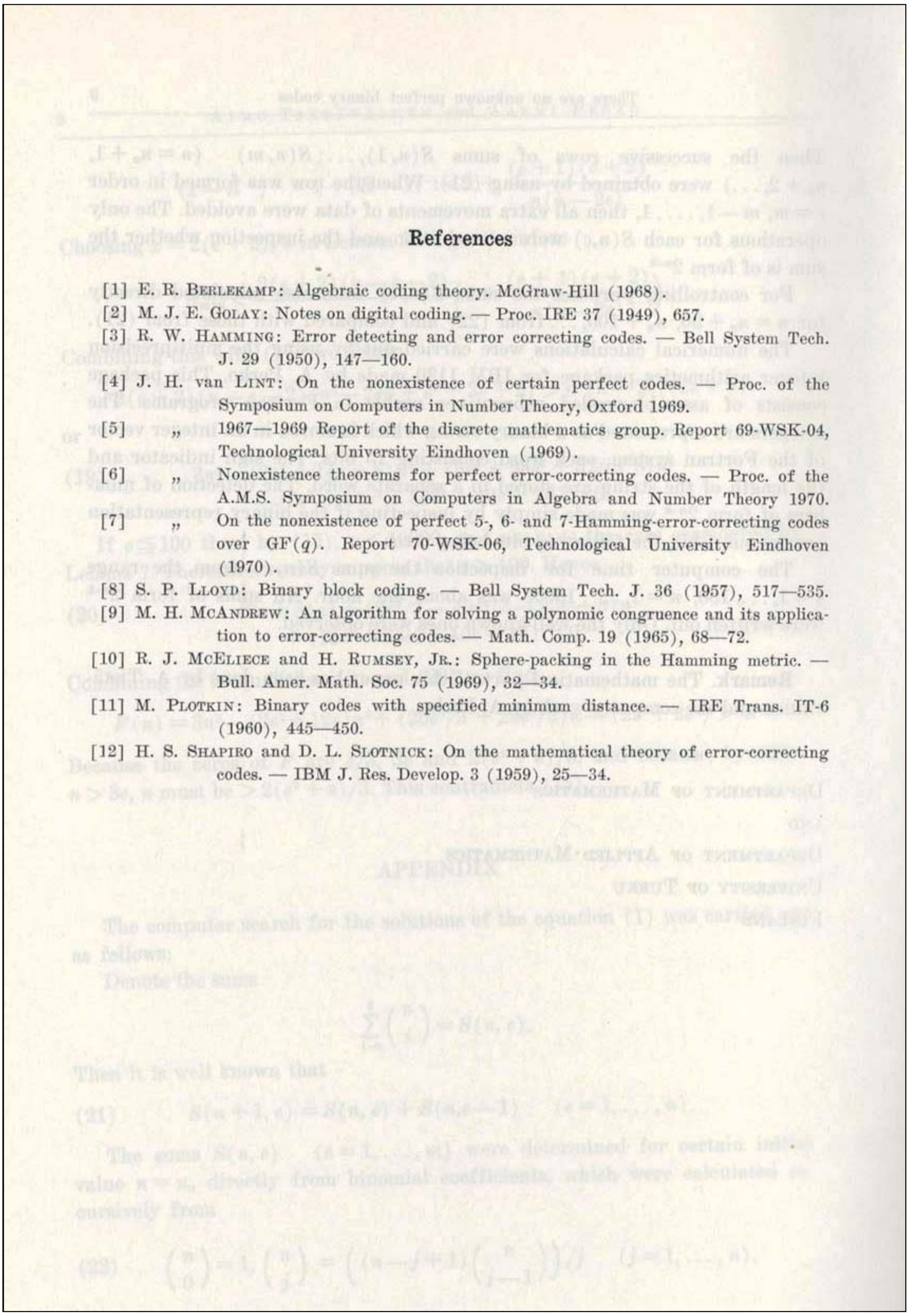
For controlling purposes the sums $S(n, e)$ were also calculated directly for $n = n_a + 50, n_a + 100, \ldots$ from (22) and compared with those from (21).

The numerical calculations were carried out by using the multiprecision integer arithmetics package for IBM 1130 made by A. Perko. This package consists of assembler-coded subroutines usable in Fortran programs. The integers are represented as a binary string which is stored in an integer vector of the Fortran system, each word consisting 15 bits. The sign indicator and the length of the string are stored in a separate word. The detection of numbers of form 2^{n-k} was made simply by inspecting if the binary representation carries one 1-bit, the following bits being zeros.

The computer time for inspection the sums $S(n, e)$ from the range $e = 1, \ldots, 100$, $n = 1, \ldots, 10000$ was about one hour. All sums of form 2^{n-k} were written out. Only the well-known ones were observed.

Remark. The mathematical part of this paper has been done by A. Tietäväinen and the computer work by A. Perko.

Department of Mathematics
and
Department of Applied Mathematics
University of Turku
Finland



References

[1] E. R. Berlekamp: Algebraic coding theory. McGraw-Hill (1968).

[2] M. J. E. Golay: Notes on digital coding. — Proc. IRE 37 (1949), 657.

[3] R. W. Hamming: Error detecting and error correcting codes. — Bell System Tech. J. 29 (1950), 147—160.

[4] J. H. van Lint: On the nonexistence of certain perfect codes. — Proc. of the Symposium on Computers in Number Theory, Oxford 1969.

[5] „ 1967—1969 Report of the discrete mathematics group. Report 69-WSK-04, Technological University Eindhoven (1969).

[6] „ Nonexistence theorems for perfect error-correcting codes. — Proc. of the A.M.S. Symposium on Computers in Algebra and Number Theory 1970.

[7] „ On the nonexistence of perfect 5-, 6- and 7-Hamming-error-correcting codes over GF(q). Report 70-WSK-06, Technological University Eindhoven (1970).

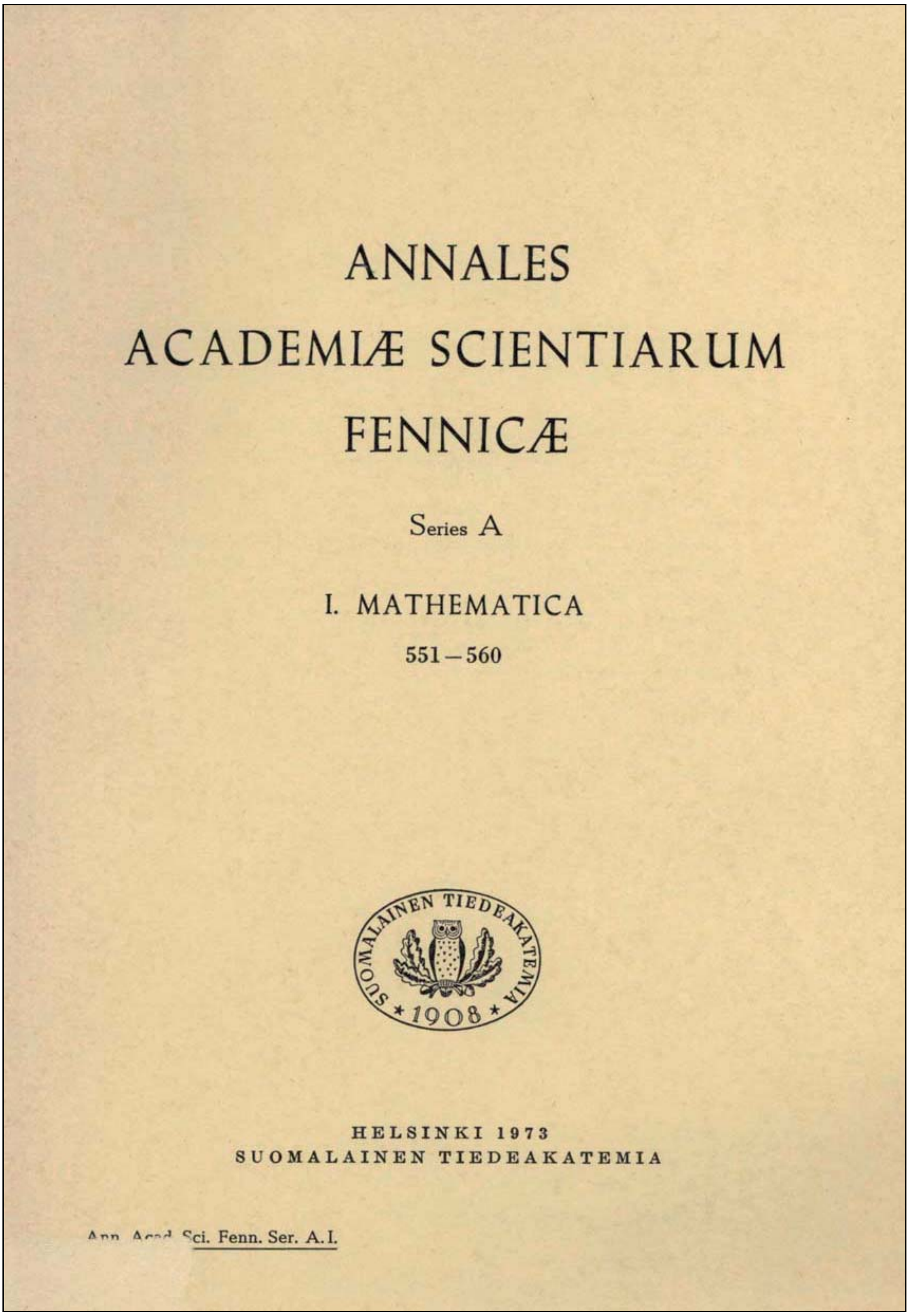
[8] S. P. Lloyd: Binary block coding. — Bell System Tech. J. 36 (1957), 517—535.

[9] M. H. McAndrew: An algorithm for solving a polynomic congruence and its application to error-correcting codes. — Math. Comp. 19 (1965), 68—72.

[10] R. J. McEliece and H. Rumsey, Jr.: Sphere-packing in the Hamming metric. — Bull. Amer. Math. Soc. 75 (1969), 32—34.

[11] M. Plotkin: Binary codes with specified minimum distance. — IRE Trans. IT-6 (1960), 445—450.

[12] H. S. Shapiro and D. L. Slotnick: On the mathematical theory of error-correcting codes. — IBM J. Res. Develop. 3 (1959), 25—34.



ANNALES ACADEMIÆ SCIENTIARUM FENNICÆ

Series A

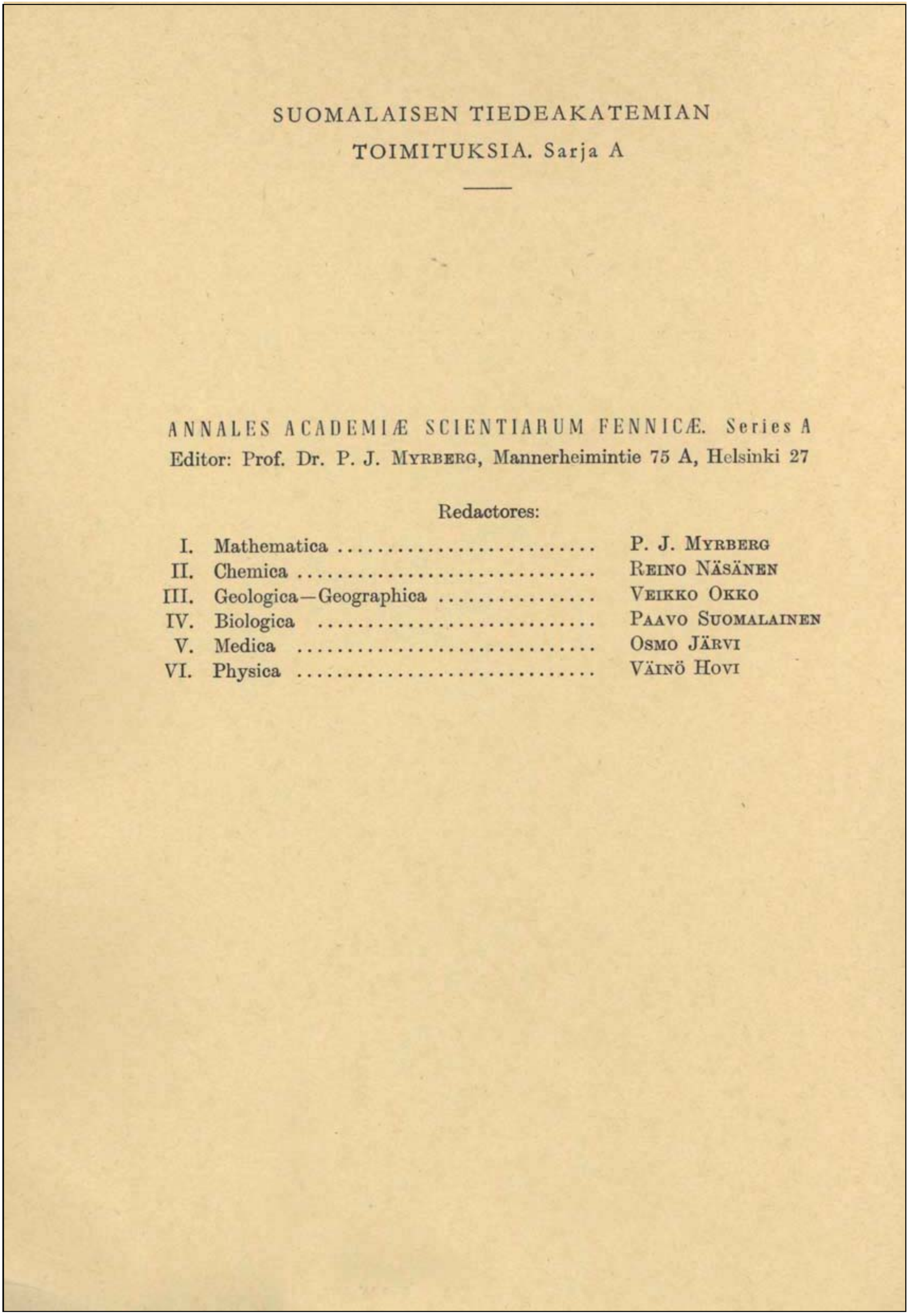
I. MATHEMATICA

551 – 560

SUOMALAINEN TIEDEAKATEMIA
1908

HELSINKI 1973
SUOMALAINEN TIEDEAKATEMIA

Ann. Acad. Sci. Fenn. Ser. A.I.



SUOMALAISEN TIEDEAKATEMIAN

TOIMITUKSIA. Sarja A

ANNALES ACADEMIÆ SCIENTIARUM FENNICÆ. Series A

ANNALES ACADEMIAE SCIENTIARUM FENNICAE

Series A

I. MATHEMATICA

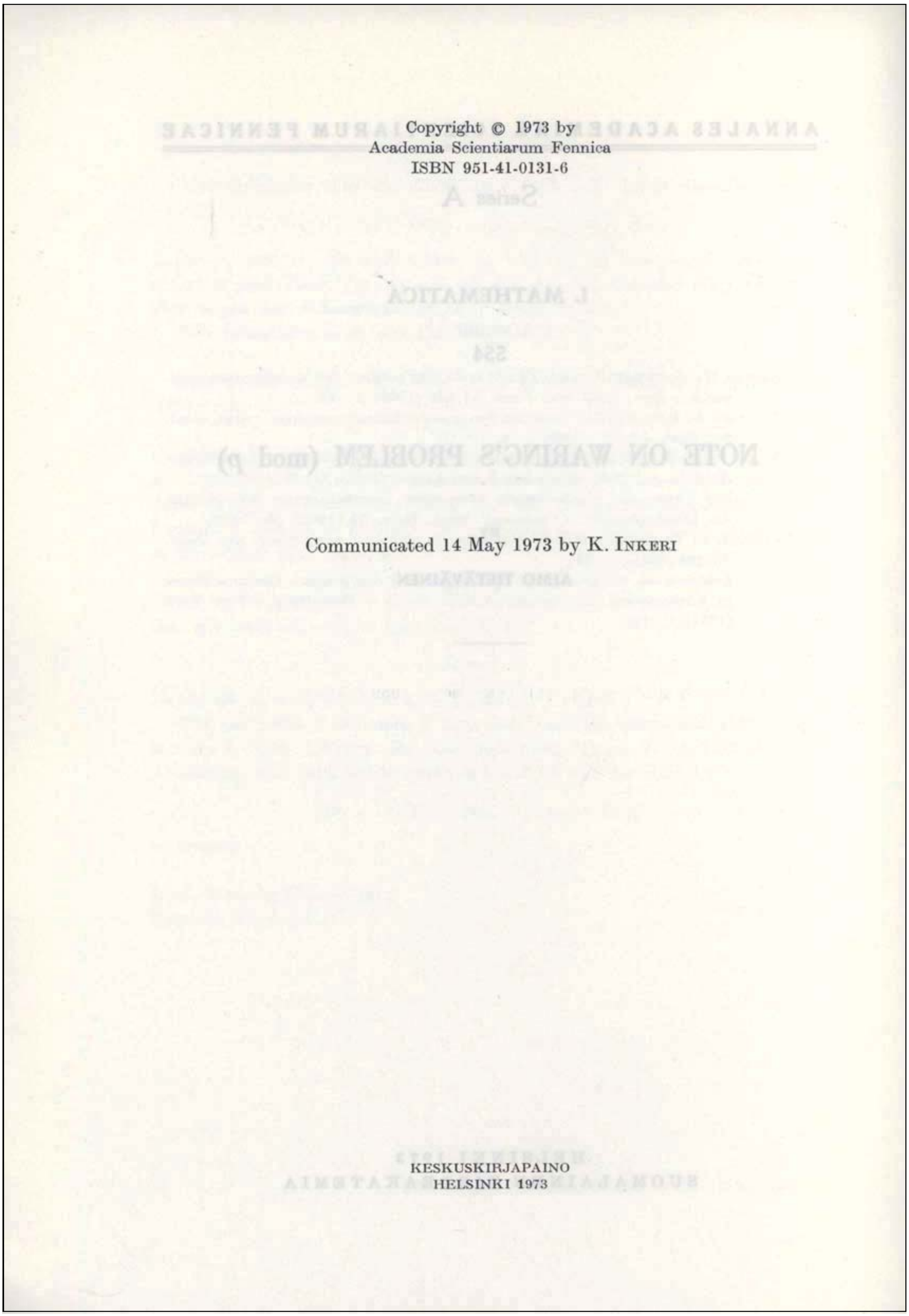
554

NOTE ON WARING'S PROBLEM (mod p)

BY

AIMO TIETÄVÄINEN

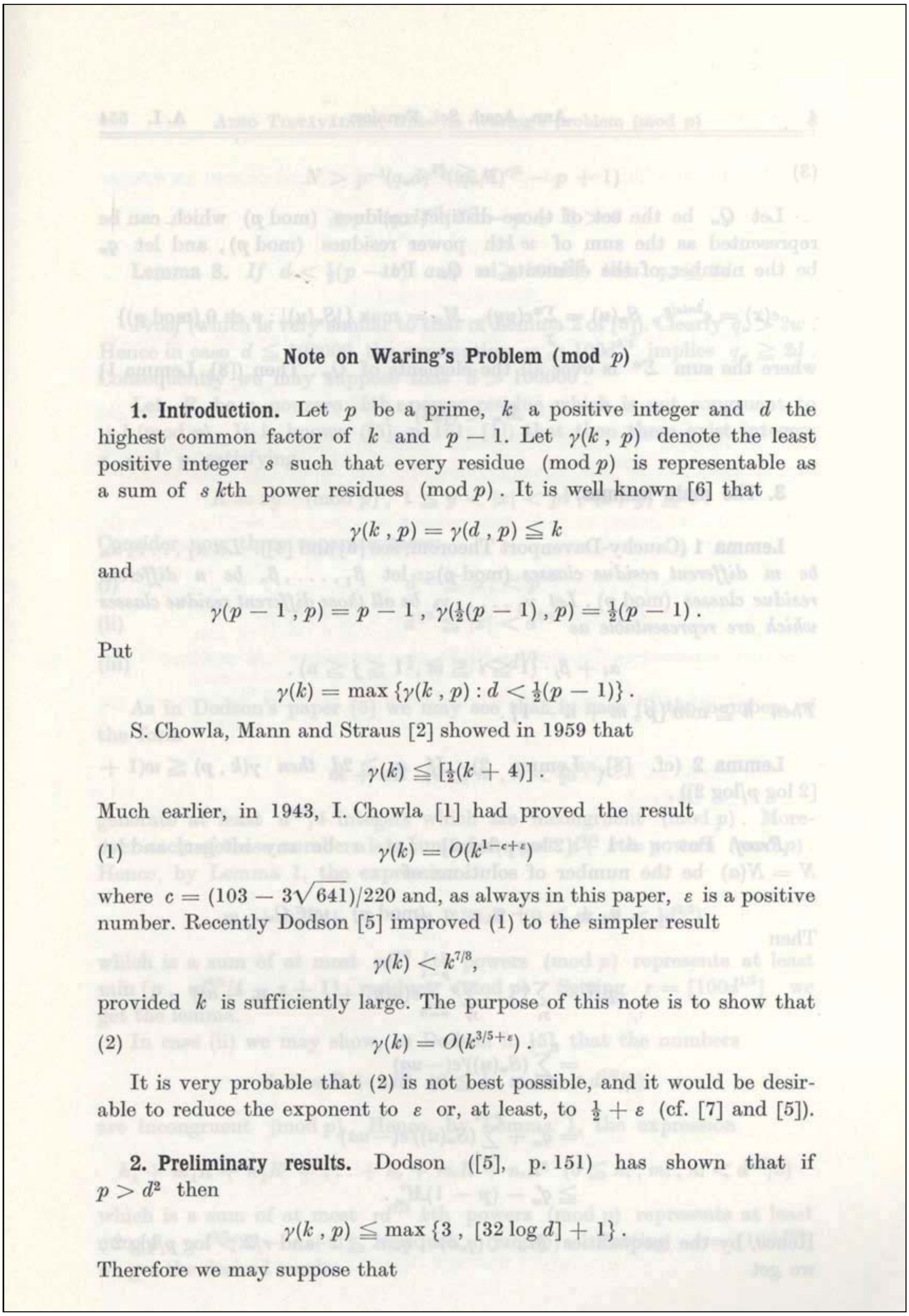
HELSINKI 1973
SUOMALAINEN TIEDEAKATEMIA



ISBN 951-41-0131-6

Communicated 14 May 1973 by K. Inkeri

KESKUSKIRJAPAINO
HELSINKI 1973



Note on Waring's Problem (mod p)

1. Introduction. Let p be a prime, k a positive integer and d the highest common factor of k and $p-1$. Let $\gamma(k\,,\,p)$ denote the least positive integer s such that every residue $(\operatorname{mod} p)$ is representable as a sum of s kth power residues $(\operatorname{mod} p)$. It is well known [6] that

$$\gamma(k\,,\,p)=\gamma(d\,,\,p)\leqq k$$

and

$$\gamma(p-1\,,\,p)=p-1\,,\ \gamma(\tfrac{1}{2}(p-1)\,,\,p)=\tfrac{1}{2}(p-1)\,.$$

Put

$$\gamma(k)=\max\{\gamma(k\,,\,p):d<\tfrac{1}{2}(p-1)\}\,.$$

S. Chowla, Mann and Straus [2] showed in 1959 that

$$\gamma(k)\leqq[\tfrac{1}{2}(k+4)]\,.$$

Much earlier, in 1943, I. Chowla [1] had proved the result

$$(1)\qquad \gamma(k)=O(k^{1-c+\varepsilon})$$

where $c=(103-3\sqrt{641})/220$ and, as always in this paper, ε is a positive number. Recently Dodson [5] improved (1) to the simpler result

$$\gamma(k)<k^{7/8},$$

provided k is sufficiently large. The purpose of this note is to show that

$$(2)\qquad \gamma(k)=O(k^{3/5+\varepsilon})\,.$$

It is very probable that (2) is not best possible, and it would be desirable to reduce the exponent to ε or, at least, to $\frac{1}{2}+\varepsilon$ (cf. [7] and [5]).

2. Preliminary results. Dodson ([5], p. 151) has shown that if $p>d^2$ then

$$\gamma(k\,,\,p)\leqq\max\{3\,,\,[32\log d]+1\}\,.$$

Therefore we may suppose that

$$p \leqq d^2 . \tag{3}$$

Let Q_w be the set of those distinct residues $(\mathrm{mod}\, p)$ which can be represented as the sum of w kth power residues $(\mathrm{mod}\, p)$, and let q_w be the number of the elements in Q_w. Put

$$e(x) = e^{2\pi i x/p} , \ S_w(u) = \sum_y{}^* e(uy) , \ M_w = \max \{ |S_w(u)| : u \not\equiv 0 \ (\mathrm{mod}\, p) \}$$

where the sum Σ^* is over all the elements of Q_w. Then ([8], Lemma 1)

$$M_w < (q_w d)^{1/2} .$$

3. The main lemmas.

Lemma 1 (Cauchy-Davenport Theorem; see [3] and [4]). *Let* $\alpha_1 , \ldots , \alpha_m$ *be* m *different residue classes* $(\mathrm{mod}\, p)$; *let* $\beta_1 , \ldots , \beta_n$ *be* n *different residue classes* $(\mathrm{mod}\, p)$. *Let* $\gamma_1 , \ldots , \gamma_h$ *be all those different residue classes which are representable as*

$$\alpha_i + \beta_j \quad (1 \leqq i \leqq m \, , \ 1 \leqq j \leqq n) .$$

Then $h \geqq \min \{ p \, , m + n - 1 \}$.

Lemma 2 (cf. [8], Lemma 2). *If* $q_w \geqq 2d$ *then* $\gamma(k \, , p) \leqq w(1 + [2 \log p/\log 2])$.

Proof. Put $r = 1 + [2 \log p/\log 2]$. Let a be any integer, and let $N = N(a)$ be the number of solutions of

$$y_1 + \ldots + y_r \equiv a \ \ (\mathrm{mod}\, p) \, , \ y_i \in Q_w .$$

Then

$$\begin{aligned} pN &= \sum_{y_1}{}^* \ldots \sum_{y_r}{}^* \sum_{u=0}^{p-1} e(u(y_1 + \ldots + y_r - a)) \\ &= \sum_{u=0}^{p-1} (S_w(u))^r e(-ua) \\ &= q_w^r + \sum_{u=1}^{p-1} (S_w(u))^r e(-ua) \\ &\geqq q_w^r - (p-1) M_w^r . \end{aligned}$$

Hence, by the inequalities $M_w < (q_w d)^{\frac{1}{2}}$, $q_w/d \geqq 2$ and $r/2 > \log p/\log 2$, we get

$$N > p^{-1}(q_w d)^{r/2}((q_w/d)^{r/2} - p + 1)$$
$$\geqq p^{-1}(q_w d)^{r/2}(2^{r/2} - p + 1) > 0\,.$$

Lemma 3. *If* $d < \frac{1}{2}(p-1)$ *and* $w \geqq 100d^{3/5}$ *then* $q_w \geqq 2d$.

Proof (which is very similar to that of Lemma 2 of [5]). Clearly $q_w > 2w$. Hence in case $d \leqq 100000$ the assumption $w \geqq 100d^{3/5}$ implies $q_w \geqq 2d$. Consequently we may suppose that $d > 100000$.

Let R be a nonzero kth power residue which is not congruent to $\pm 1 \pmod p$. It is known ([5], p. 151; [1]) that then there exist integers x and y satisfying

$$R \equiv xy^{-1} \pmod p\,,\ 1 \leqq y < |x| < p^{\frac{1}{2}}\,,\ (x\,,\ y) \leqq 1\,.$$

Consider now three separate cases:

(i) $\qquad d^{2/5} \leqq |x| < p^{1/2}$

(ii) $\qquad d^{1/5} \leqq |x| < d^{2/5}$

(iii) $\qquad 1 < |x| < d^{1/5}$.

As in Dodson's paper [5] we may see that in case (i) the numbers of the form

$$m + nR \quad (0 \leqq m\,,\ n < \tfrac{1}{2}d^{2/5})$$

generate at least $d^{4/5}/4$ integers which are incongruent $\pmod p$. Moreover each of these numbers is a sum of at most $d^{2/5}$ kth powers $\pmod p$. Hence, by Lemma 1, the expression

$$m_1 + n_1R + \ldots + m_r + n_rR \quad (0 \leqq m\,,\ n < \tfrac{1}{2}d^{2/5})$$

which is a sum of at most $rd^{2/5}$ kth powers $\pmod p$ represents at least $\min\{p\,,\ rd^{4/5}/4 - r + 1\}$ residues $\pmod p$. Setting $r = [100d^{1/5}]$ we get the lemma.

In case (ii) we may show, as Dodson in [5], that the numbers

$$h + mR + nR^2 \quad (0 \leqq h\,,\ m\,,\ n < d^{1/5}/3)$$

are incongruent $\pmod p$. Hence, by Lemma 1, the expression

$$h_1 + m_1R + n_1R^2 + \ldots + h_r + m_rR + n_rR^2 \quad (0 \leqq h_i\,,\ m_i\,,\ n_i < d^{1/5}/3)$$

which is a sum of at most $rd^{1/5}$ kth powers $\pmod p$ represents at least $\min\{p\,,\ rd^{3/5}/27 - r + 1\}$ residues $\pmod p$. Putting $r = [100d^{2/5}]$ we get the desired result.

Also in case (iii) we adopt the method of [5] and choose an integer f such that

$$d^{2/5} \leqq |x|^f < d^{3/5}.$$

Thus

$$R^f \equiv x^f y^{-f} \pmod p$$

where $(x^f, y^f) = 1$, $1 \leqq y^f < |x|^f$ and $R^f \not\equiv \pm 1 \pmod p$. Moreover (cf. [5], pp. 153—154) the numbers

$$m + nR^f \quad (0 \leqq m , n < \tfrac{1}{2}d^{2/5})$$

form at least $d^{4/5}/4$ distinct residues $(\mathrm{mod}\, p)$, each number being the sum of at most $d^{2/5}$ kth powers $(\mathrm{mod}\, p)$. The result now follows as in case (i).

4. Proof of (2). Lemma 3 implies that $q_w \geqq 2d$ if $w \geqq 100d^{3/5}$. It follows from this and Lemma 2 that

$$\gamma(k, p) < (1 + 100d^{3/5})(1 + 2 \log p/\log 2)\,. \tag{4}$$

Since we assumed in (3) that $p \leqq d^2$, the inequality (4) implies

$$\gamma(k, p) < (1 + 100d^{3/5})(1 + 4 \log d/\log 2) = O(k^{3/5+\varepsilon})\,.$$

University of Turku
Turku, Finland

References

[1] Chowla, I.: On Waring's Problem (mod p). - Proc. Indian Nat. Acad. Sci. A 13 (1943), 195–220.

[2] Chowla, S. — Mann, H. B. — Straus, E. G.: Some Applications of the Cauchy-Davenport Theorem. - Norske Vid. Selsk. Forh. (Trondheim) 32 (1959), 74–80.

[3] Davenport, H.: On the Addition of Residue Classes. - J. London Math. Soc. 10 (1935), 30–32.

[4] —»— A Historical Note. - J. London Math. Soc. 22 (1947), 100–107.

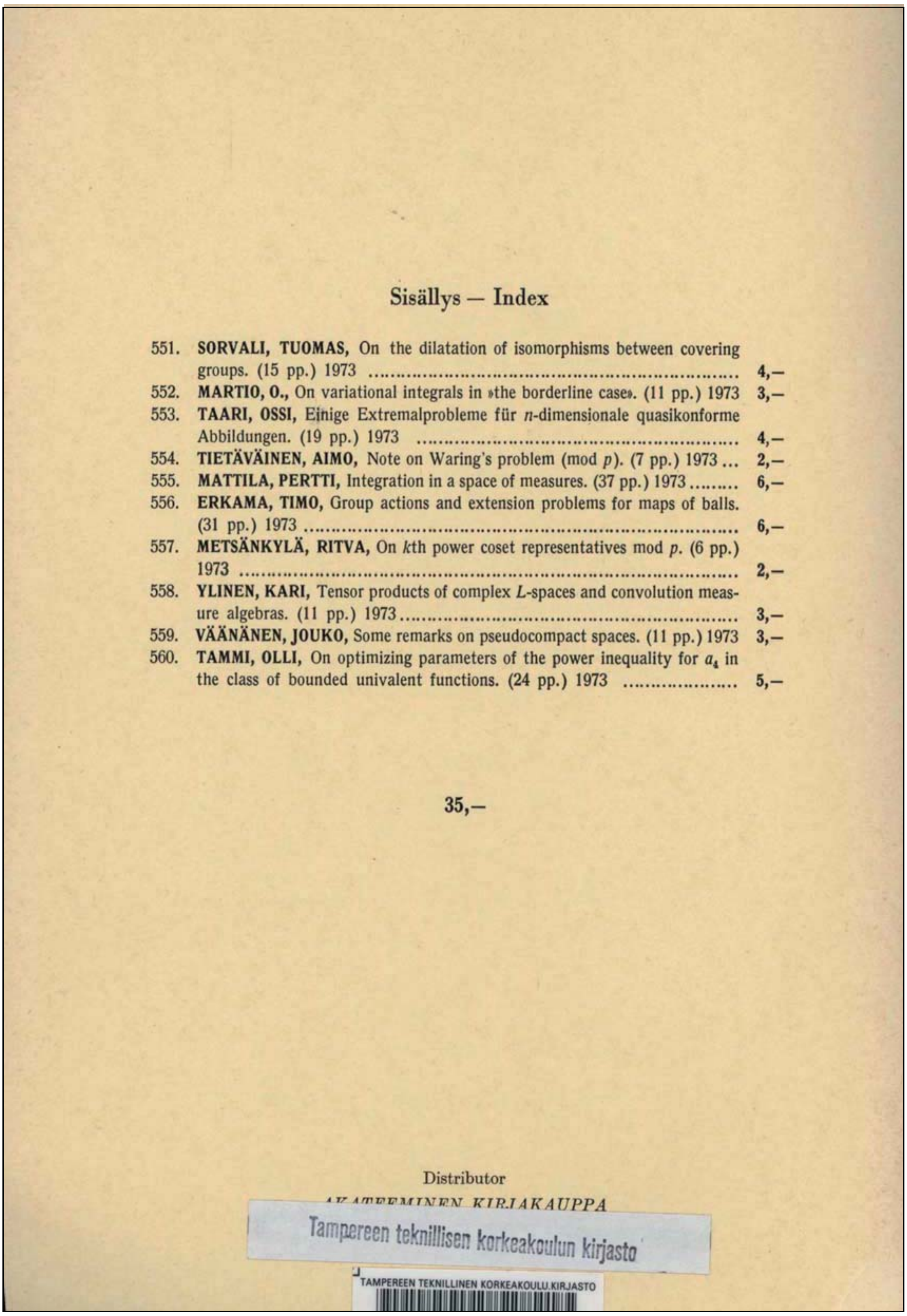
[5] Dodson, M.: On Waring's Problem in $GF(p)$. - Acta Arithmetica XIX (1971), 147–173.

[6] Hardy, G. H. — Littlewood, J. E.: Some Problems of 'Partitio Numerorum': VIII. The Number $\Gamma(k)$ in Waring's Problem. - Proc. London Math. Soc. 28 (1927), 518–542.

[7] Heilbronn, H.: Lecture Notes on Additive Number Theory mod p. California Institute of Technology, 1964.

[8] Tietäväinen, A.: Proof of a Conjecture of S. Chowla. - J. Number Theory (to appear).

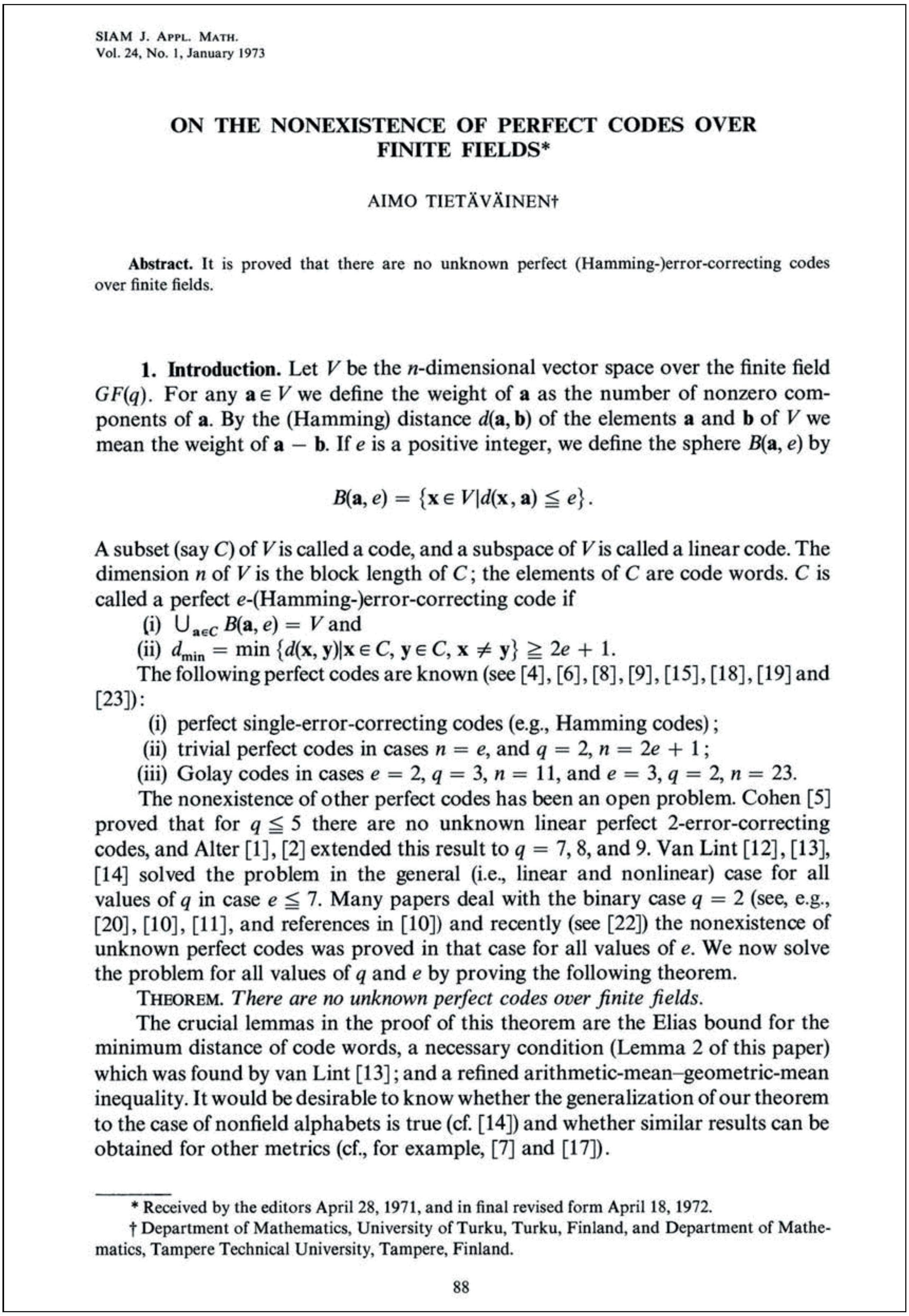
Printed September 1973



Sisällys — Index

35,—

Distributor
AKATEEMINEN KIRJAKAUPPA



SIAM J. APPL. MATH.
Vol. 24, No. 1, January 1973

ON THE NONEXISTENCE OF PERFECT CODES OVER FINITE FIELDS*

AIMO TIETÄVÄINEN†

Abstract. It is proved that there are no unknown perfect (Hamming-)error-correcting codes over finite fields.

1. Introduction. Let V be the n-dimensional vector space over the finite field $GF(q)$. For any $\mathbf{a} \in V$ we define the weight of $\mathbf{a}$ as the number of nonzero components of $\mathbf{a}$. By the (Hamming) distance $d(\mathbf{a}, \mathbf{b})$ of the elements $\mathbf{a}$ and $\mathbf{b}$ of V we mean the weight of $\mathbf{a} - \mathbf{b}$. If e is a positive integer, we define the sphere $B(\mathbf{a}, e)$ by

$$B(\mathbf{a}, e) = \{\mathbf{x} \in V | d(\mathbf{x}, \mathbf{a}) \leqq e\}.$$

A subset (say C) of V is called a code, and a subspace of V is called a linear code. The dimension n of V is the block length of C; the elements of C are code words. C is called a perfect e-(Hamming-)error-correcting code if

(i) $\bigcup_{\mathbf{a} \in C} B(\mathbf{a}, e) = V$ and

(ii) $d_{\min} = \min \{d(\mathbf{x}, \mathbf{y}) | \mathbf{x} \in C, \mathbf{y} \in C, \mathbf{x} \neq \mathbf{y}\} \geqq 2e + 1$.

The following perfect codes are known (see [4], [6], [8], [9], [15], [18], [19] and [23]):

(i) perfect single-error-correcting codes (e.g., Hamming codes);

(ii) trivial perfect codes in cases $n = e$, and $q = 2$, $n = 2e + 1$;

(iii) Golay codes in cases $e = 2$, $q = 3$, $n = 11$, and $e = 3$, $q = 2$, $n = 23$.

The nonexistence of other perfect codes has been an open problem. Cohen [5] proved that for $q \leqq 5$ there are no unknown linear perfect 2-error-correcting codes, and Alter [1], [2] extended this result to $q = 7, 8$, and 9. Van Lint [12], [13], [14] solved the problem in the general (i.e., linear and nonlinear) case for all values of q in case $e \leqq 7$. Many papers deal with the binary case $q = 2$ (see, e.g., [20], [10], [11], and references in [10]) and recently (see [22]) the nonexistence of unknown perfect codes was proved in that case for all values of e. We now solve the problem for all values of q and e by proving the following theorem.

THEOREM. *There are no unknown perfect codes over finite fields.*

The crucial lemmas in the proof of this theorem are the Elias bound for the minimum distance of code words, a necessary condition (Lemma 2 of this paper) which was found by van Lint [13]; and a refined arithmetic-mean–geometric-mean inequality. It would be desirable to know whether the generalization of our theorem to the case of nonfield alphabets is true (cf. [14]) and whether similar results can be obtained for other metrics (cf., for example, [7] and [17]).

* Received by the editors April 28, 1971, and in final revised form April 18, 1972.

† Department of Mathematics, University of Turku, Turku, Finland, and Department of Mathematics, Tampere Technical University, Tampere, Finland.

2. Preliminaries. Extending a result of Hamming [8], van Lint [12] proved the following lemma.

LEMMA 1. *If a perfect e-error-correcting code of block length n over GF(q) exists, then there is an integer k such that*

$$\sum_{i=0}^{e} \binom{n}{i} (q-1)^i = q^{n-k}, \tag{1}$$

and q^k is the cardinality of this code.

Lloyd [16] proved a theorem which gave a necessary condition for the existence of a binary perfect e-error-correcting code. This theorem was later generalized by F. J. MacWilliams and A. M. Gleason (see [3] and [15, pp. 103–112]). Using this generalization, van Lint [13], [15, Lemma 5.5.1] proved the next lemma.

LEMMA 2. *If a perfect e-error-correcting code of block length n over GF(q) exists, then there are distinct positive integers $x_1, x_2, \cdots, x_e$ such that*

$$x_1 + x_2 + \cdots + x_e = \frac{e(n-e)(q-1)}{q} + \frac{e(e+1)}{2} \tag{2}$$

and

$$x_1 x_2 \cdots x_e = e! q^{n-k-e}. \tag{3}$$

The Elias bound for the minimum distance may be given as the following two lemmas [4, Lemmas 13.61 and 13.62].

LEMMA 3. *Given an integer t and a code of block length n and cardinality q^k, there exists a critical sphere of radius t which includes K code words, where*

$$K \geqq q^{k-n} \sum_{i=0}^{t} \binom{n}{i} (q-1)^i.$$

By suitable translation of the code, this critical sphere may be centered at $(0, 0, \cdots, 0)$.

LEMMA 4. *If each of K code words has weight $\leqq (q-1)xn/q$, where $0 \leqq x \leqq 1$, then the distance between some pair of these K code words must be no greater than $x(2-x)(q-1)n/q(1-K^{-1})$.*

The special case $q = 2$ of our theorem, stated in the next lemma, was proved in [22].

LEMMA 5. *There are no unknown perfect codes over GF(2)*

Furthermore, van Lint ([11]–[14], [15, pp. 95, 96 and 116–118], see also [21]) proved the following lemmas.

LEMMA 6. *If a perfect e-error-correcting code of block length n over GF(q) exists $(e < n)$, then*

$$q \leqq (n-1)/e.$$

LEMMA 7. *If $e \geqq 4$ and $q = p^v$ with $p > e$, then there is no nontrivial perfect e-error-correcting code over GF(q).*

LEMMA 8. *If there exists an unknown perfect e-error-correcting code of block length n over GF(q), then*

(i) $e \geqq 8$ *and*

(ii) $q > 100$ *or* $n > 1000$ *or* $e > 1000$.

We also need the following refinement of the arithmetic-mean–geometric-mean inequality.

LEMMA 9. *Let $y_1, y_2, \cdots, y_s$ and p be positive integers such that $y_{i+1}/y_i \geqq p$, for every i. Then*

$$y_1 y_2 \cdots y_s \leqq R^{s-1}((y_1 + y_2 + \cdots + y_s)/s)^s, \tag{4}$$

where

$$R = 4p/(p+1)^2. \tag{5}$$

Proof (by induction). The assertion (4) is trivial for $s = 1$. Suppose now that $h \geqq 1$, $y_1 < y_2 < \cdots < y_h$,

$$y_1 y_2 \cdots y_h \leqq R^{h-1}((y_1 + y_2 + \cdots + y_h)/h)^h$$

and $y_{h+1}/y_h \geqq p$. Let $(y_1 + y_2 + \cdots + y_h)/h = Y$ and $y_{h+1} = zY$, whence $z \geqq p$. Then

$$y_1 y_2 \cdots y_{h+1} \leqq R^{h-1} z Y^{h+1}. \tag{6}$$

Let

$$f(x) = xY^{h+1}((hY + xY)/(h+1))^{-h-1} = x(h+1)^{h+1}(h+x)^{-h-1}.$$

Then f decreases on $[1, \infty)$, and hence

$$f(z) \leqq f(p) = p(1 + (p-1)/(h+1))^{-h-1} \leqq 4p(p+1)^{-2} = R.$$

Consequently,

$$zY^{h+1} \leqq R((hY + y_{h+1})/(h+1))^{h+1} = R((y_1 + y_2 + \cdots + y_{h+1})/(h+1))^{h+1}.$$

Combining this with (6), we get the assertion (4) in case $s = h + 1$.

3. Proof of Theorem in case $n \geqq \frac{1}{2}e^2 + e$. Assume the contrary: There exists an unknown perfect code with parameters e, n, and q, where $q = p^v$, p a prime, and

$$n \geqq \tfrac{1}{2}e^2 + e. \tag{7}$$

By Lemmas 5, 7 and 8, we may restrict ourselves to

$$q \geqq 3, \quad e \geqq p, \quad e \geqq 8. \tag{8}$$

For a positive integer m, define $A(m) = p^{-u}m$, where p^u is the highest power of p dividing m. Let the x_j, $1 \leqq j \leqq e$, be the numbers mentioned in Lemma 2. Denote $x_j \sim x_h$ if $A(x_j) = A(x_h)$. This relation $\sim$ defines a partition of the set $\{x_1, x_2, \cdots, x_e\}$ into disjoint subsets $X_1, \cdots, X_r$. It was proved in [22] that

$$e - r > (5e \log 2)/(4 \log e) - 1 \quad \text{for } p = 2. \tag{9}$$

We now show that generally

$$e - r \geqq [e/p] \log p/\log e. \tag{10}$$

Since $e \geqq p$ and $e - r$ is an integer, this implies

$$e - r \geqq 1. \tag{11}$$

It follows from (3) that

$$A(x_1x_2 \cdots x_e) = A(e!). \tag{12}$$

For a real number a, let $Q(a)$ be the product of the positive integers not exceeding a and not divisible by p, and, furthermore, let $[a]$ be the largest integer not exceeding a. Then

$$\begin{aligned} A(e!) &\leqq Q(e) \cdot [e/p]! \\ &\leqq Q(e)(e/p)^{[e/p]} \\ &= Q(e)e^{[e/p](1-(\log p)/(\log e))}. \end{aligned} \tag{13}$$

On the other hand, $A(x_1x_2 \cdots x_e)$ is greater than or equal to the product of those r least positive integers which are not divisible by p. Hence

$$A(x_1x_2 \cdots x_e) \geqq Q(e)e^{r-e+[e/p]}. \tag{14}$$

Collecting the results (12), (13) and (14), we get the assertion (10).

Let X_i be any one of the sets $X_1, \cdots, X_r$, let $s(i)$ be the cardinality of X_i, and let

$$R_i = \left(\prod_{x\in X_i} x\right) \bigg/ \left(\sum_{x \in X_i} \frac{x}{s(i)}\right)^{s(i)}.$$

Now we may apply Lemma 9 which gives the result

$$R_i \leqq R^{s(i)-1},$$

where R is defined by (5). It follows from this that

$$R_1 \cdots R_r \leqq \prod_{i=1}^{r} R^{s(i)-1} = R^{e-r}$$

or

$$x_1x_2 \cdots x_e \leqq R^{e-r} \prod_{i=1}^{r} \left(\sum_{x\in X_i} \frac{x}{s(i)}\right)^{s(i)},$$

which implies, by the arithmetic-mean–geometric-mean inequality, that

$$x_1x_2 \cdots x_e \leqq R^{e-r}((x_1 + x_2 + \cdots + x_e)/e)^e.$$

Using Lemma 2 and recalling (1), we get therefore

$$q^{-e}e! \sum_{i=0}^{e} \binom{n}{i}(q-1)^i \leqq R^{e-r}\left(\frac{(n-e)(q-1)}{q} + \frac{e+1}{2}\right)^e,$$

and, consequently,

$$R^{e-r} > (n-b)^{-e}e!\binom{n}{e} = \prod_{j=0}^{e-1}\left(1 + \frac{b-j}{n-b}\right), \tag{15}$$

where

$$b = e - \frac{q(e+1)}{2(q-1)}. \tag{16}$$

Let $c = [b] + 1$. Then

$$\begin{aligned}\prod_{j=0}^{e-1}\left(1+\frac{b-j}{n-b}\right) &= \prod_{j=0}^{c-1}\left(1+\frac{b-j}{n-b}\right)\prod_{j=c}^{e-1}\left(1+\frac{b-j}{n-b}\right)\\ &> \left(1+\sum_{j=0}^{c-1}\frac{b-j}{n-b}\right)\left(1+\sum_{j=c}^{e-1}\frac{b-j}{n-b}\right)\\ &= 1-\frac{e(e-2b-1)}{2(n-b)}-\frac{c(2b-c+1)(e-c)(c+e-2b-1)}{4(n-b)^2}\\ &\geqq 1-\frac{e(e-2b-1)}{2(n-b)}-\frac{(2b+1)^2(2e-2b-1)^2}{16(n-b)^2}.\end{aligned}$$

Using (15) and (16) and recalling that $n \geqq \frac{1}{2}e^2 + e$, we thus obtain

$$\begin{aligned}R^{e-r} &> 1-\frac{e^2+e}{2(q-1)n-(q-2)e+q}-\frac{e^2(q-2)(e+1)^2q}{16(2(q-1)n-(q-2)e+q)^2}\\ &> 1-\frac{1}{q-1}-\frac{(q-2)q}{16(q-1)^2} > \frac{15}{16}-\frac{1}{q-1}.\end{aligned}\tag{17}$$

If $p \geqq 5$ then, by (5), (11) and (17), $5/9 \geqq R^{e-r} > 11/16$, a contradiction.

Suppose now that $p = 3$. Then $q = 3$, for in case $q \geqq 9$ (17) implies $3/4 > 13/16$, a contradiction. For $q = 3$, (17) takes the form

$$(3/4)^{[e/3]\log 3/\log e} > 29/64$$

or

$$\frac{[e/3]\log 3}{\log e} < \frac{\log(64/29)}{\log(4/3)}.$$

Hence

$$e \leqq 26, \tag{18}$$

and it follows, by Lemma 8, that

$$n > 1000. \tag{19}$$

The inequalities (18) and (19) imply in case $q = 3$ that

$$1-\frac{e^2+e}{2(q-1)n-(q-2)e+q}-\frac{e^2(q-2)(e+1)^2q}{16(2(q-1)n-(q-2)e+q)^2} > \frac{3}{4}.$$

Substituting this and the equation $R = 3/4$ in (17) and recalling that $e - r \geqq 1$, we get an impossibility.

Suppose finally that $p = 2$, whence $q \geqq 4$. Because $e \geqq 8$, we know, by (9), that $e - r \geqq 3$, and, using a similar argument as in case $p = 3$, we see that $q = 4$. Thus, by (9), we may write the inequality (17) in the form

$$(8/9)^{(5e\log 2)/(4\log e)-1} > 11/18$$

or

$$\frac{e\log 2}{\log e} < \frac{4}{5}\,\frac{\log(18/11)}{\log(9/8)} + 1 < 5.$$

Hence $e < 32$ and, by Lemma 8, $n > 1000$. Consequently we get, by (17), the impossibility

$$\left(\frac{8}{9}\right)^3 > 1 - \frac{e^2+e}{6n-2e+4} - \frac{e^2(e+1)^2}{2(6n-2e+4)^2} > 1 - \frac{1}{5} - \frac{1}{50}.$$

4. Proof of Theorem in case $n < \frac{1}{2}e^2 + e$. Suppose that, contrary to our assertion, there exists an unknown perfect code C with parameters e, n and q such that

$$n < \tfrac{1}{2}e^2 + e, \quad q \geqq 3, \quad e \geqq 8. \tag{20}$$

Then we know, by the definition of e-error-correcting codes, that

$$n \geqq d_{\min} \geqq 2e + 1. \tag{21}$$

Put $t = e + 1$ in Lemma 3. Then, by (1),

$$\begin{aligned}
K &\geqq q^{k-n}\left(\sum_{i=0}^{e}\binom{n}{i}(q-1)^i + \binom{n}{e+1}(q-1)^{e+1}\right)\\
&= 1 + \binom{n}{e+1}(q-1)^{e+1}\left(\sum_{i=0}^{e}\binom{n}{i}(q-1)^i\right)^{-1}\\
&= 1 + \binom{n}{e+1}(q-1)^{e+1}\binom{n}{e}^{-1}(q-1)^{-e}\left(1 + \frac{e}{(n-e+1)(q-1)} + \cdots\right)^{-1}\\
&> 1 + \frac{(n-e)((n-e+1)(q-1)-e)}{(e+1)(n-e+1)},
\end{aligned}\tag{22}$$

and hence, by (21),

$$K > 1 + \frac{(q-2)(n-e)}{e+1}.$$

Consequently,

$$(1-K^{-1})^{-1} = 1 + (K-1)^{-1} < 1 + \frac{e+1}{(q-2)(n-e)}. \tag{23}$$

Choosing $x = (q-1)^{-1}n^{-1}(e+1)q$ in Lemma 4, we get therefore

$$d_{\min} < \frac{(e+1)(2(q-1)n-(e+1)q)}{(q-1)n}\left(1 + \frac{e+1}{(q-2)\,(n-e)}\right). \tag{24}$$

Using the same method as above, but choosing $t = e + 2$, $q = 3$ and $x = 3(e+2)/2n$, we get

$$d_{\min} < \frac{(e+2)(4n-3e-6)}{2n}\left(1 + \frac{(e+2)^2}{(2n-e)(2n-3e+2)}\right) \quad \text{for } q = 3. \tag{25}$$

Consider first the case $q \geqq 5$. We shall show that the inequalities (24) and (21) imply

$$F(n) = n^2 - (\tfrac{1}{2}e^2 + 3e)n + e^3 + 2e^2 > 0. \tag{26}$$

Since the zeros of F are $2e$ and $\frac{1}{2}e^2 + e$ and since we know, by (21), than $n > 2e$, n must be greater than $\frac{1}{2}e^2 + e$. This contradicts (20).

If $q \geqq 7$, then

$$1 + \frac{e+1}{(q-2)(n-e)} \leqq \frac{5n - 4e + 1}{5(n-e)}. \tag{27}$$

Furthermore, for all q,

$$\frac{2(q-1)n - (e+1)q}{(q-1)n} < \frac{2n - e - 1}{n}. \tag{28}$$

Combining the inequalities (24), (21), (27) and (28), we get

$$(e+1)(2n - e - 1)(5n - 4e + 1) > 5(n-e)n(2e+1)$$

or

$$5n^2 - (3e^2 + 11e + 3)n + 4e^3 + 7e^2 + 2e - 1 > 0.$$

Since

$$(\tfrac{1}{2}e^2 - 4e + 3)n + e^3 + 3e^2 - 2e + 1 > 0,$$

this implies (26).

If $q = 5$, the inequalities (24) and (21) imply

$$12n^2 - (7e^2 + 26e + 7)n + 10e^3 + 15e^2 - 5 > 0. \tag{29}$$

If $e \leqq 40$, then, by (20), $n < 1000$, and we have the case considered by Lemma 8. Therefore $e > 40$ and hence

$$(e^2 - 10e + 7)n + 2e^3 + 9e^2 + 5 > 0. \tag{30}$$

Now (26) follows from (29) and (30).

Suppose now that $q = 4$. Then, by Lemma 6, $n > 4e$; and it follows that we may replace the assertion (26) by

$$F_1(n) = 2n^2 - (e^2 + 10e)n + 4e^3 + 8e^2 > 0, \tag{31}$$

because the zeros of F_1 are $4e$ and $\frac{1}{2}e^2 + e$. For the proof of (31) we use (22) and get

$$(1 - K^{-1})^{-1} < 1 + \frac{(e+1)(n-e+1)}{(n-e)(3n-4e+3)} \leqq \frac{3n - 3e + 5}{3n - 4e + 3}.$$

Therefore,

$$2e + 1 < \frac{(e+1)(6n - 4e - 4)(3n - 3e + 5)}{3n(3n - 4e + 3)}$$

or

$$9n^2 - (6e^2 + 18e - 9)n + 12e^3 + 4e^2 - 28e - 20 > 0. \tag{32}$$

Since we may suppose, as in case $q = 5$, that $e > 40$, we have

$$(3e^2/2 - 27e - 9)n + 6e^3 + 32e^2 + 28e + 20 > 0. \tag{33}$$

The inequalities (32) and (33) imply the assertion (31).

Suppose finally that $q = 3$. Combining the inequalities (25) and (21), we find

$$\begin{aligned} &12n^3 - (6e^2 + 48e + 12)n^2 + (14e^3 + 63e^2 + 42e - 8)n \\ &\qquad\qquad - 6e^4 - 27e^3 - 42e^2 - 36e - 24 > 0. \end{aligned} \tag{34}$$

Since we may suppose, as in case $q = 5$, that $e > 40$, we have

$$(6e + 12)n^2 + (e^3 - 21e^2 - 42e + 8)n + 15e^3 + 42e^2 + 36e + 24 > 0. \tag{35}$$

Combining the inequalities (34) and (35), we obtain

$$F_2(n) = 12n^3 - (6e^2 + 42e)n^2 + (15e^3 + 42e^2)n - (6e^4 + 12e^3) > 0.$$

Because the zeros of F_2 are $\frac{1}{2}e$, $2e$ and $\frac{1}{2}e^2 + e$, and because, by (21), $n > 2e$, n must be greater than $\frac{1}{2}e^2 + e$. This contradicts (20).

Acknowledgment. The author would like to thank Professor J. H. van Lint and the referee for many very valuable suggestions and remarks.

REFERENCES

[1] R. Alter, *On the nonexistence of close-packed double Hamming-error-correcting codes on $q = 7$ symbols*, J. Comput. System Sci., 2 (1968), pp. 169–176.

[2] ———, *On the nonexistence of perfect double Hamming-error-correcting codes on $q = 8$ and $q = 9$ symbols*, Information and Control, 13 (1968), pp. 619–627.

[3] E. F. Assmus Jr., H. F. Mattson Jr. and R. Turyn, *Cyclic codes*, Rep. AFCRL-66-348, Applied Research Laboratory, Sylvania Electronic Systems, 1966.

[4] E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

[5] E. L. Cohen, *A note on perfect double error-correcting codes on q symbols*, Information and Control, 7 (1964), pp. 381–384.

[6] M. J. E. Golay, *Notes on digital coding*, Proc. IRE, 37 (1949), p. 657.

[7] S. W. Golomb and L. R. Welch, *Algebraic coding and the Lee metric*, Error Correcting Codes, H. B. Mann, ed., John Wiley, New York, 1968, pp. 175–194.

[8] R. W. Hamming, *Error detecting and error correcting codes*, Bell System Tech. J., 29 (1950), pp. 147–160.

[9] B. Lindström, *On group and nongroup perfect codes in q symbols*, Math. Scand., 25 (1969), pp. 149–158.

[10] J. H. van Lint, *On the nonexistence of certain perfect codes*, Computers in Number Theory, Proc. Science Research Council Atlas Symposium (Oxford, 1969), 1971.

[11] ———, *1967–1969 report of the discrete mathematics group*, Rep. 69-WSK-04, Technological University, Eindhoven, 1969.

[12] ———, *On the nonexistence of perfect 2- and 3-Hamming-error-correcting codes over GF(q)*, Information and Control, 16 (1970), pp. 396–401.

[13] ———, *Nonexistence theorems for perfect error-correcting codes*, Computers in Algebra and Number Theory, SIAM-AMS, vol. 3, to appear.

[14] ———, *On the nonexistence of perfect 5-, 6- and 7-Hamming-error-correcting codes over GF(q)*, Rep. 70-WSK-06, Technological University, Eindhoven, 1970.

[15] ———, *Coding Theory*, Lecture Notes in Mathematics 201, Springer-Verlag, Berlin, 1971.

[16] S. P. Lloyd, *Binary block coding*, Bell System Tech. J., 36 (1957), pp. 517–535.

[17] F. J. MacWilliams, *Error-correcting codes—a historical survey*, Error-Correcting Codes, H. B. Mann, ed., John Wiley, New York, 1968, pp. 3–13.

[18] V. Pless, *On the uniqueness of the Golay codes*, J. Combinatorial Theory, 5 (1968), pp. 215–228.

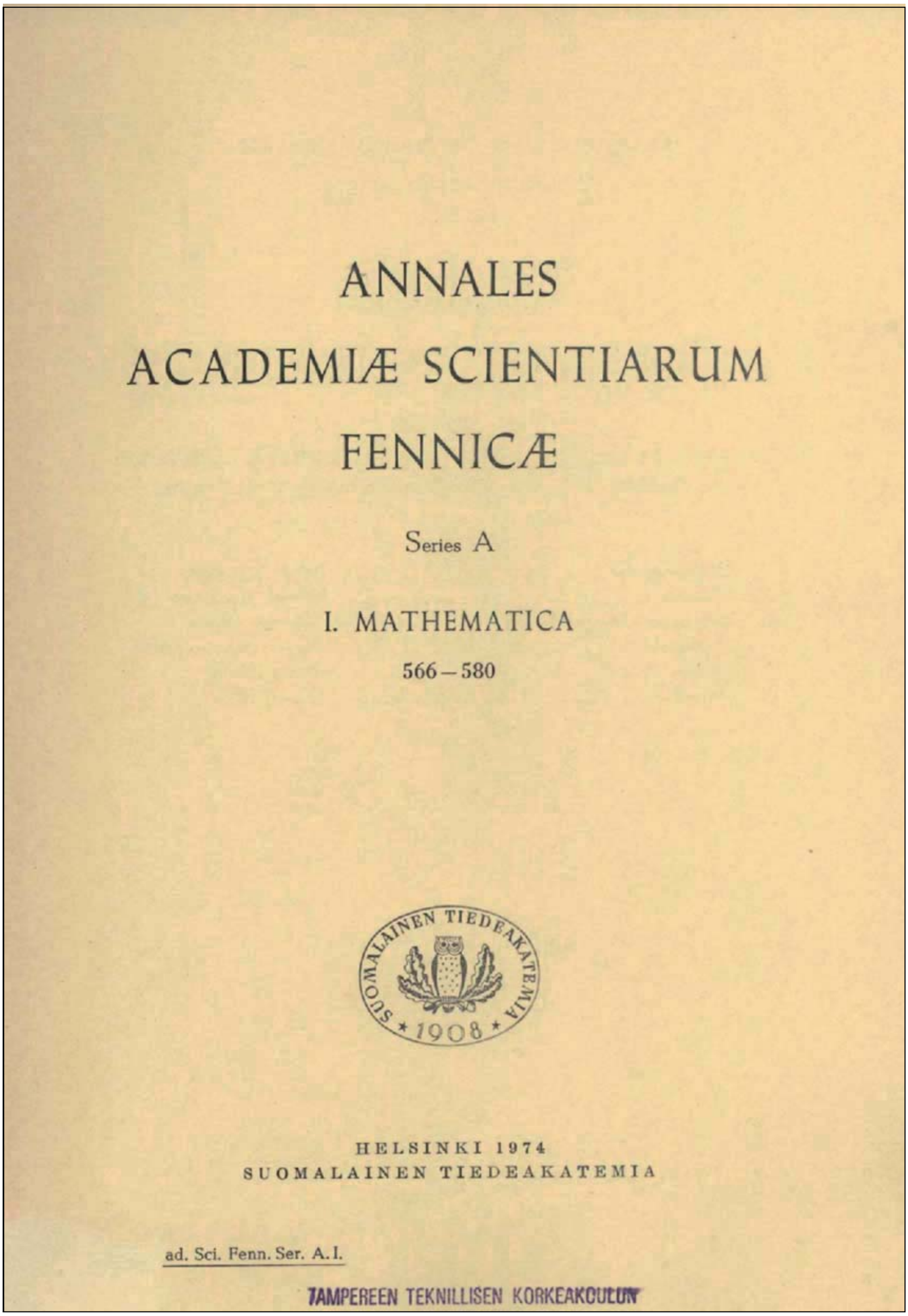
[19] J. Schönheim, *On linear and nonlinear single-error-correcting q-nary perfect codes*, Information and Control, 12 (1968), pp. 23–26.

[20] H. S. Shapiro and D. L. Slotnick, *On the mathematical theory of error-correcting codes*, IBM J. Res. Develop., 3 (1959), pp. 25–34.

[21] A. Tietäväinen, *On the nonexistence of perfect 4-Hamming-error-correcting codes*, Ann. Acad. Sci. Fenn., Ser. A I, 485 (1970), pp. 3–6.

[22] A. Tietäväinen and A. Perko, *There are no unknown perfect binary codes*, Ann. Univ. Turku., Ser. A I, 148 (1971), pp. 3–10.

[23] Ju. L. Vasilev, *On non-group close-packed codes*, Problemy Kibernet., 8 (1962), pp. 337–339 (in Russian); Probleme der Kybernetik, 8 (1965), pp. 375–378. (In German.)



ANNALES ACADEMIÆ SCIENTIARUM FENNICÆ

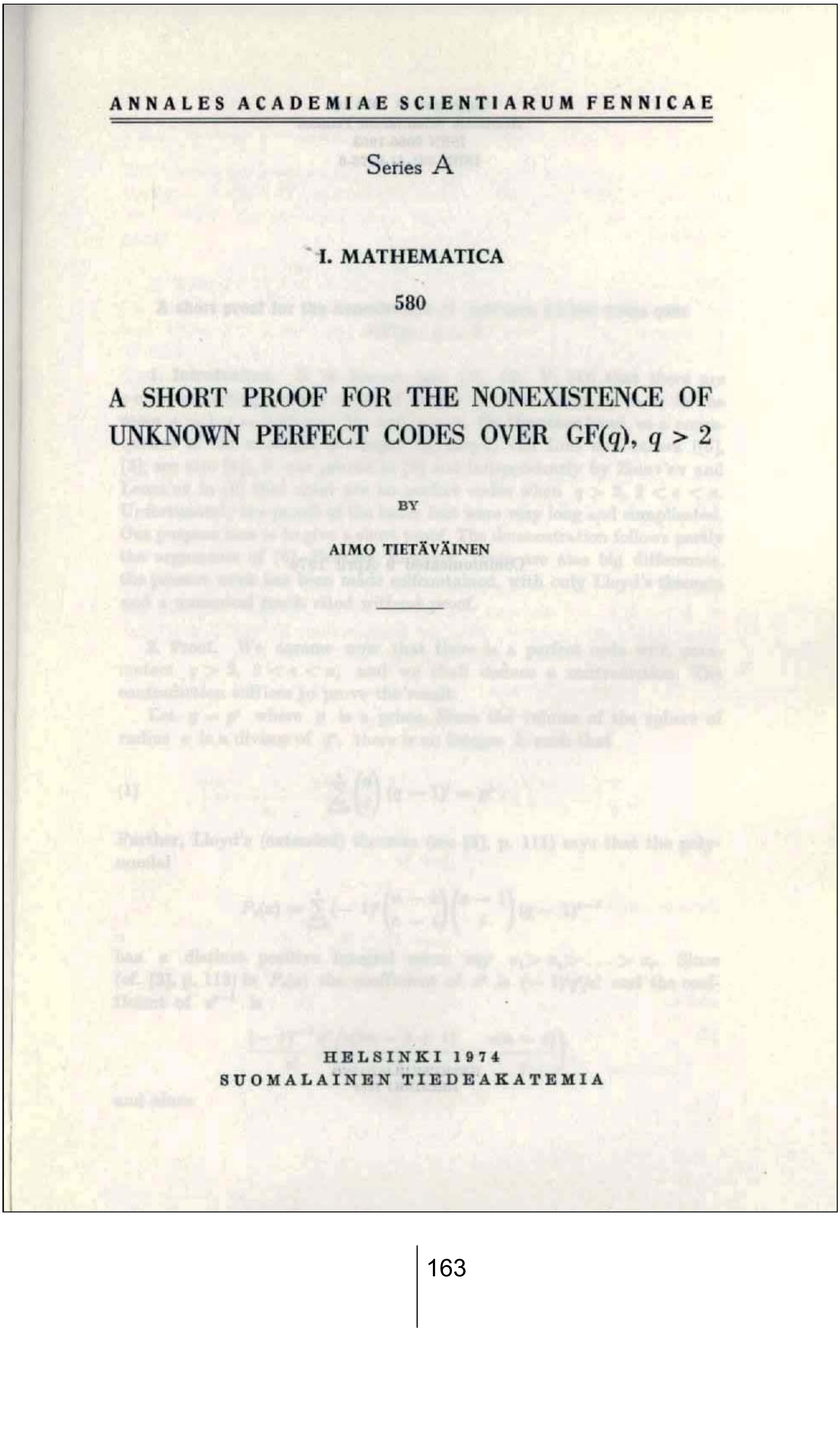
Series A

I. MATHEMATICA

566 – 580

SUOMALAINEN TIEDEAKATEMIA
* 1908 *

HELSINKI 1974
SUOMALAINEN TIEDEAKATEMIA



ANNALES ACADEMIAE SCIENTIARUM FENNICAE

Series A

I. MATHEMATICA

580

A SHORT PROOF FOR THE NONEXISTENCE OF UNKNOWN PERFECT CODES OVER GF(q), $q > 2$

BY

AIMO TIETÄVÄINEN

HELSINKI 1974
SUOMALAINEN TIEDEAKATEMIA

ISSN 0066-1953
ISBN-951-41-0176-6

Communicated 9 April 1974

KESKUSKIRJAPAINO
HELSINKI 1974

A short proof for the nonexistence of unknown perfect codes over GF(q), $q > 2$

1. Introduction. It is known (see [3], Ch. V; [4]) that there are e-error-correcting perfect codes of length n over GF(q), $q > 2$, in the cases $e = 1$, $e = 2$ (if $q = 3$), and $e = n$. On the other hand, as a consequence of the methods developed by Lloyd, van Lint and others ([5], [3]; see also [4]), it was proved in [6] and independently by Zinov'ev and Leont'ev in [9] that there are no perfect codes when $q > 2$, $2 < e < n$. Unfortunately the proofs of the latter fact were very long and complicated. Our purpose here is to give a short proof. The demonstration follows partly the arguments of [6]. Because, however, there are also big differences, the present work has been made selfcontained, with only Lloyd's theorem and a numerical result cited without proof.

2. Proof. We assume now that there is a perfect code with parameters $q > 2$, $2 < e < n$, and we shall deduce a contradiction. The contradiction suffices to prove the result.

Let $q = p^r$ where p is a prime. Since the volume of the sphere of radius e is a divisor of q^n, there is an integer k such that

$$(1) \qquad \sum_{i=0}^{e} \binom{n}{i} (q-1)^i = p^k .$$

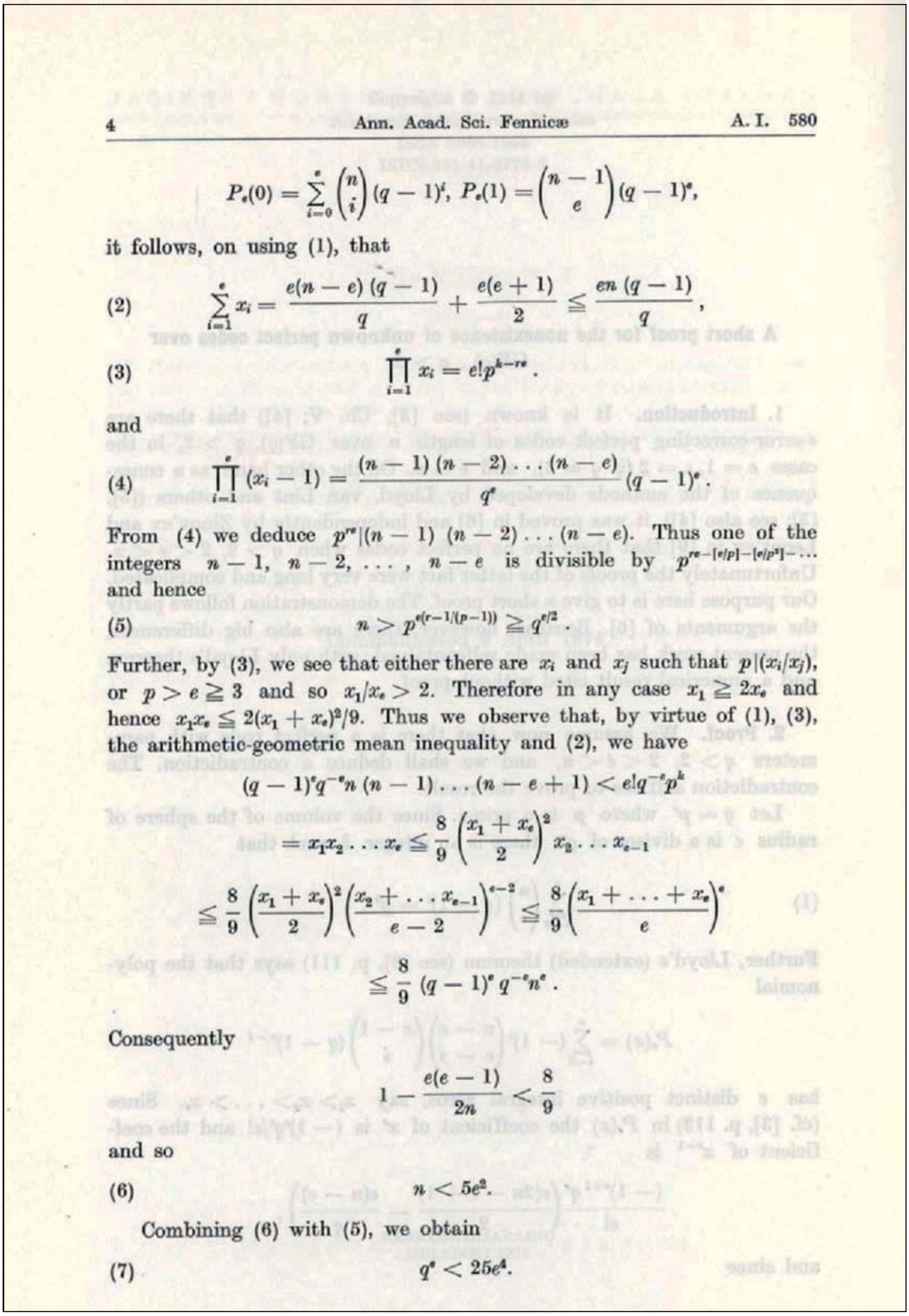
Further, Lloyd's (extended) theorem (see [3], p. 111) says that the polynomial

$$P_e(x) = \sum_{i=0}^{e} (-1)^i \binom{n-x}{e-i} \binom{x-1}{i} (q-1)^{e-i}$$

has e distinct positive integral zeros, say $x_1 > x_2 > \ldots > x_e$. Since (cf. [3], p. 113) in $P_e(x)$ the coefficient of x^e is $(-1)^e q^e / e!$ and the coefficient of x^{e-1} is

$$\frac{(-1)^{e+1} q^e}{e!} \left(\frac{e(2n-e+1)}{2} - \frac{e(n-e)}{q} \right),$$

and since



$$P_e(0) = \sum_{i=0}^{e} \binom{n}{i} (q-1)^i, \ P_e(1) = \binom{n-1}{e} (q-1)^e,$$

it follows, on using (1), that

$$\sum_{i=1}^{e} x_i = \frac{e(n-e)(q-1)}{q} + \frac{e(e+1)}{2} \leqq \frac{en(q-1)}{q}, \tag{2}$$

$$\prod_{i=1}^{e} x_i = e!p^{k-re}. \tag{3}$$

and

$$\prod_{i=1}^{e} (x_i - 1) = \frac{(n-1)(n-2)\ldots(n-e)}{q^e} (q-1)^e. \tag{4}$$

From (4) we deduce $p^{re} | (n-1)(n-2)\ldots(n-e)$. Thus one of the integers $n-1, n-2, \ldots, n-e$ is divisible by $p^{re-[e/p]-[e/p^2]-\ldots}$ and hence

$$n > p^{e(r-1/(p-1))} \geqq q^{e/2}. \tag{5}$$

Further, by (3), we see that either there are x_i and x_j such that $p | (x_i/x_j)$, or $p > e \geqq 3$ and so $x_1/x_e > 2$. Therefore in any case $x_1 \geqq 2x_e$ and hence $x_1 x_e \leqq 2(x_1 + x_e)^2/9$. Thus we observe that, by virtue of (1), (3), the arithmetic-geometric mean inequality and (2), we have

$$(q-1)^e q^{-e} n(n-1)\ldots(n-e+1) < e!q^{-e}p^k$$

$$= x_1 x_2 \ldots x_e \leqq \frac{8}{9}\left(\frac{x_1 + x_e}{2}\right)^2 x_2 \ldots x_{e-1}$$

$$\leqq \frac{8}{9}\left(\frac{x_1 + x_e}{2}\right)^2 \left(\frac{x_2 + \ldots x_{e-1}}{e-2}\right)^{e-2} \leqq \frac{8}{9}\left(\frac{x_1 + \ldots + x_e}{e}\right)^e$$

$$\leqq \frac{8}{9}(q-1)^e q^{-e} n^e.$$

Consequently

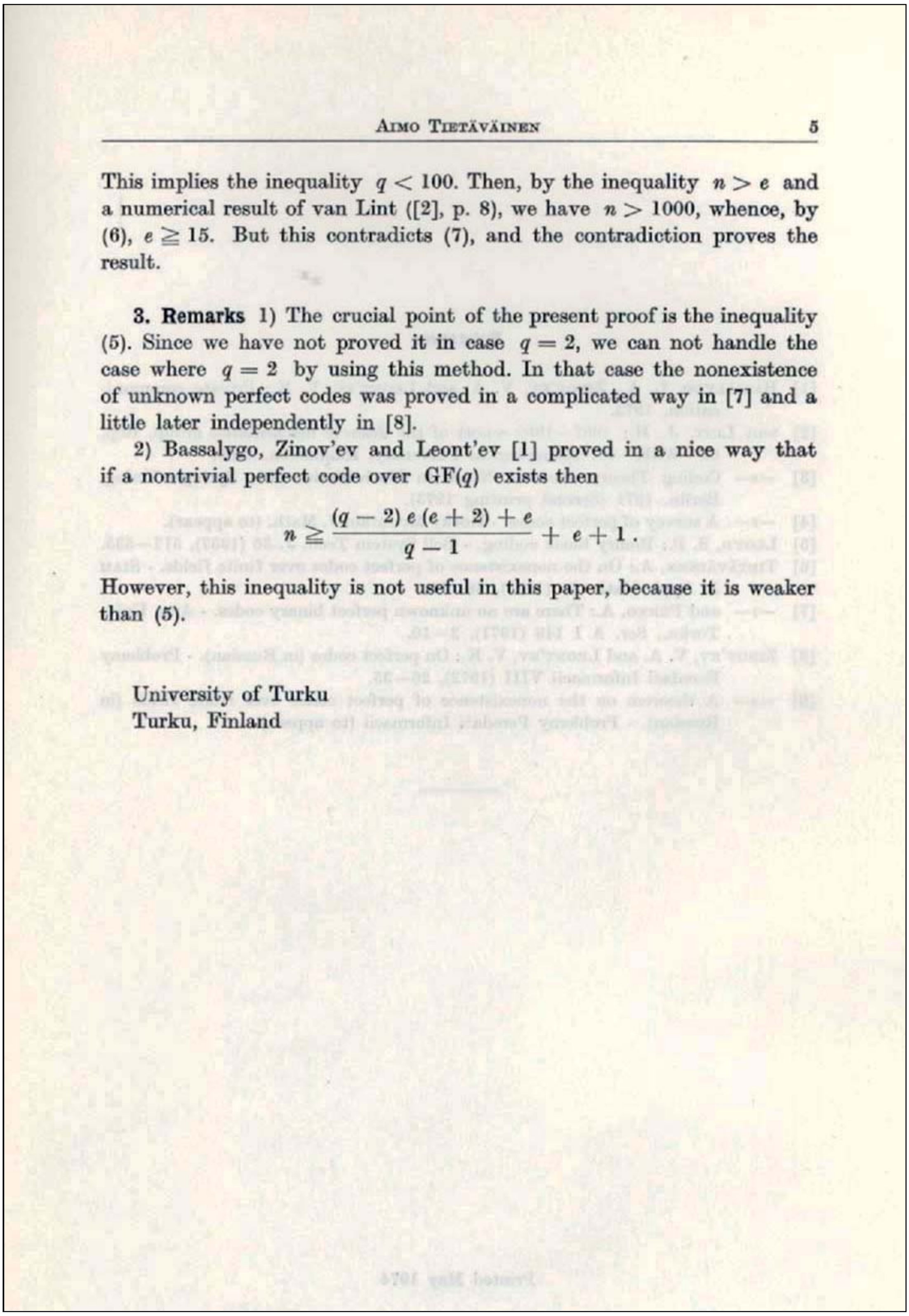
$$1 - \frac{e(e-1)}{2n} < \frac{8}{9}$$

and so

$$n < 5e^2. \tag{6}$$

Combining (6) with (5), we obtain

$$q^e < 25e^4. \tag{7}$$



This implies the inequality $q < 100$. Then, by the inequality $n > e$ and a numerical result of van Lint ([2], p. 8), we have $n > 1000$, whence, by (6), $e \geqq 15$. But this contradicts (7), and the contradiction proves the result.

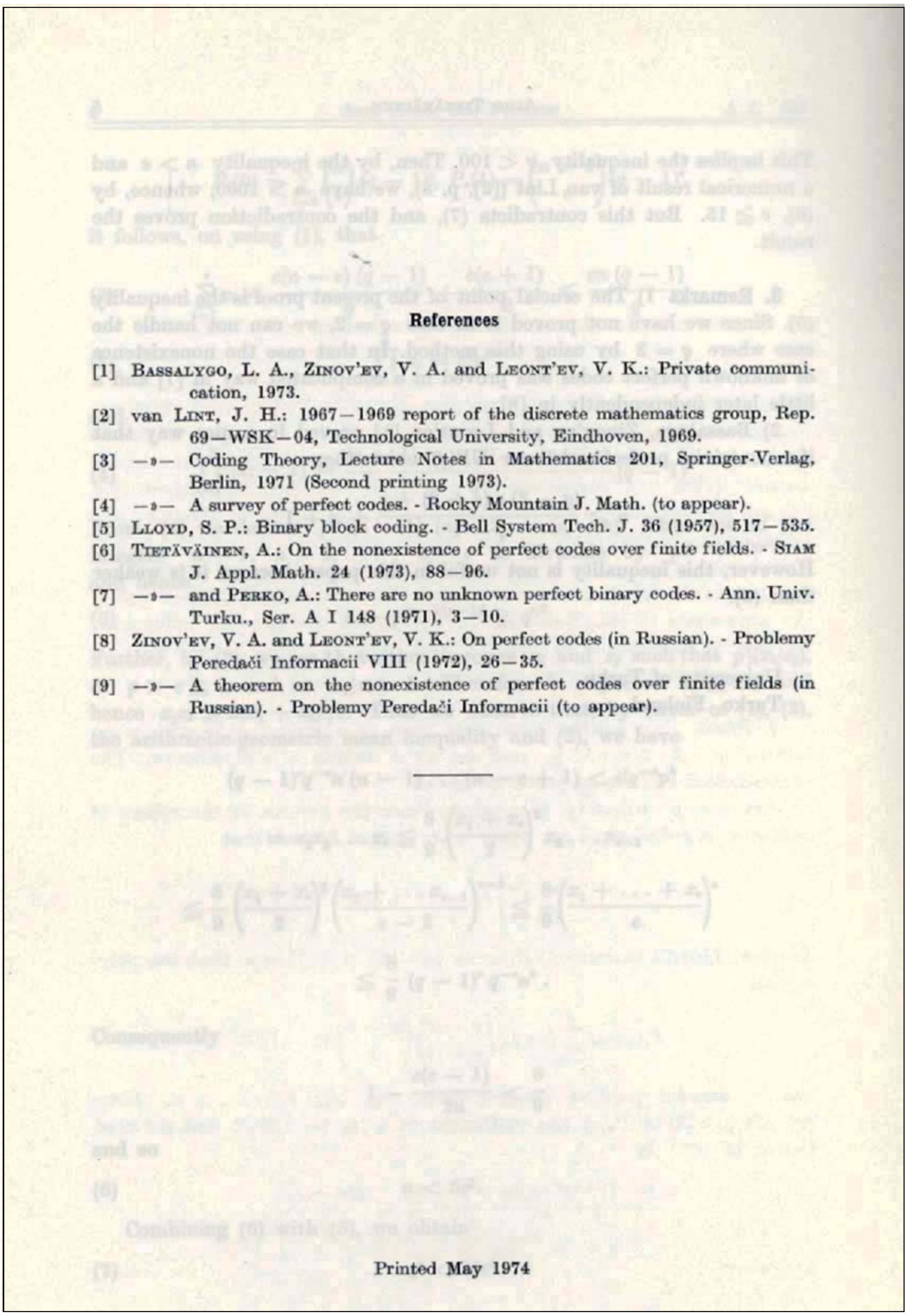
3. Remarks 1) The crucial point of the present proof is the inequality (5). Since we have not proved it in case $q = 2$, we can not handle the case where $q = 2$ by using this method. In that case the nonexistence of unknown perfect codes was proved in a complicated way in [7] and a little later independently in [8].

2) Bassalygo, Zinov'ev and Leont'ev [1] proved in a nice way that if a nontrivial perfect code over GF(q) exists then

$$n \geqq \frac{(q-2)\, e\, (e+2) + e}{q-1} + e + 1 \, .$$

However, this inequality is not useful in this paper, because it is weaker than (5).

University of Turku
Turku, Finland



References

[1] Bassalygo, L. A., Zinov'ev, V. A. and Leont'ev, V. K.: Private communication, 1973.

[2] van Lint, J. H.: 1967—1969 report of the discrete mathematics group, Rep. 69—WSK—04, Technological University, Eindhoven, 1969.

[3] —»— Coding Theory, Lecture Notes in Mathematics 201, Springer-Verlag, Berlin, 1971 (Second printing 1973).

[4] —»— A survey of perfect codes. - Rocky Mountain J. Math. (to appear).

[5] Lloyd, S. P.: Binary block coding. - Bell System Tech. J. 36 (1957), 517—535.

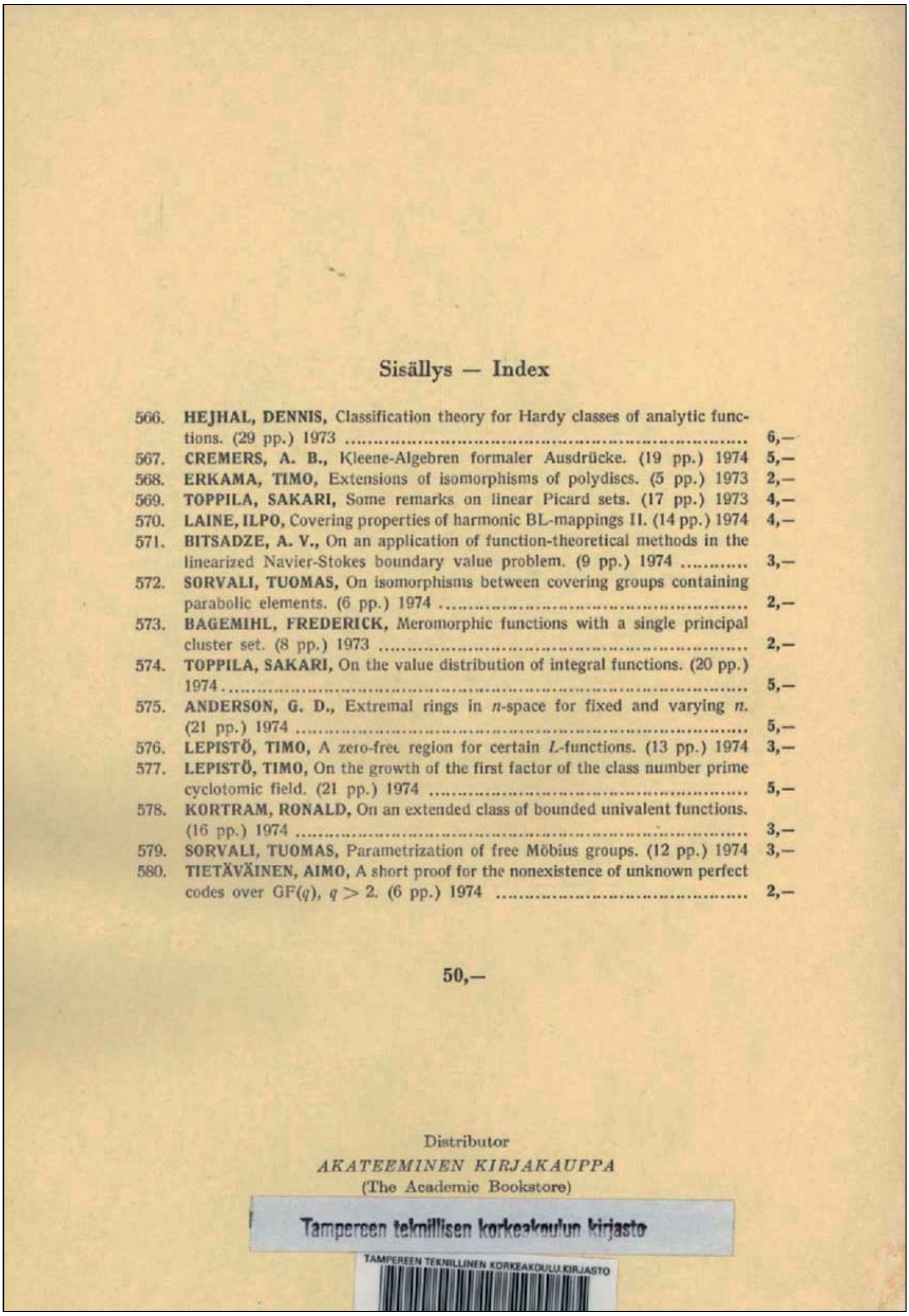
[6] Tietäväinen, A.: On the nonexistence of perfect codes over finite fields. - Siam J. Appl. Math. 24 (1973), 88—96.

[7] —»— and Perko, A.: There are no unknown perfect binary codes. - Ann. Univ. Turku., Ser. A I 148 (1971), 3—10.

[8] Zinov'ev, V. A. and Leont'ev, V. K.: On perfect codes (in Russian). - Problemy Peredači Informacii VIII (1972), 26—35.

[9] —»— A theorem on the nonexistence of perfect codes over finite fields (in Russian). - Problemy Peredači Informacii (to appear).

Printed May 1974



Sisällys — Index

50,—

Distributor
AKATEEMINEN KIRJAKAUPPA
(The Academic Bookstore)

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O. Box 527
FIN-33101 Tampere, Finland

ISBN 978-952-15-2817-0
ISSN 1456-2774