

Krishnaswami Alladi  
John R. Klauder  
Calyampudi R. Rao

# The Legacy of Alladi Ramakrishnan in the Mathematical Sciences



ALLADI RAMAKRISHNAN (1923-2008): Picture taken in Madras, India in 1958 upon his return from the Institute for Advanced Study in Princeton when he was full of visions to create a similar center for advanced learning in Madras



# The Legacy of Alladi Ramakrishnan in the Mathematical Sciences





Krishnaswami Alladi • John R. Klauder  
Calyampudi R. Rao  
Editors

# The Legacy of Alladi Ramakrishnan in the Mathematical Sciences

 Springer

*Editors*

Krishnaswami Alladi  
Department of Mathematics  
University of Florida  
Gainesville, FL 32611, USA  
alladik@ufl.edu

Calyampudi R. Rao  
Department of Statistics  
The Pennsylvania State University  
University Park, PA 16802, USA  
crr1@psu.edu

John R. Klauder  
Department of Physics  
University of Florida  
Gainesville, FL 32611, USA  
klauder@phys.ufl.edu

ISBN 978-1-4419-6262-1 e-ISBN 978-1-4419-6263-8  
DOI 10.1007/978-1-4419-6263-8  
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010932321

Mathematics Subject Classification (2010): 11-XX, 81-XX, 83-XX, 60-XX

© Springer Science+Business Media, LLC 2010

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

*Cover illustration:* Photo of Alladi Ramakrishnan courtesy of Krishnaswami Alladi.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Preface

Alladi Ramakrishnan (1923–2008) was an eminent scientist who had a wide range of research interests in theoretical and mathematical physics. Professor Ramakrishnan made significant contributions to probability and statistics, elementary particle physics, cosmic rays and astrophysics, matrix theory, and the special theory of relativity. Ramakrishnan believed strongly that in addition to doing fundamental research, one must contribute to the advancement of the profession. Inspired by his visit to the Institute for Advanced Study in Princeton in 1957–1958, he returned to Madras and began the *Theoretical Physics Seminar* at his family home *Ekamra Nivas*. These seminars were ultimately responsible for the creation of MATSCIENCE, The Institute of Mathematical Sciences in 1962. This institute, of which he was the Director for its first 21 years, has grown steadily in size and stature, and is his monumental contribution to the profession. In a distinguished scientific life that has spanned more than five decades, Professor Ramakrishnan has come into close contact with, and was influenced by, several eminent mathematicians and physicists, and has moulded the careers of his several students and young researchers. This volume, which is a tribute to his great legacy, not only deals with his significant contributions to research and the profession, but also contains a fine collection of research and survey papers by leading physicists and mathematicians that cover a broad range of areas in the mathematical sciences.

The first part of this volume is about Professor Alladi Ramakrishnan and his contributions. The book begins with an article entitled “Contributions of Alladi Ramakrishnan to the Mathematical Sciences” in which the remarkable career and contributions of Ramakrishnan are described by his son Krishnaswami Alladi who was very close to his father and accompanied Professor Ramakrishnan regularly on his worldwide scientific trips. Included in Krishna’s article is a description of Ramakrishnan’s visit to the Institute for Advanced Study in Princeton in 1957–1958, and the subsequent exciting series of events in Madras which led to the creation of MATSCIENCE. This is immediately followed by an article on Alladi Ramakrishnan’s (now famous) Theoretical Physics Seminar. The list of eminent speakers at the seminar and the list of students who attended are provided.

The creation of MATSCIENCE was heralded with enthusiasm by scientists around the world. Telegrams and letters came pouring in for the inauguration. A few sample congratulatory telegrams from world famous physicists and mathematicians

are reproduced and brief comments are made about the person sending the telegram or letter and Ramakrishnan's association with that scientist.

The creation of MATSCIENCE was like a dream come true! The next item in this volume is Alladi Ramakrishnan's speech *The Miracle has Happened* which he gave at the inauguration of MATSCIENCE on 3 January 1962. In his inimitable style, Professor Ramakrishnan describes the series of incredible events, each as improbable as the other, that took place in rapid succession. Ramakrishnan was charged with emotion as he gave this most inspiring speech, which is actually a model in English diction!

Professor Ramakrishnan believed in maintaining close contact with the international scientific community. Just as he invited eminent scientists regularly to the Theoretical Physics Seminar and to MATSCIENCE, he traveled across the globe annually to disseminate the work of his group. In these travels, he made new contacts and that invigorated not only his own research, but also the visiting scientists program at MATSCIENCE. Thus, we have included a brief description of some of Ramakrishnan's significant overseas trips.

Part I of the volume concludes with the list of scientific publications of Professor Alladi Ramakrishnan, and the list of his Ph.D. students.

Parts II–IV of the volume constitute research and survey papers by physicists and mathematicians who got to know Professor Alladi Ramakrishnan very well over the years. The range of topics covered by these papers is broad as were the research interests of Professor Ramakrishnan. The papers have been grouped as follows – Part II: pure mathematics, Part III: probability and statistics, and Part IV: applied mathematics and theoretical physics. Some of Professor Ramakrishnan's former Ph.D. students and grand students have contributed papers included in Parts III and IV. Within each of the parts of the volume, the papers are listed alphabetically by author's names.

## Part II: Pure Mathematics

Shreeram Abhyankar, a leading algebraic geometer, admired Ramakrishnan not only for his research, but also for creating an institute for advanced study in the mathematical sciences in India. Abhyankar, who takes great pride in India's intellectual past, himself created and directed a mathematics institute in Pune, Maharashtra, called the *Bhaskaracharya Prathistama*. In a massive paper jointly dedicated to Professor Ramakrishnan and his father, Abhyankar discusses extensions of his important work of 1967 on "gap invariance" with the intention of applying these ideas to a famous unsolved problem in algebraic geometry, namely, the *Jacobian Conjecture*.

Alladi Ramakrishnan was very much interested in using combinatorics to provide elegant proofs of identities and to use combinatorial insight to obtain generalizations and extensions. In this spirit, Krishna Alladi studies partitions into nonrepeating odd parts in a novel combinatorial way using 2-modular Ferrers graphs and their under-

lying Durfee squares to provide a unified treatment of several important identities in the theory of partitions and  $q$ -series.

Professor Ramakrishnan was fascinated by the symmetries and properties of the Pascal triangle which he often used to explain various enumeration problems arising in the theory of probability. Catalan numbers, which are defined using the middle binomial coefficients of the Pascal triangle, arise in a variety of settings. In a charming article, George Andrews investigates a  $q$ -analogue of the Catalan numbers and establishes several identities for these  $q$ -analogues, from which classical identities for the Catalan numbers fall out as special cases.

Another lifelong passion for Ramakrishnan was Euclidean geometry which he used to explain difficult concepts in the theory of special relativity. Richard Askey's paper deals with the beautiful theorem of Ptolemy on cyclic quadrilaterals and the extension of this result by the Indian mathematician Brahmagupta.

Alladi Ramakrishnan was also very proud of India's cultural and intellectual heritage. Naturally he was a great admirer of Ramanujan. In a joint paper with his student Atul Dixit, Bruce Berndt, one of the greatest authorities on Ramanujan's work, discusses a transformation formula of Ramanujan and how this leads to transformations involving the Gamma and Riemann zeta functions. This transformation formula of Ramanujan may be found in the book "Ramanujan's Lost Notebook and other unpublished papers" that was released during the Ramanujan Centennial in 1987. But this particular transformation formula is not in Ramanujan's lost notebook discovered by George Andrews in 1976 at the Wren Library in Cambridge University, but is in the "loose papers" that were located in Oxford University Library.

The area of quadratic forms has witnessed dramatic progress in the last few years including the resolution of a problem on universal quadratic forms stemming from Ramanujan. Alexander Berkovich and William Jagy show how certain modular identities of degree 3 discovered by Ramanujan can be used to establish some very appealing positivity results for some integral ternary quadratic forms.

Asking questions of an additive nature for integers defined multiplicatively leads to very intriguing problems. Jean-Marc Deshouillers and Florian Luca, leading authorities in additive number theory, discuss the frequency of integers for which  $n!$  is a sum of three squares, and show that the density of such integers is at least  $7/8$ . This result is extremely interesting in the light of the classical theorem of Lagrange which asserts that every positive integer is a sum of (at most) four squares, and the simple observation that integers of the form  $8k + 7$  cannot be represented as a sum of three squares.

Alladi Ramakrishnan was a great admirer of Euler for his many fundamental contributions. He once wrote an article on the charms of Euler's  $e$ . So it is only appropriate that there is a paper in this volume emphasizing Euler's work. Dominique Foata's paper "Eulerian polynomials: from Euler's time to the present" provides a beautiful survey of the topic. Foata starts with Euler's memoir of 1755 to find out Euler's motivation to study these polynomials. He then describes how these polynomials emerged in a  $q$ -generalized form in the work of Carlitz in the twentieth century and describes the underlying combinatorics. The contents of this paper were

delivered by Professor Foata in the Tenth Ulam Colloquium at the University of Florida in February 2008 and Professor Alladi Ramakrishnan attended that lecture.

The interaction between number theory and physics has attracted a lot of attention in recent years. Continuing his earlier investigations on connections between the Epstein zeta function and crystal symmetries, Shigeru Kanemitsu in joint work with Haruo Tsukada discusses several interesting examples, showing how crystal symmetry may be understood via zeta symmetry.

One of the main conjectures in the theory of linear forms is due to Minkowski on products of linear forms. Minkowski's conjecture has been proved for six dimensions or less, but the general result is still unproven. In his paper, Raghavan treats a modified problem in a novel fashion and obtains similar results to Minkowski's conjecture.

The penultimate paper of Part II is the seminal work of Peter Sin and John Thompson on the divisor matrix, Dirichlet series, and  $SL(2, Z)$ . Although divisors of integers have been studied since antiquity, no one has done a systematic study of the infinite upper triangular matrix  $[a_{i,j}]$ , where  $a_{i,j} = 1$  if  $i$  divides  $j$  and 0 otherwise. Thompson and Sin explore connections between this matrix, Dirichlet series, and  $SL(2, Z)$ . The subject matter of this paper was Professor Thompson's talk in Oslo in May 2008 after he received the Abel Prize. We are honored that this fundamental paper is included in this volume.

The final article in Part II is a letter by Michel Waldschmidt in which he proves a conjecture of Alladi Ramakrishnan on circulants. Professor Ramakrishnan was intrigued by the Lorentz Transformation in Special Relativity and provided new and elegant derivations of it. He wrote a paper "Pythagoras to Lorentz via Fermat" in which instead of considering the Fermat equation as the generalization of the Pythagorean equation, he studied an  $n$ -dimensional circulant generalization of the Pythagorean equation. Alladi Ramakrishnan connected this to the Lorentz transformation and determined its rational solutions. In this context, he made a conjecture regarding circulants and the proof of this conjecture is provided by Michel Waldschmidt.

### Part III: Probability and Statistics

Alladi Ramakrishnan did fundamental work in the theory of probability. Thus, it is appropriate that this volume contains excellent papers in probability and statistics.

Alladi Ramakrishnan, along with Homi Bhabha, made pioneering contributions to the theory of nuclear cascades by the use of stochastic processes. The opening paper of Part II by Krishna Athreya deals with the Galton–Watson branching processes and the associated branching random walk. The limiting behavior of the spatial distribution of points in certain point processes is investigated and an application to the photon–electron cascade is described.

The next paper by Malay Ghosh, Kwok Pui Choi, and Jialiang Li provides a smooth treatment of the logistic distribution without the use of contour integration. The authors show how to calculate the moments, the moment generating function, and the characteristic function.

The paper by C.R. Rao is a comprehensive review of entropy and cross-entropy. He discusses their characterizations and indicates possible applications. Entropy has been used in characterizing probability distributions in theoretical physics to which Professor Ramakrishnan has made fundamental contributions. Entropy has been used as a measure of diversity in environmental studies. Cross-entropy has emerged as a useful tool in solving stochastic and nonstochastic optimization problems.

The father and son team of Jayaram Sethuraman and Sunder Sethuraman discuss connections between Bernoulli strings and random permutations. In this regard, they point out very elegantly the connection between marked Poisson processes and Bernoulli strings.

Professor Ramakrishnan's grand student P.R. Vittal, S. Jaisankar, and V. Muralidhar investigate storage models. Storage theory has received considerable attention, and two of the leading contributors to this field are Joe Gani and Pap Moran, contemporaries of Alladi Ramakrishnan. Vittal, Jaisankar, and Muralidhar discuss storage problems for a class of one-dimensional master equations with separable kernels.

The problem of testing equality of survival distributions has received considerable attention. In the final paper of Part III, S.S. Wu, P.V. Rao, and Aparna Raychaudhry address this problem on the basis of paired censored survival data. They utilize test statistics that consist of linear combinations of two appropriately chosen statistics. In addition, they present a method for estimating optimal weights for such linear combinations.

## **Part IV: Theoretical Physics and Applied Mathematics**

Professor Alladi Ramakrishnan worked and directed students in several areas of theoretical physics and applied mathematics and the breadth of his interests is reflected in the topics covered by the authors of this section.

Imaging science has become one of the most active areas of research owing to applications ranging from determining the size of tumors to detection of tanks under foliage. Yunmei Chen, an authority in imaging science, and her student Xiaojing Ye present a novel variational model for inverse consistent deformable image registration. Their model is formulated as an energy minimization model and experimental results indicate the efficiency of their model.

Alladi Ramakrishnan's former Ph.D. student V. Devanathan (who later became the head of the nuclear physics department at the University of Madras) discusses a statistical model for the quark structure of the nucleon in a joint paper with his Ph.D. student S. Karthiyayini. Their paper contains a good description of both the static



and dynamic properties of the nucleon. A thermodynamic bag model is proposed to obtain realistic distribution functions that correctly yield the nucleon structure functions.

Alladi Ramakrishnan produced a technique called the  $\sigma$ -operation to construct the  $4 \times 4$  Dirac matrices from the  $2 \times 2$  anticommuting Pauli matrices. This led him to study the more general  $\omega$ -commutation ( $\omega$  is a root of unity) and the hierarchy of matrices satisfying the  $\omega$ -commutation. That was the evolution of Ramakrishnan's  $L$ -matrix theory which he pursued in depth by himself and with his Ph.D. students. R. Jagannathan, a former Ph.D. student of Alladi Ramakrishnan who later became a professor at MATSCIENCE, provides a nice review of generalized Clifford algebras and their applications to physics. In doing so, he discusses various ramifications of the work of Alladi Ramakrishnan and his group and the extensions that he himself has obtained. The fact that generalized Clifford algebras are so pervasive is well brought out in this paper.

Finding  $q$ -analogs of classical functions and identities has proved to be extremely fruitful because the scope of applications is considerably broadened with the introduction of  $q$ -analogs. R. Jagannathan and R. Sridhar, former student and grand student of Alladi Ramakrishnan who later became professors at MATSCIENCE, discuss a  $(p, q)$ -analog of the Rogers-Szegö polynomial and the  $(p, q)$  oscillator in physics. Just as the Rogers-Szegö polynomial is associated with the  $q$ -oscillator algebra, the authors show that the  $(p, q)$ -Rogers-Szegö polynomial is associated with the  $(p, q)$ -oscillator algebra.

John Klauder's paper "Rethinking renormalization" is a critical re-examination of the notion of renormalizability for several extreme types of quantum field theory. Normally, counter terms that are needed to remove divergences which arise in quantum field-theoretic calculations are introduced on a term-by-term basis after evaluation of suitable functional integrals. Klauder's approach differs by excising divergence-causing terms in the integrand of functional integrals, thereby eliminating divergences altogether. Ultimately, the aim is to apply this technique to the difficult task of quantizing the gravitational field.

The father and daughter team of A.N. Mitra and Gargi-Mitra Delmotte present a rather broad description of pattern formation in crystals and crystal-like structures under the influence of magnetic fields. The ability of these structures for self-replication, compartmental organization, and fractionalization serves as a basis for theoretical speculations that organic life may have originated utilizing similar mechanisms.

The final paper in the volume is by R. Parthasarathy, a grand student of Alladi Ramakrishnan who later became a professor at MATSCIENCE. This is a review of the work of the Ehrenfest theorem in Abelian and non-Abelian quantum field theories. The theorem is shown to be valid in appropriately defined physical subspaces.

This volume which contains a fine collection of papers covering a broad range of topics in number theory, algebra, geometry, probability, statistics, theoretical, nuclear, and mathematical physics, and certain topics in applied mathematics is a fitting tribute to the memory of Alladi Ramakrishnan who had such a profound influence on the scientific profession. The contributors include some of his students

and grand students who themselves went on to pursue highly successful academic careers, and eminent mathematicians, physicists, probabilists, and statisticians, who got know Professor Ramakrishnan and his work very well over the years. Our thanks to all the contributors of this volume. A special thanks to Professor Frank Garvan of the University of Florida who provided crucial help in assembling the TeX files of the papers for production. Felix Portnoy of Springer, New York, and Ejaz Ahmad in Chennai, India, did a fine job in typesetting the entire volume. Finally we wish to express our appreciation to Elizabeth Loew, Ann Kostant, Joachim Heinze and Hans Koelsch of Springer for their interest in producing this volume and their support throughout this venture.

University of Florida  
University of Florida  
The Pennsylvania State University

*Krishnaswami Alladi*  
*John Klauter*  
*C.R. Rao*



# Contents

<b>Preface</b> .....	vii
<b>Part I The Legacy of Alladi Ramakrishnan</b>	
<b>Contributions of Alladi Ramakrishnan to the Mathematical Sciences</b> .....	3
Krishnaswami Alladi	
<b>Alladi Ramakrishnan’s Theoretical Physics Seminar</b> .....	11
Krishnaswami Alladi	
<b>Telegrams Received for the MATSCIENCE Inauguration</b> .....	25
Krishnaswami Alladi	
<b>The Miracle has Happened</b> .....	67
Alladi Ramakrishnan	
<b>Overseas Trips of Alladi Ramakrishnan</b> .....	73
Krishnaswami Alladi	
<b>List of Publications of Alladi Ramakrishnan</b> .....	81
<b>List of PhD Students of Alladi Ramakrishnan</b> .....	89
<b>Part II Pure Mathematics</b>	
<b>Inversion and Invariance of Characteristic Terms: Part I</b> .....	93
Shreeram S. Abhyankar	
<b>Partitions with Non-Repeating Odd Parts and Q-Hypergeometric Identities</b> .....	169
Krishnaswami Alladi	

<b><i>q</i>-Catalan Identities</b> .....	183
George E. Andrews	
<b>Completing Brahmagupta's Extension of Ptolemy's Theorem</b> .....	191
Richard Askey	
<b>A Transformation Formula Involving the Gamma and Riemann Zeta Functions in Ramanujan's Lost Notebook</b> .....	199
Bruce C. Berndt and Atul Dixit	
<b>Ternary Quadratic Forms, Modular Equations, and Certain Positivity Conjectures</b> .....	211
Alexander Berkovich and William C. Jagy	
<b>How Often is <math>n!</math> a Sum of Three Squares?</b> .....	243
Jean-Marc Deshouillers and Florian Luca	
<b>Eulerian Polynomials: From Euler's Time to the Present</b> .....	253
Dominique Foata	
<b>Crystal Symmetry Viewed as Zeta Symmetry II</b> .....	275
Shigeru Kanemitsu and Haruo Tsukada	
<b>Positive Homogeneous Minima for a System of Linear Forms</b> .....	293
Srinivasacharya Raghavan	
<b>The Divisor Matrix, Dirichlet Series, and <math>SL(2, \mathbb{Z})</math></b> .....	299
Peter Sin and John G. Thompson	
<b>Proof of a Conjecture of Alladi Ramakrishnan on Circulants</b> .....	329
Michel Waldschmidt	
<b>Part III Probability and Statistics</b>	
<b>Branching Random Walks</b> .....	337
K.B. Athreya	
<b>A Commentary on the Logistic Distribution</b> .....	351
Malay Ghosh, Kwok Pui Choi, and Jialiang Li	
<b>Entropy and Cross Entropy: Characterizations and Applications</b> .....	359
C.R. Rao	

**Optimal Weights for a Class of Rank Tests for Censored Bivariate Data** .....369  
 Samuel S. Wu, P.V. Rao, and Aparna Raychaudhuri

**Connections Between Bernoulli Strings and Random Permutations**.....389  
 Jayaram Sethuraman and Sunder Sethuraman

**Storage Models for a Class of Master Equations with Separable Kernels** .....401  
 P. R. Vittal, S. Jayasankar, and V. Muralidhar

**Part IV Theoretical Physics and Applied Mathematics**

**Inverse Consistent Deformable Image Registration** .....419  
 Yunmei Chen and Xiaojing Ye

**A Statistical Model for the Quark Structure of the Nucleon** .....441  
 V. Devanathan and S. Karthiyayini

**On Generalized Clifford Algebras and their Physical Applications**.....465  
 Ramaswamy Jagannathan

**$(p, q)$ -Rogers-Szegö Polynomial and the  $(p, q)$ -Oscillator** .....491  
 Ramaswamy Jagannathan and Raghavendra Sridhar

**Rethinking Renormalization** .....503  
 John R. Klauder

**Magnetism, FeS Colloids, and Origins of Life** .....529  
 Gargi Mitra-Delmotte and A.N. Mitra

**The Ehrenfest Theorem in Quantum Field Theory** .....565  
 Ragavachariar Parthasarathy



**Part I**  
**The Legacy of Alladi Ramakrishnan**



# Contributions of Alladi Ramakrishnan to the Mathematical Sciences

Krishnaswami Alladi

Professor Alladi Ramakrishnan, my father, belonged to a small eminent group of Indian scientists who made fundamental contributions to several fields of study and sustained a high level of productivity over a significant period of time. If among this select versatile group of researchers we seek those who also have made monumental contributions to the profession by creating leading institutions of advanced research, then we are down to a mere handful such as Professors Raman, Bhabha, Mahalanobis, Ramakrishnan, and a few more. In an illustrious scientific career that began in 1947, Professor Ramakrishnan has published about 150 influential research papers in leading journals on topics ranging over Stochastic Processes, Elementary Particle Physics, Matrix Algebra, and the Special Theory of Relativity, has guided 24 PhD students, lectured on his research at over 200 institutions of higher learning the world over and at numerous international conferences, and created MATSCIENCE, The Institute of Mathematical Sciences in Madras. It is amazing that even after his retirement, and indeed until the very end, his passion for science and his spirit of enquiry remained unabated. Here I shall briefly describe some of his significant contributions including his most recent ones, and the circumstances that led to them.

**Early life and career choice:** Right from his school days, my father demonstrated his originality both in mathematics and physics. In Loyola College, Madras, he was awarded a special prize by his mathematics teacher Adivarahan, who was much impressed by my father's unusual originality in classical geometry. I have witnessed this facility with geometry in the past few years, when my father applied simple but ingenious geometrical arguments to explain and unravel new features of difficult concepts in Special Relativity.

My paternal grandfather Sir Alladi Krishnaswami Iyer was one of the greatest lawyers of India during the first half of the twentieth century. He played a crucial role in drafting the Constitution of India. Naturally, to young Ramakrishnan, his father was a great influence. Indeed my father enrolled in law, passed the exams

---

K. Alladi  
Department of Mathematics, University of Florida, Gainesville, FL 32611, USA  
e-mail: [alladik@ufl.edu](mailto:alladik@ufl.edu)

in flying colors, and secured a Gold Medal in Hindu Law. He even assisted my grandfather by taking notes to his dictation concerning the Indian Constitution. My father's early contacts through my grandfather were men of the calibre of Dr. S. Radhakrishnan, the great philosopher statesman who became the President of India, C. Rajagopalachari, and others. In spite of all this exposure and contact with lawyers and statesmen, my father had this inner desire to pursue science as a career. Although my grandfather was immersed in the field of law, he used to say "compare the nationalism of politics to the internationalism of science," a sentence that profoundly influenced my father to change his career. Actually, the desire of my father to take to science as a career was kindled in 1943 when he heard a magnificent lecture on Meson Theory at the Presidency College, Madras, by Professor Homi Bhabha who had just returned to India from England as one of the youngest Fellows of the Royal Society (FRS). However, it was only 4 years later after a brief stint in law that my father decided to eschew a lucrative legal career and take science as a profession. It was at this instance that my paternal grandmother Lady Venkalakshmi convinced my grandfather to let her young son pursue his dreams and goals.

**Work with Bhabha:** In 1947, my father joined what was at that time the fledgeling Tata Institute of Fundamental Research that functioned under Bhabha's direct supervision in Kenilworth, Bhabha's aunt's home. Thus my father was one of the first members of the Tata Institute and worked closely in contact with Bhabha himself. My father always stressed that the greatest gift a teacher can give a research student is a good problem, and in this sense he was very fortunate that Bhabha introduced him to Cascade Theory and the Fluctuation Problem of Cosmic Radiation. The study of this problem required the probabilistic analysis of the distribution of a discrete number of particles in continuous energy space. My father soon realized that it was possible to attack this problem directly by noting that the contribution to the density comes from the probability of a single particle in an infinitesimal interval, which is proportional to the length of that interval, the coefficient representing the density. He named the correlation densities as *Product Densities*, a name that is still in vogue today. Bhabha, who was a master of limiting processes, had also an idea of how to solve this problem, but by a longer method.

**Product Densities and related work:** In August 1949, my father left the Tata Institute and sailed to England with my mother to complete his PhD under the direction of Professor M. S. Bartlett at the University of Manchester. Professor Bartlett consulted his distinguished friend Professor D. G. Kendall, who was then at Magdalene College at Oxford University. Professor Kendall not only confirmed the correctness of my father's work but also approved the name product densities. Kendall had previously arrived at such functions up to the second order in his pioneering studies on population growth and called them cumulant densities. In my father's work on product densities, the more general  $n$ -th order functions were considered. Thus, within two months of arrival in England, my father had completed his work for the PhD. But he had to stay for two years there to complete his residency requirements. My father's PhD work on product densities appeared in the Proceedings of the Cambridge Philosophical Society (1950), and Bhabha's alternate approach appeared around the same time in the Proceedings of the Royal Society.

Some years later, my father had the opportunity to give a talk on applications of stochastic processes to cascade theory at the Max Planck Institute in Göttingen. The German Nobel Laureate Werner Heisenberg who heard this lecture made very complimentary comments. On the basis of Heisenberg's comments, Professor S. Flugge of Springer Verlag invited my father to write a comprehensive article on stochastic processes with emphasis on product densities. This authoritative article, the first of its kind on these topics, appeared in the *Handbuch der Physik* (Springer). It had a significant influence and resulted in a flood of papers in the area, most notably by Professor S. K. Srinivasan. A book by A.T. Barucha Reid on Markov Processes makes ample references to product densities and the work of Bhabha-Ramakrishnan. The method of product densities became very well-known and is considered by many to be perhaps my father's most significant contribution.

In the 1950s, my father worked on the problem of the Fluctuating Density Field that came up in studies of the Milky Way by the great Indian astrophysicist Subramaniam Chandrasekhar. My father wrote a series of eight papers on this subject. Chandrasekhar was so impressed that he communicated all of them to the *Astrophysical Journal*.

Another notable contribution was my father's work on *Inverse Probability* in Stochastic Processes leading to the concept of the origin of a stochastic process. This paper was presented to the Indian Academy of Sciences in 1955. It was on the basis of this presentation that Sir. C. V. Raman had my father elected immediately as a Fellow of the Indian Academy of Sciences.

The work on inverse probability had other implications. It led my father to interpret the Feynman observation of a negative energy electron travelling back in time as actually tracing back in the inverse probability sense. This yielded a simple proof of the equivalence of the Feynman and the field theoretic formulation by splitting the Feynman propagator into positive and negative energy parts. The first person to establish this equivalence rigorously was Dyson, but only a few have really understood Dyson's deep and difficult derivation. My father's paper on this topic appeared in the *Journal of Mathematical Analysis and Applications* (1967). In addition, at the invitation of Professor Heitler, he published his work on stochastic processes and the Feynman propagator as a book entitled *Elementary particles and cosmic rays* published by the Pergamon Press (1962).

**Visit to the Institute for Advanced Study:** The year 1957–1958 was another turning point in my father's career when he visited the Institute for Advanced Study in Princeton at the invitation of its Director, Robert Oppenheimer. At the Institute, my father had the opportunity to listen to the lectures of, and discuss with, the leading young physicists of that generation, like T. D. Lee and C. N. Yang, who soon afterwards won the Nobel Prize in Physics. My father returned to India filled with the desire to induct talented students into theoretical physics and expose them to the latest advances in this field.

**The Theoretical Physics Seminar:** Not satisfied with the curriculum at the University of Madras where my father was a professor, he gave lectures on quantum mechanics and other advanced topics at our family home Ekamra Nivas during the period 1958–1961 and named this the *Theoretical Physics Seminar*. As the

daughter of a professor of mathematics Dr. H. Subramani Iyer, and as one who had accompanied my father to England and to Princeton, my mother Mrs. Lalitha Ramakrishnan had a full understanding of the significance of such efforts by my father. With enthusiasm she hosted the eager students who gathered in Ekamra Nivas for the Theoretical Physics Seminar, and many eminent scientists who lectured at our home. Among the luminaries who addressed the students at the Theoretical Physics Seminar were Nobel Laureate Donald Glaser, and Professors Murray Gell Mann and Abdus Salam, both of whom won Nobel Prizes later.

**Creation of MATSCIENCE:** In 1960, Nobel Laureate Professor Niels Bohr visited India as the guest of Prime Minister Jawaharlal Nehru. When Bohr came to Madras, there was only one group of students who could understand his lectures, namely, those trained by my father. Bohr spent a leisurely evening at Ekamra Nivas discussing with my father and his students. When Bohr returned to Delhi, he was asked what his impressions about science in India were. Professor Bohr said that two things impressed him most – the massive Tata Institute of Fundamental Research in Bombay, and the small group of students trained by Alladi Ramakrishnan in Madras! This statement by Bohr was flashed in the newspapers like *The Hindu* and sparked Nehru's interest to contact my father. Mr. C. Subramaniam arranged for a meeting at the Governor's Residence, Raj Bhavan, in Madras, between Nehru and my father, in which the students of the Theoretical Physics Seminar were introduced to the Prime Minister. At this meeting, Nehru asked my father what he wanted. Here was the Prime Minister of India asking you what you want! At such an instance, you do not ask for anything meagre. So my father asked for an institute for advanced fundamental research in the mathematical sciences like the Institute for Advanced Study in Princeton. The rest is history. With the recommendation of Niels Bohr, the support of C. Subramaniam, and the benevolence of Jawaharlal Nehru, MATSCIENCE, The Institute of Mathematical Sciences was created in 1962 with my father as the Director. Subramaniam Chandrasekhar was invited to inaugurate the institute. I remember sitting in the English Lecture Hall of the Presidency College, Madras that day and listening to a magnificent lecture by my father – perhaps the finest he has delivered in his life.

My father served as the Director of MATSCIENCE for 21 years until his retirement in 1983. He conceived it in his family home, nurtured it in its infancy, and saw it grow in size and stature. During his tenure as Director, hundreds of eminent mathematicians and physicists visited the Institute, including Nobel Laureates Hans Bethe, Hans Jensen, Linus Pauling and John Bardeen, Fields Medallists Laurent Schwarz and Rene Thom, the mathematical giant Marshall Stone, the eminent Indian statistician C. R. Rao, the Ramanujan expert and partition authority George Andrews, and the legendary mathematician Paul Erdős. It was also during these 21 years that he travelled widely, lecturing at about 200 centers of learning all over the world. My mother and I accompanied him on these trips. The constant contact with outstanding academicians during these foreign tours as well as those who visited MATSCIENCE, and the experience of visiting several great centers of learning, made a deep impression on me. I thought of nothing else but an academic career and was naturally led into it.

**Work in quantum mechanics:** From the early sixties on, my father's research shifted to theoretical and particle physics. His most significant work in this area was the prescription he gave to make the transition from Pauli to Dirac Matrices. He called this the  $\sigma$ -operation. This work was part of a more comprehensive study of  $\omega$ -commutation relations among matrices, generalizing the anticommuting property of the Pauli matrices. All this occupied him and his students for about a decade, when a series of about 50 papers were published under the banner of *L-matrix theory*, mostly in the Journal of Mathematical Analysis and Applications. Subsequently, these papers were also published collectively in the form of a book entitled *L-matrix theory or the grammar of Dirac matrices* by Tata McGraw Hill (1972) and released by His Excellency V. V. Giri, the President of India.

**PhD students:** My father's work on product densities and L-matrix theory provided food for thought for talented students who worked under his guidance. Over a period of a quarter century (1958–1983), he produced about 24 PhD students. He provided opportunities for all of them to go abroad to visit centers of learning and to participate in international conferences. He was extremely generous in providing ample leave for them to travel, much to the envy of scientists in other institutions in India where leave and travel rules were much stricter. My father believed that young researchers would profit by contact with experts at institutions worldwide, and he therefore provided opportunities for them to travel. There was of course the risk of losing some of these talented students to other institutions. But he was convinced that science is an international enterprise and therefore did not want the students to feel stifled due to lack of travel. Such large heartedness among senior administrators is hard to find. Some of the students who went abroad did not return but made successful careers in the United States. Some others joined the faculty at MATSCIENCE. Four students accepted positions at educational institutions in Madras and all four not only developed schools of research at their respective institutions, but also became heads of their departments. They were Professors P. M. Mathews at the Department of Theoretical Physics of the University of Madras, S. K. Srinivasan of the Department of Mathematics at IIT Madras, V. Devanathan of the Department of Nuclear Physics of the University of Madras, and A. Vijayakumar of the Mathematics Department of Anna University, Madras.

**Work in Special Relativity:** My father had a fascination for Special Relativity since his college days inspired by the book of Joos on Theoretical Physics that he read at the suggestion of Sir C. V. Raman. His first significant piece of work in this area were a series of papers on the theme *Einstein is a natural completion of Newton* that appeared in the Journal of Mathematical Analysis and Applications. In a paper entitled *Ramakrishnan's approach to the theory of relativity* that also appeared in same journal (1974), the famous analyst Norman Levinson of MIT rigorously established some of the postulates my father made in his papers.

My father was always intrigued why only Einstein received the credit for the theory of relativity when so much of the theory depended on the Lorentz Transformation. In his years after his retirement in 1983, he came back time and again to the Lorentz transformation, offering new and elegant derivations of it using simple but ingenious geometric arguments. He felt that although the Lorentz transformation

is over a hundred years old, it still bears a youthful countenance. His work on the Lorentz transformation reached a peak in his paper, ‘*A rod approach to the theory of relativity*’ in which he clarified the distinction between Space-like and Time-like intervals. This paper appeared in the Special Millennium Issue of the Journal of Mathematical Analysis and Applications in September 2000 in honor of its Founding Editor Professor Richard Bellman with whom my father had close scientific contact since 1956.

**Recent training of students:** Although my father retired in 1983, he continued to inspire and influence talented students. Especially in the last few years, several brilliant high school and undergraduate students have come to his home in Madras to learn from him. They have profited immensely by his instruction and encouragement because every one of them has come to the United States to pursue higher studies in order to take a career of research. I have met only one other person, the mathematical legend Paul Erdős, who had such a passion to meet talented young minds and encourage them to pursue mathematics. My father’s appetite for research and the desire to train students had not diminished with time.

**Intellectually active till the end:** After his retirement in 1983, my father visited me in Florida every year during the Spring term with my mother. During my ten year term as Chairman of the Mathematics Department at the University of Florida, I ran a vibrant visiting program. I was inspired as a young boy watching my father organize an outstanding visiting program at MATSCIENCE and indeed even before that with his Theoretical Physics Seminar. During his visits to Florida, my father never missed a single featured colloquium talk in the mathematics department. He enjoyed discussions with the distinguished speakers at the seminars on campus and more informally at parties at our home graciously hosted by my wife Mathura\*. He was in good health and spirits till the very end and attended scientific lectures on campus just a few weeks before he died. My father passed away peacefully at our home in Gainesville, Florida, on June 7, 2008, with his entire family by his side. In fact, just two hours before he died, he attended a dance program organized by Mathura, in which he especially enjoyed the final item presented by Mathura and my daughters Lalitha and Amritha.

I close with the dictum that dominated his life:

*The pursuit of science is at its best  
when it is a part of a way of life.*

This is the motto of MATSCIENCE and is inscribed at the entrance to the institute.

---

\*A selection of photographs of my father taken recently in Gainesville is included in this book as part of a list of overseas photos. These photographs show him in discussion with various eminent mathematicians at our home in Florida.

## Note

This is an updated version of an article by me under the same title that appeared in *“Point Processes and Product Densities”* (S. K. Srinivasan and A. Vijayakumar Eds.), Narosa, New Delhi (2003), pp. ix–xiv brought out in connection with a conference in Madras, India, on the same topic in honor of Professor Alladi Ramakrishnan for his 80-th birthday. Also, an abridged version of this article appeared as an obituary note by me in the newsletter of the Institute for Advanced Study. See

*“Alladi Ramakrishnan (1923–2008) – Institute visit inspired creation of the Institute of Mathematical Sciences in Madras”, The Institute Letter, Institute for Advanced Study, Princeton (Spring 2009), p. 7*

I was invited to write this obituary note by Professor Peter Goddard, Director of the Institute for Advanced Study, Princeton.

Finally, I should add that my father has given a detailed account of his life, including all significant scientific events, in his inimitable style in

*“The Alladi Diary”, Vol 1, East–West Books, Madras, India (2000)*

*“The Alladi Diary”, Vol 2, East–West Books, Madras, India (2003)*

# Alladi Ramakrishnan's Theoretical Physics Seminar

Krishnaswami Alladi

After completing his PhD at the University of Manchester, my father returned to India and joined the physics department at the University of Madras as a Reader in 1952. He was later promoted as Professor. He was developing *the theory of product densities* that he had initiated in his PhD thesis and studying applications of it by himself and with his students. He availed every possible opportunity to invite eminent scientists to the University of Madras and to our family home *Ekamra Nivas* and encouraged his students to listen to their lectures and engage in discussions with them. When my father visited the Institute for Advanced Study in Princeton in 1957–1958 at the invitation of its Director Robert Oppenheimer, he had the opportunity to listen to over one hundred seminars on theoretical physics by the leading researchers of that generation. My father returned to India filled with a desire to expose students to the latest developments in modern physics. Not satisfied with the curriculum at the Madras University, he gave advanced lectures in theoretical physics to students at *Ekamra Nivas*. Eager students gathered at the seminar to hear his lectures, and this was formally called *The Theoretical Physics Seminar*. He invited eminent scientists to lecture in this seminar. My mother Mrs. Lalitha Ramakrishnan graciously hosted the foreign speakers and the students by arranging lavish South Indian dinners after the seminars. I was a very young boy, but I had the privilege of meeting the eminent visitors. The seminars were held in the upstairs lecture hall of *Ekamra Nivas*, and the dinners were either on the lawns or in the rear building, and were often served on plantain leaves as is the custom in South India. Some of these eminent scientists were our house guests. The magic moment was when Nobel Laureate Niels Bohr visited *Ekamra Nivas* in 1960 and lectured in my father's seminar. Bohr was visiting India as the guest of Prime Minister Jawaharlal Nehru. When Bohr returned to Delhi on completing his visit, he told reporters that two things impressed him the most: the massive set up of the Tata Institute in Bombay, and the small group of students trained by my father in Madras. This statement by Bohr was flashed in national newspapers like *The Hindu*, and it

---

K. Alladi  
Department of Mathematics, University of Florida, Gainesville, FL 32611, USA  
e-mail: [alladik@ufl.edu](mailto:alladik@ufl.edu)



attracted the attention of Prime Minister Nehru, who wanted to meet my father and his students. With the assistance of Mr. C. Subramaniam, the Minister for Education, such a meeting was arranged at the Raj Bhavan, the residence of the Governor of Madras. After meeting my father and his students, Nehru asked my father what he wanted. Here was the Prime Minister of India asking what you want! At such an instance, you do not ask for anything meagre. My father asked for an institute for advanced research like the Institute for Advanced Study in Princeton. With the support of Mr. C. Subramaniam and the benevolence of Jawaharlal Nehru, MATSCIENCE, The Institute of Mathematical Sciences, was inaugurated on 3 January, 1962, with my father as its Director.

Attached is the list of eminent visitors to Ekamra Nivas and the Theoretical Physics Seminar from 1954 to 1961 prepared from my father's documents. Dates of the visits are given in parenthesis. The list of students in the seminar is also given.

### **Distinguished Scientists Who Visited Alladi Ramakrishnan's Home and the Theoretical Physics Seminar, 1954–1961**

- 1) **Professor P. A. M. Dirac**, F.R.S., Nobel Laureate  
Lucasian Professor, Cambridge University, England (Dec 1954)
- 2) **Professor Mark Oliphant**, F.R.S. (he was later knighted and became Governor of South Australia)  
Australian National University, Canberra (Jan 1955)
- 3) **Professor C. F. Powell**, F.R.S., Nobel Laureate  
Melville Wills Professor of Physics, University of Bristol, England (Dec 1955)
- 4) **Professor Cherry**  
University of Melbourne, Australia
- 5) **Professor Harry Messel**  
Nuclear Science Foundation, University of Sydney, Australia (Jan 1957)
- 6) **Professor W. W. Bruechner**  
Massachusetts Institute of Technology, USA
- 7) **Professor T. G. Room**, F.R.S.  
University of Sydney, Australia
- 8) **Professor Laurent Schwartz**, Fields Medalist  
University of Paris, France
- 9) **Professor H. Pitt**, F.R.S.  
University of Leeds, England
- 10) **Sir C. G. Darwin**, F.R.S.  
Former President, Royal Society, England
- 11) **Professor L. Janossy**  
Director, Eotvos Institute, Budapest, Hungary (Jan 1959)
- 12) **Professor S. Koba**  
Yukawa Hall, Kyoto, Japan

- 13) **Dr. T. Kotani**  
University of Tokyo, Japan
- 14) **Professor Andre Mercier**  
University of Berne, Switzerland (Sept 1959)
- 15) **Professor N. Dallaporta**  
University of Padova, Italy (Oct 1959)
- 16) **Professor A. M. Lane**  
A.E. Research Establishment, Harwell, England (Dec 1959)
- 17) **Professor George Gamow**  
University of Colorado, USA (Dec 1959)
- 18) **Professor Abdus Salam**, F.R.S. (Salam later won the Nobel Prize)  
Imperial College, London, England (Jan 1960)
- 19) **Professor Niels Bohr**, Nobel Laureate  
Bohr Institute of Theoretical Physics, Copenhagen, Denmark (Jan 1960)
- 20) **Professor Christoff**  
University of Sofia, Bulgaria (Feb 1960)
- 21) **Professor Phillip Morrison**  
Cornell University, Ithaca, USA (Mar 1960)
- 22) **Professor A. H. Copeland**  
University of Michigan, Ann Arbor, USA (Oct 1960)
- 23) **Professor Kamp-de-Feriet**  
University of Lille, France (Jan 1961)
- 24) **Professor W. Heitler**, F.R.S.  
University of Zurich, Switzerland (Feb 1961)
- 25) **Professor Marshall H. Stone**  
Distinguished Service Professor, University of Chicago, USA (Apr 1961)
- 26) **Professor Hlavaty**  
Institute of Fluid Dynamics, Indiana, USA
- 27) **Professor Murray Gell-Mann** (Gell-Mann later won the Nobel Prize)  
California Institute of Technology, USA (Summer 1961)
- 28) **Professor R. A. Dalitz**  
University of Chicago, USA (Summer 1961)
- 29) **Professor Sandstrom**  
Uppsala University, Sweden (July 1961)
- 30) **Professor Donald Glaser**, Nobel Laureate  
University of California, Berkeley, USA (Aug 1961)
- 31) **Dr. Maurice Shapiro**  
Naval Research Lab., Washington, USA (Aug 1961)
- 32) **Professor S. Chandrasekhar** (he later won the Nobel Prize)  
Distinguished Service Professor, University of Chicago, USA (Nov 1961)
- 33) **Professor M. J. Lighthill** (he was later knighted)  
Director, Royal Aircraft Establishment, Farnborough, England (Nov 1961)
- 34) **Professor McCrea Hazlett**  
Vice-President, University of Rochester, USA (Dec 1961)

## **Students of Professor Alladi Ramakrishnan Who Attended the Theoretical Physics Seminar**

- 1) K. Ananthanarayanan
  - 2) A. P. Balachandran
  - 3) G. Bhamathi
  - 4) V. Devanathan (*acted as the Seminar Secretary*)
  - 5) N. G. Deshpande
  - 6) S. Indumathi
  - 7) P. M. Mathews
  - 8) T. K. Radha
  - 9) V. Radhakrishnan
  - 10) P. Rajagopal
  - 11) B. Ramachandran
  - 12) G. Ramachandran
  - 13) K. Raman
  - 14) N. R. Ranganathan
  - 15) M. Srinivasan
  - 16) S. K. Srinivasan
  - 17) R. Thunga
  - 18) R. K. Umerjee
  - 19) R. Vasudevan
  - 20) K. Venkatesan
  - 21) V. K. Viswanathan
- \* E. T. Nambi Iyengar (*helped with all academic correspondence.*)



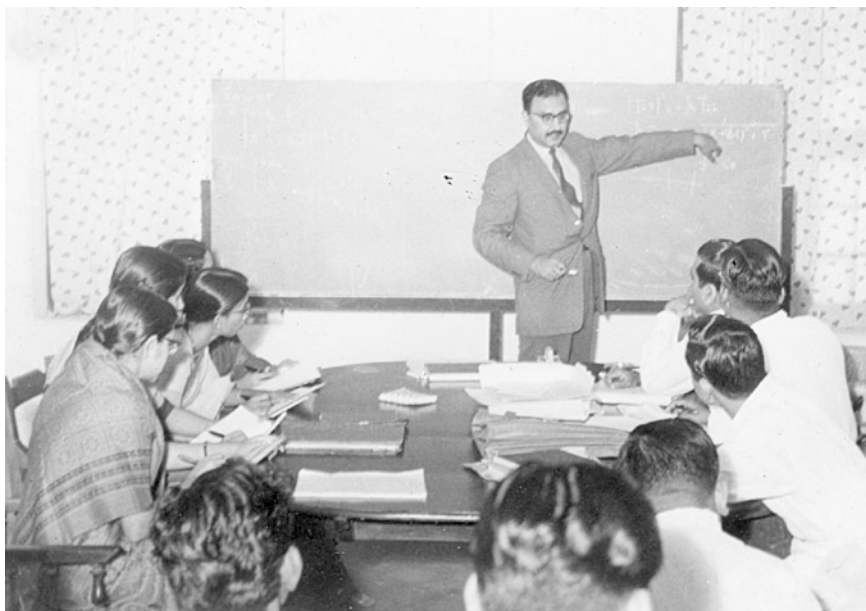
“Ekamra Nivas”, the family home of Alladi Ramakrishnan, was the venue of the Theoretical Physics Seminar, which was the genesis of MATSCIENCE



Prof. and Mrs. Alladi Ramakrishnan (seated) at a meeting in Madras of the Asian Students at which Education Minister C. Subramaniam (speaking) was the chief guest. It was at this meeting, after listening to Prof. Ramakrishnan, that Mr. Subramaniam decided to meet the students of the Theoretical Physics Seminar – Oct 1959



Education Minister Mr. C. Subramaniam in conversation with Prof. Ramakrishnan during a dinner at Ekamra Nivas where he met the students of the Theoretical Physics Seminar – Oct 1959



Prof. Abdus Salam, FRS (Imperial College, London), lecturing at Alladi Ramakrishnan's Theoretical Physics Seminar at Ekamra Nivas, Jan 1960



Abdus Salam (left) with Alladi Ramakrishnan, Mrs. Lalitha Ramakrishnan and students at Ekamra Nivas, Jan 1960



Alladi Ramakrishnan (R) hosted a dinner at his home Ekamra Nivas for Nobel Laureate Niels Bohr (Center), Jan 1960

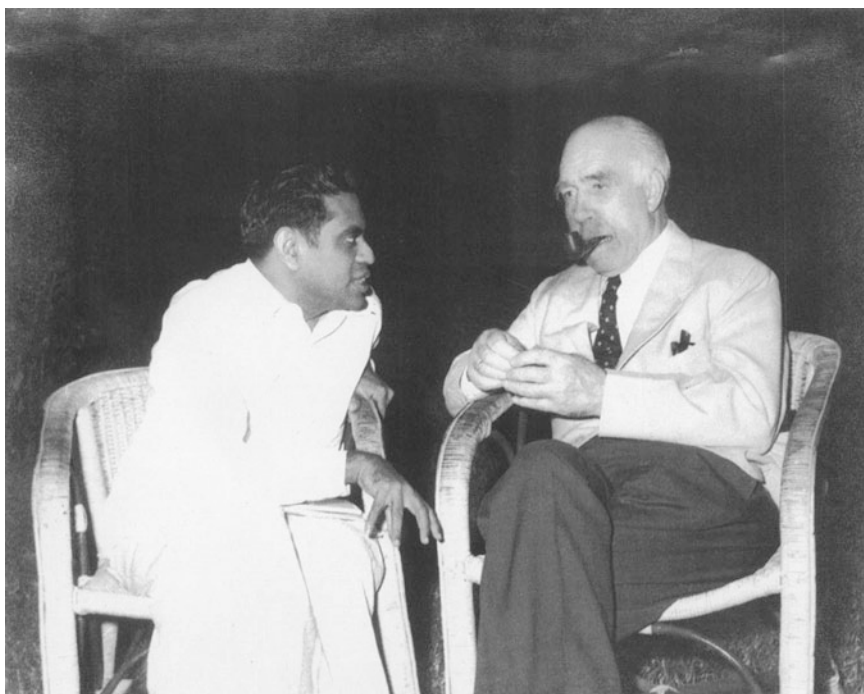




Nobel Laureate Niels Bohr and Mrs. Bohr with Alladi and Lalitha Ramakrishnan at Ekamra Nivas, Jan 1960



Krishna was always with his father in Madras when distinguished visitors were present, and on many trips abroad that Alladi Ramakrishnan went on. Here is Krishna with his father Alladi Ramakrishnan and Nobel Laureate Niels Bohr at Ekamra Nivas, Jan 1960



Nobel Laureate Niels Bohr in discussion with Professor Alladi Ramakrishnan at Ekamra Nivas, January 1960

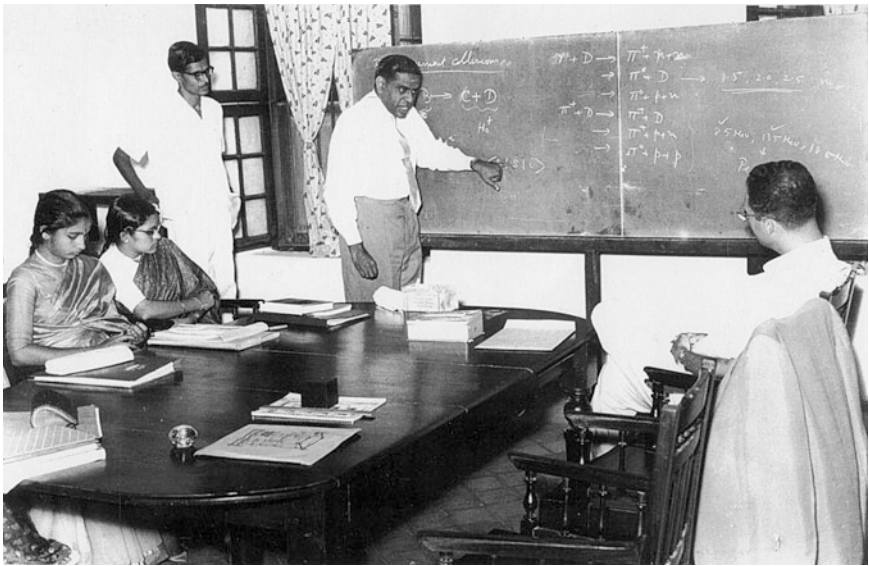


Alladi Ramakrishnan, Mrs. Lalitha Ramakrishnan and their son Krishna with Professor Marshall Stone at Madras Airport – April 1961





Alladi Ramakrishnan welcoming and introducing Nobel Laureate Donald Glaser (Berkeley) to the students of his Theoretical Physics Seminar, August 1961



Professor Alladi Ramakrishnan explaining the work of his group to Nobel Laureate Donald Glaser (Berkeley) at the Theoretical Physics Seminar – August 1961



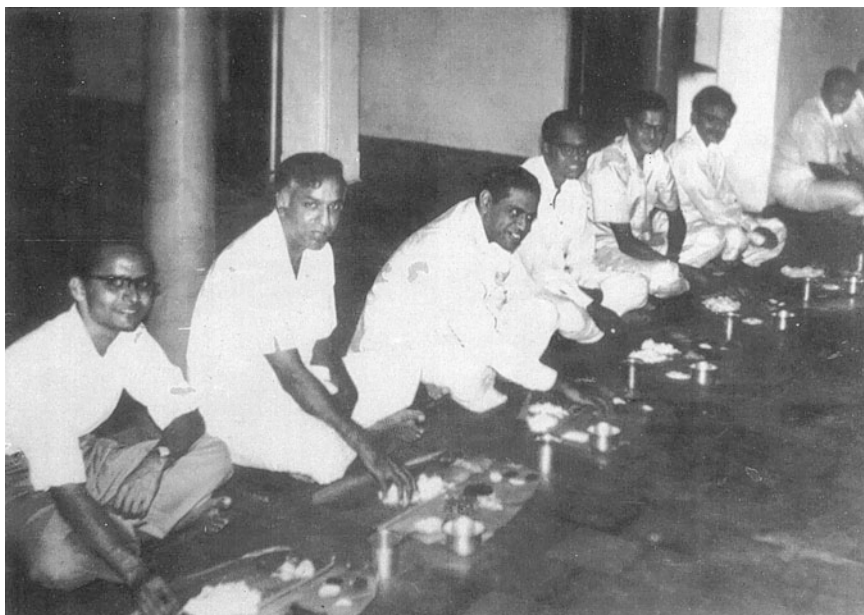
Professor Alladi Ramakrishnan with Nobel Laureate Donald Glaser at Ekamra Nivas – August 1961



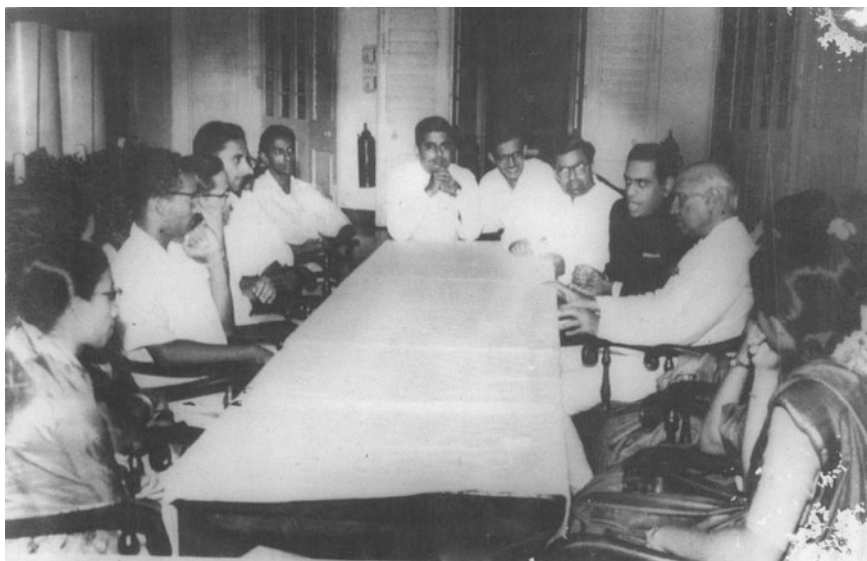
Prof. Alladi Ramakrishnan and Mrs. Lalitha Ramakrishnan with Sir James Lighthill (FRS) who visited “Ekamra Nivas” in November 1961. Also in the picture is Mr. C. Subramaniam, Minister of Education



Richard Dalitz (L) and Murray Gellmann (R) with Mrs. Lalitha Ramakrishnan at Ekamra Nivas, 1961



Astrophysicist Subrahmanyam Chandrasekar (second from left) enjoying a South Indian Style dinner served on a banana leaf at Ekamra Nivas. Also in the picture – Alladi Ramakrishnan (third from left) and students of the Theoretical Physics Seminar, Nov 1961



Professor Alladi Ramakrishnan and the students of his Theoretical Physics Seminar with Prime Minister Jawaharlal Nehru at the Raj Bhavan (Governor's Residence) in Madras on October 8, 1961

# Telegrams Received for the MATSCIENCE Inauguration

Krishnaswami Alladi

The inauguration of MATSCIENCE, The Institute of Mathematical Sciences, Madras, India, on 3 January 1962, was greeted with great enthusiasm by scientists from around the world. My father, Professor Alladi Ramakrishnan, in his inaugural speech as the Director of the new Institute, referred to the creation of MATSCIENCE as a miracle, because a series of unexpected pleasant circumstances came in rapid succession to bear fruit. About a week before the inauguration of MATSCIENCE, congratulatory telegrams and letters started pouring in from scientists around the world. I was just past my sixth birthday at that time, but I remember the sense of excitement at our family home *Ekamra Nivas* as my father was preparing for that sensational event. I remember a dinner at the roof garden of the Dasaprakash Hotel in Madras a few days before the inauguration at which my father's students who attended his *Theoretical Physics Seminar* were present. At this dinner, my father asked each student to predict the number of congratulatory telegrams and messages that would be received by 3 January 1962. Such was the mood at that magic moment! My father had preserved these telegrams and had them photocopied and bound in two volumes. In the following pages I have presented photo copies of a selection of these telegrams and letters. I have made some observations about the person sending the message/telegram and my father's association with that scientist.

---

K. Alladi  
Department of Mathematics, University of Florida, Gainesville, FL 32611, USA  
e-mail: [alladik@ufl.edu](mailto:alladik@ufl.edu)



INVEST WISELY Buy NATIONAL SAVINGS CERTIFICATES  
 1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53. 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65. 66. 67. 68. 69. 70. 71. 72. 73. 74. 75. 76. 77. 78. 79. 80. 81. 82. 83. 84. 85. 86. 87. 88. 89. 90. 91. 92. 93. 94. 95. 96. 97. 98. 99. 100.

90 (8) repeat

1329  
3-1-62  
C

INDIAN POSTS AND TELEGRAPHS DEPARTMENT

**INVEST WISELY**  
Buy NATIONAL SAVINGS CERTIFICATES

Class }  
Prefix } Code } No. }

Recd. from } Sent at } H. } M. }  
By } To }  
By } By }

Handed in at (Office of Origin) } Date } Hour } Minute } Service }  
Recd. here at } H. } M. }

TO }  
XF MF R 214 LONDON 2 OCS 26  
PROFESSOR ALLADI RAMAKRISHNAN EKAMRANIVAS  
27 LUZ MYLAPORE MADRAS MSX  
= WARMEST WISHES THEORETICAL INSTITUTE CONFIDENCE UNDER  
YOUR INSPIRING LEADERSHIP INSTITUTE WILL BECOME A GREAT  
CENTRE FOR RESEARCH = SALAM =

Telegram from ABDUS SALAM, Imperial College, London, Dec. 28, 1961

*Comments.* Alladi Ramakrishnan and Abdus Salam were close friends and admired each other not only for their research, but also for the efforts they both made for the scientific profession. Salam was interested in starting an institute of fundamental research in Pakistan, but he eventually created the *International Center for Theoretical Physics (ICTP)* in Trieste, Italy, in 1964, with the support of UNESCO. Salam was Director of ICTP since its inception until his death, but he retained his position at Imperial College, London. Salam visited Madras in 1960 as Ramakrishnan's guest and lectured at the Theoretical Physics Seminar. He invited Ramakrishnan to a conference in Italy in 1960 to have discussions with a small group of visionary scientists when he (Salam) was planning the creation of ICTP. Thus, it was natural that Salam admired and supported Ramakrishnan's efforts to create an institute for fundamental research in India. Ramakrishnan visited ICTP several times in the 1960s at the invitation of Salam, and his wife and son accompanied him for these extended stays in Trieste, the first being in 1965 when the ICTP was located in Piazza Oberdan before it moved to its magnificent permanent location in Grignano, near the Castle Miramare outside Trieste. After Salam won the 1979 Nobel Prize, he visited MATSCIENCE in 1980 which had by then moved to its permanent home on Taramani Campus in Madras.

INVEST WISELY BUY NATIONAL SAVINGS CERTIFICATES

INDIAN POSTS AND TELEGRAPH DEPARTMENT

3 Pages 4 34

20 3rd

Class }  
 Prefix } Code } No. }  
 Recd. from }  
 By }  
 Sent at } H. } M. }  
 Date. } Hour. } Minute. }  
 Remarks }  
 Words. }

TO XF PL 300 KOBENHAVN 2 OCS 115

PROFESSOR RAMAKRISHNAN EKAMRA NIVAS 27 LUZMYLAPORE MADRAS  
 AT INAUGURATION OF THE INSTITUTE OF MATHEMATICAL SCIENCES  
 IN MADRAS THE WHOLE GROUP OF THE COPENHAGEN INSTITUTE FOR  
 THEORETICAL PHYSICS WANTS TO SEND ITS HEARTIEST  
 FELICITATIONS STOP THE COMMUNITY OF PHYSICISTS HAS BEEN  
 IMPRESSED BY THE VIGOUR AND ZEAL WITH WHICH PROF RAMAKRISHNAN  
 HAS BEEN ABLE TO EDUCATE AND INSPIRE HIS YOUNG PUPILS AND  
 COLLABORATORS AND THE WORK IN THE NEW INSTITUTE WILL BE  
 FOLLOWED WITH KEEN EXPECTATIONS STOP INDEED AS AN IMPORTANT  
 ASSET TO SCIENTIFIC RESEARCH IN INDIA THE CREATION OF THE  
 MADRAS INSTITUTE IS EAGERLY WELCOMED IN THAT WORLDWIDE  
 COOPERATION IN SCIENCE WHICH OFFERS SO GREAT OPPORTUNITIES  
 FOR PROMOTING THE UNDERSTANDING BETWEEN ALL PEOPLES  
 NIELS BOHR

This is the three page telegram from NIELS BOHR reassembled as one page

Comments. Nobel Laureate Niels Bohr was one of the greatest and most influential physicists. Bohr founded and directed the Copenhagen Institute of Theoretical Physics. Bohr visited India in January 1960 as the personal guest of Prime Minister Jawaharlal Nehru. Alladi Ramakrishnan had corresponded with Professor Bohr earlier, and Bohr graciously agreed to visit the Theoretical Physics Seminar at Alladi Ramakrishnan's home Ekamra Nivas in Madras. Bohr and his wife had dinner on the lawns of Ekamra Nivas and stayed until midnight



talking to Ramakrishnan, the students, and other guests. Upon return to Delhi at the end of his visit, Bohr expressed the opinion that two things impressed him the most on his trip to India – the massive setup of the Atomic Energy Commission and the Tata Institute founded by Homi Bhabha in Bombay, and the group of students being trained by Alladi Ramakrishnan in Madras. Here is a quote from the *The Hindu*, India's National Newspaper, about Bohr's statement:

Dr. Bohr said that the Atomic Energy Establishment was a mighty endeavor where research is being conducted in the best way under the leadership of Dr. H.J. Bhabha, a great scientist and at the same time a very good administrator.

Asked about the place mathematics should occupy in the pursuit of theoretical physics, the professor said that in Bombay and Madras energetic efforts were being made for the promotion of knowledge of physics which demanded new mathematical methods of education of young people to be able to fruitfully contribute to such work. Wonderful work was being done in the field of theoretical physics by Professor Alladi Ramakrishnan of the Madras University.

This statement by Bohr, which was flashed in the newspapers, sparked the attention of Prime Minister Nehru, and ultimately led to the creation of MATSCIENCE.


Later on an academic trip to Europe in 1960, Alladi Ramakrishnan visited the Copenhagen Institute at Bohr's invitation and attended the Symposium on Nuclear Structure. Ramakrishnan was invited for dinner at Bohr's residence where he met the whole family. Nobel Laureate Jensen was at this dinner. Ramakrishnan was also invited to a party at the home of Aage Bohr (son of Niels Bohr) who later became a Nobel laureate!

Bohr's happiness in the creation of MATSCIENCE and his total support can be seen from both the length of his telegram and its contents.

---


63 ✓      7  
61

**INVESTWISELY**  
Buy NATIONAL SAVINGS CERTIFICATES



INDIAN POSTS AND TELEGRAPHS DEPARTMENT

4.1.62

|  |                           |   |      |         |                      |       |
|--|---------------------------|---|------|---------|----------------------|-------|
| Class }<br>Prefix }  | Code _____                | No. _____   | C.   |         |                      |       |
| Recd. from _____   | Sent at _____ H. _____ M. |  |      |         |                      |       |
| By _____   | To _____<br>By _____      |   |      |         |                      |       |
| Landed in at (Office of Origin)  |                           | Date  | Time | Address | Service instructions | Words |
| LT 1320 DR 1Q2 MUENCHEN 4 OCS 31   |                           | Recd. here at _____ H. _____ M.   |      |         |                      |       |
| LT PROFESSOR ALLADI RAMAKRISHNAN KAMRA NIVAS 27 LUZ MYLAPORE   |                           |   |      |         |                      |       |
| MADRAS :   |                           |   |      |         |                      |       |
| MY BEST WISHES TO THE INAUGURATION OF THE INSTITUTE OF MATHEMATICAL SCIENCES AND MUCH SUCCESS IN ITS FUTURE WORK : WERNER HEISENBERG : |                           |   |      |         |                      |       |

M. B.—The name of the sender, if telegraphed, should be written after but separated from, the text

Telegram from WERNER HEISENBERG, Munich, Germany, Jan 4, 1962

*Comments.* Alladi Ramakrishnan first met Professor Nobel Laureate Werner Heisenberg in 1949 at a Conference on Modern Physics in Edinburgh, Scotland. At that time, Ramakrishnan was doing his Ph.D. at the University of Manchester on the topic of product densities in the area of probability. Heisenberg invited Ramakrishnan to give a seminar in Gottingen during Ramakrishnan's round-the-world academic tour of 1956. Heisenberg was very much impressed with Ramakrishnan's talk in the theory of probability and stochastic processes. Professor Flugge, Editor of the *Handbuch der Physik* of Springer-Verlag, was also at this seminar. Based on Heisenberg's recommendation, Flugge invited Ramakrishnan to write a comprehensive article focussing on his *theory of product densities* and its relationship with central ideas in the theory of probability and stochastic processes. This was published in the *Handbuch der Physik* in 1959.\*

---

\*Alladi Ramakrishnan, "Probability and stochastic processes", in *Handbuch der Physik*, 3 (1959) Springer, Berlin, 524-651.

8 (b)

**INVEST WISELY**  
Buy NATIONAL SAVINGS CERTIFICATES

**INDIAN POSTS AND TELEGRAPHS DEPARTMENT**

Class XF Code KB No. 34

Recd. from 8 Sent at ..... H ..... M. Office MADRAS

By AS To ..... By ..... MADRAS D. 19. 29/12

|                                 |           |           |           |                      |           |
|---------------------------------|-----------|-----------|-----------|----------------------|-----------|
| Handed in at (Office of Origin) | Date      | Hour      | Minute    | Service Instructions | Words     |
| <u>R30 Pasadena call</u>        | <u>25</u> | <u>00</u> | <u>00</u> | <u>00</u>            | <u>22</u> |

TO Prof Alladi Ramakrishnan Ekamra  
no 27 Luz Mylapore mdr  
= Best wishes for the success  
of the institute Murray Gellman  
Professor of physics

Telegram from MURRAY GELL-MANN, Caltech, Dec 29, 1961

C-3.

**INVEST WISELY**  
Buy NATIONAL SAVINGS CERTIFICATES

**INDIAN POSTS AND TELEGRAPHS DEPARTMENT**

No. 11

Received here at ..... H ..... M. 27

LT V 276 CHICAGO ILL 28 06S 42 30/12

== LT PROF ALLADI RAMAKRISHAN 27 LUZ MYLAPORE MADRAS INDIA ==  
= DELIGHTED TO LEARN OF INAUGURATION OF INSTITUTE OF  
MATHEMATICAL SCIENCES STOP I CONGRATULATE MADRAS STATE  
AND SEND BEST WISHES AND CONFIDENT HOPES FOR A BRILLIANT  
FUTURE FOR THE INSTITUTE RICHARD DALITZ UNIVERSITY OF CHICAGO =

The sequence of entries at the beginning of this telegram is—class of telegram, time handed in, serial number (In the case of Foreign Telegrams only), office of origin, date, service instructions (if any) and number of words.  
This form must accompany any inquiry respecting this telegram.  
L. C. & Sons, Calcutta—No. C 6/57 (MFP. Regn. No. 11/3/P-661—20-1-65)—(P-1/240A/55-56)—18-2-57—2,04,000—Es.

Telegram from RICHARD DALITZ, University of Chicago, Dec 29, 1961

*Comments.* Professor Murray Gell-Mann (Caltech) and Professor Richard Dalitz (University of Chicago) were visiting India in 1961 in connection with a summer school in theoretical physics conducted by the Tata Institute in Bangalore that Ramakrishnan also attended. Gell-Mann's work in physics was already creating a sensation. Ramakrishnan invited Gell-Mann and Dalitz to Madras to speak at the *Theoretical Physics Seminar* and both of them stayed at *Ekamra Nivas* prior to the Bangalore summer school. Gell-Mann and Dalitz thoroughly enjoyed their visit to Madras, both academically and socially. Gell-Mann invited Ramakrishnan to give a Colloquium in Caltech. That visit to Caltech was one of the highlights of Ramakrishnan's round-the-world tour in 1962. His wife Lalitha and son Krishna accompanied Ramakrishnan on that magnificent trip, the first for Krishna overseas, and the beginning of several such trips with his father during the next ten years. Gell-Mann hosted a party at his home in honor of Ramakrishnan. The legendary physicist Richard Feynman (who later won the Nobel Prize) was also at this party.


Gell-Mann won the 1969 *Nobel Prize* in physics for his *quark model of the atom*. Inspired by Gell-Mann's revolutionary work, Ramakrishnan studied Gell-Mann's ideas closely; subsequently, Ramakrishnan obtained an elegant generalization of the Gell-Mann–Nishijima relation.

---

Special Telegrams: 1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53. 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65. 66. 67. 68. 69. 70. 71. 72. 73. 74. 75. 76. 77. 78. 79. 80. 81. 82. 83. 84. 85. 86. 87. 88. 89. 90. 91. 92. 93. 94. 95. 96. 97. 98. 99. 100.

392-10  
40-  
12  
3.1.62  
C

INVEST WISELY  
Buy NATIONAL SAVINGS CERTIFICATES



INDIAN POSTS AND TELEGRAPHS DEPARTMENT

|                                 |                            |              |         |                      |       |
|---------------------------------|----------------------------|--------------|---------|----------------------|-------|
| Class                           |                            | No.          |         |                      |       |
| Postage                         | Cash                       |              |         |                      |       |
| Send from                       | Send at                    | H.           | M.      |                      |       |
| By                              | To                         |              |         |                      |       |
| By                              | By                         |              |         |                      |       |
| Handed in at (Office of Origin) | Date                       | Hour         | Minutes | Service Instructions | Words |
| TO                              | XF LH 6 PRINCETON 2 OCS 31 | Send here at | H.      | M.                   |       |

PROFF A RAMAKRISHNAN EKAMBRA NIVAS 27 LUZ MYLAPORE MMN

BEST WISHES FOR THE FUTURE SUCCESS OF YOUR INSTITUTE

AS A CENTER OF SCIENTIFIC RESEARCH = T D LEE AND C N YANG =

\* The name of the sender of telegrams, should be written after, but enclosed from, the text


Telegram from T.D. LEE and C.N. YANG, Institute for Advanced Study, Princeton, Jan 3, 1962


Comments. When Ramakrishnan visited the Institute for Advanced Study in Princeton in 1957-1958, Lee and Yang were in residence at the Institute. Everyone was talking about Lee and Yang's theory of nonconservation of parity and excited about the possibility of a Nobel Prize in Physics which they were awarded later that year.

56

C-3. 000075

INVEST WISELY  
Buy NATIONAL  
SAVINGS  
CERTIFICATES

  
**INDIAN POSTS AND TELEGRAPHS DEPARTMENT**

No.  4.1.62

Received here at \_\_\_\_\_ H. \_\_\_\_\_ M. \_\_\_\_\_

XF DL 67 NEWYORKNY 3 OCS 22

DR ALLADI RAMAKRISHNAN EKAMRA NIVAS 27 LUZ MYLAPORE MADRAS

= BEST WISHES FOR SUCCESS OF THE NEW INSTITUTE OF  
MATHEMATICAL SCIENCES = MARK KAC =


The sequence of entries at the beginning of this telegram is—class of telegram, time handed in, serial number (in the case of Foreign Telegrams only), office of origin, date, service instructions (if any) and number of words.  
This form must accompany any inquiry respecting this telegram.  
L. C. & Seas, Calcutta—No. G 6/57 (MFP. Regn. No. 11/3/P-561—30-1-56)—(P-1/249A/55-56)—18-2-55—3,04,000 Hrs.

Telegram from MARK KAC, Rockefeller University, New York, Jan 4, 1962

*Comments.* Mark Kac was a very famous probabilist of Polish descent. He was born in Ukraine in 1914. He received his Ph.D. in 1937 in Lwow, Poland, under the direction of Hugo Steinhaus. Kac immigrated to the United States in 1938. He was at Cornell University until 1931 when he moved to Rockefeller University in New York where he stayed for 20 years before finally moving to the University of Southern California. Kac was mainly interested in probability. His question *Can you hear the shape of a drum?* spurred enormous research activity. He is also known for the *Erdős-Kac Theorem* which led to the creation of Probabilistic Number Theory.

322

INVEST WISELY  
 Buy NATIONAL SAVINGS DEBIT CARDS



INDIAN POSTS AND TELEGRAMS DEPARTMENT

900190 2/1/62

|                                 |         |      |             |
|---------------------------------|---------|------|-------------|
| Class                           | Code    | No.  |             |
| Prof. Code                      |         |      |             |
| Recd. from                      | Sent at | H    | M.          |
| By                              | To      |      |             |
| Handed in at (Office of Origin) |         | Date | Hour Minute |
| Service Instructions.           |         |      |             |

OFFICE STAMP  
 2/1/62  
 TELEGRAMS DEPARTMENT

LT 2145 95 SANTIAGOCHILE 1 OCS 21 LT PROFESSOR ALLADI  
 RAMAKRISHNAN 27 LUZ MYLAPORE MADRAS ==  
 BEST WISHES FOR INAUGURATION AND SUBSEQUENT SUCCESS INSTITUTE OF  
 MATHEMATICAL SCIENCES == MARSHALL STONE =

Telegram from MARSHALL STONE, University of Chicago, Jan 4, 1962

*Comments.* Marshall Stone was one of the most influential mathematicians of the twentieth century. The *Stone–Weierstrass theorem* is so fundamental that everyone going through graduate school in mathematics sees this in a course on real analysis. Stone was more than just a great mathematician. He believed in making contributions to the profession. Under his dynamic leadership as Chairman, the Mathematics Department at the University of Chicago grew to great heights in the 1950s. The period when Stone was chairman at Chicago has been often referred to as *The Stone Age*! Stone was also President of the American Mathematical Society. Thus with his own desire to mould the shape of mathematics education and research in America, he could understand and appreciate Alladi Ramakrishnan’s interests in creating a stimulating atmosphere for scientific research in Madras. Also, Marshall Stone’s father was a Justice of the US Supreme Court; thus, Professor Stone could appreciate and understand Alladi Ramakrishnan’s family background very well.


Professor Stone visited India regularly in the 1960s and 1970s because he served on committees for the improvement of mathematics in India. During some of these visits, he lectured at MATSCIENCE. In 1963 he was *Ramanujan Visiting Professor* at MATSCIENCE. Later in January 1969, he was present when MATSCIENCE moved into its new buildings on Taramani Campus. Stone attended the inauguration of the new building and gave the first lecture there.

Stone visited Madras regularly in December/January to attend the annual festival of the Madras Music Academy with Alladi and Lalitha Ramakrishnan. Professor Stone died in Madras in January 1987 after attending the 1986–1987 music season there.

Stone traveled extensively. This telegram was sent from Chile!

41 ✓ 42

**INVEST WISELY**  
Buy NATIONAL SAVINGS CERTIFICATES



**INDIAN POSTS AND TELEGRAPH DEPARTMENT**

No. **26 C**

Class }  
Prefix } Code \_\_\_\_\_

Recd. from \_\_\_\_\_ Sent at \_\_\_\_\_ H. \_\_\_\_\_ M. \_\_\_\_\_  
To \_\_\_\_\_  
By \_\_\_\_\_

Handed to at (Office of Origin) \_\_\_\_\_ Date \_\_\_\_\_ Hour \_\_\_\_\_ Minute \_\_\_\_\_  
By \_\_\_\_\_

TO \_\_\_\_\_ Recd. here at \_\_\_\_\_ H. \_\_\_\_\_ M. \_\_\_\_\_

LT ED 68 LONDON 2 OCS 28 LT ALLADI


RAMAKRISHNAN EKAMRA NIVAS 27 LUZ MYLAPORE MMN

BEST WISHES TO INSTITUTE OF MATHEMATICAL SCIENCES  
REPRESENTING IMPORTANT NEW ENCOURAGEMENT TO INDIAN RESEARCH  
MAURICE BARTLETT UNIVERSITY COLLEGE LONDON

Telegram from MAURICE BARTLETT, University College, London, Jan 3, 1962

22 ✓  
46

**INVEST WISELY**  
Buy NATIONAL SAVINGS CERTIFICATES



**INDIAN POSTS AND TELEGRAPHS DEPARTMENT**

No. **77**

Received here at \_\_\_\_\_ H. \_\_\_\_\_ M. \_\_\_\_\_

GLT. No. DR 78 Oxford 30 Oct 14  
GLT Remakrishnan Ekamra Nivas  
27 Luz mylapore madras

Warmest felicitations and best wishes  
= David Kendall

The sequence of entries at the beginning of this telegram is—class of telegram, time handed in, serial number (in the case of Foreign Telegrams only), office of origin, date, service instructions (if any) and number of words. This form must accompany any inquiry respecting this telegram.  
L. C. & Sons, Calcutta—No. G 6/67 (MFP. Regn. No. 11/3/P-561-80-1-56)-(P-1/249A/56-56)-12-2-57-2,04,000 Eka.

Telegram from DAVID KENDALL, Oxford University, Dec 31, 1961



*Comments.* Professor M.S. Bartlett, a highly reputed statistician, was the Ph.D. advisor of Alladi Ramakrishnan at the University of Manchester. Bartlett is widely known for his work in multivariate analysis, stochastic processes, and applications of statistics to genetics, and wrote a number of influential papers and books. Among his honors are the Guy Silver (1952) and Gold (1969) Medals of the Royal Statistical Society, Fellowship of the Royal Society (1961), his election as Foreign Associate to the US National Academy of Sciences (1993). He was President of the Royal Statistical Society (1966).

Bartlett was appointed as Professor at the University of Manchester in 1947. When Alladi Ramakrishnan arrived at the University of Manchester in 1949, Bartlett was much impressed with Ramakrishnan's *method of product densities* and communicated Ramakrishnan's papers\* to the Proceedings of the Cambridge Philosophical Society.

David Kendall (FRS), another very eminent statistician, was on Alladi Ramakrishnan's Ph.D. committee. Kendall at that time was at Oxford, and Ramakrishnan visited Oxford regularly to have discussions with Kendall at Magdalene College. Alladi Ramakrishnan's product density method extended Kendall's work to higher orders.\*\* When Alladi Ramakrishnan visited Oxford University in 1960 during his trip to Europe, Kendall invited Ramakrishnan to a High Table dinner. Kendall later was appointed as professor at Cambridge University.

---

---

\* Alladi Ramakrishnan, "Stochastic processes relating to particles distributed in a continuous infinity of states", *Proc. Cambridge Phil. Soc.*, **46** (1950), 595–602.


Alladi Ramakrishnan, "Stochastic processes associated with random divisions of a line", *Proc. Cambridge Phil. Soc.*, **49** (1953), 473–485.

\*\* Alladi Ramakrishnan, "Stochastic processes and their applications to physical problems", *Ph.D. Thesis, Univ. Manchester* (1951).


56 154  
24 3015

G-3.

**INVEST WISELY**  
Buy NATIONAL SAVINGS CERTIFICATES



**INDIAN POSTS AND TELEGRAPHS DEPARTMENT**

No. 

Received here at \_\_\_\_\_ H. \_\_\_\_\_ M.

LT PC DR -111 HEIDELBERG 29 QCS 41

LT PROF ALLADI RAMAKRISHNAN EKAMRA NIVAS  
27 LUZ MYLAPORE MADRAS ===:

MY GREETINGS AND BEST WISHES ON THE OCCASION OF THE  
FOUNDATION OF THE NEW INSTITUTE OF MATHEMATICAL  
SCIENCES WHICH REPRESENTS THE NASCENT SPRIT OF THE NEW  
SCIENTIFIC GENERATION IN INDIA = MAASS =

The sequence of entries at the beginning of this telegram is—class of telegram, time handed in, serial number (in the case of Foreign Telegrams only), office of origin, date, service instructions (if any) and number of words.  
This form must accompany any inquiry respecting this telegram.  
L. C. & Sons, Calcutta—No. G 8/57 (MFP. Regn. No. 11/2/P-561—20-1-56)—(P-1/240A/55-34)—18-2-57—2,04,000 Bks.

Telegram from HANS MAASS, Max Planck Institute, Heidelberg, Germany, Dec. 30, 1961

*Comments.* Hans Maass, a very eminent German mathematician, made notable contributions to Number Theory in the area of modular forms. He is now most known for introducing in 1949 what are now called *Maass wave forms* which in the past few years have become crucial in understanding the relationship between Ramanujan's mock theta-functions and the theory of modular forms. Alladi Ramakrishnan met Maass during his first round-the-world academic tour of 1956 when Maass invited Ramakrishnan for a talk at the Max Planck Institute in Heidelberg.

---

UNIVERSITY OF CAMBRIDGE DEPARTMENT OF PHYSICS

TELEPHONE  
CAMBRIDGE 54481

CAVENDISH L  
FREE SCHOOL  
CAMBRIDGE

Professor N. F. Mott, F.R.S.

3rd January, 1962.

70

Dear Professor Ramakrishnan,

Unfortunately your letter only reached me today and it seems too late to send you the cable. I would, however, like to send you the best wishes from all of us in the Cavendish Laboratory for your new venture.

Yours sincerely,

N. F. Mott

Professor A. Ramakrishnan,  
'Ekamra Nivas,  
27 Luz,  
Mylapore,  
Madras, INDIA.

Letter from N.F. Mott, F.R.S., University of Cambridge, Jan 3, 1962

*Comments.* Sir Nevill Francis Mott, F.R.S., won the Nobel Prize for physics in 1977 (along with Philip W. Anderson and J.H. van Vleck) for his work on the electronic structure of magnetic and disordered systems. Mott held a lectureship at the University of Manchester in 1929 but moved to Cambridge in 1930. He then was at Bristol where he was Wills Professor of Physics and Director of the Wills Physical Laboratories before being appointed Cavendish Professor of Physics at Cambridge in 1954. Mott was elected Fellow of the Royal Society in 1936 and served as President of the Physical Society in 1957. Alladi Ramakrishnan met Mott in England while doing his Ph.D. in Manchester during 1949-1951.

---

72  
7 Cavendish  
Ca  
6  
72  
11.1.62  
Dear Ramakrishnan,

I was very glad to hear that there is to be an Institute of Advanced Mathematics in Madras and that you are to be its Director. I send you my warmest congratulations.

I did not get your letter until Jan 4<sup>th</sup>. I suppose it was delayed by the Christmas rush.

Wishing every success to you and your new Institute

Yours sincerely

P A M Dirac

Letter from P.A.M. Dirac, Cambridge University, England, Jan 11, 1962

*Comments.* Professor Dirac was one of the greatest physicists of the twentieth century. He received the Nobel Prize in Physics in 1933 along with Erwin Schroedinger. He is the one who predicted the positron. Many things are named after him, such as the Dirac delta-function and the Dirac equation. He was Lucasian Professor at Cambridge University, England. Among his notable students were Homi Bhabha who later founded the Tata Institute, and Harish-Chandra of Lie theory fame. Professor Dirac was a guest of Alladi Ramakrishnan at Ekamra Nivas in December 1954.

Dirac was a great influence on Ramakrishnan for several reasons. After Ramakrishnan visited the Institute for Advanced Study in Princeton in 1957–1958, the focus of his research shifted to elementary particle physics. One of the problems that engaged Ramakrishnan's attention was why Dirac used only a set of four anticommuting matrices and discarded the fifth (denoted as  $\gamma_5$ ) in his theory. In understanding this, Ramakrishnan came up with a new idea, namely that of a  $\sigma$ -operation, which explained first how the anticommuting  $4 \times 4$  Dirac matrices can be built from the  $2 \times 2$  Pauli matrices. Then with the  $\sigma$ -operation, he constructed\*

\*Alladi Ramakrishnan, "The Dirac Hamiltonian as a member of a hierarchy of matrices", *J. Math. Anal. Appl.*, **20** (1967), 9–16.

an algebra of  $2n \times 2n$  matrices. In a sequence of papers by himself and with his students, he studied various ramifications of this algebra and its connections with Clifford algebras. He published a book (*L-Matrix theory, or the grammar of Dirac matrices*, Tata McGraw-Hill, 1972) which is a compilation of his papers on this topic. In 1980 on a visit to Florida State University for a lecture in the statistics department, Ramakrishnan called on Dirac (who had moved to the physics department at Florida State University after retirement from Cambridge) and had a discussion with him on the  $\sigma$ -operation.

---

UNIVERSITÉ DE PARIS  
FACULTÉ DES SCIENCES

71

DÉPARTEMENT DE MATHÉMATIQUES  
11 rue Pierre Curie  
PARIS 5<sup>e</sup>

10.1.62

Paris le 5 janvier 1962

Tél. : MEDICIS 22-50

77

Mon cher Ramakrishnan,

J'ai été très heureux de la nouvelle que vous m'annoncez, de la création d'un Institut de recherche à Madras. C'était infiniment souhaitable et je suis sûr que vous lui assurerez le meilleur succès. Je pense que cela vous donnera un lourd travail, mais dont l'utilité est certaine.

J'aurais aimé pouvoir vous envoyer un câble pour l'inauguration mais je n'ai pas reçu votre lettre à temps.

Avec mes meilleurs vœux et mes meilleurs sentiments.



Paris. I.A.C.

Laurent SCHWARTZ  
37 rue Pierre Nicole  
PARIS (5<sup>ème</sup>)

Letter from Laurent SCHWARTZ, Paris V, Jan 5, 1962

*Comments.* Laurent Schwartz, a great French mathematician, was awarded the Fields Medal in 1950 for creating the theory of distributions through which one gets, for example, a clearer understanding of the Dirac delta-function. He was for many years at the Ecole Polytechnique in Paris. 1966 Fields Medalist Alexander Grothendieck was a student of Laurent Schwartz. Alladi Ramakrishnan's interest in stochastic processes and probability motivated his interest in the fundamental work of Schwartz on distributions. Laurent Schwartz visited Alladi Ramakrishnan's Theoretical Physics Seminar in 1957.

RESEARCH INSTITUTE FOR FUNDAMENTAL PHYSICS  
(YUKAWA HALL)  
Kyoto University  
Kyoto, Japan

81

January 8, 1962

Professor Alladi Ramakrishnan  
Department of Physics  
University of Madras  
Madras  
India

78

Dear Professor Ramakrishnan:

Since I was absent from our Institute during the closing days and new year's days, I had the chance to see your letter only when it was too late to send a cable in time for the inauguration of your Institute. But, please accept my hearty congratulations. I am very pleased to know that a new institute for advanced research in mathematics and theoretical physics is established in Madras and that you are the first director. I am sure that your Institute will contribute a great deal not only to the advancement of mathematics and theoretical physics, but also will serve for closer cooperation between Indian and Japanese scientists, in particular, and Asian scientists more generally.

Sincerely yours,

*Hideki Yukawa*

Hideki Yukawa, Director  
Research Institute for  
Fundamental Physics  
Kyoto University, Kyoto  
Japan

Letter from Hideki Yukawa, Director Research Institute for Fundamental Physics Kyoto University, Kyoto, Japan, Jan 8, 1962

*Comments.* Hideki Yukawa won the 1949 Nobel Prize in physics for predicting the existence of the pion which was discovered in 1947. After briefly serving as professor at Columbia University, Yukawa became the First Director of the Yukawa Institute of Theoretical Physics (=Yukawa Hall) in Kyoto in 1953. Alladi Ramakrishnan visited Yukawa Hall for two weeks during his momentous first round-the-world tour of 1956. Meeting Professor Yukawa and the new generation of Japanese physicists in the post-World War II era in Japan made a big impression on Ramakrishnan, and gave him a desire to start create a similar institute and atmosphere in Madras. Alladi Ramakrishnan has acknowledged the effect of the visit to Yukawa Hall in his speech "A Miracle Has Happened" that he delivered at the inauguration of MATSCIENCE.

---

MASSACHUSETTS INSTITUTE OF TECHNOLOGY CAMBRIDGE 39, M.

Professor Alladi Ramakrishnan,  
University of Madras,  
Ekamra Nivas,  
27, Luz, Mylapore,  
Madras, India

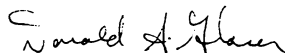
January 9, 1962

Dear Professor Ramakrishnan:

Congratulations on your success in convincing the Government of Madras to sponsor an Institute of Mathematical Sciences there. I wish you the very best of luck as the first director of this Institute in pursuing advanced research in mathematics and theoretical physics, in which you and your students and colleagues have already had considerable success.

I am very sorry that, because of the holidays, I did not have the opportunity to write you in time for the inauguration ceremony on January 3, but wish to send you anyway my warmest personal regards and best wishes for your continued success in creative work.

Sincerely yours,



Donald A. Glaser  
Visiting Professor of  
Biophysics

DAG-pc

Letter from Donald A. Glaser, Visiting Professor of Biophysics, MIT, Jan 9, 1962

*Comments.* Donald Glaser won the Nobel Prize in Physics for the invention of the *bubble chamber*. Glaser joined the faculty of the University of California, Berkeley, in 1959, and received the Nobel Prize when he was there. Glaser lectured at Alladi Ramakrishnan's Theoretical Physics Seminar at Ekamra Nivas in August 1961 and had long discussions with the students. This letter was sent from MIT where Glaser was visiting in January 1962.

---



---

CALIFORNIA INSTITUTE OF TECHNOLOGY  
PASADENA

NORMAN BRIDGE LABORATORY OF PHYSICS

January 5, 1962

80

15:162

88

Professor Alladi Ramakrishnan  
University of Madras  
27, Luz, Mylapore  
Madras, India

Dear Professor Ramakrishnan:

I am sorry to say that I was away for the holidays, so your letter was only opened today, too late for me to cable you. Rest assured, however, that you have my good wishes and I hope you have great success.

Sincerely,



R. P. Feynman

RPF:n

Letter from R.P. Feynman, California Institute of Technology, Jan 5, 1962

*Comments.* Nobel Laureate Richard Feynman was one of the most eminent physicists of the twentieth century. His penetrating insight was admired by all. Alladi Ramakrishnan had the opportunity to meet Feynman in his office Caltech in 1956 and hear from the great man himself about how the electron travels back in time. Ramakrishnan was visiting the RAND Corporation in 1956 at the invitation of Richard Bellman, and it was Bellman who arranged this meeting with Feynman. Inspired by this meeting, and guided by his own intuition in probability, Ramakrishnan subsequently obtained a new simple proof of the equivalence of the Feynman and field-theoretic formalism by splitting the Feynman propagator into its real and imaginary parts. This paper\* appeared in the *Journal of Mathematical Analysis and Applications*, of which Bellman was the Editor-in-Chief.


In the Fall of 1962, at the invitation of Murray Gellmann, Ramakrishnan gave a colloquium at Caltech. Feynman attended Ramakrishnan's talk and the party in honor of Ramakrishnan at Gellmann's home in the hills of Altadena.

---


\*Alladi Ramakrishnan, "Some new topological features of Feynman graphs", *J. Math. Anal. Appl.*, **17** (1967), 68-71.

C-3.

**INVEST WISELY**  
Buy NATIONAL  
SAVINGS  
CERTIFICATES



**INDIAN POSTS AND TELEGRAPHS DEPARTMENT**



9 35 16

No. \_\_\_\_\_

Received here at \_\_\_\_\_ H. \_\_\_\_\_ M. 30/12

LT NIL 212 PRINCETON NJER 29 OCS 47 LT DR  
 ALLADI RAMAKRISHNAN EKAMRA NIVAS 27 LUZ MYLAPORE MADRAS

-- ON THE OCCASION OF THE INAUGURATION OF THE  
 INSTITUTE OF MATHEMATICAL SCIENCES I AM SENDING YOU  
 ALL GOOD WISHES FOR YOUR FUTURE WORK AS DIRECTOR OF A  
 NEW AND HIGHLY IMPORTANT CENTER OF SCIENTIFIC  
 RESEARCH-- BENGT STROMGREN--

The sequence of entries at the beginning of this telegram is—class of telegram, time handed in, serial number (in the case of Foreign Telegrams only), office of origin, date, service instructions (if any) and number of words.  
 This form must accompany any inquiry respecting this telegram.  
 L. C. & Sons, Calcutta—No. G 6/87 (M.F.F. Regn. No. 11/3/P-561—20-1-56)—(P-1/249A/65-66)—18-2-57—2,04,000—ks.

Telegram from BENGT STROMGREN, Institute for Advanced Study, Princeton, Dec 30, 1961

Comments. Bengt Stromgren, a noted Danish astronomer and astrophysicist, was appointed as the first Professor of Astrophysics at the Institute for Advanced Study in Princeton in 1957. There, he occupied the office of Albert Einstein who had died a little earlier. When Ramakrishnan visited the Institute for Advanced Study in Princeton in 1957–1958, he heard over 100 seminars, including those of Stromgren on astrophysics. Stromgren’s lectures were of particular interest to Ramakrishnan who a few years earlier had started publishing papers in the *Astrophysical Journal*, all of which were communicated by the great Indian astrophysicist Subrahmanyam Chandrasekhar.

C-3.



 INDIAN POSTS AND TELEGRAPHS DEPARTMENT
No. 361Received here at H M.

LT LF DR268 ROCHESTER NY 2 OCS 44

LT PROFESSOR ALLAD RAMAKRISHNA EXKAMARA NIVAS

27 LUZ MYLAPORE MADRASX

ON BEHALF OF UNIVERSITY OF ROCHESTER DEPT OF PHYSICS AND  
 ASTRONOMU EXTEND TO YOU PERSONALLY AND COLAGES IN NEW  
 INSTITUTE VERY BEST WISHES FOR DISTINGUSHIED CONRIBUTIONS  
 TO SCIENCE AND AHIMAN WELFARE EVERYWHERE = PROF MARSHAK =

The sequence of entries at the beginning of this telegram is—class of telegram, time handed in, serial number (in the case of Foreign Telegrams only), office of origin, date, service instructions (if any) and number of words.

This form must accompany any inquiry respecting this telegram.

L. C. & Sons, Calcutta—No. G 6/57 (M.F.P. Requ. No. 11/3/P-561—30-1-66)—(P-1/249A/55-56)—18-2-57—2,04,000 Bks.

Telegram from ROBERT MARSHAK, Professor of Physics, University of Rochester, Jan 3, 1962

*Comments.* Alladi Ramakrishnan's first contact with Professor Marshak was through the High Energy Physics Conference at Rochester that Marshak organized in 1956. Marshak invited Ramakrishnan to this conference during Ramakrishnan's first world tour of 1956. Marshak, a very eminent physicist, was also a great statesman for the discipline. He launched this successful series of conferences in high energy physics, and these were called the Rochester Conferences because there were initially held at the University of Rochester. Subsequently, these high energy physics conferences were held in different parts of the globe, and Marshak continued to be a key component in the conferences.

The 1956 Rochester conference had an enormous impact on Ramakrishnan. It exposed him to the latest advances in particle physics, and the significant research done in the United States. It was at this conference that Ramakrishnan met Robert Oppenheimer, Director of the Institute for Advanced Study, Princeton, and as a consequence of this meeting, Ramakrishnan received an invitation from Oppenheimer to visit the Institute for Advanced Study in 1957–1958. Finally, through this first meeting at the Rochester Conference, Ramakrishnan got to know Marshak quite well, and their friendship and mutual admiration grew over the years. Marshak who was not only an eminent scientist, but also someone who contributed to the profession with his administrative, organizational, and leadership skills, very much admired and appreciated Ramakrishnan's efforts in creating and leading MATSCIENCE. Marshak visited MATSCIENCE in January 1963 as the First Niels Bohr Visiting Professor and attended the First Anniversary Symposium of MATSCIENCE which was in his honor. He was very much impressed with the atmosphere of the new institute, vibrant with several eminent visiting scientists from abroad, and an enthusiastic group of faculty and students. In return, Marshak invited Ramakrishnan to the University of Rochester several times in the 1960s.

C-3.

**INVEST WISELY**  
Buy NATIONAL  
SAVINGS  
CERTIFICATES



INDIAN POSTS AND TELEGRAPHS DEPARTMENT

96

No. ....

Received here at 12/ H. .... M.

LT 1013 DR171 ROCHESTER NY 2 EST 57  
 LT PROFESSOR ALLADI RAMAKRISHNAN EKAMARA NIVAS  
 27 LUZMYLAPORE MADRASINDIA-  
 DELIGHTED TO HEAR OF CREATION OF INSTITUTE FOR MATHEMATICAL  
 SCIENCES WHICH WILL BE IMPORTANT STEP IN FURTHER DEVELOPMENT  
 OF INDIANA SCIENCE YOUR APPOINTMENT AS DIRECTOR PROMISES WELL FOR  
 SUCCESS OF INSTITUTE BE ASSURED OF MY BEST WISHES AND WILLINGNESS  
 TO HELP WHERE POSSIBLE - MCREA HAZLETT UNIVERSITY OF ROCHESTER


The sequence of entries at the beginning of this telegram is—class of telegram, time handed in, serial number (in the case of Foreign Telegrams only), office of origin, date, service instructions (if any) and number of words.  
 This form must accompany any inquiry respecting this telegram.  
 L. C. & Sons, Calcutta—No. G 617 (MFP. Regn. No. 11/312-561—36-1-56)—(P-1249A/55-56)—18-2-57—2,04,000

Telegram from MCREA HAZLETT, University of Rochester, Jan 3, 1962

*Comments.* McCrea Hazlett was Provost at the University of Rochester from 1961 to 1968. He was one of the last visitors to Alladi Ramakrishnan's Theoretical Physics Seminar in Madras in November 1961 just before the creation of MATSCIENCE. When Ramakrishnan visited the University of Rochester at the invitation of Professor Marshak in 1963, 1966, and 1967, Hazlett graciously hosted Ramakrishnan and his family. Hazlett visited Madras again in January 1964 with his family after MATSCIENCE was created and delivered the Second Anniversary Address of the Institute.

C.-3.

**INVEST WISELY**  
Buy NATIONAL  
SAVINGS  
CERTIFICATES



**INDIAN POSTS AND TELEGRAPHS DEPARTMENT**

No. .... 31

Received here at 4/57 H. .... M. 3-60

... LT 1155 DR172 SANTAMONIA CALIF 2 OCS 54  
LT ALLADI RAMAKRISHNAN EKAMRA NIVAS 27 LUZ MYLAPORE MADRAS

- WARMEST CONGRATULATIONS ON THE FOUNDING OF THE INSTITUTE  
OF MATHEMATICAL SCIENCES AND ON YOUR APPOINTMENT AS DIRECTOR  
I AM SURE THAT MUCH FRUITFUL AND CREATIVE WORK WILL EMERGE AND THAT  
IT WILL GREATLY CONTRIBUTE TO THE CULTURAL GLORY OF THE NEW AND  
DYNAMIC INDIA - BELLMAN -

The sequence of figures at the beginning of this telegram is—class of telegram, time handed in, serial number (in the case of Foreign Telegrams only), office of origin, date, service instructions (if any) and number of words.  
This form must accompany any inquiry respecting this telegram.  
L. C. & Indian Government. No. G 6/57 (MFP. Rem. No. 11/3/7-561-30-1-56)- (P-1/2492/55-56)- 13-2-57-2,04,000. K.

Telegram from RICHARD BELLMAN, Rand Corporation, Santa Monica, California, Jan 3, 1962

*Comments.* Richard Bellman, one of most well-known applied mathematicians, was a senior scientist at the famous RAND (acronym for Research and Development) Corporation, located in Santa Monica, a lovely suburb of Los Angeles. In 1949, when Alladi Ramakrishnan was doing his Ph.D. at the University of Manchester, he became aware of the fundamental work of Richard Bellman and Ted Harris. Subsequently, Bellman got interested in Ramakrishnan's work on *product densities* and invited Ramakrishnan to the RAND Corporation during Ramakrishnan's first round-the-world tour of 1956; Bellman was much impressed with Ramakrishnan's work on probability and suggested that Ramakrishnan contact the brilliant applied mathematician Peter Lax at the Courant Institute. Indeed, Lax invited Ramakrishnan for a colloquium at Courant that year. Bellman also arranged a meeting for Ramakrishnan with Richard Feynman at Caltech that year. Bellman and Ramakrishnan were very close. Bellman invited Ramakrishnan several times to California – on a major assignment to Rand in 1962, and subsequently to the University of Southern California in the 1970s after he (Bellman) moved there. Bellman founded the *Journal of Mathematical Analysis and Applications*, and Ramakrishnan contributed several fundamental papers to the journal from the 1950s to 2000 (the Millennium Bellman Memorial issue). Bellman had Ramakrishnan appointed as one of the Editors of the *Journal of Mathematical Analysis and Applications*. Ramakrishnan's only regret was that Bellman never visited Madras and he (Ramakrishnan) did not have an opportunity to host Bellman at MATSCIENCE and Ekamra Nivas.



INDIAN POSTS AND



TELEGRAPHIC DEPARTMENT

2/16/62 33-15

000294 2/1/62  
OFFICE STAMP  
INDIAN POSTS AND TELEGRAPH DEPARTMENT  
MADRAS

|                                      |                                 |                         |
|--------------------------------------|---------------------------------|-------------------------|
| Class }<br>Prefix } _____ Code _____ | N.o. _____                      |                         |
| Send. from _____                     | Sent at _____ H. _____ M. _____ |                         |
| By _____                             | To _____                        |                         |
|                                      | By _____                        |                         |
| Handed in at (Office of Origin)      | Date _____                      | Hour _____ Minute _____ |
|                                      | Service Instructions _____      |                         |
|                                      | Words _____                     |                         |

TO \_\_\_\_\_  
*correction to problem* H. \_\_\_\_\_ M. \_\_\_\_\_


LT I DR 96 STANFORDCALIF 1 OCS ~~OTD~~ WDS 46 I

+ LT PROF ALLADI RAMAKRISHNAN EKAMRA  
 NIVAS 27 LUZ MYLAPORE MADRAS MYLAPORE  
 === THE STANFORD UNIVERSITY DEPARTMENT  
 OF PHYSICS SENDS GREETINGS AND GOOD WISHES comp. the text  
 ON THE OCCASION OF THE INSTITUTE OF MATHEMATICAL  
 SCIENCES AND HOPES FOR A GREAT FUTURE OF  
 CREATIVE SCIENTIFIC WORK = K I SCHIFF =

W. D. The name of the recipient telegraphed, should be written after, but separated from, the text


Two page telegram from L.I. SCHIFF, Department of Physics, Stanford University, Jan 2, 1962

Comments. Leonard Schiff was one of the very few physicists who made notable contributions to almost every branch of physics. His book on *Quantum Mechanics* became the Bible in the field and was used by professors the world over to train their students. Indeed, Ramakrishnan lectured out of Schiff's Quantum Mechanics in Madras to his students. Schiff was Chairman of the Physics Department at Stanford University from 1948 to 1966, and so this telegram he sent was on behalf of the whole physics department. Schiff invited Ramakrishnan to visit Stanford for two weeks in 1962. Schiff came to MATSCIENCE in February 1963 and gave a series of lectures on gravitation. Ramakrishnan visited Stanford at the invitation of Schiff later in the 1960s on his annual scientific round-the-world trips.



INDIAN POSTS AND TELEGRAPH DEPARTMENT

37 38



60

42 3.1.62

|   |         |      |              |                      |              |
|---|---------|------|--------------|----------------------|--------------|
| Class   |         | Code |              | No.                  | C.           |
| Read. from  | Sent at |      | H            | M                    | Office Stamp |
| By  | To      |      |              |                      |              |
| By  |         |      |              |                      |              |
| Handed in at (Office of Origin)   | Date    | Hour | Minute       | Special Instructions | Words        |
| TO  |         |      | Send here as |                      |              |
| LT MD 17 CANBERRA SUB 2 OCS 72 LT PROF ALLADI RAMAKRISHNAN<br>EKAMRA NIVAS 27 LUZ MYLAPORE MMN<br>= HEARTIEST CONGRATULATIONS UPON THE CREATION OF YOUR NEW<br>INSTITUTE STOP THE VISION OF THE GOVERNMENT OF MADRAS AND<br>YOUR APPOINTMENT AS DIRECTOR WILL STIMULATE THOSE<br>BRANCHES OF MATHEMATICS AND THEORETICAL PHYSICS TO WHICH<br>INDIA HAS ALREADY CONTRIBUTED SO MUCH STOP I SEND<br>GREETINGS AND WARMEST GOOD WISHES FOR THE SUCCESS OF YOUR<br>GREAT VENTURE FROM MY COLLEAGUES AND MYSELF :<br>= MARK OLIPHANT NATUNIV = |         |      |              |                      |              |


Two page telegram from MARK OLIPHANT, Australian National University, Canberra, Jan 3, 1962

*Comments.* During his visit to Australia in 1954, Alladi Ramakrishnan met the eminent physicist Sir Mark Oliphant, a former associate of Lord Ernest Rutherford. Professor Oliphant was soon going to be visiting India under the auspices of The Royal Society, and so Alladi Ramakrishnan invited Oliphant to Madras to deliver the Rutherford Memorial Lecture. Oliphant came to Madras in January 1955 and stayed in Ekamra Nivas.

Some years later, Oliphant was appointed Governor of South Australia. In 1973, when Alladi Ramakrishnan was visiting different universities in Australia, he made a trip to Adelaide where he was the guest of Oliphant in the Governor's Mansion!


**INVEST WISELY**  
Buy NATIONAL SAVINGS CERTIFICATE

INDIAN POSTS AND



TELEGRAPHS DEPARTMENT

73 63



11-1-62

Class }  
Prefix } ..... Code (1)

Recd. from..... Sent at.....H.....M.....  
To.....  
By.....

No. C  
Office-stamp  
**058**

|                                 |      |      |        |                      |       |
|---------------------------------|------|------|--------|----------------------|-------|
| Handed in at (Office of Origin) | Date | Hour | Minute | Service Instructions | Words |
|---------------------------------|------|------|--------|----------------------|-------|

TO 5  
12/1/62

Recd. here at ..... H. .... M.

LT MK DR 111 BUDAPEST 110CS 69 LT PROF ALLADI RAMAKRISHNAN  
EKMARA NIVAS 27 LUZ MYLAPORE MADRAS =

I AM VERY GLAD THAT THE INDIA GOVERNMENT HAS ESTABLISHED IN  
MADRAS AN INSTITUTE TO FURTHER RESEARCH IN MATHEMATICS AND PHYSICA  
IN WHICH I AM SURE IMPORTANT RESULTS WILL BE OBTAINED STOP MAY I  
WISH YOU AND YOUR COLLEGUES GREAT SUCCESS AND IN PARTICULAR MAY I  
CONGRATULATE YOU ON YOUR APPOINTMENT AS FIRST DIRECTOR STOP

SINCERELY YOURS LAJOS = YOURS = LAJOS JANOSSY=

Two page telegram from LAJOS JANOSSY, Eotvos Institute, Budapest, Jan 12, 1962

*Comments.* Alladi Ramakrishnan first met Professor Janossy in the Winter in 1949 in Edinburgh, Scotland, at a conference on modern physics. Ramakrishnan was doing his Ph.D. at the University of Manchester under Professor Bartlett at that time. In Edinburgh, Ramakrishnan heard the lecture of Janossy and noticed strong connections between his own work on product densities and that of Janossy.\* Ramakrishnan was invited to talks at Dublin where Janossy was at that time. Janossy later returned to Hungary and became the Director of the Eotvos Institute in Budapest. After Ramakrishnan began the Theoretical Physics Seminar in Madras, he invited Professor Janossy to address the seminar and meet his students.

\*Alladi Ramakrishnan, "A note on Janossy's model of a nucleon cascade", *Proc. Cambridge Phil. Soc.*, **48** (1952), 451-456.



INVEST WISELY Buy NATIONAL SAVINGS CERTIFICATES



INDIAN POSTS AND



TELEGRAPHS DEPARTMENT

62/ 25

4-1-62

Class }  
 Prefix } \_\_\_\_\_ Code \_\_\_\_\_  
 No. 000308

Recd. from \_\_\_\_\_ Sent at \_\_\_\_\_ H \_\_\_\_\_ M. \_\_\_\_\_  
 To \_\_\_\_\_  
 By \_\_\_\_\_

Handed in at (Office of Origin) \_\_\_\_\_ Date \_\_\_\_\_ Hour \_\_\_\_\_ Minute \_\_\_\_\_ Service Instructions \_\_\_\_\_



LT NIL DR 186 CLEVELAND OHIO 3 OCS 54  
 T PROFESSOR ALLADI RAMAKRISHNAN EKAMARANIVAS 27 LUZ  
 MYLAPORE-MADRAS X :::

I LOOK FORWARD WITH GREAT INTEREST TO THE NEW AND CREATIVE IDEAS THAT WILL BE DISCOVERED BY YOUR COUNTRYMEN AT THE NEW INSTITUTE OF MATHEMATICS SCIENCES MY CONGRATULATIONS TO YOU ALLADI AND TO THOSE OF YOUR GOVERNMENT WHO HAVE MADE THE INSTITUTE POSSIBLE :: BAYARD RANKIN

Telegram from BAYARD RANKIN, Case Western University, Cleveland, Ohio, Jan 4, 1962

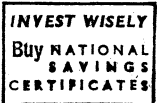
C-8.



31



14/ 18



INDIAN POSTS AND TELEGRAPHS DEPARTMENT

No. \_\_\_\_\_

Received here at \_\_\_\_\_ H \_\_\_\_\_ M.

XF R 13 CAMBRIDGE MASS 29 OCS 55 P EST  
 ALADI RAMAKRISHNAN EKAMARANIVAS  
 27 LUZ MYLAPORE MADRAS

MY WARMEST CONGRATULATIONS ON YOUR APPOINTMENT AS DIRECTOR OF THE INSTITUTE OF MATHEMATICAL SCIENCES AND MY BEST WISHES FOR THE SUCCESS OF THIS INSTITUTE WHICH I AM SURE WILL PLAY A MOST IMPORTANT ROLE IN THE DEVELOPMENT OF SCIENTIFIC RESEARCH IN INDIA SINCERELY

The sequence of  
 of Foreign Telegrams  
 This form must  
 L. C. & S.

BRUNO ROSSI MIT

Telegram from BRUNO ROSSI, Department of Physics, MIT, Dec 30, 1961

*Comments.* Alladi Ramakrishnan met Professors Rankin and Rossi at the Massachusetts Institute of Technology in 1956 where he gave talks in the Norbert Weiner Seminar. Rankin had received his Ph.D. in 1955 from Berkeley and his thesis was on stochastic processes and its uses in cascade theory which was an area in which Ramakrishnan had done considerable work. Thus it was natural for Rankin to be much interested in Ramakrishnan's work and invite him to a seminar at MIT. Rankin subsequently moved to Case Western University which is where he was when he sent the telegram. Rankin is known for the book "Differential space, quantum systems and prediction" that he edited with Norbert Weiner and published by the MIT Press in 1966.

Bruno Rossi was a famous experimental physicist who made notable contributions to cosmic rays and particle physics. He was interested in Ramakrishnan's work on cosmic rays. Rossi made his first major discoveries on cosmic rays in Florence, Italy in 1928. He moved to the United States in 1939 to escape the persecution of the fascist regime. He was first at the University of Chicago but then was appointed at MIT in 1946 as Professor of Physics. He was subsequently made Institute Professor at MIT in 1965. He was a Member of the National Academy of Sciences and won the Wolf Prize in Physics in 1987.

---

C-3

INVEST WISELY  
Buy NATIONAL  
SAVINGS  
CERTIFICATES

INDIAN POSTS AND TELEGRAPHS DEPARTMENT

No. 40

Received here at.....H.....M.

XF NIL R -190 PASADENA CALIF 29 OCS 64

PROFESSOR ALLADI RAMAKRISHNAN EKAMRA NIVAS  
27 LUZ MYLAPORE MADRAS INDIA -

PLEASE ACCEPT MY CONGRATULATIONS ON OCCASION OF THE  
FOUNDATION OF INSTITUTE OF MATHEMATICAL SCIENCES OF  
UNIVERSITY OF MADRAS AND ON YOUR APPOINTMENT AS ITS  
DIRECTOR MY COLLEAGUES AT THE MT WILSON AND  
PALOMAR OBSERVATORIES AND THE CALIFORNIA INSTITUTE OF  
TECHNOLOGY JOIN ME IN WISHING YOU THE BEST OF SUCCESS  
GUIDO MUNCH CALIFORNIA INSTITUTE OF TECHNOLOGY -

Telegram from GUIDO MUNCH, CALTECH, Pasadena, California, Dec 30, 1961

*Comments.* Guido Munch was an astrophysicist who worked with the great Subrahmanyam Chandrasekhar at the famous Yerkes Observatory outside Chicago in the 1940s. Munch hailed from Mexico. He received his Ph.D. from Chicago in 1946. Munch was on the faculty of the California Institute of Technology (Caltech) in Pasadena from 1953 onwards and was associated with both the Mt. Wilson and Palomar observatories outside Los Angeles. Munch was elected to the American Academy of Arts and Sciences in 1962.

Alladi Ramakrishnan became interested in applications of stochastic processes to astrophysics and therefore corresponded with Chandrasekhar in the early 1950s. During his first academic world tour of 1956, Ramakrishnan was traveling east bound, and therefore he met Guido Munch first in California and later Chandrasekhar in Chicago. Ramakrishnan published a series of papers in the *Astrophysical Journal* all communicated by Chandrasekhar who was the Managing Editor of that journal. Two of the papers\* were on an integral equation of Chandrasekhar and Munch. In his second academic round-the-world tour of 1962, when Ramakrishnan had an extended stay at the RAND Corporation in Santa Monica, he visited Mt. Wilson again at the invitation of Guido Munch.

\*A. Ramakrishnan and P.M. Mathews, "On an integral equation of Chandrasekhar and Munch", *Astrophys. J.*, **115** (1952), 141-144.

A. Ramakrishnan and P.M. Mathews, "On the solution of an integral equation of Chandrasekhar and Munch", *Astrophys. J.*, **119** (1954), 81-90.

In his telegram, Munch refers to the Institute of Mathematical Sciences as being part of the University of Madras. This is incorrect. MATSCIENCE was a separate institute, and still is, funded by the Department of Atomic Energy and the Government of Madras. Alladi Ramakrishnan was Professor of Physics at the University of Madras prior to being appointed Director of MATSCIENCE, and Munch must have been misled by that connection.

---

G-3

INVEST WISELY  
Buy NATIONAL  
SAVINGS  
CERTIFICATES



BRITTING  
GAMOW 33



INDIAN POSTS AND TELEGRAPHS DEPARTMENT

No. 99

24  
3  
25

mp/mg  
On  
16/12

Received here at \_\_\_\_\_ H. \_\_\_\_\_ M.

31/12

LT (81) DR 149 Boulder Colorado 30 Dec 33  
Prof. Alladi Ramakrishnan  
Ekamra Nivas 27 Luz Mysore Madras  
Best wishes to madras institute  
of mathematical, sciences and  
its first director from George  
Gamow and all physicists  
of the university of Colorado

The sequence of entries at the beginning of this telegram is—class of telegram, time handed in, serial number (in the case of Foreign Telegrams only), office of origin, date, service instructions (if any) and number of words.

This form must accompany any inquiry respecting this telegram.

L. C. & Sons, Calcutta—No. G 6/87 (MFP, Regn. No. 11/3/P-861—30-1-56)—(P-1/249A/55-56)—18-2-57—2,04,000 Bks.

Telegram from GEORGE GAMOW, Department of Physics, University of Colorado at Boulder, Dec 31, 1961

Comments. Professor George Gamow lectured in Alladi Ramakrishnan's Theoretical Physics Seminar in December 1959. Gamow was not only an eminent researcher in physics, but also a great expositor, who by his books reached out to students of all ages.

Ramakrishnan visited the University of Colorado several times in the 1960s to participate in physics conferences and also to deliver colloquia in the physics department there. Thus he had several contacts in the physics department at Boulder.

D. C. Chandrai & Co.-PI/67A/85-56-(Part II)-8-5-56-No. II P-176-28-7-58-1,00,000 Bks.



INDIAN POSTS AND



TELEGRAPHS DEPARTMENT

3/20

|                                      |                           |                                 |        |
|--------------------------------------|---------------------------|---------------------------------|--------|
| Class }<br>Prefix } _____ Code _____ | No. _____                 |                                 | C.     |
| Recd. from _____                     | Sent at _____ H. _____ M. | Stamp.                          |        |
| By _____                             | To _____                  | By _____                        |        |
| Handed in at (Office of Origin)      | Date                      | Hour                            | Minute |
| Service Instructions                 |                           | Words                           |        |
| TO                                   |                           | Recd. here at _____ H. _____ M. |        |



LT EE DR 347 WASHINGTONDC 27 DEC 61

LT PROFESSOR ALLADI RAMAKRISHNAN EKAMRA NIVAS  
27 LUZ MYLAPORE MADRAS INDIA

HEARTIEST CONGRATULATIONS ON THE INAUGURATION OF  
THE INSTITUTE OF MATHEMATICAL SCIENCES IN MADRAS PERIOD  
I AM CONFIDENT THAT IT WILL SHINE LIKE A BEACON  
ILLUMINATING THE PROGRESS OF PHYSICS IN INDIA PERIOD  
THE PROVINCE AND THE UNIVERSITY ARE TO BE  
CONGRATULATED ON PERSUADING SO ABLE A SCIENTIST TO  
SERVE AS DIRECTOR PERIOD = DOCTOR MORRIS M SHAPIRO  
SUPERINTENDENT NUCLEONICS DIVISION NAVAL RESEARCH  
LABORATORY

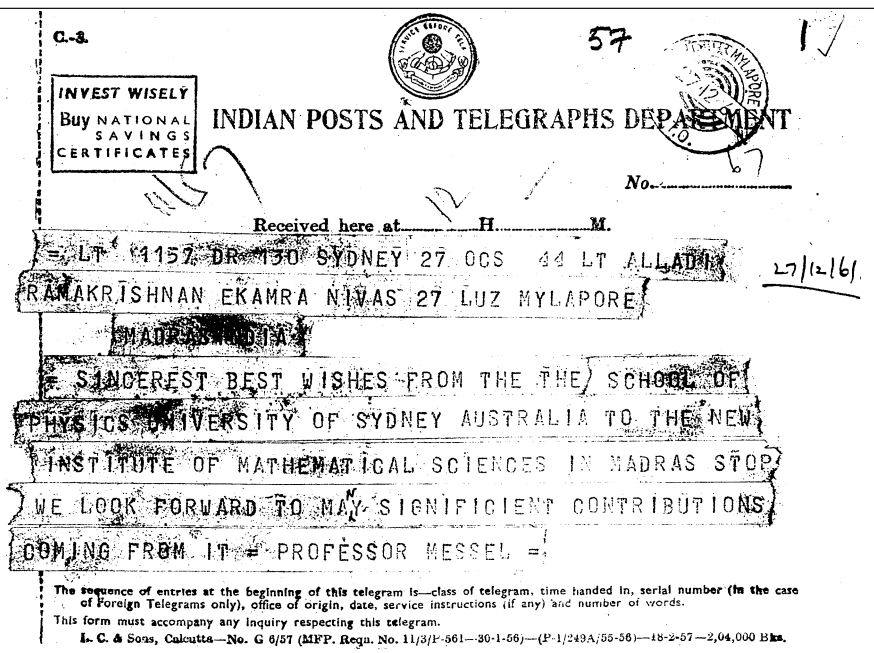
Two page telegram from MAURICE M. SHAPIRO, Superint – Nucleonics Division, Naval Research Laboratory, Dec 28, 1961

*Comments.* Maurice Shapiro was a veteran of the famous Manhattan Project directed by Robert Oppenheimer at Los Alamos. He had a long and distinguished career in the field of cosmic rays and neutrino astrophysics. He founded the Cosmic Ray Laboratory at the Naval Research Laboratories in Washington, DC, and was there for the remainder of his life. Shapiro was very much interested in Alladi Ramakrishnan's work on cosmic rays. When Ramakrishnan was on a round-the-world scientific trip in 1956, Shapiro invited him to lecture at the Naval Research Labs. Thus began the fruitful contact with Shapiro. In 1957–1958, when Ramakrishnan was visiting the Institute for Advanced Study in Princeton, Shapiro invited him to lecture in Washington, DC.

Shapiro visited India in 1961 and lectured at the Theoretical Physics Seminar. When Alladi Ramakrishnan introduced Shapiro to the Minister for Education Mr. C. Subramaniam, Shapiro told the Minister how impressed he was with the theoretical physics seminar, and that it would be in the best interests of Indian science to start a new institute as envisioned by

Ramakrishnan. Shapiro told Subramaniam that watching the students at work in Ekamra Nivas reminded him of the manner in which scientists gathered round Oppenheimer at Los Alamos! That was a high and generous tribute which made a great impression on Subramaniam. Shapiro went on to suggest that the students should meet the Prime Minister of India. Thus Shapiro's input was crucial in the launching of MATSCIENCE. Shapiro visited MATSCIENCE in December 1963 and also in the 1970s. After the creation of MATSCIENCE, Alladi Ramakrishnan visited the United States annually, and Shapiro regularly invited him to lecture at the Naval Research Labs.

---



Telegram from HARRY MESSEL, University of Sydney, Dec 28, 1961

*Comments.* Alladi Ramakrishnan first met Harry Messel in Dublin in the winter of 1949 when he went there at the invitation of Professor Janossy to deliver a lecture on stochastic processes. Ramakrishnan was, at that time, a Ph.D. student at the University of Manchester working under Professor M.S. Bartlett. Messel was working under Janossy. Ramakrishnan and Messel became very good friends and had common research interests on cosmic rays. Messel then went to Australia where he took a permanent position in Sydney. After Ramakrishnan returned to India from England and was at the University of Madras, Messel invited him to Sydney in 1954. In return, Ramakrishnan invited Messel to Madras and to the Theoretical Physics Seminar in 1957.

It was the notes for the lectures that Alladi Ramakrishnan gave at Sydney that became the basis of his *Handbuch der Physik* article of 1959. Also, the lectures in Sydney led Alladi Ramakrishnan to novel interpretations of integrals of random functions.\*

\*Alladi Ramakrishnan, "Phenomenological interpretation of the integrals of a class of random functions", *Proc. Koninkl. Netherlands Akad.* **58** (= *Indag. Math.*, **17**) (1955), 470-482.  
Alladi Ramakrishnan, "Phenomenological interpretation of the integrals of a class of random functions - II", *Proc. Koninkl. Netherlands Akad.* **58** (= *Indag. Math.*, **17**) (1955), 634-645.  
Alladi Ramakrishnan, "Processes represented as integrals of a class of random functions", *Proc. Koninkl. Netherlands Akad.* **59** (= *Indag. Math.*, **18**) (1956), 121-127.



C-3.

INVEST WISELY  
Buy NATIONAL  
SAVINGS  
CERTIFICATES



INDIAN POSTS AND TELEGRAPHS DEPARTMENT

No. 360

Received here at ..... H. .... M.

XF OL R397 GENEVE RSQ 3 OCS 41

ALLADI RAMAKRISHNAN 27 LUZ MADRAS-4 INDIA=

DELIGHTED HEAR TO CREATION INSTITUTE OF MATHEMATICAL SCIENCES AND  
YOUR APPOINTMENT STOP THIS IS GOOD NEWS FOR FUTURE OF SCIENCE IN  
INDIA STOP CONGRATULATIONS AND BEST WISHES FOR FUTURE FROM ALL  
AT CERN = WEISSKOPF CERNLAB =

The sequence of entries at the beginning of this telegram is—class of telegram, time handed in, serial number (In the case of Foreign Telegrams only), office of origin, date, service instructions (if any) and number of words. This form must accompany any inquiry respecting this telegram.  
L. C. & Sons, Calcutta—No. G 6/57 (MFP. Requ. No. 11/3/1-561—30-1-56)—(P-1/249A/55-56)—18-2-57—2,04,000 lks.

Telegram from VICTOR WEISSKOPF, CERNLAB, Dec 28, 1961

Comments. Victor Weisskopf, a world renowned physicist, did his Ph.D. in 1931 under the guidance of Nobel Laureates Max Born and Eugene Wigner. He then proceeded to do his postdoctoral work with Nobel Laureates Werner Heisenberg, Erwin Schrodinger, Wolfgang Pauli, and Niels Bohr. He worked on the Manhattan project that produced the atom bomb. After World War II, he was Professor at MIT. Among his students, there was Murray Gell-Mann who later won the Nobel Prize. During 1961-1966, Weisskopf was Director-General of CERN outside Geneva in Switzerland where one of the famous accelerators is located. Weisskopf was awarded the National Medal of Science in 1972 and the Wolf Prize in 1981.

Alladi Ramakrishnan interacted with Weisskopf when he attended international high energy physics conferences in Rochester in 1956 and elsewhere. This telegram was sent by Weisskopf when he was Director-General in CERN. Alladi Ramakrishnan visited CERN several times in the 1960s, first in 1960, and later in 1965 when Weisskopf was Director-General there. Weisskopf visited MATSCIENCE in January 1964 and inaugurated its Second Anniversary Symposium.

---

THE INSTITUTE FOR ADVANCED STUDY  
PRINCETON, NEW JERSEY

SCHOOL OF MATHEMATICS

January 9, 1962

81  
Professor Alladi Ramakrishnan  
Ekamra Nivas  
27, Luz, Mylapore  
Madras, India

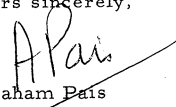
Dear Professor Ramakrishnan:

As I have been away on vacation for a few weeks, it is only now that I find your kind letter of December 23, and I profoundly regret not to have been able to answer you earlier.

I would like to send you all my good wishes for the new Institute of Mathematical Sciences of which you are to be the first Director. I have every hope and expectation that this Institute may further increase the important role which mathematicians and mathematical physicists of your country have played and are playing in science. I would expect that there are vast and untapped sources of spiritual energy in your country which can be made to flourish by Institutes of the kind which the Government of Madras has now decided to sponsor.

I shall be glad to be of help in any way to your Institute. With kindest regards,

Yours sincerely,

  
Abraham Pais

AP:jp




P.S. Would you kindly inform me of the precise address of the Institute so that scientific material of our Institute can be sent there.

Letter from Abraham Pais, The Institute for Advanced Study, Princeton, Jan 9, 1962

*Comments.* Abraham Pais, a very eminent physicist, was also a science historian. Born in the Netherlands, Pais did his doctoral work under the world renowned L. Rosenfeld at Utrecht. His Ph.D. work attracted the attention of Niels Bohr and Pais served as Bohr's assistant for a few years. In 1947, Pais moved to the Institute for Advanced Study in Princeton where he became Albert Einstein's colleague. Among his major contributions to physics was his explanation of certain puzzling properties of strange particles, which together with the ideas of Murray Gell-Mann led to formulation of the quantum number called *strangeness*.

Alladi Ramakrishnan visited the Institute for Advanced Study in 1957-1958 and got to know Abraham Pais quite well. Ramakrishnan attended over 100 seminars at the Institute including those of Pais and Sam Treiman on weak interactions.

---

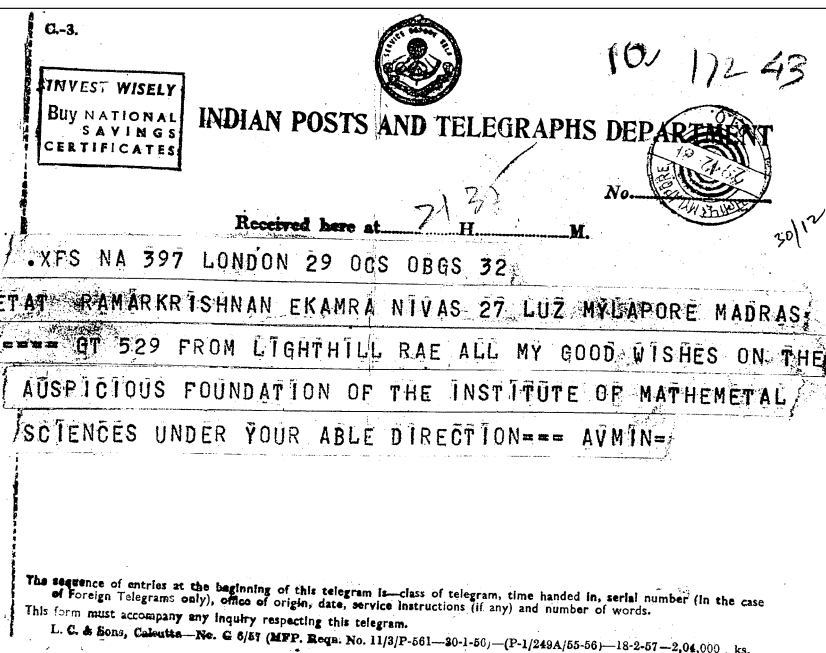
|  |  |   |      |  |
|--|--|---|------|--|
|   |  |  |      | 50<br>44   |
| Class Prefix <b>LT</b> Code <b>DD</b>  |  | No.   |      |  |
| Recd. from <b>18/12/61</b>   |  | Sent at ..... H. .... M.  |      | Office-stamp   |
| By .....   |  | To .....  |      | By .....   |
| Handed in at (Office of Origin) <b>DR 168 Kobenhavn</b>  |  | Date <b>28</b>  | Hour | Minute   |
|  |  | Service Instructions <b>005</b>   |      | Words <b>25</b>  |
| TO   |  | Recd. here at 9 H. — M.   |      |  |
| <b>LT Prof Ramakrishnan 27 Luz</b><br><b>Villa Mylapore Madras</b><br><b>= heartiest congratulations and</b><br><b>very best wishes of long</b><br><b>fruitful activity to your</b><br><b>Institute and its <del>dear</del> director</b><br><b>= Rosenfeld</b> |  |   |      |  |

MGFN 80 P&amp;T NK/61-25-5-61-4, 62, 500 B.H.

Telegram from L. ROSENFELD, Dec 29, 1961

*Comments.* Leon Rosenfeld, a Belgian physicist from Liege, was a collaborator of Niels Bohr. He did very fundamental work in quantum electrodynamics predating Dirac. Rosenfeld succeeded George Uhlenbeck as professor of theoretical physics at the University of Utrecht in Holland in 1940. Among his notable students at Utrecht was physicist Abraham Pais. In 1947, he was appointed as Professor of Theoretical Physics at the University of Manchester. After serving in Manchester, he moved to Copenhagen.

As a Ph.D. student at the University of Manchester during 1949–1951, Alladi Ramakrishnan heard a course of lectures by Professor Rosenfeld on nuclear physics. Later Ramakrishnan met Rosenfeld in Copenhagen during a visit to the Copenhagen Institute in 1960 at the invitation of its Director Niels Bohr. Rosenfeld visited MATSCIENCE as the First Niels Bohr Visiting Professor in 1963–1964. Upon arrival in Madras, Rosenfeld said that he had not attended a single conference in Europe in the last year without meeting someone or the other who had not visited, or was planning to visit, MATSCIENCE. Such was the flow of visiting scientists even in the very first years of the Institute. Professor Rosenfeld expressed surprise at the very small and humble accommodations of the Institute which had such an outstanding program of visitors. Rosenfeld delivered the opening lecture of the Second Anniversary Symposium of MATSCIENCE and spoke about Bohr's contribution to twentieth century physics.



Telegram from M.J. LIGHTHILL, Royal Aircraft Establishment, Farnborough, England, Dec 29, 1961

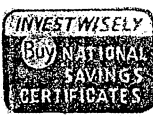
*Comments.* Sir James Lighthill (FRS) was one of the most eminent and productive applied mathematicians in England. He held the Beyer Chair at the University of Manchester during 1946–1949. As a Ph.D. student in 1949–1951 at Manchester, Alladi Ramakrishnan heard over 100 lectures of Lighthill on methods of mathematical physics. In return, Lighthill appreciated Ramakrishnan’s new method of product densities in the theory of probability. Thus began a long friendship between Ramakrishnan and Lighthill and the two had a mutual admiration for their research and professional contributions.

Lighthill was one of the last visitors to the Theoretical Physics Seminar in November 1961, and endorsed the creation of MATSCIENCE when he met Education Minister C. Subramaniam at Ekamra Nivas. He was at that time Director of the Royal Aircraft Establishment in Farnborough from where this telegram was sent. AVMIN in the telegram probably refers to Aviation Ministry.

Lighthill was an acknowledged world authority on aeroacoustics and fluid mechanics. His recognitions include the Royal Medal (1964) and the Copley Medal (1998). In 1964, he was appointed as the Royal Society Resident Professor at Imperial College, London. At his invitation, Alladi Ramakrishnan visited Imperial College in 1965 and 1969. Lighthill later served as Lucasian Professor of Mathematics at Trinity College, Cambridge. Ramakrishnan visited Lighthill in Cambridge in 1975 and he (Lighthill) arranged a meeting for Ramakrishnan with Professor Alan Baker at Cambridge; Ramakrishnan wanted to meet Baker to get his advice for Krishna who in 1975 was going to UCLA for his Ph.D. When Lighthill retired from the Lucasian Professorship in 1979, that Chair was filled by Stephen Hawking. Lighthill then became Provost of the University College, London.

Lighthill founded the Institute of Mathematics and its Applications (IMA) in 1964, a professional body for mathematicians, and a learned society in England.

31 ✓ 47  
2.1.62



INDIAN POSTS AND TELEGRAPHS DEPARTMENT



000101

Class }  
 Prefix } Code \_\_\_\_\_ No. \_\_\_\_\_

Recd. from \_\_\_\_\_ Sent at \_\_\_\_\_ H. \_\_\_\_\_ M. \_\_\_\_\_  
 By \_\_\_\_\_ To \_\_\_\_\_  
 By \_\_\_\_\_

Handed to at (Office of Origin) \_\_\_\_\_ Date \_\_\_\_\_ Hour \_\_\_\_\_ Minutes \_\_\_\_\_



LT R DR 88 OXFORD 1 OCS CORRECTION TO FOLLOW 106

LT PROFESSOR ALLADI RAMAKRISHNAN DIRECTOR THE INSTITUTE  
 OF MATHEMATICAL SCIENCES EKAMRA NIVAS 27 LUZ  
 MYLAPORE MADRAS X

ON BEHALF OF THE PRESS AND OUR MANY DISTINGUISHED

N. B.—The name of the sender, if telegraphed, should be written after, but separated from, the text.

INTERNATIONAL EDITORIAL BOARDS I SEND YOU OUR  
 CONGRATULATIONS AND BEST WISHES FOR THE FUTURE ON THE  
 OCCASION OF THE INAUGURATION OF YOUR NEW AND IMPORTANT  
 INSTITUTE STOP THE ESTABLISHMENT OF THIS INSTITUTE SHOWS  
 GREAT VISION AND BODES WELL FOR THE NEW SCIENTIFIC GENERATION  
 OF INDIA STOP I AM CERTAIN THAT THE CREATIVE WORK

N. B.—The name of the sender, if telegraphed, should be written after, but separated from, the text.

THAT WILL BE DONE AT YOUR INSTITUTE WILL NOT ONLY BENEFIT  
 INDIA BUT THE WORLD STOP WILL KINDEST REGARDS AND BEST WISHES  
 S MAXWELL PUBLISHER AT PERGAMON PRESS

N. B.—The name of the sender, if telegraphed, should be written after, but separated from, the text.

Three page telegram from CAPTAIN MAXWELL, Pergamon Press, Oxford, England, Jan 2, 1962

Comments. With so much work done by Alladi Ramakrishnan and his students on the method of product densities and its applications, as well as on the theory of elementary particles and cosmic rays starting from 1950, it was only natural for Ramakrishnan to consider writing a book. The first was a comprehensive 1959 survey by Ramakrishnan in the *Handbuch der Physik* (Springer) on probabilistic methods stemming from product densities. This was

followed by the book “Elementary particles and cosmic rays” published by the Pergamon Press, Oxford, in late 1962. During his 1960 trip to Europe, Alladi Ramakrishnan visited Oxford University at the invitation of Professor D.G. Kendall, and during this visit he met Captain Maxwell, Publisher of the Pergamon Press in London. Maxwell then extended a book contract to Ramakrishnan. The book came to print in late 1962 after the birth of MATSCIENCE.

---

# The Miracle has Happened

Alladi Ramakrishnan

*My father Professor Alladi Ramakrishnan was a master of exposition, both in written and spoken form. He was a dynamic speaker, an orator in every sense. Right from my boyhood, I had the pleasure to listen to many of his scientific lectures and speeches, and was inspired by his manner of speaking and the power of his oratory. The finest speech he ever gave was perhaps at the inauguration of MATSCIENCE, the Institute of Mathematical Sciences, on January 3, 1962, when his thoughts came pouring out at that very exciting and memorable occasion. As a six year old boy, I was in the front row of the English Lecture Hall at the Presidency College in Madras when he delivered that speech extempore, as was his custom. The speech was later written up from a tape recording. This speech was printed in the appendix to The Alladi Diary, Vol I, East–West Books, Madras, (2000). It is reprinted here with the permission of East–West Books, Madras, Pvt. Ltd.*

*Krishnaswami Alladi*

## Alladi Ramakrishnan's Speech on the Inauguration of the MATSCIENCE Institute

So the miracle has happened. By the Grace of God and the will of man, a new situation has been brought into being which augurs to be the starting point of an intellectual renaissance, the nature and magnitude of which cannot be foreseen at the present time. It is incredible that a series of events, each as improbable as the other, should have taken place in such steady and rapid succession. It is as though a chapter of a book of fairy tales has been transmuted into real life, and I feel like one who wakes up from a dream to find reality stranger than fantasy.

---

K. Alladi

Department of Mathematics, University of Florida, Gainesville, FL 32611, USA

e-mail: [alladik@ufl.edu](mailto:alladik@ufl.edu)

The dream is so chaste that I have the courage to ask all those present here to share it with me. It originated five years ago in the exotic atmosphere of the quaint old town of Kyoto in Japan where I spent six weeks at the invitation of Professor Yukawa. In the 'domestic' environment of the Yukawa Hall, young Japanese physicists, the hope and pride of their country, just resurrected from the second World War, gathered together in enlightened leisure to discuss the most abstruse problems of modern physics. That strange enchantment drew me into the domain of elementary particle physics, and I played with the idea of creating something like the Yukawa Hall in my own home town where my great father made his legendary reputation in another field of intellectual activity.

The enchantment became a passion when a fortuitous circumstance took me to the New World, and I had the opportunity to attend the Conference on High Energy Physics at the University of Rochester in the spring of 1956. Within four days, I was brought face to face with the rising generation of American physicists. One had only to listen to Gell-Mann and Chew, Feynman and Goldberger, to realise that a new era in American physics had been ushered in. American institutions no longer depended on the guidance of European scientists as they did a decade ago, when due to the chance of war, they were able to offer hospitality to European physicists like Fermi, Segre, and Bethe. American physics leapt from infancy to manhood within this decade, and it has now become almost a necessity for European physicists to spend some time in the great American institutions and in the laboratories where things are happening every day and every hour. I felt that such a transformation needs to come in my own country which despite its organised efforts in scientific research has yet to take a place in creative science.

I therefore tried to analyse the causes for our failure. There has always been the conventional argument that there was not enough talent in the country which is not borne out by facts. It is a tragedy too deep for tears that we do not take cognisance of talent or creative work unless it has received recognition outside our frontiers. Sometimes the wait is too long, the response so cold, that it freezes up the all too frail impulses for academic life in our country. What we need is a new generation of scientists, impatient for opportunities, intolerant of mediocrity, full of action, full of manly pride, and friendship like their compeers in the new world, who have not only faith in their powers, but in the scientific progress of their country.

I was strengthened in this faith during my stay, at the kind invitation of Professor Oppenheimer, in Princeton, where the most gifted minds in mathematical sciences gather together every year in an atmosphere exhilarating for creative work. It was a momentous year when the work of Yang and Lee marked the greatest advance in physical thought since the birth of quantum mechanics in 1926. I held a watching brief as a representative of our unborn Institute and I returned from Princeton with no other thought dominating my mind except to reproduce, in a small measure at least, the atmosphere for such creative work. Chance and circumstance came to my favour when a small band of students, stricken by the same splendid sickness, gathered round me in goodly friendship. We had no resource at our command except the love of common excitement for doing something new. To this fraternity, we gave the name – "Theoretical Physics Seminar". It was located in my family



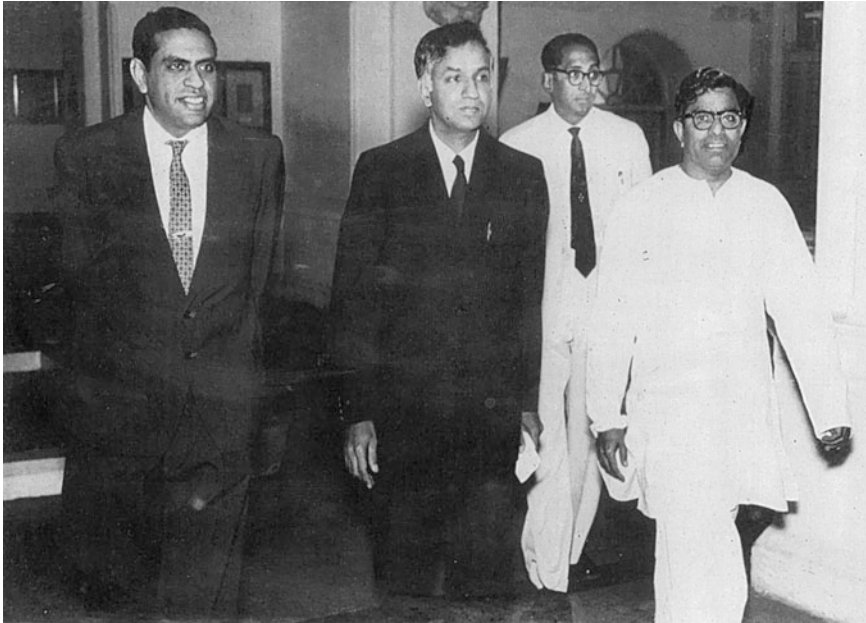
home with the consent of my gracious wife. We met in leisured comfort and indulged in the impertinence of attempting to work on the same type of problems as are engaging the attention of theoretical physicists elsewhere. We were encouraged in our efforts by the frequent visits of famous physicists whose friendliness and cooperation were our only sources of strength and sustenance. What a fine hour it was when Bohr and Salam who span the growth of modern physics from atomic physics to gauge theories of elementary particles, evinced an interest which gave us the strength to hope when we were all alone and everything seemed so near despair. We waited and watched for something to happen.

It was one of the fortunate moments of my life when I met the Finance Minister one evening at a gathering of international students. It puzzled me beyond comprehension to find the Minister, who must be more concerned with building dams and bridges, getting interested in the development of mathematical research. I felt a trifle guilty that I had inveigled him into this domain which had intoxicated me and my associates beyond reason. Soon I realised that it became almost a faith with him, a faith which was strengthened by his recent visit to the United States. He returned with the conviction that creative science needed the noble heat of youthful ambition and not the tepid caution of unfeeling mediocrity. Before proceeding to take steps for the creation of an institute for advanced learning, he was anxious to have the blessings and active support of our Prime Minister. It occurred to him in a discussion with my esteemed and genial friend Dr. M. M. Shapiro that all students associated with me should be introduced to the Prime Minister during his visit to Madras. The impression they made on our Prime Minister was more due to his generosity than to their own achievements. It was his wide humanity and deep concern for the prosperity of our country that made him see the light of hope even in the feeble efforts of smaller men. His support by agreeing to be our Patron, gave that final impulse which resulted in the setting up of this Institute.

The final act in this strange dream is even more fantastic than the events that preceded it. I approached Professor S. Chandrasekhar, one of the greatest astrophysicists of our time, who stands so high above the rest of our own common mould, with a request that he should associate himself with the new Institute. It was an insolence on my part to do so when I was assuming the Directorship of the Institute. I suppose you will excuse me for this if I assure you that the spirit in which I did so was animated by that in the greatest of legends when Arjuna approached Lord Krishna for his support. It was accepted with that same legendary grace, and the Institute has honoured itself by his association with it. This band of students, this firstlings of the fold, must consider themselves to be the happy few to have chosen him as their guide.

This then is the genesis of this new Institute, which symbolises the hopes and ideals of the entire scientific community in India. The Government of Madras and in particular the Chief and Finance Ministers ably assisted by the Education Secretary, another victim of the splendid sickness, must be congratulated for the most gracious gesture that has ever been made by any administrative authority to the academic community in our country. The best tribute we can pay to our government is to say, "it does not seem to be the red tape – it is the blue riband." Is it not natural that

greetings have poured from scientists all over the world, from California in the west to Sydney in the east? To those scientists who visited Madras, whose very presence had introduced the heady atmosphere of Berkeley into the placid environs of my family home, we are deeply grateful, for they kept alive the state of hope till the moment of its realisation. As for myself, it is a period of thanksgiving to my great teachers Professor Bhabha and Professor Bartlett, who initiated me into theoretical physics. My only regret is that my parents whose home nursed the happy breed, are not alive today at the crucial moment of my academic life. In recompense, I shall pass on to my students their message that the pursuit of science is at its best when it is a part of a way of life. That is the ideal to which this institute is dedicated.



(l to r) Alladi Ramakrishnan, Subrahmanyam Chandrasekar (University of Chicago), and Minister of Education Mr. C. Subramaniam walking to the dais for the inauguration of MATSCIENCE, The Institute of Mathematical Sciences – 3 Jan 1962



Professor Alladi Ramakrishnan delivering his speech “The Miracle has happened” at the inauguration of MATSCIENCE, The Institute of Mathematical Sciences – 3 Jan 1962. The inaugural function was held at the Old English Lecture Hall of the Presidency College of the University of Madras

# Overseas Trips of Alladi Ramakrishnan

## Krishnaswami Alladi

Professor Alladi Ramakrishnan believed that close interaction with the leading scientists around the world was essential for fundamental research. He traveled annually to present his work at conferences and universities worldwide. He has lectured at more than 30 international conferences, and given talks at about 200 centers of higher learning in North America, Europe, Asia, and Australia. In doing so, not only did he disseminate his research work and those of his group, but also used it as an opportunity to make new contacts and invite active researchers to Madras. Among his many trips abroad, we mention a few that were especially significant in terms of his career.

**1949–1951** Visit to England for PhD at the University of Manchester under M. S. Bartlett and D. G. Kendall (Oxford); method of product densities recognized in England.

**1954** Trip to Australia: Lectures at the University of Sydney at the invitation of Harry Messel became the basis of the *Handbuch der Physik* article five years later.

**1956** Round-the-World Trip: Visit to Yukawa Hall, Kyoto; met Richard Bellman and Nobel Laureate Richard Feynman in California (the beginning of a long association with Richard Bellman); met astrophysicist Subrahmanyan Chandrasekar in Chicago; attended High Energy Physics Conference in Rochester organized by Robert Marshak and met Robert Oppenheimer there; lectured at the Max Planck Institute in Göttingen, Germany at the invitation of Nobel Laureate Werner Heisenberg, which resulted in a contract with Springer for an article on probability in the *Handbuch der Physik*.

**1957–1958** Visit to the Institute for Advanced Study in Princeton at the invitation of its Director Robert Oppenheimer, after which the focus of Alladi Ramakrishnan's research turned from stochastic processes to elementary particle physics.

**1960** Trip to Europe: visited the Copenhagen Institute of Physics at the invitation of Nobel Laureate Niels Bohr; series of lectures on stochastic processes at the University of Berne at the invitation of Andre Mercier; visited Imperial College,

---

K. Alladi

Department of Mathematics, University of Florida, Gainesville, FL 32611, USA

e-mail: [alladik@ufl.edu](mailto:alladik@ufl.edu)

London, as the guest of Abdus Salam; while in London was offered a contract by the Pergamon Press to write a book on Elementary Particles and Cosmic Rays which was published in 1963.

**1962** First Round-the-World trip after assuming Directorship of MATSCIENCE: two month visit to Rand Corporation in California at the invitation of Richard Bellman; lectured at Caltech at the invitation of Murray Gellmann; spent two weeks each at Stanford University at the invitation of Leonard Schiff and at the University of California, Berkeley.

**1964** Visit to Russia: lectured at the High Energy Physics Conference in Dubna.

**1965** Trip to Europe: two month visit to the International Centre for Theoretical Physics (ICTP) in Trieste, Italy at the invitation of its Director Abdus Salam; visits to CERN in Geneva, Saclay, Orsay, and Institute Henri Poincare in Paris.

**1968** Round-the-World trip, and visit to Europe: Lectured at Cornell University at the invitation of Nobel Laureate Hans Bethe (contact with Bethe resulted in Bethe visiting MATSCIENCE in 1969); visited Moscow and Leningrad as guest of the Russian Academy of Sciences; lectured at the High Energy Physics Conference in Vienna.

**1971** Visit to New Zealand: talk at the Rutherford Centennial Conference, Christchurch.

On most of his trips abroad from 1962 onwards, his wife Lalitha and his son Krishna accompanied him. It was thus a great academic experience for Krishna from an early age. We present here a sample of some overseas photographs of Professor Alladi Ramakrishnan.



Alladi Ramakrishnan in front of Fulld Hall, Institute for Advanced Study, Princeton, Fall 1957





Alladi Ramakrishnan in Europe during a trip in 1960



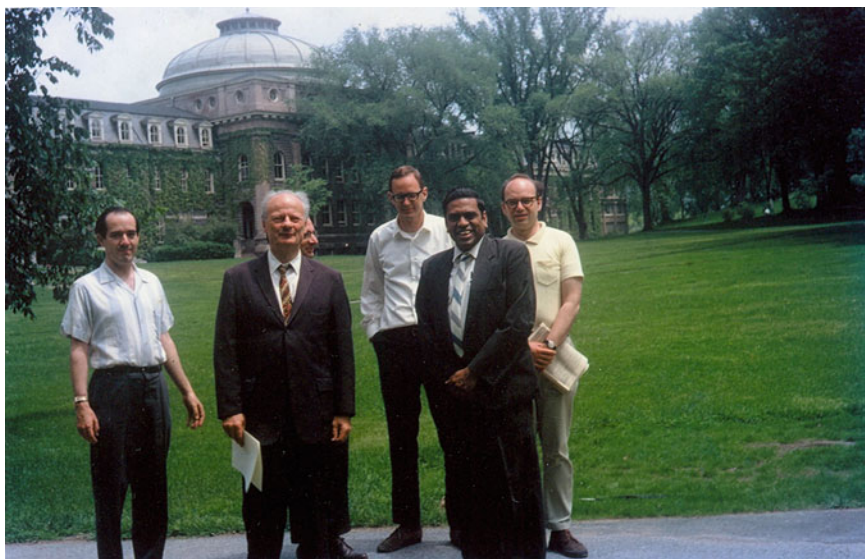
Alladi Ramakrishnan with Nobel Laureate Niels Bohr at Bohr's home in Copenhagen – 1960



Prof. Alladi Ramakrishnan at a conference at the International Centre for Theoretical Physics in Trieste, Italy – 1968



Mrs. Lalitha Ramakrishnan and Krishna who accompanied Prof. Ramakrishnan on his scientific tours, often attended the conference lectures as well – 1968



Professor Alladi Ramakrishnan with Nobel Laureate Hans Bethe and other physicists at Cornell University – 1968



Alladi Ramakrishnan in attendance at the International Conference on High Energy Physics in Vienna – July 1968



**The following are recent pictures taken at the home of Mathura and Krishna Alladi in Gainesville, Florida**



Mathura Alladi, Lalitha Ramakrishnan and Alladi Ramakrishnan in discussion with National Academy of Sciences Member Richard Askey (University of Wisconsin) – March 2005



Alladi Ramakrishnan in discussion with 1994 Fields Medalist Efim Zelmanov (UC San Diego) who delivered the Tenth Erdos Colloquium – January 2008



Alladi Ramakrishnan in discussion with the eminent combinatorialist Dominique Foata (Univ. Strasbourg, France) who delivered the Tenth Ulam Colloquium. Mrs. Ramakrishnan looks on – February 2008



Alladi Ramakrishnan and Krishna Alladi with National Academy of Science Member and Ramanujan Expert George Andrews (Penn. State Univ.) who is Distinguished Visiting Professor at Florida each spring – February, 2008



Alladi Ramakrishnan in discussion with 1970 Fields Medalist John Thompson during a party at the Alladi House in Gainesville – April 2008. Thompson, who is Graduate Research Professor at the University of Florida, won the Abel Prize in May 2008



Alladi Ramakrishnan in discussion with Professor Bertram Kostant (MIT), who delivered the Center for Applied Mathematics Colloquium – April 2008



# List of Publications of Alladi Ramakrishnan

- 1) (with H. J. Bhabha) “The mean-square deviation of the number of electrons and quanta in cascade theory”, *Proc. Indian Acad. Sci.*, **32** (1950), 141–153.
- 2) “Stochastic processes relating to particles distributed in a continuous infinity of states”, *Proc. Camb. Phil. Soc.*, **46** (1950), 595–602.
- 3) “A note on the size frequency distribution of penetrating showers”, *Proc. Phys. Soc. Lond.*, **63A** (1950), 861–863.
- 4) “Stochastic processes and their applications to physical problems” *PhD Thesis, Univ. Manchester* (1951).
- 5) “Some simple stochastic processes”, *J. Roy. Stat. Soc.*, **13** (1951), 131–140.
- 6) “A note on Janossy’s mathematical model of a nucleon cascade”, *Proc. Camb. Phil. Soc.*, **48** (1952), 451–456.
- 7) “On an integral equation of Chandrasekhar and Munsch”, *Astrophys. J.*, **115** (1952), 141–144.
- 8) “Stochastic processes associated with random divisions of a line”, *Proc. Camb. Phil. Soc.*, **49** (1953), 473–485.
- 9) (with P. M. Mathews) “A stochastic problem relating to counters”, *Phil. Mag.*, **44** (1953), 1122–1127.
- 10) (with P. M. Mathews) “Numerical work on the fluctuation problem of electron cascades”, *Prog. Theor. Phys.*, **9** (1953), 679–681.
- 11) (with P. M. Mathews) “On a class of stochastic integro-differential equations”, *Proc. Indian Acad. Sci. Ser. A*, **38** (1953), 450–466.
- 12) (with P. M. Mathews) “On the solution of an integral equation of Chandrasekhar and Munsch”, *Astrophys. J.*, **119** (1954), 81–90.
- 13) “A stochastic model of a fluctuating density field”, *Astrophys. J.*, **119** (1954), 443–455.
- 14) “A stochastic model of a fluctuating density field – II” *Astrophys. J.*, **119** (1954), 682–685.
- 15) “On the molecular distribution functions of a one dimensional fluid – I”, *Phil. Mag.*, **45** (1954), 401–410.
- 16) “On counters with random dead time” *Phil. Mag.*, **45** (1954), 1050–1052.
- 17) (with P. M. Mathews) “On the molecular distribution functions of a one dimensional fluid – II”, *Phil. Mag.*, **45** (1954), 1053–1058.
- 18) (with P. M. Mathews) “Studies in the stochastic problem of electron–photon cascades”, *Prog. in Theor. Phys.*, **11** (1954), 95–117.
- 19) (with S. K. Srinivasan) “Two simple stochastic models of cascade multiplication”, *Prog. in Theor. Phys.*, **11** (1954), 595–603.
- 20) “On stellar statistics” *Astrophys. J.*, **122** (1955), 24–31.

- 21) "Inverse probability and evolutionary Markov stochastic processes", *Proc. Indian Acad. Sci.*, **41** (1955), 145–153. (Read at the Annual Meeting of the Academy in Belgaum, Dec 1954.)
- 22) (with S. K. Srinivasan) "Fluctuations in the number of photons in an electron–photon cascade", *Prog. Theor. Phys.*, **13** (1955), 93–99.
- 23) (with P. M. Mathews) "Straggling of the range of fast particles as a stochastic process", *Proc. Indian Acad. Sci.*, **41** (1955), 202–209. (Read at the Annual Meeting of the Academy in Belgaum, Dec 1954.)
- 24) "Phenomenological interpretation of the integrals of a class of random functions", *Proc. Koninkl. Neth. Akad.* **58** (= *Indag. Math.*, **17**) (1955), 470–482.
- 25) "Phenomenological interpretation of the integrals of a class of random functions – II", *Proc. Koninkl. Neth. Akad.* **58** (= *Indag. Math.*, **17**) (1955), 634–645.
- 26) (with S. K. Srinivasan) "Correlation problems in the study of brightness of the Milky Way", *Astrophys. J.*, **123** (1956), 479–485.
- 27) "Processes represented as integrals of a class of random functions", *Proc. Koninkl. Neth. Akad.* **59** (= *Indag. Math.*, **18**) (1956), 121–127.
- 28) (with P. M. Mathews) "Stochastic processes associated with a symmetric oscillatory Poisson process", *Proc. Indian Acad. Sci.*, **43A** (1956), 84–98.
- 29) (with S. K. Srinivasan) "A new approach to cascade theory", *Proc. Indian Acad. Sci. Ser. A*, **44** (1956), 263–273.
- 30) "A physical approach to stochastic processes", *Proc. Indian Acad. Sci. Ser. A*, **44** (1956), 428–450.
- 31) (with S. K. Srinivasan) "Stochastic integrals associated with point processes" (in French), *Publ. Inst. Stat. Univ. Paris*, **5** (1956), 95–106.
- 32) (with R. Vasudevan) "On the distribution of visible stars", *Astrophys. J.*, **126** (1957), 573–578.
- 33) "Ergodic properties of some simple stochastic processes", *Z. angew. Math. Mech.*, **37** (1957), 336–344. (Read at the GAMM Conference in May 1956 in Stuttgart.)
- 34) (with S. K. Srinivasan) "A note on cascade theory with ionisation loss", *Proc. Indian Acad. Sci., Ser. A*, **45** (1957), 133–138.
- 35) (with N. R. Ranganathan, S. K. Srinivasan and R. Vasudevan) "Multiple processes in electron–photon cascades", *Proc. Indian Acad. Sci. Ser. A*, **45** (1957), 311–326.
- 36) (with S. K. Srinivasan) "On age distribution in population growth", *Bull. Math. Biophys.*, **20** (1958), 289–303.
- 37) "Theoretical physics in the USA", *Curr. Sci.* **27** (1958), 469–471.
- 38) "Ambigenous stochastic processes", *Z. angew. Math. Mech.*, **39** (1959), 389–390.
- 39) (with N. R. Ranganathan and S. K. Srinivasan) "Meson production in nucleon–nucleon collisions", *Nucl. Phys.*, **10** (1959), 160–165.
- 40) (with N. R. Ranganathan, S. K. Srinivasan and K. Venkatesan) "Photo-mesons from polarized nucleons" *Proc. Indian Acad. Sci. Ser. A*, **49** (1959), 302–306.
- 41) (with N. R. Ranganathan and S. K. Srinivasan) "A note on the interaction between nucleon and anti-nucleon", *Proc. Indian Acad. Sci.*, **50** (1959), 91–94.
- 42) "Probability and stochastic processes" in *Handbuch der Physik* (S. Flugge, Ed) **III/2** (1959), Springer, Berlin (1959), 524–651.
- 43) (with N. R. Ranganathan, S. K. Srinivasan, and R. Vasudevan) "A note on dispersion relations", *Nucl. Phys.*, **15** (1960), 516–518.
- 44) "Perturbation expansions and kernel functions associated with single particle wave functions" in *Studies in Theor. Phys., Proc. 1959 Mussoorie Summer School* **1** (1960), 1–14.
- 45) "Quantum mechanics of the photon" in *Studies in Theor. Phys., Proc. 1959 Mussoorie Summer School* **1** (1960), 15–18.

- 46) "Applications of the theory of stochastic processes to physical problems", in *Studies in Theor. Phys., Proc. 1959 Mussoorie Summer School* **2** (1960), 239–253.
- 47) (with R. Vasudevan) "A physical approach to some limiting stochastic operation", *J. Indian Math. Soc.* **XXIV** (Golden Jubilee Volume, 1960), 458–477. (work done at Institute for Advanced Study in 1957–1958; presented at the Int'l Congress of Mathematicians, Edinburgh, 1958).
- 48) (with A. P. Balachandran and N. R. Ranganathan) "Some remarks on the structure of elementary particle interactions", *Proc. Indian Acad. Sci.*, **52** (1960), 1–11.
- 49) (with T. K. Radha and R. Thunga) "On the decomposition of the Feynman propagator" *Proc. Indian Acad. Sci. Ser. A*, **52** (1960), 228–239.
- 50) (with P. Rajagopal and R. Vasudevan) "Ambigenous stochastic processes", *J. Math. Anal. Appl.*, **1** (1960), 145–162. (Read by AR at the GAMM Conf., Hanover in May 1959).
- 51) (with T. K. Radha) "Correlation problems in evolutionary stochastic processes", *Proc. Camb. Phil. Soc.*, **57** (1961), 843–847.
- 52) (with A. P. Balachandran, N. G. Deshpande, and N. R. Ranganathan) "On an isobaric spin scheme for leptons and leptonic decays of strange particles", *Nucl. Phys.*, **26** (1961), 52–56.
- 53) (with R. Vasudevan) "A physical approach to limiting stochastic operations", *J. Indian Math. Soc. (N.S)* **24** (1961), 457–477.
- 54) (with G. Bhamathi and S. Indumathi) "A limiting process in quantum electrodynamics", *Proc. Indian Acad. Sci. Ser. A*, **53** (1961), 206–213.
- 55) (with V. Devanathan and G. Ramachandran) "A time dependent approach to rearrangement collisions", *Il Nuovo Cimento*, **21** (1961), 145–154.
- 56) (with S. K. Srinivasan) "A note on electron photon showers", *Nucl. Phys.*, **25** (1961), 152–154.
- 57) (with G. Bhamathi, S. Indumathi, T. K. Radha, and R. Thunga) "Some consequences of spin  $\frac{3}{2}$  for  $\Xi$ ", *Il Nuovo Cimento*, **22** (1961), 604–609.
- 58) (with V. Devanathan and G. Ramachandran) "Elastic photo production of neutral pions from deuterium", *Nucl. Phys.*, **24** (1961), 163–168.
- 59) (with N. R. Ranganathan) "Stochastic models in quantum mechanics" *J. Math. Anal. Appl.*, **3** (1961), 261–294. (Presented by AR at the Conference on Elementary Particles, Trieste 1960).
- 60) (with K. Venkatesan) "Some new stochastic aspects in cascade theory", in *Proc. 7-th Annual Cosmic Ray Symposium, Chandigarh* (1961), 59–61.
- 61) (with T. K. Radha) "Essay on symmetries" *Lectures at the Kodaikanal Summer School*, **2** (1961), 1–77.
- 62) (with N. R. Ranganathan) "Stochastic methods in quantum mechanics", *J. Math. Anal. Appl.*, **3** (1961), 261–294.
- 63) (with T. K. Radha and R. Thunga) "The physical basis of quantum field theory", *J. Math. Anal. Appl.*, **4** (1962), 494–526.
- 64) (with T. K. Radha and R. Thunga) "On the concept of virtual states", *J. Math. Anal. Appl.*, **5** (1962), 225–236.
- 65) (with G. Ramachandran) "Magnetic bremsstrahlung in nucleon–electron collisions", *Rand Corporation Preprint*, Los Angeles (1962).
- 66) "New perspectives on the Dirac Hamiltonian and the Feynman propagator", in *High Energy Phys. and Fundamental Particles*, Gordon and Breach, NY (1962), 665–672.
- 67) "A new form of the Feynman propagator", *J. Math. Phys. Sci.*, **1** (1967), 57–64.
- 68) (with A. P. Balachandran and K. Raman) "Low energy  $K^+$ -nucleon scattering", *Il Nuovo Cimento*, **24** (1962), 369–378.
- 69) (with V. Devanathan and K. Venkatesan) "On the scattering of pions by deuterons", *Nucl. Phys.*, **29** (1962), 680–686.

- 70) (with T. K. Radha and R. Thunga) "Possible resonances in  $\Xi_p$  reactions", *Nucl. Phys.*, **29** (1962), 517–523.
- 71) (with A. P. Balachandran) "Partial wave dispersion relations for  $\Lambda$ -nucleon scattering", *Il Nuovo Cimento*, **24** (1962), 980–999.
- 72) (with A. P. Balachandran, T. K. Radha, and R. Thunga) "On the  $Y^*$  resonances", *Il Nuovo Cimento*, **24** (1962), 1006–1012.
- 73) (with A. P. Balachandran, T. K. Radha, and R. Thunga) "On the spin and parity of  $Y^*$  resonances", *Il Nuovo Cimento*, **25** (1962), 723–729.
- 74) (with A. P. Balachandran, T. K. Radha, and R. Thunga) "Photo production of pions and  $\Lambda$ -hyperons", *Il Nuovo Cimento*, **25** (1962), 939–942.
- 75) (with G. Bhamathi, S. Indumathi, T. K. Radha, and R. Thunga) "Dispersion analysis of  $\Xi$  production in  $KN$  collisions" *Nucl. Phys.*, **37** (1962), 585–593.
- 76) (with T. K. Radha, K. Raman, and R. Thunga) "Quantum numbers and decay models of resonances" *Rand Corp. Preprint*, Los Angeles (1962).
- 77) "An unconventional view of perturbation expansions", in *Proc. Seminar on Unified Theories of Elem. Particles, Univ. Rochester, D. Lurie and N. Mukunda, Eds.* (1963), 411–421.
- 78) (With V. Devanathan and K. Venkatesan) "A note on the use of Wick's theorem", *J. Math. Anal. Appl.*, **8** (1964), 345–349.
- 79) (with K. Raman and R. K. Umergee) "Isobar production in nucleon–nucleon scattering", *Nucl. Phys.*, **60** (1964), 401–426.
- 80) (with K. Raman and R. K. Umergee) "Isobar production in nucleon–nucleon scattering – II, Polarization effects", *Nucl. Phys.*, **66** (1965), 609–631.
- 81) (with S. K. Srinivasan and R. Vasudevan) "Some new mathematical features in cascade theory", *J. Math. Anal. Appl.*, **11** (1965), 278–289. (Presented at the Int'l Conf. on Cosmic Rays, **5** (1964), TIFR, Bombay, 458–501.)
- 82) (with T. S. Shankara and K. Venkatesan) "Sensitivity of the vector coupling constant to  $\mu$ -neutrino mass and T-invariance", *Il Nuovo Cimento*, **37** (1965), 1046–1048.
- 83) (with R. Vasudevan and S. K. Srinivasan) "Scattering phase shifts in stochastic fields", *Z. Phys.*, **196** (1966), 112–122.
- 84) "Fundamental Multiplets" in *Symp. in Theor. Phys. and Maths., Alladi Ramakrishnan Ed.* **5** Plenum, NY (1967), 85–92.
- 85) "New perspectives on the Dirac Hamiltonian and the Feynman propagator", in *High energy physics and fundamental particles* (1967), Gordon and Breach, New York, 665–672. (Presented at the Theoretical Physics Institute, University of Colorado, Boulder, 1967.)
- 86) (with S. K. Srinivasan and R. Vasudevan) "Angular correlations in the brightness of the Milky Way", *J. Math. Phys. Sci.*, **1** (1967), 75–84.
- 87) "Some new topological features in Feynman graphs", *J. Math. Anal. Appl.*, **17** (1967), 68–91.
- 88) "A new form of the Feynman propagator", *J. Math. Phys. Sci.*, **1** (1967), 57–64.
- 89) "L-matrix hierarchy and the higher dimensional Dirac Hamiltonian", *J. Math. Phys. Sci.*, **1** (1967), 190–193.
- 90) (with S. K. Srinivasan and R. Vasudevan) "Multiple product densities", *J. Math. Phys. Sci.*, **1** (1967), 275–279.
- 91) "Graphical representation of CPT", *J. Math. Anal. Appl.*, **17** (1967), 147–150.
- 92) (with I. V. V. Raghavacharyulu) "A new combinatorial feature of Feynman graphs", *J. Math. Anal. Appl.*, **18** (1967), 175–181.
- 93) "The Dirac Hamiltonian as a member of a hierarchy of matrices", *J. Math. Anal. Appl.*, **20** (1967), 9–16.
- 94) "Helicity and energy as members of a hierarchy of eigenvalues", *J. Math. Anal. Appl.*, **20** (1967), 397–401.
- 95) "Symmetry operations on a hierarchy of matrices", *J. Math. Anal. Appl.*, **21** (1968), 39–42.

- 96) "On the relationship between L-matrix hierarchy and Cartan spinors", *J. Math. Anal. Appl.*, **22** (1968), 570–576.
- 97) (with P. S. Chandrasekharan, N. R. Ranganathan, and R. Vasudevan) "A generalization of the L-Matrix hierarchy", *J. Math. Anal. Appl.*, **23** (1968), 10–14.
- 98) "L-Matrices, quaternions and propagators", *J. Math. Anal. Appl.*, **23** (1968), 250–253.
- 99) (with I. V. V. Raghavacharyulu) "A note on the representation of Dirac groups", in *Symposia on Theor. Phys. and Math.* (Alladi Ramakrishnan, Ed.), **8** (1968), Plenum, NY, 25–32.
- 100) "Should we revise our notions about spin and parity in relativistic quantum theory?" *J. Math. Phys. Sci.*, **3** (1969), 213–219.
- 101) (with P. S. Chandrasekharan and T. S. Santhanam) "On representations of generalised Clifford algebras", *J. Math. Phys. Sci.*, **3** (1969), 301–313.
- 102) "Symmetries associated with roots of the unit matrix", *J. Math. Phys. Sci.*, **3** (1969), 317–318.
- 103) "Generalized helicity matrices" *J. Math. Anal. Appl.*, **26** (1969), 88–91.
- 104) (with P. S. Chandrasekharan, T. S. Santhanam, and A. Sundaram) "Helicity matrices for generalized Clifford algebra", *J. Math. Anal. Appl.*, **26** (1969), 275–278.
- 105) (with P. S. Chandrasekharan, N. R. Ranganathan, T. S. Santhanam, and R. Vasudevan) "The generalized Clifford algebra and the unitary groups", *J. Math. Anal. Appl.*, **27** (1969), 164–170.
- 106) (with P. S. Chandrasekharan, N. R. Ranganathan, T. S. Santhanam, and R. Vasudevan) "Idempotent matrices from a generalized Clifford algebra", *J. Math. Anal. Appl.*, **27** (1969), 563–564.
- 107) (with P. S. Chandrasekharan, N. R. Ranganathan, and R. Vasudevan) "Kemmer algebra from generalized Clifford elements", *J. Math. Anal. Appl.*, **28** (1969), 108–110.
- 108) "On the algebra of L-matrices", in *Symposia on Theor. Phys. and Math.* (Alladi Ramakrishnan, Ed.), **9** (1969), Plenum, NY, 73–78.
- 109) "L-matrices and propagators with imaginary parameters", in *Symposia on Theor. Phys. and Math.* (Alladi Ramakrishnan, Ed.) **9** (1969), Plenum Press, NY, 79–84.
- 110) (with R. Vasudevan) "A hierarchy of idempotent matrices", in *Symposia on Theor. Phys. and Math.* (Alladi Ramakrishnan, Ed.), **9** (1969), Plenum, NY, 85–88.
- 111) (with P. S. Chandrasekharan and R. Vasudevan) "Representation of para-Fermi rings and generalized Clifford algebra", *J. Math. Anal. Appl.*, **31** (1970), 1–5.
- 112) "On the composition of generalized helicity matrices", *J. Math. Anal. Appl.*, **31** (1970), 254–258.
- 113) (with R. Vasudevan) "On generalized idempotent matrices", *J. Math. Anal. Appl.*, **32** (1970), 414–423.
- 114) "New generalizations of Pauli matrices", in *Proc. Int'l Conf. on Symmetries and Quark Models, Wayne State Univ. 1969*, Gordon and Breach, NY (1970), 133–138.
- 115) "Unitary generalization of Pauli matrices", in *Symposia on Theor. Phys. and Math.* (Alladi Ramakrishnan, Ed.), **10** (1970), Plenum, NY, 51–57.
- 116) (with I. V. V. Raghavacharyulu) "Generalized Clifford basis and infinitesimal generators of the unitary group", in *Symposia on Theor. Phys. and Math.* (Alladi Ramakrishnan, Ed.), **10** (1970), Plenum, NY, 59–62.
- 117) (with P. S. Chandrasekharan and T. S. Santhanam) "L-matrices and the fundamental theorem of spinor theory", in *Symposia on Theor. Phys. and Math.* (Alladi Ramakrishnan, Ed.), **10** (1970), Plenum, NY, 63–68.
- 118) "Stochastic theory of evolutionary processes (1937–1971)" in *Stochastic Point Processes, Proc. IBM Conf. on Point Processes*, (J. Lewis Ed.), Interscience, John Wiley (1971), 533–548.
- 119) "A new approach to quantum numbers in elementary particle physics", in *Proc. Rutherford Centennial Conf., Christchurch* (1971), 150–156.



- 120) (with P. S. Chandrasekharan and R. Vasudevan) "Algebras derived from polynomial conditions", *J. Math. Anal. Appl.*, **35** (1971), 131–134.
- 121) (with P. S. Chandrasekharan and R. Vasudevan) "Para-Fermi operators and special unitary algebras", *J. Math. Anal. Appl.*, **35** (1971), 249–254.
- 122) "The weak interaction Hamiltonian in L-Matrix theory", *J. Math. Anal. Appl.*, **37** (1972), 432–434.
- 123) "On the shell structure of an L-Matrix", *J. Math. Anal. Appl.*, **38** (1972), 106–108.
- 124) "A matrix decomposition theorem", *J. Math. Anal. Appl.*, **40** (1972), 36–38.
- 125) "Einstein – a natural completion of Newton", *J. Math. Anal. Appl.*, **42** (1973), 377–380.
- 126) "The generalized Gell-Mann – Nishijima relation", in *Proc. Conf. on Nucl. Phys., MATSCIENCE Report*, **78** (1973), 1–4.
- 127) (with R. Jaganathan) "A new approach to matrix theory or many facets of the matrix decomposition theorem", in *Topics in Numerical Analysis, Proc. 1974 Int'l Conf. on Numerical Analysis, Dublin* (J. H. Miller Ed.) **133** (1976), 133–139.
- 128) "New concepts in matrix theory" *J. Math. Anal. Appl.*, **60** (1977), 255–258.
- 129) "Unnoticed symmetries in Einstein's special relativity", *J. Math. Anal. Appl.*, **63** (1978), 335–338.
- 130) "Lorentz to Gell-Mann – can the masters be retouched?" *Proc. Tamil Nadu Acad. Sci.* **1** (1978), 29–33.
- 131) "A new concept in special relativity – exterior relative velocity", *Proc. Tamil Nadu Acad. Sci.* **1** (1978) 67–69.
- 132) "A new look at matrix operations" in *Proc. Oberwolfach Conf. on Math. Phys. Dec. 1977* (Methoden und Verfahren der Mathematischen Physik – B. Brosowski and E. Martenson, Eds.), Peter Lang Publ. (1978), 49–53.
- 133) "New facets and new concepts in the special theory of relativity", in *Proc. Oberwolfach Conf. on Math. Phys. Dec. 1977* (Methoden und Verfahren der Mathematischen Physik – B. Brosowski and E. Martenson, eds.), Peter Lang Publ. (1978), 55–61.
- 134) "Approach to stationarity in stochastic processes", *Proc. Tamil Nadu Acad. Sci.* **2** (1979), 189–190.
- 135) "On the generalization of the Gellmann-Nishijima relation", in *Symmetries in Science* (B. Gruber and R. S. Millman, Eds.), Plenum Publ. Corp. (1980), 323–325.
- 136) "Mathematical features of evolutionary stochastic processes", *Methods of Operations Research*, **36** (1980), 239–240.
- 137) "A new concept in probability theory", *J. Math. Anal. Appl.*, **83** (1981), 408–410.
- 138) "Duality in stochastic processes", *J. Math. Anal. Appl.*, **84** (1981), 483–485.
- 139) "Further unnoticed symmetries in special relativity", *J. Math. Anal. Appl.*, **94** (1983), 237–241.
- 140) "Tantalizing asymmetries in special relativity", *J. Math. Anal. Appl.*, **110** (1985), 222–224.
- 141) "Decomposition of intervals in special relativity", *J. Math. Anal. Appl.*, **110** (1985), 225–226.
- 142) "A reflection principle", *Phys. Educ.*, **30** (1995), 204–205.
- 143) "A remarkable unnoticed theorem in special relativity", *J. Math. Anal. Appl.*, **213** (1997), 155–159.
- 144) "Theorem on non-simultaneity – extension to an external observer", *J. Math. Anal. Appl.*, **213** (1997), 354–356.
- 145) "Cubic and general extensions of the Lorentz transformation", *J. Math. Anal. Appl.*, **229** (1999), 88–92.
- 146) "A new Rod approach to the special theory of relativity", *J. Math. Anal. Appl.*, **249** (2000), 243–251. (Special Millennium issue dedicated to Richard Bellman)

**Books/Monographs/Handbuch articles**

- 1) “Probability and stochastic processes”, in *Handbuch der Physik*, **3** (1959) Springer, Berlin, 524–651.
- 2) “Elementary particles and cosmic rays”, Pergamon Press, Oxford (1962), 580 pp
- 3) “L-Matrix theory or the grammar of Dirac matrices”, Tata-McGraw Hill, Bombay-New Delhi (1972)
- 4) “Special relativity”, East West Books, Madras, India (2005)

**Books edited**

*Symposia in Theor. Phys. and Math.* (Alladi Ramakrishnan, Ed.) Vol I – Vol X, (1967–1970), Plenum, New York.

# List of PhD Students of Alladi Ramakrishnan

*The following is the list of students who obtained their PhD under the supervision of Professor Alladi Ramakrishnan at the University of Madras and at MATSCIENCE, The Institute of Mathematical Sciences. The year PhD was granted is given at the beginning. All PhD degrees were granted by the University of Madras.*

- 1) (1956) **P. M. Mathews**
- 2) (1957) **S. K. Srinivasan**
- 3) (1960) **R. Vasudevan**
- 4) (1961) **N. R. Ranganathan**
- 5) (1962) **T. K. Radha**
- 6) (1962) **R. Thunga**
- 7) (1962) **A. P. Balachandran**
- 8) (1963) **V. Devanathan**
- 9) (1963) **K. Venkatesan**
- 10) (1963) **G. Bhamathi**
- 11) (1963) **S. Indumathi**
- 12) (1963) **G. Ramachandran**
- 13) (1964) **K. Raman**
- 14) (1964) **R. K. Umerjee**
- 15) (1965) **K. Ananthanarayanan**
- 16) (1969) **T. S. Shankara**
- 17) (1970) **T. S. Santhanam**
- 18) (1970) **K. Srinivasa Rao**
- 19) (1971) **P. S. Chandrasekharan**
- 20) (1971) **A. Sundaram**
- 21) (1972) **Nalini B. Menon**
- 22) (1975) **A. R. Tekumalla**
- 23) (1976) **R. Jagannathan**
- 24) (1979) **A. Vijayakumar**

**Part II**  
**Pure Mathematics**

# Inversion and Invariance of Characteristic Terms: Part I

Shreeram S. Abhyankar

*Fondly dedicated to the 104th birthday of my father  
Professor S.K. Abhyankar who was born on 4/4/1904  
Also dedicated to the memory of  
Professor Alladi Ramakrishnan in great admiration*

**Summary** In my 1967 paper with almost the same title which appeared in volume 89 of the American Journal of Mathematics, I proved the invariance of the characteristic terms in the fractional power series expansion of a branch of an algebraic plane curve over fields of characteristic zero. Now I extend the results by a more generous interpretation of the characteristic terms, and by relaxing the characteristic zero hypothesis.

**Mathematics Subject Classification (2000)** Primary 14H20; Secondary 13F30

**Key words and phrases** Invariance · Valuations

## 1 Introduction

A branch of an algebraic or analytic plane curve can be parametrized by expressing both the variables as power series in a parameter; we call this the MT (= Maclaurin–Taylor) expansion. In case of zero characteristic, by Hensel’s Lemma or by Newton’s Theorem on fractional power series expansion, one of the variables can be arranged to be a power of the parameter, and then certain divisibility properties of the exponents in the expansion of the other variable lead to the characteristic terms whose importance was first pointed out by Smith [34] and Halphen [23] as noted in Zariski’s famous book [35].

---

S.S. Abhyankar  
Mathematics Department, Purdue University, West Lafayette, IN 47907, USA  
e-mail: [ram@cs.purdue.edu](mailto:ram@cs.purdue.edu)

In my 1967 paper [4] I showed that, as long as the variable which is a power of the parameter is nontangential, the characteristic terms remain invariant. This I did by first showing that if I flip the variables, then the characteristic terms change by a definite inversion formula whose proof essentially depends on the binomial theorem. This will be reviewed in Sect. 4. In Sect. 5, I shall relate this to quadratic transformations and establish the invariance of another type of characteristic term, namely, the first exponent whose coefficient is transcendental over a certain subfield of the ground field. While doing this, I shall reorganize the NT (= newtonian) expansion into the ED (= euclidean) expansion, which is a generalized form of the so called HN (= Hamburger-Noether) expansion. The reorganization will partly make things work even in the mixed characteristic meromorphic case.

As basic references for this paper, the reader may profitably consult my Rambling Article [5], Tata Notes [6], Engineering Book [9], and Algebra Book [12].

After fixing the notation in Sect. 2, a host of Remarks and Lemmas will be collected in Sect. 3. These deal with Euclidean Sequences (3.1), Characteristic Sequences (3.2), Binomial Lemmas (3.3) and (3.4), Special Subfields (3.5), Gap Lemmas (3.6) and (3.7), Valuation Expansions (3.8) and (3.9), and Uniqueness of Power Series Rings (3.10).

In Sect. 6, I shall show how the above mentioned first transcendental coefficient is related to a generator of the residue field of the branch. Moreover, the generator can be chosen so that the said coefficient is a polynomial in it. This leads to an algebraic incarnation of the topological theory of dicritical divisors which I shall describe. In Sect. 7, I shall relate field generators to dicritical divisors.

In Sect. 8, I shall preview Part II which will include various topics from algebraic curve theory such as the conductor and genus formulas of Dedekind and Noether, and the automorphism theorems of Jung and Kulk. In Part II, I shall also relate all this to the Jacobian problem which conjectures that if the Jacobian of  $n$  polynomials in  $n$  variables over a characteristic zero field equals a nonzero constant, then the variables can be expressed as polynomials in the given polynomials; see [13–15].

As hinted in the Note following Lemma (3.4) of Sect. 3, Newton's Binomial Theorem For Fractional Exponents is the real heart of this paper. I was very lucky in having studied this in the hand-written manuscript of my father's book [1] two years before it was published when I was 11 years old. Very relevant is the following comment which he makes on page 235 of his book:

From

$$(a + b)^n = a^n + \dots$$

we get the standard form

$$(1 + x)^n = 1 + \dots$$

by writing 1 for  $a$  and  $x$  for  $b$ ; the standard form is simpler and is more convenient to use; all problems regarding binomial expansions can be solved by using the standard form.

Coming to the idea of Inversion in the title of this paper, let me repeat from page 194 of my Engineering Book [9] the following quotation from page 323 of the chapter on Abel in Bell's Men of Mathematics [17]:

Instead of assuming that people are depraved because they drink to excess, Galton inverted this hypothesis ... For the moment we need note only that Galton, like Abel, inverted his problem – turned it upside-down, and inside-out, back-end-to and foremost-end-backward ... ‘you must always invert,’ as Jacobi said when asked the secret of his mathematical discoveries. He was recalling what Abel and he had done.

On page 309 of the chapter on Abel, Bell says: One of his (= Abel’s) classics in this direction is the first proof of the general binomial theorem, special cases of which had been stated by Newton and Euler.

In other words, Bell disagrees with my viewpoint that Newton stated and proved the most general form of the Binomial Theorem.

In this connection, let me repeat what I said on page 417 of my Ramblings Article [5]: Generally speaking, from Newton to Cauchy, mathematicians used power series without regard to convergence. They were criticized for this and the matter was rectified by the analysts Cauchy and Abel who developed a rigorous theory of convergence. After another hundred years or so we were taught, say by Hensel, Krull, and Chevalley, that it really didn’t matter, i.e., we may disregard convergence after all! So the algebraist was freed from the shackles of analysis, or rather (as in Vedanta philosophy) he was told that he always was free but had only forgotten it temporarily.

Now one good way to study the rest of this paper is to INVERT it by first reading the last section called EPILOGUE, which is sort of an extended Introduction or a Birds Eye View of the entire paper. Another idea is to start with Sect. 4 and refer to Sects. 2 and 3 as necessary. More precisely, start by reading definition (••) of a valuation sequence given at the beginning of Sect. 4. Our goal in Sect. 4 is to show that the newtonian expansion of the first two terms of that sequence partly determines the newtonian expansion of any two consecutive terms.

## 2 Notation

We shall mostly follow the notation and terminology of my Kyoto paper [7] and my books [9, 12]. In particular:  $\mathbb{N}$  = the set of all nonnegative integers,  $\mathbb{N}_+$  = the set of all positive integers,  $\widehat{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ , and  $R^\times$  = the set of all nonzero elements in a ring  $R$ . The GCD of a set of integers  $S$  is the unique nonnegative generator of the ideal  $S\mathbb{Z}$  in the ring of integers  $\mathbb{Z}$  generated by  $S$ ; if the set  $S$  contains a noninteger then  $\text{GCD}(S) = \infty$ . A set of integers  $J$  is bounded from below means for some integer  $e$  we have  $e \leq j$  for all  $j \in J$ , and we write  $\min J$  for the smallest element of such a set, with the convention that if  $J$  is the empty set  $\emptyset$  then  $\min J = \infty$ .

To fix some more notation: Recall that a quasilocal ring is a (commutative with identity) ring  $R$  having a unique maximal ideal  $M(R)$ ; we let  $H(R)$  stand for its residue field  $R/M(R)$ , and by  $H_R : R \rightarrow R/M(R)$  we denote the residue class epimorphism; note that then  $H(R) = H_R(R)$ . By a coefficient set of  $R$ , we mean a subset  $k$  of  $R$  with  $0 \in k$  and  $1 \in k$  such that  $H_R$  maps  $k$  bijectively onto

$H(R)$ . By a coefficient field of  $R$ , we mean a coefficient set  $k$  of  $R$  such that  $k$  is a subfield of  $R$ . For any subfield  $K$  of  $R$ , we note that  $H_R$  maps  $K$  isomorphically onto the subfield  $H_R(K)$  of  $H(R)$  and we let  $\text{trdeg}_K H(R)$  and  $[H(R) : K]$  stand for  $\text{trdeg}_{H_R(K)} H(R)$  and  $[H(R) : H_R(K)]$ , respectively. Given an element  $z$  in an overring of  $R$ , we say that  $z$  is residually transcendental over  $K$  at  $R$  to mean that  $z \in R$  and  $H_R(z)$  is transcendental over  $H_R(K)$ .

Recall that a field extension  $L/K$  is algebraic (resp: finite algebraic, transcendental, simple transcendental, pure transcendental) means  $[K(w) : K] < \infty$  for all  $w \in L$  (resp:  $[L : K] < \infty$ ,  $\text{trdeg}_K L > 0$ ,  $\text{trdeg}_K L = 1$  and  $L = K(t)$  for some  $t \in L$ ,  $\text{trdeg}_K L = v \in \mathbb{N}$  and  $L = K(t_1, \dots, t_v)$  for some  $t_1, \dots, t_v$  in  $L$ ). Recall that an affine domain over a field is a domain which is a finitely generated ring extension of that field. The characteristic of a field  $K$  is denoted by  $\text{ch}(K)$ . The dimension  $\text{dim}(R)$  of a ring  $R$  is the maximum length  $n$  of a chain of prime ideals

$$P_0 \subsetneq P_1 \subsetneq \dots \subsetneq P_n$$

in  $R$ .

A noetherian quasilocal ring  $R$  is called a local ring. The smallest number of generators of  $M(R)$  is called the embedding dimension of  $R$  and is denoted by  $\text{emdim}(R)$ . We always have  $\text{emdim}(R) \geq \text{dim}(R)$  and  $R$  is regular means equality holds; a regular local ring is always a domain. A DVR is a one-dimensional regular local domain; Alternatively, a DVR is the valuation ring of a real discrete valuation in the following sense. A valuation is a map  $W : L \rightarrow G \cup \{\infty\}$ , where  $L$  is a field and  $G$  is an ordered abelian group, such that for all  $u, u'$  in  $L$  we have  $W(uu') = W(u) + W(u')$  and  $W(u + u') \geq \min(W(u), W(u'))$  and for any  $u$  in  $L$  we have:  $W(u) = \infty \Leftrightarrow u = 0$ . We put  $G_W = W(K^\times)$  and  $R_W = \{u \in K : W(u) \geq 0\}$  and call these the value group and the valuation ring of  $W$ . Now  $R_W$  is a ring with the unique maximal ideal  $M(R_W) = \{u \in K : W(u) > 0\}$ . Thus  $R_W$  is a quasilocal ring. If  $G_W = \mathbb{Z}$  then  $W$  is said to be real discrete.

A quasilocal ring  $V$  dominates a quasilocal ring  $S$  means  $S$  is a subring of  $V$  with  $M(S) \subset M(V)$ , and then:  $\text{restrdeg}_S V$  denotes the residual transcendence degree of  $V$  over  $S$ , i.e., the transcendence degree of  $H(V)$  over  $H_V(S)$ ; we say that  $V$  is residually rational over  $S$  to mean that  $H(V) = H_V(S)$ ; we say that  $V$  is residually algebraic (resp: residually finite algebraic, residually transcendental, residually simple transcendental, residually pure transcendental) over  $S$  to mean that the field extension  $H(V)/H_V(S)$  is algebraic (resp: finite algebraic, transcendental, simple transcendental, pure transcendental). Given any subring  $A$  of a quasilocal ring  $V$ , upon letting  $S$  to be the localization of  $A$  at the prime ideal  $A \cap M(V)$ , we put  $\text{restrdeg}_A V = \text{restrdeg}_S V$  and call it the residual transcendence degree of  $V$  over  $A$ , and we say that  $V$  is residually rational (resp: residually algebraic, residually finite algebraic, residually transcendental, residually simple transcendental, residually pure transcendental) over  $A$  to mean that  $V$  is residually rational (resp: residually algebraic, residually finite algebraic, residually transcendental, residually simple transcendental, residually pure transcendental) over  $S$ .



For any local domain  $R$  and any  $z \in R^\times$ , we define  $\text{ord}_R z$  to be the largest nonnegative integer  $e$  such that  $z \in M(R)^e$ ; if  $z = 0$  then we put  $\text{ord}_R z = \infty$ . If  $R$  is regular then we extend this to the quotient field  $\text{QF}(R)$  of  $R$  by putting

$$\text{ord}_R(x/y) = \text{ord}_R x - \text{ord}_R y$$

for all  $x, y$  in  $R^\times$ ; if  $\dim(R) > 0$  then this gives a real discrete valuation of  $\text{QF}(R)$  whose valuation ring  $V$  dominates  $R$  and is residually pure transcendental over  $R$  of residual transcendence degree  $\dim(R) - 1$ . See (Q35.5) on pages 559–577 of [12].

Given any subring  $K$  of a domain  $L$ , by the transcendence degree of  $L$  over  $K$  we mean the transcendence degree of  $\text{QF}(L)$  over  $\text{QF}(K)$ , and we continue to denote it by  $\text{trdeg}_K L$ ; note that by convention, if  $\text{trdeg}_K L = \infty$  then  $(\text{trdeg}_K L) - 1 = \infty$ . Given any subring  $K$  of a field  $L$ , by  $\overline{D}(L/K)$  we denote the set of all valuation rings  $V$  with  $\text{QF}(V) = L$  such that  $K \subset V$ , and by  $D(L/K)$  we denote the set of all  $V \in \overline{D}(L/K)$  such that  $\text{trdeg}_{H_V(K)} H(V) = (\text{trdeg}_K L) - 1$ ; we call these  $V$  the *valuation rings* and *prime divisors* of  $L/K$  respectively. Note that if  $L$  is a finitely generated field extension of a field  $K$  then every member of  $D(L/K)$  is a DVR; moreover if  $\text{trdeg}_K L = 1$  then  $L$  is the only member of  $\overline{D}(L/K)$  which does not belong to  $D(L/K)$ .

Given any affine domain  $A$  over a field  $K$  with  $\text{QF}(A) = L$ , by  $\overline{I}(A/K)$  and  $I(A/K)$  we denote the set of all  $V \in \overline{D}(L/K)$  and  $V \in D(L/K)$ , respectively, such that  $A \not\subset V$ ; we call these  $V$  the *infinity valuation rings* and *infinity divisors* of  $A/K$  respectively. Note that all members of  $D(L/K)$ , and hence all members of  $I(A/K)$ , are DVRs. Also note that if  $\text{trdeg}_K L = 1$  then  $I(A/K)$  is a nonempty finite set, and for every  $V \in D(L/K)$  we have  $[H(V) : K] \in \mathbb{N}_+$ . Let us recall that DD = Dedekind Domain = normal noetherian domain of dimension at most one. Note that the localizations of a DD at the various nonzero prime ideals in it are DVRs whose intersection is the given DD. Note that a domain is a PID iff it is a DD as well as a UFD. Also note that a domain is a PID iff it is a noetherian UFD of dimension at most one. Let us say that a domain is proper to mean that it is not a field. In particular, a proper PID is a PID which is not a field.

Given any local domain  $R$ , by  $\overline{D}(R)^\Delta$  we denote the set of all  $V \in \overline{D}(\text{QF}(R)/R)$  such that  $V$  dominate  $R$ , and we let  $D(R)^\Delta$  denote the set of all  $V \in \overline{D}(R)^\Delta$  such that  $\text{restrdeg}_R V = \dim(R) - 1$ ; we call these  $V$  the *valuation rings* of  $\text{QF}(R)$  dominating  $R$  and *prime divisors* of  $R$  respectively; note that then for every  $V \in \overline{D}(R)^\Delta$  we have  $\text{restrdeg}_R V \leq \dim(R)$ , and for every  $V \in D(R)^\Delta$  we have that  $V$  is a DVR.

The habitat for most of the Remarks and Lemmas of the next section will be a DVR  $V$  with its quotient field  $\text{QF}(V) = L$ , its completion  $\widehat{V}$ , a coefficient field  $K$ , and a uniformizing parameter  $T$ , i.e., an element of  $V$  of order 1. Note that  $\widehat{V}$  can be identified with the power series ring  $K[[T]]$  and  $L$  with a subfield of the meromorphic series field  $K((T))$ . For any

$$y = y(T) = \sum_{i \in \mathbb{Z}} A_i T^i \in K((T)) \quad \text{with} \quad A_i \in K$$

we define the  $T$ -support  $\text{Supp}_T y(T)$  of  $y(T)$  to be the set of all  $i \in \mathbb{Z}$  with  $A_i \neq 0$ , and then we define the  $T$ -order and  $T$ -initial-coefficient of  $y(T)$  by putting

$$\text{ord}_T y(T) = \min \text{Supp}_T y(T)$$

and

$$\text{inco}_T y(T) = A_e \quad \text{where} \quad e = \text{ord}_T y(T)$$

with the understanding that if  $y(T) = 0$  then  $\text{ord}_T y(T) = \infty$  and  $\text{inco}_T y(T) = 0$ ; note that in case of  $\widehat{V} = K[[T]]$  we have  $\text{ord}_V y = \text{ord}_T y(T)$  for all  $y \in L$ . By a *special subfield*  $S$  of  $K((T))$  we mean either the null ring  $S = \{0\} \subset K$  or a subfield  $S$  of  $K((T))$  such that: if  $a \in S \cap K^\times$  and  $b \in K^\times$  with  $b^q = a^p$  for some  $p \in \mathbb{Z}$  and  $q \in \mathbb{N}_+$  then  $b \in S$ ; if  $S \subset K$  then we may call  $S$  a *special subfield* of  $K$ . Observe that if  $k$  is any special subfield of  $K$  then  $k$  as well as  $k((T))$  are special subfields of  $K((T))$ ; by convention, if  $k = \{0\}$  then  $k((T)) = \{0\}$ . We put

$$\text{sub}_T y(T) = \begin{cases} \text{the smallest special subfield } k \text{ of } K \\ \text{such that } A_i \in S \text{ for all } i \in \mathbb{Z} \\ \text{(with the note that } k = \{0\} \Leftrightarrow y(T) = 0) \end{cases}$$

We call  $\text{sub}_T y(T)$  the  $T$ -subfield of  $y(T)$ .

As a weaker version of algebraic closedness, we say that a field  $K$  is *root-closed* to mean that for every  $a \in K$  and  $n \in \mathbb{N}_+$  we have  $X^n - a = (X - a_1) \dots (X - a_n)$  for some  $a_1, \dots, a_n$  in  $K$ . Both the notions of a special subfield and a root-closed field are inspired by root extraction, i.e., the finding of square-roots, cube-roots, and so on. The process of root extraction also inspires the concept of a *quasiroot-closed* domain which we shall introduce in Remark (3.10).

### 3 Remarks and Lemmas

We start off by codifying the euclidean algorithm (= method of long division) of finding the GCD of a pair of integers.

Remark on Euclidean algorithm (3.1). By a *euclidean sequence pair*, we mean a pair  $((e_j)_{0 \leq j \leq l}, (p_j)_{0 \leq j < l})$  of sequences of integers  $e_j \in \mathbb{Z}$  and  $p_j \in \mathbb{Z}$  with  $l \in \mathbb{N}_+$  such that:

$$e_1 \neq 0 = p_0 = 0 \neq p_j \text{ for } 2 \leq j < l \text{ with } p_j > 0 \text{ for } 3 \leq j < l, \quad (1)$$

$$e_{j-1} = p_j e_j + e_{j+1} \text{ with } 0 < e_{j+1} < |e_j| \text{ for } 1 \leq j \leq l-1, \quad (2)$$

$$|e_j| > |e_l| = \text{GCD}(e_0, e_1) = \text{GCD}(e_0, \dots, e_l) \text{ for } 1 \leq j \leq l-1, \quad (3)$$

$$l = 1 \Leftrightarrow e_0 \equiv 0 \pmod{(e_1)}. \quad (4)$$

The usual euclidean algorithm implies that any pair of integers  $(e_0, e_1)$  with  $e_1 \neq 0$  can be embedded in a unique euclidean sequence pair  $((e_j)_{0 \leq j \leq l}, (p_j)_{0 \leq j < l})$  which we call the *euclidean extension* of  $(e_0, e_1)$ .

To apply this construction to orders of elements, let  $V$  be a DVR with

$$V \subset \widehat{V} = \text{the completion of } V \quad \text{and} \quad \text{QF}(V) = L \subset \widehat{L} = \text{QF}(\widehat{V})$$

and let  $K$  be a coefficient set of  $V$ .

By a  $(V, K)$ -*protosequence* we mean a sequence

$$(z_j, e_j, p_j, A_l^*(v), e_l^*, z_l^*)_{v \in \mathbb{Z}, 0 \leq j \leq l+1}$$

where

$$((e_j)_{0 \leq j \leq l}, (p_j)_{0 \leq j < l})$$

is a euclidean sequence pair and

$$\begin{cases} z_j \in L^\times \text{ with } \text{ord}_V z_j = e_j \text{ for } 0 \leq j \leq l, \\ \text{and } z_{j-1} = z_j^{p_j} z_{j+1} \text{ for } 1 \leq j \leq l-1 \end{cases} \quad (5)$$

and

$$\begin{cases} z_{l+1} \in L \text{ with } \text{ord}_V z_{l+1} = e_{l+1} \text{ and } p_l = p_{l+1} \in \mathbb{Z} \cup \{\infty\} \\ \text{and } z_l^* \in L \text{ with } \text{ord}_V z_l^* = e_l^* \\ \text{such that } z_{l+1} = 0 \Leftrightarrow p_l = \infty \Leftrightarrow z_l^* = 0 \\ \text{and } z_{l+1} \neq 0 \Rightarrow (e_{l-1}/e_l) \leq p_l(e_l/|e_l|) \text{ with } 0 < e_{l+1} < |e_l| \end{cases} \quad (6)$$

and

$$\begin{cases} A_l^*(v) \in K \text{ for all } v \in \mathbb{Z} \text{ such that} \\ \left. \begin{cases} = 0 & \text{if } v < (e_{l-1}/|e_l|) \\ \neq 0 & \text{if } v = (e_{l-1}/|e_l|) \\ = 0 & \text{if } v > p_l(e_l/|e_l|) \text{ and } z_{l+1} \neq 0 \end{cases} \right\} \end{cases} \quad (7)$$

such that in  $\widehat{L}$  we have

$$z_l^* = z_{l-1} - \sum_{(e_{l-1}/|e_l|) \leq v < \infty} A_l^*(v) z_l^{v(|e_l|/e_l)} = \begin{cases} 0 & \text{if } z_{l+1} = 0 \\ z_l^{p_l} z_{l+1} & \text{if } z_{l+1} \neq 0. \end{cases} \quad (8)$$

Any pair of elements  $(z_0, z_1)$  in  $L^\times$  with  $\text{ord}_V z_1 \neq 0$  can clearly be embedded in a unique  $(V, K)$ -protosequence

$$(z_j, e_j, p_j, A_l^*(v), e_l^*, z_l^*)_{v \in \mathbb{Z}, 0 \leq j \leq l+1}$$

which we call the  $(V, K)$ -*protoexpansion* of  $(z_0, z_1)$ .

To contrast the above expansion (8) with the usual expansions in terms of a uniformizing parameter  $T$  of  $\widehat{V}$ , we note that

$$\text{for } 0 \leq j \leq l + 1$$

there exist unique

$$\begin{cases} A_j(v) \in K \text{ for all } v \in \mathbb{Z} \\ \text{with } A_j(v) = 0 \text{ for } v < e_j \\ \text{and if } e_j \in \mathbb{Z} \text{ then } A_j(e_j) \neq 0 \end{cases} \quad (9)$$

such that

$$z_j = z_j(T) = \sum_{e_j \leq v < \infty} A_j(v) T^v. \quad (10)$$

In case  $z_0, z_1$  belong to  $V$ , we may visualize  $x = z_1(T), y = z_0(T)$  as giving a parametrization of a branch of a curve in the  $(x, y)$ -plane centered at the point  $(z_1(0), z_0(0))$ .

To continue our construction by a  $(V, K)$ -presequence we mean a sequence

$$(z_{ij}, e_{ij}, p_{ij}, A_{il(i)}^*(v), e_{il(i)}^*, z_{il(i)}^*)_{v \in \mathbb{Z}, 0 \leq j \leq l(i)+1, 0 \leq i \leq \kappa} \quad \text{with } \kappa \in \mathbb{N}$$

where

$$(z_{ij}, e_{ij}, p_{ij}, A_{il(i)}^*(v), e_{il(i)}^*, z_{il(i)}^*)_{v \in \mathbb{Z}, 0 \leq j \leq l(i)+1}$$

is a  $(V, K)$ -protosequence for  $0 \leq i \leq \kappa$  with

$$(z_{il(i)}^*, z_{il(i)}) = (z_{i+1,0}, z_{i+1,1}) \text{ for } 0 \leq i < \kappa \quad (11)$$

and

$$z_{\kappa, l(\kappa)+1} = 0. \quad (12)$$

Now given any pair of elements  $(z_0, z_1)$  in  $L^\times$  with  $\text{ord}_V z_1 \neq 0$ , clearly there exists a unique  $(V, K)$ -presequence

$$(z_{ij}, e_{ij}, p_{ij}, A_{il(i)}^*(v), e_{il(i)}^*, z_{il(i)}^*)_{v \in \mathbb{Z}, 0 \leq j \leq l(i)+1, 0 \leq i \leq \kappa}$$

with  $(z_{00}, z_{01}) = (z_0, z_1)$ , and we call this the  $(V, K)$ -preexpansion of  $(z_0, z_1)$ .

In a moment we shall relate the above expansion with the characteristic terms coming out of Newton's fractional power series expansion. To do this we start off with a string of definitions in the following Remark.

Remark on GCD dropping sequence (3.2). A GCD sequence is a system  $d$  consisting of its length  $h(d) \in \mathbb{N}$  and its sequence  $(d_i)_{0 \leq i \leq h(d)+2}$  where  $d_0 = 0, d_i \in \mathbb{N}_+$  for  $1 \leq i \leq h(d) + 1, d_i \in d_{i+1}\mathbb{Z}$  for  $0 \leq i \leq h(d)$ , and  $d_{h(d)+2} \in \widehat{\mathbb{R}}$ . A *charseq* (= characteristic sequence) is a system  $m$  consisting of its length  $h(m) \in \mathbb{N}$  and its sequence  $(m_i)_{0 \leq i \leq h(m)+1}$  where  $m_0 \in \mathbb{Z}^\times, m_i \in \mathbb{Z}$  for  $1 \leq i \leq h(m)$ , and

$m_{h(m)+1} \in \widehat{\mathbb{R}}$ . Given any charseq  $m$  with  $h = h(m)$ , its GCD sequence is the GCD sequence  $d = d(m)$  obtained by putting  $h(d) = h$ , and  $d_i = \text{GCD}(m_0, \dots, m_{i-1})$  for  $0 \leq i \leq h + 2$ ; its reciprocal sequence  $n(m)$  is the sequence  $n = (n_i)_{1 \leq i \leq h+1}$  obtained by putting  $n_i = d_1/d_i$  for  $1 \leq i \leq h + 1$ ; its difference sequence is the charseq  $q = q(m)$  obtained by putting  $h(q) = h$  with  $q_i = m_i$  for  $0 \leq i \leq 1$  and  $q_i = m_i - m_{i-1}$  for  $2 \leq i \leq h + 1$ ; note that clearly  $d(q) = d(m)$ . Given any charseq  $q$  with  $h = h(q)$  and  $d = d(q)$ , its inner product sequence is the charseq  $s = s(q)$  obtained by putting  $h(s) = h$  with  $s_0 = q_0$  and  $s_i = \sum_{1 \leq j \leq i} q_j d_j$  for  $1 \leq i \leq h + 1$ , and its normalized inner product sequence is the charseq  $r = r(q)$  obtained by putting  $h(r) = h$  with  $r_0 = s_0$  and  $r_i = s_i/d_i$  for  $1 \leq i \leq h + 1$ . Note that then  $d(r) = d(q)$ .

Let us also note that if  $m_{h+1} = \infty$  then  $q_{h+1} = s_{h+1} = r_{h+1} = d_{h+2} = \infty$  by the infinity convention according to which: for all  $c \in \mathbb{R}$  we have  $\infty \pm c = \infty$  and  $-\infty \pm c = -\infty$ , for all  $c \in \mathbb{R}_+$  = the set of all positive real numbers we have  $\infty c = \infty/c = \infty$  and  $-\infty c = -\infty/c = -\infty$ , and we have  $\infty + \infty = \infty$ .

It is worth observing that any one of the four sequences  $m, q(m), s(q(m)), r(q(m))$  determines the other three.

Given any charseq  $m$ , by the characteristic pair sequence of  $m$  we mean the sequence  $(\widehat{m}_i(m), \widehat{n}_i(m))_{1 \leq i \leq h(m)}$  defined by putting  $\widehat{m}_i(m) = m_i/d_{i+1}(m)$  and  $\widehat{n}_i(m) = d_i(m)/d_{i+1}(m)$  for  $1 \leq i \leq h(m)$ ; we call  $\widehat{m}(m) = \widehat{m}_i(m)_{1 \leq i \leq h(m)}$  the derived numerator sequence of  $m$ , and we call  $\widehat{n}(m) = \widehat{n}_i(m)_{1 \leq i \leq h(m)}$  the derived denominator sequence of  $m$ .

A charseq  $m$  is upper-unbounded means  $m_{h(m)+1} = \infty$ .

For any set of integers  $J$  which is bounded from below and for any nonzero integer  $l$ , we define the GCD-dropping sequence  $m = m(J, l)$  of  $J$  relative to  $l$  by saying that  $m$  is the unique upper-unbounded charseq with  $m_0 = l$  and  $m_1 = \min J$  such that for  $2 \leq i \leq h(m) + 1$  we have

$$m_i = \min\{j \in J : j \text{ is nondivisible by } \text{GCD}(m_0, \dots, m_{i-1})\}.$$

If  $F(X, Y)$  is a monic polynomial of positive degree  $N$  in  $Y$  with coefficients in the univariate meromorphic series field  $K((X))$  over an algebraically closed field  $K$  with  $N \not\equiv 0 \pmod{\text{ch}(K)}$  such that  $F$  is a power of a monic irreducible member of  $K((X))[Y]$ , then by Newton's Theorem on fractional meromorphic series expansion, we can factor

$$F(T^N, Y) = \prod_{1 \leq i \leq N} [Y - \eta_i(T)] \quad \text{with} \quad \eta_i(T) \in K((T)).$$

Clearly  $\text{Supp}_T \eta_i(T)$  is independent of  $i$ ; we denote this common support by  $\text{Supt}(F)$  and call it the newtonian support of  $F$ . We define the newtonian charseq  $m(F, l)$  of  $F$  relative to a nonzero integer  $l$  by putting  $m(F, l) = m(\text{Supt}(F), l)$ .

In [4], we had assumed  $F(X, Y) \in K[[X]][Y]$  and  $l = N$ . In Sect. 4, we shall now reprove the assertions of [4] for the somewhat more general case of  $F(X, Y) \in K((X))[Y]$  and  $l = \pm N$ .

First binomial lemma (3.3). Let us consider the univariate meromorphic series field  $K((T))$  over a field  $K$ . Let  $y \in K((T))^\times$  and  $z \in K((T))$  be such that

$$\text{ord}_T y = v < w = \text{ord}_T z$$

with

$$\text{inco}_T y = \eta \quad \text{and} \quad \text{inco}_T z = \zeta.$$

Then for any  $n \in \mathbb{Z}$  we have

$$\left\{ \begin{array}{l} (y+z)^n = y' + z' \text{ where } y' = y^n \in K((T))^\times \text{ and } z' \in K((T)) \\ \text{with } \text{inco}_T y' = \eta' \text{ and } \text{inco}_T z' = \zeta' \\ \text{are such that } \text{sub}_T y' \subset \text{sub}_T y \\ \text{with } \text{ord}_T y' = nv = v' < w' = (n-1)v + w \leq \text{ord}_T z' \\ \text{and } w' - v' = w - v \text{ with } \eta' = \eta^n \\ \text{and if } n \not\equiv 0 \pmod{\text{ch}(K)} \text{ then } \text{ord}_T z' = w' \text{ with } \zeta' = n\zeta\eta^{n-1}. \end{array} \right. \quad (\bullet)$$

Moreover, assuming  $n \not\equiv 0 \pmod{\text{ch}(K)}$ , we have that:

$$\left\{ \begin{array}{l} \text{if } \text{ord}_T(y - \eta T^v) = w \\ \text{then } \text{ord}_T(y' - \eta' T^{v'}) = w' \end{array} \right. \quad (1^*)$$

whereas

$$\left\{ \begin{array}{l} \text{if } y \in K((T^d)) \text{ and } w/d \notin \mathbb{Z} \text{ for some } d \in \mathbb{N}_+ \\ \text{then for that } d \text{ we have } y' \in K((T^d)) \text{ and } w'/d \notin \mathbb{Z} \end{array} \right. \quad (2^*)$$

while

$$\left\{ \begin{array}{l} \text{if } y \in k((T)) \text{ and } \zeta \notin k \text{ for some subfield } k \text{ of } K \\ \text{then for that } k \text{ we have } y' \in k((T)) \text{ and } \text{inco}_T z' \notin k. \end{array} \right. \quad (3^*)$$

*Proof.* (1\*)–(3\*) follow from (•). So it suffices to prove (•). If  $n \in \mathbb{N}$  then we are done because by the binomial theorem we have

$$z' = \sum_{1 \leq i \leq n} \binom{n}{i} z^i y^{n-i} = nzy^{n-1} + \cdots + z^n.$$

If  $-n \in \mathbb{N}_+$  then we are done by applying the geometric series identity to the previous case. In greater detail, if  $-n \in \mathbb{N}_+$  then by the previous case we can write

$$(y+z)^{-n} = \bar{y} + \bar{z} \quad \text{where } \bar{y} = y^{-n} \in K((T))^\times \quad \text{and} \quad \bar{z} \in K((T))$$

with

$$\text{inco}_T \bar{y} = \bar{\eta} \quad \text{and} \quad \text{inco}_T \bar{z} = \bar{\zeta}$$

are such that  $\text{sub}_T \bar{y} \subset \text{sub}_T y$  with

$$\text{ord}_T \bar{y} = -nv = \bar{v} < \bar{w} = (-n - 1)v + w \leq \text{ord}_T \bar{z}$$

and

$$\bar{w} - \bar{v} = w - v \quad \text{with} \quad \bar{\eta} = \eta^{-n}$$

and

$$\text{if } n \not\equiv 0 \pmod{\text{ch}(K)} \text{ then } \text{ord}_T \bar{z} = \bar{w} \text{ with } \bar{\zeta} = -n\zeta\eta^{-n-1}.$$

By the geometric series identity  $(1 + X)^{-1} = 1 - X + X^2 - \dots$  we get

$$(\bar{y} + \bar{z})^{-1} = \bar{y}^{-1} (1 + (\bar{z}/\bar{y}))^{-1} = \bar{y}^{-1} - \bar{z}\bar{y}^{-2} + \bar{z}^2\bar{y}^{-3} - \dots$$

and

$$\bar{y}^{-1} = (\bar{\eta}^{-1} T^{-\bar{v}}) (1 - (\bar{y}\bar{\eta}^{-1} T^{-\bar{v}} - 1) + (\bar{y}\bar{\eta}^{-1} T^{-\bar{v}} - 1)^2 - \dots)$$

with  $\text{sub}_T \bar{y}^{-1} \subset \text{sub}_T \bar{y}$ , and therefore the desired result follows by taking

$$y' = \bar{y}^{-1} \quad \text{and} \quad z' = -\bar{z}\bar{y}^{-2} + \bar{z}^2\bar{y}^{-3} - \dots$$

Second Binomial Lemma (3.4). Let  $K$  be a root-closed field and let us consider the univariate meromorphic series field  $K((T))$ . Let  $y \in K((T))^\times$  and  $z \in K((T))$  be such that

$$\text{ord}_T y = v < w = \text{ord}_T z$$

with

$$\text{inco}_T y = \eta \quad \text{and} \quad \text{inco}_T z = \zeta.$$

Let  $n = p/q$  where  $p$  and  $q$  are integers with  $q > 0$  and  $\text{GCD}(p, q) = 1$  such that  $q \not\equiv 0 \pmod{\text{ch}(K)}$  and  $pv \equiv 0 \pmod{q}$ . Then *mantrawise* (= briefly suggestively) we have (•) of (3.3), i.e., *bashyawise* (= precisely detailwise) we have that:

$$\left\{ \begin{array}{l} \text{there exists } y' \in K((T))^\times \text{ with } (y')^q = y^p \text{ and } \text{sub}_T y' \subset \text{sub}_T y \\ \text{and for any such } y' \text{ there exists } z' \in K((T))^\times \text{ such that} \\ \text{upon letting } x = y' + z' \text{ with } \text{inco}_T y' = \eta' \text{ and } \text{inco}_T z' = \zeta' \\ \text{we have } x^q = (y + z)^p \\ \text{with } \text{ord}_T y' = nv = v' < w' = (n - 1)v + w \leq \text{ord}_T z' \\ \text{and } w' - v' = w - v \text{ with } (\eta')^q = \eta^p \\ \text{and if } p \not\equiv 0 \pmod{\text{ch}(K)} \text{ then } \text{ord}_T z' = w' \text{ with } q\zeta' = p\zeta\eta'/\eta. \end{array} \right. \quad (\dagger)$$

Moreover, assuming  $p \not\equiv 0 \pmod{\text{ch}(K)}$ , we have (1\*)–(3\*) of (3.3).

*Proof.* (1\*)–(3\*) follow from (†). So it suffices to prove (†). Since the field  $K$  is root-closed, we have  $\xi^q = \eta^p$  for some  $\xi \in K$ . Applying Hensel's Lemma to  $Y^q - (1 + X)^p$  we get an identity in  $K[[X]]$  saying that

$$(1 + b_1X + b_2X^2 + \dots)^q = (1 + X)^p \quad \text{with } b_1, b_2, \dots \text{ in } K. \quad (1)$$

Differentiating both sides with respect to  $X$  and then putting  $X = 0$  we get

$$qb_1 = p. \quad (2)$$

Substituting  $X = y\eta^{-1}T^{-v} - 1$  in (1) and letting

$$y' = \xi (1 + b_1(y\eta^{-1}T^{-v} - 1) + b_2(y\eta^{-1}T^{-v} - 1)^2 + \dots) T^{(pv)/q}$$

we get  $y' \in K((T))^\times$  with  $(y')^q = y^p$  and  $\text{sub}_T y' \subset \text{sub}_T y$ .

Now let  $y'$  be any element of  $K((T))^\times$  such that

$$(y')^q = y^p \quad \text{and} \quad \text{sub}_T y' \subset \text{sub}_T y. \quad (3)$$

Then letting  $\text{inco}_T y' = \eta'$  we clearly get

$$(\eta')^q = \eta^p. \quad (4)$$

For any  $x \in K((T))$  we have

$$x^q = (y + z)^p \Leftrightarrow (x/y')^q = (1 + (z/y))^p$$

which follows by dividing the LHS by (3), and hence substituting  $X = z/y$  in (1) and letting

$$x = y'(1 + b_1(z/y) + b_2(z/y)^2 + \dots)$$

we get  $x \in K((T))^\times$  such that  $x^q = (y + z)^p$  and

$$x - y' = y'(b_1(z/y) + b_2(z/y)^2 + \dots). \quad (5)$$

Now letting  $z' = x - y'$  and  $\text{inco}_T z' = \zeta'$ , by (3)–(5) we see that  $z' \in K((T))$  and  $x = y' + z'$  with

$$\text{ord}_T y' = nv = v' < w' = (n - 1)v + w \leq \text{ord}_T z' \quad \text{and} \quad w' - v' = w - v$$

and

$$\text{if } p \not\equiv 0 \pmod{\text{ch}(K)} \text{ then } \text{ord}_T z' = w' \text{ with } q\zeta' = p\zeta\eta'/\eta.$$



*Note.* Lemma (3.3) was not used in Lemma (3.4). So the former was reproved in the latter. The former used the Binomial Theorem for integer exponents, while the latter used the Binomial Theorem for fractional exponents in disguise. Removing the disguise, Mantrawise, Lemma (3.4) follows by saying that by the Binomial Theorem for fractional exponents we have

$$(y + z)^n = y^n \left( 1 + n(y/z) + (n(n - 1)/2)(y/z)^2 + \dots \right)$$

and so we are done by taking  $y' = y^n$  and  $z' = (y + z)^n - y^n$ ; but care has to be taken when  $\text{ch}(K) \neq 0$ . In spite of what was said in the Introduction, we shall not directly use (3.4), i.e., we shall not explicitly use the Binomial Theorem for fractional exponents, but really it is lurking everywhere!!

Remark on special subfields (3.5). The essence of the above two Binomial Lemmas (3.3) and (3.4) is the Invariance of the Gap, i.e., the equation  $w' - v' = w - v$ , which underlies all the claims of [4] as well as their generalization in the present paper.

Now consider the univariate meromorphic series field  $K((T))$  over a field  $K$ .

The two cases (2\*) and (3\*) of (3.3) and (3.4) can be unified by introducing the notion of the  $(T, S)$ -gap  $v$  of  $y(T) = T^e \sum_{0 \leq i < \infty} A_i T^i$  with  $A_i \in K$  and  $A_0 \neq 0$ , where  $S$  is any subfield of  $K((T))$ , by putting  $v = \min\{i \in \mathbb{N} : A_i T^i \notin S\}$ , in case (2\*) we take  $S = K((T^d))$  and in case (3\*) we take  $S = k((T))$ . To include the ordinary gaps as in case (1\*), like the gap of length 4 between  $T$  and  $T^5$  in  $T + T^5 + T^6 + \dots$ , we have to allow  $S$  to be the null ring, which is not a subfield of  $K((T))$  under the usual convention. This is why we introduced the notion of a special subfield.

More generally, we define a *quasispecial subfield*  $S$  of  $K((T))$  to be either the nullring  $S = \{0\} \subset K((T))$  or a subfield  $S$  of  $K((T))$ ; if  $S \subset K$  then we may call  $S$  a *quasispecial subfield* of  $K$ . Now let  $S$  be a quasispecial subfield of  $K((T))$ . Given any  $y = y(T) \in K((T))^\times$  let

$$y(T) = T^e \sum_{0 \leq i < \infty} A_i T^i \quad \text{with } \text{ord}_{T,y}(T) = e \text{ and } A_i \in K \text{ with } A_0 \neq 0$$

and

$$v = \begin{cases} \min\{i \in \mathbb{N}_+ : A_i T^i \notin S\} & \text{if } S = \{0\} \\ \min\{i \in \mathbb{N} : A_i T^i \notin S\} & \text{if } S \neq \{0\} \end{cases}$$

with the convention that the minimum of the empty set of integers is  $\infty$ . We define the  $(T, S)$ -gap and the  $(T, S)$ -coefficient of  $y(T)$  by putting

$$\text{gap}_{(T,S)}y(T) = v \quad \text{and} \quad \text{coef}_{(T,S)}y(T) = \begin{cases} A_v & \text{if } v \neq \infty \\ 0 & \text{if } v = \infty. \end{cases}$$

We are particularly interested in the following cases (1<sup>#</sup>), (2<sup>#</sup>), (3<sup>#</sup>) of a quasispecial subfield  $S$  of  $K((T))$ ; note that in each of these cases  $S$  is a special subfield of  $K((T))$ .

$$\left\{ \begin{array}{l} S = \{0\}; \\ \text{note that then } v = \text{ord}_T(y(T)T^{-e} - A_0). \end{array} \right. \quad (1^\#)$$

$$\left\{ \begin{array}{l} S = K((T^d)) \text{ where } d \in \mathbb{N}_+; \\ \text{note that then } v = \min(\text{Supp}_T(y(T)T^{-e} - A_0) \setminus d\mathbb{Z}). \end{array} \right. \quad (2^\#)$$

$$\left\{ \begin{array}{l} S = k((T)) \text{ where } k \text{ is a nonnull special subfield of } K; \\ \text{note that then } v = \begin{cases} \min\{i \in \mathbb{N}_+ : A_i \notin k\} & \text{if } A_0 \in k \\ 0 & \text{if } A_0 \notin k. \end{cases} \end{array} \right. \quad (3^\#)$$

To prepare for proving the next Lemma (3.6), let  $S$  be a quasispecial subfield of  $K((T))$  and let  $y(T), z(T), x(T)$  in  $K[[T]]^\times$  be such that

$$y(T) = T \sum_{0 \leq i < \infty} A_i T^i \text{ with } \text{ord}_T y(T) = 1 \text{ and } \text{gap}_{(T,S)} y(T) = v$$

and

$$z(T) = T \sum_{0 \leq j < \infty} B_j T^j \text{ with } \text{ord}_T z(T) = 1 \text{ and } \text{gap}_{(T,S)} z(T) = w$$

and

$$x(T) = y(z(T)) = T \sum_{0 \leq l < \infty} C_l T^l \text{ with } \text{ord}_T x(T) = 1 \text{ and } \text{gap}_{(T,S)} x(T) = \pi$$

where  $A_i, B_j, C_l$  are in  $K$  with

$$A_0 \neq 0 \neq B_0 \neq 0 \neq C_0$$

and where we note that now  $e = 1$ . For  $0 \leq l < \infty$  we clearly have

$$C_l T^l = \sum_{0 \leq i \leq l} \left( A_i T^i \times \text{the term of } T\text{-degree } l - i \text{ in } \left( \sum_{0 \leq j \leq l-i} B_j T^j \right)^{i+1} \right)$$

and hence

$$C_l T^l = \begin{cases} A_0 B_l T^l + B_0 A_l T^l + \sum_{0 < i < l} A_i T^i D_{il} & \text{if } l \neq 0 \\ A_0 B_0 & \text{if } l = 0 \end{cases} \quad (I)$$

with

$$D_{il} = \sum^* M_\lambda \prod_{0 \leq j \leq l-i} (B_j T^j)^{\lambda_j} \tag{II}$$

where  $\sum^*$  indicates summation over all  $(\lambda_0, \dots, \lambda_{l-i}) \in \mathbb{N}^{l-i+1}$  for which

$$\sum_{0 \leq j \leq l-i} j \lambda_j = l - i \tag{III}$$

and  $M_\lambda$  is the multinomial coefficient

$$M_\lambda = \frac{(i + 1)!}{\lambda_0! \dots \lambda_{l-i}!}.$$

We shall now prove the following assertions:

- $$\left\{ \begin{array}{l} (1) \text{ If } 0 < i < l < \infty \text{ with } l \leq \min(v, w) \text{ then } A_i T^i D_{il} \in S. \\ (2) \pi \geq \min(v, w). \\ (3) \text{ If } v < w \text{ then } \pi = v \text{ and } C_v T^v - B_0 A_v T^v \in S. \\ (4) \text{ If } w < v \text{ then } \pi = w \text{ and } C_w T^w - A_0 B_w T^w \in S. \\ (5) C_0 - A_0 B_0 = 0 \in S. \\ (6) \text{ If } 0 \neq v = w \neq \infty \text{ then } C_v T^v - (A_0 B_v T^v + B_0 A_v T^v) \in S. \\ (7) \text{ If } x(T) = T \text{ then } v = w \text{ and } A_0 B_0 = 1. \\ (8) \text{ If } x(T) = T \text{ and } 0 \neq v = w \neq \infty \text{ then } A_0 B_v T^v + B_0 A_v T^v \in S. \end{array} \right. \tag{IV}$$

To prove (1) let  $0 < i < l < \infty$  with  $l \leq \min(v, w)$ . Since  $0 < i < l \leq v$ , we get  $A_i T^i \in S$ . If  $S = \{0\}$  then  $A_i T^i = 0$  and hence  $A_i T^i D_{il} = 0 \in S$ . If  $S \neq \{0\}$  then  $1 \in S$  and because  $i < l \leq w$ , every term in each product involved in (II) belongs to  $S$ , and hence again  $A_i T^i D_{il} \in S$ .

To prove (2) let  $0 \leq l < \min(v, w)$ . Then  $A_0 B_l T^l \in S$  and  $B_0 A_l T^l \in S$ , and hence by (I) and (1) we get  $C_l T^l \in S$ . It follows that  $\pi \geq \min(v, w)$ .

To prove (3) let  $v < w$ . If  $v \neq 0$  then  $A_0 B_v T^v \in S$  and hence by (I) and (1) we get  $C_v T^v - B_0 A_v T^v \in S$ ; but  $B_0 \in S^\times$  with  $A_v T^v \notin S$  and therefore  $C_v T^v \notin S$ ; consequently by (2) we see that  $\pi = v$ . If  $v = 0$  then  $A_0 \notin S$  with  $B_0 \in S$  and hence by (I) we get  $C_0 - B_0 A_0 = 0 \in S$  with  $C_0 \notin S$  and therefore  $\pi = 0 = v$ .

To prove (4) let  $w < v$ . If  $w \neq 0$  then  $B_0 A_w T^w \in S$  and hence by (I) and (1) we get  $C_w T^w - A_0 B_w T^w \in S$ ; but  $A_0 \in S^\times$  with  $B_w T^w \notin S$  and therefore  $C_w T^w \notin S$ ; consequently by (2) we see that  $\pi = w$ . If  $w = 0$  then  $B_0 \notin S$  with  $A_0 \in S$  and hence by (I) we get  $C_0 - A_0 B_0 = 0 \in S$  with  $C_0 \notin S$  and therefore  $\pi = 0 = w$ .

By (I) we obviously get (5). By (I) and (1) we see that if  $0 \neq v = w \neq \infty$  then  $C_v T^v - (A_0 B_v T^v + B_0 A_v T^v) \in S$ , which proves (6).

If  $x(T) = T$  then  $\pi = \infty$  with  $C_0 = 1$ , and hence by (3) and (4) we get  $v = w$  and by (5) we get  $A_0 B_0 = 1$ , which proves (7).

If  $x(T) = T$  and  $0 \neq v = w \neq \infty$  then  $\pi = \infty$ , and hence by (6) we get  $A_0 B_v T^v + B_0 A_v T^v \in S$ , which proves (8).

Gap Lemma (3.6). Consider the univariate meromorphic series field  $K((T))$  over a root-closed field  $K$ . Let  $y(T)$  and  $z(T)$  in  $K((T))^\times$  with

$$\text{ord}_T y(T) = e \neq 0 \neq \epsilon = \text{ord}_T z(T)$$

be such that

$$y(T) = T^e \sum_{0 \leq i < \infty} A_i T^i \quad \text{and} \quad z(T) = T^\epsilon \sum_{0 \leq j < \infty} B_j T^j$$

where

$$A_i \text{ and } B_j \text{ are in } K \text{ with } A_0 \neq 0 \neq B_0.$$

Assume that  $e \not\equiv 0 \pmod{\text{ch}(K)}$  and  $\epsilon \not\equiv 0 \pmod{\text{ch}(K)}$ . Then by Hensel's Lemma, there exist  $\hat{y}(T)$  and  $\hat{z}(T)$  in  $K[[T]]^\times$  with

$$\text{ord}_T \hat{y}(T) = 1 = \text{ord}_T \hat{z}(T)$$

such that

$$\hat{y}(T)^e = y(T) \quad \text{and} \quad \hat{z}(T)^\epsilon = z(T)$$

and

$$\hat{y}(T) = T \sum_{0 \leq i < \infty} \hat{A}_i T^i \quad \text{and} \quad \hat{z}(T) = T \sum_{0 \leq j < \infty} \hat{B}_j T^j$$

where

$$\hat{A}_i \text{ and } \hat{B}_j \text{ are in } K \text{ with } \hat{A}_0^e = A_0 \neq 0 \neq B_0 = \hat{B}_0^\epsilon.$$

Given any *special subfield*  $S$  of  $K((T))$  let

$$\text{gap}_{(T,S)} y(T) = v \quad \text{with} \quad \text{gap}_{(T,S)} \hat{y}(T) = \hat{v}$$

and

$$\text{gap}_{(T,S)} z(T) = w \quad \text{with} \quad \text{gap}_{(T,S)} \hat{z}(T) = \hat{w}.$$

Assume that

$$v \neq 0 \neq w.$$

Then

$$v = \hat{v} \quad \text{with} \quad \text{coef}_{(T,S)} y(T) = (\text{coef}_{(T,S)} \hat{y}(T)) e \hat{A}_0^{e-1} \quad (1)$$

and

$$w = \hat{w} \quad \text{with} \quad \text{coef}_{(T,S)} z(T) = (\text{coef}_{(T,S)} \hat{z}(T)) \epsilon \hat{B}_0^{\epsilon-1}. \quad (2)$$

Moreover,

$$\text{if } \widehat{y}(\widehat{z}(T)) = T \text{ then } v = w \text{ and } \widehat{A}_0 \widehat{B}_0 = 1 \quad (3)$$

and

$$\left\{ \begin{array}{l} \text{if } \widehat{y}(\widehat{z}(T)) = T \text{ and } \infty \neq v = w \neq \infty \text{ then} \\ (\text{coef}_{(T,S)} z(T)) e \widehat{A}_0^\epsilon T^v + (\text{coef}_{(T,S)} y(T)) \epsilon \widehat{B}_0^\epsilon T^v \in S. \end{array} \right. \quad (4)$$

*Proof.* (1) follows from (3.3) by noting that  $\widehat{v} > 0$  and taking

$$(n, v, w, y, z) = \left( e, 1, \widehat{v} + 1, T \sum_{0 \leq i < \widehat{v}} \widehat{A}_i T^i, T \sum_{\widehat{v} \leq i < \infty} \widehat{A}_i T^i \right)$$

and (2) follows from (3.3) by noting that  $\widehat{w} > 0$  and taking

$$(n, v, w, y, z) = \left( \epsilon, 1, \widehat{w} + 1, T \sum_{0 \leq i < \widehat{w}} \widehat{B}_i T^i, T \sum_{\widehat{w} \leq i < \infty} \widehat{B}_i T^i \right).$$

By (3.5)(IV)(7), we see that

$$\text{if } \widehat{y}(\widehat{z}(T)) = T \text{ then } \widehat{v} = \widehat{w} \text{ and } \widehat{A}_0 \widehat{B}_0 = 1 \quad (')$$

and by (3.5)(IV)(8) we see that

$$\left\{ \begin{array}{l} \text{if } \widehat{y}(\widehat{z}(T)) = T \text{ and } \widehat{v} = \widehat{w} \neq \infty \text{ then} \\ (\text{coef}_{(T,S)} \widehat{z}(T)) \widehat{A}_0 T^{\widehat{v}} + (\text{coef}_{(T,S)} \widehat{y}(T)) \widehat{B}_0 T^{\widehat{v}} \in S. \end{array} \right. \quad (')$$

Now, in view of (1) and (2), by (') we get (3), and by (') we get (4).

**Remark on gap lemma (3.7).** We shall now paraphrase (3.6) by using the language of DVRs.

So let  $V$  be a DVR with

$$V \subset \widehat{V} = \text{the completion of } V \text{ and } \text{QF}(V) = L \subset \widehat{L} = \text{QF}(\widehat{V}).$$

Let  $T$  be a uniformizing parameter of  $\widehat{V}$ . Assume that  $\text{ch}(L) = \text{ch}(H(V))$  and let  $K$  be a coefficient field of  $\widehat{V}$ . Note that then  $\widehat{V} = K((T))$ . Assume that  $H(V)$ , and hence  $K$ , is root-closed.

Given any  $y = y(T) \in K((T))^\times$  and  $z = z(T) \in K((T))^\times$  let

$$\text{ord}_T y = e \text{ with } \text{inco}_T y = A \text{ and } \text{ord}_T z = \epsilon \text{ with } \text{inco}_T z = B.$$

Since  $K$  is root-closed, we can choose

$$\widehat{A} \in K^\times \text{ with } (\widehat{A})^e = A \text{ and } \widehat{B} \in K^\times \text{ with } (\widehat{B})^\epsilon = B.$$

Assuming  $e \not\equiv 0 \pmod{\text{ch}(K)}$ , with the chosen  $\widehat{A}$ , by Hensel's Lemma there exists a unique  $\widehat{y} = \widehat{y}(T) \in K((T))^\times$  such that

$$(\widehat{y})^e = y \quad \text{and} \quad \text{ord}_T \widehat{y} = 1 \quad \text{with} \quad \text{inco}_T \widehat{y} = \widehat{A}.$$

Clearly  $\theta(T) \mapsto \theta(\widehat{y}(T))$  gives an automorphism  $K((T)) \rightarrow K((T))$  and hence there exists a unique  $\tilde{z} = \tilde{z}(T) \in K((T))^\times$  such that

$$\tilde{z}(\widehat{y}(T)) = z(T).$$

We call  $\tilde{z} = \tilde{z}(T)$  the  $(V, K, T)$ -*expansion* of  $z$  in terms of  $y$  relative to  $\widehat{A}$ , or briefly we call  $\tilde{z} = \tilde{z}(T)$  the  $(V, K, T)$ -*expansion* of  $(z, y, \widehat{A})$ . Concerning the dependence of this expansion on  $\widehat{A}$ , let us note that

$$\left\{ \begin{array}{l} \text{if } \widehat{A}^* \text{ is any other member of } K \text{ with } (\widehat{A}^*)^e = A \\ \text{then } \omega = \widehat{A}^*/\widehat{A} \text{ is an } e\text{-th root of 1 in } K \\ \text{and for the } (V, K, T)\text{-expansion } \tilde{z}^* \text{ of } (z, y, \widehat{A}^*) \\ \text{we have } \tilde{z}^*(T) = \tilde{z}(\omega T) \\ \text{and hence } \text{Supp}_T \tilde{z}^*(T) = \text{Supp}_T \tilde{z}(T). \end{array} \right. \quad (\text{b})$$

Assuming  $e \not\equiv 0 \pmod{\text{ch}(K)}$  but without assuming any condition on  $\epsilon$ , with the chosen  $\widehat{A}$ , in view of (b) we may put

$$m(z, y, V, K) = m(\text{Supp}_T \tilde{z}(T), e)$$

(because  $\text{Supp}_T \tilde{z}(T)$  is independent of  $\widehat{A}$ ) and call it the  $(V, K)$ -*charseq* of  $(z, y)$ .

Also assuming  $\epsilon \not\equiv 0 \pmod{\text{ch}(K)}$ , with the chosen  $\widehat{B}$ , by Hensel's Lemma there exists a unique  $\widehat{z} = \widehat{z}(T) \in K((T))^\times$  such that

$$(\widehat{z})^\epsilon = z \quad \text{and} \quad \text{ord}_T \widehat{z} = 1 \quad \text{with} \quad \text{inco}_T \widehat{z} = \widehat{B}.$$

Clearly  $\theta(T) \mapsto \theta(\widehat{z}(T))$  gives an automorphism  $K((T)) \rightarrow K((T))$  and hence there exists a unique  $\widetilde{y} = \widetilde{y}(T) \in K((T))^\times$  such that

$$\widetilde{y}(\widehat{z}(T)) = y(T).$$

Note that now  $\widetilde{y} = \widetilde{y}(T)$  is the  $(V, K, T)$ -*expansion* of  $(y, z, \widehat{B})$ .

Again clearly there exist unique  $z^\dagger(T)$  and  $y^\dagger(T)$  in  $K((T))$  such that

$$z^\dagger(\widehat{y}(T)) = \widehat{z}(T) \quad \text{and} \quad y^\dagger(\widehat{z}(T)) = \widehat{y}(T). \quad (\bullet)$$

Substituting the first equation of (●) in its second, we get

$$y^\dagger(z^\dagger(\widehat{y}(T))) = \widehat{y}(T)$$

and hence

$$y^\dagger(z^\dagger(T)) = T \tag{1}$$

Raising the second equation of (●) to the  $e$ -th power and the first to the  $\epsilon$ -th power we get

$$y^\dagger(T)^e = \widetilde{y}(T) \quad \text{and} \quad z^\dagger(T)^\epsilon = \widetilde{z}(T). \tag{2}$$

By the first equation of (●) we get

$$\text{ord}_T z^\dagger(T) = 1 \quad \text{with} \quad \text{inco}_T z^\dagger(T) = \widehat{B}/\widehat{A} \tag{3}$$

and by the second equation of (●) we get

$$\text{ord}_T y^\dagger(T) = 1 \quad \text{with} \quad \text{inco}_T y^\dagger(T) = \widehat{A}/\widehat{B}. \tag{4}$$

Now we claim the FIRST INVERSION THEOREM which says that:

$$\left\{ \begin{array}{l} \text{Given any special subfield } S \text{ of } K((T)), \\ \text{upon letting } \text{gap}_{(T,S)} \widetilde{y}(T) = v \text{ and } \text{gap}_{(T,S)} \widetilde{z}(T) = w, \\ \text{we have the following.} \\ (1^*) \text{ If } v \neq 0 \neq w \text{ then } v = w. \\ (2^*) \left\{ \begin{array}{l} \text{If } \infty \neq v \neq 0 \neq w \neq \infty \text{ then} \\ (\text{coef}_{(T,S)} \widetilde{z}(T)) e \widehat{A}^{e+\epsilon} T^v + (\text{coef}_{(T,S)} \widetilde{y}(T)) \epsilon \widehat{B}^{\epsilon+e} T^v \in S. \end{array} \right. \\ (3^*) \text{ If } S = K((T^d)) \text{ for some } d \in \mathbb{N}_+ \text{ then } 0 \neq v = w \neq 0. \end{array} \right. \tag{I}$$

Namely, in view of (1)–(4), by (3.6)(3) and (3.6)(4) we obtain (1\*) and (2\*) respectively. If  $S \neq \{0\}$  then by definition  $v \neq 0 \neq w$  and hence by (1\*) we get (3\*).

Next we claim the SECOND INVERSION THEOREM which says that:

$$\left\{ \begin{array}{l} \text{Upon letting } m = m(z, y, V, K) \text{ and } m' = m(y, z, V, K) \\ \text{we have } 0 \neq h(m) = h(m') \neq 0 \\ \text{and } e = m_0 = m'_1 \text{ with } \epsilon = m'_0 = m_1 \\ \text{and } m_\mu - \epsilon = m'_\mu - e \text{ for } 2 \leq \mu \leq h(m) + 1 \\ \text{and } d_1(m) = |e| \text{ with } d_1(m') = |\epsilon| \\ \text{and } d_2(m) = d_2(m') = \text{GCD}(e, \epsilon) \\ \text{and } d_\mu(m) = d_\mu(m') \text{ for } 2 \leq \mu \leq h(m) + 2. \end{array} \right. \tag{II}$$

Namely, everything is obvious except the assertion  $h(m) = h(m')$  together with the assertions that for  $2 \leq \mu \leq h(m) + 1$  we have

$$m_\mu - \epsilon = m'_\mu - e \quad \text{and} \quad d_{\mu+1}(m) = d_{\mu+1}(m').$$

Clearly the assertions about  $h(m)$  and  $d_{\mu+1}(m)$  follow from the assertion about  $m_\mu - \epsilon$ . By induction on  $\mu$  let us prove that for  $1 \leq \mu \leq h(m) + 1$  we have

$$m_\mu - \epsilon = m'_\mu - e.$$

For  $\mu = 1$  this is line 3 of (II). To go from  $\mu$  to  $\mu + 1$  can be achieved by taking

$$d = d_{\mu+1}(m)$$

in (I)(3\*). This completes the proof of (II).

Remark on valuation protoexpansions (3.8). To merge Remarks (3.1), (3.2), and (3.7), let  $V$  be a DVR with

$$V \subset \widehat{V} = \text{the completion of } V \quad \text{and} \quad \text{QF}(V) = L \subset \widehat{L} = \text{QF}(\widehat{V}).$$

Let  $T$  be a uniformizing parameter of  $\widehat{V}$ . Assume that  $\text{ch}(L) = \text{ch}(H(V)) = 0$  and let  $K$  be a coefficient field of  $\widehat{V}$ . Note that then  $\widehat{V} = K((T))$ . Assume that  $H(V)$ , and hence  $K$ , is root-closed.

Given any pair of elements  $(z_0, z_1)$  in  $L^\times$  with  $\text{ord}_V z_1 \neq 0$ , by (3.1) and (3.7) there exists a system

$$(z_j, e_j, p_j, A_l^*(v), e_l^*, z_l^*, A_j(v), \widehat{A}_j, \tilde{A}_j(v), m^{(j)}, \tilde{z}_j)_{v \in \mathbb{Z}, 0 \leq j \leq l+1}$$

where

$$(z_j, e_j, p_j, A_l^*(v), e_l^*, z_l^*)_{v \in \mathbb{Z}, 0 \leq j \leq l+1}$$

is the  $(V, K)$ -protoexpansion of  $(z_0, z_1)$  as described in (3.1)(1)–(3.1)(8) with  $A_j(v)$  as in (3.1)(9) and (3.1)(10), and

$$\text{for } 0 \leq j \leq l + 1$$

we have

$$\begin{cases} \widehat{A}_j \in K^\times \text{ with } (\widehat{A}_j)^{e_j} = \text{inco}_T z_j & \text{if } j \neq l + 1, \\ \widehat{A}_j \in K^\times \text{ with } (\widehat{A}_j)^{e_j} = \text{inco}_T z_j & \text{if } j = l + 1 \text{ and } z_{l+1} \neq 0, \\ \widehat{A}_j = 0 \in K & \text{if } j = l + 1 \text{ and } z_{l+1} = 0, \end{cases} \quad (1)$$



and

$$m^{(j)} = \begin{cases} m(z_j, z_{j+1}, V, K) & \text{if } l \neq j \neq l + 1 \\ m(z_j, z_{j+1}, V, K) & \text{if } j = l \text{ and } z_{l+1} \neq 0 \\ m(z_{j-1}^*, z_{j-1}, V, K) & \text{if } j = l + 1 \text{ and } z_{l+1} \neq 0 \\ m(\emptyset, 1) & \text{if } l \leq j \leq l + 1 \text{ and } z_{l+1} = 0 \end{cases} \quad (2)$$

and

$$\tilde{z}_j = \tilde{z}_j(T) = \sum_{v \in \mathbb{Z}} \tilde{A}_j(v) T^v \quad \text{with } \tilde{A}_j(v) \in K \quad (3)$$

is the  $(V, K, T)$ -expansion of  $(z_j, z_{j+1}, \widehat{A}_{j+1})$  in case  $l \neq j \neq l + 1$ , and in the remaining cases:

$$\begin{cases} \text{if } j = l \text{ and } z_{l+1} \neq 0 \\ \text{then (3) is the } (V, K, T)\text{-expansion of } (z_j, z_{j+1}, \widehat{A}_{j+1}) \end{cases} \quad (4)$$

and

$$\begin{cases} \text{if } j = l + 1 \text{ and } z_{l+1} \neq 0 \\ \text{then (3) is the } (V, K, T)\text{-expansion of } (z_{j-1}^*, z_{j-1}, \widehat{A}_{j-1}) \end{cases} \quad (5)$$

and

$$\begin{cases} \text{if } l \leq j \leq l + 1 \text{ and } z_{l+1} = 0 \text{ then in (3) we take} \\ \tilde{z}_j = \tilde{z}_j(T) = 0 = \tilde{A}_j(v) \text{ for all } v \in \mathbb{Z}; \end{cases} \quad (6)$$

we call such a system a *mixed  $(V, K, T)$ -protoexpansion* of  $(z_0, z_1)$ . It follows that

$$\begin{cases} \text{if } z_{l+1} \neq 0 \text{ then } e_l^* = p_l e_l + e_{l+1} = m_2^{(l-1)} \\ \text{and } \tilde{A}_{l-1}(v) = \begin{cases} 0 & \text{if } e_l^* > v \not\equiv 0 \pmod{e_l} \\ A_l^*(v/|e_l|) & \text{if } e_l^* > v \equiv 0 \pmod{e_l} \\ \tilde{A}_{l+1}(v - p_l e_l) & \text{if } e_l^* \leq v; \end{cases} \end{cases} \quad (7)$$

In view of (3.1)(5),

$$\text{for } 0 \leq j \leq l - 2$$

upon letting

$$\begin{cases} \check{A}_{j+2}(v) \in K \text{ for all } v \in \mathbb{Z} \text{ such that} \\ \check{A}_{j+2}(v) = 0 \text{ for } v < e_{j+2} \text{ and } \check{A}_{j+2}(e_{j+2}) \neq 0 \\ \text{and } \check{z}_{j+2} = \check{z}_{j+2}(T) = \sum_{e_{j+2} \leq v < \infty} \check{A}_{j+2}(v) T^v \\ \text{is the } (V, K, T)\text{-expansion of } (z_{j+2}, z_{j+1}, \widehat{A}_{j+1}) \end{cases} \quad (1_j)$$

we have

$$\check{A}_{j+2}(v + e_{j+2}) = \check{A}_j(v + e_j) \text{ for all } v \in \mathbb{Z} \quad (2_j)$$

and hence

$$\left\{ \begin{array}{l} \text{upon letting } \check{m}^{(j+2)} = m(z_{j+2}, z_{j+1}, V, K) \\ \text{we have } 0 \neq h(m^{(j)}) = h(\check{m}^{(j+2)}) \neq 0 \\ \text{and } e_{j+1} = m_0^{(j)} = \check{m}_0^{(j+2)} \text{ and } e_j = m_1^{(j)} \text{ with } e_{j+2} = \check{m}_1^{(j+2)} \\ \text{and } m_\mu^{(j)} - e_j = \check{m}_\mu^{(j+2)} - e_{j+2} \text{ for } 1 \leq \mu \leq h(m^{(j)}) + 1 \\ \text{and } d_1(m^{(j)}) = d_1(\check{m}^{(j+2)}) = |e_{j+1}| \\ \text{and } d_2(m^{(j)}) = d_2(\check{m}^{(j+2)}) = \text{GCD}(e_j, e_{j+1}) = \text{GCD}(e_{j+1}, e_{j+2}) \\ \text{and } d_\mu(m^{(j)}) = d_\mu(\check{m}^{(j+2)}) \text{ for } 1 \leq \mu \leq h(m^{(j)}) + 2. \end{array} \right. \quad (3_j)$$

and, in view of (3<sub>j</sub>), by taking  $(z, y) = (z_{j+1}, z_{j+2})$  in (2.7)(II) we see that

$$\left\{ \begin{array}{l} 0 \neq h(m^{(j)}) = h(m^{(j+1)}) \neq 0 \\ \text{and } e_{j+1} = m_0^{(j)} = m_1^{(j+1)} \text{ and } e_j = m_1^{(j)} \text{ with } e_{j+2} = m_0^{(j+1)} \\ \text{and } m_\mu^{(j)} - e_j = m_\mu^{(j+1)} - e_{j+1} \text{ for } 2 \leq \mu \leq h(m^{(j)}) + 1 \\ \text{and } d_1(m^{(j)}) = |e_{j+1}| \text{ with } d_1(m^{(j+1)}) = |e_{j+2}| \\ \text{and } d_2(m^{(j)}) = d_2(m^{(j+1)}) = \text{GCD}(e_j, e_{j+1}) = \text{GCD}(e_{j+1}, e_{j+2}) \\ \text{and } d_\mu(m^{(j)}) = d_\mu(m^{(j+1)}) \text{ for } 2 \leq \mu \leq h(m^{(j)}) + 2. \end{array} \right. \quad (4_j)$$

Now (4<sub>0</sub>) + ⋯ + (4<sub>j-1</sub>) ⇒

$$\left\{ \begin{array}{l} \text{for } 0 \leq j \leq l-1 \text{ we have } 0 \neq h(m^{(0)}) = h(m^{(j)}) \neq 0 \\ \text{and } e_1 = m_0^{(0)} \text{ and } e_0 = m_1^{(0)} \text{ with } e_{j+1} = m_0^{(j)} \text{ and } e_j = m_1^{(j)} \\ \text{and } m_\mu^{(0)} - e_0 = m_\mu^{(j)} - e_j \text{ for } 2 \leq \mu \leq h(m^{(0)}) + 1 \\ \text{and } d_1(m^{(0)}) = |e_1| \text{ with } d_1(m^{(j)}) = |e_{j+1}| \\ \text{and } d_2(m^{(0)}) = d_2(m^{(j)}) = \text{GCD}(e_0, e_1) = \text{GCD}(e_j, e_{j+1}) \\ \text{and } d_\mu(m^{(0)}) = d_\mu(m^{(j)}) \text{ for } 2 \leq \mu \leq h(m^{(0)}) + 2. \end{array} \right. \quad (I)$$

Moreover, in view of (2)–(7) we see that

$$\left\{ \begin{array}{l} \text{if } z_{l+1} \neq 0 \text{ then } h(m^{(l+1)}) = h(m^{(l-1)}) - 1 \\ \text{and } e_l = m_0^{(l+1)} = m_0^{(l-1)} \\ \text{and } p_l e_l + e_{l+1} = m_1^{(l+1)} = m_2^{(l-1)} \text{ with } e_{l-1} = m_1^{(l-1)} \\ \text{and } m_\mu^{(l+1)} = m_{\mu+1}^{(l-1)} \text{ for } 2 \leq \mu \leq h(m^{(l+1)}) + 1 \\ \text{and } d_1(m^{(l+1)}) = |e_l| \\ \text{and } d_1(m^{(l-1)}) = |e_l| = d_2(m^{(l-1)}) = \text{GCD}(e_l, e_{l-1}) \\ \text{and } d_2(m^{(l+1)}) = d_3(m^{(l-1)}) = \text{GCD}(e_l, e_{l+1}) \\ \text{and } d_\mu(m^{(l+1)}) = d_{\mu+1}(m^{(l-1)}) \text{ for } 2 \leq \mu \leq h(m^{(l+1)}) + 2. \end{array} \right. \quad (II)$$

Preamble for next lemma. Having dealt with case (3.5)(2<sup>#</sup>), turning to case (3.5)(3<sup>#</sup>), let

$$\left\{ \begin{array}{l} S = k((T)) \text{ where } k \text{ is a nonnull special subfield of } K, \\ \theta = \text{an unspecified member of } k^\times \\ (\theta \text{ is called Abhyankar's nonzero and may be read as } \theta), \\ \theta' = \text{an unspecified member of } k \\ (\theta' \text{ is called Abhyankar's constant and may be read as } \theta'), \\ \text{gap}_{(T,S)} \tilde{z}_j(T) = v_j \text{ with } \text{coef}_{(T,S)} \tilde{z}_j(T) = \bar{A}_j \text{ for } 0 \leq j \leq l \\ \text{with the understanding that if } z_{l+1} = 0 \text{ then } v_l = \infty \text{ and } \bar{A}_l = 0, \end{array} \right. \quad (8)$$

and let

$$z_l^\dagger = \sum_{(e_{l-1}/e_l) \leq v < (v_{l-1} + e_{l-1})/e_l} A_l^*(v) z_l^{v(e_l/e_l)} \in K[z_l, z_l^{-1}] \quad (9)$$

and

$$z_l^b = z_{l-1} - z_l^\dagger \in L \text{ with } \text{ord}_V z_l^b = e_l^b \quad (10)$$

and let

$$z_l^b = z_l^b(T) = \sum_{v \in \mathbb{Z}} A_l^b(v) T^v \text{ with } A_l^b(v) \in K \quad (11)$$

be the usual expansion in  $K((T))$  and

$$\left\{ \begin{array}{l} \text{if } z_l^b \neq 0 \\ \text{then let } \tilde{z}_l^b = \tilde{z}_l^b(T) \text{ be the } (V, K, T)\text{-expansion of } (z_l^b, z_l, \widehat{A}_l) \\ \text{and let } \text{gap}_{(T,S)} \tilde{z}_l^b(T) = v_l^b \text{ with } \text{coef}_{(T,S)} \tilde{z}_l^b(T) = \bar{A}_l^b. \end{array} \right. \quad (12)$$

Finally let

$$z_l^{bb} = z_{l-1}/z_l^{(e_{l-1}-e_l)/e_l} \in L^\times \text{ with } \text{ord}_V z_l^{bb} = e_l^{bb} \quad (13)$$

and note that then

$$e_l^{bb} = e_l. \quad (14)$$

With the above notation at hand, we shall now prove the:

Coefficient lemma (III). We have the following.

(1\*) If  $\widehat{A}_1 \in k$  and  $v_0 > 0$  then for  $0 \leq j \leq l$  we have  $\widehat{A}_j \in k$ , and for  $0 \leq j \leq l-1$  we have  $v_j = v_0$  with  $\bar{A}_j = \theta \bar{A}_0 + \theta'$ .

(2\*)  $z_{l+1} \neq 0 \Leftrightarrow m_2^{(l-1)} \neq \infty \Rightarrow m_2^{(l-1)} = p_l e_l + e_{l+1}$ .

(3\*) If  $\widehat{A}_l \in k$  and  $v_{l-1} = \infty$  then  $v_l = \infty$  and  $\widehat{A}_{l+1} \in k$  with  $\tilde{z}_{l-1}(T) \in k((T))$  and  $A_l^*(v) \in k$  for all  $v \in \mathbb{Z}$ .

(4\*) If  $\widehat{A}_l \in k$  and  $v_{l-1} + m_1^{(l-1)} > m_2^{(l-1)}$  then  $z_{l+1} \neq 0$  with  $\widehat{A}_{l+1} \in k$  and  $v_l + m_2^{(l-1)} = v_{l-1} + m_1^{(l-1)}$  with  $\bar{A}_l = \theta \bar{A}_{l-1} + \theta'$ .

- (5\*) If  $\widehat{A}_l \in k$  and  $v_{l-1} + m_1^{(l-1)} < m_2^{(l-1)}$  then  $z_l^b \neq 0$  and  $v_l^b = 0$  with  $\bar{A}_l^b = \bar{A}_{l-1}$  and  $e_l^b = v_{l-1} + e_{l-1} \equiv 0 \pmod{(e_l)}$ .
- (6\*) If  $\widehat{A}_l \in k$  and  $v_{l-1} \neq \infty$  with  $v_{l-1} + m_1^{(l-1)} = m_2^{(l-1)}$  then  $z_{l+1} \neq 0 \neq z_l^b$  and  $v_l^b = 0$  with  $\bar{A}_l^b = \bar{A}_{l-1}$  and  $e_l^b = v_{l-1} + e_{l-1} \not\equiv 0 \pmod{(e_l)}$ .
- (7\*) If  $\widehat{A}_1 \in k$  and  $v_0 = 0$  then  $\text{inco}_T z_0 = \theta \bar{A}_0 \in K \setminus k$  and

$$\frac{\text{inco}_T z_l^{\text{bb}}}{\text{inco}_T z_l} = \theta \bar{A}_0^E \quad \text{with } E = (-1)^{l+1} (e_1/e_l) \in \mathbb{Z}^\times.$$

Prenote. In the statements as well as proofs of (1\*)–(7\*), some quantity such as  $v_{l-1}$  may take the value  $\infty$ , and then the reader is advised to follow the infinity convention described in the second paragraph of (3.2).

*Note.* In the following proofs of (1\*)–(7\*), we shall frequently invoke two obvious but very useful principles which in the context of (8) may be stated thus. The MP = MULTIPLICATIVE PRINCIPLE says that if  $\widehat{A}_{j+1} \in k$  and  $z^\sharp \in L$  is such that  $z^\sharp = \theta z_j z_{j+1}^p$  with  $p \in \mathbb{Z}$  then:  $z^\sharp \neq 0$  and upon letting  $\tilde{z}^\sharp(T)$  be the  $(V, K, T)$ -expansion of  $(z^\sharp, z_{j+1}, \widehat{A}_{j+1})$  and putting

$$\text{gap}_{(T,S)} \tilde{z}^\sharp(T) = v^\sharp \quad \text{with } \text{coef}_{(T,S)} \tilde{z}^\sharp(T) = \bar{A}^\sharp$$

we have

$$v^\sharp = v_j \quad \text{with } \bar{A}^\sharp = \theta \bar{A}_j \quad \text{and: if } v_j > 0 \text{ then } \{\text{inco}_T z^\sharp, \text{inco}_T z_j\} \subset k.$$

The AP = ADDITIVE PRINCIPLE says that if  $\widehat{A}_{j+1} \in k$  and  $z^{\sharp\sharp} \in L$  is such that

$$z^{\sharp\sharp} = z_j - \sum_{v \leq \theta} \tilde{A}_j(v) \widehat{z}_{j+1}^v \quad \text{where } \theta \in \mathbb{Z} \text{ with } \theta < v_j + e_j$$

and where  $\widehat{z}_{j+1} = \widehat{z}_{j+1}(T) \in K((T))$  is such that

$$\widehat{z}_{j+1}^{e_{j+1}} = z_{j+1} \quad \text{and } \text{inco}_T \widehat{z}_{j+1} = \widehat{A}_{j+1}$$

then:  $z^{\sharp\sharp} \neq 0$  and upon letting  $\text{ord}_V z^{\sharp\sharp} = e^{\sharp\sharp}$  and upon letting  $\tilde{z}^{\sharp\sharp}(T)$  be the  $(V, K, T)$ -expansion of  $(z^{\sharp\sharp}, z_{j+1}, \widehat{A}_{j+1})$  and putting

$$\text{gap}_{(T,S)} \tilde{z}^{\sharp\sharp}(T) = v^{\sharp\sharp} \quad \text{with } \text{coef}_{(T,S)} \tilde{z}^{\sharp\sharp}(T) = \bar{A}^{\sharp\sharp}$$

we have

$$v^{\sharp\sharp} + e^{\sharp\sharp} = v_j + e_j \quad \text{with } \bar{A}^{\sharp\sharp} = \bar{A}_j.$$

Proof of (1\*). In view of an obvious induction, it suffices to show that, given any integer  $j \in \{0, \dots, l-2\}$  with  $\widehat{A}_{j+1} \in k$  and  $v_j > 0$ , we have  $\{\widehat{A}_j, \widehat{A}_{j+2}\} \subset k$  and  $v_j = v_{j+1}$  with  $\bar{A}_j = \theta \bar{A}_{j+1} + \theta'$ . Now for any  $j \in \{0, \dots, l-2\}$  with  $\widehat{A}_{j+1} \in k$  and  $v_j > 0$ , by (2<sub>j</sub>) and MP we see that

$$\{\widehat{A}_j, \widehat{A}_{j+2}\} \subset k \text{ and } \text{gap}_{(T,S)} \check{z}_{j+2}(T) = v_j \text{ with } \text{coef}_{(T,S)} \check{z}_{j+2}(T) = \theta \bar{A}_j$$

and by taking  $(z, y) = (z_{j+1}, z_{j+2})$  in (3.7)(I) we see that

$$\text{gap}_{(T,S)} \check{z}_{j+2}(T) = v_{j+1} \text{ with } \text{coef}_{(T,S)} \check{z}_{j+2}(T) = \theta \bar{A}_{j+1} + \theta'$$

and by combining the above two displays we get  $\{\widehat{A}_j, \widehat{A}_{j+2}\} \subset k$  and  $v_j = v_{j+1}$  with  $\bar{A}_j = \theta \bar{A}_{j+1} + \theta'$ .

Proof of (2\*). In view of (2), this follows from (3.1)(6) to (3.1)(8).

Proof of (3\*). In view of (1)–(7), this follows from (3.1)(6) to (3.1)(8) together with (3.7)(I).

Proof of (4\*). Assuming  $\widehat{A}_l \in k$  and  $v_{l-1} + m_1^{(l-1)} > m_2^{(l-1)}$ , by (2\*) we have

$$\widehat{A}_l \in k \text{ with } z_{l+1} \neq 0 \text{ and } m_2^{(l-1)} = p_l e_l + e_{l+1} \neq \infty. \quad (0^\#)$$

To prove that

$$\widehat{A}_{l+1} \in k \text{ and } v_l + m_2^{(l-1)} = v_{l-1} + m_1^{(l-1)} \text{ with } \bar{A}_l = \theta \bar{A}_{l-1} \quad (\#)$$

we proceed thus. In view of (3.1)(8), by (0<sup>#</sup>) we have

$$\widehat{A}_l \in k \text{ with } z_{l+1} \neq 0 \text{ and } z_l^* = z_l^{p_l} z_{l+1}. \quad (1^\#)$$

Let  $\check{z}_l^* = \check{z}_l^*(T)$  be the  $(V, K, T)$ -expansion of  $(z_l^*, z_l, \widehat{A}_l)$  and let

$$\text{gap}_{(T,S)} \check{z}_l^*(T) = v_l^* \text{ with } \text{coef}_{(T,S)} \check{z}_l^*(T) = \bar{A}_l^*. \quad (2^\#)$$

Clearly

$$\text{ord}_V z_{l-1} = m_1^{(l-1)} \quad (3^\#)$$

and by (0<sup>#</sup>) and (1<sup>#</sup>) we see that

$$\text{ord}_V z_l^* = m_2^{(l-1)}. \quad (4^\#)$$

In view of (3.1)(8), by (2<sup>#</sup>)–(4<sup>#</sup>) and AP with  $z_j = z_{l-1}$  and  $z^{\#\#} = z_l^*$  it follows that

$$v_l^* + m_2^{(l-1)} = v_{l-1} + m_1^{(l-1)} \quad \text{with } \bar{A}_l^* = \bar{A}_{l-1}. \quad (5^\#)$$

Let  $\check{z}_{l+1}(T)$  be the  $(V, K, T)$ -expansion of  $(z_{l+1}, z_l, \widehat{A}_l)$  and let

$$\text{gap}_{(T,S)}\check{z}_{l+1}(T) = \check{v}_{l+1} \quad \text{with } \text{coef}_{(T,S)}\check{z}_{l+1}(T) = \check{A}_{l+1}. \quad (6^\#)$$

By (5<sup>#</sup>) we see that  $v_l^* > 0$ , and hence by (1<sup>#</sup>), (6<sup>#</sup>) and MP with  $(z^{\#}, z_j, z_{j+1}) = (z_l^*, z_{l+1}, z_l)$  we get

$$\widehat{A}_{l+1} \in k \quad \text{and } \check{v}_{l+1} = v_l^* \quad \text{with } \check{A}_{l+1} = \Theta \bar{A}_l^*. \quad (7^\#)$$

In view of (6<sup>#</sup>) and (7<sup>#</sup>), by taking  $(y, z) = (z_l, z_{l+1})$  in (3.7)(I) we see that

$$\check{v}_{l+1} = v_l \quad \text{with } \check{A}_{l+1} = \Theta \bar{A}_l + \Theta'. \quad (8^\#)$$

Combining (5<sup>#</sup>), (7<sup>#</sup>) and (8<sup>#</sup>) we get (‡).

Proof of (5\*). Assuming  $\widehat{A}_l \in k$  and  $v_{l-1} + m_1^{(l-1)} < m_2^{(l-1)}$ , in view of (9)–(12) and (3.1)(6)–(3.1)(8), by (2\*) we see that

$$z_l^b \neq 0 \quad \text{with } e_l^b = v_{l-1} + e_{l-1} \equiv 0 \pmod{e_l}$$

and hence, in view of (9)–(12) and (3.1)(6)–(3.1)(8), by AP with  $z_j = z_{l-1}$  and  $z^{\#\#} = z_l^b$  we see that

$$v_l^b = 0 \quad \text{with } \bar{A}_l^b = \bar{A}_{l-1}.$$

Proof of (6\*). Assuming  $\widehat{A}_l \in k$  and  $v_{l-1} \neq \infty$  with  $v_{l-1} + m_1^{(l-1)} = m_2^{(l-1)}$ , in view of (9)–(12) and (3.1)(6)–(3.1)(8), by (2\*) we see that

$$z_{l+1} \neq 0 \neq z_l^b \quad \text{with } e_l^b = v_{l-1} + e_{l-1} \not\equiv 0 \pmod{e_l}$$

and hence, in view of (9)–(12) and (3.2)(6)–(3.2)(8), by AP with  $z_j = z_{l-1}$  and  $z^{\#\#} = z_l^b$  we see that

$$v_l^b = 0 \quad \text{with } \bar{A}_l^b = \bar{A}_{l-1}.$$

Proof of (7\*). Assuming  $\widehat{A}_1 \in k$  and  $v_0 = 0$ , we clearly have

$$\text{inco}_T z_0 = \Theta \bar{A}_0 \in K \setminus k. \quad (0')$$

To prove the equation

$$\frac{\text{inco}_T z_l^{\text{bb}}}{\text{inco}_T z_l} = \ominus \bar{A}_0^E \quad \text{with } E = (-1)^{l+1}(e_1/e_l) \in \mathbb{Z}^\times \quad (')$$

we define the euclidean postextension of the integer pair  $(e_0, e_1)$  with  $e_1 \neq 0$  to be the sequence pair  $((e'_j)_{0 \leq j \leq l+1}, (p'_j)_{0 \leq j \leq l})$  obtained by putting  $e'_j = e_j$  or 0 accordings as  $0 \leq j \leq l$  or  $j = l + 1$ , and  $p'_j = p_j$  or  $e_{l-1}/e_l$  accordings as  $0 \leq j \leq l - 1$  or  $j = l$ . Note that now

$$e'_{j-1} = p'_j e'_j + e'_{j+1} \quad \text{for } 1 \leq j \leq l. \quad (1')$$

Given any integers  $e''_0, e''_1$ , let us define integers  $e''_2, \dots, e''_{l+1}$  by requiring that

$$e''_{j-1} = p'_j e''_j + e''_{j+1} \quad \text{for } 1 \leq j \leq l. \quad (2')$$

Let  $M_j = \begin{pmatrix} e'_{j-1} & e'_j \\ e''_{j-1} & e''_j \end{pmatrix}$  for  $1 \leq j \leq l + 1$ , and  $N_j = \begin{pmatrix} 0 & 1 \\ 1 & -p'_j \end{pmatrix}$  for  $1 \leq j \leq l$ .

Then

$$M_j N_j = M_{j+1} \quad \text{with } \det(N_j) = -1 \quad \text{for } 1 \leq j \leq l \quad (3')$$

and hence

$$\det(M_{l+1}) = (-1)^l \det(M_1). \quad (4')$$

Clearly  $\det(M_{l+1}) = e''_{l+1} e'_l$  and if  $(e''_0, e''_1) = (1, 0)$  then  $\det(M_1) = -e'_1$ . Therefore

$$\text{if } (e''_0, e''_1) = (1, 0) \quad \text{then } e''_{l+1} = (-1)^{l+1}(e_1/e_l). \quad (5')$$

Let the sequence  $(z'_j)_{0 \leq j \leq l+1}$  be defined by putting  $z'_j = z_j$  or  $z_{l-1}/z_l^{p'_j}$  according as  $0 \leq j \leq l$  or  $j = l + 1$ . Then

$$z'_{j-1} = (z'_j)^{p'_j} z'_{j+1} \quad \text{for } 1 \leq j \leq l. \quad (6')$$

Assuming  $A \in K^\times$  to be such that  $\text{inco}_T z'_j = \ominus A^{e''_j}$  for  $0 \leq j \leq 1$ , by (1'), (2'), and (6') we see that

$$\text{inco}_T z'_j = \ominus A^{e''_j}$$

for  $0 \leq j \leq l + 1$ ; consequently by (5') we conclude that

$$\begin{cases} \text{if } (e''_0, e''_1) = (1, 0) \text{ and } A \in K^\times \text{ is such that} \\ \text{inco}_T z'_j = \ominus A^{e''_j} \text{ for } 0 \leq j \leq 1, \\ \text{then } \text{inco}_T z'_{l+1} = \ominus A^E \text{ with } E = (-1)^{l+1}(e_1/e_l) \in \mathbb{Z}^\times. \end{cases} \quad (7')$$

By (13) and (14) we have

$$\text{inco}_T z'_{l+1} = \frac{\text{inco}_T z'_l{}^{\text{bb}}}{\text{inco}_T z_l}$$

and hence by taking  $A = \bar{A}_0$  in (7') we get (').

Remark on valuation preexpansions (3.9). For further merging of Remarks (3.1), (3.2), and (3.7), let  $V$  be a DVR with

$$V \subset \widehat{V} = \text{the completion of } V \quad \text{and} \quad \text{QF}(V) = L \subset \widehat{L} = \text{QF}(\widehat{V}).$$

Let  $T$  be a uniformizing parameter of  $\widehat{V}$ . Assume that  $\text{ch}(L) = \text{ch}(H(V)) = 0$  and let  $K$  be a coefficient field of  $\widehat{V}$ . Note that then  $\widehat{V} = K((T))$ . Assume that  $H(V)$ , and hence  $K$ , is root-closed.

Given any pair of elements  $(z_0, z_1)$  in  $L^\times$  with  $\text{ord}_V z_1 \neq 0$ , by (3.1) and (3.8) there exists a system

$$(z_{ij}, e_{ij}, p_{ij}, A_{il(i)}^*(v), e_{il(i)}^*, z_{il(i)}^*, A_{ij}(v), \widehat{A}_{ij}, \tilde{A}_{ij}(v), m^{(ij)}, \tilde{z}_{ij})_{v \in \mathbb{Z}, 0 \leq j \leq l(i)+1, 0 \leq i \leq \kappa}$$

such that

$$(z_{ij}, e_{ij}, p_{ij}, A_{il(i)}^*(v), e_{il(i)}^*, z_{il(i)}^*)_{v \in \mathbb{Z}, 0 \leq j \leq l(i)+1, 0 \leq i \leq \kappa}$$

is the  $(V, K)$ -preexpansion of  $(z_0, z_1)$  and

$$(z_{ij}, e_{ij}, p_{ij}, A_{il(i)}^*(v), e_{il(i)}^*, z_{il(i)}^*, A_{ij}(v), \widehat{A}_{ij}, \tilde{A}_{ij}(v), m^{(ij)}, \tilde{z}_{ij})_{v \in \mathbb{Z}, 0 \leq j \leq l(i)+1}$$

is a mixed  $(V, K, T)$ -protoexpansion of  $(z_{i0}, z_{i1})$  for  $0 \leq i \leq \kappa$ ; in analogy with a mixed protoexpansion, we call such a system a *mixed  $(V, K, T)$ -preexpansion* of  $(z_0, z_1)$ .

Let us record that, for  $0 \leq i \leq \kappa \in \mathbb{N}$ , by (3.1)(1)–(3.1)(5) we now have a pair of sequences

$$((e_{ij})_{0 \leq j \leq l(i)}, (p_{ij})_{0 \leq j < l(i)})$$

of integers  $e_{ij} \in \mathbb{Z}$  and  $p_{ij} \in \mathbb{Z}$  with  $l(i) \in \mathbb{N}_+$  such that:

- (1)  $p_{i0} = 0 \neq e_{il(i)}$ ,
- (2)  $e_{i,j-1} = p_{ij}e_{ij} + e_{i,j+1}$  with  $p_{ij} \neq 0 < e_{i,j+1} < |e_{ij}|$  for  $1 \leq j \leq l(i) - 1$ ,
- (3)  $|e_{ij}| > |e_{il(i)}| = \text{GCD}(e_{i0}, e_{i1}) = \text{GCD}(e_{i0}, \dots, e_{il(i)})$  for  $1 \leq j \leq l(i) - 1$ ,
- (4)  $l(i) = 1 \Leftrightarrow e_{i0} \equiv 0 \pmod{(e_{i1})}$ ,
- (5)  $z_{i,j-1} = z_{ij}^{p_{ij}} z_{i,j+1}$  for  $1 \leq j \leq l(i) - 1$ .

Let us also record that (3.1)(6)–(3.1)(12) and (3.8)(1)–(3.8)(7) hold with obvious modifications. Note that (3.1)(11) is used in proving (II) below.



Now rewriting (3.8)(I) and (3.8)(II) in terms of the difference sequence  $q(m)$  defined in (3.2) we respectively see that

$$\left\{ \begin{array}{l} \text{for } 0 \leq j \leq l(i) - 1 \text{ and } 0 \leq i \leq \kappa \\ \text{we have } 0 \neq h(m^{(i0)}) = h(m^{(ij)}) \neq 0 \\ \text{and } e_{i1} = m_0^{(i0)} \text{ and } e_{i0} = m_1^{(i0)} \\ \text{with } e_{i,j+1} = m_0^{(ij)} \text{ and } e_{ij} = m_1^{(ij)} \\ \text{and } q_\mu(m^{(i0)}) = q_\mu(m^{(ij)}) \text{ for } 2 \leq \mu \leq h(m^{(i0)}) + 1 \\ \text{and } d_1(m^{(i0)}) = |e_{i1}| \text{ with } d_1(m^{(ij)}) = |e_{i,j+1}| \\ \text{and } d_2(m^{(i0)}) = d_2(m^{(ij)}) = \text{GCD}(e_{i0}, e_{i1}) = \text{GCD}(e_{ij}, e_{i,j+1}) \\ \text{and } d_\mu(m^{(i0)}) = d_\mu(m^{(ij)}) \text{ for } 2 \leq \mu \leq h(m^{(i0)}) + 2 \end{array} \right. \quad \text{(I)}$$

and

$$\left\{ \begin{array}{l} \text{for } 0 \leq i < \kappa \\ \text{we have } h(m^{(i+1,0)}) = h(m^{(i,l(i)-1)}) - 1 \\ \text{and } e_{i+1,1} = m_0^{(i+1,0)} = m_0^{(i,l(i)-1)} = e_{il(i)} \\ \text{and } e_{i+1,0} = m_1^{(i+1,0)} = m_2^{(i,l(i)-1)} = p_{il(i)}e_{il(i)} + e_{i,l(i)+1} \\ \text{and } e_{i,l(i)-1} = m_1^{(i,l(i)-1)} \\ \text{and } q_\mu(m^{(i+1,0)}) = q_{\mu+1}(m^{(i,l(i)-1)}) \text{ for } 2 \leq \mu \leq h(m^{(i+1,0)}) + 1 \\ \text{and } d_1(m^{(i+1,0)}) = |e_{i+1,1}| \\ \text{and } d_1(m^{(i,l(i)-1)}) = |e_{il(i)}| = d_2(m^{(i,l(i)-1)}) = \text{GCD}(e_{il(i)}, e_{i,l(i)-1}) \\ \text{and } d_2(m^{(i+1,0)}) = d_3(m^{(i,l(i)-1)}) = \text{GCD}(e_{i+1,0}, e_{i+1,1}) \\ \text{and } d_\mu(m^{(i+1,0)}) = d_{\mu+1}(m^{(i,l(i)-1)}) \text{ for } 2 \leq \mu \leq h(m^{(i+1,0)}) + 2. \end{array} \right. \quad \text{(II)}$$

Combining (I) and (II) we get the concise **THIRD INVERSION THEOREM** which shows the power of the difference sequence and which says that:

$$\left\{ \begin{array}{l} \text{for } 0 \leq j \leq l(i) - 1 \text{ and } 0 \leq i \leq \kappa \\ \text{we have } h(m^{(ij)}) = h(m^{(00)}) - i \\ \text{and } q_0(m^{(ij)}) = e_{i,j+1} = \text{ord}_V z_{i,j+1} \text{ with } z_{i,j+1} \in L^\times \\ \text{and } q_1(m^{(ij)}) = e_{ij} = \text{ord}_V z_{ij} \text{ with } z_{ij} \in L^\times \\ \text{and } q_\mu(m^{(ij)}) = q_{\mu+i}(m^{(00)}) \text{ for } 2 \leq \mu \leq h(m^{(ij)}) + 1 \\ \text{and } d_\mu(m^{(ij)}) = d_{\mu+i}(m^{(00)}) \text{ for } 2 \leq \mu \leq h(m^{(ij)}) + 2. \end{array} \right. \quad \text{(III)}$$

Remark on root-closed fields (3.10). The concepts of root-closed fields and special subfields, as well as Newton's Binomial Theorem for fractional exponents, all lead to the idea of root extraction, which in turn inspires the following generalization (I)

of a 1936 result of F. K. Schmidt, where we use the terminology according to which: By a *quasiroot-closed pair* we mean a pair  $(R, I)$  consisting of a domain  $R$  and a nonzero ideal  $I$  in it such that

$$\left\{ \begin{array}{l} \text{for every } a \in I \text{ we have } b_n^n = (1 + a) \text{ for some } b_n \in R \\ \text{for infinitely many } n \in \mathbb{N}_+. \end{array} \right.$$

By a *quasiroot-closed domain* we mean a domain  $R$  such that  $(R, I)$  is a quasiroot-closed pair for some nonzero ideal  $I$  in  $R$ . By  $\mathcal{N}(R)$  we denote the *normalization* of a domain  $R$ , i.e., the integral closure of  $R$  in  $\text{QF}(R)$ .

(I) Let  $(R, I)$  be any quasiroot-closed pair.

- (1) Then for every DVR  $V$  with  $\text{QF}(R) = \text{a subfield of } \text{QF}(V)$  we have  $R \subset V$ .
- (2) More generally, for every noetherian domain  $W$  with  $\text{QF}(R) = \text{a subfield of } \text{QF}(W)$  we have  $R \subset \mathcal{N}(W)$ .
- (3) Moreover, if  $R$  is noetherian and  $W$  is any quasiroot-closed noetherian domain with  $\text{QF}(R) = \text{QF}(W)$  then  $\mathcal{N}(R) = \mathcal{N}(W)$ .
- (4) Finally, if  $R$  is a DVR then for every normal noetherian domain  $W$  with  $\text{QF}(R) = \text{QF}(W) \neq W$  we have  $R = W$ .

*Proof of (1).* If  $R \not\subset V$  then for some  $x \in R$  we will have  $\text{ord}_V(x) = -q$  with  $q \in \mathbb{N}_+$ . Since  $I \neq \{0\}$ , we can take  $0 \neq y \in I$ . Upon letting  $a = yx^m$  for large  $m \in \mathbb{N}_+$  we get  $a \in I$  and  $\text{ord}_V a = -p$  with  $p \in \mathbb{N}_+$ . Clearly  $\text{ord}_V(1 + a) = -p$ . Now taking  $n > p$ , the equation  $b_n^n = (1 + a)$  implies  $\text{ord}_V b_n = p/n \notin \mathbb{Z}$  which is a contradiction. Therefore,  $R \subset V$ .

*Proof of (2).* Follows from (1) by noting that by Theorem (4.10) on page 118 of Nagata [28]  $\mathcal{N}(W)$  is the intersection of all DVRs  $V$  with  $\text{QF}(W) = \text{QF}(V)$  and  $W \subset V$ .

*Proof of (3).* By (2) we get  $\mathcal{N}(R) \subset \mathcal{N}(W)$  with  $\mathcal{N}(W) \subset \mathcal{N}(R)$  and hence  $\mathcal{N}(R) = \mathcal{N}(W)$ .

*Proof of (4).* Follows from (2) by noting that there is no subring strictly between a DVR and its quotient field.

Recall that a quasilocal domain  $R$  is *henselian* means it satisfies the following condition: If  $f(Y)$  is any monic polynomial of degree  $n > 0$  with coefficients in  $R$  such that, letting  $\tilde{f}(Y)$  denote the polynomial obtained by applying  $H_R$  to the coefficients of  $f(Y)$ , we have  $\tilde{f}(Y) = g^*(Y)h^*(Y)$  where  $g^*(Y)$  and  $h^*(Y)$  are monic coprime polynomials in  $H(R)[Y]$ , then there exists unique monic  $\underline{g}(Y)$  in  $h(Y)$  in  $R[Y]$  such that  $f(Y) = \underline{g}(Y)\underline{h}(Y)$  with  $\underline{g}(Y) = g^*(Y)$  and  $\underline{h}(Y) = h^*(Y)$ . In order to apply (I) to this case, by taking

$$f(Y) = Y^n - (1 + a) \text{ and } n \not\equiv 0 \pmod{\text{ch}(H(R))}$$

we see that:

(II) If  $R$  is a henselian quasilocal domain which is not a field then  $(R, M(R))$  is a quasisroot-closed pair.

By (I) and (II) we get the following:

(III) If  $R$  and  $S$  are henselian local domains with  $R \neq \text{QF}(R) = \text{QF}(S) \neq S$  then  $R = S$ .

In this connection, referring to [12], we note that:

(IV) Every complete local domain is henselian. The  $r$ -variable power series ring  $K[[X_1, \dots, X_r]]$  over a field  $K$  with  $r \in \mathbb{N}_+$  is an  $r$ -dimensional complete local domain which is normal and unequal to its quotient field  $K((X_1, \dots, X_r))$ .

By (III) and (IV) we see that:

(V) If  $r$  and  $s$  are positive integers and  $K$  and  $L$  are fields for which we have  $K((X_1, \dots, X_r)) = L((Y_1, \dots, Y_s))$ , then we have  $K[[X_1, \dots, X_r]] = L[[Y_1, \dots, Y_s]]$  and  $r = s$ .

## 4 Newtonian Expansion

In Remarks (3.1) and (3.9), we organized the valuation data in  $\kappa + 1$  blocks of sizes  $l(0), l(1), \dots, l(\kappa)$ . Now we shall reorganize it in a single sequence of length  $l(0) + l(1) + \dots + l(\kappa)$ . To be more precise, the blocks were of sizes  $l(0) + 2, \dots, l(\kappa) + 2$  where the last two members of a block essentially coincided with the second and third members of the next block. Likewise the reorganized single sequence will more precisely be of length  $l(0) + \dots + l(\kappa) - \kappa + 1$ . In Sect. 5, we shall give a brief review of quadratic transformations and discuss invariance properties of newtonian characteristic sequences. In PART II, we shall revisit Newton's polygonal method and thereby deduce certain integral dependence properties of the coefficients of fractional power series expansions.

Let  $V$  be a DVR with

$$V \subset \widehat{V} = \text{the completion of } V \quad \text{and} \quad \text{QF}(V) = L \subset \widehat{L} = \text{QF}(\widehat{V})$$

and let  $K$  be a coefficient set of  $V$ . In (3.1) we have defined what we mean by a  $(V, K)$ -presequence

$$(z_{ij}, e_{ij}, p_{ij}, A_{il(i)}^*(v), e_{il(i)}^*, z_{il(i)}^*)_{v \in \mathbb{Z}, 0 \leq j \leq l(i)+1, 0 \leq i \leq \kappa}. \quad (\bullet)$$

Note that then

$$\left\{ \begin{array}{l} \kappa \in \mathbb{N} \text{ with } l(\kappa) \in \mathbb{N}_+, \\ 2 \leq l(i) \in \mathbb{N}_+ \text{ for } 0 \leq i < \kappa, \\ z_{ij} \in L \text{ with } \text{ord}_V z_{ij} = e_{ij}, \\ z_{il(i)}^* \in L \text{ with } \text{ord}_V z_{il(i)}^* = e_{il(i)}^*, \\ p_{ij} \in \mathbb{Z} \cup \{\infty\} \text{ with } A_{il(i)}^*(v) \in K, \end{array} \right. \quad (1^\dagger)$$

where the quantities  $z_{ij}, z_{il(i)}^*, p_{ij}, A_{il(i)}^*(v)$  satisfy the conditions described in (3.1). In particular we have  $\text{ord}_V z_{00} = e_{00} \in \mathbb{Z}$  with  $\text{ord}_V z_{01} = e_{01} \in \mathbb{Z}^\times$ . Moreover, having noted that the pair  $(z_{00}, z_{01})$  uniquely determines  $(\bullet)$ , we have called  $(\bullet)$  the  $(V, K)$ -preexpansion of  $(z_{00}, z_{01})$ .

Now we define a  $(V, K)$ -sequence to be a sequence

$$(z_j, e_j, p_j, B_j(v), \epsilon(i), t(j))_{v \in \mathbb{Z}, 0 \leq j \leq \lambda, 0 \leq i \leq \kappa} \quad (\bullet\bullet)$$

where

$$\left\{ \begin{array}{l} \kappa \in \mathbb{N} \text{ with } \lambda = \epsilon(\kappa) \in \mathbb{N}_+, \\ \epsilon(i) \in \mathbb{N}_+ \text{ for } 0 \leq i \leq \kappa, \\ \epsilon(i) < \epsilon(i+1) \text{ for } 0 \leq i < \kappa, \end{array} \right. \quad (1^\ddagger)$$

and

$$t(j) = \begin{cases} \max\{i : 1 \leq i \leq \kappa + 1 \text{ with } \epsilon(i-1) \leq j\} & \text{if } j \geq \epsilon(0) \\ 0 & \text{if } j < \epsilon(0) \end{cases} \quad (2^\ddagger)$$

and

$$\left\{ \begin{array}{l} z_j \in L^\times \text{ with } \text{ord}_V z_j = e_j \in \mathbb{Z} \text{ for } 0 \leq j \leq \lambda, \\ e_1 \neq 0 < e_{j+1} < |e_j| \text{ for } 1 \leq j < \lambda, \\ p_j \in \mathbb{Z} \text{ for } 0 \leq j < \lambda \text{ with } p_\lambda = \infty, \\ p_0 = 0 \neq p_j \text{ for } 2 \leq j < \lambda \text{ with } p_j > 0 \text{ for } 3 \leq j < \lambda, \\ B_j(v) \in K \text{ for } 0 \leq j \leq \lambda \text{ and } v \in \mathbb{Z}, \\ B_0(v) = 0 \text{ for all } v \in \mathbb{Z}, \end{array} \right. \quad (3^\ddagger)$$

with

$$z_{j-1} - \sum_{(e_{j-1}/|e_j|) \leq v < \infty} B_j(v) z_j^{v(|e_j|/e_j)} = \begin{cases} 0 \text{ in } \widehat{L} & \text{if } j = \lambda \\ z_j^{p_j} z_{j+1} & \text{if } 1 \leq j < \lambda \end{cases} \quad (4^\ddagger)$$

are such that

$$\left\{ \begin{array}{l} \text{if } j \in \{1, \dots, \lambda\} \setminus \{\epsilon(0), \dots, \epsilon(\kappa)\} \\ \text{then } B_j(v) = 0 \text{ for all } v \in \mathbb{Z} \\ \text{and } e_{j-1}/e_j \notin \mathbb{Z} \text{ with } e_{j-1} = p_j e_j + e_{j+1}, \\ \text{and } z_{j-1} = z_j^{p_j} z_{j+1}, \end{array} \right. \quad (5^\ddagger)$$

and

$$\left\{ \begin{array}{l} \text{if } j \in \{\epsilon(0), \dots, \epsilon(\kappa)\} \\ \text{then } e_{j-1}/e_j \in \mathbb{Z} \text{ with } e_{j-1}/e_j \leq p_j(e_j/|e_j|), \\ \text{and } B_j(v) \begin{cases} = 0 & \text{if } v < (e_{j-1}/|e_j|) \\ \neq 0 & \text{if } v = (e_{j-1}/|e_j|) \\ = 0 & \text{if } v > p_j(e_j/|e_j|) \text{ and } j \neq \lambda, \end{cases} \end{array} \right. \quad (6^\ddagger)$$

and we make the *convention* that

$$\epsilon(-1) = 0 \quad \text{and} \quad \epsilon(\kappa + 1) = \infty. \quad (7^\ddagger)$$

Any pair of elements  $(z_0, z_1)$  in  $L^\times$  with  $\text{ord}_V z_1 \neq 0$  can clearly be embedded in a unique  $(V, K)$ -sequence  $(\bullet\bullet)$  which we call the  $(V, K)$ -*expansion* of  $(z_0, z_1)$ .

Note that if  $(z_{00}, z_{01}) = (z_0, z_1)$  then  $(\bullet)$  and  $(\bullet\bullet)$  determine each other by the relations

$$\left\{ \begin{array}{l} \lambda = l(0) + \dots + l(\kappa) - \kappa, \\ \epsilon(i) = l(0) + \dots + l(i) - i \text{ for } 0 \leq i \leq \kappa, \\ z_j = z_{0j} \text{ for } 0 \leq j \leq \epsilon(0), \\ z_j = z_{i,j+1-\epsilon(i-1)} \text{ for } 1 \leq i \leq \kappa \text{ and } \epsilon(i-1) \leq j \leq \epsilon(i), \\ p_j = p_{0j} \text{ for } 0 \leq j \leq \epsilon(0), \\ p_j = p_{i,j+1-\epsilon(i-1)} \text{ for } 1 \leq i \leq \kappa \text{ and } \epsilon(i-1) \leq j \leq \epsilon(i), \end{array} \right. \quad (2^\ddagger)$$

and

$$\left\{ \begin{array}{l} B_j(v) = A_{i(i)}^*(v) \text{ for } 1 \leq i \leq \kappa \text{ and } j = \epsilon(i), \\ z_j^{p_j} z_{j+1} = z_{i(i)}^* \text{ for } 1 \leq i < \kappa \text{ and } j = \epsilon(i). \end{array} \right. \quad (3^\ddagger)$$

Descriptive Note  $(8^\ddagger)$ . In a more descriptive manner, the  $i$ -th row of  $(\bullet)$  as a “matrix” looks like

$$z_{i0}, z_{i1}, \dots, z_{i,l(i)+1}$$

and a slight trimming converts it into the  $i$ -th =  $\iota$ -th subsequence of  $(\bullet\bullet)$  which looks like

$$z_{\epsilon(i-1)}, z_{\epsilon(i-1)+1}, \dots, z_{\epsilon(i)-1}$$

with the convention (7 $\ddagger$ ) that  $\epsilon(-1) = 0$ ; namely, for  $i = 0$ , delete the last two terms of the  $i$ -th row whereas, for  $i > 0$ , delete the first and the last two terms of the  $i$ -th row. Moreover, at the  $\epsilon(i)$ -th spot of  $(\bullet\bullet)$  with  $0 \leq i < \kappa$  we put the following expansion with nonempty support:

$$z_{\epsilon(i)-1} = \left( \sum_{\nu} B_{\epsilon(i)}(\nu) z_{\epsilon(i)}^{\nu} \right) + z_{\epsilon(i)}^{p_{\epsilon(i)}} z_{\epsilon(i)+1}.$$

In  $(\bullet\bullet)$ , the basic sequence is  $(z_j, e_j, p_j, B_j(\nu))_{\nu \in \mathbb{Z}, 0 \leq j \leq \lambda}$ . The remaining two quantities  $\epsilon(i)$  and  $\iota(j)$  are determined by the basic sequence thus. The  $\epsilon(i)$  are those values of  $j$  at which the support of the function  $\nu \mapsto B_j(\nu)$  is nonempty; we label the  $\epsilon(i)$  so that they increase with  $i$ . The  $\iota(j)$  are the counters to locate  $\epsilon(i)$ . In other words, if  $j = 0, 1, 2, \dots, \lambda$  are the markers of the train stations as we march along the basic sequence, then  $\epsilon(i)$  is the label of a crowded station (say, a junction), and for  $0 \leq j \leq \lambda$  we have  $\iota(j) = i \Leftrightarrow \epsilon(i-1) \leq j < \epsilon(i)$ , i.e., we have

$$\epsilon(\iota(j) - 1) \leq j < \epsilon(\iota(j))$$

with the convention (7 $\ddagger$ ) that  $\epsilon(-1) = 0$  and  $\epsilon(\kappa + 1) = \infty$ . With this convention we can write

$$\epsilon(-1) = 0 < \epsilon(0) < \epsilon(1) < \dots < \epsilon(\kappa) = \lambda < \infty = \epsilon(\kappa + 1).$$

**Definition.** Let  $T$  be a uniformizing parameter of  $\widehat{V}$ . Assume that  $\text{ch}(L) = \text{ch}(H(V)) = 0$  and  $K$  is a coefficient field of  $\widehat{V}$ . Note that then  $\widehat{V} = K((T))$ . Assume that  $H(V)$ , and hence  $K$ , is root-closed. Given any pair  $(z_0, z_1)$  in  $L^\times$  with  $\text{ord}_V z_1 \neq 0$ , in view of (3.7) and what we have said above, there exists a system

$$(z_j, e_j, p_j, B_j(\nu), \epsilon(i), \iota(j), A_j(\nu), \widehat{A}_j, \widetilde{A}_j(\nu), m^{(j)}, \widetilde{z}_j)_{\nu \in \mathbb{Z}, 0 \leq j \leq \lambda, 0 \leq i \leq \kappa} \tag{\bullet\bullet\bullet}$$

such that  $(\bullet\bullet)$  is the  $(V, K)$ -expansion of  $(z_0, z_1)$  and

$$\text{for } 0 \leq j \leq \lambda$$

we have

$$\widehat{A}_j \in K^\times \text{ with } (\widehat{A}_j)^{e_j} = \text{inco}_{T,z_j} \tag{1}$$

and

$$m^{(j)} = m(z_j, z_{j+1}, V, K) \tag{2}$$

and

$$\begin{cases} A_j(\nu) \in K \text{ for all } \nu \in \mathbb{Z} \\ \text{with } A_j(\nu) = 0 \text{ for } \nu < e_j \text{ and } A_j(e_j) \neq 0 \end{cases} \tag{3}$$

and

$$\begin{cases} \widetilde{A}_j(\nu) \in K \text{ for all } \nu \in \mathbb{Z} \\ \text{with } \widetilde{A}_j(\nu) = 0 \text{ for } \nu < e_j \text{ and } \widetilde{A}_j(e_j) \neq 0 \end{cases} \tag{4}$$

such that

$$z_j = z_j(T) = \sum_{e_j \leq v < \infty} A_j(v)T^v \tag{5}$$

is the usual expansion of  $z_j$  in  $K((T))$  and

$$\tilde{z}_j = \tilde{z}_j(T) = \sum_{e_j \leq v < \infty} \tilde{A}_j(v)T^v \tag{6}$$

is the  $(V, K, T)$ -expansion of  $(z_j, z_{j+1}, \hat{A}_{j+1})$  with the proviso that

$$\left\{ \begin{array}{l} \text{if } j = \lambda \text{ then } m^{(j)} = m(\emptyset, 1) \\ \text{and } \tilde{z}_j = \tilde{z}_j(T) = 0 = \tilde{A}_j(v) \text{ for all } v \in \mathbb{Z}; \end{array} \right. \tag{7}$$

we call such a system a *mixed  $(V, K, T)$ -expansion* of  $(z_0, z_1)$ .

Since  $(\bullet)$  and  $(\bullet\bullet)$  determine each other, referring to (3.2) for notation, (3.9)(III) may be paraphrased as the:

First invariance theorem (I). For  $0 \leq j \leq \lambda - 1$  we have

$$\left\{ \begin{array}{l} h(m^{(j)}) = h(m^{(0)}) - \iota(j) \\ \text{and } q_0(m^{(j)}) = e_{j+1} = \text{ord}_V z_{j+1} \text{ with } z_{j+1} \in L^\times \\ \text{and } q_1(m^{(j)}) = e_j = \text{ord}_V z_j \text{ with } z_j \in L^\times \\ \text{and } q_\mu(m^{(j)}) = q_{\mu+\iota(j)}(m^{(0)}) \text{ for } 2 \leq \mu \leq h(m^{(j)}) + 1 \\ \text{and } d_\mu(m^{(j)}) = d_{\mu+\iota(j)}(m^{(0)}) \text{ for } 2 \leq \mu \leq h(m^{(j)}) + 2. \end{array} \right.$$

Moreover, we have

$$h(m^{(\lambda)}) = h(m^{(0)}) - \iota(\lambda) = 0 \quad \text{with } \iota(\lambda) = \kappa + 1.$$

Preamble for next theorem. Referring to (3.5) for notation, having just dealt with case (3.5)(2<sup>#</sup>), turning to case (3.5)(3<sup>#</sup>) let

$$\left\{ \begin{array}{l} S = k((T)) \text{ where } k \text{ is a nonnull special subfield of } K, \\ \theta = \text{an unspecified member of } k^\times \\ (\theta \text{ is called Abhyankar's nonzero and may be read as } \theta), \\ \theta' = \text{an unspecified member of } k \\ (\theta' \text{ is called Abhyankar's constant and may be read as } \theta'), \\ \text{gap}_{(T,S)} \tilde{z}_j(T) = v_j \text{ with } \text{coef}_{(T,S)} \tilde{z}_j(T) = \bar{A}_j \text{ for } 0 \leq j \leq \lambda \\ \text{with the understanding that } v_\lambda = \infty \text{ and } \bar{A}_\lambda = 0 \end{array} \right. \tag{8}$$

and

$$\text{for } 1 \leq l \leq \lambda$$

let

$$z_l^\dagger = \sum_{(e_{l-1}/|e_l|) \leq \nu < (\nu_{l-1} + e_{l-1})|e_l|^{-1}} B_l(\nu) z_l^{\nu(e_{l1}/e_l)} \in K[z_l, z_l^{-1}] \quad (9)$$

and

$$z_l^b = z_{l-1} - z_l^\dagger \in L \text{ with } \text{ord}_V z_l^b = e_l^b \quad (10)$$

and let

$$z_l^b = z_l^b(T) = \sum_{\nu \in \mathbb{Z}} A_l^b(\nu) T^\nu \text{ with } A_l^b(\nu) \in K \quad (11)$$

be the usual expansion in  $K((T))$  and

$$\begin{cases} \text{if } z_l^b \neq 0 \\ \text{then let } \tilde{z}_l^b = \tilde{z}_l^b(T) \text{ be the } (V, K, T)\text{-expansion of } (z_l^b, z_l, \widehat{A}_l) \\ \text{and let } \text{gap}_{(T,S)} \tilde{z}_l^b(T) = \nu_l^b \text{ with } \text{coef}_{(T,S)} \tilde{z}_l^b(T) = \widehat{A}_l^b \end{cases} \quad (12)$$

and finally let

$$z_l^{bb} \in L \text{ with } \text{ord}_V z_l^{bb} = e_l^{bb} \quad (13)$$

and

$$z_l^{bbb} \in L \text{ with } \text{ord}_V z_l^{bbb} = e_l^{bbb} \quad (14)$$

be defined by putting

$$z_l^{bb} = \begin{cases} 0 & \text{if } e_{l-1}/e_l \notin \mathbb{Z} \\ z_{l-1}/z_l^{(e_{l-1}-e_l)/e_l} & \text{if } e_{l-1}/e_l \in \mathbb{Z} \end{cases} \quad (15)$$

and

$$z_l^{bbb} = \begin{cases} 0 & \text{if } z_l^b = 0 \\ 0 & \text{if } z_l^b \neq 0 \text{ and } e_l^b/e_l \notin \mathbb{Z} \\ z_l^b/z_l^{(e_l^b-e_l)/e_l} & \text{if } z_l^b \neq 0 \text{ and } e_l^b/e_l \in \mathbb{Z} \end{cases} \quad (16)$$

and let

$$z_l^{bb} = z_l^{bb}(T) = \sum_{\nu \in \mathbb{Z}} A_l^{bb}(\nu) T^\nu \text{ with } A_l^{bb}(\nu) \in K \quad (17)$$

and

$$z_l^{bbb} = z_l^{bbb}(T) = \sum_{\nu \in \mathbb{Z}} A_l^{bbb}(\nu) T^\nu \text{ with } A_l^{bbb}(\nu) \in K \quad (18)$$

be the usual expansion in  $K((T))$ .



With the above notation at hand, we shall now prove the:

Second invariance theorem (II). For  $0 \leq j \leq \lambda - 1$  we have the following.

- (1\*) If  $\widehat{A}_{j+1} \in k$  with  $v_j > 0$  and  $l = \epsilon(\iota(j))$  with  $v_j + m_1^{(l-1)} > m_2^{(l-1)}$  then:  $\widehat{A}_l \in k$  and  $v_{l-1} = v_j$  with  $\bar{A}_{l-1} = \Theta \bar{A}_j + \Theta'$  and we have  $l < \lambda$  with  $\widehat{A}_{l+1} \in k$  and  $v_l = v_j + m_1^{(l-1)} - m_2^{(l-1)} > 0$  with  $\bar{A}_l = \Theta \bar{A}_j + \Theta'$ .
- (2\*) If  $\widehat{A}_{j+1} \in k$  with  $v_j > 0$  and  $l = \epsilon(\iota(j))$  with  $v_j + m_1^{(l-1)} < m_2^{(l-1)}$  then:  $\widehat{A}_l \in k$  and  $v_{l-1} = v_j$  with  $\bar{A}_{l-1} = \Theta \bar{A}_j + \Theta'$  and we have  $z_l^b \neq 0$  and  $v_l^b = 0$  with  $\bar{A}_l^b = \bar{A}_{l-1}$  and  $e_l^b = v_{l-1} + e_{l-1} \equiv 0 \pmod{e_l}$ .
- (3\*) If  $\widehat{A}_{j+1} \in k$  with  $\infty \neq v_j > 0$  and  $l = \epsilon(\iota(j))$  with  $v_j + m_1^{(l-1)} = m_2^{(l-1)}$  then:  $\widehat{A}_l \in k$  and  $v_{l-1} = v_j$  with  $\bar{A}_{l-1} = \Theta \bar{A}_j + \Theta'$  and  $l - \lambda \neq 0 \neq z_l^b$  and we have  $v_l^b = 0$  with  $\bar{A}_l^b = \bar{A}_{l-1}$  and  $e_l^b = v_{l-1} + e_{l-1} \not\equiv 0 \pmod{e_l}$ .
- (3\*\*) Notation. For stating the following generalization (4\*)–(6\*) of (1\*)–(3\*) we introduce the quantities  $\mu(j)$ ,  $\mu^*(j)$ , and  $\mu(j, j')$  thus. We put

$$\mu(j) = \max\{\mu \in \{1, \dots, h(m^{(j)}) + 1\} : v_j + m_1^{(j)} \geq m_\mu^{(j)}\}$$

and we note that if  $v_j = \infty$  then  $\mu(j) = h(m^{(j)}) + 1$ , whereas if  $v_j \neq \infty$  then  $\mu(j)$  is the unique integer with  $1 \leq \mu(j) \leq h(m^{(j)})$  such that

$$m_{\mu(j)}^{(j)} \leq v_j + m_1^{(j)} < m_{\mu(j)+1}^{(j)}.$$

If  $v_j = \infty$  then we put  $\mu^*(j) = \infty$ , whereas if  $v_j \neq \infty$  then we put

$$\mu^*(j) = v_j + m_1^{(j)} - m_{\mu(j)}^{(j)}.$$

For  $j \leq j' \leq \lambda - 1$  we put

$$\mu(j, j') = \iota(j') - \iota(j) + 1$$

and we note that then  $\mu(j, j) = 1$  and hence  $\mu^*(j) = v_j + m_{\mu(j, j)}^{(j)} - m_{\mu(j)}^{(j)}$ . The proofs of (4\*)–(6\*) will be by induction on  $\mu(j, j')$  starting with the

$$\text{ground case of } \mu(j, j') = 1,$$

i.e., the case when

$$\iota(j') = \iota(j) \quad \text{and} \quad \epsilon(\iota(j) - 1) \leq j \leq j' < \epsilon(\iota(j)).$$

- (4\*) If  $\widehat{A}_{j+1} \in k$  with  $\mu^*(j) = \infty$  then for  $j \leq j' \leq \lambda - 1$  we have  $\{\widehat{A}_{j'}, \widehat{A}_{j'+1}\} \subset k$  with  $v_{j'+1} = v_{j'} = \infty$  and  $\tilde{z}_{j'}(T) \in k((T))$  with  $B_{j'+1}(v) \in k$  for all  $v \in \mathbb{Z}$ .
- (5\*) If  $\widehat{A}_{j+1} \in k$  with  $\infty \neq \mu^*(j) > 0$  then, letting  $l = \epsilon(\iota(j) + \mu(j) - 1)$ , for  $j \leq j' \leq l - 1$  we have  $\{\widehat{A}_{j'}, \widehat{A}_{j'+1}\} \subset k$  and  $1 \leq \mu(j, j') \leq \mu(j, l - 1) = \mu(j)$  with  $\mu(j, j') + \mu(j') = 1 + \mu(j)$  and  $\infty \neq \mu^*(j') = \mu^*(j) > 0$  with  $\bar{A}_{j'} = \Theta \bar{A}_j + \Theta'$ , and moreover:  $z_l^b \neq 0$  and  $v_l^b = 0$  with  $\bar{A}_l^b = \bar{A}_{l-1}$  and  $e_l^b = v_{l-1} + e_{l-1} \equiv 0 \pmod{(e_l)}$ , and finally:  $\bar{A}_j \in K \setminus k$  and

$$\frac{\text{incor}_T z_l^{\text{bbb}}}{\text{incor}_T z_l} = \Theta \bar{A}_l^b = \Theta \bar{A}_j + \Theta'.$$

- (6\*) If  $\widehat{A}_{j+1} \in k$  with  $\mu^*(j) = 0$  and  $\mu(j) \neq 1$  then, letting  $l = \epsilon(\iota(j) + \mu(j) - 2)$ , for  $j \leq j' \leq l - 1$  we have  $\{\widehat{A}_{j'}, \widehat{A}_{j'+1}\} \subset k$  and  $1 \leq \mu(j, j') \leq \mu(j, l - 1) = \mu(j) - 1$  with  $\mu(j, j') + \mu(j') = 1 + \mu(j)$  and  $\mu^*(j') = \mu^*(j) = 0$  with  $\bar{A}_{j'} = \Theta \bar{A}_j + \Theta'$ , and moreover:  $\bar{A}_l \in k$  and  $v_{l-1} = v_j$  with  $\bar{A}_{l-1} = \Theta \bar{A}_j + \Theta'$  and  $l - \lambda \neq 0 \neq z_l^b$  and  $v_l^b = 0$  with  $\bar{A}_l^b = \bar{A}_{l-1}$  and  $e_l^b = v_{l-1} + e_{l-1} \not\equiv 0 \pmod{(e_l)}$ .
- (6\*\*\*) Notation. To facilitate stating claim (7\*), we supplement the definition of the derived denominator sequence  $\widehat{n}_i(m)_{1 \leq i \leq h(m)}$  of a charseq  $m$  with  $h(m) > 0$  given in (3.2) by introducing its signed version

$$n_i^{\text{bb}}(m) = (-1)^{n_i^b(m)} \widehat{n}_i(m)$$

where the positive integer  $n_i^b(m)$  is defined thus. Let  $\left( (e_j^{(i)})_{0 \leq j \leq l^{(i)}} \right)$ ,  $\left( (p_j^{(i)})_{0 \leq j < l^{(i)}} \right)$  be the euclidean extension of  $(e_0^{(i)}, e_1^{(i)})$  where

$$(e_0^{(i)}, e_1^{(i)}) = \begin{cases} (q_i(m), d_i(m)) & \text{if } 2 \leq i \leq h(m) \\ (q_1(m), q_0(m)) & \text{if } i = 1. \end{cases}$$

Now (paying special attention to the  $j = 0$  case) we put

$$n_i^b(m) = \begin{cases} l^{(i)} + 1 & \text{if } e_1^{(i)} > 0 \\ l^{(i)} & \text{if } e_1^{(i)} \leq 0. \end{cases}$$

- (7\*) If  $\widehat{A}_{j+1} \in k$  with  $\mu^*(j) = 0$  then, letting  $l = \epsilon(\iota(j) + \mu(j) - 1)$ , we have  $\bar{A}_j \in K \setminus k$  and

$$\frac{\text{incor}_T z_l^{\text{bbb}}}{\text{incor}_T z_l} = \Theta (\bar{A}_j + \Theta')^E \quad \text{with } E = n_{\mu(j)}^{\text{bb}}(m^{(j)}) \in \mathbb{Z}^\times.$$

*Note.* In proving Theorem (II), we shall be using the following Reincarnated Version of Lemma (3.8)(III). The said Reincarnated Version says that the Original Version remains valid when for  $0 \leq j \leq \lambda - 1$ , upon letting  $l = \epsilon(\iota(j))$ , we substitute the subsequence  $(z_j, z_{j+1}, \dots, z_l)$  and its associated quantities  $(e_j, \dots, e_l), \dots$  for the sequence  $(z_0, z_1, \dots, z_l)$  together with its associated quantities considered in (3.8). Note that in the said substitution we put  $A_j^*(\nu) = B_l(\nu)$ .

Reincarnated coefficient lemma (III). For  $0 \leq j \leq \lambda - 1$ , upon letting  $l = \epsilon(\iota(j))$ , we have the following.

- (1\*) If  $\widehat{A}_{j+1} \in k$  with  $\nu_j > 0$  then for  $j \leq j' \leq l$  we have  $\widehat{A}_{j'} \in k$ , and for  $j \leq j' \leq l - 1$  we have  $\nu_{j'} = \nu_j$  with  $\bar{A}_{j'} = \theta \bar{A}_j + \theta'$ .
- (2\*)  $l < \lambda \Leftrightarrow l - \lambda \neq 0 \Leftrightarrow m_2^{(l-1)} \neq \infty \Rightarrow m_2^{(l-1)} = p_l e_l + e_{l+1}$ .
- (3\*) If  $\widehat{A}_l \in k$  and  $\nu_{l-1} = \infty$  then  $\nu_l = \infty$  and  $\widehat{A}_{l+1} \in k$  with  $\bar{z}_{l-1}(T) \in k((T))$  and  $B_l(\nu) \in k$  for all  $\nu \in \mathbb{Z}$ .
- (4\*) If  $\widehat{A}_l \in k$  and  $\nu_{l-1} + m_1^{(l-1)} > m_2^{(l-1)}$  then  $l < \lambda$  with  $\widehat{A}_{l+1} \in k$  and  $\nu_l + m_2^{(l-1)} = \nu_{l-1} + m_1^{(l-1)}$  with  $\bar{A}_l = \theta \bar{A}_{l-1} + \theta'$ .
- (5\*) If  $\widehat{A}_l \in k$  and  $\nu_{l-1} + m_1^{(l-1)} < m_2^{(l-1)}$  then  $z_l^b \neq 0$  and  $\nu_l^b = 0$  with  $\bar{A}_l^b = \bar{A}_{l-1}$  and  $e_l^b = \nu_{l-1} + e_{l-1} \equiv 0 \pmod{e_l}$ .
- (6\*) If  $\widehat{A}_l \in k$  and  $\nu_{l-1} \neq \infty$  with  $\nu_{l-1} + m_1^{(l-1)} = m_2^{(l-1)}$  then  $l - \lambda \neq 0 \neq z_l^b$  and  $\nu_l^b = 0$  with  $\bar{A}_l^b = \bar{A}_{l-1}$  and  $e_l^b = \nu_{l-1} + e_{l-1} \not\equiv 0 \pmod{e_l}$ .
- (7\*) If  $\widehat{A}_{j+1} \in k$  and  $\nu_j = 0$  then  $\text{incor}_T z_j = \theta \bar{A}_j \in K \setminus k$  and

$$\frac{\text{incor}_T z_l^{bb}}{\text{incor}_T z_l} = \theta \bar{A}_j^E \quad \text{with } E = (-1)^{l+1-j} (e_{j+1}/e_l) \in \mathbb{Z}^\times.$$

*Proof of (II)(1\*).* Now if  $\widehat{A}_{j+1} \in k$  with  $\nu_j > 0$  and  $l = \epsilon(\iota(j))$  with  $\nu_j + m_1^{(l-1)} > m_2^{(l-1)}$  then by (III)(1\*) we get  $\widehat{A}_l \in k$  with  $\nu_{l-1} = \nu_j$  and also  $\nu_{l-1} + m_1^{(l-1)} > m_2^{(l-1)}$  with  $\bar{A}_{l-1} = \theta \bar{A}_j + \theta'$  and hence by (III)(4\*) we conclude that  $l < \lambda$  with  $\widehat{A}_{l+1} \in k$  and  $\nu_l = \nu_j + m_1^{(l-1)} - m_2^{(l-1)} > 0$  with  $\bar{A}_l = \theta \bar{A}_{l-1} + \theta'$ .

*Proof of (II)(2\*).* Now if  $\widehat{A}_{j+1} \in k$  with  $\nu_j > 0$  and  $l = \epsilon(\iota(j))$  with  $\nu_j + m_1^{(l-1)} < m_2^{(l-1)}$  then by (III)(1\*) we get  $\widehat{A}_l \in k$  with  $\nu_{l-1} + m_1^{(l-1)} < m_2^{(l-1)}$  and  $\nu_{l-1} = \nu_j$  with  $\bar{A}_{l-1} = \theta \bar{A}_j + \theta'$  and hence by (III)(5\*) we conclude that  $z_l^b \neq 0$  and  $\nu_l^b = 0$  with  $\bar{A}_l^b = \bar{A}_{l-1}$  and

$$e_l^b = \nu_{l-1} + e_{l-1} \equiv 0 \pmod{e_l}.$$

*Proof of (II)(3\*).* Now if  $\widehat{A}_{j+1} \in k$  with  $\infty \neq \nu_j > 0$  and  $l = \epsilon(\iota(j))$  with  $\nu_j + m_1^{(l-1)} = m_2^{(l-1)}$  then by (III)(1\*) we get  $\widehat{A}_l \in k$  with  $\nu_{l-1} + m_1^{(l-1)} = m_2^{(l-1)}$  and  $\nu_{l-1} = \nu_j$  with  $\bar{A}_{l-1} = \theta \bar{A}_j + \theta'$  and hence by (III)(6\*) we conclude that  $l - \lambda \neq 0 \neq z_l^b$  and  $\nu_l^b = 0$  with  $\bar{A}_l^b = \bar{A}_{l-1}$  and

$$e_l^b = \nu_{l-1} + e_{l-1} \not\equiv 0 \pmod{e_l}.$$

*Proof of (II)(4\*).* Assuming  $\widehat{A}_{j+1} \in k$  with  $\mu^*(j) = \infty$ , and given any  $j'$  with  $j \leq j' \leq \lambda - 1$ , by induction on  $\mu(j, j')$  we shall show that  $\{\widehat{A}_{j'}, \widehat{A}_{j'+1}\} \subset k$  with  $v_{j'+1} = v_{j'} = \infty$  and  $\tilde{z}_{j'}(T) \in k((T))$  with  $B_{j'+1}(v) \in k$  for all  $v \in \mathbb{Z}$ . In the ground case we are done by (III)(1\*) and (III)(3\*). So let  $\mu(j, j') > 1$  and assume true for all smaller values of  $\mu(j, j')$ . Now letting  $j_1 = \epsilon(\iota(j))$  and  $j'' = j_1 - 1$  we have  $j \leq j'' < j'' + 1 = j_1 \leq j' \leq \lambda - 1$  with (i)  $\mu(j, j'') = 1$  and (ii)  $\mu(j_1, j') = \mu(j, j') - 1$ . In view of (i), by (III)(1\*) and (III)(3\*) we get (iii)  $\widehat{A}_{j_1+1} \in k$  and (iv)  $\mu^*(j_1) = \infty$ . In view of (ii) to (iv) we are done by the induction hypothesis.

Note on proofs of (II)(5\*)–(II)(7\*). In the following arguments we may tacitly use (I) together with the fact that for  $1 \leq j \leq \lambda - 1$  we have  $m_0^{(j)} = q_0(m^{(j)})$  and  $m_\mu^{(j)} = q_1(m^{(j)}) + \cdots + q_\mu(m^{(j)})$  for  $1 \leq \mu \leq h(m^{(j)}) + 1$ . This is particularly relevant for comparing  $\mu(j)$  and  $\mu(j')$  with  $j \neq j'$ . Similarly for  $\mu^*(j)$  and  $\mu^*(j')$ .

*Proof of (II)(5\*).* Assume that  $\widehat{A}_{j+1} \in k$  with  $\infty \neq \mu^*(j) > 0$ , and let us put  $l = \epsilon(\iota(j) + \mu(j) - 1)$ .

In case of  $\mu(j) = 1$  everything follows from (III)(1\*) and (III)(5\*).

In the general case, given any  $j'$  with  $j \leq j' \leq l - 1$ , by induction on  $\mu(j, j')$  we shall show that  $\{\widehat{A}_{j'}, \widehat{A}_{j'+1}\} \subset k$  and  $1 \leq \mu(j, j') \leq \mu(j, l - 1) = \mu(j)$  with  $\mu(j, j') + \mu(j') = 1 + \mu(j)$  and  $\infty \neq \mu^*(j') = \mu^*(j) > 0$  with  $\bar{A}_{j'} = \ominus \bar{A}_j + \ominus'$ . In the ground case we are done by (III)(1\*). So let  $\mu(j, j') > 1$  and assume true for all smaller values of  $\mu(j, j')$ . Now upon letting  $j_1 = \epsilon(\iota(j))$  and  $j'' = j_1 - 1$  we see that  $j \leq j'' < j'' + 1 = j_1 \leq j' \leq \lambda - 1$  with (i)  $\mu(j, j'') = 1$  and (ii)  $\mu(j_1, j') = \mu(j, j') - 1$ . Assuming  $\mu(j) > 1$ , in view of (i), by (III)(1\*) and (III)(4\*) we also conclude that (iii)  $\widehat{A}_{j_1+1} \in k$  and (iv)  $\infty \neq \mu^*(j_1) > 0$  and (v)  $\iota(j_1) + \mu(j_1) = \iota(j) + \mu(j)$ . In view of (ii) to (v) we are done by the induction hypothesis.

In view of what we have proved in the above paragraph, by (III)(5\*) we get the “moreover” and the “finally.”

*Proof of (II)(6\*).* Assume that  $\widehat{A}_{j+1} \in k$  with  $\mu^*(j) = 0$  and  $\mu(j) \neq 1$ , and let us put  $l = \epsilon(\iota(j) + \mu(j) - 2)$ .

In case of  $\mu(j) = 2$  everything follows from (III)(1\*) and (III)(6\*).

In the general case, given any  $j'$  with  $j \leq j' \leq l - 1$ , by induction on  $\mu(j, j')$  we shall show that  $\{\widehat{A}_{j'}, \widehat{A}_{j'+1}\} \subset k$  and  $1 \leq \mu(j, j') \leq \mu(j, l - 1) = \mu(j) - 1$  with  $\mu(j, j') + \mu(j') = 1 + \mu(j)$  and  $\mu^*(j') = \mu^*(j) = 0$  with  $\bar{A}_{j'} = \ominus \bar{A}_j + \ominus'$ . In the ground case we are done by (III)(1\*). So let  $\mu(j, j') > 1$  and assume true for all smaller values of  $\mu(j, j')$ . Now upon letting  $j_1 = \epsilon(\iota(j))$  and  $j'' = j_1 - 1$  we see that  $j \leq j'' < j'' + 1 = j_1 \leq j' \leq \lambda - 1$  with (i)  $\mu(j, j'') = 1$  and (ii)  $\mu(j_1, j') = \mu(j, j') - 1$ . Assuming  $\mu(j) > 2$ , in view of (i), by (III)(1\*) and (III)(4\*) we also conclude that (iii)  $\widehat{A}_{j_1+1} \in k$  and (iv)  $\mu^*(j_1) = 0$  and (v)  $\iota(j_1) + \mu(j_1) = \iota(j) + \mu(j)$ . In view of (ii) to (v) we are done by the induction hypothesis.

In view of what we have proved in the above paragraph, by (III)(6\*) we get the “moreover.”

*Proof of (II)(7\*).* This follows from (II)(6\*) and (III)(7\*). In greater detail, the case of  $\mu(j) = 1$  is done by (III)(7\*). So assume that  $\mu(j) \neq 1$  and let

$$(z'_0, z'_1) = (z_L^b, z_L) \quad \text{where} \quad L = \epsilon(\iota(j) + \mu(j) - 2) \quad (')$$

and let

$$(z'_J, e'_J, p'_J, B'_J(v), \epsilon'(i), l'(J))_{v \in \mathbb{Z}, 0 \leq J \leq \lambda', 0 \leq i \leq \kappa'} \quad (\bullet\bullet')$$

be the  $(V, K)$ -expansion of  $(z'_0, z'_1)$ . Also let

$$(z'_J, e'_J, p'_J, B'_J(v), \epsilon'(i), l'(J), A'_J(v), \widehat{A}'_J, \dots)_{v \in \mathbb{Z}, 0 \leq J \leq \lambda', 0 \leq i \leq \kappa'} \quad (\bullet\bullet\bullet')$$

be the mixed  $(V, K, T)$ -expansion of  $(z'_0, z'_1)$ , and let

$$v'_J, \bar{A}'_J, (z'_J)^{bb}, \dots$$

have the corresponding meanings. Then assuming

$$\widehat{A}_{j+1} \in k \quad \text{with} \quad \mu^*(j) = 0$$

by (II)(6\*) we see that

$$\bar{A}'_0 = \theta \bar{A}_j + \theta' \quad \text{with} \quad e'_0 \not\equiv 0 \pmod{e'_1} \quad (i)$$

and

$$\widehat{A}'_1 \in k \quad \text{with} \quad v'_0 = 0. \quad (ii)$$

In view of (i) and (ii), upon letting

$$l' = \epsilon'(l'(0)) \quad (iii)$$

and applying (III)(7\*) with  $j = 0$  to the “primed” system we see that

$$\text{incot}_T z'_0 = \theta \bar{A}'_0 \in K \setminus k \quad (iv)$$

and

$$\frac{\text{incot}_T (z'_{l'})^{bb}}{\text{incot}_T z'_{l'}} = \theta (\bar{A}'_0)^{E'} \quad \text{with} \quad E' = (-1)^{l'+1} (e'_1/e'_{l'}) \in \mathbb{Z}^\times. \quad (v)$$

Now clearly

$$z_L^b = z_L^{pL} z_{L+1}. \quad (vi)$$

In view of (vi), upon letting

$$l = \epsilon(\iota(j) + \mu(j) - 1) \quad (vii)$$

we see that

$$z'_{l'} = z_l \quad \text{and} \quad (z'_{l'})^{\text{bb}} = z_l^{\text{bb}} \quad \text{with} \quad E' = n_{\mu(j)}^{\text{bb}}(m^{(j)}).$$

By (i)–(vii) we conclude that

$$\bar{A}_j \in K \setminus k \tag{i*}$$

and

$$\frac{\text{incor}_{T'} z_l^{\text{bb}}}{\text{incor}_T z_l} = \theta (\bar{A}_j + \theta')^E \quad \text{with} \quad E = n_{\mu(j)}^{\text{bb}}(m^{(j)}) \in \mathbb{Z}^\times. \tag{ii*}$$

Note on the proof of (II)(7\*). To get a clearer picture of the above proof remember that, as explained in the Descriptive Note (8<sup>‡</sup>), the  $(V, K)$ -sequence  $(\bullet\bullet)$  is obtained by straightening the  $(V, K)$ -presequence  $(\bullet)$ , and while doing this we drop the first element of each row, except the first; the dropped element is reinstated by the concept of  $z_l^b$  where we observe that  $z_l^b = z_l^{p_l} z_{l+1}$ . Also remembering (3.1)(8) and (3.1)(11) we observe that

$$\frac{\text{incor}_{T'} z_l^{\text{bb}}}{\text{incor}_T z_l} = A_l^*(e_{l-1}/|e_l|) = \text{the first coefficient of the summation in (3.1)(8)}.$$

At any rate,  $(\bullet\bullet')$  is obtained by chopping off the initial  $0 \leq j \leq L - 1$  piece of  $(\bullet\bullet)$  and replacing the chopped off piece by  $z_L^b = z_L^{p_L} z_{L+1}$ . Finally observe that the  $j = 0$  case of (II)(7\*) requires special treatment which is taken care of in (II)(6\*\*).

## 5 Quadratic Transformations

For details referring to [2, 3, 9–11] in general, and specifically to (Q35.8) on pages 569–577 of [12], let us recall some basic facts about QDTs = Quadratic Transformations.

Recall that,  $\text{spec}(S)$  denotes the set of all prime ideals in a ring  $S$ . If  $S$  is a domain then the modelic  $\mathfrak{W}(S) =$  the set of all localizations of  $S$  at various prime ideals in  $S$ , and if  $J$  is an ideal in  $S$  then the modelic blowup

$$\mathfrak{W}(S, J) = \bigcup_{0 \neq x \in J} \mathfrak{W}(S[Jx^{-1}])$$

where  $Jx^{-1} = \{yx^{-1} : y \in J\}$ ; if  $S$  is quasilocal then the dominating modelic blowup  $\mathfrak{W}(S, J)^\Delta =$  the set of all those members of  $\mathfrak{W}(S, J)$  which dominate  $S$ .

Let  $R$  be a positive dimensional local domain. By a QDT of  $R$  we mean a member of  $\mathfrak{W}(R, M(R))^\Delta$ . For any QDT  $S$  of  $R$  we have  $0 < \dim(S) \leq \dim(R)$  with  $\dim(R) - \dim(S) = \text{restrdeg}_R S$ , and  $S/M(S)$  is a finitely generated field extension of  $R/M(R)$ . We have  $\dim(S) = 1$  for at least one and at most a finite number of QDTs  $S$  of  $R$ . If  $R$  is regular then every QDT  $S$  of  $R$  is regular, and  $\dim(S) = 1$

for exactly one  $S$  which then coincides with the valuation ring of the real discrete valuation  $\text{ord}_R$  mentioned in Sect. 2, and hence in particular it is residually pure transcendental over  $R$ . Some QDT of  $R$  coincides with  $R$  iff  $R$  is a DVR. If  $V$  is any valuation ring dominating  $R$  then  $V$  dominates exactly one QDT  $S$  of  $R$ , and we call  $S$  the QDT of  $R$  along  $V$ .

A QDT of a positive dimensional local domain  $R$  may also be called a first QDT of  $R$ ; by a second QDT of  $R$  we mean a first QDT of a first QDT of  $R, \dots$ , by a  $j$ -th QDT of  $R$  we mean a first QDT of a  $(j - 1)$ -th QDT of  $R$ . We declare  $R$  to be the only zeroth QDT of  $R$ . By a QDT sequence of  $R$  we mean a sequence  $(R_j)_{0 \leq j < \infty}$  with  $R_0 = R$  such that  $R_j$  is a first QDT of  $R_{j-1}$  for  $0 < j < \infty$ .

If  $V$  is any valuation ring dominating a positive dimensional local domain  $R$  then, for any nonnegative integer  $j$ , there is a unique  $j$ -th QDT  $R_j$  of  $R$  which is dominated by  $V$  and we call it the  $j$ -th QDT of  $R$  along  $V$ ; we call  $(R_j)_{0 \leq j < \infty}$  the QDT sequence of  $R$  along  $V$ . To get a concrete set of generators of  $M(R_j)$  for all  $j$ , we proceed thus.

**Definition (#).** Let  $V$  be the valuation ring of a valuation  $W : L \rightarrow G \cup \{\infty\}$  of a field  $L$  and let  $K$  be a coefficient set of  $V$ . Let

$$\bar{L} = \{z \in L : W(z) = 0 \text{ or } \infty\}.$$

Given any  $(z_0, \dots, z_\tau) \in L^{\tau+1} \setminus \bar{L}^{\tau+1}$  where  $\tau$  is a positive integer, we shall define its QDT sequence  $(0^\#)$  along  $(V, K)$ . The reader may prefer to first study the  $\tau = 1$  case starting in Note (III\*). Now clearly there exists a unique sequence

$$(z_{0j}, \dots, z_{\tau j}, c_{0j}, \dots, c_{\tau j}, t(j))_{0 \leq j < \infty} \tag{0^\#}$$

with  $(z_{00}, \dots, z_{\tau 0}) = (z_0, \dots, z_\tau)$  and

$$(z_{0j}, \dots, z_{\tau j}, c_{0j}, \dots, c_{\tau j}, t(j)) \in (L^{\tau+1} \setminus \bar{L}^{\tau+1}) \times K^{\tau+1} \times \{0, \dots, \tau\}$$

for  $0 \leq j < \infty$  such that for  $0 \leq j < \infty$  and  $0 \leq t \leq \tau$  we have

$$z_{t,j+1} = \begin{cases} z_{tj} \text{ with } c_{tj} = 1 & \text{if } t = t(j) \\ \bar{z}_{tj} \text{ with } c_{tj} = 0 & \text{if } t \neq t(j) \text{ and } W(\bar{z}_{tj}) \neq 0 \\ \bar{z}_{tj} - c_{tj} \in M(V) \text{ with } c_{tj} \neq 0 & \text{if } t \neq t(j) \text{ and } W(\bar{z}_{tj}) = 0 \end{cases} \tag{1^\#}$$

where

$$\bar{z}_{tj} = \begin{cases} \frac{z_{tj}}{z_{t(j)j}} & \text{if } 0 < W(z_{t(j)j}) \leq W(z_{tj}) \\ \frac{z_{tj}}{z_{t(j)j}} & \text{if } W(z_{t(j)j}) < 0 > W(z_{tj}) \\ \frac{z_{tj}}{1/z_{t(j)j}} & \text{if } W(z_{tj}) < 0 < W(z_{t(j)j}) \\ \frac{z_{tj}}{1/z_{t(j)j}} & \text{if } W(z_{t(j)j}) < 0 < |W(z_{t(j)j})| \leq W(z_{tj}) \\ z_{tj} & \text{if } 0 = W(z_{tj}) < |W(z_{t(j)j})| \end{cases} \tag{2^\#}$$

and where, upon letting

$$\begin{cases} t^*(j) = \{0 \leq t \leq \tau : 0 \neq W(z_{tj}) \neq \infty\} \\ t^{**}(j) = \max(t^*(j)) \\ t_+^*(j) = \{t \in t^*(j) : 0 < W(z_{tj}) < \infty\} \\ t_+^{**}(j) = \max\{t \in t_+^*(j) : W(z_{tj}) \leq W(z_{t'j}) \forall t' \in t_+^*(j)\} \\ t_-^*(j) = \{t \in t^*(j) : t \geq t_+^{**}(j)\} \\ t_-^{**}(j) = \max\{t \in t_-^*(j) : |W(z_{tj})| \leq |W(z_{t'j})| \forall t' \in t_-^*(j)\} \end{cases} \quad (3^\#)$$

with the understanding that if  $t_+^*(j) = \emptyset$  then  $t_+^{**}(j) = 0 = t_-^{**}(j)$ , we put

$$t(j) = \begin{cases} t_-^{**}(j) & \text{if } t_+^*(j) \neq \emptyset \\ t^{**}(j) & \text{if } t_+^*(j) = \emptyset. \end{cases} \quad (4^\#)$$

Noting that for all  $j \in \mathbb{N}$  we have  $0 \neq W(z_{t(j)j}) < \infty$ , for  $0 \leq t \leq \tau$  we put

$$\pi(t, j) = \begin{cases} 1 & \text{if } 0 < W(z_{t(j)j}) \leq W(z_{tj}) \\ 1 & \text{if } W(z_{(j)j}) < 0 > W(z_{tj}) \\ -1 & \text{if } W(z_{tj}) < 0 < W(z_{t(j)j}) \\ -1 & \text{if } W(z_{t(j)j}) < 0 < |W(z_{t(j)j})| \leq W(z_{tj}) \\ 0 & \text{if } 0 = W(z_{tj}) < |W(z_{t(j)j})| \end{cases} \quad (5^\#)$$

and we observe that  $z_{t(j)j}^{\pi(t,j)}$  is the denominator in each line of (2<sup>#</sup>).

Let us define the flipping set  $\Phi^\#$  of (0<sup>#</sup>) by putting

$$\Phi^\# = \text{the set of all } j \in \mathbb{N}_+ \text{ such that } t(j-1) \neq t(j). \quad (6^\#)$$

Let  $p(u)_{1 \leq u < \widehat{\lambda}}$  be the unique sequence such that  $\{p(u) : 1 \leq u < \widehat{\lambda}\} = \Phi^\#$  with  $p(u) < p(u+1)$  whenever  $1 \leq u < u+1 < \widehat{\lambda}$  where  $\widehat{\lambda} = \infty$  or  $\text{card}(\Phi^\#) + 1$  according as the cardinality  $\text{card}(\Phi^\#)$  is infinite or finite.

Let us define the translation set  $\Psi^\#$  of (0<sup>#</sup>) by putting

$$\Psi^\# = \begin{cases} \text{the set of all } j \in \mathbb{N} \text{ such that} \\ \text{for every } t \in \{0, \dots, \tau\} \text{ with } z_{tj} \neq 0 \text{ we have} \\ \frac{z_{tj}}{z_{t(j)j}^{\pi(t,j)}} \in V \setminus M(V) \text{ for some } n(t, j) \in \mathbb{Z} \end{cases} \quad (7^\#)$$

and let us note that this defines  $n(t, j)$  uniquely. Let  $u(i)_{0 \leq i < \widehat{\kappa}}$  be the unique sequence such that  $\{p(u(i)) : 0 \leq i < \widehat{\kappa}\} = \Phi^\# \cap \Psi^\#$  with  $u(i) < u(i+1)$  whenever



$0 \leq i < i + 1 < \widehat{\kappa}$  where  $\widehat{\kappa} = \infty$  or  $\text{card}(\Phi^\# \cap \Psi^\#)$  according as the cardinality  $\text{card}(\Phi^\# \cap \Psi^\#)$  is infinite or finite.

We call  $(0^\#)$  the QDT *sequence* of  $(z_0, \dots, z_\tau)$  along  $(V, K)$  and we call

$$(\pi(0, j), \dots, \pi(\tau, j), p(u), u(i))_{0 \leq j < \infty, 1 \leq u < \widehat{\lambda}, 0 \leq i < \widehat{\kappa}} \quad (8^\#)$$

the *supplement* of the QDT sequence.

*Note (I<sup>\*</sup>).* The proofs of the following Lemmas (I) and (II) are straightforward. Lemma (II) deals with a situation when  $(z_{0j}, \dots, z_{\tau j})$  are generators of the maximal ideal  $M(R_j)$  of a local domain  $R_j$  dominated by  $V$ ; in that situation clearly  $j$  belongs to  $N^\#$  where  $N^\# = \{j \in \mathbb{N} : W(z_{tj}) > 0 \text{ for } 0 \leq t \leq \tau\}$ . Note that if  $j \in N^\#$  then only the first line of  $(2^\#)$  is relevant. Also note that:

- (i)  $j \in N^\#$  for a certain value of  $j$  implies  $j \in N^\#$  for all bigger values of  $j$ .
- (ii)  $t_+^*(j) \neq \emptyset$  for a certain value of  $j$  implies  $t_+^*(j) \neq \emptyset$  for all bigger values of  $j$ .
- (iii) If  $W$  is real, i.e., if the value group  $G_W$  is order isomorphic to an additive subgroup of  $\mathbb{R}$  then  $j \in N^\#$  for all sufficiently large values of  $j$ .
- (iv) If  $j < j^*$  in  $\mathbb{N} \cup \{\infty\}$  are such that  $t(j) = t(j')$  for all  $j \leq j' < j^*$  then  $z_{t(j)j} = z_{t(j')j'}$  whenever  $j \leq j' < j^*$ .

Finally note that by definition

$$|W(z)| = W(z) \text{ or } -W(z) \text{ according as } W(z) \geq 0 \text{ or } W(z) < 0.$$

**Lemma (I).** *Let  $j \in \Phi^\#$  and  $j < j^* \in \mathbb{N}_+ \cup \{\infty\}$  be such that for all  $j' \in \mathbb{N}$  with  $j < j' < j^*$  we have  $j' \notin \Phi^\#$ , and if  $j^* \neq \infty$  then we have  $j^* \in \Phi^\#$ . Then we have the following.*

- (I.1) For all  $j' \in \mathbb{N}$  with  $j \leq j' < j^*$  we have  $t(j') = t(j)$  and  $z_{t(j')j'} = z_{t(j)j}$ . If  $j^* \neq \infty$  then we have  $t(j^*) \neq t(j)$ .
- (I.2) If  $j^* = \infty$  then we have  $1 < \widehat{\lambda} < \infty$  and  $p(\widehat{\lambda} - 1) = j$ . If  $j^* \neq \infty$  then for a unique integer  $u$  with  $1 \leq u < u + 1 < \widehat{\lambda}$  we have  $p(u) = j < j^* = p(u + 1)$ .
- (I.3) Assume  $j \notin \Psi^\#$ . Then for  $0 \leq i < \widehat{\kappa}$  we have  $j \neq p(u(i))$ . Moreover, either: for all  $t \in \{0, \dots, \tau\} \setminus \{t(j)\}$  we have  $z_{tj} = 0$ , or: for some  $t \in \{0, \dots, \tau\} \setminus \{t(j)\}$  we have  $z_{tj} \neq 0$  with  $z_{tj}/z_{t(j)j}^n \notin V \setminus M(V)$  for all  $n \in \mathbb{Z}$ . In the “either” case, for all  $t \in \{0, \dots, \tau\} \setminus \{t(j)\}$  and for all  $j' \in \mathbb{N}$  with  $j \leq j' < j^*$ , we have  $c_{tj'} = 0 = z_{tj'}$ . Furthermore, for every  $t$  of the “or” case and for all  $j' \in \mathbb{N}$  with  $j \leq j' < j^*$ , we have

$$c_{tj'} = 0 \text{ with } \pi(t, j') = \pi(t, j) \text{ and } z_{tj} = z_{t(j)j}^{\pi(t,j)(j'-j)} z_{tj'}.$$

- (I.4) Assume  $j \in \Psi^\#$ . Then  $j = p(u(i))$  for a unique  $i$  with  $0 \leq i < \widehat{\kappa}$ . Moreover, if  $j^* = \infty$  then for any  $t \in \{0, \dots, \tau\} \setminus \{t(j)\}$ , whereas if  $j^* \neq \infty$  then for  $t = t(j^*)$ , for all  $j' \in \mathbb{N}$  with  $j \leq j' < j^*$  we have

$$z_{tj} - \sum_{j \leq v < j'} c_{tv} z_{t(j)j}^{\pi(t,j)(v-j+1)} = z_{t(j)j}^{\pi(t,j)(j'-j)} z_{tj'} \quad \text{with } \pi(t, j') = \pi(t, j) \quad \text{(i)}$$

which may be viewed as a Taylor Expansion with Remainder discussed in (9.5). If  $j^* = \infty$  then (i) gives rise to the equation

$$z_{tj} = \sum_{j \leq v < j^*} c_{tv} z_{t(j)j}^{\pi(t,j)(v-j+1)} \quad \text{(ii)}$$

which may be thought of as an infinite Taylor Expansion discussed in (9.2), with a suitable interpretation of the equality; see (9.3) for the case when  $V$  is a DVR.

(I.5) Assuming  $j \in \Psi^\#$  and letting  $\bar{v} = \{v \in \mathbb{N} : j \leq v < j^* \text{ with } c_{tv} \neq 0\}$  we have the following. If  $j^* = \infty$  and  $t \in \{0, \dots, \tau\} \setminus \{t(j^*)\}$  then letting  $v_1 < \dots < v_w$  or  $v_1 < v_2 < \dots$  be the finitely many or infinitely many values of  $v \in \bar{v}$  and putting  $v_0 = j - 1$  we have

$$n(t, v_v + 1) = v_{v+1} - v_v$$

for  $0 \leq v < w$  or  $0 \leq v < \infty$  respectively. If  $j^* \neq \infty$  and  $t = t(j^*)$  then letting  $v_1 < \dots < v_w$  be the values of  $v \in \bar{v}$  and putting  $v_0 = j - 1$  we have

$$n(t, v_v + 1) = v_{v+1} - v_v$$

for  $0 \leq v < w$ .

Sketch proof of (I.5). Letting

$$X = z_{t(j)j} \quad \text{and} \quad (N(q), Z(q)) = (n(t, q), z_{tq})$$

for all  $q \in \mathbb{N}$ , we have

$$Z(q) = C(q)X^{N(q)} + X^{N(q)}Z(q + N(q)) \quad \text{where } 0 \neq C(q) \in K. \quad \text{[q]}$$

Comparing  $[v_0 + 1]$  and (i) with  $j' = v_1 + 1$  we see that

$$N(v_0 + 1) = v_1 - v_0 \quad \text{and} \quad C(v_0 + 1) = c_{tv_1}$$

with

$$Z(v_0 + 1) = c_{tv_1}X^{v_1-v_0} + X^{v_1-v_0}Z(v_1 + 1).$$

Substituting  $[v_1 + 1]$  in the last equation and comparing the resulting equation and (i) with  $j' = v_2 + 1$  we see that

$$N(v_1 + 1) = v_2 - v_1 \quad \text{and} \quad C(v_1 + 1) = c_{tv_2}$$

with

$$Z(\nu_0 + 1) = c_{t\nu_1} X^{\nu_1 - \nu_0} + c_{t\nu_2} X^{\nu_2 - \nu_1} + X^{\nu_2 - \nu_1} Z(\nu_2 + 1).$$

And so on. Thus by induction on  $\nu$  we get

$$N(\nu_\nu + 1) = \nu_{\nu+1} - \nu_\nu \text{ and } C(\nu_\nu + 1) = c_{\nu_{\nu+1}}$$

with

$$Z(\nu_0 + 1) = c_{t\nu_1} X^{\nu_1 - \nu_0} + c_{t\nu_2} X^{\nu_2 - \nu_1} + \dots + c_{t\nu_{\nu+1}} X^{\nu_{\nu+1} - \nu_\nu} + X^{\nu_{\nu+1} - \nu_\nu} Z(\nu_{\nu+1} + 1)$$

for all relevant values of  $\nu$ .

**Lemma (II).** *Assume that  $V$  dominates a positive dimensional local domain  $R$  for which  $M(R) = (z_0, \dots, z_\tau)R$ . Let  $(R_j)_{0 \leq j < \infty}$  be the QDT sequence of  $R$  along  $V$ . Then we have the following.*

(II.1) If  $n \in \mathbb{N}$  is such that  $K$  contains a coefficient set  $K_j$  of  $R_n$  for  $0 \leq j \leq n$ , then for  $0 \leq j \leq n$  we have

$$\{c_{0j}, \dots, c_{\tau j}\} \in K_j = K \cap R_j \text{ and } M(R_j) = (z_{0j}, \dots, z_{\tau j})R_j.$$

(II.2) If  $V$  is a DVR then  $\widehat{\lambda} \in \mathbb{N}_+$  and for all integers  $n \geq \widehat{\lambda}$  we have  $t(n) = t(\widehat{\lambda})$ . If  $V$  is a DVR and  $\widehat{QF}(R) = \widehat{QF}(V)$  then  $\widehat{\lambda} \in \mathbb{N}_+$  and for all integers  $n \geq \widehat{\lambda}$  we have  $t(n) = t(\widehat{\lambda})$  and  $W(z_{t(n)n}) = 1$ .

(II.3) If  $R$  is regular of dimension  $\tau + 1$  and  $V$  is a prime divisor of  $R$  then there exists a unique positive integer  $n$  such that for all integers  $0 \leq j < n \leq \mu$  we have  $R_j \neq R_n = R_\mu = V$  and  $\dim(R_j) > \dim(R_n) = \dim(R_\mu) = 1$ . Moreover,  $R_n$  is residually pure transcendental over  $R_{n-1}$  of residual transcendence degree  $\dim(R_{n-1}) - 1$ . Finally  $n$  is the essential length of the QDT sequence  $(R_j)_{0 \leq j < \infty}$  in the sense of Note (II\*\*) below.

(II.4) If  $R$  is one dimensional and  $V$  is a prime divisor of  $R$  then  $V$  is residually finite algebraic over  $R$  and there exists  $n \in \mathbb{N}$  such that for all integers  $\mu \geq n$  we have  $M(V) = M(R_\mu)V$  with  $V/M(V) = R_\mu/M(\mu)$ .

*Note (II\*).* For (II.3) see Proposition 3 of [2] and its proof. The first part of (II.4) is proved in Theorem 1(4) of [2], and the rest of (II.4) follows from it by (II.2). It may be tempting to think that (II.4) implies  $V = R_\mu$  for large  $\mu$ , but Example (E3.2) on page 206 of Nagata [28] shows this to be untrue.

*Note (II\*).* Given any positive dimensional local domain  $R$  and any QDT sequence  $(R_j)_{0 \leq j < \infty}$  of  $R$ , by the *essential length* of the QDT sequence we mean the unique  $n \in \mathbb{N} \cup \{\infty\}$  such that if  $n = \infty$  then for all  $j \in \mathbb{N}$  we have  $R_j \neq R_{j+1}$ , whereas if  $n \in \mathbb{N}$  then for all  $j \in \mathbb{N}$  with  $j < n$  we have  $R_j \neq R_{j+1}$  and for all  $j \in \mathbb{N}$  with  $j \geq n$  we have  $R_j = R_{j+1}$ . Note that  $R_j = R_{j+1}$  iff  $R_j$  is a DVR.

**Lemma (III).** Assume that  $\tau = 1$  and  $V$  dominates a two dimensional regular local domain  $R$  with quotient field  $L$  and  $M(R) = (z_0, z_1)R$ . Let  $(R_j)_{0 \leq j < \infty}$  be the QDT sequence of  $R$  along  $V$ . Then we have the following.

- (III.1) The essential length of the QDT sequence  $(R_j)_{0 \leq j < \infty}$  is finite or infinite according as  $V$  is residually transcendental or algebraic over  $R$ .
- (III.2) If  $V$  is residually transcendental over  $R$  then  $V$  is a prime divisor of  $R$ .
- (III.3) Assume that  $V$  is residually algebraic over  $R$ . Then the value group  $G_W$  is order isomorphic to either (i) the set of all lexicographically ordered pairs of integers or (ii) the additive group of all integers or (iii) a non-cyclic additive subgroup of  $\mathbb{Q}$  or (iv) an additive subgroup of  $\mathbb{R}$  of the form  $\{a_1\pi_1 + a_2\pi_2 : (a_1, a_2) \in \mathbb{Z}^2\}$  for some positive real numbers  $\pi_1, \pi_2$  which are linearly independent over  $\mathbb{Q}$ . In these cases, we shall respectively say that  $V$  is nonreal discrete or real discrete or rational nondiscrete or irrational. Now assume that  $K$  contains a coefficient set  $K_j$  of  $R_j$  for all  $j \in \mathbb{N}$ . Then:

- (i\*)  $\text{card}(\Phi^\#) \neq \infty \neq \text{card}(\Psi^\#)$  iff  $V$  is nonreal discrete;
- (ii\*)  $\text{card}(\Phi^\#) \neq \infty = \text{card}(\Psi^\#)$  iff  $V$  is real discrete;
- (iii\*)  $\text{card}(\Phi^\#) = \infty = \text{card}(\Psi^\#)$  iff  $V$  is rational nondiscrete;
- (iv\*)  $\text{card}(\Phi^\#) = \infty \neq \text{card}(\Psi^\#)$  iff  $V$  is irrational.

*Proof.* In view of Lemma (II) this follows from [2, 3].

*Note (III\*).* In the next two Lemmas we continue to give special attention to the  $\tau = 1$  case. Here we make some definitions for that case. For any nonnegative integer  $j$  we let  $t'(j)$  be the unique member of  $\{0, 1\}$  different from  $t(j)$ . By the quadratic expansion of any  $(z_0, z_1) \in L^2 \setminus \overline{L}^2$  along  $(V, K)$  we mean the sequence

$$(z_{0j}, z_{1j}, c_{0j}, c_{1j}, t(j), t'(j))_{0 \leq j < \infty} \tag{9^\#}$$

where  $(z_{0j}, z_{1j}, c_{0j}, c_{1j}, t(j))_{0 \leq j < \infty}$  is the  $\tau = 1$  version of (0<sup>#</sup>); moreover, by the supplement of the quadratic expansion we mean the  $\tau = 1$  version of (8<sup>#</sup>), i.e.,

$$(\pi(0, j), \pi(1, j), p(u), u(i))_{0 \leq j < \infty, 1 \leq u < \widehat{\lambda}, 0 \leq i < \widehat{\kappa}} \tag{10^\#}$$

Since the euclidean algorithm played a crucial role in it, the  $(V, K)$ -expansion

$$(z_j, e_j, p_j, B_j(v), \epsilon(i), t(j))_{v \in \mathbb{Z}, 0 \leq j \leq \lambda, 0 \leq i \leq \kappa} \tag{\bullet\bullet}$$

introduced in (4.1) is called the euclidean expansion of  $(z_0, z_1)$  along  $(V, K)$ , and (•••) is called the mixed euclidean expansion of  $(z_0, z_1)$  along  $(V, K, T)$ . In Lemma (IV) we shall give a stand alone description of the quadratic expansion. In Lemma (V) we shall restate the  $\tau = 1$  case of Lemma (I). In Part II, we shall compare the quadratic expansion with the euclidean expansion.

**Lemma (IV).** Assuming  $\tau = 1$ , for the quadratic expansion (9<sup>#</sup>) of  $(z_0, z_1)$  along  $(V, K)$  with  $(z_{0j}, z_{1j}) \in L^2 \setminus \overline{L}^2$  for  $0 \leq j < \infty$ , we have the following.

(IV.1) Recalling that for every  $j \in \mathbb{N}$  we have  $(z_{0j}, z_{1j}) \in L^2 \setminus \overline{L}^2$ , we can paraphrase the characterizations (3<sup>#</sup>)–(5<sup>#</sup>) of  $t(j)$  and  $\pi(t, j)$  by saying that  $t(j) \in \{0, 1\}$  with  $z_{t(j)j} \notin \overline{L}$  and with  $\pi(t(j), j) = 1$  with  $\pi(t'(j), j) \in \{0, 1, -1\}$  satisfy (1)–(8) stated below.

- (1) If  $0 < W(z_{1j}) \leq W(z_{0j})$  then  $t(j) = 1$  and  $\pi(t'(j), j) = 1$ .
- (2) If  $0 < W(z_{0j}) < W(z_{1j})$  then  $t(j) = 0$  and  $\pi(t'(j), j) = 1$ .
- (3) If  $W(z_{1j}) < 0 < W(z_{0j})$  then  $t(j) = 1$  and  $\pi(t'(j), j) = 1$ .
- (4) If  $W(z_{1j}) > 0 > W(z_{0j})$  then  $t(j) = 1$  and  $\pi(t'(j), j) = -1$ .
- (5) If  $W(z_{1j}) < 0 < -W(z_{1j}) \leq W(z_{0j})$  then  $t(j) = 1$  and  $\pi(t'(j), j) = -1$ .
- (6) If  $W(z_{1j}) < 0 < W(z_{0j}) < -W(z_{1j})$  then  $t(j) = 0$  and  $\pi(t'(j), j) = -1$ .
- (7) If  $W(z_{1j}) \neq 0 = W(z_{0j})$  then  $t(j) = 1$  and  $\pi(t'(j), j) = 0$ .
- (8) If  $W(z_{1j}) = 0 \neq W(z_{0j})$  then  $t(j) = 0$  and  $\pi(t'(j), j) = 0$ .

(IV.2) Next the definitions (1<sup>#</sup>) and (2<sup>#</sup>) can be paraphrased by saying that for  $0 \leq j < \infty$  and  $0 \leq t \leq 1$  we have

$$z_{t,j+1} = \begin{cases} z_{tj} \text{ with } c_{tj} = 1 & \text{if } t = t(j) \\ \bar{z}_{tj} \text{ with } c_{tj} = 0 & \text{if } t = t'(j) \text{ and } W(\bar{z}_{tj}) \neq 0 \\ \bar{z}_{tj} - c_{tj} \in M(V) \text{ with } c_{tj} \neq 0 & \text{if } t = t'(j) \text{ and } W(\bar{z}_{tj}) = 0 \end{cases}$$

where

$$\bar{z}_{tj} = \frac{z_{tj}}{z_{t(j)j}^{\pi(t,j)}}.$$

(IV.3) To paraphrase definition (6<sup>#</sup>) of the flipping set  $\Phi^\#$ , recalling that

$$\Phi^\# = \text{the set of all } j \in \mathbb{N}_+ \text{ such that } t(j-1) \neq t(j).$$

and

$$\widehat{\lambda} = \text{card}(\Phi^\#) + 1 \in \mathbb{N}_+ \cup \{\infty\}$$

we supplement the definition of  $p(u)_{1 \leq u < \widehat{\lambda}}$  by the *convention*

$$p(-1) = p(0) = 0$$

and we note that now the members of  $\Phi^\# \cup \{0\}$  are labelled as

$$p(-1) = p(0) = 0 < p(1) < p(2) < \dots \quad \text{if } \widehat{\lambda} = \infty$$

and

$$p(-1) = p(0) = 0 < p(1) < \dots < p(\widehat{\lambda} - 1) \quad \text{if } \widehat{\lambda} \in \mathbb{N}_+.$$

(IV.4) To paraphrase definition (7<sup>#</sup>) of the translation set  $\Psi^\#$ , recalling that

$$\Psi^\# = \begin{cases} \text{the set of all } j \in \mathbb{N} \text{ such that} \\ \text{for every } t \in \{0, 1\} \text{ with } z_{tj} \neq 0 \text{ we have} \\ \frac{z_{tj}}{z_{t(j)j}^n} \in V \setminus M(V) \text{ for a (unique) } n(t, j) \in \mathbb{Z} \end{cases}$$

and

$$\widehat{\kappa} = \text{card}(\Phi^\# \cap \Psi^\#) \in \mathbb{N} \cup \{\infty\}$$

we supplement the definition of  $u(i)_{0 \leq i < \widehat{\kappa}}$  by the *convention*

$$u(-1) = 0$$

and we note that now we have the integer sequences

$$u(-1) = 0 < u(0) < u(1) < \dots \quad \text{if } \widehat{\kappa} = \infty$$

and

$$u(-1) = 0 < u(0) < \dots < u(\widehat{\kappa} - 1) \quad \text{if } \widehat{\kappa} \in \mathbb{N}$$

while the members of  $(\Phi^\# \cap \Psi^\#) \cup \{0\}$  are labelled as

$$p(u(-1)) = 0 < p(u(0)) < p(u(1)) < \dots \quad \text{if } \widehat{\kappa} = \infty$$

and

$$p(u(-1)) = 0 < p(u(0)) < \dots < p(u(\widehat{\kappa} - 1)) \quad \text{if } \widehat{\kappa} \in \mathbb{N}.$$

**Lemma (V).** *Assume  $\tau = 1$ . Let  $j \in \Phi^\#$  and  $j < j^* \in \mathbb{N}_+ \cup \{\infty\}$  be such that for all  $j' \in \mathbb{N}$  with  $j < j' < j^*$  we have  $j' \notin \Phi^\#$ , and if  $j^* \neq \infty$  then we have  $j^* \in \Phi^\#$ . Then we have the following.*

- (V.1) For all  $j' \in \mathbb{N}$  with  $j \leq j' < j^*$  we have  $t(j') = t(j)$  and  $z_{t(j')j'} = z_{t(j)j}$ . If  $j^* \neq \infty$  then we have  $t(j^*) = t'(j)$ .
- (V.2) If  $j^* = \infty$  then we have  $1 < \widehat{\lambda} < \infty$  and  $p(\widehat{\lambda} - 1) = j$ . If  $j^* \neq \infty$  then for a unique integer  $u$  with  $1 \leq u < u + 1 < \widehat{\lambda}$  we have  $p(u) = j < j^* = p(u + 1)$ .
- (V.3) Assume  $j \notin \Psi^\#$ . Then for  $0 \leq i < \widehat{\kappa}$  we have  $j \neq p(u(i))$ . Moreover, either:  $z_{t'(j)j} = 0$ , or:  $z_{t'(j)j} \neq 0$  with  $z_{t'(j)j}/z_{t'(j)j}^n \notin V \setminus M(V)$  for all  $n \in \mathbb{Z}$ . In the “either” case, for  $t = t'(j)$  and for all  $j' \in \mathbb{N}$  with  $j \leq j' < j^*$ , we have  $c_{tj'} = 0 = z_{tj'}$ . In the “or” case, for  $t = t'(j)$  and for all  $j' \in \mathbb{N}$  with  $j \leq j' < j^*$ , we have

$$c_{tj'} = 0 \quad \text{with} \quad \pi(t, j') = \pi(t, j) \quad \text{and} \quad z_{tj} = \frac{\pi(t, j)(j'-j)}{z_{t(j)j}} z_{tj'}.$$

(V.4) Assume  $j \in \Psi^\sharp$ . Then  $j = p(u(i))$  for a unique  $i$  with  $0 \leq i < \widehat{\kappa}$ . Moreover, if  $j^* = \infty$  then for  $t = t'(j)$ , whereas if  $j^* \neq \infty$  then for  $t = t(j^*)$ , for all  $j' \in \mathbb{N}$  with  $j \leq j' < j^*$  we have

$$z_{tj} - \sum_{j \leq v < j'} c_{tv} z_{t(j)j}^{\pi(t,j)(v-j+1)} = z_{t(j)j}^{\pi(t,j)(j'-j)} z_{tj'} \quad \text{with } \pi(t, j') = \pi(t, j) \tag{i}$$

which may be viewed as a Taylor Expansion with Remainder discussed in (9.5). If  $j^* = \infty$  then (i) gives rise to the equation

$$z_{tj} = \sum_{j \leq v < j^*} c_{tv} z_{t(j)j}^{\pi(t,j)(v-j+1)} \tag{ii}$$

which may be thought of as an infinite Taylor Expansion discussed in (9.2), with a suitable interpretation of the equality; see (9.3) for the case when  $V$  is a DVR.

(V.5) Assuming  $j \in \Psi^\sharp$  and letting  $\bar{v} = \{v \in \mathbb{N} : j \leq v < j^* \text{ with } c_{tv} \neq 0\}$  we have the following. If  $j^* = \infty$  and  $t = t'(j^*)$  then letting  $v_1 < \dots < v_w$  or  $v_1 < v_2 < \dots$  be the finitely many or infinitely many values of  $v \in \bar{v}$  and putting  $v_0 = j - 1$  we have

$$n(t, v_v + 1) = v_{v+1} - v_v$$

for  $0 \leq v < w$  or  $0 \leq v < \infty$  respectively. If  $j^* \neq \infty$  and  $t = t(j^*)$  then letting  $v_1 < \dots < v_w$  be the values of  $v \in \bar{v}$  and putting  $v_0 = j - 1$  we have

$$n(t, v_v + 1) = v_{v+1} - v_v$$

for  $0 \leq v < w$ .

Note on inversion and invariance (VI). The three Inversion Theorems of Sects. (3.7) and (3.9), the two Invariance Theorems of Sect. 4, and the above quadratic transformation Lemmas (I)–(V) of this section are refinements of the results of my papers [2, 4]. More about all this in Part II.

## 6 Dicritical Divisors

The concept of dicritical divisors arose in the topological study of a map  $\mathbb{C}^2 \rightarrow \mathbb{C}$  given by a polynomial  $f \in k[X, Y] \setminus k$  when  $k$  is the field of complex numbers. The term dicritical divisor seems to have been introduced by Mattei and Moussu [27], and was then used by Artal-Bartolo [16], Eisenbud–Neumann [21], Fourier [22], Le–Weber [26], Neumann [29], Rudolph [31], and others. On the other hand, Pierrette Cassou-Noguès [18, 19] and Neumann–Norbury [30] use the alternative term horizontal divisors.

In Definition (6.1) we introduce the algebraic incarnation of dicritical divisors. In Note (6.2) we pay a heuristic visit to the original topological version.

The dicritical divisors may be viewed as a nonempty finite set of univariate polynomials strategically (and quite algebraically) located inside the belly of a randomly chosen bivariate polynomial. It is certainly amazing that, until 1980, no endoscopic examination of bivariate polynomial bellies (=affine plane curve bellies) revealed their existence. We have stressed “and quite algebraically” to indicate that in our treatment we do not use any topology or analysis which, under the pretext of geometric viewpoint, only muddies the water. Of course, it may be admitted that one person’s clarity can be another person’s muddying of waters and vice versa. Positively speaking, muddying may amount to stirring!!

In Note (6.6) I shall introduce the dicritical divisor theory of local rings and compare it to the analogous theory of quasirational and nonquasirational surface singularities coming out of my papers [2, 8].

Preamble for (6.1)–(6.4). Let us consider the bivariate polynomial ring  $B = k[X, Y]$  over a field  $k$  and let  $L = k(X, Y) = \text{QF}(B)$  where  $\text{QF}(B)$  denotes the quotient field of  $B$ . Given any

$$f = f(X, Y) \in B \setminus k$$

of (total) degree  $N$ , by  $B_f$  we denote the localization of  $B$  at the multiplicative set  $k[f]^\times$ , and we note that then  $B_f$  is the affine domain  $k(f)[X, Y]$  over the field  $k(f)$  with  $\text{QF}(B_f) = k(X, Y) = L$  and we have  $\text{trdeg}_{k(f)} L = 1$ . Now a localization of a UFD is a UFD, and irreducibles in the localization are essentially the same as irreducibles in the original UFD except that the localization has more units. Consequently  $B_f$  is a one-dimensional UFD and hence it is a DD as well as a PID. It follows that  $B_f$  is the affine coordinate ring of an irreducible nonsingular affine plane curve over  $k(f)$ .

Note that  $D(L/k)$  is the set of all valuation rings  $V$  with  $\text{QF}(V) = L$  and  $k \subset V$  such that  $\text{trdeg}_k H(V) = 1$  where  $H_V : V \rightarrow H(V) = V/M(V)$  is the residue class epimorphisms; moreover, every member of  $D(L/k)$  is a DVR, and  $I(B/k)$  is the set of all  $V \in D(L/k)$  with  $B \not\subset V$ . Also note that  $D(L/k(f))$  is the set of all valuation rings  $V$  with  $\text{QF}(V) = L$  and  $k(f) \subset V \neq L$ ; moreover, every member of  $D(L/k(f))$  is a DVR, and  $I(B_f/k(f))$  is the set of all  $V \in D(L/k(f))$  with  $B_f \not\subset V$ .

**Definition (6.1).** For every  $V \in D(L/k(f))$  we put

$$\deg(V) = \deg_f V = [H(V) : k(f)] \in \mathbb{N}_+$$

and we call this the  $f$ -degree of  $V$ , or briefly the degree of  $V$ . Moreover, for every  $V \in I(B_f/k(f))$  we put

$$\text{ind}(V) = \text{ind}_f V = -\min(\text{ord}_V X, \text{ord}_V Y) \in \mathbb{N}_+$$



and we call this the  $f$ -index of  $V$ , or briefly the index of  $V$ . Finally we put

$$I(B/k, f) = \left\{ \begin{array}{l} \text{the set of all } V \in I(B/k) \text{ at which} \\ f \text{ is residually transcendental over } k \end{array} \right.$$

and we observe that

$$I(B/k, f) = I(B_f/k(f)) = \text{a nonempty finite set.} \tag{†}$$

Now labelling the distinct members of  $I(B/k, f)$  as  $V_1, \dots, V_m$ , we call them the *dicritical divisors* of  $f$  (in  $B$ ). In Part II we shall show that by the “sigma-eyee-feye” formula from extension theory of DVRs we have

$$\sum_{1 \leq i \leq m} \text{ind}(V_i) \text{deg}(V_i) = N \tag{•}$$

and, if  $k$  is algebraically closed and is of characteristic zero, then for  $1 \leq i \leq m$

$$\left\{ \begin{array}{l} H(V_i) = H_{V_i}(k(t_i)) \text{ for some } t_i \in V_i \text{ so that } H_{V_i}(f) = H_{V_i}(P_i(t_i)) \\ \text{where } P_i(Z) \in k[Z] \setminus k \text{ is a univariate polynomial} \\ \text{whose } Z\text{-degree equals } \text{deg}(V_i) \end{array} \right. \tag{••}$$

In the proof we shall use Newton’s fractional power series expansion. In Part II we shall also show that the characteristic zero hypothesis can be removed by replacing Newton expansion by Hamburger-Noether expansion. Note that the integers  $m$  and  $\text{deg}(V_1), \dots, \text{deg}(V_m)$  depend only on  $f$  as a element of the ring  $B$  and not on the particular generators  $X, Y$  of that ring, but the integers  $\text{ind}(V_1), \dots, \text{ind}(V_m)$  do depend on  $X, Y$ , as will be shown in Example (6.5).

*Note (6.2).* Momentarily assuming  $k$  to be the complex number field  $\mathbb{C}$ , the dicritical divisors may be “heuristically explained” thus. The polynomial map  $\mathbb{C}^2 \rightarrow \mathbb{C}^1$  which is given by  $(a, b) \mapsto f(a, b)$  can be extended to a rational map  $P^2 \rightarrow P^1$  of the complex projective plane to the complex projective line. But as a “rational map” it may have points of indeterminacy. We get rid of these by “blowing up”  $P^2$  to get a compact complex nonsingular surface  $W$  on which the map  $f$  extends to a well defined map  $\phi : W \rightarrow P^1$ . Just as  $P^2$  is obtained by adding one projective line (called the line at infinity) to  $\mathbb{C}^2$ , the surface  $W$  is obtained by adding a finite number of projective lines  $P_1^1, \dots, P_n^1$  to  $\mathbb{C}^2$ . Consideration of connectivity tells us that, depending on the particular line  $P_i^1$ , the restriction of the map  $\phi$  to  $P_i^1$  maps it either onto the entire target line  $P^1$  or to a single point of it, i.e., it is either surjective or constant. Those  $P_i^1$  for which it is surjective are called dicritical divisors. By suitably relabelling, we may assume that  $P_1^1, \dots, P_m^1$  are dicritical while  $P_{m+1}^1, \dots, P_n^1$  are not. It can be shown that  $m$  is positive. Moreover, it can also be shown that by deleting a suitable point from a dicritical  $P_i^1$  and also deleting a

suitable point from the target  $P^1$ , the resulting map  $\mathbb{C}^1 \rightarrow \mathbb{C}^1$  is given by a univariate polynomial  $P_i(Z)$  of some degree  $d_i$ ; note that  $d_i$  is the degree of the ramified covering  $P_i^1 \rightarrow P^1$ . By rotating the axes, i.e., by making a homogeneous linear transformation, we may assume that  $f$  is monic of degree  $N$  in  $Y$ . It turns out that then

$$\sum_{1 \leq i \leq m} e_i d_i = N$$

where the positive integer  $e_i$  is the ramification index coming out of the Dedekind Domain theory which is the same thing as the Riemann Surface theory.

As a side remark recall that  $f$  is a field generator means  $k(f, g) = k(X, Y)$  for some rational function  $g \in k(X, Y)$ ; it turns out that if the polynomial  $f$  is a field generator then the complementary generator  $g$  can be chosen to be a polynomial iff  $d_i = 1$  for some dicritical  $P_i^1$ . Without assuming  $f$  to be a field generator, how do we show that the dicritical divisors are independent of the particular blow up  $W$  and how do we algebraize them?

To consider the independence, let  $\bar{\phi} : \bar{W} \rightarrow P^1$  be any other blow up, and label the projective lines in  $\bar{W} \setminus \mathbb{C}^2$  as  $\bar{P}_1^1, \dots, \bar{P}_m^1, \bar{P}_{m+1}^1, \dots, \bar{P}_n^1$  so that the first  $\bar{m}$  are dicritical while the remaining ones are not. It can be shown that there exists a blow up  $\tilde{\phi} : \tilde{W} \rightarrow P^1$  together with maps  $\theta : \tilde{W} \rightarrow W$  and  $\bar{\theta} : \tilde{W} \rightarrow \bar{W}$  such that  $\tilde{\phi} \theta = \bar{\phi} \bar{\theta} = \phi \bar{\theta}$ . Label the projective lines in  $\tilde{W} \setminus \mathbb{C}^2$  as  $\tilde{P}_1^1, \dots, \tilde{P}_m^1, \tilde{P}_{m+1}^1, \dots, \tilde{P}_n^1$  so that the first  $\tilde{m}$  are dicritical while the remaining ones are not. It can be shown that  $\tilde{m} = \bar{m} = m$  and after suitable labelling, for  $1 \leq i \leq m$ , we have  $\theta(\tilde{P}_i^1) = P_i^1$  and  $\bar{\theta}(\tilde{P}_i^1) = \bar{P}_i^1$  with induced bijections  $\tilde{P}_i^1 \rightarrow P_i^1$  and  $\tilde{P}_i^1 \rightarrow \bar{P}_i^1$ .

Now let us proceed to the algebraization which will actually reprove the independence. Recall that: for any finitely generated field extension  $L$  of a field  $K$  we have put  $D(L/K) =$  the set of all prime divisors of  $L/K$ , i.e., the set of all DVRs  $V$  with quotient field  $\text{QF}(V) = L$  such that  $K \subset V$  and  $\text{trdeg}_K H(V) = (\text{trdeg}_K L) - 1$  where  $H(V) = V/M(V) =$  the residue field of  $V$ ; for any affine domain  $A$  over  $K$  with  $\text{QF}(A) = L$  we have put  $I(A/K) =$  the set of all infinity divisors of  $A/K$ , i.e., the set of all  $V \in D(L/K)$  such that  $A \not\subset V$ . Henceforth, we consider the bivariate polynomial ring  $B = k[X, Y]$  over a field  $k$  and we let  $\text{QF}(B) = L = k(X, Y)$  and we put  $I(B/k, f) =$  the set of all those members  $V$  of  $I(B/k)$  for which  $f$  is residually transcendental over  $k$ . Let  $V_i$  be the local ring of  $P_i^1$  on  $W$ . Then clearly  $V_i \in I(B/k)$  for  $1 \leq i \leq n$ , and we have:  $V_i \in I(B/k, f) \Leftrightarrow 1 \leq i \leq m$ .

It can also be shown that  $I(B/k) =$  the totality of the local rings of the projective lines on various blow ups of  $P^2$  which are in the complements of  $\mathbb{C}^2$ . At any rate,  $I(B/k, f)$  is a nonempty finite set which we have defined without any aid of blowing ups, and this is our algebraic definition of dicritical divisors of  $f$ . Since  $I(B/k, f)$  does correspond to the geometrically defined dicritical divisors on any blow up of  $P^2$  on which the rational map  $P^2 \rightarrow P^1$  becomes well-defined, this reproves the independence in a more succinct manner; the geometric proof sketched in the paragraph before last was rather fuzzy at best. This is the beauty of the approach by “models” which are collections of local rings and so on; for details see the Algebra and Geometry books [9, 12].

Now the  $I(B/k, f)$  from surface theory coincides with the  $I(B_f/k(f))$  from curve theory, where we have put  $B_f = k(f)[X, Y]$ . Note that  $B_f$  can be identified with the affine coordinate ring of the generic curve  $f^\# = 0$  where we take an indeterminate  $u$  over  $k$  and put  $f^\# = f - u$ . Substituting  $f$  for  $u$ , this generic curve acquires the confusing equation  $f = f$ . The confusion (like the Maya covering the Brahma) can be removed by using two sets of variables giving  $f(\bar{X}, \bar{Y}) = f(X, Y)$ . Indeed, experience shows that such  $f = f$  arguments provide exceptionally powerful tools! Although the curve  $f = 0$  may be reducible and may even have multiple components and may be full of singularities, but miraculously the curve  $f^\# = 0$  is irreducible and nonsingular. The best way to see this is to realize  $B_f$  as the localization of  $B$  at the multiplicative subset  $k[f]^\times =$  the set of all nonzero elements in  $k[f]$ . Of course, the nonsingularity of  $f^\#$  is only at finite distance, i.e., in general it will have singularities at infinity.

In any case,  $I(B_f/k(f))$  is nothing but the set of all branches of  $f^\#$  at infinity. To deal with them we put  $F(X, Y) = f(X^{-1}, Y)$  and  $F^\#(X, Y) = F(X, Y) - u$ . Now

$$F(X, Y) = Y^N + \sum_{1 \leq j \leq N} A_j(X)Y^{N-j} \text{ where } A_j(X) \in k(X) \subset k((X)).$$

The branches of  $f^\#$  at infinity are the branches of  $F^\#$  which in turn are the irreducible factors in  $k(u)((X))[Y]$  written as

$$F^\#(X, Y) = \prod_{1 \leq i \leq m} F_i^\#(X, Y) \text{ with } F_i^\#(X, Y) = Y^{N_i} + \sum_{1 \leq j \leq N_i} A_{ij}^\#(X)Y^{N_i-j}$$

where  $A_{ij}^\#(X) \in k(u)((X))$ . Yes, it is not an accident that this is the same  $m$  as the number of dicritical divisors  $V_1, \dots, V_m$ . Indeed, after suitable labelling, there is a natural isomorphism  $\sigma_i$  of  $V_i$  onto the DVR  $V_i^\dagger$  given by the branch  $F_i^\#$ .

Basically, assuming  $k$  to be an algebraically closed field of characteristic zero, we shall end up finding  $t_i^\dagger$  in an algebraic closure of  $k(u)$  such that  $H(V_i^\dagger) = k(t_i^\dagger)$  and  $u = P_i(t_i^\dagger)$  where  $P_i(Z) \in k[Z]$  is the univariate polynomial of degree  $d_i$  we spoke of in the first paragraph of this Note. Upon letting  $t_i = \sigma^{-1}(t_i^\dagger)$  we would then get  $t_i \in V_i$  such that  $H(V_i) = k(t_i)$  and  $f = P_i(t_i)$ .

To find  $t_i^\dagger$  we use Newton's polygonal method to solve the equation  $F_i^\#(X, Y) = 0$  and thereby expand  $Y$  as a fractional meromorphic series  $\tilde{Y}$  in  $X$ , and also to expand  $X$  as a fractional meromorphic series  $\tilde{X}$  in  $Y$ . Now we use the inversion formula given in [4] to compare these two expansions. Details in Part II.

Philosophy (6.3). The importance of polynomials derives from the fact that they can be viewed as functions in two different ways. To the algebraist, a bivariate polynomial

$$f = f(X, Y) = \sum_{i+j \leq N} a_{ij}X^iY^j \in k[X, Y] \setminus k \text{ with } a_{ij} \in k$$

of (total) degree  $N$  is a function  $\mathbb{N}^2 \rightarrow k$  given by  $(i, j) \mapsto a_{ij}$ . To the analyst, who prefers his field to be the complex number field  $\mathbb{C}$ , it is a map  $\mathbb{C}^2 \rightarrow \mathbb{C}$  given by  $(\alpha, \beta) \mapsto f(\alpha, \beta)$ . Finally, to the geometer, who is an animal linking the analyst with the algebraist, it defines a plane curve  $C : f(X, Y) = 0$ ; if  $k$  is algebraically closed then the points of  $C$  belong to  $k^2$ ; if  $k$  is not algebraically closed then it is better to let the points of  $C$  live in  $\text{spec}(k[X, Y])$ .

Before he proceeds to “compactify”  $\mathbb{C}^2$  and  $\mathbb{C}$ , the analyst thinks of the “fibers” of the map  $\mathbb{C}^2 \rightarrow \mathbb{C}$  above various values  $c$  of  $f$ , and then he may perform catastrophic tortuous surgery, and so on.

In place of this, as algebraists (or algebraic-geometers) we take an indeterminate  $u$  over  $k(X, Y)$  and think of the “generic curve”  $f^\# = 0$  where

$$f^\# = f^\#(X, Y) = f(X, Y) - u \in k(u)[X, Y].$$

By “identifying”  $u$  with  $f$ , i.e., by the shocking (= absurd sounding but surprisingly correct and extremely useful) equation  $f = f$ , we can take  $B_f$  to be the affine coordinate ring of  $f^\#$ . As noted above,  $B_f$  is a PID and hence  $f^\#$  is an irreducible nonsingular affine plane curve. Instead of saying that we can take  $B_f$  to be the affine coordinate ring of  $f^\#$ , let us be more pedantic and set up an isomorphism between the two. Now the affine coordinate ring  $B_f^\#$  of  $f^\#$  is given by

$$H_f : B^\# = k(u)[X, Y] \rightarrow k(u)[X^\#, Y^\#] = B_f^\#$$

where  $H_f$  is a  $k(u)$ -epimorphism which sends  $(X, Y)$  to  $(X^\#, Y^\#)$  and for whose kernel we have

$$\ker(H_f) = f^\# B^\#.$$

Taking indeterminates  $(\bar{X}, \bar{Y})$  over  $k(X, Y)$ , we view  $B_f$  as an affine coordinate ring by considering the  $k(f)$ -epimorphism

$$\bar{H}_f : \bar{B}_f = k(f)[\bar{X}, \bar{Y}] \rightarrow k(f)[X, Y] = B_f$$

which sends  $(\bar{X}, \bar{Y})$  to  $(X, Y)$  and for whose kernel we have

$$\ker(\bar{H}_f) = (f(\bar{X}, \bar{Y}) - f(X, Y))\bar{B}_f.$$

Also we have an obvious  $k$ -isomorphism

$$\hat{H}_f : B^\# = k(u)[X, Y] \rightarrow k(f)[\bar{X}, \bar{Y}] = \bar{B}_f$$

which sends  $(u, X, Y)$  to  $(f, \bar{X}, \bar{Y})$ . Now the said isomorphism

$$\tilde{H}_f (= \text{restriction of } H_f^\#) : B_f \rightarrow B_f^\#$$

is the unique isomorphism such that  $\widetilde{H}_f \overline{H}_f \widehat{H}_f = H_f$ , i.e., such that the obvious rectangle

$$\begin{array}{ccc} B_f = k(f)[X, Y] & \xrightarrow{\widetilde{H}_f (= \text{restriction of } H_f^\#)} & B_f^\# = k(u)[X^\#, Y^\#] \\ \overline{H}_f \uparrow & & H_f \uparrow \\ \overline{B}_f = k(f)[\overline{X}, \overline{Y}] & \xleftarrow{\widehat{H}_f} & B^\# = k(u)[X, Y]. \end{array}$$

commutes. Moreover, the said isomorphism extends to an isomorphism

$$H_f^\# : L = \text{QF}(B_f) = k(X, Y) \rightarrow k(u)(X^\#, Y^\#) = \text{QF}(B_f^\#) = L_f^\#$$

of the function fields.

To distinguish between  $B_f/k(f)$  (resp:  $L/k(f)$ ) and  $B_f^\#/k(u)$  (resp:  $L_f^\#/k(u)$ ) we may call them the affine coordinate ring (resp: function field) of the intrinsic generic curve and the extrinsic generic curve, respectively.

The affine coordinate ring  $B_{f,k}$  of  $f$  is given by the  $k$ -epimorphism

$$H_{f,k} : B = k[X, Y] \rightarrow B_{f,k} = k[x, y] = B_{f,k} \subset k(x, y) = L_{f,k}$$

which sends  $(X, Y)$  to  $(x, y)$  and for whose kernel we have

$$\ker(H_{f,k}) = fB$$

where  $L_{f,k}$  is the total quotient ring of  $B_{f,k}$ , which means the quotient field if  $f$  is irreducible (in  $B$ ).

Assuming  $f$  to be irreducible,  $I(B_{f,k}/k)$  is a nonempty finite subset of  $D(L_{f,k}/k)$  which is a set of DVRs; for every  $V \in D(L_{f,k}/k)$  we put

$$\deg_{f,k}(V) = [H(V) : k] \in \mathbb{N}_+$$

and we call this is the  $(f, k)$ -degree of  $V$ ; for every  $V \in I(B_{f,k}/k)$  we put

$$\text{ind}_{f,k} V = -\min(\text{ord}_V x, \text{ord}_V y) \in \mathbb{N}_+$$

and we call this the  $(f, k)$ -index of  $V$ .

Note that, without assuming  $f$  to be irreducible, for every  $V \in D(L/k(f))$ , upon letting  $V^\# = H_f^\#(V)$ , we have

$$V^\# \in D(L_f^\#/k(u)) \quad \text{with} \quad \deg(V) = \deg_{(f^\#, k(u))} V^\#$$

and if  $V \in I(B_f/k(f))$  then we have

$$V^\# \in I(B_f^\# / k(u)) \quad \text{with} \quad \text{ind}(V) = \text{ind}_{(f^\#, k(u))} V^\#.$$

Remark on infinity (6.4). Continuing the discussion of (6.3), without assuming  $f$  to be irreducible, to take care of points at infinity, we introduce two different incarnations  $\dot{f} = \dot{f}(\dot{X}, \dot{Y})$  and  $\ddot{f} = \ddot{f}(\ddot{X}, \ddot{Y})$  of  $f$  thus.

We write

$$f(X, Y) = \sum_{0 \leq l \leq N} f_l(X, Y) \quad \text{with} \quad f_l(X, Y) = \sum_{i+j=l} a_{ij} X^i Y^j$$

where  $f_l$  is either zero or is homogeneous of degree  $l$ . We call  $f_N = f_N(X, Y)$  the degree form of  $f$  which we denote by  $\text{defo}(f)$  or  $f^+$ . Now we let

$$(\dot{X}, \dot{Y}) = (1/X, Y/X) \quad \text{and} \quad \dot{B} = k[\dot{X}, \dot{Y}]$$

with

$$\dot{f}(\dot{X}, \dot{Y}) = \dot{X}^N f(1/\dot{X}, \dot{Y}/\dot{X}) = \sum_{0 \leq l \leq N} \dot{X}^{N-l} f_l(1, \dot{Y}) \in k[\dot{X}, \dot{Y}]$$

and

$$(\ddot{X}, \ddot{Y}) = (X/Y, 1/Y) \quad \text{and} \quad \ddot{B} = k[\ddot{X}, \ddot{Y}]$$

with

$$\ddot{f}(\ddot{X}, \ddot{Y}) = \dot{Y}^N f(\dot{X}/\dot{Y}, 1/\dot{Y}) = \sum_{0 \leq l \leq N} \dot{Y}^{N-l} f_l(\dot{X}, 1) \in k[\ddot{X}, \ddot{Y}]$$

and we note that  $\dot{f}$  and  $\ddot{f}$  are polynomials of degree  $N$ .

Let  $L_\infty$  consist of  $X$  together with all irreducible homogeneous polynomials in  $k[X, Y] \setminus k$  which are monic in  $Y$ . We call  $L_\infty$  the line at infinity (over  $k$ ). If  $Q \in L_\infty \setminus \{X\}$  is of degree 1 then  $Q = Y - \beta X$  where  $\beta \in k$  and with  $Q$  we associate the triple  $(1, \beta, 0) \in k^3$  by putting  $Q(1, \beta, 0) = Q$ . With  $X$  associate the triple  $(0, 1, 0)$  by putting  $Q(0, 1, 0) = X$ ; note that  $Q(1, 0, 0) = Y$ . Thinking of the usual projective line (over  $k$ ) as consisting of all triples  $(\alpha, \beta, 0) \in k^3$  such that if  $\alpha \neq 0$  then  $\alpha = 1$  and if  $\alpha = 0$  then  $\beta = 1$ , the mapping which sends  $(\alpha, \beta, 0)$  to  $Q(\alpha, \beta, 0)$  gives a bijection of the said line onto the set of degree 1 points of  $L_\infty$ . For any  $Q \in L_\infty$ , we let  $e(f, Q)$  be the largest nonnegative integer such that  $Q^{e(f, Q)}$  divides  $f^+$  in  $B$ ; we call  $e(f, Q)$  the exponent of  $Q$  in  $f$ . Clearly  $\{Q \in L_\infty : e(f, Q) > 0\}$  is a nonempty finite set and labelling its distinct members which are different from  $X$  as  $\{Q_1, \dots, Q_p\}$  and letting  $Q_0 = X$  we have

$$f^+ = \theta \prod_{0 \leq i \leq p} Q_i^{e_i} \quad \text{with} \quad e_i = e(f, Q_i)$$

and hence, as a case of Bézout's theorem, we get the obvious equation

$$\sum_{0 \leq i \leq p} e_i d_i = N \quad \text{with} \quad d_i = \deg(Q_i)$$

which says that  $f$  and  $L_\infty$  meet in  $N$  points counted properly.

Recall that for any finite number of elements  $x_1, \dots, x_r$  in an overfield of  $k$  we have defined

$$\mathfrak{W}(k; x_1, \dots, x_r) = \bigcup_{1 \leq j \leq r \text{ with } x_j \neq 0} \mathfrak{W}(k[x_1/x_j, \dots, x_r/x_j])$$

and for any subset  $J$  of a domain  $S$  let us put

$$\mathfrak{W}(S, J) = \{R \in \mathfrak{W}(S) : JR \neq R\}.$$

Also recall that any  $V \in \overline{D}(L/k)$  dominates a unique member of  $\mathfrak{W}(k; x_1, \dots, x_r)$  which is called the *center* of  $V$  on  $\mathfrak{W}(k; x_1, \dots, x_r)$ .

We define the projective plane and the projective line over  $k$  by putting

$$\mathcal{P}_k^2 = \mathfrak{W}(k; X, Y, 1) \quad \text{with} \quad \mathcal{P}_k^1 = \mathfrak{W}(k; X, 1)$$

and we define the affine plane and the affine line over  $k$  by putting

$$\mathcal{A}_k^2 = \mathfrak{W}(B) \quad \text{with} \quad \mathcal{A}_k^1 = \mathfrak{W}(k[X])$$

and we define the projective point and the affine point over  $k$  by putting

$$\mathcal{P}_k^0 = \mathcal{A}_k^0 = \{k\}$$

and we note that then

$$\mathcal{P}_k^2 = \mathfrak{W}(B) \cup \mathfrak{W}(\dot{B}) \cup \mathfrak{W}(\ddot{B})$$

and by putting

$$\dot{\mathcal{A}}_k^1 = \mathfrak{W}(\dot{B}, (\dot{X} \dot{B})) \quad \text{with} \quad \ddot{\mathcal{A}}_k^0 = \mathfrak{W}(\ddot{B}, (\ddot{X}, \ddot{Y}) \ddot{B})$$

we have the disjoint unions

$$\mathcal{P}_k^2 = \mathcal{A}_k^2 \coprod \dot{\mathcal{A}}_k^1 \coprod \ddot{\mathcal{A}}_k^0 \quad \text{with} \quad \mathcal{P}_k^1 = \mathcal{A}_k^1 \coprod \mathcal{A}_k^0.$$

Informally speaking,  $\ddot{\mathcal{A}}_k^0$  is the set consisting only of the local ring of the origin in the  $(\ddot{X}, \ddot{Y})$ -plane, and so we may identify  $\ddot{\mathcal{A}}_k^0$  with  $\mathcal{A}_k^0$ . Again informally speaking,  $\dot{\mathcal{A}}_k^1$  is the line  $\dot{X} = 0$  in the plane  $\mathfrak{W}(\dot{B})$ ; formally speaking, to identify  $\dot{\mathcal{A}}_k^1$  with the  $X$ -line  $\mathcal{A}_k^1 = \mathfrak{W}(k[X])$ , considering the  $k$ -epimorphism  $\dot{B} \rightarrow k[X]$  given

by  $(\dot{X}, \dot{Y}) \mapsto (0, X)$ , and remembering the commutativity of epimorphism and localization, we note that  $R \mapsto R/(\dot{X}R)$  gives a bijection  $\dot{\mathcal{A}}_k^1 \rightarrow \mathcal{A}_k^1$ . Thus  $\dot{B}$  is the preferred chart to study the line at infinity in  $\mathcal{P}_k^2$ , i.e.,

$$\mathcal{P}_k^2 \setminus \mathcal{A}_k^2 = \dot{\mathcal{A}}_k^1 \coprod \ddot{\mathcal{A}}_k^0.$$

To match this line at infinity with  $L_\infty$ , first we define the local ring  $R(L_\infty)$  of  $L_\infty$  by putting

$$R(L_\infty) = \dot{B}_{\dot{X}\dot{B}}$$

and noting that this is the unique one-dimensional member of  $\dot{\mathcal{A}}_k^1$ ; it can also be characterized as the DVR  $R_\infty$  of  $L/k$  for which

$$\text{ord}_{R_\infty} g = -\text{deg}(g) \text{ for all } g \in B.$$

Next we define the local ring  $R(Q)$  of  $Q \in L_\infty$  by putting

$$R(Q) = \begin{cases} \ddot{B}_{(\dot{X}, \dot{Y})\dot{B}} & \text{if } Q = X \\ \dot{B}_M \text{ where } M = (\dot{X}, Q/X^{\text{deg}(Q)})\dot{B} & \text{if } Q \neq X \end{cases}$$

and we note that  $Q \mapsto R(Q)$  gives bijections  $\{X\} \rightarrow \ddot{\mathcal{A}}_k^0$  and  $L_\infty \setminus \{X\} \rightarrow \dot{\mathcal{A}}_k^1$ . To complete the picture, we define the local ring  $R(Q)$  of any  $Q \in \text{spec}(B)$  by putting

$$R(Q) = B_Q$$

so that  $Q \mapsto R(Q)$  gives a bijection  $\text{spec}(B) \rightarrow \mathcal{A}_k^2$ . Thus,

$$Q \mapsto R(Q) \quad \text{gives a bijection} \quad SP_k^2 \rightarrow \mathcal{P}_k^2$$

where by definition

$$\text{the spectral projective plane } SP_k^2 = \text{spec}(B) \coprod L_\infty \coprod \{L_\infty\}.$$

Moreover, for any  $(\alpha, \beta, 1) \in k^3$  we put

$$Q(\alpha, \beta, 1) = (X - \alpha, Y - \beta)B \in \text{spec}(B)$$

and we note that then

$$(\alpha, \beta, \gamma) \mapsto R(Q(\alpha, \beta, \gamma)) \quad \text{gives a bijection} \quad UP_k^2 \rightarrow RP_k^2$$



where by definition

$$\text{the usual projective plane } UP_k^2 = \begin{cases} \text{the set of all } (\alpha, \beta, \gamma) \in k^3 \\ \text{such that: if } \gamma \neq 0 \text{ then } \gamma = 1, \\ \text{if } \gamma = 0 \neq \alpha \text{ then } \alpha = 1, \\ \text{if } \gamma = 0 = \alpha \text{ then } \beta = 1, \end{cases}$$

and

$$\text{the rational projective plane } RP_k^2 = \begin{cases} \text{the set of rational points of } \mathcal{P}_k^2, \\ \text{i.e., 2-dimensional members of } \mathcal{P}_k^2 \\ \text{which are residually rational over } k. \end{cases}$$

To summarize, we have maps

$$UP_k^2 \xrightarrow{Q} SP_k^2 \xrightarrow{R} \mathcal{P}_k^2 \quad \text{with} \quad \text{im}(QR) = RP_k^2$$

where the first injective map is  $(\alpha, \beta, \gamma) \mapsto Q(\alpha, \beta, \gamma)$  and the second bijective map is  $Q \mapsto R(Q)$ .

Let us observe that  $I(B/k, f) \subset I(B/k) \setminus \{R_\infty\}$ , and moreover the center of any  $V \in I(B/k) \setminus \{R_\infty\}$  on  $\mathcal{P}_k^2$  is the two dimension regular local domain  $R$ , with quotient field  $L$  and  $[H(R) : k] < \infty$ , described thus:

$$(\dagger) R = k[x, y]_J \text{ with } x \in M(R) \setminus M(R)^2 \text{ where}$$

$$(x, y) = (1/X, Y/X) \text{ or } (x, y) = (1/Y, X/Y) \text{ according as } X \notin V \text{ or } x \in V$$

and  $J$  is the maximal ideal in  $k[x, y]$  generated by  $x$  and a nonconstant irreducible monic polynomial  $\zeta(y) \in k[y]$ . Furthermore, if  $V \in I(B/k, f)$  then  $V$  is a dicritical divisor of  $f$  in  $R$  with  $fx^N \in R$  and we have

$$F_N(1, y) \in \zeta(y)k[y] \text{ or } F_N(y, 1) \in \zeta(y)k[y] \text{ according as } X \notin V \text{ or } x \in V.$$

By Lemma (II) of Sect. 5, it follows that if  $V \in I(B/k, f)$  then the relative algebraic closure  $k'$  of  $k$  in  $H(V)$  is a finite algebraic extension of  $k$  and  $H(V)$  is a simple transcendental extension of  $k'$ ; we say that  $f$  is *residually a polynomial* over  $B$  relative to  $V$  to mean that  $f \in V$  and  $H_V(f) \in k'[t] \setminus k'$  for some  $t \in H(V)$  with  $H(V) = k'(t)$ .

Further discussion in Part II.

*Example (6.5).* To indicate the dependence of  $N$  and  $m$  on  $f$ , let us write  $N_f$  and  $m_f$  for them. Then clearly  $m_f$  and  $\deg_f(V_1), \dots, \deg_f(V_m)$  depend only on  $f$  as an element of  $B$  and not on the particular generators  $X, Y$  of  $B$ . This can be paraphrased by letting  $\text{Aut}_k(B)$  be the group of all  $k$ -automorphisms of  $B$  and saying

that for every  $\tau$  in  $\text{Aut}_k(B)$  we have that: (i)  $m_{\tau(f)} = m_f$ ; (ii)  $\tau(V_i)_{1 \leq i \leq m}$  are the dicritical divisors of  $\tau(f)$ ; and (iii)  $\deg_{\tau(f)}(\tau(V_i)) = \deg_f(V_i)$  for  $1 \leq i \leq m$ . Let us call  $f$  a ring generator to mean that  $B = k[f, g]$  for some  $g$  in  $B$ . Then it is clear that  $f$  is a ring generator iff  $N_{\tau(f)} = 1$  for some  $\tau$  in  $\text{Aut}_k(B)$ . Therefore by (6.1)( $\bullet$ ), it follows that:

$$f \text{ is a ring generator} \Leftrightarrow m_f = 1 = \deg_f(V_1) \Rightarrow \text{ind}_f(V_1) = N_f.$$

Now to exhibit the dependence of  $\text{ind}_f(V_i)$  on  $X, Y$ , it suffices to take  $f$  to be the ring generator  $Y - X^N$  with any  $N \in \mathbb{N}_+$  and noting that  $\text{ind}_f(V_1) = N_f = N$  but  $\text{ind}_{\tau(f)}(\tau(V_1)) = N_{\tau(f)} = 1$  where  $\tau$  in  $\text{Aut}_k(B)$  is given by  $(X, Y) \mapsto (X, Y + X^N)$ .

*Note (6.6).* Let  $R$  be a two dimensional regular local domain. Now given any  $z \in \text{QF}(R)^\times$ , by a *dicritical divisor* of  $z$  in  $R$  we mean a prime divisor  $V$  of  $R$  such that  $z$  is residually transcendental over  $R$  relative to  $V$ . By Lemma (II) of Sect. 5, we know that the residue field  $K^* = H(V)$  of any prime divisor  $V$  of  $R$  is of the form  $K^* = K'(t)$  where the finite algebraic field extension  $K'$  of  $K = H_V(R)$  is the relative algebraic closure of  $K$  in  $K^*$  and the element  $t$  is not algebraic over  $K'$ . Assuming  $z \in \text{QF}(R)$  to be residually transcendental over  $R$  relative to  $V$ , after writing

$$H_V(z) = \frac{P(t)}{Q(t)}$$

where  $P(t), Q(t)$  are nonzero members of  $K'[t]$  having no nonconstant common factor in  $K'[t]$ , we define the *relative polar degree*  $\text{rpdeg}_{(V,t)}z$  of  $z$  relative to  $(V, t)$  to be the number of distinct nonconstant irreducible monic factors of  $Q(t)$  in  $K'[t]$ . Note that

$$\max(\deg_t P(t), \deg_t Q(t))$$

is a positive integer which is independent of  $t$  as long as  $K^* = K'(t)$ ; we denote this positive integer by  $\text{resdeg}_{(V,R)}z$  and call it the *residue degree* of  $z$  relative to  $(V, R)$ . We also define the *polar degree*  $\text{pdeg}_V z$  of  $z$  relative to  $V$  to be the minimum of  $\text{rpdeg}_{(V,t)}z$  taken over all  $t \in K^*$  with  $K^* = K'(t)$ . We say that  $z$  is *residually a polynomial* over  $R$  relative to  $V$  to mean that  $\text{pdeg}_V z = 0$ , i.e., to mean that  $H_V(z) \in K'[t] \setminus K'$  for some  $t \in K^*$  with  $K^* = K'(t)$ ; note that for any such  $t$  we have  $\text{resdeg}_{(V,R)}z = \deg_t P(t)$ ; moreover if  $t'$  and  $P'(t')$  are any other such values of  $t$  and  $P(t)$  then  $P'(t') = aP(bt + c)$  for some  $a, b, c$  in  $K'$  with  $a \neq 0 \neq b$ .

( $\dagger^*$ ) As an analogue of (6.1)( $\dagger$ ) we note that any  $z \in \text{QF}(R)^\times$  has at most a finite number of dicritical divisors in  $R$ . Moreover, this number is zero iff either  $z \in R$  or  $1/z \in R$ . [To see this, first observe that if  $z$  has a dicritical divisor in  $R$  then obviously  $z \notin R$  and  $1/z \notin R$ . So henceforth assume that  $z \notin R$  and  $1/z \notin R$ . Now  $R$  is normal because it is regular, and hence by the bracketed proof on pages 75–76 of [3] we find an epimorphism  $h : R[z] \rightarrow H(R)[Z]$  with indeterminate  $Z$  such that  $h(z) = Z$  and  $h(x) = H_R(x)$  for all  $x \in R$ . It follows that  $M(R)R[z]$  is a prime ideal in  $R[z]$  with  $(M(R)R[z]) \cap R = M(R)$ . Let  $S$  be the localization of  $R[z]$  at  $M(R)R[z]$  and let  $T$  be the integral closure of  $S$  in  $\text{QF}(R)$ . By Lemma (T54) on

page 268 of [12] we have  $\dim(S) = 1$  and hence by Theorem (4.10) on page 118 of Nagata [28] we see that

$$T = V_1 \cap \dots \cap V_e$$

where  $e$  is a positive integer and  $V_1, \dots, V_e$  are pairwise distinct DVRs with quotient field  $\text{QF}(R)$ . Clearly  $V_1, \dots, V_e$  are exactly all the dicritical divisors of  $z$  in  $R$ .]

Given any  $F, G$  in  $R^\times$ , by a *dicritical divisor* of  $(F, G)$  in  $R$  we mean a dicritical divisor of  $F/G$  in  $R$ . The above terms relative polar degree  $\text{rpdeg}$ , residue degree  $\text{resdeg}$ , polar degree  $\text{pdeg}$ , and residually a polynomial, are now applicable with  $z$  replaced by  $(F, G)$ .

Geometrically speaking, we may visualize  $R$  to be the local ring of a simple point of an algebraic or arithmetical surface, and think of  $z$  as a *rational function* at that simple point, and  $(F, G)$  as the *pencil* of curves  $F = uG$  at that point. Let us call the pencil *special* to mean that  $G$  equals a unit times a power of a regular parameter, i.e.,  $GR = x^m R$  for some  $x \in M(R) \setminus M(R)^2$  and  $m \in \mathbb{N}$ .

By (6.4)(‡) we see that a bivariate polynomial  $f \in B \setminus k$  gives rise to a special pencil in each relevant  $R$ , and hence the following Local Ring Proposition LRP would imply the following Polynomial Ring Proposition PRP.

LRP says that if  $(F, G)$  is any special pencil in a two dimensional regular local ring  $R$  then  $F/G$  is residually a polynomial over  $R$  relative to any dicritical divisor  $V$  of  $F/G$  in  $R$ .

PRP says that if  $f$  is any nonconstant member of a bivariate polynomial ring  $B = k[X, Y]$  then  $f$  is residually a polynomial over  $B$  relative to any dicritical divisor of  $f$  in  $R$ .

Let  $A$  be a two-dimensional affine domain over an algebraically closed field and let  $R$  be the localization of  $A$  at a maximal ideal. Now (†\*) says that if  $R$  is regular then, for any rational function

$$z = F/G$$

with  $F \neq 0 \neq G$  in  $R$ ,  $z$  has only a finite number of dicritical divisors in  $R$ ; moreover, if the pencil  $(F, G)$  is special then  $z$  is residually a polynomial over  $R$  relative to every dicritical divisor of  $z$  in  $R$ . In view of the results of [2, 8], it can be shown that all except a finite number of prime divisors  $V$  of  $R$  are residually simple transcendental over  $R$ ; moreover, if  $R$  is regular then the said finite number is zero. This is the analogue from the theory of quasirational singularities we spoke of in the preamble of this section. Thus a (possibly singular) point of a surface in the quasirational theory is replaced by a rational function at a simple point of a surface in the dicritical theory.

Needless to say that a simple point in the former theory is replaced by a special pencil in the latter theory. Likewise, residually simple transcendental in the former theory is replaced by residually a polynomial in the latter theory.

As a final philosophical comment, I wish to observe that the LHS  $I(B/k, f)$  of the equation (6.1)(†) represents points at infinity of the projective plane while its RHS  $I(B_f, /k(f))$  represents the branches at infinity of a generic plane curve. Thus the LHS stands for the projective viewpoint while the RHS stands for the

meromorphic viewpoint. Although, in [6, 7, 13–15], I have been beating the drums of the meromorphic viewpoint, it has suddenly dawned on me that the difference between these two methods is merely a matter of semantics!!

More discussion in Part II.

## 7 Field Generators

Consider the bivariate polynomial ring  $k[X, Y]$  over a field  $k$ . A polynomial  $f(X, Y) \in k[X, Y]$  is a field generator means for some  $g = g(X, Y) \in k(X, Y)$  we have  $k(X, Y) = k(f, g)$ ; here the complementary generator  $g$  may or may not be a polynomial. In his 1974 Purdue Ph.D. Thesis [25], Jan gave an example of a field generator which has no complementary polynomial field generator. In Theorem (7.6) I shall give a criterion for the existence of a complementary polynomial field generator. Recently, Pierrette Cassou-Noguès [18, 19] ascribed this criterion to Russell [32, 33], and she used it to revisit Jan's example. However, I shall give a short, almost obvious, proof of (7.6) which is completely independent of the rest of this paper. The criterion (7.6) can be paraphrased by saying that a field generator  $f$  has a complementary polynomial field generator iff  $f$  has a dicritical divisor of degree 1.

Note that if a polynomial  $f$  is a field generator then the generic curve  $f = u$ , where  $u$  is an indeterminate, is a curve of genus zero having a rational place over  $k(u)$ , and conversely. In Example (7.7), I shall discuss the circle to illustrate this fact. It was conjectured by me and proved by my student Jan in his Thesis [25] that a field generator has at most two points at infinity. Without assuming  $f$  to be a field generator, in Part II I shall generalize this by giving a bound on the number of points at infinity of  $f$  in terms of the genus of  $f = u$ .

Preamble for (7.1)–(7.5). Let  $L$  be a finitely generated field extension of a field  $K$  with  $\text{trdeg}_K L = \epsilon$ . Let  $A$  be an affine domain over  $K$  with  $\text{QF}(A) = L$  where  $\text{QF}(A)$  denotes the quotient field of  $A$ . Note that  $D(L/K)$  is the set of all valuation rings  $V$  with  $\text{QF}(V) = L$  and  $K \subset V$  such that  $\text{trdeg}_K H(V) = \epsilon - 1$  where

$$H_V : V \rightarrow H(V) = V/M(V)$$

is the residue class epimorphisms and we are identifying  $H(K)$  with  $K$ ; moreover, every member of  $D(L/K)$  is a DVR, and  $I(A/K)$  is the set of all  $V \in D(L/K)$  with  $A \not\subset V$ .

**Lemma (7.1).** *Assume that  $L = K(x)$  where  $x$  is transcendental over  $K$ . Let  $V$  be the  $(1/x)$ -adic valuation, i.e., let  $V$  be the localization of  $K[x]$  at the prime ideal generated by  $1/x$ . Then  $V \in D(L/K)$  with  $H(V) = K$ .*

*Proof.* Obvious.

**Lemma (7.2).** *Assume that  $L = K(y)$  where  $y$  is transcendental over  $K$ . Let  $V \in D(L/K)$  be such that  $H(V) = K$ . Then  $L = K(x)$  for some  $x \in L$  such that  $V$  is the  $(1/x)$ -adic valuation. Moreover, if  $K$  is infinite and  $V_2, \dots, V_m$  are any finite number of members of  $D(L/K) \setminus \{V\}$ , then  $x$  can be chosen so that we also have  $x \notin M(V_2) \cup \dots \cup M(V_m)$ .*

*Proof.* If  $V$  is the  $(1/y)$ -adic valuation then taking  $z = y$  we see that  $L = K(z)$  and  $V$  is the  $(1/z)$ -adic valuation. If not then  $V$  must be the localization of  $K[y]$  at the prime ideal generated by  $y - a$  for some  $a \in K$ , and taking  $z = 1/(y - a)$  we see that  $L = K(z)$  and  $V$  is the  $(1/z)$ -adic valuation. Now without the “Moreover” it suffices to take  $x = z$ . With the “Moreover” we clearly have  $z \in V_2 \cup \dots \cup V_m$  and, since  $K$  is infinite, for all except a finite number of  $c \in K$  we must have

$$z + c \notin M(V_2) \cup \dots \cup M(V_m)$$

and it suffices to take  $x = z + c$ .

**Lemma (7.3).** *Assume that  $L = K(x)$  where  $x$  is transcendental over  $K$ . Let  $V$  be the  $(1/x)$ -adic valuation and assume that  $V \in I(A/K)$ . Let  $\{V_2, \dots, V_m\}$  be the distinct elements of  $I(A/K) \setminus \{V\}$ , and note that for  $2 \leq i \leq m$  we clearly have  $K[x] \subset V_i$  and  $V_i$  is the localization of  $K[x]$  at the prime ideal generated by an irreducible element  $x_i$  in  $k[x]$ . Now assume that  $A$  is a UFD. Then  $A$  is a proper PID, and  $A$  is the localization of  $K[x]$  at the multiplicative set consisting of all monomials in  $x_2, \dots, x_m$ . Moreover, if  $x \notin M(V_2) \cup \dots \cup M(V_m)$  then clearly  $x$  is an irreducible element in  $A$ .*

*Proof.* To see that  $A$  equals the said localization, note that  $A$  is normal because it is a UFD, and hence  $A$  is the intersection of all the members of  $D(A/K) \setminus I(A/K)$ , but this intersection clearly equals the said localization.

**Lemma (7.4).** *Assume that  $\epsilon = 1$  and  $L = K(x)$  for some  $x \in A$ . Then  $H(V) = K$  for some  $V \in I(A/K)$ .*

*Proof.* Take  $V$  to be the  $(1/x)$ -adic valuation and apply (7.1).

**Lemma (7.5).** *Assume that  $\epsilon = 1$  and  $L = K(y)$  for some  $y \in L$ . Also assume that,  $K$  is infinite,  $A$  is a UFD, and  $H(V) = K$  for some  $V \in I(A/K)$ . Then  $A$  is a proper PID and  $L = K(x)$  for some irreducible  $x \in A$ .*

*Proof.* Take  $\{V_2, \dots, V_m\} = I(A/K) \setminus \{V\}$  and apply (7.2) and (7.3).

Preamble for (7.6). Consider the bivariate polynomial ring  $B = k[X, Y]$  over a field  $k$  and let  $L = k(X, Y) = \text{QF}(B) =$  the quotient field of  $B$ . Given any

$$f = f(X, Y) \in B \setminus k,$$

by  $B_f$  we denote the localization of  $B$  at the multiplicative set  $k[f]^\times$ , and we note that then  $B_f$  is the affine domain  $k(f)[X, Y]$  over the field  $k(f)$  with  $\text{QF}(B_f) = k(X, Y) = L$  and we have  $\text{trdeg}_{k(f)} L = 1$ . Note that a localization of a UFD is a

UFD, and irreducibles in the localization are essentially the same as irreducibles in the original UFD except that the localization has more units. Hence we get:

**Theorem (7.6).** *In the above setup we have the following.*

- (1) If  $L = k(f, g)$  for some  $g \in B$ , then  $H(V) = k(f)$  for some  $V \in I(B_f/k(f))$ .
- (2) If  $L = k(f, l)$  for some  $l \in L$  and  $H(V) = k(f)$  for some  $V \in I(B_f/k(f))$ , then  $L = k(f, g)$  for some  $g \in B$ .

*Proof.* Taking  $(K, L, A) = (k(f), k(X, Y), B_f)$ , (1) follows from (7.4). Likewise (2) follows from (7.5) after noting that the irreducible  $x \in B_f$  when multiplied by a suitable  $b \in k[f]^\times$  produces an irreducible  $bx \in B$  and we obviously have  $k(f, bx) = k(f, x) = L$ .

*Example (7.7).* We illustrate the above theorem by showing that the circle is a field generator over  $\mathbb{C}$  but not over  $\mathbb{R}$ . The underlying obvious fact behind this is that  $f$  is a field generator of  $L = k(X, Y)$  iff the general curve  $f^\# = f(X, Y) - u$ , where  $u$  is an indeterminate, is of genus zero and has a rational place over  $k(u)$ , i.e., a  $V \in D(L_f^\#/k(u))$  which is residually rational over  $k(u)$ ; here  $L_f^\#$  is the function field of  $f^\#$ , i.e., the quotient field of the residue class ring of  $k(u)[X, Y]$  modulo the ideal generated by  $f^\#$ . For the circle  $f = X^2 + Y^2 - 1$  with  $k = \mathbb{R}$ , if  $f^\#$  had a rational place then we can find a nonzero triple  $(a(u), b(u), c(u))$  in  $k[u]$  such that

$$a(u)^2 + b(u)^2 = c(u)^2 + uc(u)^2.$$

Since the equation  $x^2 + y^2 = 0$  has no solution in  $\mathbb{R}$  other than  $(0, 0)$ , it follows that if  $(a(u), b(u)) \neq 0$  then the LHS of the above equation is a nonzero polynomial of even degree. But if  $c(u) \neq 0$  then the RHS of the equation is a nonzero polynomial of odd degree. Therefore, the circle is not a field generator over  $\mathbb{R}$ . Over  $k = \mathbb{C}$  it is a field generator because  $k(f, X + iY) = k(X, Y)$ .

## 8 Preview of Part II

As said in the Introduction, Part II will include various topics from algebraic curve theory such as the conductor and genus formulas of Dedekind and Noether, and the automorphism theorems of Jung and Kulk. In Part II, I shall also relate all this to the Jacobian problem which conjectures that if the Jacobian of  $n$  polynomials in  $n$  variables over a characteristic zero field equals a nonzero constant then the variables can be expressed as polynomials in the given polynomials; see [13–15]. As indicated in the preamble of Sect. 4, in Part II, I shall revisit Newton's polygonal method. As said at the end of Sect. 5, in Part II, I shall say more about the Inversion and Invariance Theorems and about quadratic transformations. As said in Sect. 6, in Part II, I shall discuss Dicritical Divisors some more. Finally, as said in the beginning

of Sect. 7, in Part II, I shall give a bound on the number of points at infinity of an algebraic plane curve.

## 9 Epilogue

Let me close with a chatty survey of the paper which can also serve as an alternative Introduction.

### 9.1 Trigonometry

In high-school we learn the expansion

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots = x \sum_{0 \leq i < \infty} a_i x^i$$

where  $a_i = 0$  or  $\frac{(-1)^{i/2}}{(i+1)!}$  according as  $i$  is odd or even. The fact that in the expansion of  $\sin x$  there is no  $x^2$  term but there is an  $x^3$  term, may be codified by saying that  $\sin x$  has a gap of size 2, i.e., 2 is the smallest positive value of  $i$  for which  $a_i \neq 0$ . Now

$$\sin^{-1} x = x + \frac{x^3}{3!} +$$

and so the inverse function has a gap of the same size 2.

It was around 1665 that Newton gave the above two expansions and Gregory gave the expansion

$$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots$$

and from this it follows that

$$\tan x = x + \frac{x^3}{3} + \dots$$

but the full expansion of  $\tan x$  is rather complicated and was obtained by Bernoulli only in the next century. At any rate the size of the gap in  $\tan x$  as well as  $\tan^{-1} x$  is again 2. All these formulas can be found in Chrystal's Algebra [20] published in 1886 and Hobson's Trigonometry [24] published in 1891. I was lucky to have

studied these two excellent books towards the end of my high-school years at the suggestion of my father. After hundred years, they are still being reprinted and I highly recommend them to all students of mathematics.

Renaming the above type of gap as absolute gap, given any positive integer  $d$ , let us define the  $d$ -gap to be the smallest value of  $i$  which is nondivisible by  $d$  and for which  $a_i \neq 0$ . Then in all the above examples, the value of the  $d$ -gap is 2 for every  $d > 2$ . As an example of a function with 3-gap 7, we can consider the power series

$$x + x^4 + x^7 + x^8 + x^9 + \dots = x(1 + x^3 + x^6 + x^7 + x^8 + \dots).$$

To illustrate yet another type of gap, consider the power series

$$x(1 + x^2 + x^3 + \theta x^5 + \theta^2 x^6 + x^7 + \dots)$$

where  $\theta$  is a transcendental number. This has a transcendentality gap of size 5, i.e., after factoring out  $x$ , the smallest power with transcendental coefficient is  $x^5$ .

Formalizing all this, in (3.5) we were led to the definition of the  $(T, S)$ -gap  $\nu$  of a nonzero meromorphic series

$$y(T) = T^e \sum_{0 \leq i < \infty} A_i T^i \text{ with } \text{ord}_{T_y}(T) = e \text{ and } A_i \in K \text{ with } A_0 \neq 0$$

over a field  $K$ , where  $S$  is a subfield of the meromorphic series field  $K((T))$  and  $\nu = \min\{i \in \mathbb{N} : A_i T^i \notin S\}$ . For the definitions of meromorphic series, ord, field, etc., see pages 25–32 and 67–88 of [9], or pages 1–39 of [12]. In particular see the first paragraph of Sect. 2 for the symbols  $\mathbb{N}, \mathbb{N}_+, \mathbb{Z}$ , and so on.

In the above examples we wrote  $x$  for  $T$ , and let  $e = 1$ . In the  $d$ -gap case we take  $S = K((T^d))$ , and in the transcendentality gap case we take  $S = k((T))$  where  $k$  is an algebraically closed subfield of  $K$ . In the absolute gap case we take  $S$  to be the null ring  $\{0\}$  although technically speaking it is not a subfield.

Assuming  $e = 1$ , let  $z(T) \in K((T))$  be the inverse of  $y(T)$ , i.e.,  $\text{ord}_T z(T) = 1$  with  $y(z(T)) = T$ ; note that if  $y(T) = \sin T$  then  $z(T) = \sin^{-1} T$ , and if  $y(T) = \tan^{-1} T$  then  $z(T) = \tan T$ . In (3.5)(IV)(7) we show that the  $(T, S)$ -gap of  $z(T)$  equals the  $(T, S)$ -gap of  $y(T)$ . We prove this gap invariance by relating the coefficients of  $y(T)$  and  $z(T)$ . Applying the said relating of coefficients to  $\tan^{-1} x$  we can recover the Bernoulli expansion of  $\tan x$ .

Actually, in (3.5) we prove something which is more general than gap invariance. Namely, for any  $z(T) \in K((T))$  with  $\text{ord}_T z(T) = 1$ , without assuming  $y(z(T)) = T$  but considering the composition  $x(T) = y(z(T))$ , by using the multinomial theorem

$$(X_1 + \dots + X_r)^n = \sum \frac{n!}{t_1! \dots t_r!} X_1^{t_1} \dots X_r^{t_r} \text{ with } r \text{ and } n \text{ in } \mathbb{N} \quad (1)$$



where the summation is over all  $t = (t_1, \dots, t_r) \in \mathbb{N}^r$  with  $t_1 + \dots + t_r = n$ , we express the coefficients of  $x$  as polynomials in the coefficients of  $y$  and  $z$ . As a consequence we show that the  $(T, S)$ -gaps  $v, w, \pi$  of  $x, y, z$  satisfy the relations

$$\begin{cases} \pi \geq \min(v, w) \\ v < w \Rightarrow \pi = v \\ w < v \Rightarrow \pi = w. \end{cases} \tag{2}$$

The  $r = 2$  case of (1) is Newton’s Binomial Theorem for positive integer exponents which he obtained around 1665. Soon after he generalized it to fractional exponents which led him to his famous theorem on fractional meromorphic series expansion of algebraic functions. For Newton’s Theorem and the related result called Hensel’s Lemma see pages 89–108 of [9].

In (3.6)(1) and (3.6)(2) we prove some properties of the  $(T, S)$ -gap by using the Binomial Lemma (3.3). It should be stressed that in this usage the full force of (3.3) has to be brought into play including the information about the relationship between the initial coefficients of the various meromorphic series.

### 9.2 Taylor Expansion and Valuations

A power series

$$f(T) = \sum_{0 \leq i < \infty} \alpha_i T^i \in K[[T]] \text{ with } \alpha_i \in K \tag{1}$$

over a field  $K$  is a meromorphic series without negative degree terms, i.e., with  $\text{ord}_T f(T) \geq 0$ . Differentiating both sides  $i$ -times and then putting  $T = 0$  we get

$$\alpha_i = \frac{f^{(i)}(0)}{i!} \tag{2}$$

where  $f^{(i)}(T)$  denotes the  $i$ -th  $T$ -derivative of  $f(T)$ . Formula (1) with the value of  $\alpha_i$  as in Formula (2), is called the Taylor expansion of  $f(T)$ . Sometimes it is called the Maclaurin expansion. Maclaurin and Taylor were disciples of Newton. We can use this to deduce the expansions

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots \quad \text{and} \quad \cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots$$

from the identities

$$\frac{d \sin x}{dx} = \cos x \text{ with } \frac{d \cos x}{dx} = -\sin x \text{ and } \sin 0 = 0 \text{ with } \cos 0 = 1.$$

The definitions of  $\sin x$  and  $\cos x$  give the last two identities while the first two follow from the equation  $\sin^2 x + \cos^2 x = 1$  by implicit differentiation.

For further commentary on Taylor Expansion see pages 104–105 of [9]. There, and on pages 39–43 of [12], you will also find the definition of a (real) discrete valuation of a field  $L$  as a surjective (= onto) map  $W : L \rightarrow \mathbb{Z} \cup \{\infty\}$  such that for all  $u, u'$  in  $L$  we have  $W(uu') = W(u) + W(u')$  and

$$W(u + u') \geq \min(W(u), W(u')) \quad (1)$$

and for any  $u$  in  $L$  we have:  $W(u) = \infty \Leftrightarrow u = 0$ . Replacing  $\mathbb{Z}$  by any ordered abelian group and deleting the adjective “surjective” we get the definition of a (general) valuation. Note that

$$\begin{cases} W(u) < W(u') \Rightarrow W(u + u') = W(u) \\ W(u') < W(u) \Rightarrow W(u + u') = W(u') \end{cases} \quad (2)$$

Writing  $v, w, \pi$  for  $W(u), W(u'), W(u + u')$  and then comparing (1) and (2) with (9.1)(2) we observe an analogy between valuations under sums and gaps under compositions. See pages 65–70 of [9] for the fact that, in case  $G$  is subgroup of  $\mathbb{R}$ , (1) and (2) may be reformulated by saying that sometimes the usual triangle inequality can be replaced by a stronger inequality which requires all triangles to be isosceles.

For any  $W$  we put  $G_W = W(K^\times)$  and  $R_W = \{u \in K : W(u) \geq 0\}$  and call these the value group and the valuation ring of  $W$ . Now  $R_W$  is a ring with the unique maximal ideal  $M(R_W) = \{u \in K : W(u) > 0\}$ . Thus  $R_W$  is a quasilocal ring to which the second paragraph of Sect. 2 is applicable. More generally, by a valuation ring of a field  $L$  we mean the valuation ring of some valuation of  $L$ . Finally, by a valuation ring we mean a valuation ring of some field. It can be shown that a ring  $V$  is a valuation ring iff  $V$  is domain such that:  $x \neq 0 \neq y$  in  $V \Rightarrow$  either  $x/y \in V$  or  $y/x \in V$ .

This would be a good time to read the rest of Sect. 2. An ambitious reader may also gradually look up the material on pages 43–201 of [12].

### 9.3 Discrete Valuation Rings or DVRs

As a supplement to the reading of Sect. 2, let us add some details about DVRs = discrete valuation rings.

We defined a DVR to be a one-dimensional regular local domain. If  $V$  is any DVR then  $u \mapsto \text{ord}_V u$  gives a discrete valuation of the field  $\text{QF}(V)$  whose valuation ring coincide with  $V$ . Conversely, the valuation ring  $R_W$  of any discrete valuation  $W$  of a field  $L$  is a DVR and for all  $u \in L$  we have  $\text{ord}_{R_W} u = W(u)$ . As another characterization of a DVR we note that a domain  $V$  is a DVR iff  $V$  is a PID such that  $V$  has exactly a nonzero prime ideal  $P$  and  $P^0, P^1, P^2, P^3, \dots$  are exactly

all the distinct nonzero ideals in  $V$ . As yet another characterization of a DVR we note that a domain  $V$  is a DVR iff  $V$  is a DD with exactly one nonzero prime ideal, where DD = Dedekind Domain = a normal noetherian domain of dimension at most one. Here noetherian ring means a ring in which every ideal is finitely generated. Normal domain means a domain which is integrally closed in its quotient field, i.e., every element of its quotient field which is integral over it (i.e., satisfies a monic polynomial equation over the domain) over the domain belongs to the domain. We note that the valuation ring of any valuation is normal.

Recall that a multiplicative set in a domain  $E$  is subset  $M$  of  $E^\times$  with  $1 \in M$  such that the product of any two elements in  $M$  belongs to  $M$ , and the localization  $E_M$  of  $E$  at  $M$  is defined by putting  $E_M = \{u/v : u \in E \text{ and } v \in M\}$ ; note that  $E_M$  is a subdomain of  $\text{QF}(E)$ , and if  $E$  is noetherian (resp: UFD) then  $E_M$  is noetherian (resp: UFD). In case  $M = E \setminus P$  for a prime ideal  $P$  in  $E$ , we may write  $E_P$  in place of  $E_{E \setminus P}$ ; note that  $E_P$  is a quasilocal domain with  $M(E_P) = PE_P$ .

A typical example of a DVR  $V$  is provided by taking a UFD  $E$  and letting  $V = E_{pE}$  where  $p$  is a nonzero nonunit irreducible element in  $E$ . For instance, take  $E = \mathbb{Z}$  and let  $p$  = a prime number, or take  $E$  to be the polynomial ring  $K[X_1, \dots, X_n]$  in a finite number of variables over a field  $K$  and  $p = p(X_1, \dots, X_n)$  = a nonconstant irreducible polynomial, or take  $E$  to be the power series ring  $K[[X_1, \dots, X_n]]$  in a finite number of variables over a field  $K$  and  $p = p(X_1, \dots, X_n)$  = a nonzero nonunit irreducible power series.

In the one variable power series case,  $K[[X]]$  is itself a DVR. In the one variable polynomial case of  $E = K[X]$ , for every  $a \in K$ , the localization  $E_a = E_{(X-a)E}$  is a DVR. Moreover,

$$E_\infty = K[1/X]_{(1/X)K[1/X]}$$

is also a DVR; this is the valuation ring of the discrete valuation  $W$  of  $K(X)$  with  $W(X) = -1$  which we call the  $(1/X)$ -adic valuation of  $K(X)$ . If  $K$  is algebraically closed, then  $E_\infty$  together with  $(E_a)_{a \in K}$  are exactly all the distinct DVRs with  $K \subset V$  and  $\text{QF}(V) = K(X)$ . In case  $K$  is not algebraically closed, we have to replace  $(E_a)_{a \in K}$  by  $(E_{pE})$  with  $p$  varying over all nonconstant monic irreducible polynomials in  $X$  over  $K$ .

Let  $V$  be a DVR with quotient field  $L$ , let  $H_V : V \rightarrow H(V)$  be the residue class epimorphism, let  $T$  be a uniformizing parameter of  $V$ , and let  $k$  be a coefficient set of  $V$ . The passage from  $\mathbb{Q}$  to  $\mathbb{R}$  suggests the definition of the completion  $\widehat{V}$  of  $V$  together with the quotient field  $\widehat{L}$  of  $\widehat{V}$  thus. A sequence  $y = (y_i)_{1 \leq i < \infty}$  in  $L$  is Cauchy means for every  $\epsilon \in \mathbb{N}_+$  there exists  $N_\epsilon \in \mathbb{N}_+$  such that for all  $i > N_\epsilon$  and  $j > N_\epsilon$  we have  $\text{ord}_V(y_i - y_j) > \epsilon$ . This is equivalent to the Cauchy sequence  $y' = (y'_i)_{1 \leq i < \infty}$  if for every  $\epsilon \in \mathbb{N}_+$  there exists  $M_\epsilon \in \mathbb{N}_+$  such that for all  $i > M_\epsilon$  we have  $\text{ord}_V(y_i - y'_i) > \epsilon$ . Now  $\widehat{L}$  may be defined to be the set of all equivalence classes of Cauchy sequences. Moreover  $\widehat{V}$  may be defined to be the set of those members of  $\widehat{L}$  which contain a Cauchy sequence consisting of elements of  $V$ . Sums and products in  $\widehat{L}$  in an obvious manner. This makes  $\widehat{L}$  an overfield of  $L$  and  $\widehat{V}$  an overdomain of  $V$  in such a manner that  $\widehat{L}$  is the quotient field of  $\widehat{V}$ . Now  $\widehat{V}$  is a DVR and for all  $x \in L$  we have  $\text{ord}_{\widehat{V}}x = \text{ord}_Vx$ . Given a sequence  $z_1, z_2, \dots$

and an element  $z$  in  $\widehat{L}$  we say that  $z_i$  tend to  $z$ , in symbols  $z_i \rightarrow z$ , to mean that  $\text{ord}_{\widehat{V}}(z - z_i) \rightarrow \infty$ , and we put  $\sum_{1 \leq i < \infty} z_i = z$  to mean that  $\sum_{1 \leq j \leq i} z_j \rightarrow z$ . Taking any uniformizing parameter  $T$  and coefficient set  $k$  of  $\widehat{V}$ , by mimicking the idea of Taylor expansion, we can show that any  $z \in \widehat{L}^\times$  with  $\text{ord}_{\widehat{V}} z = e$  can uniquely be expressed as

$$z = \sum_{e \leq i < \infty} a_i T^i$$

where  $a_i \in k$  with  $a_e \neq 0$ ; we may call this the Taylor expansion of  $z$  in  $k((T))$ ; we can extend the sum to the left of  $e$  by putting  $a_i = 0$  for all  $i < e$ ; if  $z = 0$  then we can take  $a_i = 0$  for all  $i \in \mathbb{Z}$ . If  $k$  is a coefficient field then  $k((T))$  is the usual power series ring.

Let us sketch a proof of the observation made in Sect. 2 to the effect that if  $A$  is an affine domain over a field  $K$  such that the transcendence degree of the quotient field  $L$  of  $A$  over  $K$  is 1, then  $I(A/K)$  is a nonempty finite set where  $I(A/K)$  is defined to be the set of all DVRs  $V$  with  $\text{QF}(V) = L$  such that  $A \not\subset V$ . For any  $x \in A$ , let  $J(x)$  be the set of all DVRs  $V$  with  $\text{QF}(V) = L$  such that  $x \notin V$ . If  $x$  is algebraic over  $K$  then clearly  $J(x)$  is empty. If  $x$  is transcendental over  $K$  then  $J(x)$  is a nonempty finite set because now  $L/K(x)$  is a finite algebraic field extension and the members of  $J(x)$  are the valuation rings of the extensions to  $L$  of the  $(1/x)$ -adic valuation of  $K(x)$ . We can write  $A = K[x_1, \dots, x_n]$  where  $x_1, \dots, x_n$  is a finite set of elements in  $A$  at least one of which is transcendental over  $K$ . It only remains to note that  $I(A/K) = \cup_{1 \leq i \leq n} J(x_i)$ . Geometrically speaking,  $A$  represents the affine coordinate ring of a curve  $C$  in  $\mathbb{A}_K^n =$  the affine  $n$ -space over  $K$ , and  $I(A/K)$  represents the set of branches of  $C$  at infinity. Recall that

$$I(A/K) \subset D(L/K) = \begin{cases} \text{the set of all DVRs } V \\ \text{with } K \subset V \text{ and } \text{QF}(V) = L. \end{cases}$$

$D(L/K)$  represents the set of all branches of  $C$ , and  $D(L/K) \setminus I(A/K)$  represents the set of all branches of  $C$  at finite distance.

To talk more about the branches of  $C$  in case  $n = 2$  and  $K$  is algebraically closed, let  $f(X, Y)$  be the bivariate irreducible polynomial in  $K[X, Y]$  such that  $f(x, y) = 0$  where  $(x, y) = (x_1, x_2)$ . Note that  $f(X, Y)$  is unique up to multiplication by a nonzero element of  $K$ , and  $f(X, Y) = 0$  is an affine equation  $C$ . To use homogeneous coordinates, let  $F(X, Y, Z) = Z^d f(X/Z, Y/Z)$  where  $d$  is the degree of  $f$ . Now a point of  $C$  at finite distance is of the form  $(a, b, 1)$  where  $a, b$  in  $K$  with  $f(a, b) = 0$ , and at infinity it is either of the form  $(a, 1, 0)$  where  $a \in K$  with  $F(a, 1, 0) = 0$  or of the form  $(1, 0, 0)$  with  $F(1, 0, 0) = 0$ . Let  $I_y$  be the set of all those members  $V$  of  $I(A/K)$  for which  $\text{ord}_V y \leq \text{ord}_V x$  and let  $I_x$  be the set of all the remaining members of  $I(A/K)$ . We define the center of any  $V \in D(L/K)$  on  $C$  thus: if  $V \notin I(A/K)$  then it is the point  $(a, b, 1)$  of  $C$  such that  $\text{ord}_V(x - a) > 0 < \text{ord}_V(y - b)$ ; if  $V \in I_y$  then it is the point  $(a, 1, 0)$  of  $C$  such that  $\text{ord}_V((x/y) - a) > 0$ ; if  $V \in I_x$  then it is the point  $(1, 0, 0)$  of  $C$ . It can be shown that every point of  $C$  is the center of at least one and at most a finite number

of branches of  $C$ . For  $V \in D(L/K) \setminus I(A/K)$  and its center  $(a, b, 1)$  on  $C$ , taking a uniformizing parameter  $T$  of  $\widehat{V}$ , we get the Taylor expansions

$$x = z_1(T) \in K[[T]] \quad \text{and} \quad y = z_0(T) \in K[[T]]$$

with  $z_1(0) = a$  and  $z_0(0) = b$ . We call this a parametrization of  $C$  at the point  $(a, b, 1)$ . It elucidates the material in the short paragraph of (3.1) just before the definition of  $(V, K)$ -presequence.

### 9.4 Newton Expansion and Hamburger-Noether Expansion

Having elucidated a part of (3.1), let us elucidate parts of (3.2) and (3.7). So consider

$$x = z_1(T) \in K((T)) \quad \text{and} \quad y = z_0(T) \in K((T))$$

with

$$\text{ord}_T z_1(T) = \epsilon \in \mathbb{Z}^\times \quad \text{and} \quad \text{ord}_T z_0(T) = e \in \mathbb{Z}^\times$$

where  $K$  is an algebraically closed field of characteristic zero. Following Newton, we can expand  $y$  in terms of  $x$  by first taking an  $\epsilon$ -th root  $\delta(T)$  of  $x$ , i.e.,

$$\delta(T) \in K((T)) \text{ with } \delta(T)^\epsilon = z_1(T)$$

and then rewriting  $y$  in terms of it as

$$y = \eta(T) \in K((T)) \quad \text{with} \quad \eta(\delta(T)) = z_0(T)$$

Let  $J$  be the  $T$ -support of  $\eta(T)$ . The charseq (= characteristic sequence)  $m(J, \epsilon)$  is, roughly speaking, a record of the members of  $J$  where the GCD with  $\epsilon$  drops. This is introduced in (3.2) and studied in (3.7). Here the main tool is the concept of  $d$ -gap mentioned in (9.1).

We call  $\eta(T)$  the Newton expansion of  $z_0$  in terms of  $z_1$ . In (3.1) we replicate this without taking roots, and call it the  $(V, K)$ -preexpansion which we develop further in (3.8), (3.9), and (4.1) where it culminates into the Valuation Theoretic expansion, i.e., the  $(V, K)$ -expansion; here  $V$  is a certain DVR. The avoidance of roots motivates items (6)–(8) of (3.1).

The Valuation Theoretic expansion is a generalized version of the so called Hamburger-Noether expansion. The Mixed Valuation Theoretic expansion, i.e., the  $(V, K, T)$ -expansion of (4.1) is a mixture of the Newton expansion and the Valuation Theoretic expansion.

Let us now further describe the organization of these numerous expansions.

In (3.1) we introduce the  $(V, K)$ -protoexpansion as a simple sequence, and the  $(V, K)$ -preexpansion as a double sequence consisting of several sequences each of

which is a  $(V, K)$ -protoexpansion. In (3.1) we reorganize the  $(V, K)$ -preexpansion as a simple sequence which we call the  $(V, K)$ -expansion. This reorganization is something like reorganizing an  $m$  by  $n$  matrix  $(a_{ij})$  as the simple sequence

$$a_{11}, \dots, a_{1m}, a_{21}, \dots, a_{2n}, \dots, a_{m1}, \dots, a_{mn}$$

of length  $mn$ . Actually, the rows of the  $(V, K)$ -preexpansion may have different lengths. Namely, the  $i$ -th row looks like  $z_{i0}, \dots, z_{i,l(i)+1}$  and has length  $l(i) + 2$ . We chop off its first term and then the first two terms of the chopped off version coincide with the last two terms of the previous row, i.e.,  $(z_{i-1,l(i-1)}, z_{i-1,l(i-1)+1}) = (z_{i1}, z_{i2})$ , and so we glue the two rows at the coincidental terms. Doing this for all except the first row, the  $(V, K)$ -preexpansion converts into a single sequence which we call the  $(V, K)$ -expansion.

In (3.8), (3.9), and (4.1), we inject some newtonian expansions into the  $(V, K)$ -protoexpansion, the  $(V, K)$ -preexpansion, and the  $(V, K)$ -expansion, and then we call the resulting object the mixed  $(V, K, T)$ -protoexpansion, the mixed  $(V, K, T)$ -preexpansion, and the mixed  $(V, K, T)$ -expansion, respectively.

### 9.5 Taylor Series with Remainder

The Taylor formula (9.2)(1) may be truncated at some value of  $i$ , say  $i = j$ , and then the last term  $\alpha_j$  need not equal  $\frac{f^{(j)}(0)}{j!}$ . The resulting formula is called Taylor series with remainder. This is illustrated by the crucial formula (3.1)(8) which explains the avoidance of roots mentioned in (9.4). Note that in (3.1), the quantity  $p_l$  is not defined until items (6)–(8), and in case of  $z_{l+1} \neq 0$ , the summation in (8) terminates at  $v = p_l(e_l/|e_l|)$ , i.e., (8) is reduced to the equation

$$z_{l-1} = \left( \sum_{(e_{l-1}/|e_l|) \leq v \leq p_l(e_l/|e_l|)} A_l^*(v) z_l^{v(e_l/|e_l|)} \right) + z_l^* \quad \text{with} \quad z_l^* = z_l^{p_l} z_{l+1}.$$

Also note that in (3.1) we have  $e_j > 0$  and  $p_j > 0$  for all  $j > 1$  and hence, in case of  $l \neq 1$ , items (6)–(8) become more transparent by putting  $|e_l| = e_l$ . Finally note that formula (4.1)(4<sup>‡</sup>) is another incarnation of (3.1)(8).

To illustrate (3.1)(8) by an example, consider the DVR  $V = K[[T]]$  having uniformizing parameter  $T$  with coefficient field  $K$ , and let

$$z_l = T^3 \text{ and } z_{l-1} = T^6 + T^{9+u} \text{ with } 0 \leq u < 3.$$

Then

$$z_{l-1} = \begin{cases} z_l^2 + z_l^3 + z_l^* & \text{with } z_l^* = z_{l+1} = 0 \text{ \& } p_l = \infty & \text{if } u = 0 \\ z_l^2 + z_l^* & \text{with } z_l^* = z_l^3 z_{l+1} \text{ \& } z_{l+1} = T^u \text{ \& } (p_l, e_{l+1}) = (3, 2) & \text{if } u \neq 0. \end{cases}$$

Let us now further comment on the formation of the mixed  $(V, K, T)$ -expansion we talked about in (9.4) above. In (3.8) we consider the sequence

$$(z_0, z_1, \dots, z_l, z_{l+1}, z_l)$$

of meromorphic series in  $K((T))$ , and we expand each term of the sequence relative to the next term in the newtonian manner, i.e., as a  $(V, K, T)$ -expansion. For the last two pairs, this is possible only if  $z_{l+1} \neq 0$ . The flipping of  $z_{l+1}$  and  $z_l$  in the end of the sequence is meant for connecting it smoothly to the next sequence of the presequence as achieved in (3.9). Think of two wagons of a railway train being connected at the smooth round buffers. Thus in (3.8), we are constructing a perfect wagon which in (3.9) gets joined to other wagon to form a whole train. In (4.1), the whole train is thought of as a single very long wagon which is called the mixed  $(V, K, T)$ -expansion.

### 9.6 Polynomials and Power Series

The field  $K(X_1, \dots, X_n)$  of rational functions over a field  $K$  does not determine the polynomial ring  $K[X_1, \dots, X_n]$  as can be seen by noting that clearly we have  $K[1/X_1, \dots, 1/X_n] \neq K[X_1, \dots, X_n]$  but  $K(1/X_1, \dots, 1/X_n) = K(X_1, \dots, X_n)$ . However, the quotient field  $K((X_1, \dots, X_n))$  of the power series ring  $K[[X_1, \dots, X_n]]$  does determine the said ring. See (3.10).

**Acknowledgments** On the algebraic side, my thanks are to Pierrette Cassou-Noguès, Bill Heinzer, Giulio Peruginelli, Avinash Sathaye, and Dave Shannon for numerous useful discussions. On the topological side, my thanks are to Dung Trang Le, Walter Neumann, Stepan Orevkov, and Claude Weber for many stimulating discussions. But above all, several private lectures which were given to me by Enrique Artal-Bartolo and Arnaud Bodin in Lille in July 2008 have been most helpful for clarifying the theory of dicritical divisors.

### References

- [1] S. K. Abhyankar, *Intermediate Algebra*, First Edition, Indian Press, Allahabad, 1943; Reprinted in 1960 by Karnatak Printing Press, Bombay
- [2] S. S. Abhyankar, *On the valuations centered in a local domain*, American Journal of Mathematics, 78 (1956), 321–348
- [3] S. S. Abhyankar, *Ramification Theoretic Methods in Algebraic Geometry*, Princeton University Press, Princeton, 1959
- [4] S. S. Abhyankar, *Inversion and invariance of characteristic pairs*, American Journal of Mathematics, 89 (1967), 363–372
- [5] S. S. Abhyankar, *Historical ramblings in algebraic geometry and related algebra*, American Mathematical Monthly, 83 (1976), 409–448
- [6] S. S. Abhyankar, *Expansion Techniques in Algebraic Geometry*, Tata Institute of Fundamental Research, Bombay, 1977

- [7] S. S. Abhyankar, *On the semigroup of a meromorphic curve, Part I*, Proceedings of the International Symposium on Algebraic Geometry, Kyoto (1977), pp. 240–414
- [8] S. S. Abhyankar, *Quasirational singularities*, American Journal of Mathematics, 101 (1979), 276–300
- [9] S. S. Abhyankar, *Algebraic Geometry for Scientists and Engineers*, American Mathematical Society, Providence, 1990
- [10] S. S. Abhyankar, *Some remarks on the Jacobian question*, Purdue Lecture Notes, pp. 1–20 (1971); Published in the Proceedings of the Indian Academy of Sciences, 104 (1994), 515–542
- [11] S. S. Abhyankar, *Resolution of Singularities of Embedded Algebraic Surfaces*, First Edition of 1966 Published by Academic, New York, Second Enlarges Edition of 1998 Published by Springer, Berlin
- [12] S. S. Abhyankar, *Lectures on Algebra I*, World Scientific, Singapore, 2006
- [13] S. S. Abhyankar, *Some thoughts on the Jacobian conjecture, Part I*, Journal of Algebra, 319 (2008), 493–548
- [14] S. S. Abhyankar, *Some thoughts on the Jacobian conjecture, Part II*, Journal of Algebra, 319 (2008), 1154–1248
- [15] S. S. Abhyankar, *Some thoughts on the Jacobian conjecture, Part III*, Journal of Algebra, 319 (2008), 2720–2826
- [16] E. Artal-Bartolo, *Une démonstration géométrique du théorème d'Abhyankar-Moh*, Crelle Journal, 464 (1995), 97–108
- [17] E. T. Bell, *Men of Mathematics*, Simon & Schuster, New York, 1937
- [18] P. Cassou-Noguès, *The effect of rational maps on polynomial maps*, Annales Polonici Mathematici, 76 (2001), 21–31
- [19] P. Cassou-Noguès, *Bad field generators*, Contemporary Mathematics, 369 (2005), 77–83
- [20] G. Chrystal, *Textbook of Algebra I and II*, Cambridge University Press, Cambridge, 1886
- [21] D. Eisenbud and W. D. Neumann, *Three-dimensional link theory and invariants of plane curve singularities*, Annals of Mathematics Studies, 101 (1985)
- [22] L. Fourier, *Topologie d'un polynome de deux variables complexes au voisinage de l'infini*, Annales de l'institut Fourier, 46 (1996), 645–687
- [23] G. Halphen, *Études sur les points singuliers des courbes algébriques planes*, Appendix to the French Edition of Salmon's Higher Plane Curves, 1894, pp. 535–648
- [24] E. W. Hobson, *Plane Trigonometry*, Cambridge University Press, Cambridge, 1891
- [25] C. J. Jan, *On polynomial generators of  $k(x, y)$* , Ph.D. Thesis, Purdue University, 1974
- [26] D. T. Le and C. Weber *A geometrical approach to the Jacobian conjecture for  $n = 2$* , Kodai Journal of Mathematics, 17 (1994), 375–381
- [27] J. F. Mattei and R. Moussu, *Holonomie et intégrales premières*, Annales scientifiques de l'école Normale Supérieure, 13 (1980), 469–523
- [28] M. Nagata, *Local Rings*, Wiley, New York, 1962
- [29] W. D. Neumann, *Complex algebraic plane curves via their links at infinity*, Inventiones Mathematicae, 98 (1989), 445–489
- [30] W. D. Neumann and P. Norbury, *Rational polynomials of simple type*, Pacific Journal of Mathematics, 204 (2002), 177–207
- [31] L. Rudolph, *Embeddings of the line in the plane*, Crelle Journal, 337 (1982), 113–118
- [32] P. Russell, *Field generators in two variables*, Journal of Mathematics of Kyoto University, 15 (1975), 555–571
- [33] P. Russell, *Good and bad field generators*, Journal of Mathematics of Kyoto University, 17 (1977), 319–331
- [34] H. J. S. Smith, *On the higher singularities of plane curves*, Proceedings of the London Mathematical Society, 6 (1873), 153–182
- [35] O. Zariski, *Algebraic Surfaces*, Springer, Berlin, 1934



# Partitions with Non-Repeating Odd Parts and Q-Hypergeometric Identities

Krishnaswami Alladi\*

*Dedicated to the memory of my father Professor Alladi Ramakrishnan*

**Summary** We obtain a series expansion for the product generating function of partitions in which the odd parts do not repeat. This is done by studying the 2-modular Ferrers graphs of such partitions via Durfee squares. This provides a unified approach to several fundamental identities in the theory of partitions and q-series such as those of Sylvester, Lebesgue, Gauss, and Rogers-Fine, and provides links with Göllnitz's deep theorem.

**Mathematics Subject Classification (2000)** 05A17, 05A15, 05A19

**Key words and phrases** Partitions · Non-repeating odd parts · 2-modular Ferrers graphs · q-series

## 1 Introduction

One of the most fundamental identities in the theory of partitions and q-series is

$$\sum_{n=0}^{\infty} \frac{q^{n(n+1)/2} (1+bq)(1+bq^2)\dots(1+bq^n)}{(1-q)(1-q^2)\dots(1-q^n)} = \prod_{m=1}^{\infty} \frac{(1+bq^{2m})}{(1-q^{2m-1})} \quad (1.1)$$

due to Lebesgue (see Andrews [9], Chap. 2). The right hand side of (1.1) is the generating function of partitions in which the even parts do not repeat, where the power of  $b$  keeps track of the number of even parts. The case  $b = -1$  yields the famous Gauss identity

---

\* Research supported in part by NSA grants MSPF-06G-150 and MSPF-08G-154

K. Alladi

Department of Mathematics, University of Florida, Gainesville, FL 32611, USA

e-mail: [alladik@ufl.edu](mailto:alladik@ufl.edu)

$$\prod_{m=1}^{\infty} \frac{(1 - q^{2m})}{(1 - q^{2m-1})} = \sum_{n=0}^{\infty} q^{n(n+1)/2} \quad (1.2)$$

which can be compared to Euler's celebrated Pentagonal Numbers Theorem.

In view of the many implications of Lebesgue's identity including (1.2), partitions with non-repeating even parts have evinced considerable interest. In contrast, partitions with non-repeating odd parts have not attracted much attention, one reason being that no series expansion similar to (1.1) is known for their product generating function

$$\prod_{m=1}^{\infty} \frac{(1 + q^{2m-1})}{(1 - q^{2m})}. \quad (1.3)$$

Our purpose is to derive identity (2.10) below, which provides such a series expansion for a two parameter refinement of the product in (1.3). We achieve this using 2-modular Ferrers graphs and their Durfee square classification (see Sect. 2). As a consequence, the classical identities of Sylvester, Lebesgue, and Rogers-Fine, fall out as special cases – see Sects. 3–6. Thus our approach shows that partitions with non-repeating odd parts are worthy of a closer study. Recently there have been some investigations of partitions with non-repeating odd parts: (a) Berkovich and Garvan [10] considered these partitions and others in the course of obtaining extensions of Dyson's rank statistic for partitions; (b) Hirschhorn and Sellers [12] have established congruences modulo powers of 3 satisfied by these partitions. But our approach and results are different from those in [10] and [12].

## 2 The Series Expansion

The 2-modular Ferrers graph of a partition is one in which each part is represented by a left justified row of twos, with a one at the end on the right if a part is odd. Thus, the parts of the partition correspond to the row sums of the entries at each node.

Partitions with non-repeating odd parts are especially convenient to study using 2-modular graphs because the ones will occur only in the corners of the graph. Therefore, conjugation would also yield a partition with non-repeating odd parts, and the number of odd parts would remain invariant under conjugation.

Our goal is to study partitions into non-repeating odd parts by keeping track of the number of odd parts and the number of even parts. That is, we wish to obtain an expansion for the product generating function

$$\prod_{m=1}^{\infty} \frac{(1 + bq^{2m-1})}{(1 - cq^{2m})}, \quad (2.1)$$

where the powers of  $b$  and  $c$  keep track of the number of odd and even parts respectively. In practice, it turns to be more convenient to keep track of the total number

of parts by a parameter  $z$ , and the number of (distinct) odd parts by a parameter  $b$ , thereby leading to the product generating function

$$\prod_{m=1}^{\infty} \frac{(1 + z b q^{2m-1})}{(1 - z q^{2m})}. \tag{2.2}$$

This is equivalent to (2.1) because the number of even parts is the total number of parts minus the number of odd parts. That is, the series expansion for (2.1) can be obtained from the series for (2.2) by the substitutions  $z \mapsto c$  and  $b \mapsto bc^{-1}$ .

We will use standard notation

$$(a)_n = (a; q)_n = \prod_{j=0}^{n-1} (1 - a q^j), \tag{2.3}$$

for any complex number  $a$  and base  $q$ . Also

$$(a)_{\infty} = \lim_{n \rightarrow \infty} (a)_n = \prod_{j=0}^{n-1} (1 - a q^j), \tag{2.4}$$

when  $|q| < 1$ . As in (2.3) and (2.4), when the base is  $q$ , we might write  $(a)_n$  and  $(a)_{\infty}$  without displaying  $q$ , but when the base is other than  $q$ , it will be displayed.

We will study partitions with non-repeating odd parts by considering the Durfee squares (= the largest square of nodes starting from the top left hand corner of the graph) in the 2-modular graphs. In any such graph  $\pi$ , there is a set of nodes to the right of the Durfee square which we denote by  $\pi_r$ , and a set of nodes below the Durfee square which we denote by  $\pi_b$ . Since we do not distinguish between a graph and the partition it represents, we will refer to these components also as partitions  $\pi_r$  and  $\pi_b$ .

With regard to Durfee squares of 2-modular graphs, there are two cases to consider:

Case 1: The bottom right hand node has a 2

Case 2: The bottom right hand node has a 1

Generating function of Case 1: Consider partitions with non-repeating odd parts whose 2-modular Ferrers graphs have a  $k \times k$  Durfee square  $D$ . The sum of the entries in the nodes is  $2k^2$  and so we have the term

$$z^k q^{2k^2} \tag{2.5}$$

as the generating of  $D$ . The generating function of the partition  $\pi_b$  is

$$\frac{(-z b q; q^2)_k}{(z q^2; q^2)_k}. \tag{2.6}$$

In computing the generating function of  $\pi_r$ , we do not need to keep track of the number of parts. Thus the parameter  $z$  will be absent in this generating function.

Since we are only after the number of odd parts in  $\pi_r$ , all of which are non-repeating, we consider the conjugate partition  $\pi_r^*$  to get the generating function which is

$$\frac{(-bq; q^2)_k}{(q^2; q^2)_k}. \tag{2.7}$$

Thus, the generating function for partitions in Case 1 with a  $k \times k$  Durfee square is the product of the expressions in (2.5), (2.6), and (2.7):

$$z^k q^{2k^2} \cdot \frac{(-zbq; q^2)_k}{(zq^2; q^2)_k} \cdot \frac{(-bq; q^2)_k}{(q^2; q^2)_k}. \tag{2.8}$$

Generating function of Case 2: Here again we consider partitions whose 2-modular graph have a  $k \times k$  Durfee square. In this case, the sum of the nodes inside the Durfee square is  $2k^2 - 1$ . The analysis of the partitions  $\pi_b$  and  $\pi_r$  is as above with the only difference being that the largest parts of  $\pi_b$  and  $\pi_r^*$  are  $\leq 2k - 2$ . Thus, the generating function for such partitions in Case 2 would be

$$bz^k q^{2k^2-1} \cdot \frac{(-zbq; q^2)_{k-1}}{(zq^2; q^2)_{k-1}} \cdot \frac{(-bq; q^2)_{k-1}}{(q^2; q^2)_{k-1}}. \tag{2.9}$$

The sum of the generating functions of Cases 1 and 2 is

$$\begin{aligned} & \frac{z^k q^{2k^2-1} (-zbq; q^2)_{k-1} (-bq; q^2)_{k-1}}{(zq^2; q^2)_k (q^2; q^2)_k} \left\{ q(1 + zbq^{2k-1})(1 + bq^{2k-1}) \right. \\ & \qquad \qquad \qquad \left. + b(1 - zq^{2k})(1 - q^{2k}) \right\}. \\ & = \frac{z^k q^{2k^2-1} (-zbq; q^2)_{k-1} (-bq; q^2)_{k-1}}{(zq^2; q^2)_k (q^2; q^2)_k} \cdot (b + q)(1 + zbq^{4k-1}). \end{aligned}$$

The desired series expansion for the product in (2.2) is obtained by summing the above expression over  $k$  and adding one, namely,

$$\begin{aligned} & 1 + \sum_{k=1}^{\infty} \frac{z^k q^{2k^2-1} (-zbq; q^2)_{k-1} (-bq; q^2)_{k-1} (b + q)(1 + zbq^{4k-1})}{(zq^2; q^2)_k (q^2; q^2)_k} \\ & = \frac{(-zbq; q^2)_{\infty}}{(zq^2; q^2)_{\infty}}. \end{aligned} \tag{2.10}$$

From (2.10), the corresponding series expansion for the product in (2.1) is obtained by means of the substitutions  $z \mapsto c$  and  $b \mapsto bc^{-1}$ :

$$\begin{aligned} & 1 + \sum_{k=1}^{\infty} \frac{c^k q^{2k^2-1} (-bq; q^2)_{k-1} (-bc^{-1}q; q^2)_{k-1} (bc^{-1} + q)(1 + bq^{4k-1})}{(cq^2; q^2)_k (q^2; q^2)_k} \\ & = \frac{(-bq; q^2)_{\infty}}{(cq^2; q^2)_{\infty}}. \end{aligned} \tag{2.11}$$

The left hand side of (2.10) can be expressed in a more elegant form as a sum starting at  $k = 0$ :

$$\sum_{k=0}^{\infty} \frac{z^k q^{2k^2} (-zbq; q^2)_k (-bq^{-1}; q^2)_k (1 + zbq^{4k-1})}{(zq^2; q^2)_k (q^2; q^2)_k (1 + zbq^{2k-1})} = \frac{(-zbq; q^2)_{\infty}}{(zq^2; q^2)_{\infty}}. \tag{2.12}$$

The advantage in our identity (2.10) over Lebesgue’s identity (1.1) is that in (2.10), there are two free parameters  $z$  and  $b$ , thereby providing more flexibility in choosing specializations. In what follows, we will use these identities to derive several fundamental identities in the theory of partitions and q-series.

### 3 Sylvester’s Identity

By analyzing partitions into distinct parts via Durfee squares, Sylvester [14] showed combinatorially that

$$(-bq)_{\infty} = 1 + \sum_{k=1}^{\infty} \frac{b^k q^{(3k^2-k)/2} (-bq)_{k-1} (1 + bq^{2k})}{(q)_k}. \tag{3.1}$$

The case  $b = -1$  (3.1) is Euler’s celebrated Pentagonal Numbers Theorem:

$$(q)_{\infty} = \sum_{k=-\infty}^{\infty} (-1)^k q^{(3k^2-k)/2}.$$

In (2.11), note that

$$c^k (-bc^{-1}q; q^2)_{k-1} (bc^{-1} + q) = (c + bq)(c + bq^3) \dots (c + bq^{2k-3})(b + cq). \tag{3.2}$$

So by letting  $c \rightarrow 0$ , the expression in (3.2) becomes

$$b^k q^{(k-1)^2},$$

yielding

$$(-bq; q^2)_{\infty} = 1 + \sum_{k=1}^{\infty} \frac{b^k q^{3k^2-2k} (-bq; q^2)_{k-1} (1 + bq^{4k-1})}{(q^2; q^2)_k}. \tag{3.3}$$

Sylvester’s identity (3.1) follows from (3.3) by the substitutions  $b \mapsto bq, q^2 \mapsto q$ , in that order. For partition interpretations of (3.1) and (3.3), see Sect. 7.

### 4 Lebesgue’s Identity

In (2.10), replace  $z \mapsto q^{-1}$ ,  $b \mapsto bq^2$  to get

$$\begin{aligned} \frac{(-bq^2; q^2)_\infty}{(q^2; q^2)_\infty} &= 1 + \sum_{k=1}^\infty \frac{q^{2k^2-k-1}(-bq^2; q^2)_{k-1}(-bq^3; q^2)_{k-1}(bq^2 + q)(1 + bq^{4k})}{(q; q^2)_k(q^2; q^2)_k} \\ &= 1 + \sum_{k=1}^\infty \frac{q^{2k^2-k}(-bq)_{2k-1}(1 + bq^{4k})}{(q)_{2k}}. \end{aligned} \tag{4.1}$$

The series on the right in (4.1) is actually the series in Lebesgue’s identity (1.1). To see this, we will add consecutive pairs of terms with odd ( $n = 2k - 1$ ) and even ( $n = 2k$ ) subscripts in (1.1) to get (4.1). More precisely, observe that

$$\begin{aligned} \sum_{n=0}^\infty \frac{q^{n(n+1)/2}(-bq)_n}{(q)_n} &= 1 + \sum_{k=1}^\infty \left\{ \frac{q^{2k^2-k}(-bq)_{2k-1}}{(q)_{2k-1}} + \frac{q^{2k^2+k}(-bq)_{2k}}{(q)_{2k}} \right\} \\ &= 1 + \sum_{k=1}^\infty \frac{q^{2k^2-k}(-bq)_{2k-1}}{(q)_{2k}} \left\{ (1 - q^{2k}) + q^{2k}(1 + bq^{2k}) \right\} \\ &= 1 + \sum_{k=1}^\infty \frac{q^{2k^2-k}(-bq)_{2k-1}(1 + bq^{4k})}{(q)_{2k}}, \end{aligned}$$

which is the right hand side of (4.1). This proves Lebesgue’s identity as a consequence of (2.10).

This amalgamation of the terms with odd and even subscripts is also present in the special case  $b = -1$  in (4.1), which yields Gauss’ identity (1.2):

$$\frac{(q^2; q^2)_\infty}{(q; q^2)_\infty} = 1 + \sum_{k=1}^\infty q^{2k^2-k}(1 + q^{2k}) = \sum_{k=-\infty}^\infty q^{2k^2-k} = \sum_{n=0}^\infty q^{n(n+1)/2}. \tag{4.2}$$

### 5 Three Parameter Extension

Our combinatorial approach to (2.10) permits the introduction of one more parameter  $\zeta$ , which would keep track of the size of the largest part. In that case,  $\zeta$  would enter into the series (2.10), but we would lose the product representation on the right in (2.10); the product would be replaced by a series which would be simpler than the series on the left in (2.10). Actually, instead of keeping track of the largest part, it is combinatorially more convenient to keep track of the number of columns by a parameter  $w$  in the 2-modular graphs of the partitions with non-repeating odd parts.

If the number of columns is  $k$ , the largest part is either  $2k$  or  $2k - 1$ . Thus we have two cases:

Case E: Largest part is  $2k$

In this case, the three parameter generating function of partitions with non-repeating odd parts is:

$$\frac{w^k z q^{2k} (-zbq; q^2)_k}{(zq^2; q^2)_k}. \tag{5.1}$$

Case O: Largest part is  $2k - 1$

In this case, the three parameter generating function is:

$$\frac{w^k z b q^{2k-1} (-zbq; q^2)_{k-1}}{(zq^2; q^2)_{k-1}}. \tag{5.2}$$

The sum of the generating functions in (5.1) and (5.2) is:

$$\begin{aligned} & \frac{w^k z q^{2k-1} (-zbq; q^2)_{k-1}}{(zq^2; q^2)_k} \left\{ (q(1 + zbq^{2k-1}) + b(1 - zq^{2k})) \right\} \\ &= \frac{w^k z q^{2k-1} (-zbq; q^2)_{k-1} (b + q)}{(zq^2; q^2)_k}. \end{aligned} \tag{5.3}$$

We sum the expression in (5.3) over  $k$  and add 1 to get

$$g(b, z, w; q) = 1 + \sum_{k=1}^{\infty} \frac{w^k z q^{2k-1} (-zbq; q^2)_{k-1} (b + q)}{(zq^2; q^2)_k}. \tag{5.4}$$

as the three parameter generating function of partitions with non-repeating odd parts that replaces the product (2.2) and has an extra parameter  $w$ .

Next we try to get a series expansion for  $g(b, z, w; q)$  that extends (2.10) by the Durfee square analysis of 2-modular graphs of partitions with non-repeating odd parts. We have the two cases as in section 2.

Case1: The bottom right node of the Durfee square has a 2.

Case2: The bottom right node of the Durfee square has a 1.

Generating function of Case 1: As before, consider partitions with non-repeating odd parts whose 2-modular Ferrers graphs have a  $k \times k$  Durfee square. The arguments in Sect. 2 carry over with the extra parameter  $w$ , and we get the three parameter generating function of such partitions to be

$$w^k z^k q^{2k^2} \cdot \frac{(-zbq; q^2)_k}{(zq^2; q^2)_k} \cdot \frac{(-wbq; q^2)_k}{(wq^2; q^2)_k}, \tag{5.5}$$

which generalizes (2.8).

**Generating function of Case 2:** In this case, if we consider the relevant partitions with a fixed  $k \times k$  Durfee square, we get

$$bw^k z^k q^{2k^2-1} \cdot \frac{(-zbq; q^2)_{k-1}}{(zq^2; q^2)_{k-1}} \cdot \frac{(-wbq; q^2)_{k-1}}{(wq^2; q^2)_{k-1}}, \tag{5.6}$$

which extends (2.9).

We sum the expressions in (5.5) and (5.6) over  $k$  to get

$$\frac{w^k z^k q^{2k^2-1} (-zbq; q^2)_{k-1} (-wbq; q^2)_{k-1}}{(zq^2; q^2)_k (wq^2; q^2)_k} \left\{ q(1 + zbq^{2k-1})(1 + wbq^{2k-1}) + b(1 - zq^{2k})(1 - wq^{2k}) \right\} \tag{5.7}$$

With the extra parameter  $w$ , the expression within {...} in (5.7) simplifies as

$$(b + q)(1 + wzbq^{4k-1}).$$

Finally, summing the expression in (5.7) over  $k$  and adding 1 we get

$$\begin{aligned} & 1 + \sum_{k=1}^{\infty} \frac{w^k z^k q^{2k^2-1} (-zbq; q^2)_{k-1} (-wbq; q^2)_{k-1} (b + q)(1 + wzbq^{4k-1})}{(zq^2; q^2)_k (wq^2; q^2)_k} \\ &= 1 + \sum_{k=1}^{\infty} \frac{w^k zq^{2k-1} (-zbq; q^2)_{k-1} (b + q)}{(zq^2; q^2)_k}. \end{aligned} \tag{5.8}$$

This is a three parameter extension of (2.10) with the product on the right in (2.10) replaced by a series on the right in (5.8).

In the next section we show how (5.8) is connected to the Rogers-Fine identity.

## 6 The Rogers-Fine Identity

The Rogers-Fine identity in the form obtained by Fine [11] is

$$\sum_{n=0}^{\infty} \frac{(\alpha q)_n \tau^n}{(\beta q)_n} = \sum_{n=0}^{\infty} \frac{(\alpha q)_n (\alpha \tau q / \beta)_n \beta^n \tau^n q^{n^2} (1 - \alpha \tau q^{2n+1})}{(\beta q)_n (\tau)_{n+1}}. \tag{6.1}$$

Fine [11] studied the function

$$F(\alpha, \beta, \tau; q) = \sum_{n=0}^{\infty} \frac{(\alpha q)_n \tau^n}{(\beta q)_n} \tag{6.2}$$



in detail under various transformations and iterations, and obtained a number of results involving this function, one of which was identity (6.1). Subsequently, Andrews [8] gave a combinatorial proof of (6.1) by studying partitions in terms of  $n \times 2n$  Durfee rectangles. More recently, another proof of (6.1) has been given in [15] using in the course Sylvester’s bijection connecting partitions into odd parts with partitions into distinct parts. Identity (6.1) is a special case of Watson’s transformation formula for a terminating very well poised  ${}_8\phi_7$  as a multiple of a terminating balanced  ${}_4\phi_3$ . Recently, Rowell and Yee [13] have given a combinatorial proof of a special case of a  ${}_4\phi_3$  from which the Rogers-Fine identity (6.1) follows.

In [5], we provided the simplest and the most direct derivation of the Rogers-Fine identity by studying the following three parameter generating function of unrestricted partitions (this was first announced in [3] without proof):

$$f(a, b, c; q) = \sum_{\pi} (1 - a)^{v_d(\pi)} b^{v(\pi)} c^{\lambda(\pi)} q^{\sigma(\pi)}, \tag{6.3}$$

where the sum is over all partitions  $\pi$ , and where

$\sigma(\pi)$  = the sum of the parts of  $\pi$ ,

$\lambda(\pi)$  = the largest part of  $\pi$ ,

$v(\pi)$  = the number of parts of  $\pi$ ,

and

$v_d(\pi)$  = the number of different parts of  $\pi$ .

It turns out that

$$f(a, b, c; q) = 1 + \sum_{n=1}^{\infty} \frac{(1 - a)(abq)_{n-1} bc^n q^n}{(bq)_n}, \tag{6.4}$$

by a straightforward analysis of the defining sum in (6.3). Our function  $f$  is related to Fine’s function  $F$  by the equation

$$\frac{(1 - bq)}{(1 - a)bcq} \{f(a, b, c; q) - 1\} = F(ab, bq, cq; q). \tag{6.5}$$

as can be seen by comparing the series in (6.2) and (6.4), but it is  $f$  that has such a natural partition interpretation. Under conjugation, the largest part and the number of parts are interchanged, and the number of different parts is invariant. Thus from the definition of  $f$  in (5.3), it follows that

$$f(a, b, c; q) = f(a, c, b; q), \tag{6.6}$$

although this symmetry is not seen in the series in (6.4). So in [3], we sought a series representation for  $f$  that would render this symmetry explicit. An analysis of the three parameter generating function of unrestricted partitions via Durfee squares provided the desired expansion, namely,

$$\begin{aligned}
 & 1 + \sum_{n=1}^{\infty} \frac{(1-a)(abq)_{n-1}bc^nq^n}{(bq)_n} \\
 &= 1 + \sum_{n=1}^{\infty} \frac{b^n c^n q^{n^2} (1-a)(abq)_{n-1}(acq)_{n-1}(1-abcq^{2n})}{(bq)_n(cq)_n}. \tag{6.7}
 \end{aligned}$$

Identity (6.7) is equivalent to the Rogers-Fine identity because the series on the right in (6.1) and (6.7) are also related via the transformation (6.5). Identity (6.7) was first stated in [2] without proof.

Although the function  $f$  defined in (6.3) is so natural and simple, it has not attracted much attention, perhaps because its representation is only in the form of a series as in (6.4) and not as a product. But if one of the parameters  $b$  or  $c$  is set equal to 1, we do get a product. This special case is actually Cauchy’s identity (see Andrews [6])

$$\sum_{n=0}^{\infty} \frac{(a)_n c^n q^n}{(q)_n} = f(a, 1, c; q) = f(a, c, 1; q) = \frac{(acq)_{\infty}}{(cq)_{\infty}}. \tag{6.8}$$

Thus, choosing  $b = 1$  in the series on the right in (6.7), we have

$$1 + \sum_{k=1}^{\infty} \frac{c^k q^{k^2} (a)_k (acq)_{k-1} (1 - acq^{2k})}{(q)_k (cq)_k} = \frac{(acq)_{\infty}}{(cq)_{\infty}}, \tag{6.9}$$

which is the product form of the Rogers-Fine identity in this special case.

The three parameter identity (5.8) is equivalent to the Rogers-Fine identity. More precisely, the substitutions

$$q \mapsto q^2, \quad b \mapsto z, \quad c \mapsto w, \quad \text{and} \quad a \mapsto bq^{-1}, \tag{6.10}$$

convert (6.7) to (5.8). Although (5.8) and (6.7) are equivalent, there are crucial differences in the combinatorics underlying them for two reasons: (a) In discussing (5.8) combinatorially, we enumerated only the corners having a 1 in the 2-modular graphs and not the corners having a 2 and (b) the last substitution  $a \mapsto bq^{-1}$  changes the combinatorics because  $q$  is present in the substitution.

Just as (6.7) corresponds to (5.8), the product form (6.9) of the Rogers-Fine identity corresponds to (2.10).

## 7 Partition Interpretations

The identities of Sylvester, Lebesgue, and Rogers-Fine have partition interpretation, the first two in the form of weighted partition identities.

In [1], I showed that Sylvester’s identity (3.1) is equivalent to the following:

**Theorem 1.** *Let  $D$  denote the set of partitions into distinct parts, and  $D_3$  the set of partitions into parts differing by at least 3. Let  $\sigma(\pi)$ ,  $\nu(\pi)$ , and  $\nu_d(\pi)$  be as in section 4. Also, for  $\pi^* \in D_3$ ,  $\pi^* : h_1 + h_2 + \dots + h_k$ , let  $\nu_3(\pi^*)$  denote the number of strict inequalities  $h_i - h_{i+1} > 3$ , for  $i = 1, 2, \dots, k$ , where  $h_{k+1} = -1$ . Then we have*

$$\sum_{\pi \in D, \sigma(\pi)=n} c^{\nu(\pi)} = \sum_{\pi^* \in D_3, \sigma(\pi^*)=n} c^{\nu(\pi^*)} (1 + c)^{\nu_3(\pi^*)}.$$

I also studied the combinatorics underlying Theorem 1. This involved the study of Ferrers graphs of partitions into distinct parts and the hooks of these graphs which lead to partitions into parts differing by  $\geq 3$ . By going from partitions into distinct parts to partitions in  $D_3$  via hooks, we get a surjective map between the sets  $D$  and  $D_3$ , which yields Theorem 1 directly without recourse to Sylvester’s identity (3.1). In [2], I showed that Theorem 1 has a three parameter refinement in which we can have parameters  $a, b, c$  keep track on the number of parts in residue classes 1, 2,  $0(mod\ 3)$ , respectively. It was also noted in [2] that this three parameter refinement of Theorem 1 is equivalent to the three parameter generalization and refinement of Göllnitz’s (Big) partition theorem in Alladi-Andrews-Gordon [6].

Subsequently, I studied [4] partitions into distinct odd parts by representing these partitions as 2-modular Ferrer’s graphs. By considering hooks in these graphs, we get partitions into parts that differ by  $\geq 6$ , with strict inequality when a part is even, and with 2 not as a part. Interestingly, the correspondence via hooks in such 2-modular graphs yields a bijection and not a surjection<sup>1</sup> as in the case of Theorem 1. More specifically, we get the following elegant result [4]:

**Theorem 2.** *The number of partitions  $\pi$  of an integer  $n$  into distinct odd parts is equal to the number of partitions  $\tilde{\pi}$  of  $n$  into parts that differ by  $\geq 6$ , where the inequality is strict if a part is even, and 2 is not a part.*

**Refinement.** This hook operation immediately yields the following refinement of Theorem 2: *The number of parts of  $\pi$  is equal to the number of parts of  $\tilde{\pi}$ , with the convention that the even parts of  $\tilde{\pi}$  are counted twice.* We now show that this refinement of Theorem 2 is in fact the partition interpretation of (3.3).

To this end, we note the decomposition

$$1 + bq^{4k-1} = (1 + bq^{2k-1}) - bq^{2k-1}(1 - q^{2k}). \tag{7.1}$$

---

<sup>1</sup> It is worthwhile to note that under dilations (in this case  $q \mapsto q^2$ ) and translations, the underlying combinatorics can change and this is non-trivial; in going from Theorem 1 to Theorem 2, the surjection changed to a bijection.

Using the decomposition in (7.1), we rewrite (3.3) as

$$\begin{aligned}
 (-bq; q^2)_\infty &= \sum_{k \geq 0} \frac{b^k q^{3k^2 - 2k} (-bq; q^2)_k}{(q^2; q^2)_k} - \sum_{k \geq 1} \frac{b^k q^{3k^2 - 2k} b q^{2k-1} (-bq; q^2)_{k-1}}{(q^2; q^2)_{k-1}} \\
 &= \Sigma_1 - \Sigma_2, \quad \text{respectively.} \tag{7.2}
 \end{aligned}$$

In  $\Sigma_1$ , the term  $3k^2 - 2k$  represents the minimal partition into parts that differ by  $\geq 6$ , namely  $1 + 7 + 13 + \dots + (6k - 5)$ . Represent this minimal partition as a Ferrer’s graph. Then the term  $(q^2; q^2)_k$  may be interpreted as the imbedding of pairs of columns of length  $j$  into this graph for  $1 \leq j \leq k$ , thereby yielding all partitions into  $k$  distinct odd parts that differ by  $\geq 6$ . Now in the graphs of such partitions, the term  $(-bq; q^2)_k$  may be interpreted as imbedding at most one pair of columns of length  $j, j - 1$ , for each  $j \in [1, k]$ . Each imbedding of a pair of columns  $j, j - 1$  creates an even part, and makes the gap  $> 6$ . Also this is exactly how even parts are produced in the graph. Thus,  $\Sigma_1$  is the generating function of partitions into parts that differ by  $\geq 6$  with strict inequality if a part is even.

Regarding  $\Sigma_2$ , we interpret  $(3k^2 - 2k) + (2k - 1)$  as the minimal partition with 2 as a part, and with distinct odd parts such that all parts differ by  $\geq 6$ , namely the minimal partition  $2 + 9 + 15 + \dots + (6k - 3)$  for  $k \geq 2$ , and 2 for  $k = 1$ . To such partitions, the term  $(q^2; q^2)_{k-1}$  imbeds pairs of columns of length  $j$  for  $1 \leq j \leq (k - 1)$  only and not for  $j = k$ . Thus the smallest part remains as 2, the rest of the parts are odd, and all parts differ by  $\geq 6$ . Finally the term  $(-q; q^2)_{k-1}$  imbeds at most one pair of columns of lengths  $j, j - 1$  for each  $j \in [1, k - 1]$ . Each imbedding creates an even part, and the inequality becomes  $> 6$ . The smallest part 2 remains untouched. Thus  $\Sigma_2$  is the generating function of the same type of partitions as those enumerated by  $\Sigma_1$  but with the extra condition that the smallest part is 2. Hence the right hand side of (7.2) represents the generating function of partitions into parts that differ by  $\geq 6$  with strict inequality if a part is even, and not having 2 as a part. When an even part is created by an imbedding, an extra factor  $b$  is introduced, and so this even part is counted twice. By keeping track of the powers of  $b$  on both sides of (7.2) and from the correspondence given above, we have shown that the refinement stated immediately after Theorem 2 is the partition interpretation of (3.3).

Actually, just as we noted a three parameter refinement of Theorem 1 in [2], we have a three parameter refinement of Theorem 2 in [4].

With regard to Lebesgue’s identity (1.1), Gordon and I [7] gave a combinatorial proof after showing that it has the following weighted partition interpretation:

**Theorem 3.** *Let  $G$  denote the set of partitions with non-repeating even parts, and let  $D$  be as in Theorem 1. For  $\pi \in G$ , let  $v_e(\pi)$  denote the number of even parts in  $\pi$ . For  $\pi^* \in D, \pi^* = m_1 + m_2 + \dots + m_k$ , let  $v_1(\pi^*)$  denote the number of strict inequalities  $m_i - m_{i+1} > 1$ , where  $m_{i+1} = 0$ . Then we have*

$$\sum_{\pi \in G, \sigma(\pi) = n} b^{v_e(\pi)} = \sum_{\pi^* \in D, \sigma(\pi^*) = n} b^{v_1(\pi^*)} (1 + b)^{v_1(\pi^*)}.$$

There are two reasons for including a discussion of Theorem 2 in this paper: (a) Theorem 2 is the partition interpretation of (3.3) and (b) the 2-modular graph approach to Theorem 2 in [4] is after all the special case of the main consideration here. That is, in this paper we are considering 2-modular graphs of partitions in which the odd parts do not repeat and separately keep track of the number of odd and even parts using two parameters. If we set equal to 0 the parameter keeping track of the number of even parts, we connect to Theorem 2. As noted in this section, the various special cases of (2.10) leading to the identities of Sylvester and Lebesgue, have interesting weighted partition interpretations. Similarly, it would be worthwhile to have a nice weighted partition interpretation of the more general identity (2.10).

**Acknowledgments** My father, Professor Alladi Ramakrishnan, was the greatest source of strength for me. He encouraged me in every aspect of my academic and research career. I was inspired by his passion for fundamental research and his grand scientific vision; indeed it was because of the exposure I had by meeting the eminent scientists he brought to our family home *Ekanra Nivas* in Madras, India, and those whom I met by being with him on his worldwide scientific tours that I decided on a research career. The main idea for this paper came in summer 2007 while I was visiting my parents in Madras. Hence, it is a privilege for me to dedicate this paper to his memory.

## References

1. K. Alladi, "Partition identities involving gaps and weights", *Trans. Am. Math. Soc.*, **349** (1997), 5001–5019.
2. K. Alladi, "A combinatorial correspondence related to Göllnitz's (Big) partition theorem and applications", *Trans. Am. Math. Soc.*, **349** (1997), 2721–2735.
3. K. Alladi, "A fundamental but unexploited partition invariant", in *Number Theory and its Applications* (S. Kanemitsu, K. Gyory, Eds.), *Developments in Math.*, **2**, Kluwer, Dordrecht (1999), 19–23.
4. K. Alladi, "A variation on a theme of Sylvester – a smoother road to Göllnitz's partition theorem", *Discrete Math.* **196** (1999), 1–11.
5. K. Alladi, "A new combinatorial study of the Rogers-Fine identity and a related partial theta series", *Int'l J. Num. Th.*, **5** (2009), 1311–1320.
6. K. Alladi, G. E. Andrews and B. Gordon, "Generalizations and refinements of a partition theorem of Göllnitz" *J. Reine Ang. Math.*, **460** (1995), 165–188.
7. K. Alladi and B. Gordon, "Partition identities and a continued fraction of Ramanujan", *J. Comb. Th. Ser. A*, **63** (1994), 214–245.
8. G. E. Andrews, "Two theorems of Gauss and allied identities proved arithmetically", *Pacific J. Math.*, **41** (1972), 563–578.
9. G. E. Andrews, "The theory of partitions", *Encyclopedia of Math. and its Applications*, **2**, Addison Wesley, Reading (1976).
10. A. Berkovich and F. Garvan, "Some observations on Dyson's new symmetries of partitions", *J. Comb. Th. Ser. A* **100** (2002), 61–93.
11. N. J. Fine, "Basic hypergeometric series and applications", *Math. Surveys and Monographs* **27**, Amer. Math. Soc., Providence (1988).
12. M. Hirschhorn and J. Sellers, "Arithmetic properties of partitions with odd parts distinct", *Ramanujan J.*, **22** (2010, to appear).

13. M. J. Rowell and A. J. Yee, “A bijective proof of a limiting case Watson’s  ${}_8\phi_7$  transformation formula”, *Ramanujan J.*, **20** (2009), 267–280.
14. J. J. Sylvester, “A constructive theory of partitions in three Acts, an Interact, and an Exodion”, *Am. J. Math.*, **5** (1882), 251–330.
15. J. Zeng, “The  $q$ -variations of Sylvester’s bijection between odd and strict partitions”, *Ramanujan J.*, **9** (2005), 289–303.

# ***q*-Catalan Identities**

**George E. Andrews**

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** *q*-Analogues of the Catalan number identities of Touchard, Jonah, and Koshy are derived.

**Mathematics Subject Classification (2000)** 05A10, 05A16, 05A30

**Key words and phrases** Catalan numbers · *q*-Catalan numbers · *q*-Analogues

## **1 Introduction**

Alladi Ramakrishnan was a grand man. I mostly knew him in the last two decades of his life. One always took away from conversation with him a sense of his joy in living and his excitement over mathematics and physics.

In my visits with him in the winter of 2008 shortly before his death, he was enthralled with the implications of and extensions of Pascal’s triangle. He had prepared an expository article titled: “Magic Lattice Imbedding Pascal Triangles.” I was a very receptive audience. Now that Professor Alladi Ramakrishnan is gone, I propose to remember him with some observations about the Catalan numbers:

$$C_n = \frac{1}{n + 1} \binom{2n}{n}, \tag{1.1}$$

a topic closely related to Pascal’s triangle. These famous integers are, by their very definition, slight variations on the central binomial coefficients. In addition, Koshy [10] has just published a 422-page book titled “Catalan Numbers.”

---

Partially supported by National Science Foundation Grant DMS-0801184.

G.E. Andrews

Department of Mathematics, The Pennsylvania State University, University Park, PA 16802, USA  
e-mail: [andrews@math.psu.edu](mailto:andrews@math.psu.edu)

Stanley [11–13] has devoted extensive attention to Catalan numbers and Gould [8] has provided an extensive bibliography. These are just a few of the many works on the Catalan numbers.

My interest in the Catalan numbers has arisen from looking at various  $q$ -analogs (cf. [6]), that is, polynomials or rational functions in a variable  $q$  that reduce naturally to the Catalan numbers when  $q = 1$ .

To provide the flavor of  $q$ -analogs, we recall Lagrange’s identity for the sum of the squares of the binomial coefficients [10, p. 89]:

$$\sum_{j=0}^n \binom{n}{j}^2 = \binom{2n}{n}. \tag{1.2}$$

Let us recall the Gaussian polynomials (a.k.a.  $q$ -binomial coefficients):

$$\begin{bmatrix} n \\ j \end{bmatrix}_q = \begin{cases} 0, & \text{if } j < 0 \text{ or } j > n, \\ \frac{(q; q)_n}{(q; q)_j (q; q)_{n-j}}, & 0 \leq j \leq n, \end{cases} \tag{1.3}$$

where

$$(a; q)_N = (1 - a)(1 - aq) \cdots (1 - aq^{N-1}). \tag{1.4}$$

The  $q$ -analog of (1.2) is well known [2, p. 37, (33.10),  $m = n = h, k \rightarrow n - k$ ]

$$\sum_{j=0}^n q^{j^2} \begin{bmatrix} n \\ j \end{bmatrix}_q^2 = \begin{bmatrix} 2n \\ n \end{bmatrix}_q. \tag{1.5}$$

While there are a number of  $q$ -analogs of the Catalan numbers (cf. [5]), we shall be primarily interested in the following two:

First,

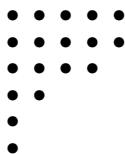
$$C_n(q) = \frac{(1 - q)}{(1 - q^{n+1})} \begin{bmatrix} 2n \\ n \end{bmatrix}_q. \tag{1.6}$$

Clearly by l’Hôpital’s rule,

$$C_n(1) = C_n.$$

(Actually  $C_n(q)$  is a polynomial in  $q$  so that we may take  $q = 1$  directly without invoking l’Hôpital.)  $C_n(q)$  was shown [4] to be related to partitions as follows:

The partition  $5 + 5 + 4 + 2 + 1 + 1$  has Ferrers graph



and conjugate  $6 + 4 + 3 + 3 + 2$  (read columns instead of rows). The largest square of nodes in a partition (in this case, a  $3 \times 3$  square) is called the *Durfee square*.



We say that a partition

$$\lambda_1 + \lambda_2 + \dots + \lambda_r \quad (\lambda_i \geq \lambda_{i+1})$$

with conjugate

$$\lambda'_1 + \lambda'_2 + \dots + \lambda'_t \quad (\lambda'_i \geq \lambda'_{i+1})$$

is *Catalan* provided  $\lambda_i < \lambda'_i$  for  $1 \leq i \leq s$ , where  $s$  is the side of the Durfee square.

It was proved in [4, Corollary 1] that  $C_N(q)$  is the generating function for Catalan partitions with largest part  $< N$  and number of parts  $\leq N$ .

For example,  $C_3(q) = 1 + q^2 + q^3 + q^4 + q^6$  and the partitions being generated are  $1 + 1, 1 + 1 + 1, 2 + 1 + 1,$  and  $2 + 2 + 2$ .

In another paper [3], we considered

$$C_n(\lambda, q) = \frac{q^{2n}(-\lambda/q; q^2)_n}{(q^2; q^2)_n}. \tag{1.7}$$

There, it was shown that

$$\lim_{q \rightarrow 1} C_n(-1, q) = \lim_{q \rightarrow 1} C_n(1, -q) = -2^{1-2n} C_{n-1}.$$

In this case [3, (3.2)],  $C_n(\lambda, q)$  is the two-variable generating function for partitions without repeated odd parts whose total number of parts is  $n$  with the exponent on  $\lambda$  counting the number of odd parts and the exponent on  $q$  exhibiting the number being partitioned.

The overarching object of this chapter is to emphasize the methods for finding  $q$ -analogs [1, Sect. 5]. Succinctly put, this method reduces binomial coefficient identities to identities for the generalized hypergeometric function [5, p. 8]:

$${}_{n+1}F_n \left[ \begin{matrix} a_0, a_1, \dots, a_n; t \\ b_1, \dots, b_n \end{matrix} \right] = \sum_{j=0}^{\infty} \frac{[a_0]_j [a_1]_j \dots [a_n]_j t^j}{j! [b_1]_j \dots [b_n]_j}, \tag{1.8}$$

where

$$[A]_j = A(A + 1) \dots (A + j - 1).$$

(We note that the symbol  $[A]_j$  is unconventional but is necessary in a paper where the symbol  $(A; q)_n$  also appears.)

Once this first step is complete, there is generally a canonical  $q$ -analog from the world of generalized  $q$ -hypergeometric functions [7, p. 4]:

$${}_{n+1}\phi_n \left( \begin{matrix} A_0, A_1, \dots, A_n; q, t \\ B_1, \dots, B_n \end{matrix} \right) = \sum_{j=0}^{\infty} \frac{(A_0; q)_j (A_1; q)_j \dots (A_n; q)_j t^j}{(q; q)_j (B_1; q)_j \dots (B_n; q)_j}. \tag{1.9}$$

The final step involves reversing the evaluation in step 1 to provide the perfect  $q$ -analog.

We have chosen three identities. First is Touchard’s identity [10, p. 319]:

$$C_{n+1} = \sum_{r \geq 0} \binom{n}{2r} 2^{n-2r} C_r. \tag{1.10}$$

We shall prove

**Theorem 1.**

$$C_{n+1}(q) = \sum_{r \geq 0} q^{2r^2+2r} \begin{bmatrix} n \\ 2r \end{bmatrix}_q C_r(q) \frac{(-q^{r+2}; q)_{n-r}}{(-q; q)_r}. \tag{1.11}$$

Note how, in this instance, the  $q$ -analog of  $2^{n-2r}$  is  $(-q^{r+2}; q)_{n-r}/(-q; q)_r$ . This is surely not something easily guessed.

Koshy provides another recursive formula for Catalan numbers [10, p. 322]:

$$C_n = \sum_{r=1}^{\infty} (-1)^{r-1} \binom{n-r+1}{r} C_{n-r}. \tag{1.12}$$

We shall prove

**Theorem 2.**

$$C_n(q) = \sum_{r=1}^n (-1)^{r-1} q^{r^2-r} \begin{bmatrix} n-r+1 \\ r \end{bmatrix}_q C_{n-r}(q) \frac{(-q^{n-r+1}; q)_r}{(-q; q)_r}. \tag{1.13}$$

Note that in this  $q$ -analog, the factor  $(-q^{n-r+1}; q)_r/(-q; q)_r$  is equal to 1 when we set  $q = 1$ .

Both Theorems 1 and 2 are deduced from the  $q$ -analog of the Chu-Vandermonde summation [7, p. 236, (II.6) and (II.7)].

Finally, we consider Jonah’s identity [10, p. 325]

$$\binom{n+1}{r} = \sum_{j=0}^r \binom{n-2j}{r-j} C_j, \tag{1.14}$$

provided  $2r \leq n$ .

We shall prove

**Theorem 3.**

$$\frac{(1 + q^{n-r+1})}{(1 + q^{r+1})} \begin{bmatrix} n+1 \\ r \end{bmatrix}_{q^2} = -(-q; q)_{n+1} \sum_{j=0}^r \begin{bmatrix} n-2j \\ r-j \end{bmatrix}_{q^2} \frac{C_{j+1}(-1; q)}{(-q; q)_{n-2j}} q^{-j-1}. \tag{1.15}$$

Here, we must rely on the  $q$ -analog of the Pfaff-Saalschütz summation [7, p. 237, (II.12)].

In our final section, we discuss some of the combinatorial questions that arise from these considerations.

### 2 $q$ -Touchard's Identity

Following the standard reduction rules given in [1, Sect. 5], we see that (1.10), Touchard's identity, may be reduced to the equivalent assertion:

$$2^n {}_2F_1 \left[ \begin{matrix} -\frac{n}{2}, -\frac{n}{2} + \frac{1}{2}; 1 \end{matrix} \right] = \frac{2^{2n+2} \left[ \frac{1}{2} \right]_{n+1}}{(n+2)!} = C_{n+1}. \tag{2.1}$$

Identity (2.1) is a specialization of the classic Chu-Vandermonde summation [5, p. 3]. Now, we choose the natural  $q$ -analog of (2.1) [7, p. 236, (II.7)]

$$(-q^2; q)_{n2} \phi_1 \left( \begin{matrix} q^{-n}, q^{-n+1}; q^2, q^2 \\ q^4 \end{matrix} \right) = C_{n+1}(q), \tag{2.2}$$

and the standard reduction of the left-hand side of (2.2) following the rules given in [1, Sect. 5] yields (1.11) which is Theorem 1.

### 3 Koshy's Identity

First, we rewrite (1.2) as follows:

$$\sum_{r=0}^n (-1)^r \binom{n-r+1}{r} C_{n-r} = 0. \tag{3.1}$$

This identity is equivalent to the assertion that

$${}_2F_1 \left( \begin{matrix} -\frac{n-1}{2}, -\frac{n}{2}; 1 \\ -n + \frac{1}{2} \end{matrix} \right) = 0 \tag{3.2}$$

and (3.2) is also a specialization of the Chu-Vandermonde summation [5, p. 3]. The corresponding  $q$ -Chu-Vandermonde summation [7, p. 236, (II.7)] is

$${}_2\phi_1 \left( \begin{matrix} q^{-n-1}, q^{-n}; q^2, q^2 \\ q^{1-2n} \end{matrix} \right) = 0. \tag{3.3}$$

After reversing the steps from (3.2) to (3.1) in the  $q$ -analogous procedure, we obtain

$$\sum_{r=0}^n (-1)^r q^{r^2-r} \begin{bmatrix} n-r+1 \\ r \end{bmatrix}_q C_{n-r}(q) \frac{(-q^{n-r+1}; q)_r}{(-q; q)_r} = 0. \tag{3.4}$$

Equation (3.4) reduces to (1.13) once we move the  $r = 0$  term to the other side of the equation.

### 4 Jonah’s Identity

Identity (1.15) is deeper than the previous results. In this case, assuming  $0 \leq 2r \leq n$ ,

$$\begin{aligned} \sum_{j=0}^r \binom{n-2j}{r-j} C_j &= -\frac{1}{2} \binom{n+2}{r+1} \left( {}_3F_2 \left( \begin{matrix} -r-1, -n+r-1, -\frac{1}{2}; 1 \\ -\frac{n}{2}-1, \frac{n}{2}-\frac{1}{2} \end{matrix} \right) - 1 \right) \\ &= -\frac{1}{2} \binom{n+2}{r+1} \left( \frac{n-2r}{n+2} - 1 \right) \quad (\text{by [5, Sect. 2.2]}) \\ &= \binom{n+1}{r}. \end{aligned} \tag{4.1}$$

The summation identity used was Pfaff-Saalschütz. The related  $q$ -analog [5, Sect. 8.4] is

$${}_3\phi_2 \left( \begin{matrix} q^{-2r-2}, q^{-2n+2r-2}, q^{-1}; q^2, q^2 \\ q^{-n-2}, q^{-n-1} \end{matrix} \right) - 1 = -\frac{(1-q^{r+1})(1+q^{n-r+1})}{(1-q^{n+2})}. \tag{4.2}$$

Finally using (4.2) to produce the  $q$ -analog of (4.1), we obtain (1.15) which is Theorem 3.

### 5 Conclusion

First, we note along with Koshy [10, p. 327] that Jonah’s theorem was generalized by Hilton and Pedersen [9] to remove the restrictions  $2r \leq n$ . A  $q$ -analog of the Hilton–Pedersen extension can be obtained in exactly the way that the  $q$ -analog of Jonah’s theorem was proved. Indeed, the following identity is equivalent to a  $q$ -analog of the Hilton–Pedersen identity [10, p. 327]:

$$\begin{aligned} \sum_{j \geq 0} \frac{(a^2 q^{2-2r-2j}; q^2)_{r-j} (-a q^{1-2j}; q)_{2j+1} C_{j+1} (-1; q) q^{-j}}{(q^2; q^2)_{r-j}} \\ = \frac{q^{1+2r-r^2} a^{2r} (a^{-2} q^{-2}; q^2)_r (a q + q^r) (-1)^{r-1}}{(1 + q^{r+1})(q^2; q^2)_r}. \end{aligned} \tag{5.1}$$

More intriguing are some obvious combinatorial questions that lie within some of our  $q$ -analogs. Suppose we rewrite (1.13) as

$$C_n(q) = \sum_{r=1}^n (-1)^{r-1} T_r(n, q), \tag{5.2}$$

where

$$T_r(n, q) = q^{r^2-r} \begin{bmatrix} n-r+1 \\ r \end{bmatrix}_q C_{n-r}(q) \frac{(-q^{n-r+1}; q)_r}{(-q; q)_r}. \tag{5.3}$$

**Problem 1.** Show that  $T_r(n, q)$  is a polynomial.

**Problem 2.** If  $2r \leq n$ , show that all the coefficients in  $T_r(n, q)$  are nonnegative.

**Problem 3.** Show that  $T_{r+1}(2r + 1, -q)$  has nonnegative coefficients.

**Problem 4.** Provide a partition-theoretic interpretation of  $T_r(n, q)$  for  $2r \leq n$  and for  $T_{n+1}(2n + 1, -q)$ .

**Problem 5.** In light of the fact that  $C_n(q)$  generates the Catalan partitions with largest part  $< n$  and number of parts  $\leq n$ , show by using Problem 4 to interpret the right-hand side of (5.2) that a sieve process eliminates all non-Catalan partitions.

## References

- [1] G. E. Andrews, Applications of basic hypergeometric functions, *SIAM Rev.*, **16** (1974), 441–484
- [2] G. E. Andrews, The theory of partitions. *Encyclopedia of Mathematics and Its Applications*, Vol. 2. (G. C. Rota, ed.) Addison-Wesley, Reading, MA, 1976 (Reprinted: Cambridge University Press, Cambridge, 1998)
- [3] G. E. Andrews, Catalan numbers,  $q$ -Catalan numbers and hypergeometric series, *J. Combin. Theory A*, **44** (1987), 267–273
- [4] G. E. Andrews, On the difference of successive Gaussian polynomials, *J. Stat. Plan. Inf.*, **34** (1993), 19–22
- [5] W. N. Bailey, *Generalized Hypergeometric Series*. Cambridge Tracts in Mathematics and Mathematical Physics No. 32, Cambridge University Press, Cambridge, 1935 (Reprinted: Hafner, New York, 1964)
- [6] J. F\"urlinger and J. Hofbauer,  $q$ -Catalan numbers, *J. Combin. Theory A*, **40** (1985), 248–264
- [7] G. Gasper and M. Rahman, *Basic Hypergeometric Series*. Encyclopedia of Mathematics and Its Applications, Vol. 35, Cambridge University Press, Cambridge, 1990
- [8] H. W. Gould, *Bell and Catalan Numbers: Research Bibliography of Two Special Number Sequences*, rev. ed., Combinatorial Research Institute, Morgantown, WV, 1978
- [9] P. Hilton and J. Pedersen, The ballot problem and Catalan numbers, *Nieuw Arch. Wisk.*, **7–8** (1990), 209–216
- [10] T. Koshy, *Catalan Numbers with Applications*, Oxford University Press, New York, 2009

- [11] R. P. Stanley, *Enumerative Combinatorics*, Vol. 1, Wadsworth & Brooks/Cole, Monterey, CA, 1986
- [12] R. P. Stanley, *Enumerative Combinatorics*, Vol. 2, Cambridge University Press, New York, 1999
- [13] R. P. Stanley, Catalan Addendum, <http://www-math.mit.edu/~rstan/ec/catadd.pdf>, version 6, October 2008

# Completing Brahmagupta's Extension of Ptolemy's Theorem

Richard Askey

*In memory of Alladi Ramakrishnan with thanks for his work in mathematics and physics and his hospitality of visitors*

**Summary** Brahmagupta extended Ptolemy's theorem on cyclic quadrilaterals to find the lengths of the diagonals, the segments made when they are cut at the point of intersection of the diagonals, and the lengths of the sides of the needles, the figures formed when opposite sides of the quadrilateral are extended until they meet. Proofs of these results are given, and a derivation of the 19th century result of the length of the third diagonal is given. This "diagonal" is formed by connecting the tips of the needles with a line segment.

**Mathematics Subject Classification (2000)** 01A32, 51M04

**Key words and phrases** Ptolemy's theorem · Brahmagupta · Third diagonal of cyclic quadrilateral

## 1 Introduction

Ptolemy's theorem on cyclic quadrilaterals is one of the gems of later Greek mathematics. We do not know how this theorem was discovered, but both the statement and the synthetic proof in [11, pages 50–51] have been admired by many. This proof has been included in many books such as the textbook [5, Sect. 198], in Heath's translation of Euclid [6, page 225], and the cultural history book [7, page 162], on the web [1], and there is even an article treating it as a "desert island theorem" [3].

As background, triangles are rigid, but quadrilateral are not. However, there is an important class of quadrilaterals which are rigid, those whose vertices lie on a circle; Ptolemy was very interested in those quadrilaterals, for he needed to be able

---

R. Askey

Department of Mathematics, University of Wisconsin, Madison, WI 53706, USA

e-mail: [askey@math.wisc.edu](mailto:askey@math.wisc.edu)

to approximate lengths of chords in circles in terms of the radius of the circle and the angle cut off by the chord. Ptolemy's theorem makes this possible since it is equivalent to the later treatment of trigonometry of triangles, containing the addition formulas of sine and cosine and also following from these addition formulas. Ptolemy's theorem is

**Theorem 1.** *The product of the diagonals of a cyclic quadrilateral is equal to the sum of the product of the opposite sides, or*

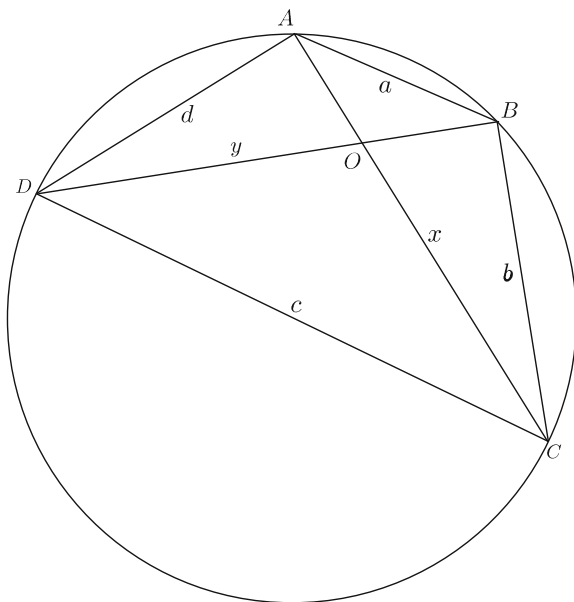
$$xy = ac + bd. \quad (1.1)$$

Since this quadrilateral is rigid, it should be possible to find the lengths of the diagonals in terms of the lengths of the sides. Brahmagupta [2, page 300, Chap. XII, Sect. IV, entry 28] stated formulas for the diagonals. Details were not given, but for the refinement he gave, which will be mentioned later, he gave a suggestion about how to derive them. There is also an interesting example in [2, pp. 301–305].

## 2 Brahmagupta's Refinements of Ptolemy's Theorem

For the cyclic quadrilaterals in Figure 1, the angles at  $B$  and  $D$  add to  $\pi$ . The length of the diagonal  $AC$  can be found by using the cosine law twice.

$$\begin{aligned} x^2 &= a^2 + b^2 - 2ab \cos B \\ &= c^2 + d^2 + 2cd \cos B \end{aligned} \quad (2.1)$$



**Figure 1** Ptolemy's theorem



since  $\cos D = \cos(\pi - B) = -\cos B$ . Eliminate  $\cos B$  by multiplication and addition to get

$$(ab + cd)x^2 = cd(a^2 + b^2) + ab(c^2 + d^2) \quad (2.2)$$

The right hand side is not attractive as it is, so first regroup the products to remove the squares, and then factor common expressions.

$$\begin{aligned} (ac)(ad) + (bc)(bd) + (ac)(bc) + (ad)(bd) &= (ac)[ad + bc] + (bd)[ad + bc] \\ &= (ac + bd)(ad + bc) \end{aligned}$$

Thus,

$$x^2 = \frac{(ac + bd)(ad + bc)}{(ab + cd)} \quad (2.3)$$

Similarly,

$$y^2 = \frac{(ac + bd)(ab + cd)}{(ad + bc)} \quad (2.4)$$

Ptolemy's theorem (1.1) follows from multiplication, and Brahmagupta's result follows from division:

$$\frac{x}{y} = \frac{(ad + bc)}{(ab + cd)} \quad (2.5)$$

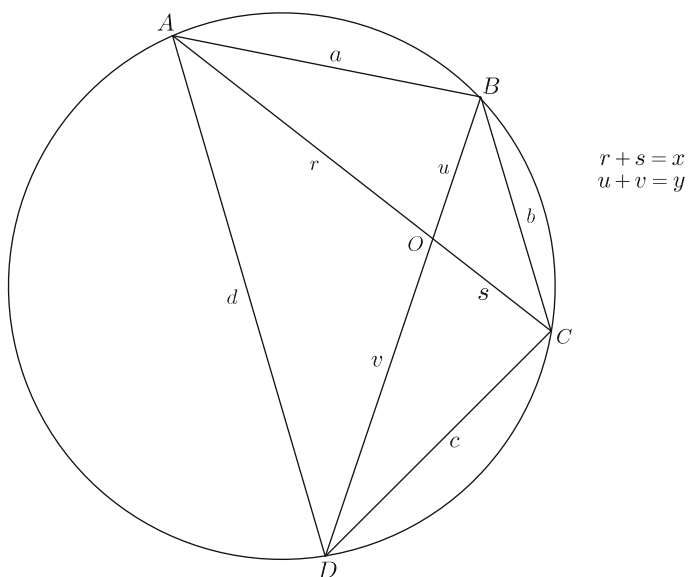
The proof above is in [8, Sect. 167], [4, p. 25], and [7, p. 219]. Heilbron also included another derivation of (2.3) found by Parameśvara from (1.1). The idea is simple and a gem. Interchange the sides  $a$  and  $d$ . One of the diagonals remains the same. The diagonals have lengths  $u$  and  $y$ . Then interchange the sides  $b$  and  $d$ , giving diagonals with lengths  $u$  and  $x$ . Then multiply to get  $x^2 y^2 u^2$  and divide by the square of Ptolemy's theorem with diagonals of lengths  $y$  and  $u$ . See [7, p. 219].

### 3 Further Results of Brahmagupta

In [2, p. 303, Sect. 32], Brahmagupta wrote: "At the intersection of the diagonals and perpendiculars, the lower segments of the diagonal and of the perpendiculars are found by proportion; those lines less these segments and the upper segments of the same. So in the needle as well as in the intersection [of the prolonged sides and perpendiculars]."

The last remark was given by the translator Colebrooke. Colebrooke also remarked: "The text relative to the method of finding those segments is irretrievably corrupt, and has been therefore omitted in the version." [2, p. 304].

Here is how to find the segments of the diagonals and the needles, which will be described below. As Brahmagupta remarked, proportions are the key (Figure 2).



**Figure 2** Brahmagupta’s theorem on segments of diagonals

Triangles  $AOB$  and  $DOC$  are similar since they each have two angles which cut off the same arcs of the circle. This gives

$$\frac{r}{v} = \frac{u}{s} = \frac{a}{c}$$

This along with  $r + s = x$  gives

$$\frac{av}{c} + \frac{cu}{a} = x$$

or

$$\frac{a}{c}(y - u) + \frac{c}{a}u = x$$

or

$$(c^2 - a^2)u = a(cx - ay)$$

Then use (2.3) and (2.4) to get

$$\begin{aligned} x &= (ad + bc)T \\ y &= (ab + cd)T \end{aligned}$$

where

$$T = \left[ \frac{ac + db}{(ab + cd)(ad + bc)} \right]^{\frac{1}{2}} \tag{3.1}$$

This gives

$$u = abT \tag{3.2}$$

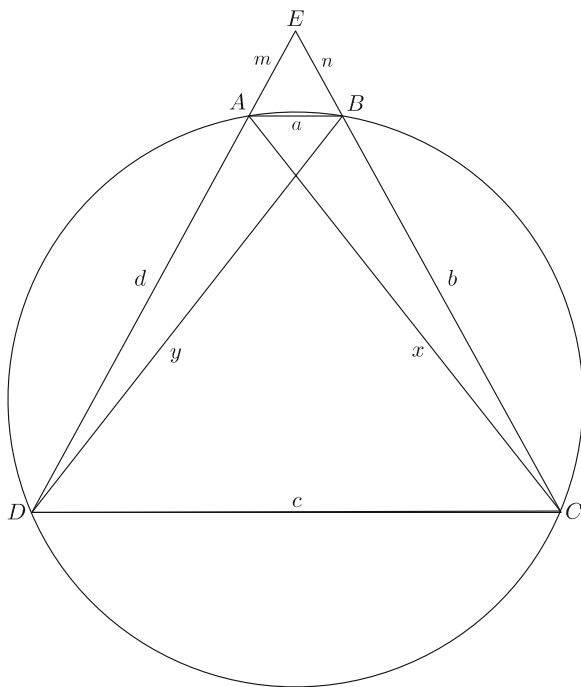


Figure 3 Brahmagupta's theorem on needles

and symmetry gives

$$v = cdT \tag{3.3}$$

$$r = adT \tag{3.4}$$

$$s = bcT \tag{3.5}$$

The needles are formed by extending opposite sides of the cyclic quadrilateral until they meet. Here is one (Figure 3).

Triangles  $AEC$  and  $BED$  are similar since angles at  $D$  and  $C$  cut off the same arc and angle  $E$  is in both triangles. Thus,

$$\frac{m}{n} = \frac{n + b}{m + d} = \frac{x}{y} = \frac{ad + bc}{ab + cd}$$

A little algebra gives

$$m = \frac{a(ad + bc)}{c^2 - a^2} \tag{3.6}$$

$$m + d = \frac{c(ab + cd)}{c^2 - a^2} \tag{3.7}$$

Symmetry then gives

$$n = \frac{a(ab + cd)}{c^2 - a^2} \quad (3.8)$$

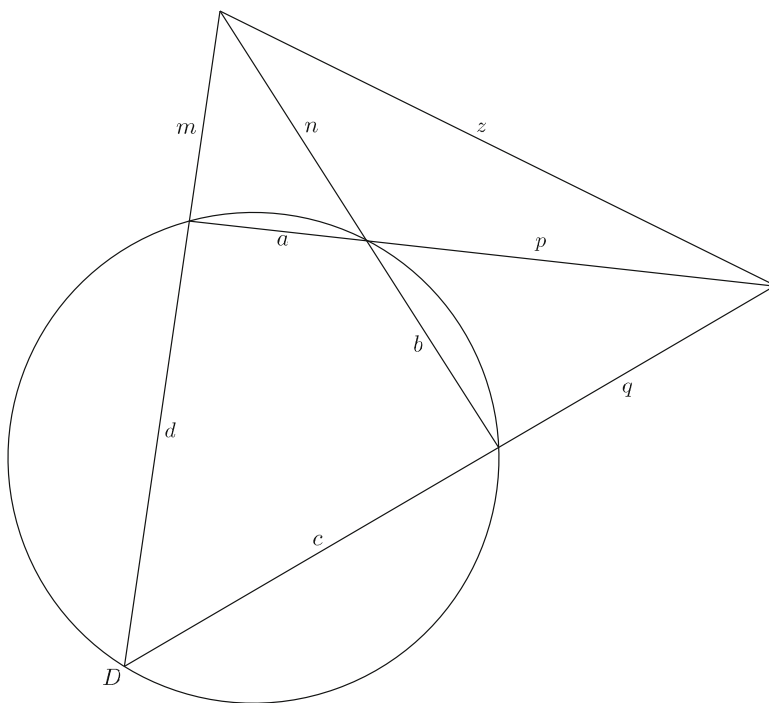
$$n + d = \frac{c(ad + bc)}{c^2 - a^2} \quad (3.9)$$

Notice that when  $a = c$ , a pair of opposite sides of the cyclic quadrilateral are parallel, so this needle does not exist.

The other needle is treated in the same way. We use all of the parts listed in (3.6)–(3.9) in the derivation of the result in the next section.

#### 4 The Third Diagonal of a Cyclic Quadrilateral

The third diagonal of a cyclic quadrilateral is formed by connecting the tips of the two needles. Brahmagupta seemingly did not consider this. Both Hobson [8, Sect. 16] and Durell and Robson [4, p. 26] dealt with this problem. Hobson gave a derivation of the formula, but one important step was just stated with a reference



**Figure 4** The third diagonal of a cyclic quadrilateral

to McDowell's Geometry. The only geometry book by McDowell that I have been able to find is [9], and the theorem in question is on pages 75 and 76, while Hobson wrote page 92. Durell and Robson gave a series of problems leading up to the length of the third diagonal. They ask the reader to prove a geometry theorem using the previous problem, which contains the essence of Sect. 3 in this paper. However, it is not necessary to do that as we will see (Figure 4).

To find the third diagonal, we again use the argument used in Sect. 2.

$$\begin{aligned} z^2 &= n^2 + p^2 + 2np \cos D \\ &= (m + d)^2 + (q + c)^2 - 2(m + d)(q + c) \cos D \end{aligned}$$

Remove  $\cos D$  and do some routine algebra to get

$$z^2 = (ab + cd)(ad + bc) \left[ \frac{ac}{(c^2 - a^2)^2} + \frac{bd}{(d^2 - b^2)^2} \right] \quad (4.1)$$

## References

- [1] Bogomolny, Alexander, *Ptolemy's Theorem*. <http://www.cut-the-knot.org/proofs/ptolemy.shtml>.
- [2] Colebrooke, Henry Thomas, *Algebra: With Arithmetic and Mensuration From The Sandskrit of Brahmagupta and Bhascara*, 1817, reprinted, Kessinger, Whitefish, MT, USA, 2008.
- [3] Crilly, Tony and Colin Fletcher, *Ptolemy's Theorem, Its Parent and Offspring*, in [10, pp. 42-49].
- [4] Durell, C.V. and A. Robson, *Advanced Trigonometry*, G. Bell, London, 1930, reprint, Dover, Mineola, N.Y., 2003.
- [5] Givental, A., translator and editor, *Kiselev's Geometry, Book 1*, Planimetry, Sumizdat, El Cerrito, CA, 2006.
- [6] Heath, Thomas L., *Euclid's Elements, Vol. 2*, second edition, Cambridge Univ. Press, 1926, reprinted, Dover, New York, 1956.
- [7] Heilbron, J.L., *Geometry Civilized*, Clarendon Press, Oxford, 1998.
- [8] Hobson, E. W., *A Treatise on Plane and Advanced Trigonometry*, Cambridge Univ. Press, Cambridge, first edition, 1891, seventh edition, 1928, reprinted, Dover, Mineola, NY, 2005.
- [9] McDowell, J., *Exercises on Euclid and in Modern Geometry*, Deighton Bell, Cambridge, 1878, available through google book search.
- [10] Pritchard, Chris, *The Changing Shape of Geometry*, Cambridge Univ. Press, Cambridge, 2003.
- [11] Toomer, G. J., translator, *Ptolemy's Almagest*, Dukworth, London, 1984, reprinted Princeton Univ. Press, princeton, 1998.

# A Transformation Formula Involving the Gamma and Riemann Zeta Functions in Ramanujan's Lost Notebook

Bruce C. Berndt<sup>1</sup> and Atul Dixit

*In Memory of Alladi Ramakrishnan*

**Summary** Two proofs are given for a series transformation formula involving the logarithmic derivative of the Gamma function found in Ramanujan's lost notebook. The transformation formula is connected with a certain integral embodying the Riemann zeta function that is similar to integrals examined by Ramanujan in his one published paper on the zeta function.

**Mathematics Subject Classification (2000)** Primary, 11M06; Secondary, 33B15

**Key words and phrases** Ramanujan's Lost Notebook · Gamma function · Riemann zeta function · Riemann  $\Xi$ -function

## 1 Introduction

Pages 219–227 in the volume [17] containing Ramanujan's lost notebook are devoted to material "Copied from the Loose Papers." We emphasize that these pages are *not* part of the original lost notebook found by George Andrews at Trinity College Library, Cambridge in the spring of 1976. These "loose papers", in the handwriting of G.N. Watson, are found in the Oxford University Library; the original manuscripts are in the library at Trinity College and have not been photocopied for publication. Most of these nine pages, which are divided into three rough, partial manuscripts, are connected with material in Ramanujan's published papers. However, there is much that is new in these fragments, which will be completely

---

<sup>1</sup> Research partially supported by grant H98230-07-1-0088 from the National Security Agency.

B.C. Berndt and A. Dixit  
Department of Mathematics, University of Illinois, 1409 West Green Street,  
Urbana, IL 61801, USA  
e-mail: [berndt@illinois.edu](mailto:berndt@illinois.edu); [aadixit2@illinois.edu](mailto:aadixit2@illinois.edu)

examined in [2]. One claim in the first manuscript on pages 219–220 is the subject of this short note and is the most interesting theorem in the manuscript. This claim provides a beautiful series transformation involving the logarithmic derivative of the gamma function and the Riemann zeta function. To state Ramanujan’s claim, it will be convenient to use the familiar notation [8, p. 952, formulas 8.360, 8.362, no. 1]

$$\psi(x) := \frac{\Gamma'(x)}{\Gamma(x)} = -\gamma - \sum_{k=0}^{\infty} \left( \frac{1}{k+x} - \frac{1}{k+1} \right), \tag{1.1}$$

where  $\gamma$  denotes Euler’s constant. We also need to recall the following functions associated with Riemann’s zeta function  $\zeta(s)$ . Let

$$\xi(s) := (s-1)\pi^{-\frac{1}{2}s}\Gamma(1+\frac{1}{2}s)\zeta(s).$$

Then Riemann’s  $\Xi$ -function is defined by

$$\Xi(t) := \xi\left(\frac{1}{2} + it\right).$$

**Theorem 1.1.** *Define*

$$\phi(x) := \psi(x) + \frac{1}{2x} - \log x. \tag{1.2}$$

*If  $\alpha$  and  $\beta$  are positive numbers such that  $\alpha\beta = 1$ , then*

$$\begin{aligned} \sqrt{\alpha} \left\{ \frac{\gamma - \log(2\pi\alpha)}{2\alpha} + \sum_{n=1}^{\infty} \phi(n\alpha) \right\} &= \sqrt{\beta} \left\{ \frac{\gamma - \log(2\pi\beta)}{2\beta} + \sum_{n=1}^{\infty} \phi(n\beta) \right\} \\ &= -\frac{1}{\pi^{3/2}} \int_0^{\infty} \left| \Xi\left(\frac{1}{2}t\right) \Gamma\left(\frac{-1+it}{4}\right) \right|^2 \frac{\cos\left(\frac{1}{2}t \log \alpha\right)}{1+t^2} dt, \end{aligned} \tag{1.3}$$

*where  $\gamma$  denotes Euler’s constant and  $\Xi(x)$  denotes Riemann’s  $\Xi$ -function.*

The first identity in (1.3) is beautiful in its elegant symmetry and surprising as well because why would subtracting the two leading terms in the asymptotic expansion of the logarithmic derivative of the Gamma function, in order to gain convergence of the infinite series on the left side, yield a “modular relation” for the resulting function? The second identity in (1.3) is also surprising, for why would the first identity foreshadow a connection with the Riemann zeta function in the second?

Although Ramanujan does not provide a proof of (1.3), he does indicate that (1.3) “can be deduced from”

$$\int_0^{\infty} (\psi(1+x) - \log x) \cos(2\pi nx) dx = \frac{1}{2} (\psi(1+n) - \log n). \tag{1.4}$$

This latter result was rediscovered by A.P. Guinand [10] in 1947, and he later found a simpler proof of this result in [11]. In a footnote at the end of his paper [11], Guinand remarks that T.A. Brown had told him that he himself had proved the self-reciprocity of  $\psi(1+x) - \log x$  some years ago, and that when he (Brown) communicated the result to G.H. Hardy, Hardy told him that the result was also given by Ramanujan in a progress report to the University of Madras, but was not published elsewhere. However, we cannot find this result in any of the three *Quarterly Reports* that Ramanujan submitted to the University of Madras [3, 18]. Therefore, Hardy's memory was perhaps imperfect; it would appear that he saw (1.4) in the aforementioned manuscript that Watson had copied. On the other hand, the only copy of Ramanujan's *Quarterly Reports* that exists is in Watson's handwriting! It could be that the manuscript on pages 219–220 of [17], which is also in Watson's handwriting, was somehow separated from the original *Quarterly Reports*, and therefore that Hardy was indeed correct in his assertion!

The first equality in (1.3) was rediscovered by Guinand in [10] and appears in a footnote on the last page of his paper [10, p. 18]. It is interesting that Guinand remarks, "This formula also seems to have been overlooked." Here then is one more instance in which a mathematician thought that his or her theorem was new, but unbeknownst to the claimant, Ramanujan had beaten her/him to the punch! We now give Guinand's version of (1.3).

**Theorem 1.2.** *For any complex  $z$  such that  $|\arg z| < \pi$ , we have*

$$\begin{aligned} & \sum_{n=1}^{\infty} \left( \frac{\Gamma'}{\Gamma}(nz) - \log nz + \frac{1}{2nz} \right) + \frac{1}{2z} (\gamma - \log 2\pi z) \\ &= \frac{1}{z} \sum_{n=1}^{\infty} \left( \frac{\Gamma'}{\Gamma} \left( \frac{n}{z} \right) - \log \frac{n}{z} + \frac{z}{2n} \right) + \frac{1}{2} \left( \gamma - \log \frac{2\pi}{z} \right). \end{aligned} \quad (1.5)$$

The first equality in (1.3) can be easily obtained from Guinand's version by multiplying both sides of (1.5) by  $\sqrt{z}$  and then letting  $z = \alpha$  and  $1/z = \beta$ . Although not offering a proof of (1.5) in [10], Guinand did remark that it can be obtained by using an appropriate form of Poisson's summation formula, namely the form given in Theorem 1 in [9]. Later, Guinand gave another proof of Theorem 1.2 in [11], while also giving extensions of (1.5) involving derivatives of the  $\psi$ -function. He also established a finite version of (1.5) in [12]. However, Guinand apparently did not discover the connection of his work with Ramanujan's integral involving Riemann's  $\Xi$ -function.

In this paper, we first provide a proof of both identities in Theorem 1.1. In Sect. 4, we construct a second proof of (1.5) along the lines suggested by Guinand in [10]. We can also provide another proof of (1.3) employing both (1.4) and

$$\int_0^{\infty} \left( \frac{1}{e^{2\pi x} - 1} - \frac{1}{2\pi x} \right) e^{-2\pi n x} dx = \frac{1}{2\pi} (\log n - \psi(1+n)), \quad (1.6)$$



which can be derived from an integral evaluation in [8, p. 377, formula 3.427, no. 7]. However, this proof is similar but slightly more complicated than the first proof that we provide below. The second author has obtained two additional proofs in [6] and [7]. In the proof in [6], (1.3) is obtained as a limiting case of a more general formula.

Although the Riemann zeta function appears at various instances throughout Ramanujan’s notebooks [15] and lost notebook [17], he only wrote one paper in which the zeta function plays the leading role [14], [16, pp. 72–77]. In fact, a result proved by Ramanujan in [14], namely (3.7) in Sect. 3 below, is a key to proving (1.3). About the integral involving Riemann’s  $\Xi$ -function in this result, Hardy [13] comments that “the properties of this integral resemble those of one which Mr. Littlewood and I have used, in a paper to be published shortly in the *Acta Mathematica*, to prove that

$$\int_{-T}^T \left| \zeta \left( \frac{1}{2} + ti \right) \right|^2 dt \sim \frac{2}{\pi} T \log T. \tag{1.7}$$

It is also interesting that on a page in the original lost notebook [17, p. 195], Ramanujan defines

$$\phi(x) := \psi(x) + \frac{1}{2x} - \gamma - \log x \tag{1.8}$$

and then concludes that (1.3) is valid. However, with the definition (1.8) of  $\phi(x)$ , the series in (1.3) do not converge. For a more complete discussion of Ramanujan’s incorrect claim, see [2].

## 2 Preliminary Results

We first collect several well-known theorems that we use in our proof. First, from [5, p. 191], for  $t \neq 0$ ,

$$\sum_{n=1}^{\infty} \frac{1}{t^2 + 4n^2\pi^2} = \frac{1}{2t} \left( \frac{1}{e^t - 1} - \frac{1}{t} + \frac{1}{2} \right). \tag{2.1}$$

Second, from [18, p. 251], we find that, for  $\text{Re } z > 0$ ,

$$\phi(z) = -2 \int_0^{\infty} \frac{t dt}{(t^2 + z^2)(e^{2\pi t} - 1)}. \tag{2.2}$$

Third, we require Binet’s integral for  $\log \Gamma(z)$ , i.e., for  $\text{Re } z > 0$  [18, p. 249], [8, p. 377, formula 3.427, no. 4],

$$\log \Gamma(z) = \left( z - \frac{1}{2} \right) \log z - z + \frac{1}{2} \log(2\pi) + \int_0^{\infty} \left( \frac{1}{2} - \frac{1}{t} + \frac{1}{e^t - 1} \right) \frac{e^{-zt}}{t} dt. \tag{2.3}$$

Fourth, from [8, p. 377, formula 3.427, no. 2], we find that

$$\int_0^{\infty} \left( \frac{1}{1 - e^{-x}} - \frac{1}{x} \right) e^{-x} dx = \gamma, \quad (2.4)$$

where  $\gamma$  denotes Euler's constant. Fifth, by Frullani's integral [8, p. 378, formula 3.434, no. 2],

$$\int_0^{\infty} \frac{e^{-\mu x} - e^{-\nu x}}{x} dx = \log \frac{\nu}{\mu}, \quad \mu, \nu > 0. \quad (2.5)$$

### 3 First Proof of Theorem 1.1

*Proof.* Our first goal is to establish an integral representation for the far left side of (1.3). Replacing  $z$  by  $n\alpha$  in (2.2) and summing on  $n$ ,  $1 \leq n < \infty$ , we find, by absolute convergence, that

$$\begin{aligned} \sum_{n=1}^{\infty} \phi(n\alpha) &= -2 \sum_{n=1}^{\infty} \int_0^{\infty} \frac{t dt}{(t^2 + n^2\alpha^2)(e^{2\pi t} - 1)} \\ &= \frac{-2}{\alpha^2} \int_0^{\infty} \frac{t dt}{(e^{2\pi t} - 1)} \sum_{n=1}^{\infty} \frac{1}{(t/\alpha)^2 + n^2}. \end{aligned} \quad (3.1)$$

Invoking (2.1) in (3.1), we see that

$$\sum_{n=1}^{\infty} \phi(n\alpha) = -\frac{2\pi}{\alpha} \int_0^{\infty} \frac{1}{(e^{2\pi t} - 1)} \left( \frac{1}{e^{2\pi t/\alpha} - 1} - \frac{\alpha}{2\pi t} + \frac{1}{2} \right) dt. \quad (3.2)$$

Next, setting  $x = 2\pi t$  in (2.4), we readily find that

$$\gamma = \int_0^{\infty} \left( \frac{2\pi}{e^{2\pi t} - 1} - \frac{e^{-2\pi t}}{t} \right) dt. \quad (3.3)$$

By Frullani's integral (2.5),

$$\int_0^{\infty} \frac{e^{-t/\alpha} - e^{-2\pi t}}{t} dt = \log \left( \frac{2\pi}{1/\alpha} \right) = \log(2\pi\alpha). \quad (3.4)$$

Combining (3.3) and (3.4), we arrive at

$$\gamma - \log(2\pi\alpha) = \int_0^{\infty} \left( \frac{2\pi}{e^{2\pi t} - 1} - \frac{e^{-t/\alpha}}{t} \right) dt. \quad (3.5)$$

Hence, from (3.2) and (3.5), we deduce that

$$\begin{aligned}
 & \sqrt{\alpha} \left( \frac{\gamma - \log(2\pi\alpha)}{2\alpha} + \sum_{n=1}^{\infty} \phi(n\alpha) \right) \\
 &= \frac{1}{2\sqrt{\alpha}} \int_0^{\infty} \left( \frac{2\pi}{e^{2\pi t} - 1} - \frac{e^{-t/\alpha}}{t} \right) dt \\
 & \quad - \frac{2\pi}{\sqrt{\alpha}} \int_0^{\infty} \frac{1}{(e^{2\pi t} - 1)} \left( \frac{1}{e^{2\pi t/\alpha} - 1} - \frac{\alpha}{2\pi t} + \frac{1}{2} \right) dt \\
 &= \int_0^{\infty} \left( \frac{\sqrt{\alpha}}{t(e^{2\pi t} - 1)} - \frac{2\pi}{\sqrt{\alpha}(e^{2\pi t/\alpha} - 1)(e^{2\pi t} - 1)} - \frac{e^{-t/\alpha}}{2t\sqrt{\alpha}} \right) dt. \tag{3.6}
 \end{aligned}$$

Now from [14, p. 260, (22)] or [16, p. 77], for  $n$  real,

$$\begin{aligned}
 & \int_0^{\infty} \Gamma\left(\frac{-1+it}{4}\right) \Gamma\left(\frac{-1-it}{4}\right) \left(\Xi\left(\frac{1}{2}t\right)\right)^2 \frac{\cos nt}{1+t^2} dt \\
 &= \int_0^{\infty} \left| \Xi\left(\frac{1}{2}t\right) \Gamma\left(\frac{-1+it}{4}\right) \right|^2 \frac{\cos nt}{1+t^2} dt \\
 &= \pi^{3/2} \int_0^{\infty} \left( \frac{1}{e^{xe^n} - 1} - \frac{1}{xe^n} \right) \left( \frac{1}{e^{xe^{-n}} - 1} - \frac{1}{xe^{-n}} \right) dx. \tag{3.7}
 \end{aligned}$$

Letting  $n = \frac{1}{2} \log \alpha$  and  $x = 2\pi t / \sqrt{\alpha}$  in (3.7), we deduce that

$$\begin{aligned}
 & -\frac{1}{\pi^{3/2}} \int_0^{\infty} \left| \Xi\left(\frac{1}{2}t\right) \Gamma\left(\frac{-1+it}{4}\right) \right|^2 \frac{\cos(\frac{1}{2}t \log \alpha)}{1+t^2} dt \\
 &= -\frac{2\pi}{\sqrt{\alpha}} \int_0^{\infty} \left( \frac{1}{e^{2\pi t} - 1} - \frac{1}{2\pi t} \right) \left( \frac{1}{e^{2\pi t/\alpha} - 1} - \frac{\alpha}{2\pi t} \right) dt \\
 &= \int_0^{\infty} \left( \frac{-2\pi/\sqrt{\alpha}}{(e^{2\pi t/\alpha} - 1)(e^{2\pi t} - 1)} + \frac{\sqrt{\alpha}}{t(e^{2\pi t} - 1)} \right. \\
 & \quad \left. + \frac{1}{t\sqrt{\alpha}(e^{2\pi t/\alpha} - 1)} - \frac{\sqrt{\alpha}}{2\pi t^2} \right) dt. \tag{3.8}
 \end{aligned}$$

Hence, combining (3.6) and (3.8), in order to prove that the far left side of (1.3) equals the far right side of (1.3), we see that it suffices to show that

$$\begin{aligned}
 & \int_0^{\infty} \left( \frac{1}{t\sqrt{\alpha}(e^{2\pi t/\alpha} - 1)} - \frac{\sqrt{\alpha}}{2\pi t^2} + \frac{e^{-t/\alpha}}{2t\sqrt{\alpha}} \right) dt \\
 &= \frac{1}{\sqrt{\alpha}} \int_0^{\infty} \left( \frac{1}{u(e^u - 1)} - \frac{1}{u^2} + \frac{e^{-u/(2\pi)}}{2u} \right) du = 0, \tag{3.9}
 \end{aligned}$$

where we made the change of variable  $u = 2\pi t/\alpha$ . In fact, more generally, we show that

$$\int_0^\infty \left( \frac{1}{u(e^u - 1)} - \frac{1}{u^2} + \frac{e^{-ua}}{2u} \right) du = -\frac{1}{2} \log(2\pi a) \tag{3.10}$$

so that if we set  $a = 1/(2\pi)$  in (3.10), we deduce (3.9).

Consider the integral, for  $t > 0$ ,

$$\begin{aligned} F(a, t) &:= \int_0^\infty \left\{ \left( \frac{1}{e^u - 1} - \frac{1}{u} + \frac{1}{2} \right) \frac{e^{-tu}}{u} + \frac{e^{-ua} - e^{-tu}}{2u} \right\} du \\ &= \log \Gamma(t) - \left( t - \frac{1}{2} \right) \log t + t - \frac{1}{2} \log(2\pi) + \frac{1}{2} \log \frac{t}{a}, \end{aligned} \tag{3.11}$$

where we applied (2.3) and (2.5). Upon the integration of (1.1), it is easily gleaned that, as  $t \rightarrow 0$ ,

$$\log \Gamma(t) \sim -\log t - \gamma t,$$

where  $\gamma$  denotes Euler’s constant. Using this in (3.11), we find, upon simplification, that, as  $t \rightarrow 0$ ,

$$F(a, t) \sim -\gamma t - t \log t + t - \frac{1}{2} \log(2\pi) - \frac{1}{2} \log a.$$

Hence,

$$\lim_{t \rightarrow 0} F(a, t) = -\frac{1}{2} \log(2\pi a). \tag{3.12}$$

Letting  $t$  approach 0 in (3.11), taking the limit under the integral sign on the right-hand side using Lebesgue’s dominated convergence theorem, and employing (3.12), we immediately deduce (3.10). As previously discussed, this is sufficient to prove the equality of the first and third expressions in (1.3), namely,

$$\begin{aligned} &\sqrt{\alpha} \left\{ \frac{\gamma - \log(2\pi\alpha)}{2\alpha} + \sum_{n=1}^\infty \phi(n\alpha) \right\} \\ &= -\frac{1}{\pi^{3/2}} \int_0^\infty \left| \Xi \left( \frac{1}{2}t \right) \Gamma \left( \frac{-1 + it}{4} \right) \right|^2 \frac{\cos \left( \frac{1}{2}t \log \alpha \right)}{1 + t^2} dt. \end{aligned} \tag{3.13}$$

Lastly, using (3.13) with  $\alpha$  replaced by  $\beta$  and employing the relation  $\alpha\beta = 1$ , we conclude that

$$\begin{aligned} &\sqrt{\beta} \left\{ \frac{\gamma - \log(2\pi\beta)}{2\beta} + \sum_{n=1}^\infty \phi(n\beta) \right\} \\ &= -\frac{1}{\pi^{3/2}} \int_0^\infty \left| \Xi \left( \frac{1}{2}t \right) \Gamma \left( \frac{-1 + it}{4} \right) \right|^2 \frac{\cos \left( \frac{1}{2}t \log \beta \right)}{1 + t^2} dt \end{aligned}$$

$$\begin{aligned}
 &= -\frac{1}{\pi^{3/2}} \int_0^\infty \left| \Xi\left(\frac{1}{2}t\right) \Gamma\left(\frac{-1+it}{4}\right) \right|^2 \frac{\cos\left(\frac{1}{2}t \log(1/\alpha)\right)}{1+t^2} dt \\
 &= -\frac{1}{\pi^{3/2}} \int_0^\infty \left| \Xi\left(\frac{1}{2}t\right) \Gamma\left(\frac{-1+it}{4}\right) \right|^2 \frac{\cos\left(\frac{1}{2}t \log \alpha\right)}{1+t^2} dt.
 \end{aligned}$$

Hence, the equality of the second and third expressions in (1.3) has been demonstrated, and so the proof is complete.  $\square$

### 4 Second Proof of (1.3)

In this section, we give our second proof of the first identity in (1.3) using Guinand’s generalization of Poisson’s summation formula in [9]. We emphasize that this route does not take us to the integral involving Riemann’s  $\Xi$ -function in the second identity of (1.3). First, we reproduce the needed version of the Poisson summation formula from Theorem 1 in [9].

**Theorem 4.1.** *If  $f(x)$  is an integral,  $f(x)$  tends to zero as  $x \rightarrow \infty$ , and  $xf'(x)$  belongs to  $L^p(0, \infty)$ , for some  $p, 1 < p \leq 2$ , then*

$$\lim_{N \rightarrow \infty} \left( \sum_{n=1}^N f(n) - \int_0^N f(t) dt \right) = \lim_{N \rightarrow \infty} \left( \sum_{n=1}^N g(n) - \int_0^N g(t) dt \right), \tag{4.1}$$

where

$$g(x) = 2 \int_0^{\rightarrow \infty} f(t) \cos(2\pi xt) dt. \tag{4.2}$$

Next, we state a lemma<sup>2</sup> that will subsequently be used in our proof of (1.3).

**Lemma 4.2.** *If  $\psi(x)$  is defined by (1.1), then*

$$\int_0^\infty \left( \psi(t+1) - \frac{1}{2(t+1)} - \log t \right) dt = \frac{1}{2} \log 2\pi. \tag{4.3}$$

*Proof.* Let  $I$  denote the integral on the left-hand side of (1.1). Then,

$$\begin{aligned}
 I &= \int_0^\infty \frac{d}{dt} \left( \log \frac{e^t \Gamma(t+1)}{t^t \sqrt{t+1}} \right) dt \\
 &= \lim_{t \rightarrow \infty} \log \frac{e^t \Gamma(t+1)}{t^t \sqrt{t+1}} - \lim_{t \rightarrow 0} \log \frac{e^t \Gamma(t+1)}{t^t \sqrt{t+1}}
 \end{aligned}$$

---

<sup>2</sup> The authors are indebted to M. L. Glasser for the proof of this lemma. The authors’ original proof of this lemma was substantially longer than Glasser’s given here.

$$\begin{aligned}
&= \log \lim_{t \rightarrow \infty} \frac{e^t \Gamma(t+1)}{t^t \sqrt{t+1}} - \log \left( \lim_{t \rightarrow 0} e^t \Gamma(t+1) \right) - \lim_{t \rightarrow 0} t \log t - \lim_{t \rightarrow 0} \frac{1}{2} \log(t+1) \\
&= \log \lim_{t \rightarrow \infty} \frac{e^t \Gamma(t+1)}{t^t \sqrt{t+1}}. \tag{4.4}
\end{aligned}$$

Next, Stirling's formula [8, p. 945, formula 8.327] tells us that

$$\Gamma(z) \sim \sqrt{2\pi} z^{z-1/2} e^{-z}, \tag{4.5}$$

as  $|z| \rightarrow \infty$  for  $|\arg z| \leq \pi - \delta$ , where  $0 < \delta < \pi$ . Hence, employing (4.5), we find that

$$\frac{e^t \Gamma(t+1)}{t^t \sqrt{t+1}} \sim \frac{\sqrt{2\pi}}{e} \left(1 + \frac{1}{t}\right)^t \tag{4.6}$$

so that

$$\lim_{t \rightarrow \infty} \frac{e^t \Gamma(t+1)}{t^t \sqrt{t+1}} = \sqrt{2\pi}. \tag{4.7}$$

Thus, from (4.4) and (4.7), we conclude that

$$I = \frac{1}{2} \log 2\pi. \tag{4.8}$$

□

Now we are ready to give our second proof of (1.3). We first prove it for  $\operatorname{Re} z > 0$ . Let

$$f(x) = \psi(xz + 1) - \log xz. \tag{4.9}$$

We show that  $f(x)$  satisfies the hypotheses of Theorem 4.1. From (1.4), we see that  $f(x)$  is an integral. Next, we need two formulas for  $\psi(x)$ . First, from [1, p. 259, formula 6.3.18], for  $|\arg z| < \pi$ , as  $z \rightarrow \infty$ ,

$$\psi(z) \sim \log z - \frac{1}{2z} - \frac{1}{12z^2} + \frac{1}{120z^4} - \frac{1}{252z^6} + \dots \tag{4.10}$$

Second, from [18, p. 250],

$$\psi'(z) = \sum_{n=0}^{\infty} \frac{1}{(z+n)^2}. \tag{4.11}$$

From (4.9) and (4.10), it follows that

$$f(x) \sim \frac{1}{2xz} - \frac{1}{12x^2z^2} + \frac{1}{120x^4z^4} - \frac{1}{252x^6z^6} + \dots \tag{4.12}$$

so that

$$\lim_{x \rightarrow \infty} f(x) = 0. \tag{4.13}$$

Next, we show that  $xf'(x)$  belongs to  $L^p(0, \infty)$  for some  $p$  such that  $1 < p \leq 2$ . Using (4.10), we find that, as  $x \rightarrow \infty$ ,

$$xf'(x) \sim -\frac{1}{2xz} \quad (4.14)$$

so that  $|xf'(x)|^p \sim (2x|z|)^{-p}$ . Thus, for  $p > 1$ , we see that  $xf'(x)$  is locally integrable near  $\infty$ . Also, using (4.11), we have

$$\begin{aligned} \lim_{x \rightarrow 0} xf'(x) &= \lim_{x \rightarrow 0} \left( xz \sum_{n=0}^{\infty} \frac{1}{(xz+n)^2} - \frac{1}{xz} - 1 \right) \\ &= \lim_{x \rightarrow 0} \left( xz \sum_{n=1}^{\infty} \frac{1}{(xz+n)^2} - 1 \right) \\ &= -1. \end{aligned} \quad (4.15)$$

This proves that  $xf'(x)$  is locally integrable near 0. Hence, we have shown that  $xf'(x)$  belongs to  $L^p(0, \infty)$  for some  $p$  such that  $1 < p \leq 2$ .

Now from (4.2) and (4.9), we find that

$$g(x) = 2 \int_0^{\infty} (\psi(tz+1) - \log tz) \cos(2\pi xt) dt.$$

Employing the change of variable  $y = tz$  and using (1.4), we find that

$$\begin{aligned} g(x) &= \frac{2}{z} \int_0^{\infty} (\psi(y+1) - \log y) \cos(2\pi xy/z) dy \\ &= \frac{1}{z} \left( \psi\left(\frac{x}{z} + 1\right) - \log\left(\frac{x}{z}\right) \right). \end{aligned} \quad (4.16)$$

Substituting the expressions for  $f(x)$  and  $g(x)$  from (4.9) and (4.16), respectively, in (4.1), we find that

$$\begin{aligned} &\lim_{N \rightarrow \infty} \left( \sum_{n=1}^N (\psi(nz+1) - \log nz) - \int_0^N (\psi(tz+1) - \log tz) dt \right) \\ &= \frac{1}{z} \left[ \lim_{N \rightarrow \infty} \left( \sum_{n=1}^N \left( \psi\left(\frac{n}{z} + 1\right) - \log \frac{n}{z} \right) - \int_0^N \left( \psi\left(\frac{t}{z} + 1\right) - \log \frac{t}{z} \right) dt \right) \right]. \end{aligned} \quad (4.17)$$

Thus,

$$\begin{aligned} & \lim_{N \rightarrow \infty} \left( \sum_{n=1}^N \left( \frac{\Gamma'}{\Gamma}(nz) + \frac{1}{2nz} - \log nz \right) + \sum_{n=1}^N \frac{1}{2nz} - \int_0^N (\psi(tz+1) - \log tz) dt \right) \\ &= \frac{1}{z} \left[ \lim_{N \rightarrow \infty} \left( \sum_{n=1}^N \left( \frac{\Gamma'}{\Gamma} \left( \frac{n}{z} + \frac{z}{2n} - \log \frac{n}{z} \right) + \sum_{n=1}^N \frac{z}{2n} \right. \right. \right. \\ & \quad \left. \left. \left. - \int_0^N \left( \psi \left( \frac{t}{z} + 1 \right) - \log \frac{t}{z} \right) dt \right) \right]. \end{aligned} \quad (4.18)$$

Now if we can show that

$$\lim_{N \rightarrow \infty} \left( \sum_{n=1}^N \frac{1}{2nz} - \int_0^N (\psi(tz+1) - \log tz) dt \right) = \frac{\gamma - \log 2\pi z}{2z}, \quad (4.19)$$

then replacing  $z$  by  $1/z$  in (4.19) will give us

$$\lim_{N \rightarrow \infty} \left( \sum_{n=1}^N \frac{z}{2n} - \int_0^N \left( \psi \left( \frac{t}{z} + 1 \right) - \log \frac{t}{z} \right) dt \right) = \frac{z(\gamma - \log(2\pi/z))}{2}. \quad (4.20)$$

Then substituting (4.19) and (4.20) in (4.18) will complete the proof of the theorem. To that end,

$$\begin{aligned} & \lim_{N \rightarrow \infty} \left( \sum_{n=1}^N \frac{1}{2nz} - \int_0^N (\psi(tz+1) - \log tz) dt \right) \\ &= \lim_{N \rightarrow \infty} \left( \frac{1}{2z} \left( \sum_{n=1}^N \frac{1}{n} - \log N \right) + \frac{\log N}{2z} - \int_0^N (\psi(tz+1) - \log tz) dt \right) \\ &= \frac{\gamma}{2z} + \lim_{N \rightarrow \infty} \left( -\frac{\log z}{2z} + \frac{\log Nz}{2z} - \int_0^N (\psi(tz+1) - \log tz) dt \right) \\ &= \frac{\gamma}{2z} - \frac{\log z}{2z} + \lim_{N \rightarrow \infty} \left( \frac{\log(Nz+1)}{2z} - \frac{1}{z} \int_0^{Nz} (\psi(t+1) - \log t) dt \right. \\ & \quad \left. - \frac{1}{2z} \log \left( 1 + \frac{1}{Nz} \right) \right) \\ &= \frac{\gamma}{2z} - \frac{\log z}{2z} + \frac{1}{z} \lim_{N \rightarrow \infty} \left( \frac{\log(Nz+1)}{2} - \int_0^{Nz} (\psi(t+1) - \log t) dt \right) \\ &= \frac{\gamma}{2z} - \frac{\log z}{2z} + \frac{1}{z} \lim_{N \rightarrow \infty} \left( \frac{1}{2} \int_0^{Nz} \frac{1}{t+1} dt - \int_0^{Nz} (\psi(t+1) - \log t) dt \right) \end{aligned}$$



$$\begin{aligned}
&= \frac{\gamma}{2z} - \frac{\log z}{2z} - \frac{1}{z} \lim_{N \rightarrow \infty} \int_0^{Nz} \left( \psi(t+1) - \frac{1}{2(t+1)} - \log t \right) dt \\
&= \frac{\gamma}{2z} - \frac{\log z}{2z} - \frac{1}{z} \int_0^{\infty} \left( \psi(t+1) - \frac{1}{2(t+1)} - \log t \right) dt \\
&= \frac{\gamma}{2z} - \frac{\log z}{2z} - \frac{\log 2\pi}{2z} \\
&= \frac{\gamma - \log 2\pi z}{2z}, \tag{4.21}
\end{aligned}$$

where in the antepenultimate line we have made use of Lemma 4.2. This completes the proof of (4.19) and hence the proof of Theorem 1.2 for  $\operatorname{Re} z > 0$ . But both sides of (1.5) are analytic for  $|\arg z| < \pi$ . Hence, by analytic continuation, the theorem is true for all complex  $z$  such that  $|\arg z| < \pi$ .

## References

- [1] M. Abramowitz and I.A. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, 1965.
- [2] G.E. Andrews and B.C. Berndt, *Ramanujan's Lost Notebook*, Part IV, Springer, New York, (in press).
- [3] B.C. Berndt, *Ramanujan's quarterly reports*, Bull. Lond. Math. Soc. **16** (1984), 449–489.
- [4] B.C. Berndt, *Ramanujan's Notebooks*, Part I, Springer, New York, 1985.
- [5] J.B. Conway, *Functions of One Complex Variable*, 2nd ed., Springer, New York, 1978.
- [6] A. Dixit, *Analogues of a transformation formula of Ramanujan*, submitted for publication.
- [7] A. Dixit, *Series transformations and integrals involving the Riemann  $\Xi$ -function*, J. Math. Anal. Appl. **368** (2010), 358–373.
- [8] I.S. Gradshteyn and I.M. Ryzhik, eds., *Table of Integrals, Series, and Products*, 5th ed., Academic, San Diego, 1994.
- [9] A.P. Guinand, *On Poisson's summation formula*, Ann. Math. (2) **42** (1941), 591–603.
- [10] A.P. Guinand, *Some formulae for the Riemann zeta-function*, J. Lond. Math. Soc. **22** (1947), 14–18.
- [11] A.P. Guinand, *A note on the logarithmic derivative of the Gamma function*, Edinb. Math. Notes **38** (1952), 1–4.
- [12] A.P. Guinand, *Some finite identities connected with Poisson's summation formula*, Proc. Edinb. Math. Soc. (2) **12** (1960), 17–25.
- [13] G.H. Hardy, *Note by G.H. Hardy on the preceding paper*, Quart. J. Math. **46** (1915), 260–261.
- [14] S. Ramanujan, *New expressions for Riemann's functions  $\xi(s)$  and  $\Xi(s)$* , Quart. J. Math. **46** (1915), 253–260.
- [15] S. Ramanujan, *Notebooks* (2 volumes), Tata Institute of Fundamental Research, Bombay, 1957.
- [16] S. Ramanujan, *Collected Papers*, Cambridge University Press, Cambridge, 1927; reprinted by Chelsea, New York, 1962; reprinted by the American Mathematical Society, Providence, RI, 2000.
- [17] S. Ramanujan, *The Lost Notebook and Other Unpublished Papers*, Narosa, New Delhi, 1988.
- [18] E.T. Whittaker and G.N. Watson, *A Course of Modern Analysis*, 4th ed., Cambridge University Press, Cambridge, 1966.

# Ternary Quadratic Forms, Modular Equations, and Certain Positivity Conjectures

Alexander Berkovich\* and William C. Jagy

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** We show that many of Ramanujan’s modular equations of degree 3 can be interpreted in terms of integral ternary quadratic forms. This way we establish that for any  $n \in \mathbf{N}$ ,

$$\left| \left\{ (x, y, z) \in \mathbf{Z}^3 : \frac{x(x+1)}{2} + y^2 + z^2 = n \right\} \right| \geq \left| \left\{ (x, y, z) \in \mathbf{Z}^3 : \frac{x(x+1)}{2} + 3y^2 + 3z^2 = n \right\} \right|,$$

just to name one among many similar “positivity” results of this type. In particular, we prove the recent conjecture of H. Yesilyurt and the first author, stating that for any  $n \in \mathbf{N}$ ,

$$\left| \left\{ (x, y, z) \in \mathbf{Z}^3 : \frac{x(x+1)}{2} + y^2 + z^2 = n \right\} \right| \geq \left| \left\{ (x, y, z) \in \mathbf{Z}^3 : \frac{x(x+1)}{2} + 7y^2 + 7z^2 = n \right\} \right|.$$

We prove a number of identities for certain ternary forms with discriminants 144, 400, 784, or 3,600 by converting every ternary identity into an identity for the appropriate  $\eta$ -quotients. In the process, we discover and prove a few new modular equations of degree 5 and 7. For any square free odd integer  $S$  with prime

---

\*Research of the first author was supported in part by NSA grant H98230-09-1-0051.

A. Berkovich  
Department of Mathematics, University of Florida, Gainesville, Florida 32611-8105  
e-mail: [alex@ufl.edu](mailto:alex@ufl.edu)

W.C. Jagy  
Math.Sci.Res.Inst., 17 Gauss Way, Berkeley, CA 94720-5070  
e-mail: [jagy@msri.org](mailto:jagy@msri.org)

factorization  $p_1 \dots p_r$ , we define the  $S$ -genus as a union of  $2^r$  specially selected genera of ternary quadratic forms, all with discriminant  $16S^2$ . This notion of  $S$ -genus arises naturally in the course of our investigation. It entails an interesting injection from genera of binary quadratic forms with discriminant  $-8S$  to genera of ternary quadratic forms with discriminant  $16S^2$ .

**Mathematics Subject Classification (2000)** Primary 11E20, 11F37, 11B65; Secondary 05A30, 33 E05

**Key words and phrases** Ternary quadratic forms ·  $S$ -genus · Modular functions · Modular equations ·  $\theta$ -functions ·  $\eta$ -quotients

## 1 Introduction

Alladi Ramakrishnan’s visits to Gainesville were always memorable. His interests were diverse, and his passion for science was truly amazing. He was a very open man, always happy to make new friends. He had so many stories to tell. His family was a true pillar of strength for him and in turn he was devoted to them.

Ramanujan’s general theta-function  $f(a, b)$  is defined by

$$f(a, b) = \sum_{n=-\infty}^{\infty} a^{\frac{(n-1)n}{2}} b^{\frac{(n+1)n}{2}}, \quad |ab| < 1. \tag{1.1}$$

In Ramanujan’s notation, the celebrated Jacobi triple product identity takes the shape

$$f(a, b) = (-a; ab)_{\infty} (-b; ab)_{\infty} (ab; ab)_{\infty}, \quad |ab| < 1, \tag{1.2}$$

with

$$(a; q)_{\infty} := \prod_{j \geq 0} (1 - aq^j).$$

It is always assumed that  $|q| < 1$ . The following four special cases will play a prominent role in our narrative

$$\phi(q) := f(q, q) = \sum_{n=-\infty}^{\infty} q^{n^2}, \tag{1.3}$$

$$\psi(q) := f(q, q^3) = \sum_{n \geq 0} q^{\frac{(n+1)n}{2}}, \tag{1.4}$$

$$f(q, q^2) = \sum_{n=-\infty}^{\infty} q^{\frac{(3n+1)n}{2}}, \tag{1.5}$$

$$f(q, q^5) = \sum_{n=-\infty}^{\infty} q^{(3n+2)n}. \tag{1.6}$$

Using (1.2), it is not hard to derive the product representation formulas

$$\phi(q) = \frac{E(q^2)^5}{E(q)^2 E(q^4)^2}, \tag{1.7}$$

$$\psi(q) = \frac{E(q^2)^2}{E(q)}, \tag{1.8}$$

$$f(q, q^2) = \frac{E(q^3)^2 E(q^2)}{E(q^6) E(q)}, \tag{1.9}$$

$$f(q, q^5) = \frac{E(q^{12}) E(q^3) E(q^2)^2}{E(q^6) E(q^4) E(q)}, \tag{1.10}$$

where

$$E(q) := \prod_{j \geq 1} (1 - q^j).$$

Combining (1.7) and (1.8), we see that

$$\psi(q)^2 = \phi(q)\psi(q^2). \tag{1.11}$$

Note that the Dedekind  $\eta(z)$  is related to  $E(q)$  as

$$\eta(z) = q^{\frac{1}{24}} E(q), \quad \text{if } q = \exp(2\pi iz) \quad \text{with } \text{Im}(z) > 0. \tag{1.12}$$

And so  $\phi(q)$ ,  $q^{\frac{1}{8}}\psi(q)$ ,  $q^{\frac{1}{24}}f(q, q^2)$ ,  $q^{\frac{1}{3}}f(q, q^5)$  all have  $\eta$ -quotient representations.

Following [2], we say that a  $q$ -series is positive if its power series coefficients are all non-negative. We define  $P[q]$  to be the set of all such series. It is plain that  $\phi(q)$ ,  $\psi(q)$ ,  $f(q, q^2)$ ,  $f(q, q^5)$ ,  $\frac{1}{E(q)}$  (and their products) are in  $P[q]$ . However, it is not at all obvious that

$$\psi(q)(\phi(q)^2 - \phi(q^7)^2) \in P[q]. \tag{1.13}$$

Motivated by their studies of 7-core partitions, H. Yesilyurt and the first author conjectured (1.13) in ((6.2), [2]). The reader should be cautioned that other similar conjectures there: (6.1), (6.3), and (6.4) are false. What makes (1.13) somewhat non-trivial is the fact that it is not true that  $\phi(q)^2 - \phi(q^7)^2 \in P[q]$ . However, in Sect. 3, we shall prove

**Theorem 1.1.** *The following identities are true*

$$\begin{aligned} \psi(q)(\phi(q)^2 - \phi(q^7)^2) &= 4q\psi(q^2)\psi(q^7)\phi(q^7) + 8q^2\psi(q^{14})\psi(q^3)\phi(q^{21}) \\ &\quad + 8q^4\psi(q^{14})f(q, q^2)f(q^7, q^{35}), \end{aligned} \tag{1.14}$$

and

$$7\phi(q^7)^2\psi(q^7) = 8q^2\psi(q^2)\psi(q^{21})\phi(q^3) + 8\psi(q^2)f(q, q^5)f(q^7, q^{14}) - 4q\psi(q^{14})\psi(q)\phi(q) - \psi(q^7)\phi(q)^2. \tag{1.15}$$

One does not have to be very perceptive to deduce that the right hand side of (1.14) is in  $P[q]$ . Hence, (1.13) follows. Our proof of (1.14) makes naive use of the theory of modular forms. We employ certain of Ramanujan’s modular equations of degree 7 [1] to derive (1.15) from (1.14). Also in Sect. 3, we provide the following beautiful interpretation of Theorem 1.1 in terms of integral ternary quadratic forms.

**Theorem 1.2.** *If  $M \equiv 1 \pmod{8}$ , then*

$$(1, 8, 8, 0, 0, 0)(M) = (1, 14, 14, 0, 0, 0)(M) + 2(2, 7, 14, 0, 0, 0)(M) + 4(3, 5, 14, 0, 0, 2)(M). \tag{1.16}$$

*If  $M \equiv 1 \pmod{8}$  and  $7|M$ , then*

$$7(1, 8, 8, 0, 0, 0) \left( \frac{M}{7^2} \right) = - (1, 14, 14, 0, 0, 0)(M) - 2(2, 7, 14, 0, 0, 0)(M) + 4(3, 5, 14, 0, 0, 2)(M), \tag{1.17}$$

where here and everywhere

$$(a, b, c, d, e, f)(M) := |\{(x, y, z) \in \mathbf{Z}^3 : ax^2 + by^2 + cz^2 + dyz + ezx + fxy = M\}|.$$

We also have the following

**Corollary 1.3.** *If  $M \equiv 1 \pmod{8}$  and  $(M|7) = 1$ , then*

$$(1, 8, 8, 0, 0, 0)(M) = (1, 14, 14, 0, 0, 0)(M) + 2(2, 7, 14, 0, 0, 0)(M).$$

*If  $M \equiv 1 \pmod{8}$ ,  $7 \parallel M$ , then*

$$(1, 8, 8, 0, 0, 0)(M) = 2(1, 14, 14, 0, 0, 0)(M) + 4(2, 7, 14, 0, 0, 0)(M).$$

A few remarks are in order. We use the convention that Jacobi’s symbol  $(M|a) = 0$ , whenever  $(M, a) > 1$ . The notation  $p \parallel n$  means that  $p|n$  but it is not true that  $p^2|n$ . A slightly different version of this corollary was communicated to us by Benjamin Kane. His observation was crucial to our investigation. We understand that Kane used Siegel’s weighted average theorem [8] together with some local calculations of Jones [9]. We note that the Corollary 1.3 has a twin:

**Corollary 1.4.** *If  $M \equiv 1 \pmod{8}$  and  $(M|7) = -1$ , then*

$$(1, 8, 8, 0, 0, 0)(M) = 4(3, 5, 14, 0, 0, 2)(M).$$

*If  $M \equiv 1 \pmod{8}$ ,  $7 \parallel M$ , then*

$$(1, 8, 8, 0, 0, 0)(M) = 8(3, 5, 14, 0, 0, 2)(M).$$

There is nothing very special about the exponent 7 in (1.13). In a future paper, we plan to prove that for any  $S \in \mathbf{N}$ ,

$$\psi(q)(\phi(q)^2 - \phi(q^S)^2) \in P[q]. \tag{1.18}$$

In this paper, we discuss in great detail  $S = 3, 5, 7, 15$ . In these cases, we will construct and prove  $\eta$ -quotient identities that imply appropriate positivity results. For  $S = 3$  and  $5$ , our identities can be written concisely as modular equations of degree 3 and 5, respectively. Ramanujan found an astounding number of modular equations of degree 3 and 5. These are collected and proven in [1]. The results there are sufficient to prove everything needed for our treatment of  $S = 3, 5$  in Sects. 2 and 3. In Sect. 4, we prove our Theorems 1.1 and 1.2. Section 5 deals with the  $S = 15$  case. In Sect. 6, we define an injective map from genera of binary quadratic forms to genera of ternary quadratic forms. This map allows us to introduce a very useful notion of  $S$ -genus.

## 2 Ramanujan’s Modular Equations of Degree 3 and Associated Identities for Ternary Quadratic Forms with Discriminant 144

Following Ramanujan, we define the multiplier  $m$  of degree  $n$  as

$$m := m(n, q) = \frac{\phi(q)^2}{\phi(q^n)^2}, \tag{2.1}$$

and

$$\alpha := \alpha(q) = 1 - \frac{\phi(-q)^4}{\phi(q)^4}, \tag{2.2}$$

$$\beta := \beta(n, q) = \alpha(q^n). \tag{2.3}$$

We often say that  $\beta$  has degree  $n$  over  $\alpha$ . We also call an algebraic relation connecting  $m$ ,  $\alpha$ , and  $\beta$  a modular equation of degree  $n$ . It is well-known, page 40, [1] that

$$\phi(q) = \phi(q^4) + 2q\psi(q^8), \tag{2.4}$$

$$\phi(q)^4 - \phi(-q)^4 = 16q\psi(q^2)^4. \tag{2.5}$$

The last equation implies another formula for  $\alpha$

$$\alpha = 16q \frac{\psi(q^2)^4}{\phi(q)^4}, \quad (2.6)$$

which will come in handy later. Pages 230–237 in [1] contain an impressive collection of 15 of Ramanujan’s modular equations of degree 3 together with succinct proofs. In particular, one can find there what amounts to the following:

**Lemma 2.1.** *If*

$$\alpha = \frac{p(2+p)^3}{(1+2p)^3},$$

*then*

$$\beta(3, q) = \frac{p^3(2+p)}{(1+2p)},$$

*and*

$$m(3, q) = 1 + 2p.$$

We comment that this lemma is a very efficient tool for verifying any modular equations of degree 3. In particular, we see that

$$m - 1 = 2 \frac{\beta^{\frac{3}{8}}}{\alpha^{\frac{1}{8}}}. \quad (2.7)$$

From (1.11), (2.1), (2.3), (2.6), and (2.7), it is readily shown that

$$\frac{\phi(q)^2}{\phi(q^3)^2} - 1 = 4q \frac{\psi(q)\psi(q^3)\psi(q^6)}{\psi(q^2)\phi(q^3)^2}. \quad (2.8)$$

Clearly, this theta-function identity can be rewritten as

$$\psi(q^2)\phi(q)^2 = \psi(q^2)\phi(q^3)^2 + 4q\psi(q)\psi(q^3)\psi(q^6). \quad (2.9)$$

Next, we multiply both sides of (2.9) by  $\frac{\psi(q)}{\psi(q^2)}$  and employ (1.11) again to deduce that

$$\psi(q)\phi(q)^2 = \psi(q)\phi(q^3)^2 + 4q\psi(q^3)\psi(q^6)\phi(q). \quad (2.10)$$

The truth of

$$\psi(q)(\phi(q)^2 - \phi(q^3)^2) \in P[q], \quad (2.11)$$

and of

$$\psi(q^2)(\phi(q)^2 - \phi(q^3)^2) \in P[q] \quad (2.12)$$

is now evident. Remarkably, we can interpret (2.9) and (2.10) in terms of integral ternary quadratic forms with discriminant 144. We remind the reader that the discriminant of a ternary form  $ax^2 + by^2 + cz^2 + dyz + ezx + fxy$  is defined as

$$\frac{1}{2} \det \begin{bmatrix} 2a & f & e \\ f & 2b & d \\ e & d & 2c \end{bmatrix}.$$

To this end, we define a sifting operator  $S_{t,s}$  by its action on power series as follows:

$$S_{t,s} \sum_{n \geq 0} c(n)q^n := \sum_{k \geq 0} c(tk + s)q^k, \tag{2.13}$$

where  $t, s$  are integers and  $0 \leq s < t$ . Observe that (2.4) implies that

$$S_{8,1}\phi(q)\phi(q^8)^2 = 2\psi(q)\phi(q)^2, \tag{2.14}$$

$$S_{8,1}\phi(q)\phi(q^6)^2 = 2\psi(q)\phi(q^3)^2, \tag{2.15}$$

$$S_{8,1}\phi(q^2)\phi(q^3)\phi(q^6) = 4q\phi(q)\psi(q^3)\psi(q^6). \tag{2.16}$$

Employing (2.14), (2.15), (2.16) together with (2.10) we see that

$$S_{8,1}(\phi(q)\phi(q^8)^2 - \phi(q)\phi(q^6)^2 - 2\phi(q^2)\phi(q^3)\phi(q^6)) = 0. \tag{2.17}$$

But the above is nothing else but the statement that

$$(1, 8, 8, 0, 0, 0)(M) = (1, 6, 6, 0, 0, 0)(M) + 2(2, 3, 6, 0, 0, 0)(M), \tag{2.18}$$

for any  $M \equiv 1 \pmod{8}$ . Actually, with the aid of (2.4), we can easily check that

$$S_{8,r}\phi(q)\phi(q^8)^2 = \frac{1}{3}S_{8,r}\phi(q)^3 \tag{2.19}$$

with  $r = 1, 7$ . Hence, (2.18) may be stated as

$$\frac{1}{3}(1, 1, 1, 0, 0, 0)(M) = (1, 6, 6, 0, 0, 0)(M) + 2(2, 3, 6, 0, 0, 0)(M), \tag{2.20}$$

for any  $M \equiv 1 \pmod{8}$ . It is very likely that modular equation (2.7) was known to Legendre and Jacobi. Surprisingly, the quadratic form interpretation given in (2.20) above appears to be new. We note that the two ternary forms  $x^2 + 6y^2 + 6z^2$  and  $2x^2 + 3y^2 + 6z^2$  on the right of (2.20) have the same discriminant = 144. Moreover, these two forms have class number = 1. This means that these forms belong to different genera and that they are both regular [7, 9, 10]. Moreover, it is easy to see that  $(-n_1|3) = -1$  for any integer  $n_1$  represented by  $x^2 + 6y^2 + 6z^2$ ,  $\gcd(n_1, 3) = 1$  and that  $(-n_2|3) = 1$  for any integer  $n_2$  represented by  $2x^2 + 3y^2 + 6z^2$ ,  $\gcd(n_2, 3) = 1$ .



We remark that the appearance of at least two genera with the same discriminant is the salient feature of all our ternary form identities. Somewhat anticipating developments in Sect. 6, we would like to comment that one can obtain the two ternary forms on the right of (2.20) starting with binary forms of discriminant  $-24$ . There are just two genera of binary quadratic forms with this discriminant: the (proper equivalence) class of  $x^2 + 6y^2$  and the class of  $2x^2 + 3y^2$  (See [6], pages 52–54). All we need to do to obtain our desired ternaries is to add  $6z^2$  to  $x^2 + 6y^2$  and  $2x^2 + 3y^2$ , respectively. Actually, we can extend (2.20) a bit as

$$\frac{1}{3}(1, 1, 1, 0, 0, 0)(M) = (1, 6, 6, 0, 0, 0)(M) + 2(2, 3, 6, 0, 0, 0)(M), \quad (2.21)$$

for any  $M \equiv 1, 2 \pmod{4}$ . To this end, we make repeated use of (2.4) and confirm that

$$S_{4,1}\phi(q)^3 = 6\psi(q^2)\phi(q)^2, \quad (2.22)$$

$$S_{4,1}\phi(q)\phi(q^6)^2 = 2\psi(q^2)\phi(q^3)^2, \quad (2.23)$$

$$S_{4,1}\phi(q^2)\phi(q^3)\phi(q^6) = 4q\psi(q)\psi(q^3)\psi(q^6) \quad (2.24)$$

$$S_{4,2}\phi(q)^3 = 12\phi(q)\psi(q^2)^2, \quad (2.25)$$

$$S_{4,2}\phi(q)\phi(q^6)^2 = 4q\psi(q^6)^2\phi(q), \quad (2.26)$$

$$S_{4,2}\phi(q^2)\phi(q^3)\phi(q^6) = 2\psi(q)\psi(q^3)\phi(q^3). \quad (2.27)$$

Next, we combine (2.9) and (2.22)–(2.24), to arrive at

$$S_{4,1} \left( \frac{1}{3}\phi(q)^3 - \phi(q)\phi(q^6)^2 - 2\phi(q^2)\phi(q^3)\phi(q^6) \right) = 0,$$

which is, essentially, the case  $M \equiv 1 \pmod{4}$  in (2.21). To see that (2.21) is also valid when  $M \equiv 2 \pmod{4}$ , we again use Lemma 2.1 to verify our next modular equation of degree 3

$$m = \frac{\beta^{\frac{1}{2}}}{\alpha^{\frac{1}{2}}} + 2\frac{\beta^{\frac{1}{8}}}{\alpha^{\frac{3}{8}}}. \quad (2.28)$$

Indeed, expressing everything in terms of  $p$  and simplifying, we obtain the trivial identity

$$1 + 2p = \frac{p(1 + 2p)}{2 + p} + 2\frac{1 + 2p}{2 + p}.$$

Hence, the proof of (2.28) is complete. The theta-function identity associated with (2.28) takes the pleasant form

$$\phi(q)\psi(q^2)^2 = \psi(q)\psi(q^3)\phi(q^3) + q\phi(q)\psi(q^6)^2. \quad (2.29)$$

Hence,

$$\phi(q)(\psi(q^2)^2 - q\psi(q^6)^2) \in P[q].$$

We observe that  $\psi(q^2)^2 - q\psi(q^6)^2 \notin P[q]$ . Again, we combine (2.25)–(2.27) and (2.29) to obtain

$$S_{4,2} \left( \frac{1}{3}\phi(q)^3 - \phi(q)\phi(q^6)^2 - 2\phi(q^2)\phi(q^3)\phi(q^6) \right) = 0,$$

which is essentially the case  $M \equiv 2 \pmod{4}$  in (2.21). This is not the end of the story, however. We discovered that (2.18) has an attractive companion

$$3(1, 8, 8, 0, 0, 0) \left( \frac{M}{3^2} \right) = -(1, 6, 6, 0, 0, 0)(M) + 2(2, 3, 6, 0, 0, 0)(M), \quad (2.30)$$

where  $M \equiv 1 \pmod{8}$ ,  $3|M$ . To prove it, we begin with the modular equation

$$\frac{3}{m} + 1 = 2 \frac{\alpha^{\frac{3}{8}}}{\beta^{\frac{1}{8}}}, \quad (2.31)$$

which can be routinely verified with the aid of Lemma 2.1. Next, we use (2.1), (2.3), and (2.6) to convert (2.31) into the theta-function identity

$$-\psi(q^3)\phi(q^2)^2 + 4\psi(q)\psi(q^2)\phi(q^3) = 3\psi(q^3)\phi(q^3)^2. \quad (2.32)$$

With a bit of labor, we can show that (2.32) is equivalent to

$$S_{24,9}(-\phi(q)\phi(q^6)^2 + 2\phi(q^2)\phi(q^3)\phi(q^6)) = 6\psi(q^3)\phi(q^3)^2.$$

Hence,

$$-(1, 6, 6, 0, 0, 0)(24n + 9) + 2(2, 3, 6, 0, 0, 0)(24n + 9) = 3 \left| \left\{ (x, y, z) \in \mathbf{Z}^3 : 3x^2 + 3y^2 + 3\frac{(1+z)z}{2} = n \right\} \right|.$$

Observe that

$$3 \left| \left\{ (x, y, z) \in \mathbf{Z}^3 : 3x^2 + 3y^2 + 3\frac{(1+z)z}{2} = n \right\} \right| = 3 \left| \left\{ (x, y, z) \in \mathbf{Z}^3 : 8x^2 + 8y^2 + z^2 = 1 + \frac{8n}{3} \right\} \right|.$$

And so we have completed the proof of (2.30). We note the following interesting corollary

$$(1, 6, 6, 0, 0, 0)(M) = 2(2, 3, 6, 0, 0, 0)(M), \quad (2.33)$$

when  $M \equiv 1 \pmod{8}$ ,  $3 \parallel M$ . Recalling (2.18), we see that

$$(1, 8, 8, 0, 0, 0)(M) = 2(1, 6, 6, 0, 0, 0)(M), \tag{2.34}$$

with  $M \equiv 1 \pmod{8}$ ,  $3 \parallel M$ . Analogously, (2.21) has its own companion identity

$$(1, 1, 1, 0, 0, 0) \left( \frac{M}{3^2} \right) = -(1, 6, 6, 0, 0, 0)(M) + 2(2, 3, 6, 0, 0, 0)(M), \tag{2.35}$$

with  $M \equiv 1, 2 \pmod{4}$ ,  $3 \mid M$ . Since the argument is pretty similar, we confine ourselves to the following diagram

Lemma 2.1

↓

Modular equation:

$$3 + m \frac{\alpha^{\frac{1}{2}}}{\beta^{\frac{1}{2}}} = 2m \frac{\alpha^{\frac{1}{8}}}{\beta^{\frac{3}{8}}}.$$

↓

Theta-function identity

$$\psi(q^2)^2 \phi(q^3) - \psi(q)\psi(q^3)\phi(q) + 3q\psi(q^6)\psi(q^3)^2 = 0.$$

↓

Ternary identity

$$(1, 1, 1, 0, 0, 0) \left( \frac{M}{3^2} \right) = -(1, 6, 6, 0, 0, 0)(M) + 2(2, 3, 6, 0, 0, 0)(M),$$

with  $M \equiv 2 \pmod{4}$ ,  $3 \mid M$ .

Lemma 2.1

↓

Modular equation

$$3 + m = 2m \frac{\alpha^{\frac{3}{8}}}{\beta^{\frac{1}{8}}}.$$

↓

Theta-function identity

$$\psi(q^6)\phi(q)^2 - 4\psi(q)\psi(q^2)\psi(q^3) + 3\psi(q^3)^2\phi(q^3) = 0.$$

↓

Ternary identity

$$(1, 1, 1, 0, 0, 0) \left( \frac{M}{3^2} \right) = -(1, 6, 6, 0, 0, 0)(M) + 2(2, 3, 6, 0, 0, 0)(M)$$

with  $M \equiv 1 \pmod{4}$ ,  $3|M$ . We conclude this section by stating (2.18), (2.30) in a way that would suggest an elegant and straightforward generalization. Let  $|\text{Aut}(a, b, c, d, e, f)|$  denote the number of integral automorphs of a ternary form  $ax^2 + by^2 + cz^2 + dyz + ezx + fxy$ . It is easy to check that

$$\frac{16}{|\text{Aut}(1,6,6,0,0,0)|} = 1,$$

$$\frac{16}{|\text{Aut}(2,3,6,0,0,0)|} = 2.$$

And so we can rewrite the right hand side of (2.18) as a weighted average over two genera. This way it becomes

$$(1, 8, 8, 0, 0, 0)(M) = \frac{16(1, 6, 6, 0, 0, 0)(M)}{|\text{Aut}(1,6,6,0,0,0)|} + \frac{16(2, 3, 6, 0, 0, 0)(M)}{|\text{Aut}(2,3,6,0,0,0)|}, \tag{2.36}$$

with  $M \equiv 1 \pmod{8}$ . Analogously, (2.30) may be stated as

$$3(1, 8, 8, 0, 0, 0) \left( \frac{M}{3^2} \right) = (-n_1|3) \frac{16(1, 6, 6, 0, 0, 0)(M)}{|\text{Aut}(1,6,6,0,0,0)|} + (-n_2|3) \frac{16(2, 3, 6, 0, 0, 0)(M)}{|\text{Aut}(2,3,6,0,0,0)|}, \tag{2.37}$$

where  $M \equiv 1 \pmod{8}$ ,  $3|M$  and  $n_1, n_2$  are any integers prime to 3 that are represented by  $x^2 + 6y^2 + 6z^2$ ,  $2x^2 + 3y^2 + 6z^2$ , respectively.

### 3 Ramanujan’s Modular Equations of Degree 5 and Associated Identities for Ternary Quadratic Forms with Discriminant 400

If all we ever wanted was to show that

$$\psi(q)(\phi(q)^2 - \phi(q^5)^2) \in P[q], \tag{3.1}$$

we could be done in a second. Indeed, using the simple identity  $5(x^2 + y^2) = (x - 2y)^2 + (y + 2x)^2$ , we find that for any  $n \in \mathbf{N}$ ,

$$\begin{aligned} &|\{(x, y) \in \mathbf{Z}^2 : x^2 + y^2 = n\}| \geq \\ &|\{(x, y) \in \mathbf{Z}^2 : 5x^2 + 5y^2 = n\}|, \end{aligned}$$

from which  $\psi(q)(\phi(q)^2 - \phi(q^5)^2) \in P[q]$  follows quickly. However, we want much more. We ask for analogues of (2.18) and for associated theta-function identities.

Where do we begin? How about if we begin with binary forms of discriminant  $-40$ . Again, there are just two genera of binary quadratic forms with this discriminant: the class of  $x^2 + 10y^2$  and the class of  $2x^2 + 5y^2$ . We now add  $10z^2$  to both forms to obtain ternaries  $x^2 + 10y^2 + 10z^2$ ,  $2x^2 + 5y^2 + 10z^2$  of discriminant 400. We observe that  $2x^2 + 5y^2 + 10z^2$  is the only form in its genus and that the genus containing  $x^2 + 10y^2 + 10z^2$  contains one more non-diagonal ternary form  $4x^2 + 5y^2 + 6z^2 + 4zx$ .

It would be wrong to assume that we constructed all genera of ternary quadratic forms of discriminant 400 this way. In fact, we are being very selective by picking just two out of twelve possible genera of discriminant 400. For the interested reader, we note that a table of genera of ternary quadratic forms, up to discriminant 1,000, is available on Neil Sloane’s website at <http://www.research.att.com/~njas/lattices/Brandt.1.html> and was computed by Alexander Schiemann. In particular, this table includes relevant discriminants 144, 400, and 784. However, it should be noted that Schiemann’s discriminants are the negative of ours. The reader should also be cautioned that the integer sextuple defining each form is preceded by an identification number and a colon, and that the identification number has no mathematical significance. The present authors use a combination of Schiemann’s software, scripts in a language called Magma, and C++ code written by the second author. For the reader with no experience of ternary forms, we heartily recommend [7], especially the tables on pages 111–113.

Again, it is easy to see that  $(-n_1|5) = 1$  for any integer  $n_1$  represented by the genus of  $x^2 + 10y^2 + 10z^2$ ,  $(n_1, 5) = 1$  and that  $(-n_2|5) = -1$  for any integer  $n_2$  represented by  $2x^2 + 5y^2 + 10z^2$ ,  $(n_2, 5) = 1$ . Also

$$|\text{Aut}(1,10,10,0,0,0)| = 16,$$

$$|\text{Aut}(4,5,6,0,4,0)| = |\text{Aut}(2,5,10,0,0,0)| = 8.$$

And so we anticipate two results similar to (2.36) and (2.37). Namely,

$$\begin{aligned} (1, 8, 8, 0, 0, 0)(M) &= (1, 10, 10, 0, 0, 0)(M) + 2(4, 5, 6, 0, 4, 0)(M) \\ &\quad + 2(2, 5, 10, 0, 0, 0)(M), \end{aligned} \tag{3.2}$$

with  $M \equiv 1 \pmod{8}$ , and

$$\begin{aligned} 5(1, 8, 8, 0, 0, 0) \left( \frac{M}{5^2} \right) &= (1, 10, 10, 0, 0, 0)(M) + 2(4, 5, 6, 0, 4, 0)(M) \\ &\quad - 2(2, 5, 10, 0, 0, 0)(M), \end{aligned} \tag{3.3}$$

with  $M \equiv 1 \pmod{8}$ ,  $5|M$ . To prove (3.2), we rewrite it as

$$S_{8,1}(\phi(q)\phi(q^8)^2 - \phi(q)\phi(q^{10})^2 - 2\phi(q^5)\chi(q) - 2\phi(q^2)\phi(q^5)\phi(q^{10})) = 0, \tag{3.4}$$

where

$$\chi(q) := \sum_{x,z \in \mathbf{Z}} q^{4x^2+4xz+6z^2}.$$

It is easy to see that

$$\chi(q) = \sum_{x \equiv z \pmod{2}} q^{x^2+5z^2}.$$

Hence,

$$\chi(q) = \phi(q^4)\phi(q^{20}) + 4q^6\psi(q^8)\psi(q^{40}). \quad (3.5)$$

Upon employing (2.4), (3.5) we obtain

$$\begin{aligned} S_{8,1}(\phi(q)\phi(q^8)^2) &= 2\psi(q)\phi(q)^2, \\ S_{8,1}(\phi(q)\phi(q^{10})^2) &= 2\psi(q)\phi(q^5)^2, \\ S_{8,1}(\phi(q^2)\phi(q^5)\phi(q^{10})) &= 8q^2\psi(q^2)\psi(q^5)\psi(q^{10}), \\ S_{8,1}(\phi(q^5)\chi(q)) &= S_{8,1}(\phi(q^5)\phi(q^4)\phi(q^{20})) \\ &= S_{8,1}(\phi(q^5)\phi(q^{16})\phi(q^{20})) \\ &\quad + S_{8,1}(2q^4\phi(q^5)\psi(q^{32})\phi(q^{20})) \\ &= 4q^3\psi(q^5)\psi(q^{20})\phi(q^2) + 4q\psi(q^4)\psi(q^5)\phi(q^{10}). \end{aligned}$$

This means that (3) and, as a result, (3.2) is equivalent to the following theta-function identity

$$\begin{aligned} \psi(q)(\phi(q)^2 - \phi(q^5)^2) &= 4q^3\psi(q^5)\psi(q^{20})\phi(q^2) \\ &\quad + 4q\psi(q^4)\psi(q^5)\phi(q^{10}) + 8q^2\psi(q^2)\psi(q^5)\psi(q^{10}). \end{aligned} \quad (3.6)$$

It is easy to convert (3.6) into a modular equation of degree 5,

$$\begin{aligned} (m-1) \frac{\alpha^{\frac{1}{8}}}{\beta^{\frac{1}{8}}} &= (1 + (1-\alpha)^{\frac{1}{2}})^{\frac{1}{2}} (1 - (1-\beta)^{\frac{1}{2}})^{\frac{1}{2}} \\ &\quad + (1 - (1-\alpha)^{\frac{1}{2}})^{\frac{1}{2}} (1 + (1-\beta)^{\frac{1}{2}})^{\frac{1}{2}} + 2(\alpha\beta)^{\frac{1}{4}}. \end{aligned} \quad (3.7)$$

Here

$$\begin{aligned} m &= \frac{\phi(q)^2}{\phi(q^5)^2}, \\ \alpha &= 1 - \frac{\phi(-q)^4}{\phi(q)^4} = 16q \frac{\psi(q^2)^4}{\phi(q)^4}, \\ \beta &= 1 - \frac{\phi(-q^5)^4}{\phi(q^5)^4} = 16q^5 \frac{\psi(q^{10})^4}{\phi(q^5)^4}. \end{aligned}$$

Our proof of (3.7) hinges upon three powerful results established in [1], pp. 285–286.

$$2(1 - (\alpha\beta)^{\frac{1}{2}} - ((1 - \alpha)(1 - \beta))^{\frac{1}{2}}) = (m - 1) \left( -1 + \frac{5}{m} \right), \tag{3.8}$$

$$\frac{\alpha^{\frac{1}{4}}}{\beta^{\frac{1}{4}}} = \frac{2m + r}{m(m - 1)}, \tag{3.9}$$

$$4(\alpha^3\beta)^{\frac{1}{8}} = \frac{r}{m} + 3 - \frac{5}{m}, \tag{3.10}$$

where  $r = (m(m^2 - 2m + 5))^{\frac{1}{2}}$ . We begin by rewriting (3.7) as

$$\begin{aligned} (m - 1) \frac{\alpha^{\frac{1}{8}}}{\beta^{\frac{1}{8}}} - 2(\alpha\beta)^{\frac{1}{4}} &= (1 + (1 - \alpha)^{\frac{1}{2}})^{\frac{1}{2}} (1 - (1 - \beta)^{\frac{1}{2}})^{\frac{1}{2}} \\ &\quad + (1 - (1 - \alpha)^{\frac{1}{2}})^{\frac{1}{2}} (1 + (1 - \beta)^{\frac{1}{2}})^{\frac{1}{2}}. \end{aligned} \tag{3.11}$$

Then we square both sides to obtain

$$(m - 1)^2 \frac{\alpha^{\frac{1}{4}}}{\beta^{\frac{1}{4}}} - 4(m - 1)\alpha^{\frac{3}{8}}\beta^{\frac{1}{8}} = 2(1 - (\alpha\beta)^{\frac{1}{2}} - ((1 - \alpha)(1 - \beta))^{\frac{1}{2}}).$$

Next, we use (3.8)–(3.10) to arrive at the trivial statement

$$\frac{(m - 1)(2m + r)}{m} - \frac{(m - 1)(r + 3m - 5)}{m} = (m - 1) \left( \frac{5}{m} - 1 \right).$$

Hence, the proof of (3.7) is complete. Consequently, (3.2) is true, as desired.

To prove (3.3), we wish to consider another modular equation of degree 5

$$\begin{aligned} \left( \frac{5}{m} - 1 \right) \frac{\beta^{\frac{1}{8}}}{\alpha^{\frac{1}{8}}} + 4(\alpha\beta)^{\frac{1}{4}} &= (1 + (1 - \alpha)^{\frac{1}{2}})^{\frac{1}{2}} (1 - (1 - \beta)^{\frac{1}{2}})^{\frac{1}{2}} \\ &\quad + (1 - (1 - \alpha)^{\frac{1}{2}})^{\frac{1}{2}} (1 + (1 - \beta)^{\frac{1}{2}})^{\frac{1}{2}}. \end{aligned} \tag{3.12}$$

Comparing it with (3.11), we see that

$$(m - 1) \frac{\alpha^{\frac{1}{8}}}{\beta^{\frac{1}{8}}} = \left( \frac{5}{m} - 1 \right) \frac{\beta^{\frac{1}{8}}}{\alpha^{\frac{1}{8}}} + 4(\alpha\beta)^{\frac{1}{4}}.$$

Next, we multiply both sides by  $\frac{\alpha^{\frac{1}{8}}}{\beta^{\frac{1}{8}}}$  to obtain

$$(m - 1) \frac{\alpha^{\frac{1}{4}}}{\beta^{\frac{1}{4}}} = \frac{5}{m} - 1 + 4(\alpha^3\beta)^{\frac{1}{8}}.$$

Employing (3.9)–(3.10), we arrive at the trivial statement

$$\frac{2m + r}{m} = \frac{5}{m} - 1 + \frac{r}{m} + 3 - \frac{5}{m}.$$

This completes the proof of (3.12). To proceed further, we rewrite (3.12) in terms of theta-functions as:

$$5\psi(q^5)\phi(q^5)^2 - \psi(q^5)\phi(q)^2 = 4\psi(q)\psi(q^4)\phi(q^{10}) - 8q\psi(q)\psi(q^2)\psi(q^{10}) + 4q^2\psi(q)\psi(q^{20})\phi(q^2). \tag{3.13}$$

Using (2.4) and (3.5) and some elbow grease, we can show that (3.13) is equivalent to

$$S_{40,25}(\phi(q)\phi(q^{10})^2 + 2\phi(q^5)\chi(q) - 2\phi(q^2)\phi(q^5)\phi(q^{10})) = 10\psi(q^5)\phi(q^5)^2.$$

This implies that

$$(1, 10, 10, 0, 0, 0)(M) + 2(4, 5, 6, 0, 4, 0)(M) - 2(2, 5, 10, 0, 0, 0)(M) = 5 \left| \left\{ (x, y, z) \in \mathbf{Z}^3 : 5x^2 + 5y^2 + 5\frac{(1+z)z}{2} = n \right\} \right|,$$

with  $M = 40n + 25$ . The last equation can be easily recognized as (3.3). As before, we can extend (3.2) and (3.3) by using  $\frac{1}{3}(1, 1, 1, 0, 0, 0)(M)$  in place of  $(1, 8, 8, 0, 0, 0)(M)$  as

$$\frac{1}{3}(1, 1, 1, 0, 0, 0)(M) = (1, 10, 10, 0, 0, 0)(M) + 2(4, 5, 6, 0, 4, 0)(M) + 2(2, 5, 10, 0, 0, 0)(M), \tag{3.14}$$

with  $M \equiv 1, 2 \pmod{4}$ , and

$$\frac{5}{3}(1, 1, 1, 0, 0, 0) \left( \frac{M}{5^2} \right) = (1, 10, 10, 0, 0, 0)(M) + 2(4, 5, 6, 0, 4, 0)(M) - 2(2, 5, 10, 0, 0, 0)(M), \tag{3.15}$$

with  $M \equiv 1, 2 \pmod{4}$ ,  $5|M$ . While we have to suppress the details for the sake of brevity, we cannot resist displaying four relevant theta-function identities.

$$\psi(q^2)(\phi(q)^2 - \phi(q^5)^2) = 2q\psi(q^5)^2\phi(q) + 2q\psi(q^{10})\phi(q^2)\phi(q^{10}) + 8q^4\psi(q^4)\psi(q^{10})\psi(q^{20}).$$



This one proves the case  $M \equiv 1 \pmod{4}$  in (3.14). Analogously,

$$\begin{aligned} \psi(q^2)^2\phi(q) &= q^2\psi(q^{10})^2\phi(q) + 2q\psi(q^2)\psi(q^5)^2 \\ &\quad + q^2\psi(q^{20})\phi(q^2)\phi(q^5) + \psi(q^4)\phi(q^5)\phi(q^{10}) \end{aligned}$$

proves the case  $M \equiv 2 \pmod{4}$  in (3.14). Finally,

$$\begin{aligned} 5q\psi(q^{10})\phi(q^5)^2 &= q\psi(q^{10})\phi(q)^2 + 2\psi(q)^2\phi(q^5) \\ &\quad - 8q^3\psi(q^2)\psi(q^4)\psi(q^{20}) - 2\psi(q^2)\phi(q^2)\phi(q^{10}) \end{aligned}$$

and

$$\begin{aligned} 5q^2\psi(q^5)^2\psi(q^{10}) &= \psi(q^2)^2\phi(q^5) + 2q\psi(q)^2\psi(q^{10}) \\ &\quad - \psi(q^4)\phi(q)\phi(q^{10}) - q^2\psi(q^{20})\phi(q)\phi(q^2) \end{aligned}$$

are required to prove (3.15).

### 4 Ternary Forms with Discriminant 784

Here we will prove Theorem 1.1 and Theorem 1.2, stated in the Introduction. It seems that the identities for theta functions in (1.14) and (1.15) correspond to modular equations of mixed degree 21. While Ramanujan had some results for modular equations of this degree [1], we could not find enough relations to handle our formulas (1.14) and (1.15). And so, it is with some reluctance that we resort to routine modular function techniques. The necessary background theory on modular functions and forms may be found in Rankin’s book [13]. Of central importance to us is the valence formula (p.98, [13]).

We begin by dividing both sides of (1.14) by  $\psi(q)\phi(q)^2$ . Making use of (1.7), (1.8), (1.9), (1.10), and (1.12) we end up with a simple identity for four  $\eta$ -quotients

$$g_1(z) + 4g_2(z) + 8g_3(z) + 8g_4(z) = 1, \tag{4.1}$$

where

$$\begin{aligned} g_1(z) &:= \frac{\eta(14z)^{10}\eta(4z)^4\eta(z)^4}{\eta(28z)^4\eta(7z)^4\eta(2z)^{10}}, \\ g_2(z) &:= \frac{\eta(14z)^7\eta(4z)^6\eta(z)^5}{\eta(28z)^2\eta(7z)^3\eta(2z)^{13}}, \\ g_3(z) &:= \frac{\eta(42z)^5\eta(28z)^2\eta(6z)^2\eta(4z)^4\eta(z)^5}{\eta(84z)^2\eta(21z)^2\eta(14z)\eta(3z)\eta(2z)^{12}}, \\ g_4(z) &:= \frac{\eta(84z)\eta(28z)\eta(21z)\eta(14z)\eta(4z)^4\eta(3z)^2\eta(z)^4}{\eta(42z)\eta(7z)\eta(6z)\eta(2z)^{11}}. \end{aligned}$$

Clearly, all our  $\eta$ -quotients in (4.1) are of the form

$$f(z) = \prod_{\delta|n} \eta(\delta z)^{r_\delta},$$

where  $n$  is some positive integer (84 in our case) and all  $\delta \geq 1$ ,  $r_\delta$  are integers. The following result was proved by Morris Newman [12].

**Theorem 4.1.** *The  $\eta$ -quotient  $f(z)$  is a modular function on*

$$\Gamma_0(n) := \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL_2(\mathbf{Z}) : c \equiv 0 \pmod{n} \right\},$$

if the following four conditions are met

$$\begin{aligned} \sum_{\delta|n} r_\delta &= 0, \\ \sum_{\delta|n} \delta r_\delta &\equiv 0 \pmod{24}, \\ \sum_{\delta|n} \frac{nr_\delta}{\delta} &\equiv 0 \pmod{24}, \\ \prod_{\delta|n} \delta^{r_\delta} &\text{ is a rational square.} \end{aligned}$$

It is now straightforward to verify that  $g_1(z)$ ,  $g_2(z)$ ,  $g_3(z)$ , and  $g_4(z)$  are modular functions on  $\Gamma_0(84)$ . Consequently,

$$h(z) := g_1(z) + 4g_2(z) + 8g_3(z) + 8g_4(z) - 1$$

is also a modular function on  $\Gamma_0(84)$ . To proceed, we will need the following observation from [3].

**Theorem 4.2.** *If  $n$  is square free integer, then a complete set of inequivalent cusps for  $\Gamma_0(4n)$  is  $\{1/s : s|4n\}$ .*

And so,  $k \cup \{\frac{1}{84}\}$  is a complete set of 12 inequivalent cusps of  $\Gamma_0(84)$  where  $k := \{1, 1/2, 1/6, 1/4, 1/12, 1/7, 1/42, 1/21, 1/3, 1/14, 1/28\}$ . From the definition of  $\eta(z)$ , it follows that the  $\eta$ -quotients have no zeros or poles in the upper-half plane (i.e.,  $Im(z) > 0$ ). Ligozat [11] calculated the order of the  $\eta$ -quotient  $f(z)$  at the cusps of  $\Gamma_0(n)$ .

**Theorem 4.3.** *If an  $\eta$ -quotient  $f$  is a modular function on  $\Gamma_0(n)$ , then at the cusp  $\frac{b}{c}$  with  $\gcd(b, c) = 1$*

$$ORD\left(f, \frac{b}{c}\right) = \frac{n}{24 \gcd(n, c^2)} \sum_{\delta|n} \frac{r_\delta \gcd(c, \delta)^2}{\delta}.$$

**Table 1** Orders of the cusps of  $\Gamma_0(84)$

| CUSP | $O_1(s)$ | $O_2(s)$ | $O_3(s)$ | $O_4(s)$ | $O_h(s)$ |
|------|----------|----------|----------|----------|----------|
| 1    | 0        | 0        | 0        | 0        | 0        |
| 1/2  | -9       | -12      | -12      | -12      | -12      |
| 1/6  | -3       | -4       | -1       | -4       | -4       |
| 1/4  | 0        | 3        | -1       | -1       | -1       |
| 1/12 | 0        | 1        | 2        | 0        | 0        |
| 1/7  | 0        | 0        | 0        | 0        | 0        |
| 1/42 | 3        | 2        | 5        | 0        | 0        |
| 1/21 | 0        | 0        | 0        | 3        | 0        |
| 1/3  | 0        | 0        | 0        | 5        | 0        |
| 1/14 | 9        | 6        | 0        | 0        | 0        |
| 1/28 | 0        | 3        | 5        | 5        | 0        |

This way we obtain Table 1, where  $O_h(s)$  is a lower bound for  $ORD(h, s)$  and  $O_i(s) := ORD(g_i, s), i = 1, 2, 3, 4$ . To prove (1.14) and (4.1), we must show that  $h(z) = 0$ . The valence formula implies that (unless  $h$  is a constant)

$$\sum_{s \in k} ORD(h, s) + ORD(h, 1/84) \leq 0.$$

Using data collected in Table 1 and keeping in mind that cusp  $1/84$  is equivalent to  $i\infty$ , we infer that (unless  $h$  is constant)

$$-17 + ORD(h, i\infty) \leq 0.$$

But direct inspection shows that  $ORD(h, i\infty) > 17$ . That is, if one expands  $h$  in powers of  $q$ , then one finds that the first 18 coefficients in this expansion are zero. Hence, one arrives at a contradiction. This contradiction implies that  $h = 0$ , as desired. This completes our proof of (1.14). Obviously, we could have proved (1.15) in a similar fashion. Instead, we choose a more painful way because there is nothing like pain for achieving excellence. In any event, our approach will shed some extra light on the relation between (1.14) and (1.15). It is not hard to verify (in term by term fashion) that

$$8q^2\psi(q^3)\psi(q^{14})\phi(q^{21}) + 8q^4\psi(q^{14})f(q, q^2)f(q^7, q^{35}) = C(q)(8q^2\psi(q^2)\psi(-q^{21})\phi(-q^3) + 8\psi(q^2)f(-q, -q^5)f(-q^7, q^{14})), \tag{4.2}$$

where

$$C(q) = q^2 \frac{E(q^{28})E(q^{14})E(q^2)^2}{E(q^7)E(q^4)^2E(q)}.$$

All one needs is a simple formula

$$E(-q) = \frac{E(q^2)^3}{E(q^4)E(q)}.$$

We will prove shortly that

$$\begin{aligned} &\psi(q)(\phi(q)^2 - \phi(q^7)^2) - 4q\psi(q^2)\psi(q^7)\phi(q^7) = \\ &C(q)(\psi(-q^7)(7\phi(-q^7)^2 + \phi(-q)^2) - 4q\psi(-q)\psi(q^{14})\phi(-q)). \end{aligned} \tag{4.3}$$

Next, we rewrite (1.14) as

$$\begin{aligned} &\psi(q)(\phi(q)^2 - \phi(q^7)^2) - 4q\psi(q^2)\psi(q^7)\phi(q^7) = \\ &8q^2\psi(q^3)\psi(q^{14})\phi(q^{21}) + 8q^4\psi(q^{14})f(q, q^2)f(q^7, q^{35}), \end{aligned} \tag{4.4}$$

and use (4) on the right and (4.3) on the left to get

$$\begin{aligned} &C(q)(\psi(-q^7)(7\phi(-q^7)^2 + \phi(-q)^2) - 4q\psi(-q)\psi(q^{14})\phi(-q)) = \\ &C(q)(8q^2\psi(q^2)\psi(-q^{21})\phi(-q^3) + 8\psi(q^2)f(-q, -q^5)f(-q^7, q^{14})). \end{aligned} \tag{4.5}$$

Dividing both sides by  $C(q)$  and replacing  $q$  by  $-q$ , we get (1.15).

But what about (4.3)? We start by rewriting it as a modular equation of degree 7. Let  $m$  be a multiplier of degree 7,  $\beta$  have degree 7 over  $\alpha$  and  $t$  be defined by

$$t = (\alpha\beta)^{1/8}.$$

Then

$$\begin{aligned} m - 1 - 2t &= \frac{7}{mt^2} \frac{\beta^{7/12}(1-\beta)^{7/12}}{\alpha^{1/12}(1-\alpha)^{1/12}} + \frac{1}{t^2} \frac{\beta^{7/12}(1-\beta)^{7/12}}{\alpha^{1/12}(1-\alpha)^{1/12}} \frac{(1-\alpha)^{1/2}}{(1-\beta)^{1/2}} \\ &\quad - 2 \frac{\beta}{t^4} \frac{\alpha^{7/24}(1-\alpha)^{7/24}}{\beta^{1/24}(1-\beta)^{1/24}}. \end{aligned} \tag{4.6}$$

The proof of the following modular equations of degree 7 can be found in (p. 314, [1])

$$(1-\alpha)^{1/8}(1-\beta)^{1/8} = 1-t, \tag{4.7}$$

$$\frac{7(1-2t)}{m} + 1 = 4 \frac{\alpha^{7/24}(1-\alpha)^{7/24}}{\beta^{1/24}(1-\beta)^{1/24}}, \tag{4.8}$$

$$\frac{(m(2t-1)+1)^2}{16} = \frac{\beta^{7/12}(1-\beta)^{7/12}}{\alpha^{1/12}(1-\alpha)^{1/12}}. \tag{4.9}$$

Observe that (4.7) implies that

$$\frac{(1 - \alpha)^{\frac{1}{2}}}{(1 - \beta)^{\frac{1}{2}}} = \frac{(1 - t)^4}{1 - \beta}.$$

And so (4.6) becomes

$$m - 1 - 2t + \beta \frac{7(1-2t)}{2t^4} + 1 - \left( \frac{7}{m} + \frac{(1-t)^4}{1-\beta} \right) \frac{(m(2t-1) + 1)^2}{16t^2} = 0. \tag{4.10}$$

Next, we recall the (19.19) in [1]

$$m = \frac{t - \beta}{t(1-t)(1-t+t^2)}.$$

We use it to eliminate  $m$  from (4.10) and obtain after some algebra that

$$(\beta^2 - \beta((1+t^8) - (1-t)^8) + t^8) \frac{P(t, \beta)}{Q(t, \beta)} = 0, \tag{4.11}$$

where  $P(t, \beta)$  and  $Q(t, \beta)$  are some polynomials in  $t$  and  $\beta$ . But  $\beta$  and  $\alpha$  are roots of the quadratic equation

$$x^2 - x((1+t^8) - (1-t)^8) + t^8 = 0,$$

as observed on page 316 in [1]. Hence, the proof of (4.3) is complete. Consequently, (1.15) is true.

We now are ready to prove Theorem 1.2. Clearly, (1.16) is equivalent to

$$S_{8,1} (\phi(q)\phi(q^8)^2 - \phi(q)\phi(q^{14})^2 - 2\phi(q^2)\phi(q^7)\phi(q^{14}) - 4\phi(q^{14})u(q)) = 0, \tag{4.12}$$

where

$$u(q) := \sum_{x,y \in \mathbf{Z}} q^{3x^2+2xy+5y^2}.$$

It is not hard to check that

$$u(q) = \sum_{x \equiv y \pmod{3}} q^{\frac{x^2+14y^2}{3}}.$$

Hence,

$$u(q) = \phi(q^3)\phi(q^{42}) + 2q^5 f(q, q^5) f(q^{14}, q^{70}). \tag{4.13}$$

We shall also require

$$S_{8,1}(q^3 f(q, q^5) f(q^{14}, q^{70})) = q^2 f(q, q^2) f(q^7, q^{35}). \tag{4.14}$$

Using (2.4) together with (4.13) and (4.14), we deduce that

$$\begin{aligned} S_{8,1}(\phi(q^{14})u(q)) &= 2q\psi(q^{14})S_{8,1}(q^6(\phi(q^3)\phi(q^{42}) + 2q^5 f(q, q^5) f(q^{14}, q^{70}))) \\ &= 4q^2\psi(q^{14})\psi(q^3)\phi(q^{21}) + 4q^4\psi(q^{14})f(q, q^2) f(q^7, q^{35}). \end{aligned} \tag{4.15}$$

Next, with the aid of (2.4), we verify that

$$S_{8,1}(\phi(q)\phi(q^{14})^2) = 2\psi(q)\phi(q^7)^2, \tag{4.16}$$

$$S_{8,1}(\phi(q^2)\phi(q^7)\phi(q^{14})) = 4q\psi(q^2)\psi(q^7)\phi(q^7). \tag{4.17}$$

Combining (2.14), (4.12), and (4.15)–(4.17) we end up with (1.14). Hence, the proof of (1.16) is complete.

Our proof of (1.17) is analogous. We verify that

$$S_{56,49}(\phi(q)\phi(q^{14})^2) = 2\psi(q^7)\phi(q)^2,$$

$$S_{56,49}(\phi(q^2)\phi(q^7)\phi(q^{14})) = 4q\psi(q)\psi(q^{14})\phi(q),$$

$$S_{56,49}(\phi(q^{14})u(q)) = 4q^2\psi(q^2)\psi(q^{21})\phi(q^3) + 4\psi(q^2) f(q, q^5) f(q^7, q^{14}).$$

These results enable us to convert (1.15) into

$$S_{56,49}(-\phi(q)\phi(q^{14})^2 - 2\phi(q^2)\phi(q^7)\phi(q^{14}) + 4\phi(q^{14})u(q)) = 14\psi(q^7)\phi(q^7)^2.$$

Hence, we have

$$\begin{aligned} &-(1, 14, 14, 0, 0, 0)(M) - 2(2, 7, 14, 0, 0, 0)(M) + 4(3, 5, 14, 0, 0, 2)(M) = \\ &7 \left| \left\{ (x, y, z) \in \mathbf{Z}^3 : 7\frac{x(x+1)}{2} + 7y^2 + 7z^2 = n \right\} \right| = \\ &7 \left| \left\{ (x, y, z) \in \mathbf{Z}^3 : x^2 + 8y^2 + 8z^2 = 1 + 8\frac{n}{7} \right\} \right|, \quad M = 56n + 49. \end{aligned}$$

The truth of (1.17) is now transparent. Again, we can extend (1.14),(1.15) by using  $\frac{1}{3}(1, 1, 1, 0, 0, 0)(M)$  instead of  $(1, 8, 8, 0, 0, 0)(M)$ . In this way we have

$$\begin{aligned} \frac{1}{3}(1, 1, 1, 0, 0, 0)(M) &= (1, 14, 14, 0, 0, 0)(M) + 2(2, 7, 14, 0, 0, 0)(M) \\ &\quad + 4(3, 5, 14, 0, 0, 2)(M), \end{aligned} \tag{4.18}$$

with  $M \equiv 1, 2 \pmod{4}$ , and

$$\begin{aligned} \frac{7}{3}(1, 1, 1, 0, 0, 0) \left( \frac{M}{7^2} \right) &= -(1, 14, 14, 0, 0, 0)(M) - 2(2, 7, 14, 0, 0, 0)(M) \\ &\quad + 4(3, 5, 14, 0, 0, 2)(M), \end{aligned} \tag{4.19}$$

with  $M \equiv 1, 2 \pmod{4}$ ,  $7|M$ . We limit ourselves to a few remarks. The theta-function identities

$$\begin{aligned} \psi(q^2)(\phi(q)^2 - \phi(q^7)^2) &= 4q^2\psi(q^4)\psi(q^{14})\phi(q^{14}) + 4q^5\phi(q^2)\psi(q^{14})\psi(q^{28}) \\ &\quad + 8q^4\psi(q^6)\psi(q^7)\psi(q^{21}) \\ &\quad + 4q\psi(q^7)f(q^2, q^4)f(q^7, q^{14}), \end{aligned}$$

and

$$\begin{aligned} \phi(q)(\psi(q^2)^2 - q^3\psi(q^{14})^2) &= q^3\psi(q^{28})\phi(q^2)\phi(q^7) + \psi(q^4)\phi(q^7)\phi(q^{14}) \\ &\quad + 2q^3\psi(q^7)\psi(q^{21})\phi(q^3) \\ &\quad + 2q\psi(q^7)f(q, q^5)f(q^7, q^{14}) \end{aligned}$$

imply the cases  $M \equiv 1 \pmod{4}$  and  $M \equiv 2 \pmod{4}$  in (4.18), respectively. Analogously,

$$\begin{aligned} 7q\psi(q^7)^2\phi(q^7) &= 8q^5\psi(q)\psi(q^3)\psi(q^{42}) + 4\psi(q)f(q, q^2)f(q^{14}, q^{28}) \\ &\quad - q\phi(q)^2\psi(q^{14}) - 4\psi(q^2)\psi(q^4)\phi(q^{14}) \\ &\quad - 4q^3\psi(q^2)\psi(q^{28})\phi(q^2) \end{aligned}$$

and

$$\begin{aligned} 7q^3\psi(q^7)^2\psi(q^{14}) &= 2\psi(q)\psi(q^3)\phi(q^{21}) + 2q^2\psi(q)f(q^7, q^{35})f(q, q^2) \\ &\quad - \psi(q^2)^2\phi(q^7) - \psi(q^4)\phi(q)\phi(q^{14}) - q^3\psi(q^{28})\phi(q)\phi(q^2) \end{aligned}$$

can be used to prove both cases in (4.19).

We conclude this section by showing how to relate the ternaries on the right of (1.16) and (1.17) to binaries with discriminant  $-56$ . Again, there are just two genera of binary quadratic forms with this discriminant. The first one contains the class of  $x^2 + 14y^2$  and the class of  $2x^2 + 7y^2$ . The second one contains the class of  $3x^2 + 2xy + 5y^2$  and the class of  $3x^2 - 2xy + 5y^2$ . We now add  $14z^2$  to both forms in the first genus of binary quadratic forms to obtain our first genus of ternary quadratic forms of discriminant 784

$$\{x^2 + 14y^2 + 14z^2, \quad 2x^2 + 7y^2 + 14z^2\}.$$

Next, we add  $14z^2$  to the first forms in the second genus of binary quadratic forms to obtain our second genus of ternary quadratic forms of discriminant 784

$$\{3x^2 + 2xy + 5y^2 + 14z^2\}.$$

Both these genera are complete as described, no other forms need be added. As a result,  $3x^2 + 2xy + 5y^2 + 14z^2$  is obviously regular. Also, it is easy to verify that

$$\frac{16}{|\text{Aut}(1, 14, 14, 0, 0, 0)|} = 1,$$

$$\frac{16}{|\text{Aut}(2, 7, 14, 0, 0, 0)|} = 2,$$

$$\frac{16}{|\text{Aut}(3, 5, 14, 0, 0, 2)|} = 4.$$

And so (1.16) can be stated as

$$\begin{aligned} (1, 8, 8, 0, 0, 0)(M) &= \frac{16(1, 14, 14, 0, 0, 0)(M)}{|\text{Aut}(1, 14, 14, 0, 0, 0)|} + \frac{16(2, 7, 14, 0, 0, 0)(M)}{|\text{Aut}(2, 7, 14, 0, 0, 0)|} \\ &\quad + \frac{16(3, 5, 14, 0, 0, 2)(M)}{|\text{Aut}(3, 5, 14, 0, 0, 2)|}, \end{aligned}$$

with  $M \equiv 1 \pmod{8}$ . The right hand side of this identity is a weighted average. In effect, this is what would occur if one extended Siegel's fundamental theorem to a sum over more than one genus. Moreover, it is easy to see that

$$(-n_1|7) = -1,$$

for any integer  $n_1$  represented by the genus of  $x^2 + 14y^2 + 14z^2$ ,  $(n_1, 7) = 1$  and that

$$(-n_2|7) = 1,$$

for any integer  $n_2$  represented by  $3x^2 + 5y^2 + 14z^2 + 2xy$ ,  $(n_2, 7) = 1$ . This allows us to rewrite (1.17) as

$$\begin{aligned} 7(1, 8, 8, 0, 0, 0) \left( \frac{M}{7^2} \right) &= (-n_1|7) \left( \frac{16(1, 14, 14, 0, 0, 0)(M)}{|\text{Aut}(1, 14, 14, 0, 0, 0)|} \right. \\ &\quad \left. + \frac{16(2, 7, 14, 0, 0, 0)(M)}{|\text{Aut}(2, 7, 14, 0, 0, 0)|} \right) \\ &\quad + (-n_2|7) \frac{16(3, 5, 14, 0, 0, 2)(M)}{|\text{Aut}(3, 5, 14, 0, 0, 2)|}, \end{aligned}$$

with  $M \equiv 1 \pmod{8}$ ,  $7|M$ .



### 5 Ternary Forms with Discriminant 3600

Up to now, all our theorems involved certain ternary forms with discriminant  $16p^2$  for prime  $p = 3, 5, 7$ . In this section, we consider ternaries with discriminant  $16S^2$  for composite  $S = 15$ . This case has all of the ingredients of the general case to be discussed later. Again, we start with the binaries with discriminant  $-120$ . There are four genera with this discriminant and each has a single class per genus:  $\{x^2 + 30y^2\}$ ,  $\{3x^2 + 10y^2\}$ ,  $\{5x^2 + 6y^2\}$ ,  $\{2x^2 + 15y^2\}$ . Following a well-trodden path, we add  $30z^2$  to each of these forms to get four ternary forms with discriminant 3,600. Next, we extend each ternary form to a genus of ternary quadratic forms. In this way, we obtain four genera of ternary quadratic forms with discriminant 3,600:

$$\begin{aligned} \{x^2 + 30y^2\} &\rightarrow TG_1 := \{x^2 + 30y^2 + 30z^2, 6x^2 + 10y^2 + 15z^2\}, \\ \{3x^2 + 10y^2\} &\rightarrow TG_2 := \{3x^2 + 10y^2 + 30z^2\}, \\ \{5x^2 + 6y^2\} &\rightarrow TG_3 := \{5x^2 + 6y^2 + 30z^2, 9x^2 + 11y^2 + 11z^2 + 2yz + 6zx + 6xy\}, \\ \{2x^2 + 15y^2\} &\rightarrow TG_4 := \{2x^2 + 15y^2 + 30z^2, 5x^2 + 12y^2 + 18z^2 + 12yz\}. \end{aligned}$$

We check that

$$\begin{aligned} \frac{16}{|\text{Aut}(1, 30, 30, 0, 0, 0)|} &= 1, \\ \frac{16}{|\text{Aut}(6, 10, 15, 0, 0, 0)|} &= \frac{16}{|\text{Aut}(3, 10, 30, 0, 0, 0)|} = 2 \\ \frac{16}{|\text{Aut}(5, 6, 30, 0, 0, 0)|} &= \frac{16}{|\text{Aut}(2, 15, 30, 0, 0, 0)|} = \frac{16}{|\text{Aut}(5, 12, 18, 12, 0, 0)|} = 2, \\ \frac{16}{|\text{Aut}(9, 11, 11, 2, 6, 6)|} &= 4. \end{aligned}$$

We now take Siegel’s weighted average over the four genera above. In this way, we are led to

$$\begin{aligned} (1, 8, 8, 0, 0, 0)(M) &= (1, 30, 30, 0, 0, 0)(M) + 2(6, 10, 15, 0, 0, 0)(M) \\ &\quad + 2(3, 10, 30, 0, 0, 0)(M) + 2(5, 6, 30, 0, 0, 0)(M) \\ &\quad + 4(9, 11, 11, 2, 6, 6)(M) + 2(2, 15, 30, 0, 0, 0)(M) \\ &\quad + 2(5, 12, 18, 12, 0, 0)(M), \end{aligned} \tag{5.1}$$

with  $M \equiv 1 \pmod{8}$ . This can be stated compactly as

$$(1, 8, 8, 0, 0, 0)(M) = \sum_{i=1}^4 W_i(M), \tag{5.2}$$

with  $M \equiv 1 \pmod{8}$ . Here,

$$W_i(M) := 16 \sum_{f \in TG_i} \frac{R_f(M)}{|\text{Aut}(f)|}, \quad i = 1, 2, 3, 4,$$

and  $R_f(M)$  denotes the number of representations of  $M$  by  $f$ . The associated theta-function identity is as follows:

$$\begin{aligned} \psi(q)\phi(q)^2 &= \psi(q)\phi(q^{15})^2 + 4q^3\psi(q^{10})\psi(q^{15})\phi(q^3) + 4q^4\psi(q^3)\psi(q^{30})\phi(q^5) \\ &\quad + 8q^5\psi(q^5)\psi(q^6)\psi(q^{30}) + 4q\psi(q^9)\phi(q^{45})^2 \\ &\quad + 8q^{11}\psi(q^9)f(q^{15}, q^{75})^2 + 8q^5\phi(q^{45})f(q^3, q^6)f(q^{15}, q^{75}) \\ &\quad + 4q^{10}f(q^3, q^6)f(q^{15}, q^{75})^2 + 4q^2\psi(q^2)\psi(q^{15})\phi(q^{15}) \\ &\quad + 4q^8\psi(q^5)\psi(q^{60})\phi(q^6) + 4q^2\psi(q^5)\psi(q^{12})\phi(q^{30}). \end{aligned} \tag{5.3}$$

Using (5.3), we easily see that

$$\psi(q)(\phi(q)^2 - \phi(q^{15})^2) \in P[q].$$

To prove (5.3), we divide both sides by  $\psi(q)\phi(q)^2$  and use (1.7)–(1.10), (1.12) to end up with an identity for eleven  $\eta$ -quotients

$$\begin{aligned} 1 &= \frac{\eta(30z)^{10}\eta(4z)^4\eta(z)^4}{\eta(60z)^4\eta(15z)^4\eta(2z)^{10}} + 4 \frac{\eta(30z)^2\eta(20z)^2\eta(6z)^5\eta(4z)^4\eta(z)^5}{\eta(15z)\eta(12z)^2\eta(10z)\eta(3z)^2\eta(2z)^{12}} \\ &\quad + 4 \frac{\eta(60z)^2\eta(10z)^5\eta(6z)^2\eta(4z)^4\eta(z)^5}{\eta(30z)\eta(20z)^2\eta(5z)^2\eta(3z)\eta(2z)^{12}} + 8 \frac{\eta(60z)^2\eta(12z)^2\eta(10z)^2\eta(4z)^4\eta(z)^5}{\eta(30z)\eta(6z)\eta(5z)\eta(2z)^{12}} \\ &\quad + 4 \frac{\eta(90z)^{10}\eta(18z)^2\eta(4z)^4\eta(z)^5}{\eta(180z)^4\eta(45z)^4\eta(9z)\eta(2z)^{12}} \\ &\quad + 8 \frac{\eta(180z)^2\eta(45z)^2\eta(30z)^4\eta(18z)^2\eta(4z)^4\eta(z)^5}{\eta(90z)^2\eta(60z)^2\eta(15z)^2\eta(9z)\eta(2z)^{12}} \\ &\quad + 8 \frac{\eta(90z)^4\eta(30z)^2\eta(9z)^2\eta(6z)\eta(4z)^4\eta(z)^5}{\eta(180z)\eta(60z)\eta(45z)\eta(18z)\eta(15z)\eta(3z)\eta(2z)^{12}} \\ &\quad + 4 \frac{\eta(180z)^2\eta(45z)^2\eta(30z)^4\eta(9z)^2\eta(6z)\eta(4z)^4\eta(z)^5}{\eta(90z)^2\eta(60z)^2\eta(18z)\eta(15z)^2\eta(3z)\eta(2z)^{12}} \\ &\quad + 4 \frac{\eta(30z)^7\eta(4z)^6\eta(z)^5}{\eta(60z)^2\eta(15z)^3\eta(2z)^{13}} + 4 \frac{\eta(120z)^2\eta(12z)^5\eta(10z)^2\eta(4z)^4\eta(z)^5}{\eta(60z)\eta(24z)^2\eta(6z)^2\eta(5z)\eta(2z)^{12}} \\ &\quad + 4 \frac{\eta(60z)^5\eta(24z)^2\eta(10z)^2\eta(4z)^4\eta(z)^5}{\eta(120z)^2\eta(30z)^2\eta(12z)\eta(5z)\eta(2z)^{12}}. \end{aligned} \tag{5.4}$$

To verify the last identity, we use the Newman theorem stated in the last section to show that all eleven quotients on the right of (5.4) are modular functions on  $\Gamma_0(360)$ .

Next, let  $H$  denote the right side of (5.4) minus 1. Obviously,  $H$  is also a modular function on  $\Gamma_0(360)$ . We will expand  $H$  in powers of  $q$  to confirm that  $H = 0$ . Just how many terms to calculate is determined by the Ligozat theorem. In this regard, we note that  $K \cup 1/360$  is a complete set of 32 inequivalent cusps of  $\Gamma_0(360)$ . Here,

$$K := \left\{ 1, \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{5}{6}, \frac{1}{8}, \frac{1}{9}, \frac{1}{10}, \frac{1}{12}, \frac{5}{12}, \frac{1}{15}, \frac{2}{15}, \frac{1}{18}, \frac{1}{20}, \frac{1}{24}, \frac{5}{24}, \frac{1}{30}, \frac{11}{30}, \frac{1}{36}, \frac{1}{40}, \frac{1}{45}, \frac{1}{60}, \frac{11}{60}, \frac{1}{72}, \frac{1}{90}, \frac{1}{120}, \frac{5}{120}, \frac{1}{180} \right\}.$$

**Table 2** Lower bounds for orders at the cusps of  $\Gamma_0(360)$

|          |                 |                |                |                |                |                 |                |                |                 |                 |                 |
|----------|-----------------|----------------|----------------|----------------|----------------|-----------------|----------------|----------------|-----------------|-----------------|-----------------|
| CUSP     | 1               | $\frac{1}{2}$  | $\frac{1}{3}$  | $\frac{2}{3}$  | $\frac{1}{4}$  | $\frac{1}{5}$   | $\frac{1}{6}$  | $\frac{5}{6}$  | $\frac{1}{8}$   | $\frac{1}{9}$   |                 |
| $O_H(s)$ | 0               | -54            | 0              | 0              | -5             | 0               | -6             | -6             | -5              | 0               |                 |
| CUSP     | $\frac{1}{10}$  | $\frac{1}{12}$ | $\frac{5}{12}$ | $\frac{1}{15}$ | $\frac{2}{15}$ | $\frac{1}{18}$  | $\frac{1}{20}$ | $\frac{1}{24}$ | $\frac{5}{24}$  | $\frac{1}{30}$  |                 |
| $O_H(s)$ | -6              | 0              | 0              | 0              | 0              | -6              | -1             | 0              | 0               | 0               |                 |
| CUSP     | $\frac{11}{30}$ | $\frac{1}{36}$ | $\frac{1}{40}$ | $\frac{1}{45}$ | $\frac{1}{60}$ | $\frac{11}{60}$ | $\frac{1}{72}$ | $\frac{1}{90}$ | $\frac{1}{120}$ | $\frac{5}{120}$ | $\frac{1}{180}$ |
| $O_H(s)$ | 0               | 0              | -1             | 0              | 0              | 0               | 0              | 0              | 0               | 0               |                 |

From Ligozat’s theorem, we derive the results in Table 2, where  $O_H(s)$  is a lower bound for  $ORD(H, s)$ . The valence formula says that (unless  $H$  is a constant)

$$\sum_{s \in K} ORD(H, s) + ORD\left(H, \frac{1}{360}\right) \leq 0.$$

Using data collected in Table 2 and keeping in mind that the cusp  $1/360$  is equivalent to  $i\infty$ , we deduce that (unless  $H$  is constant)

$$-90 + ORD(H, i\infty) \leq 0. \tag{5.5}$$

We use Maple to calculate 91 coefficients of the Fourier expansion of  $H$ . In this way we see that  $ORD(H, i\infty) > 90$ , and thus (5.5) is contradicted. Hence,  $H = 0$ . The proof of (5.3) and (5.4) is now complete. In exactly the same mechanical manner, we can prove three companion identities:

$$\begin{aligned} 3\psi(q^3)\phi(q^3)^2 &= 4q\psi(q^5)\psi(q^6)\phi(q^5) + 8q^3\psi(q^2)\psi(q^{10})\psi(q^{15}) \\ &\quad + 4\psi(q^3)\phi(q^{15})^2 + 4q^3f(q, q^2)f(q^5, q^{25})^2 \\ &\quad + 4q^2\psi(q^4)\psi(q^{15})\phi(q^{10}) + 4q^4\psi(q^{15})\psi(q^{20})\phi(q^2) \\ &\quad - \psi(q^3)\phi(q^5)^2 - 4q^4\psi(q^5)\psi(q^{30})\phi(q) \\ &\quad - 4q\psi(q)\psi(q^{10})\phi(q^{15}), \end{aligned} \tag{5.6}$$

$$\begin{aligned}
 5\psi(q^5)\phi(q^5)^2 &= \psi(q^5)\phi(q^3)^2 + 4\psi(q^2)\psi(q^3)\phi(q^{15}) \\
 &\quad + 8q^4\psi(q)\psi(q^6)\psi(q^{30}) + 4q^5\phi(q^9)^2\psi(q^{45}) \\
 &\quad + 8q^7f(q^3, q^{15})^2\psi(q^{45}) + 8qf(q^{15}, q^{30})f(q^3, q^{15})\phi(q^9) \\
 &\quad + 4q^2f(q^3, q^{15})^2f(q^{15}, q^{30}) - 4q^2\psi(q^6)\psi(q^{15})\phi(q) \\
 &\quad - 4q^7\psi(q)\psi(q^{60})\phi(q^6) - 4q\psi(q)\psi(q^{12})\phi(q^{30}) \\
 &\quad - 4q\phi(q^3)\psi(q^3)\psi(q^{10}), \tag{5.7}
 \end{aligned}$$

$$\begin{aligned}
 15q\psi(q^{15})\phi(q^{15})^2 &= -q\psi(q^{15})\phi(q)^2 - 4\psi(q)\psi(q^6)\phi(q^5) \\
 &\quad + 4\psi(q^2)\psi(q^5)\phi(q^3) + 4f(q, q^5)^2f(q^5, q^{10}) \\
 &\quad + 4q\psi(q^{15})\phi(q^3)^2 + 8q\psi(q^2)\psi(q^3)\psi(q^{10}) \\
 &\quad - 4q^3\psi(q)\psi(q^{30})\phi(q) - 4\psi(q^3)\psi(q^4)\phi(q^{10}) \\
 &\quad - 4q^2\psi(q^3)\psi(q^{20})\phi(q^2). \tag{5.8}
 \end{aligned}$$

Moreover, using some elbow grease, one checks that the above is just a generating function form of the following statements.

$$3(1, 8, 8, 0, 0, 0) \left( \frac{M}{32} \right) = -W_1(M) - W_2(M) + W_3(M) + W_4(M), \tag{5.9}$$

with  $M \equiv 1 \pmod{8}$ ,  $3|M$ ,

$$5(1, 8, 8, 0, 0, 0) \left( \frac{M}{52} \right) = W_1(M) - W_2(M) + W_3(M) - W_4(M), \tag{5.10}$$

with  $M \equiv 1 \pmod{8}$ ,  $5|M$ ,

$$15(1, 8, 8, 0, 0, 0) \left( \frac{M}{152} \right) = -W_1(M) + W_2(M) + W_3(M) - W_4(M), \tag{5.11}$$

with  $M \equiv 1 \pmod{8}$ ,  $15|M$ . To state (5.9)–(5.11) in an economical manner, we need to develop appropriate notation. Let  $n_i$  be some integer represented by  $TG_i$ , such that  $\gcd(n_i, w) = 1$  for some  $1 \leq w, w|15$ . Next, we define  $\epsilon(i, w)$  as

$$\epsilon(i, w) := (-n_i | w).$$

We also require that

$$\epsilon(i, 1) := 1.$$

It is important to realize that this definition does not depend on one’s choice of  $n_i$ . We are now well-equipped to combine (5.2), (5.9)–(5.11) into the single potent statement

$$w(1, 8, 8, 0, 0, 0) \left( \frac{M}{w^2} \right) = \sum_{i=1}^4 \epsilon(i, w)W_i(M), \tag{5.12}$$

where  $w = 1, 3, 5, 15$  and  $M \equiv 1 \pmod 8$ ,  $w|M$ . As before one can extend (5.12) by using  $\frac{1}{3}(1, 1, 1, 0, 0, 0)(M)$  instead of  $(1, 8, 8, 0, 0, 0)(M)$

$$w(1, 1, 1, 0, 0, 0) \left( \frac{M}{w^2} \right) = 3 \sum_{i=1}^4 \epsilon(i, w) W_i(M), \tag{5.13}$$

where  $w = 1, 3, 5, 15$  and  $M \equiv 1, 2 \pmod 4$ ,  $w|M$ . We forgo the proof.

## 6 S-Genus

Let  $S$  be an odd and square free number and let  $S = p_1 p_2 \dots p_r$  be the prime factorization of  $S$ . In this section, we introduce (what we believe to be new) a notion of  $S$ -genus of ternary forms. To this end, we define an injective map from genera of binary quadratic forms of discriminant  $-8S$  to genera of ternary quadratic forms of discriminant  $16S^2$ . According to Theorem 3.15 in [6], there are exactly  $2^r$  of these genera of binary quadratic forms  $BG_1, \dots, BG_{2^r}$ . Let  $ax^2 + bxy + cy^2$  be some quadratic form in  $BG_i$ , with some  $1 \leq i \leq 2^r$ . We convert it into a ternary form

$$f(x, y, z) := ax^2 + |b|xy + cy^2 + 2Sz^2.$$

Next, we extend  $f$  to a genus  $TG_i$  that contains  $f$ . It can be shown that the map

$$BG_i \rightarrow TG_i, \quad i = 1, 2, \dots, 2^r$$

does not depend on what specific binary form from  $BG_i$  we decided to start with. We can now define the  $S$ -genus as a union

$$S\text{-genus} := TG_1 \cup TG_2 \cup \dots \cup TG_{2^r}. \tag{6.1}$$

We have an elementary proof that the  $TG_i$ 's are disjoint, using all the special features of the construction. Put briefly, any possible  $r$ -tuple with entries  $\pm 1$  occurs as  $((q|p_1), (q|p_2), \dots, (q|p_r))$  for some  $BG_k$  and an odd prime  $q$  represented by a form in  $BG_k$ . Therefore, for some  $i \neq j$ , there is a prime  $p|S$  such that the forms in, say,  $BG_i$  represent only quadratic residues  $\pmod p$  among numbers not divisible by  $p$ , while the forms in  $BG_j$  represent only quadratic nonresidues  $\pmod p$  among numbers not divisible by  $p$ . This separation is carried over to  $TG_i$  and  $TG_j$ , showing disjointness. We did wonder if the structure of the  $S$ -genus were really required for the proof, and the answer seems to be yes. It is easy to show this much: if  $g_1(x, y)$  and  $g_2(x, y)$  are any positive primitive binary quadratic forms of the same discriminant and the same genus, and  $N$  is any positive integer, then  $g_1(x, y) + Nz^2$  and  $g_2(x, y) + Nz^2$  are in the same genus. The converse is

not always true; however, as shown in this example kindly supplied by Wai Kiu Chan, the binaries  $x^2 + 12y^2$  and  $3x^2 + 4y^2$  are in distinct genera, but the ternaries  $x^2 + 12y^2 + 2z^2$  and  $3x^2 + 4y^2 + 2z^2$  are in the same genus.

Let  $n_i$  be some integer represented by  $TG_i$  such that  $\gcd(n_i, w) = 1$  for some  $1 \leq w, w|S$ . For any positive divisor  $w$  of  $S$ , we define  $\epsilon(i, w)$  as

$$\epsilon(i, w) := (-n_i | w), \tag{6.2}$$

and for that matter we always take  $\epsilon(i, 1) := 1$ . Again, we remark that this definition does not depend on our choice of  $n_i$ . For those with some background in quadratic forms, we comment that for the prime divisor  $p$  of  $S$ ,  $\epsilon(i, p) = 1$  if and only if the forms of  $TG_i$  are isotropic over the  $p$ -adic numbers.

We will be using the mass of a genus. As our quadratic forms are positive, each has only a finite set of integral automorphs. On the other hand, any form is equivalent to an infinite set of forms, so when we start with a genus  $G$  and define

$$\text{Mass}(G) := \sum_{f \in G} \frac{1}{|\text{Aut}(f)|}$$

we emphasize that the summation is understood to be over the (finite) set of equivalence classes in  $G$ . Furthermore, for ternary forms we allow our automorphs to have determinants  $\pm 1$ .

We propose that for  $i = 1, 2, \dots, 2^r$

$$M_i = \prod_{j=1}^r \frac{p_j + \epsilon(i, p_j)}{2}, \tag{6.3}$$

where

$$M_i := \sum_{f \in TG_i} \frac{16}{|\text{Aut}(f)|} = 16\text{Mass}(TG_i). \tag{6.4}$$

This seems to generalize Lemma 6.6 on page 152 in [4].

One way to see why the  $S$ -genus is such an appealing construct is to consider a mass for the  $S$ -genus, defined by

$$M(S\text{-genus}) := \sum_{f \in S\text{-genus}} \frac{16}{|\text{Aut}(f)|} = M_1 + \dots + M_{2^r}. \tag{6.5}$$

Remarkably, (6.3) together with the orthogonality relation

$$\sum_{i=1}^{2^r} \epsilon(i, w) = 0, \quad 2 \leq w, w|S. \tag{6.6}$$

implies that

$$M(S\text{-genus}) = S. \tag{6.7}$$

Thus,  $2^r$  genera conspire to produce a startling simplification. Perhaps, more important is the fact that all our identities for ternary forms can be stated in laconic fashion as

$$(1, 1, 1, 0, 0, 0)(M) = \sum_{f \in S\text{-genus}} \frac{48R_f(M)}{|\text{Aut}(f)|}, \tag{6.8}$$

with  $M \equiv 1, 2 \pmod{4}$ .

Recall that

$$(a, b, c, d, e, f)(N) := |\{(x, y, z) \in \mathbf{Z}^3 : ax^2 + by^2 + cz^2 + dyz + ezx + fxy = N\}|,$$

and note that this is 0 if  $N$  is not an integer. For any  $2 \leq w, w|S$

$$w(1, 1, 1, 0, 0, 0) \left( \frac{M}{w^2} \right) = 3 \sum_{i=1}^{2^r} \epsilon(i, w) W_i(M), \tag{6.9}$$

with  $M \equiv 1, 2 \pmod{4}$ ,  $w|M$  and

$$W_i(M) := 16 \sum_{f \in TG_i} \frac{R_f(M)}{|\text{Aut}(f)|}. \tag{6.10}$$

Note that (6.10) allows us to rewrite (6.8) as

$$(1, 1, 1, 0, 0, 0)(M) = 3 \sum_{i=1}^{2^r} W_i(M),$$

with  $M \equiv 1, 2 \pmod{4}$ . We propose that (6.8) and (6.9) hold true for any square free odd  $S$ .

We pause at this point to describe orthogonality relations a bit more fully. First, it follows from properties of the Jacobi symbol that if  $uv|S$ , then  $\epsilon(i, uv) = \epsilon(i, u)\epsilon(i, v)$ . Next, the earlier brief proof that the  $TG_i$ s are disjoint shows us that for any  $i \neq j$ , there is a prime  $p|S$  such that  $\epsilon(i, p)\epsilon(j, p) = -1$ . In turn, for  $i \neq j$  this gives an easy proof that  $\sum_{w|S} \epsilon(i, w)\epsilon(j, w) = 0$ . So, fixing the divisors  $w|S$  in increasing order  $w_1 = 1, \dots, w_{2^r} = S$  and thereby constructing a  $2^r$  by  $2^r$  matrix called  $E$  with entries  $E_{ij} = \epsilon(j, w_i)$  and transpose  $E'$ , we find that  $EE' = E'E = 2^r I$ .

Now, suppose that we have some  $M \equiv 1, 2 \pmod{4}$  for which each  $p_i \parallel M$  so that  $S|M$  and  $\gcd(S^2, M) = S$ , but more to the point if  $w|S$  and  $w > 1$ , then

$M \not\equiv 0 \pmod{w^2}$ . So both sides of (6.9) are 0 in this case. Make a column vector  $W$  with the entries  $W_1(M), W_2(M), \dots, W_{2r}(M)$ , we see that the vector  $EW$  has a nonzero entry in the first position but 0 everywhere else. But if we take a column vector  $T$  with all entries equal to 1, it is also true that the vector  $ET$  has a nonzero entry in the first position but 0 everywhere else. As  $E$  is nonsingular, it follows that vectors  $W$  and  $T$  are linearly dependent, so for these values of  $M$ ,

$$W_1(M) = W_2(M) = \dots = W_{2r}(M).$$

We also point out that for  $M \equiv 1, 2 \pmod{4}$  but  $\gcd(S, M) = 1$ , only a single genus in the  $S$ -genus is allowed to have forms that represent  $M$ . This will be the genus  $TG_i$  that gives  $\epsilon(i, p_j) = (-M | p_j)$  for all  $j$ . What is remarkable about (6.8) is that it continues to be true as  $\gcd(M, S)$  increases, and indeed as  $M$  becomes divisible by high powers of several  $p_i$ . The proofs of (6.3), (6.8), and (6.9) will be given elsewhere.

**Acknowledgments** We are grateful to Benjamin Kane for his insights. His Corollary 1.3 was crucial to our investigation. We would like to thank Manjul Bhargava, Wai Kiu Chan, Frank Garvan, Jonathan Hanke, Byungchan Kim, and Michael Somos for their kind interest and helpful discussions. We are grateful to a thoughtful referee who suggested that (6.3) may be verified by appeal to [5].

## References

1. B.C. Berndt, *Ramanujan's Notebooks, Part III*, Springer, New York, 1991.
2. A. Berkovich, H. Yesilyurt, New identities for 7-cores with prescribed BG-rank, *Discrete Math.*, 308 (2008), 5246–5259.
3. A. J. F. Biagioli, A proof of some identities of Ramanujan using modular forms, *Glasgow Math. J.*, 31 (3) (1989), 271–295.
4. J. W. S. Cassels, *Rational Quadratic Forms*, Academic, London, 1968.
5. J. H. Conway, N. J. A. Sloane, Low-dimensional lattices. IV. The mass formula, *Proc. R. Soc. Lond. A* 419 (1988), 259–286.
6. D. A. Cox, *Primes of the Form  $x^2 + ny^2$ : Fermat Class Field Theory and Complex Multiplication*, Wiley, New York, 1989.
7. L. E. Dickson, *Modern Elementary Theory of Numbers*, The University of Chicago Press, Chicago, 1939.
8. H. Iwaniec, *Topics in Classical Automorphic Forms*, Graduate Studies in Mathematics, v.17, Amer. Mathematical Society, 1997.
9. B.W. Jones, *The Arithmetic Theory of Quadratic Forms*, Mathematical Association of America, 1950.
10. W. C. Jagy, I. Kaplansky, A. Schiemann, There are 913 regular ternary forms, *Mathematika* 44 (1997), 332–341.
11. G. Ligozat, Courbes modularies de genre 1, *Bull. Soc. Math. Fr. [Memoire 43]* (1972), 1–80.
12. M. Newman, Construction and application of a class of modular functions (II), *Proc. Lond. Math. Soc. (3)*, 9 (1959), 373–387.
13. R. A. Rankin, *Modular Forms and Functions*, Cambridge University Press, Cambridge, 1977.



# How Often is $n!$ a Sum of Three Squares?

Jean-Marc Deshouillers and Florian Luca

*Dedicated to the memory of Ramakrishna Alladi, with respect*

**Summary** The positive integers  $n$  such that  $n!$  is a sum of three squares have a density which is equal to  $7/8$ . The key point for the proof of this result is to show that the above sequence is *automatic* and to study the matrix associated to the underlying automaton.

**Mathematics Subject Classification (2000)** 11B35, 11B05

**Key words and phrases** Automatic sequences · Sums of three squares · Density of sequences

## 1 Introduction

In this paper, we prove the following result.

**Theorem 1.** *The estimate*

$$\#\{n \leq x : n! \text{ is a sum of three squares}\} = 7x/8 + O(x^{2/3})$$

*holds.*

Our proof is based on the so-called *automatic sequences*. We refer the reader to the excellent monography [1] for the relevant definitions and results that we use in this paper.

---

J.-M. Deshouillers

IMB, Université de Bordeaux et CNRS, 33405 Talence cedex, France

e-mail: [jean-marc.deshouillers@math.u-bordeaux1.fr](mailto:jean-marc.deshouillers@math.u-bordeaux1.fr)

F. Luca

Instituto de Matemáticas, Universidad Nacional Autónoma de México, C.P. 58089,

Morelia, Michoacán, México

e-mail: [fluca@matmor.unam.mx](mailto:fluca@matmor.unam.mx)

For a positive integer  $n$ , we let

- $\nu(n)$  be the exponent of 2 in the factorization of  $n!$ ;
- $p(n) \in \{0, 1\}$  be the parity of  $\nu(n)$ ; i.e.,  $p(n)$  is chosen such that  $\nu(n) \equiv p(n) \pmod{2}$ ;
- $\varepsilon_1(n), \varepsilon_2(n) \in \{0, 1\}$  be such that

$$n! \equiv 2^{\nu(n)} + \varepsilon_1(n)2^{\nu(n)+1} + \varepsilon_2(n)2^{\nu(n)+2} \pmod{2^{\nu(n)+3}}.$$

In other words, we can write  $n!$  in base 2 as

$$n! = \overline{\cdots \varepsilon_2(n) \varepsilon_1(n) 1 \underbrace{0 \cdots 0}_{\nu(n) \text{ times}}}. \tag{1}$$

Whenever needed, we may complete the expansion of  $n!$  by zeros on the left. For example,

$$\overline{100!} = \overline{11000} = \overline{011000},$$

and so  $\varepsilon_2(4) = 0$ . Here, we point out that it is known that the number of 1's in the binary representation of  $n!$  tends to infinity with  $n$  at a rate at least as large as the logarithm of  $n$  (see [2]).

By a celebrated result of Legendre,  $n!$  is a sum of three squares except when the  $(\varepsilon_2(n), \varepsilon_1(n), p(n)) = (1, 1, 0)$ . Thus, our main result can be reformulated as follows.

**Theorem 2.** *The number of positive integers  $n \leq x$  such that*

$$(\varepsilon_2(n), \varepsilon_1(n), p(n)) = (1, 1, 0)$$

*equals  $x/8 + O(x^{2/3})$ .*

We next introduce some more definitions. For a nonnegative integer  $k$ , we let

- $E_i(k) = (\varepsilon_i(8k), \dots, \varepsilon_i(8k + 7))$ , for  $i = 1, 2$ ;
- $P(k) = (p(8k), \dots, p(8k + 7))$ ;
- $B(k) = (E_2(k), E_1(k), P(k))$ .

In the next section, we prove that the sequence  $(B(0)B(1)\cdots)$ , regarded as  $(E_2(0)E_2(1)\cdots, E_1(0)E_1(1)\cdots, P(0)P(1)\cdots)$  is the fixed point of a certain substitution. In the last section, we study that substitution.

## 2 The Substitution

We first give names to a family of strings of eight elements from  $\{0, 1\}$  so that we can identify them later. We let

$$\begin{aligned}
 A &= 00111100, & B &= 11000011, \\
 J &= 00011101, & K &= 11101101, & L &= 00010010, & M &= 11100010, \\
 S &= 00000110, & T &= 00111010, & U &= 00001001, & V &= 11001010, \\
 W &= 11000101, & X &= 11110110, & Y &= 00110101, & Z &= 11111001.
 \end{aligned}$$

We have the following preliminary result.

**Proposition 1.** (i) *The sequence  $P(0)P(1)P(2)\cdots$  is the fixed point, starting with  $A$ , of the substitution*

$$A \rightarrow AB, \quad B \rightarrow BA.$$

(ii) *The sequence  $E_1(0)E_1(1)E_1(2)\cdots$  is the fixed point, starting with  $J$ , of the substitution*

$$J \rightarrow JK, \quad K \rightarrow MK, \quad L \rightarrow JL, \quad M \rightarrow ML.$$

(iii) *The sequence  $E_2(0)E_2(1)E_2(2)\cdots$  is the fixed point, starting with  $S$ , of the substitution*

$$\begin{aligned}
 S &\rightarrow ST, & T &\rightarrow UV, & U &\rightarrow SW & V &\rightarrow XV, \\
 W &\rightarrow XY, & X &\rightarrow ZT, & Y &\rightarrow UY, & Z &\rightarrow ZW.
 \end{aligned}$$

Indeed, the sequence  $P(0)P(1)P(2)\cdots$  is also the fixed point of the substitution

$$00 \rightarrow 0011, \quad 11 \rightarrow 1100;$$

an avatar of the well-known Thue-Siegel-Morse substitution.

In a similar way, the sequence  $E_1(0)E_1(1)E_1(2)\cdots$  can be seen as a fixed point, starting with  $00$ , of the substitution

$$00 \rightarrow 0001, \quad 01 \rightarrow 1101, \quad 10 \rightarrow 0010, \quad 11 \rightarrow 1110.$$

That is indeed what we are going to prove.

The advantage of considering substitutions acting on strings of length 8 is that Proposition 1 immediately implies that the sequence  $B(0)B(1)B(2)\cdots$  is the fixed point, starting with  $(S, J, A)$ , of the substitution

$$\begin{aligned}
 B_1 &= (S, J, A) \rightarrow (ST, JK, AB), & B_2 &= (T, K, B) \rightarrow (UV, MK, BA), \\
 B_3 &= (U, M, B) \rightarrow (SW, ML, BA), & B_4 &= (V, K, A) \rightarrow (XV, MK, AB), \\
 B_5 &= (S, M, B) \rightarrow (ST, ML, BA), & B_6 &= (W, L, A) \rightarrow (XY, JL, AB), \\
 B_7 &= (X, M, A) \rightarrow (ZT, ML, AB), & B_8 &= (V, K, B) \rightarrow (XV, MK, BA), \\
 B_9 &= (T, L, A) \rightarrow (UV, JL, AB), & B_{10} &= (X, J, A) \rightarrow (ZT, JK, AB), \\
 B_{11} &= (Y, L, B) \rightarrow (UY, JL, BA), & B_{12} &= (Z, M, A) \rightarrow (ZW, ML, AB), \\
 B_{13} &= (T, L, B) \rightarrow (UV, JL, BA), & B_{14} &= (X, M, B) \rightarrow (ZT, ML, BA), \\
 B_{15} &= (U, J, A) \rightarrow (SW, JK, AB), & B_{16} &= (V, L, B) \rightarrow (XV, JL, BA), \\
 B_{17} &= (Z, J, A) \rightarrow (ZW, JK, AB), & B_{18} &= (U, J, B) \rightarrow (SW, JL, BA), \\
 B_{19} &= (Y, L, A) \rightarrow (UY, JL, AB), & B_{20} &= (W, L, B) \rightarrow (XY, JL, BA), \\
 B_{21} &= (V, L, A) \rightarrow (XV, JL, AB), & B_{22} &= (Z, M, B) \rightarrow (ZW, ML, BA), \\
 B_{23} &= (W, K, B) \rightarrow (XY, MK, BA), & B_{24} &= (X, J, B) \rightarrow (ZT, JK, BA), \\
 B_{25} &= (S, J, B) \rightarrow (ST, JK, BA), & B_{26} &= (W, K, A) \rightarrow (XY, MK, AB), \\
 B_{27} &= (Y, K, A) \rightarrow (UY, MK, AB), & B_{28} &= (Z, J, B) \rightarrow (ZW, JK, BA), \\
 B_{29} &= (T, K, A) \rightarrow (UV, MK, AB), & B_{30} &= (Y, K, B) \rightarrow (UY, MK, BA), \\
 B_{31} &= (U, M, A) \rightarrow (SW, ML, AB), & B_{32} &= (S, M, A) \rightarrow (ST, ML, AB).
 \end{aligned}$$

One advantage of that explicit list is to number the relevant triples, numbering which will be used in the next section.

Let us now sketch the proof of Proposition 1. We shall indeed restrict ourselves to the case of the sequence  $E_1(0)E_1(1)E_1(2)\cdots$ . For any positive  $n$ , we consider its expression in the base 2 as we did in (1), namely

$$n = \overline{\cdots\alpha_2(n)\alpha_1(n)10\cdots0}$$

so that  $(\varepsilon_2(n), \varepsilon_1(n)) = (\alpha_2(n!), \alpha_1(n!))$ .

From now on, we shall work in  $\mathbb{Z}/2\mathbb{Z}$ . We leave it to the reader to check the following easy lemma.

**Lemma 1.** *Let  $n$  and  $m$  be two positive integers. We have the following relations:*

$$(\alpha_2(nm), \alpha_1(nm)) = (\alpha_2(n), \alpha_1(n)) + (\alpha_2(m), \alpha_1(m)); \tag{2}$$

$$\alpha_1(2n) = \alpha_1(n), \quad \alpha_1(4n + 1) = 0, \quad \alpha_1(4n + 3) = 1. \tag{3}$$

Of course, the sequence  $(\alpha_2(n))_{n \geq 1}$  possesses properties similar to (3). We now introduce the formal power series in  $\mathbb{Z}/2\mathbb{Z}((X))$  defined by

$$F_1(X) = \sum_{k \geq 1} \alpha_1(k)X^k. \tag{4}$$

Relation (3) leads to the following functional equation for  $F_1(X)$ :

$$\begin{aligned}
 F_1(X) &= \sum_{h=1}^4 \sum_{\substack{k \geq 1 \\ k \equiv h \pmod{4}}} \alpha_1(k) X^k \\
 &= \sum_{k \geq 1} \alpha_1(2k) X^{2k} + \sum_{k \geq 0} X^{4k+3} \\
 &= \sum_{k \geq 1} \alpha_1(k) X^{2k} + \frac{X^3}{1 - X^4} \\
 &= F_1(X^2) + \frac{X^3}{1 - X^4}.
 \end{aligned} \tag{5}$$

The usual convention  $0! = 1 = \overline{001}$  allows us to set  $\alpha_1(0) = 0$ . Relation (2) combined with the convention we just introduced implies that for  $n \geq 0$  we have

$$\varepsilon_1(n) = \sum_{0 \leq k \leq n} \alpha_1(k). \tag{6}$$

Denoting

$$G_1(X) := \sum_{n \geq 0} \varepsilon_1(n) X^n,$$

the relation (6) implies that we have

$$\begin{aligned}
 G_1(X) &= \sum_{n \geq 0} \varepsilon_1(n) X^n = \sum_{n \geq 0} \left( \sum_{0 \leq k \leq n} \alpha_1(k) \right) X^n \\
 &= \sum_{k \geq 0} \alpha_1(k) \sum_{n \geq k} X^n = \sum_{k \geq 0} \alpha_1(k) \frac{X^k}{1 - X} \\
 &= \frac{F_1(X)}{1 - X} = \frac{F_1(X)}{1 + X}.
 \end{aligned} \tag{7}$$

Thus, the functional (5) for  $F_1(X)$  leads to the following functional equation for  $G_1(X)$ :

$$(1 + X)G_1(X) = (1 + X^2)G_1(X^2) + \frac{X^3}{1 - X^4}. \tag{8}$$

The direct consideration of the first terms and that of the coefficients of  $X^k$  for the different congruence classes of  $k$  modulo 4 leads us to the following set of relations which are satisfied by the sequence  $(\varepsilon_1(n))_{n \geq 0}$  for all integers  $\ell \geq 0$ :

- (R0)  $\varepsilon_1(1) = \varepsilon_1(2) = 0, \varepsilon_1(3) = 1;$
- (R1)  $\varepsilon_1(4\ell) + \varepsilon_1(4\ell - 1) + \varepsilon_1(2\ell) + \varepsilon_1(2\ell - 1) = 0;$
- (R2)  $\varepsilon_1(4\ell + 1) + \varepsilon_1(4\ell) = 0;$

$$(R3) \quad \varepsilon_1(4\ell + 2) + \varepsilon_1(4\ell + 1) + \varepsilon_1(2\ell + 1) + \varepsilon_1(2\ell) = 0;$$

$$(R4) \quad \varepsilon_1(4\ell + 3) + \varepsilon_1(4\ell + 2) + 1 = 0.$$

We first remark that the set of relations (R) completely determines the sequence  $(\varepsilon_1(n))_{n \geq 0}$ . In order to prove the second assertion in Proposition 1, it is enough to show that the sequence  $(\varepsilon(n))_{n \geq 0}$  which is the fixed point, starting with 00, of the substitution

$$00 \rightarrow 0001, \quad 01 \rightarrow 1101, \quad 10 \rightarrow 0010, \quad 11 \rightarrow 1110,$$

is indeed the sequence  $(\varepsilon_1(n))_{n \geq 0}$ . For that purpose, it is enough to show that the sequence  $(\varepsilon(n))_{n \geq 0}$  satisfies the set of relations (R). Well, relations (R0)–(R4) are easily checked. Relation (R1) is slightly more subtle to check, so let us justify it in detail. For any integer  $m \geq 0$ , we have

$$\varepsilon(4m) + \varepsilon(4m + 1) = 0 \quad \text{and} \quad \varepsilon(4m + 2) + \varepsilon(4m + 3) = 1,$$

which is due to the fact that the only possible blocks for the string

$$\varepsilon(4m)\varepsilon(4m + 1)\varepsilon(4m + 2)\varepsilon(4m + 3) \quad \text{are} \quad \{0001, 1101, 0010, 1110\}.$$

This implies that

$$\varepsilon(2\ell - 2) + \varepsilon(2\ell - 1) + \varepsilon(2\ell) + \varepsilon(2\ell + 1) + 1 = 0.$$

But by looking at the substitution, we clearly see that

$$1 + \varepsilon(2\ell - 2) = 1 + \varepsilon(2(\ell - 1)) = \varepsilon(4(\ell - 1) + 3) = \varepsilon(4\ell - 1), \quad \varepsilon(2\ell + 1) = \varepsilon(4\ell).$$

We thus have

$$\varepsilon(4\ell - 1) + \varepsilon(2\ell - 1) + \varepsilon(2\ell) + \varepsilon(4\ell) = 0,$$

which implies that  $(\varepsilon(n))_{n \geq 0}$  satisfies also relation (R1). Thus,  $\varepsilon = \varepsilon_1$ . The second assertion of Proposition 1 is therefore proved. The first one is much easier, and the third one can be proved by arguments very similar to the above one.

### 3 Proof of Theorem 2

We introduce a  $32 \times 32$  matrix called  $M$ , with coefficients  $m_{ij}$  which are all 0 except when there is either a term  $B_j \rightarrow B_k B_i$  or a term  $B_j \rightarrow B_i B_k$  in the set of the relations defining the substitution associated to the sequence  $(B(k))_{k \geq 0}$ . For example, the first relation is  $B_1 \rightarrow B_1 B_2$ . Thus, the first column of the matrix  $M$  consists of elements which are successively 1, 1 and the remaining ones are equal to 0. A further example is the following: the only relations in which the second



**Lemma 2.** *Let  $a$  be a positive integer. Consider the finite sequence  $\mathcal{B}(a) = B(a), B(a + 1), \dots, B(2a - 1)$  with  $a$  elements. For  $1 \leq i \leq 32$ , we let  $\Pi_i(a)$  denote the number of occurrences of  $B_i$  in the sequence  $\mathcal{B}(a)$ . We also let  $\Pi(a)$  denote the column vector  $(\Pi_1(a), \dots, \Pi_{32}(a))^T$ . Then, for any integer  $h \geq 0$ , we have*

$$\Pi(2^h a) = M^h \Pi(a). \tag{9}$$

We now give the key property of the sequence  $(M^h)_{h \geq 0}$  of powers of  $M$ .

**Proposition 2.** (i)  $\lambda_1 = 2$  is an eigenvalue of  $M$  of multiplicity 1. All other eigenvalues of  $M$  are  $\leq 2^{1/2}$  in absolute value.  
(ii) The sequence  $(2^{-h} M^h)_{h \geq 0}$  tends to the matrix  $J$  the entries of which are all equal to  $1/32$ .

*Proof.* Part (i) of Proposition 2 was easily confirmed by MAPLE. Part (ii) of Proposition 2 is a simple application of the Perron–Frobenius Theorem, which is classical in the study of homogeneous Markov chains. Indeed, let us consider the matrix

$$S := (1/2) M.$$

The following properties are straightforward to verify:

- (S1)  $S$  is *stochastic* meaning that all its entries are nonnegative and the sum of the entries in each column is equal to 1;
- (S2) The transposed  $S^T$  of  $S$  is also stochastic;
- (S3)  $S^{16}$  has only positive entries.

By the Perron–Frobenius Theorem, properties (S1) and (S3) imply that the sequence  $(S^h)_{h \geq 0}$  converges to a matrix with identical columns. Properties (S2) and (S3) imply that  $(S^h)_{h \geq 0}$  converges to a matrix with identical rows. Hence, the desired result (ii). □

We are now in a position to prove Theorem 2.

*Proof of Theorem 2.* We let  $x$  be large and put  $h := \lfloor c \log x \rfloor$ , where  $c := 2/(3 \log 2)$ . By Proposition 2, we have that

$$(M^h)_{i,j} = 2^{h-5} + O(2^{h/2}) \quad \text{for all } i, j \in \{1, \dots, 32\}. \tag{10}$$

Let  $a := \lfloor 8x/2^h \rfloor$ . Observe that  $a = O(x^{1/3})$ . Furthermore, we obviously have

$$\# \left( [x, 2x] \Delta [2^h 8a, 2^{h+1} 8a] \right) = x + O(2^h) = x + O(x^{2/3}), \tag{11}$$

where for two subsets  $\mathcal{B}$  and  $\mathcal{C}$  of real numbers, we use  $\mathcal{B} \Delta \mathcal{C}$  for the symmetric difference  $\mathcal{B} \Delta \mathcal{C} = (\mathcal{B} \cup \mathcal{C}) \setminus (\mathcal{B} \cap \mathcal{C})$ .

Although we have a very limited knowledge of the vector  $\Pi(a)$  introduced in Lemma 2, namely we only know that the sum of its entries is  $a$ , relations (9) and



(10) imply that the coordinates of the vector  $\Pi(2^h a)$  are almost all equal. More precisely, we have that the estimate

$$\left| \Pi_i(2^h a) - 2^{h-5} a \right| = O(2^{h/2} a) = O(x^{2/3}) \quad \text{holds for all } i \in \{1, \dots, 32\}. \tag{12}$$

Our aim is to evaluate the number of integers  $n$  in  $[x, 2x)$  for which

$$(\varepsilon_2(n), \varepsilon_1(n), p(n)) = (1, 1, 0).$$

Via relation (11), it follows that by discarding at most  $O(x^{2/3})$  such values of  $n$ , we may assume that  $n \in [2^h 8a, 2^{h+1} 8a)$ . Our life would have been very easy if there were exactly one triple equal to  $(1, 1, 0)$  in each  $B_i$  for  $i = 1, \dots, 32$ . This is not the case. For example, in the 8-tuple  $B_3 = (U, M, B)$ , there is no triple  $(1, 1, 0)$ , whereas in the 8-tuple  $B_{14} = (X, M, B)$ , there are three such triples. However, in the totality of the 32 tuples  $B_i$  for  $i = 1, \dots, 32$ , there are exactly 32 triples  $(1, 1, 0)$  so that *on the average* there is exactly one triple  $(1, 1, 0)$  for each  $B_i$  for  $i = 1, \dots, 32$ . Combining this with the important relation (12), which states the equidistribution of the blocks  $B_i$  for  $i = 1, \dots, 32$ , implies that the counting function of the set of positive integers  $n \leq x$  such that  $(\varepsilon_2(n), \varepsilon_1(n), p(n)) = (1, 1, 0)$  is  $x/8 + O(x^{2/3})$ , which proves Theorem 2, and therefore also Theorem 1.

## References

[1] J-P. Allouche and J. Shallit, *Automatic Sequences. Theory, Applications, Generalizations*, Cambridge University Press, Cambridge, UK, (2003).  
 [2] F. Luca, “The number of nonzero digits of  $n!$ ”. *Canadian Math. Bull.* **45** (2002), 115–118.

# Eulerian Polynomials: From Euler's Time to the Present

Dominique Foata\*

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** The polynomials commonly called “Eulerian” today have been introduced by Euler himself in his famous book “*Institutiones calculi differentialis cum eius usu in analysi finitorum ac Doctrina serierum*” [5, Chap. VII], back in 1755. They have been since thoroughly studied, extended, applied. The purpose of the present paper is to go back to Euler’s memoir, find out his motivation and reproduce his derivation, surprisingly partially forgotten. The rebirth of those polynomials in a  $q$ -environment is due to Carlitz two centuries after Euler. A brief overview of Carlitz’s method is given, as well as a short presentation of combinatorial works dealing with natural extensions of the classical Eulerian polynomials.

**Mathematics Subject Classification (2000)** 01A50, 05A15, 05A30, 33B10

**Key words and phrases** Eulerian polynomials · Bernoulli numbers · Genocchi numbers · Tangent numbers ·  $q$ -Eulerian polynomials

## 1 Introduction

Before Euler’s time Jacques Bernoulli had already introduced his famous *Bernoulli numbers*, denoted by  $B_{2n}$  ( $n \geq 1$ ) in the sequel. Those numbers can be defined by their generating function as

$$\frac{u}{e^u - 1} = 1 - \frac{u}{2} + \sum_{n \geq 1} \frac{u^{2n}}{(2n)!} (-1)^{n+1} B_{2n}, \quad (1.1)$$

---

\* Invited address at the 10-th Annual Ulam Colloquium, University of Florida, Gainesville, February 18, 2008.

D. Foata

Institut Lothaire, 1, rue Murner, F-67000 Strasbourg, France

e-mail: [foata@unistra.fr](mailto:foata@unistra.fr)

their first values being shown in the table:

|          |     |      |      |      |      |          |     |
|----------|-----|------|------|------|------|----------|-----|
| $n$      | 1   | 2    | 3    | 4    | 5    | 6        | 7   |
| $B_{2n}$ | 1/6 | 1/30 | 1/42 | 1/30 | 5/66 | 691/2730 | 7/6 |

(1.2)

Note that besides the first term  $-u/2$ , there is no term of odd rank in the series expansion (1.1), a property easy to verify. On the other hand, the factor  $(-1)^{n+1}$  in formula (1.1) and the first values shown in the above table suggest that those numbers are all *positive*, which is true.

Jacques Bernoulli ([1], p. 95–97) had introduced the numbers called after his name to evaluate the sum of the  $n$ -th powers of the first  $m$  integers. He then proved the following summation formula

$$\sum_{i=1}^m i^n = \frac{m^{n+1}}{n+1} + \frac{m^n}{2} + \frac{1}{n+1} \sum_{1 \leq r \leq n/2} \binom{n+1}{2r} m^{n-2r+1} (-1)^{r+1} B_{2r}, \quad (1.3)$$

where  $n, m \geq 1$ . Once the first  $\lfloor n/2 \rfloor$  Bernoulli numbers have been determined (and there are quick ways of getting them, directly derived from (1.1)), there are only  $2 + \lfloor n/2 \rfloor$  terms to sum on the right-hand side for evaluating  $\sum_{i=1}^m i^n$ , whatever the number  $m$ .

Euler certainly had this summation formula in mind when he looked for an expression for the *alternating* sum  $\sum_{i=1}^k i^n (-1)^i$ . Instead of the Bernoulli numbers, he introduced another sequence  $(G_{2n})$  ( $n \geq 1$ ) of integers, later called *Genocchi numbers*, after the name of Peano’s mentor [11]. They are related to the Bernoulli numbers by the relation

$$G_{2n} := 2(2^{2n} - 1)B_{2n} \quad (n \geq 1), \quad (1.4)$$

their first values being shown in the next table.

|          |     |      |      |      |      |           |        |
|----------|-----|------|------|------|------|-----------|--------|
| $n$      | 1   | 2    | 3    | 4    | 5    | 6         | 7      |
| $B_{2n}$ | 1/6 | 1/30 | 1/42 | 1/30 | 5/66 | 691/2,730 | 7/6    |
| $G_{2n}$ | 1   | 1    | 3    | 17   | 155  | 2,073     | 38,227 |

Of course, it is not obvious that the numbers  $G_{2n}$  defined by (1.4) are integers and furthermore odd integers. This is a consequence of the little Fermat theorem and the celebrated von Staudt–Clausen theorem (see, for instance, the classical treatise by Nielsen [16] entirely devoted to the studies of Bernoulli numbers and related sequences) that asserts that the expression

$$(-1)^n B_{2n} - \sum_p \frac{1}{p}, \quad (1.5)$$

where the sum is over all prime numbers  $p$  such that  $(p - 1) \mid 2n$ , is an *integer*. From (1.1) and (1.4), we can easily obtain the generating function for the Genocchi numbers in the form:

$$\frac{2u}{e^u + 1} = u + \sum_{n \geq 1} \frac{u^{2n}}{(2n)!} (-1)^n G_{2n}. \quad (1.6)$$

The formula obtained by Euler for the alternating sum  $\sum_{i=1}^m i^n (-1)^i$  is quite analogous to Bernoulli's formula (1.3). It suffices to know the first  $\lfloor n/2 \rfloor$  Genocchi numbers to complete the computation. Euler's formula is the following.

**Theorem 1.1.** *Let  $(G_{2n})$  ( $n \geq 1$ ) be the sequence of numbers defined by relation (1.4) (or by (1.6)). If  $n = 2p \geq 2$ , then*

$$\sum_{k=1}^m k^{2p} (-1)^k = (-1)^m \frac{m^{2p}}{2} + \sum_{k=1}^p \binom{2p}{2k-1} (-1)^{m+k+1} \frac{G_{2k}}{4k} m^{2p-2k+1}, \quad (1.7)$$

while, if  $n = 2p + 1$ , the following holds:

$$\begin{aligned} \sum_{k=1}^m k^{2p+1} (-1)^k &= (-1)^m \frac{m^{2p+1}}{2} + \sum_{k=1}^{p+1} \binom{2p+1}{2k-1} (-1)^{m+k+1} \frac{G_{2k}}{4k} m^{2p-2k+2} \\ &\quad + (-1)^{p+1} \frac{G_{2p+2}}{4(p+1)}. \end{aligned} \quad (1.8)$$

The first values of the numbers  $G_{2n}$  do appear in Euler's memoir. However, he did not bother proving that they were odd integral numbers. The two identities (1.7) and (1.8) have not become classical, in contrast to Bernoulli's formula (1.3), but the effective discovery of the *Eulerian polynomials* made by Euler for deriving (1.7) and (1.8) has been fundamental in numerous arithmetical and combinatorial studies in modern times. Our purpose in the sequel is to present Euler's discovery by making a contemporary reading of his calculation. Two centuries after Euler, the Eulerian polynomials were given an extension in the algebra of the  $q$ -series, thanks to Carlitz [2]. Our intention is also to discuss some aspects of that  $q$ -extension with a short detour to contemporary works in Combinatorics.

It is a great privilege for me to have met Professor Alladi Ramakrishnan, the brilliant Indian physicist and mathematician, who has been influential in so many fields, from Probability to Relativity Theory. He was kind enough to listen to my 2008 University of Florida Ulam Colloquium address and told me of his great admiration for Euler. I am pleased and honored therefore to dedicate the present text to his memory.

## 2 Euler's Definition of the Eulerian Polynomials

Let  $(a_i(x))$  ( $i \geq 0$ ) be a sequence of polynomials in the variable  $x$  and let  $t$  be another variable. For each *positive* integer  $m$ , we have the banal identity:

$$\sum_{i=0}^{m-1} a_i(x) t^i = \frac{1}{t} \sum_{i=1}^m a_{i-1}(x) t^i = a_0(x) + \sum_{i=1}^m a_i(x) t^i - a_m(x) t^m. \quad (2.1)$$

Now, consider the operator  $\Delta = \sum_{k \geq 0} \frac{(-1)^k}{k!} D^k$ , where  $D$  is the usual differential operator. Starting with a given polynomial  $p(x)$  define

$$a_i(x) := \Delta^{m-i} p(x) \quad (0 \leq i \leq m);$$

$$\mathbf{S}(p(x), t) := \sum_{i=1}^m \Delta^{m-i} p(x) t^i = \sum_{i=1}^m a_i(x) t^i.$$

As  $D$  commutes with  $\Delta$ , we have  $\mathbf{S}(D^k p(x), t) = D^k \mathbf{S}(p(x), t)$  for each  $k \geq 0$  so that using (2.1), we get:

$$\begin{aligned} a_0(x) + \mathbf{S}(p(x), t) - p(x)t^m &= a_0(x) + \sum_{i=1}^m a_i(x)t^i - a_m(x)t^m \\ &= \frac{1}{t} \sum_{i=1}^m a_{i-1}(x)t^i = \frac{1}{t} \sum_{i=1}^m \Delta a_i(x) t^i \\ &= \frac{1}{t} \sum_{i=1}^m \sum_{k \geq 0} \frac{(-1)^k}{k!} D^k a_i(x) t^i \\ &= \frac{1}{t} \sum_{k \geq 0} \frac{(-1)^k}{k!} \sum_{i=1}^m D^k a_i(x) t^i \\ &= \frac{1}{t} \sum_{k \geq 0} \frac{(-1)^k}{k!} \mathbf{S}(D^k p(x), t) \\ &= \frac{1}{t} \left( \mathbf{S}(p(x), t) + \sum_{k \geq 1} \frac{(-1)^k}{k!} \mathbf{S}(D^k p(x), t) \right). \end{aligned}$$

Hence,

$$\mathbf{S}(p(x), t) = \frac{1}{t-1} \left( p(m)t^{m+1} - a_0(x)t + \sum_{k \geq 1} \frac{(-1)^k}{k!} \mathbf{S}(D^k p(x), t) \right). \tag{2.2}$$

We now work out specializations of (2.2) for the monomials  $p(x) = x^n$  ( $n \geq 0$ ) and for  $x = m$ . First, we verify that:

$$\Delta^i x^n = (x - i)^n \quad (0 \leq i \leq m). \tag{2.3}$$

It is true for  $i = 0$  and all  $n \geq 0$ . When  $i \geq 1$ , we have

$$\begin{aligned} \Delta^{i+1} x^n &= \Delta \Delta^i x^n = \Delta(x - i)^n \\ &= \sum_{k=0}^n (-1)^k \binom{n}{k} (x - i)^{n-k} = (x - (i + 1))^n. \quad \square \end{aligned}$$

Consequently,

$$S(x^n, t) = \sum_{i=1}^m \Delta^{m-i} x^n t^i = \sum_{i=1}^m (x - (m - i))^n t^i.$$

Let  $S(x^n, t) := \mathbf{S}(x^n, t) |_{\{x=m\}}$  so that

$$S(x^n, t) = \sum_{i=1}^m i^n t^i. \tag{2.4}$$

On the other hand,

$$a_0(m) = \Delta^m x^n |_{\{x=m\}} = \begin{cases} 1, & \text{si } n = 0; \\ 0, & \text{si } n \geq 1. \end{cases} \tag{2.5}$$

Identity (1.2), when  $p(x) = x^n$  and  $x = m$ , becomes:

$$\begin{aligned} S(x^n, t) &= \frac{1}{t-1} \left( m^n t^{m+1} - a_0(m)t + \sum_{k=1}^n \frac{(-1)^k}{k!} S(D^k x^n, t) \right) \\ &= \frac{1}{t-1} \left( m^n t^{m+1} - a_0(m)t \right. \\ &\quad \left. + \sum_{k=1}^n \frac{(-1)^k}{k!} S(n(n-1) \cdots (n-k+1)x^{n-k}, t) \right), \end{aligned}$$

and finally,

$$S(x^n, t) = \frac{1}{t-1} \left( m^n t^{m+1} - a_0(m)t + \sum_{k=1}^n (-1)^k \binom{n}{k} S(x^{n-k}, t) \right). \tag{2.6}$$

Another proof of identity (2.6) consists, for  $m \geq 1, 0 \leq k \leq m$  and  $n \geq 0$ , of letting  $s(k, m, n) := (-1)^k \binom{n}{k} \sum_{i=1}^m i^{n-k} t^i$  so that  $s(0, m, n) = \sum_{i=1}^m i^n t^i$  and of deriving the previous identity under the form

$$s(0, m, n) = \frac{1}{t-1} \left( m^n t^m - a_0(m)t + \sum_{k=1}^n s(k, m, n) \right),$$

still assuming that  $a_0(m) = 1$  if  $n = 0$  and  $a_0(m) = 0$  if  $n \geq 1$ . The identity is banal for  $m = 1$  and every  $n \geq 0$ . When  $m \geq 2$ , we have the relations

$$s(k, m, n) = (-1)^k \binom{n}{k} t^m + s(k, m-1, n)$$

so that

$$\sum_{k=1}^n s(k, m, n) = ((m-1)^n - m^n)t^m + \sum_{k=1}^n s(k, m-1, n).$$

Hence, by induction on  $m \geq 1$

$$\begin{aligned} s(0, m, n) &= \sum_{i=1}^m i^n t^i = m^n t^m + \sum_{i=1}^{m-1} i^n t^n = m^n t^m + s(0, m-1, n) \\ &= m^n t^m + \frac{1}{t-1} \left( (m-1)t^m - a_0(m)t + \sum_{k=1}^n s(k, m-1, n) \right) \\ &= m^n t^m + \frac{1}{t-1} \left( -a_0(m)t + \sum_{k=1}^n s(k, m, n) + m^n t^m \right) \\ &= \frac{1}{t-1} \left( m^n t^m - a_0(m)t + \sum_{k=1}^n s(k, m, n) \right). \quad \square \end{aligned}$$

As  $a_0(m) = 1$  when  $n = 0$ , identity (2.6) yields

$$S(1, t) = \sum_{i=1}^m t^i = \frac{1}{t-1} (t^{m+1} - t) = \frac{t(t^m - 1)}{t-1},$$

which is the classical formula for geometric progressions.

For discovering the polynomials, later called ‘‘Eulerian,’’ Euler further rewrites identity (2.6) for  $n = 1, 2, 3, \dots$  by reporting the expressions already derived for  $S(x^k, t)$  ( $0 \leq k \leq n-1$ ) into  $S(x^n, t)$ . As  $a_0(m) = 0$  for  $n \geq 1$ , the successive reports lead to

$$\begin{aligned} S(x, t) &= \frac{1}{t-1} (mt^{m+1} - S(1, t)) = \frac{1}{t-1} \left( mt^{m+1} - \frac{t(t^m - 1)}{t-1} \right) \\ &= \frac{mt^{m+1}}{t-1} \mathbf{1} - \frac{t(t^m - 1)}{(t-1)^2} \mathbf{1}; \\ S(x^2, t) &= \frac{1}{t-1} \left( m^2 t^{m+1} - 2S(x, t) + S(1, t) \right) \\ &= \frac{m^2 t^{m+1}}{t-1} \mathbf{1} - \frac{2mt^{m+1}}{(t-1)^2} \mathbf{1} + \frac{t(t^m - 1)}{(t-1)^3} (\mathbf{t} + \mathbf{1}); \\ S(x^3, t) &= \frac{1}{t-1} \left( m^3 t^{m+1} - 3S(x^2, t) + 3S(x, t) - S(1, t) \right) \\ &= \frac{m^3 t^{m+1}}{t-1} \mathbf{1} - \frac{3m^2 t^{m+1}}{(t-1)^2} \mathbf{1} + \frac{3mt^{m+1}}{(t-1)^3} (\mathbf{t} + \mathbf{1}) \\ &\quad - \frac{t(t^m - 1)}{(t-1)^4} (\mathbf{t}^2 + 4\mathbf{t} + \mathbf{1}); \end{aligned}$$

$$\begin{aligned}
 S(x^4, t) &= \frac{1}{t-1} \left( m^4 t^{m+1} - 4S(x^3, t) + 6S(x^2, t) - 4S(x, t) + S(1, t) \right) \\
 &= \frac{m^4 t^{m+1}}{t-1} \mathbf{1} - \frac{4m^3 t^{m+1}}{(t-1)^2} \mathbf{1} + \frac{6m^2 t^{m+1}}{(t-1)^3} (\mathbf{t} + \mathbf{1}) \\
 &\quad - \frac{4m t^{m+1}}{(t-1)^4} (\mathbf{t}^2 + \mathbf{4t} + \mathbf{1}) + \frac{t(t^m - 1)}{(t-1)^5} (\mathbf{t}^3 + \mathbf{11t}^2 + \mathbf{11t} + \mathbf{1}); \\
 S(x^5, t) &= \frac{1}{t-1} \left( m^5 t^{m+1} - 5S(x^4, t) + 10S(x^3, t) \right. \\
 &\quad \left. - 10S(x^2, t) + 5S(x, t) - S(1, t) \right) \\
 &= \frac{m^5 t^{m+1}}{t-1} \mathbf{1} - \frac{5m^4 t^{m+1}}{(t-1)^2} \mathbf{1} + \frac{10m^3 t^{m+1}}{(t-1)^3} (\mathbf{t} + \mathbf{1}) \\
 &\quad - \frac{10m^2 t^{m+1}}{(t-1)^4} (\mathbf{t}^2 + \mathbf{4t} + \mathbf{1}) + \frac{5m t^{m+1}}{(t-1)^5} (\mathbf{t}^3 + \mathbf{11t}^2 + \mathbf{11t} + \mathbf{1}) \\
 &\quad - \frac{t(t^m - 1)}{(t-1)^6} (\mathbf{t}^4 + \mathbf{26t}^3 + \mathbf{66t}^2 + \mathbf{26t} + \mathbf{1}).
 \end{aligned}$$

Let  $A_n(t)$  be the coefficient of  $\frac{t(t^m - 1)}{(t-1)^{n+2}}$  in the above expansions of  $S(x^n, t)$  for  $n = 0, 1, 2, 3, 4, 5$  so that  $A_0(t) = A_1(t) = 1, A_2(t) = t + 1, A_3(t) = t^2 + 4t + 1, A_4(t) = t^3 + 11t^2 + 11t + 1, A_5(t) = t^4 + 26t^3 + 66t^2 + 26t + 1$ . We observe that the expression of  $A_n(t)$  in the expansion of  $S(x^n, t)$  is obtained by means of the formula

$$A_n(t) = \sum_{k=0}^{n-1} \binom{n}{k} A_k(t) (t-1)^{n-1-k}. \tag{2.7}$$

We also see that  $S(x^n, t) = \sum_{i=1}^m i^n t^i$  can be expressed as

$$S(x^n, t) = \sum_{l=1}^n (-1)^{n+l} \binom{n}{l} \frac{t^{m+1} A_{n-l}(t)}{(t-1)^{n-l+1}} m^l + (-1)^n \frac{t(t^m - 1)}{(t-1)^{n+1}} A_n(t). \tag{2.8}$$

Once those two facts have been observed, it remains to prove that the new expression found for  $S(x^n, t)$  holds for every  $n \geq 0$  as stated next.

**Theorem 2.1.** *Let  $A_0(t) := 1$  and let  $(A_n(t))$  ( $n \geq 0$ ) be the sequence of polynomials inductively defined by (2.7). Then identity (2.8) holds.*

*Proof.* For proving such a theorem today we proceed by induction, reporting (2.7) into (2.6) and making an appropriate change of variables in the finite sums. This is the first proof that is now presented. At the time of Euler the “ $\sum$ ” notation does not exist, and of course not the double sums! For example, (2.6) is displayed as:



$$\begin{aligned}
 S.x^n p^x &= \frac{1}{p-1} \left( x^n p^{x+1} - A^p - nS.x^{n-1} p^x + \frac{n(n-1)}{1 \cdot 2} S.x^{n-2} p^x \right. \\
 &\quad - \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} S.x^{n-2} p^x \\
 &\quad \left. + \frac{n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3 \cdot 4} S.x^{n-4} p^x - \&c \right).
 \end{aligned}$$

Euler is then led to imagine another proof based on fact on the concept of linear independence.

For the first proof, start with identity (2.6) and apply the induction hypothesis to all the terms  $S(x^{n-k}, t)$  of the sum, with  $k$  running from 1 to  $n$ :

$$\begin{aligned}
 S(x^n, t) &= \frac{1}{t-1} \left( m^n t^{m+1} + \sum_{k=1}^n (-1)^k \binom{n}{k} S(x^{n-k}, t) \right) \\
 &= \frac{1}{t-1} \left( m^n t^{m+1} + \sum_{j=0}^{n-1} (-1)^{n-j} \binom{n}{j} S(x^j, t) \right) \\
 &= \frac{1}{t-1} \left( m^n t^{m+1} + \sum_{j=0}^{n-1} (-1)^{n-j} \binom{n}{j} \right. \\
 &\quad \left. \sum_{l=1}^j (-1)^{j+l} \binom{j}{l} \frac{t^{m+1} A_{j-l}(t)}{(t-1)^{n-l+1}} m^l + (-1)^j \frac{t(t^m-1)}{(t-1)^{j+1}} A_j(t) \right) \\
 &= \frac{m^n t^{m+1}}{t-1} + \sum_{l=1}^{n-1} \sum_{j=l}^{n-1} (-1)^{n+l} \frac{n!}{(n-j)! l! (j-l)!} \frac{t^{m+1} A_{j-l}(t)}{(t-1)^{j-l+2}} m^l \\
 &\quad + (-1)^n \frac{t(t^m-1)}{(t-1)^{n+1}} \sum_{j=0}^{n-1} \binom{n}{j} A_j(t) (t-1)^{n-j-1}.
 \end{aligned}$$

With  $j = k - l$ , we deduce

$$\begin{aligned}
 S(x^n, t) &= \frac{m^n t^{m+1}}{t-1} + \sum_{l=1}^{n-1} (-1)^{n+l} \binom{n}{l} t^{m+1} m^l \sum_{k=0}^{n-1-l} \binom{n-l}{k} \frac{A_k(t)}{(t-1)^{k+2}} \\
 &\quad + (-1)^n \frac{t(t^m-1)}{(t-1)^{n+1}} A_n(t) \\
 &= \frac{m^n t^{m+1}}{t-1} + \sum_{l=1}^{n-1} (-1)^{n+l} \binom{n}{l} \frac{t^{m+1} m^l}{(t-1)^{n-l+1}} \\
 &\quad \sum_{k=0}^{n-1-l} \binom{n-l}{k} A_k(t) (t-1)^{n-l-k-1} + (-1)^n \frac{t(t^m-1)}{(t-1)^{n+1}} A_n(t)
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{m^n t^{m+1}}{t-1} + \sum_{l=1}^{n-1} (-1)^{n+l} \binom{n}{l} \frac{t^{m+1} m^l}{(t-1)^{n-l+1}} A_{n-l}(t) \\
 &\quad + (-1)^n \frac{t(t^m-1)}{(t-1)^{n+1}} A_n(t) \\
 &= \sum_{l=1}^n (-1)^{n+l} \binom{n}{l} \frac{t^{m+1} m^l}{(t-1)^{n-l+1}} A_{n-l}(t) + (-1)^n \frac{t(t^m-1)}{(t-1)^{n+1}} A_n(t). \quad \square
 \end{aligned}$$

The original proof by Euler can be reproduced as follows. With  $p(x) = x^n$  ( $n \geq 1$ ) and  $l = 0, 1, \dots, n$  let

$$\begin{aligned}
 Y_l &:= S(D^l p(x), t); \\
 Z_l &:= \begin{cases} \frac{D^l p(m)}{t-1} t^{m+1}, & \text{si } 0 \leq l \leq n-1; \\ \frac{D^n p(m)}{t-1} (t^{m+1} - t), & \text{si } l = n. \end{cases}
 \end{aligned}$$

In formula (2.2), successively replace  $p(x)$  by  $D^l p(x)$  for  $l = 0, 1, \dots, n$ . We obtain

$$Y_l = Z_l + \sum_{j=1}^{n-l} \frac{(-1)^j}{j!} \frac{1}{t-1} Y_{l+j} \quad (0 \leq l \leq n), \tag{2.9}$$

remembering that  $a_0(x)$  is null for  $x = m$ , and 1 for  $Y_n = S(D^n p(x), t)$ . Now, each of the two sequences  $(Y_0, Y_1, \dots, Y_n)$ ,  $(Z_0, Z_1, \dots, Z_n)$  is a basis for the algebra of polynomials of degree at most equal to  $n$ , since both  $Y_l$  and  $Z_l$  are polynomials of degree  $n-l$  ( $l = 0, 1, \dots, n$ ). Accordingly, there exists a sequence  $(b_0, b_1, \dots, b_n)$  of coefficients such that

$$Y_0 = \sum_{l=0}^n b_l (-1)^l Z_l \tag{2.10}$$

so that, by using (2.9),

$$\begin{aligned}
 Y_0 &= \sum_{l=0}^n b_l (-1)^l \left( Y_l - \sum_{j=1}^{n-l} \frac{(-1)^j}{j!} \frac{1}{t-1} Y_{l+j} \right) \\
 &= \sum_{k=0}^n \left( b_k (-1)^k - \sum_{\substack{l+j=k, \\ l \geq 0, j \geq 1}} b_l (-1)^l \frac{(-1)^j}{j!} \frac{1}{t-1} \right) Y_k.
 \end{aligned}$$

Hence,  $b_0 = 1$ ; furthermore, for  $k = 1, 2, \dots, n$ ,

$$b_k = \frac{1}{t-1} \left( \frac{b_{k-1}}{1!} + \frac{b_{k-2}}{2!} + \dots + \frac{b_1}{(k-1)!} + \frac{b_0}{k!} \right)$$

or still

$$k! (t - 1)^k b_k = \sum_{l=0}^{k-1} \binom{k}{l} l! (t - 1)^l b_l (t - 1)^{k-1-l}.$$

By comparison with the induction formula for the Eulerian polynomials

$$b_n = \frac{1}{n! (t - 1)^n} A_n(t) \quad (n \geq 0).$$

Finally, reporting  $b_n$  into (2.10) yields (2.8). □

### 3 A Formulary for the Eulerian Polynomials

The polynomials  $A_n(t)$  ( $n = 0, 1, \dots$ ), inductively defined by  $A_0(t) := 1$  and identity (2.7) for  $n \geq 1$  are unanimously called *Eulerian polynomials*. It is hard to trace back the exact origin of their christening.

Form the exponential generating function

$$A(t, u) := \sum_{n \geq 0} A_n(t) \frac{u^n}{n!}.$$

Then (2.7) is equivalent to the identity

$$\begin{aligned} A(t, u) &= 1 + \sum_{n \geq 1} \sum_{\substack{k+l=n \\ 0 \leq k \leq n-1}} A_k(t) \frac{u^k}{k!} \frac{(t - 1)^{l-1}}{l!} u^l \\ &= 1 + A(t, u) \times \sum_{l \geq 1} \frac{(t - 1)^{l-1}}{l!} u^l \\ &= 1 + A(t, u) \frac{1}{t - 1} (\exp(u(t - 1)) - 1); \end{aligned}$$

hence

$$A(t, u) = \sum_{n \geq 0} A_n(t) \frac{u^n}{n!} = \frac{t - 1}{t - \exp(u(t - 1))}, \tag{3.1}$$

which is the traditional exponential generating function for the Eulerian polynomials, explicitly given by Euler himself in his memoir.

We may also write:

$$\begin{aligned} \sum_{n \geq 0} \frac{A_n(t)}{(1 - t)^{n+1}} \frac{u^n}{n!} &= \frac{e^u}{1 - te^u} = e^u \sum_{j \geq 0} (te^u)^j \\ &= \sum_{j \geq 0} t^j \sum_{n \geq 0} \frac{(u(j + 1))^n}{n!} = \sum_{n \geq 0} \frac{u^n}{n!} \sum_{j \geq 0} t^j (j + 1)^n, \end{aligned}$$

which leads to another equivalent definition of the Eulerian polynomials

$$\frac{A_n(t)}{(1-t)^{n+1}} = \sum_{j \geq 0} t^j (j+1)^n \quad (n \geq 0), \tag{3.2}$$

which is the most common starting definition of those polynomials today.

Now rewrite identity (2.8) as

$$\sum_{i=1}^m i^n t^i = -t^{m+1} \sum_{k=0}^n \binom{n}{k} \frac{A_k(t)}{(1-t)^{k+1}} m^{n-k} + \frac{t A_n(t)}{(1-t)^{n+1}}. \tag{3.3}$$

A simple argument on the order of the formal series in  $t$  shows that, when  $m$  tends to infinity, (3.3) implies (3.2). Conversely, replace each fraction  $A_k(t)/(1-t)^{k+1}$  on the right-hand side of (3.3) by  $\sum_{j \geq 0} t^j (j+1)^k$ , which is the right-hand side of (3.2). An easy calculation shows that the right-hand side of (3.3) becomes

$$-\sum_{j \geq 0} t^{m+1+j} (m+1+j)^n + t \sum_{j \geq 0} t^j (j+1)^n = \sum_{i=1}^m i^n t^i.$$

Consequently, identity (3.2) and its *finite* form (3.3) are also *equivalent*.

The relation

$$A_n(t) = (1 + (n-1)t)A_{n-1}(t) + t(1-t) \cdot DA_{n-1}(t) \quad (n \geq 1), \tag{3.4}$$

where  $D$  stands for the differential operator in  $t$ , can be proved from (3.2) as follows: the right-hand side of (3.3) is equal to:

$$\begin{aligned} & (1-t)^n \sum_{j \geq 0} t^j (j+1)^{n-1} (1 + (n-1)t - nt + (1-t)j) \\ &= (1-t)^n \sum_{j \geq 0} t^j (j+1)^{n-1} (1-t)(j+1) \\ &= (1-t)^{n+1} \sum_{j \geq 0} t^j (j+1)^n = A_n(t). \end{aligned} \quad \square$$

Finally, let

$$A_n(t) := \sum_{k \geq 0} A_{n,k} t^k.$$

By (3.2),  $A_0(t) = (1-t)/(1-t) = 1$  and the constant coefficient of each polynomial  $A_n(t)$  is  $A_{n,0} = 1$ . In (3.3), the coefficient of  $t^k$  ( $k \geq 1$ ) is equal to  $A_{n,k}$  on the left and  $A_{n-1,k} + (n-1)A_{n-1,k-1} + kA_{n-1,k} - (k-1)A_{n-1,k-1}$  on the right. As  $A_{n,k} = 0$  for  $k \geq n+1$ , we obtain the recurrence relation

$$\begin{aligned} A_{n,k} &= (k+1)A_{n-1,k} + (n-k)A_{n-1,k-1} \quad (1 \leq k \leq n-1); \\ A_{n,0} &= 1 \quad (n \geq 0); \quad A_{n,k} = 0 \quad (k \geq n) \end{aligned} \tag{3.5}$$

so that each  $A_n(t)$  is a polynomial with *positive* integral coefficients. The first values of the coefficients  $A_{n,k}$ , called *Eulerian numbers*, are shown in the next table.

| k=  | 0 | 1   | 2    | 3    | 4    | 5   | 6 |
|-----|---|-----|------|------|------|-----|---|
| n=1 | 1 |     |      |      |      |     |   |
| 2   | 1 | 1   |      |      |      |     |   |
| 3   | 1 | 4   | 1    |      |      |     |   |
| 4   | 1 | 11  | 11   | 1    |      |     |   |
| 5   | 1 | 26  | 66   | 26   | 1    |     |   |
| 6   | 1 | 57  | 302  | 302  | 57   | 1   |   |
| 7   | 1 | 120 | 1191 | 2416 | 1191 | 120 | 1 |

In summary, the Eulerian polynomials  $A_n(t) = \sum_{k=0}^n A_{n,k}t^k$  ( $n \geq 0$ ) are defined by the following relations which we restate here collectively for clarity. In doing so, we retain the equation numbers already assigned for these relations:

$$A_0(t) = 1; \quad A_n(t) = \sum_{k=0}^{n-1} \binom{n}{k} A_k(t)(t-1)^{n-1-k} \quad (n \geq 1); \tag{2.7}$$

$$\sum_{i=1}^m i^n t^i = \sum_{l=1}^n (-1)^{n+l} \binom{n}{l} \frac{t^{m+1} A_{n-l}(t)}{(t-1)^{n-l+1}} m^l + (-1)^n \frac{t(t^m-1)}{(t-1)^{n+1}} A_n(t), \tag{2.8}$$

for  $m \geq 1, n \geq 0$ ;

$$\sum_{n \geq 0} A_n(t) \frac{u^n}{n!} = \frac{t-1}{t - \exp(u(t-1))}; \tag{3.1}$$

$$\frac{A_n(t)}{(1-t)^{n+1}} = \sum_{j \geq 0} t^j (j+1)^n \quad (n \geq 0); \tag{3.2}$$

$$A_0(t) = 1, \quad A_n(t) = (1 + (n-1)t)A_{n-1}(t) + t(1-t) \cdot DA_{n-1}(t) \quad (n \geq 1); \tag{3.3}$$

$$A_{n,k} = (k+1)A_{n-1,k} + (n-k)A_{n-1,k-1} \quad (1 \leq k \leq n-1), \tag{3.4}$$

$$A_{n,0} = 1 \quad (n \geq 0); \quad A_{n,k} = 0 \quad (k \geq n);$$

$$A_{n,k} = \sum_{0 \leq i \leq k} (-1)^i (k-i+1)^n \binom{n+1}{i} \quad (0 \leq k \leq n-1); \tag{3.5}$$

$$x^n = \sum_{0 \leq k \leq n-1} \binom{x+k}{n} A_{n,k} \quad (n \geq 0). \tag{3.6}$$

The last two relations are easy to establish. For the first one make use of (3.2), starting with  $A_n(t)(1-t)^{-(n+1)}$ . For the second, called *Worpitzky identity*, simply calculate the coefficient of  $t^k$  in  $(1-t)^{n+1} \sum_{n \geq 0} t^n (j+1)^n$ . As mentioned earlier, the first three relations (2.7), (2.8), and (3.1) are due to Euler. The fourth relation (3.2) must also be attributed to him, as he used it for  $t = -1$  in a second memoir [6] to give a sense to the divergent series  $\sum_j (-1)^j (j+1)^m$ . The second (2.8) is fundamental for the next calculation involving tangent numbers. It seems to have been forgotten in today's studies.

### 4 A Relation with the Tangent Numbers

The *tangent numbers*  $T_{2n-1}$  ( $n \geq 1$ ) are defined as the coefficients of the Taylor expansion of  $\tan u$ :

$$\begin{aligned} \tan u &= \sum_{n \geq 1} \frac{u^{2n-1}}{(2n-1)!} T_{2n-1} \\ &= \frac{u}{1!} + \frac{u^3}{3!} 2 + \frac{u^5}{5!} 16 + \frac{u^7}{7!} 272 + \frac{u^9}{9!} 7936 + \frac{u^{11}}{11!} 353792 + \dots \end{aligned}$$

But, starting with (1.6), the generating function for the Genocchi numbers  $G_{2n}$  ( $n \geq 1$ ) can be evaluated as follows:

$$\begin{aligned} \sum_{n \geq 1} \frac{u^{2n}}{(2n)!} G_{2n} &= \sum_{n \geq 1} \frac{(iu)^{2n}}{(2n)!} (-1)^n G_{2n} \\ &= \frac{2iu}{e^{iu} + 1} - iu = \frac{iu(1 - e^{iu})}{1 + e^{iu}} = u \tan(u/2). \end{aligned}$$

Hence,

$$\sum_{n \geq 1} \frac{u^{2n}}{(2n)!} G_{2n} = u \tan(u/2) = \sum_{n \geq 1} \frac{u^{2n}}{2^{2n-1}(2n-1)!} T_{2n-1},$$

and then

$$n T_{2n-1} = 2^{2n-2} G_{2n} = 2^{2n-1} (2^{2n} - 1) B_{2n} \quad (n \geq 1). \tag{4.1}$$

The first values of the tangent numbers  $T_{2n-1}$  ( $n \geq 1$ ), compared with the Genocchi numbers, are shown in the next table.

|            |   |   |    |     |      |         |            |
|------------|---|---|----|-----|------|---------|------------|
| $n$        | 1 | 2 | 3  | 4   | 5    | 6       | 7          |
| $G_{2n}$   | 1 | 1 | 3  | 17  | 155  | 2073    | 38227      |
| $T_{2n-1}$ | 1 | 2 | 16 | 272 | 7936 | 353.792 | 22.368.256 |

In the exponential generating function (3.1) for the Eulerian polynomials  $A_n(t)$ , replace  $t$  by  $-1$  and  $u$  by  $iu$  with  $i = \sqrt{-1}$ . We get:

$$\sum_{n \geq 0} A_n(-1) \frac{(iu)^n}{n!} = \frac{2}{1 + e^{-2iu}}.$$

Hence

$$\begin{aligned} \sum_{n \geq 1} i^{n-1} A_n(-1) \frac{u^n}{n!} &= \frac{1}{i} \left( \frac{2}{1 + e^{-2iu}} - 1 \right) = \frac{1}{i} \frac{1 - e^{-2iu}}{1 + e^{-2iu}} = \tan u \\ &= \sum_{n \geq 1} T_{2n-1} \frac{u^{2n-1}}{(2n-1)!}. \end{aligned}$$

Accordingly,

$$A_{2n}(-1) = 0, \quad A_{2n-1}(-1) = (-1)^{n-1} T_{2n-1} \quad (n \geq 1). \tag{4.2}$$

With  $t = -1$  identity (2.8) becomes

$$\sum_{k=1}^m k^n (-1)^k = (-1)^{m+1} \sum_{k=0}^n \binom{n}{k} \frac{A_k(-1)}{2^{k+1}} m^{n-k} + \frac{(-1)A_n(-1)}{2^{n+1}}.$$

When  $n = 2p \geq 2$  we get:

$$\sum_{k=1}^m k^{2p} (-1)^k = (-1)^{m+1} \frac{m^{2p}}{2} + \sum_{k=1}^p \binom{2p}{2k-1} (-1)^{m+k} \frac{T_{2k-1}}{2^{2k}} m^{2p-2k+1}. \tag{4.3}$$

Using the relation  $n T_{2n-1} = 2^{2n-2} G_{2n}$  ( $n \geq 1$ ), this can be rewritten in terms of the Genocchi numbers as:

$$\sum_{k=1}^m k^{2p} (-1)^k = (-1)^{m+1} \frac{m^{2p}}{2} + \sum_{k=1}^p \binom{2p}{2k-1} (-1)^{m+k} \frac{G_{2k}}{4k} m^{2p-2k+1}. \tag{4.4}$$

When  $n = 2p + 1 \geq 1$  we get:

$$\begin{aligned} &\sum_{k=1}^m k^{2p+1} (-1)^k \\ &= (-1)^{m+1} \frac{m^{2p+1}}{2} + \sum_{k=1}^{p+1} \binom{2p+1}{2k-1} (-1)^{m+k} \frac{T_{2k-1}}{2^{2k}} m^{2p-2k+2} \\ &\quad + (-1)^{p+1} \frac{T_{2p+1}}{2^{2p+2}}, \end{aligned} \tag{4.5}$$

so that in terms of the Genocchi numbers

$$\begin{aligned} & \sum_{k=1}^m k^{2p+1} (-1)^k \\ &= (-1)^{m+1} \frac{m^{2p+1}}{2} + \sum_{k=1}^{p+1} \binom{2p+1}{2k-1} (-1)^{m+k} \frac{G_{2k}}{4k} m^{2p-2k+2} \\ & \quad + (-1)^{p+1} \frac{G_{2p+2}}{4(p+1)}. \end{aligned} \tag{4.6}$$

Both identities (4.4) and (4.6) were established by Euler. This achieves the proof of Theorem 1.1.

### 5 The Carlitz $q$ -Eulerian Polynomials

Recall the traditional  $q$ -ascending factorial defined for each ring element  $\omega$  and each variable  $q$  by

$$\begin{aligned} (\omega; q)_k &:= \begin{cases} 1, & \text{if } k = 0; \\ (1 - \omega)(1 - \omega q) \cdots (1 - \omega q^{k-1}), & \text{if } k \geq 1; \end{cases} \\ (\omega; q)_\infty &:= \prod_{k \geq 0} (1 - \omega q^k); \end{aligned}$$

the  $q$ -binomial coefficients

$$\begin{bmatrix} n \\ k \end{bmatrix}_q := \frac{(q; q)_n}{(q; q)_k (q; q)_{n-k}} \quad (0 \leq k \leq n)$$

and the  $q$ -analogs of integers and factorials

$$\begin{aligned} [n]_q &:= \frac{(q; q)_n}{(1 - q)^n} = 1 + q + q^2 + \cdots + q^{n-1}; \\ [n]!_q &:= [n]_q [n - 1]_q \cdots [1]_q. \end{aligned}$$

As  $\lim_{q \rightarrow 1} (t; q)_{n+1} = (1 - t)^{n+1}$  and  $\lim_{q \rightarrow 1} [j + 1]_q = j + 1$ , definition (2.10) suggests that a new sequence of polynomials  $A_n(t, q)$ , called the  $q$ -Eulerian polynomials, can be defined by the identity

$$\frac{A_n(t, q)}{(t; q)_{n+1}} = \sum_{j \geq 0} t^j ([j + 1]_q)^n \quad (n \geq 0), \tag{5.1}$$



as was done by Carlitz [2] in his seminal paper, thereby entering the  $q$ -series environment initiated by Heine [13]. Also see Gasper and Rahman [12].

By analogy with the Eulerian polynomials, we can replace the infinite series (5.1) by a *finite* sum and try to express the left-hand side as a linear combination of fractions  $A_k(t, q)/(t; q)_{k+1}$  ( $0 \leq k \leq n$ ). From (5.1), we get:

$$\begin{aligned} \sum_{j=1}^m t^j ([j]_q)^n &= \sum_{k=0}^{m-1} t^{k+1} ([k+1]_q)^n \\ &= t \sum_{j \geq 0} t^{j+1} ([j+1]_q)^n - \sum_{j \geq 0} t^{m+1+j} ([m+1+j]_q)^n \\ &= t \frac{A_n(t, q)}{(t; q)_{n+1}} - t^{m+1} \sum_{j \geq 0} t^j (1 + q + \dots + q^j + q^{j+1} + \dots + q^{j+m})^n \\ &= t \frac{A_n(t, q)}{(t; q)_{n+1}} \\ &\quad - t^{m+1} \sum_{j \geq 0} t^j \sum_{k=0}^n \binom{n}{k} (1 + q + \dots + q^j)^k (q^{j+1} + \dots + q^{j+m})^{n-k} \\ &= t \frac{A_n(t, q)}{(t; q)_{n+1}} - t^{m+1} \sum_{j \geq 0} t^j \sum_{k=0}^n \binom{n}{k} [j+1]_q^k q^{(j+1)(n-k)} [m]_q^{n-k} \\ &= t \frac{A_n(t, q)}{(t; q)_{n+1}} - t^{m+1} \sum_{k=0}^n \binom{n}{k} q^{n-k} [m]_q^{n-k} \sum_{j \geq 0} (tq^{(n-k)})^j [j+1]_q^k. \end{aligned}$$

This establishes the identity

$$\sum_{j=1}^m t^j ([j]_q)^n = t \frac{A_n(t, q)}{(t; q)_{n+1}} - t^{m+1} \sum_{k=0}^n \binom{n}{k} q^{n-k} [m]_q^{n-k} \frac{A_k(tq^{n-k}, q)}{(tq^{n-k}; q)_{k+1}}, \tag{5.2}$$

apparently new, that  $q$ -generalizes (2.11). Naturally, both definitions (5.1) and (5.2) for the polynomials  $A_n(t, q)$  are *equivalent*.

Other equivalent definitions can be derived as follows. Starting with (5.1), we can express  $A_n(t, q)$  as a polynomial in  $t$  as follows:

$$\begin{aligned} \frac{A_n(t, q)}{(t; q)_{n+1}} &= \sum_{j \geq 0} t^j ([j+1]_q)^n = \sum_{j \geq 0} t^j \left( \frac{1 - q^{j+1}}{1 - q} \right)^n \\ &= \frac{1}{(1 - q)^n} \sum_{j \geq 0} t^j (1 - q^{j+1})^n \frac{1}{(1 - q)^n} \sum_{j \geq 0} t^j \sum_{k=0}^n \binom{n}{k} (-1)^k q^{jk+k} \\ &= \frac{1}{(1 - q)^n} \sum_{k=0}^n \binom{n}{k} (-1)^k q^k \sum_{j \geq 0} (tq^k)^j = \frac{1}{(1 - q)^n} \sum_{k=0}^n \binom{n}{k} \frac{(-1)^k q^k}{1 - tq^k}, \end{aligned}$$

so that

$$A_n(t, q) = \frac{1}{(1 - q)^n} \sum_{k=0}^n \binom{n}{k} (-1)^k q^k (t; q)_k (tq^{k+1}; q)_{n-k}. \tag{5.3}$$

By examining (5.3), we can see that  $A_n(t, q)$  is a polynomial in  $t$  of degree at most equal to  $(n - 1)$  since the coefficient of  $t^n$  in  $A_n(t, q)$  must be equal to

$$\begin{aligned} & - \frac{1}{(1 - q)^n} \sum_{k=0}^n \binom{n}{k} (-1)^k (-1)^{n+1} q^{n(n+1)/2} \\ & = - \frac{(-1)^{n+1} q^{n(n+1)/2}}{(1 - q)^n} \sum_{k=0}^n \binom{n}{k} (-1)^k = 0. \end{aligned}$$

To see that  $A_n(t, q)$  is a polynomial both in  $t$  and  $q$ , we can use identity (5.3) and write

$$\begin{aligned} (1 - q)A_n(t, q) &= \sum_{k=0}^n \binom{n}{k} (-1)^k q^k \frac{(t; q)_k}{(1 - q)^{n-1}} (tq^{k+1}; q)_{n-k}; \\ (1 - tq^n)A_{n-1}(t, q) &= \sum_{k=0}^{n-1} \binom{n-1}{k} (-1)^k q^k \frac{(t; q)_k}{(1 - q)^{n-1}} (tq^{k+1}; q)_{n-k}. \end{aligned}$$

Hence

$$\begin{aligned} & (1 - q)A_n(t, q) - (1 - tq^n)A_{n-1}(t, q) \\ &= \sum_{k=1}^{n-1} \binom{n-1}{k-1} (-q)^k \frac{(t; q)_k}{(1 - q)^{n-1}} (tq^{k+1}; q)_{n-k} + (-q)^n \frac{(t; q)_n}{(1 - q)^{n-1}} \\ &= \sum_{j=0}^{n-2} \binom{n-1}{j} (-q)^{j+1} \frac{(1-t)(tq; q)_j}{(1 - q)^{n-1}} (tqq^{j+1}; q)_{n-1-j} \\ &\quad - q(1-t)(-q)^{n-1} \frac{(tq; q)_{n-1}}{(1 - q)^{n-1}} \\ &= -q(1-t) \sum_{j=0}^{n-1} \binom{n-1}{j} (-q)^j \frac{(tq; q)_j}{(1 - q)^{n-1}} (tqq^{j+1}; q)_{n-1-j} \end{aligned}$$

and consequently

$$(1 - q)A_n(t, q) = (1 - tq^n)A_{n-1}(t, q) - q(1 - t)A_{n-1}(tq, q). \tag{5.4}$$

With  $A_n(t, q) := \sum_{j=0}^{n-1} t^j A_{n,j}(q)$  we deduce from (5.4) that the coefficients  $A_{n,j}(q)$  satisfy the recurrence

$$A_{n,j}(q) = [j + 1]_q A_{n-1,j}(q) + q^j [n - j]_q A_{n-1,j-1}(q). \tag{5.5}$$

This shows that each polynomial  $A_n(t, q)$  is a polynomial in  $t, q$  with *positive integral* coefficients.

The first values of the polynomials  $A_n(t, q)$  are reproduced in the following table:

$$\begin{aligned} A_0(t, q) &= A_1(t, q) = 1; \quad A_2(t, q) = 1 + tq; \quad A_3(t, q) = 1 + 2tq(q + 1) + t^2q^3; \\ A_4(t, q) &= 1 + tq(3q^2 + 5q + 3) + t^2q^3(3q^2 + 5q + 3) + t^3q^6; \\ A_5(t, q) &= 1 + tq(4q^3 + 9q^2 + 9q + 4) + t^2q^3(6q^4 + 16q^3 + 22q^2 + 16q + 6) \\ &\quad + t^3q^6(4q^3 + 9q^2 + 9q + 4) + t^4q^{10}. \end{aligned}$$

Finally, from the identity  $1/(t; q)_{n+1} = \sum_{j \geq 0} \begin{bmatrix} n+j \\ n \end{bmatrix}_q t^j$ , we get

$$\sum_{i=0}^{n-1} A_{n,i}(q) t^i \sum_{j \geq 0} \begin{bmatrix} n+j \\ n \end{bmatrix}_q t^j = \sum_{k \geq 0} ([k + 1]_q)^n t^k,$$

and obtain for each  $k$  the formula *à la Worpitzky*

$$\sum_{i=0}^{n-1} A_{n,i}(q) \begin{bmatrix} k+n-i \\ n \end{bmatrix}_q = ([k + 1]_q)^n. \tag{5.6}$$

Surprisingly, it took another twenty years to Carlitz [3] to construct a full combinatorial environment for the polynomials  $A_n(t, q)$  he introduced in 1954.

## 6 A Detour to Combinatorics

In contemporary combinatorics, the following integral-valued statistics for permutations have been widely used. For each permutation  $\sigma = \sigma(1)\sigma(2)\cdots\sigma(n)$  of  $12\cdots n$  define:

$$\begin{aligned} \text{exc } \sigma &:= \#\{i : 1 \leq i \leq n, \sigma(i) > i\}; \\ \text{des } \sigma &:= \#\{i : 1 \leq i \leq n - 1, \sigma(i) > \sigma(i + 1)\}; \\ \text{maj } \sigma &:= \sum_{\sigma(i) > \sigma(i+1)} i; \\ \text{inv } \sigma &:= \#\{(\sigma(i), \sigma(j)) : i < j, \sigma(i) > \sigma(j)\}; \end{aligned}$$

called number of *excedances*, number of *descents*, *major index*, and *inversion number*, respectively. Those four statistics can also be defined for each permutation with repetitions (“multiset”).

For  $r \geq 1$  and each sequence  $\mathbf{m} = (m_1, m_2, \dots, m_r)$  of nonnegative integers, let  $R(\mathbf{m})$  denote the class of all  $\binom{m_1 + \dots + m_r}{m_1, \dots, m_r}$  permutations of the multiset  $1^{m_1} 2^{m_2} \dots r^{m_r}$ . It was already known and proved by MacMahon [15] that “exc” and “des,” on the one hand, “maj” and “inv,” on the other hand, were equidistributed on each class  $R(\mathbf{m})$ , accordingly on each symmetric group  $\mathfrak{S}_m$ .

However, we had to wait for Riordan [19] for showing that if  $A_{n,k}$  is defined to be the number of permutations  $\sigma$  from  $\mathfrak{S}_n$  having  $k$  descents (i.e., such that  $\text{des } \sigma = k$ ), then  $A_{n,k}$  satisfies recurrence (3.4):

$$A_{n,k} = (k + 1)A_{n-1,k} + (n - k)A_{n-1,k-1} \quad (1 \leq k \leq n - 1);$$

$$A_{n,0} = 1 \quad (n \geq 0); \quad A_{n,k} = 0 \quad (k \geq n).$$

This result provides the following combinatorial interpretations for the Eulerian polynomials:

$$A_n(t) = \sum_{\sigma \in \mathfrak{S}_n} t^{\text{exc } \sigma} = \sum_{\sigma \in \mathfrak{S}_n} t^{\text{des } \sigma},$$

the second equality being due in fact to MacMahon! The latter author [15, p. 97, and p. 186] knew how to calculate the generating function for the classes  $R(\mathbf{m})$  by “exc” by using his celebrated Master Theorem, but did not make the connection with the Eulerian polynomials. A thorough combinatorial study of those polynomials was made in the monograph [10] in 1970.

In 1974, Carlitz [3] completes his study of his  $q$ -Eulerian polynomials by showing that

$$A_n(t, q) = \sum_{\sigma \in \mathfrak{S}_n} t^{\text{des } \sigma} q^{\text{maj } \sigma} \quad (n \geq 0).$$

As “inv” has the same distribution over  $\mathfrak{S}_n$  as “maj,” it was very tantalizing to make a full statistical study of the pair (des, inv). Let  $e_q(u) := \sum_{n \geq 0} u^n / (q; q)_n$  be the (first)  $q$ -exponential. First, a straightforward calculation leads to the identity

$$1 + \sum_{n \geq 1} t A_n(t) \frac{u^n}{n!} = \frac{1 - t}{1 - t \exp((1 - t)u)}.$$

In the above fraction, make the substitution  $\exp(u) \leftarrow e_q(u)$  and express the fraction thereby transformed as a  $q$ -series:

$$\sum_{n \geq 0} {}^{\text{inv}}A_n(t, q) \frac{u^n}{(q; q)_n} = \frac{1 - t}{1 - t e_q((1 - t)u)}.$$

The new coefficients  ${}^{\text{inv}}A_n(t, q)$  are to be determined. They were characterized by Stanley [21] who proved the identity:

$${}^{\text{inv}}A_n(t, q) = t \sum_{\sigma \in \mathfrak{S}_n} t^{\text{des } \sigma} q^{\text{inv } \sigma} \quad (n \geq 1).$$

Now rewrite the exponential generating function for the Eulerian polynomials  $A_n(s)$  (see (3.1)) as

$$\sum_{n \geq 0} A_n(s) \frac{u^n}{n!} = \frac{(1 - s) \exp u}{\exp(su) - s \exp u}.$$

In the right-hand side, make the substitutions  $s \leftarrow sq$ ,  $\exp(u) \leftarrow e_q(u)$ . Again, express the fraction thereby transformed as a  $q$ -series:

$$\sum_{n \geq 0} {}^{\text{exc}}A_n(s, q) \frac{u^n}{(q; q)_n} = \frac{(1 - sq)e_q(u)}{e_q(squ) - sq e_q(squ)}.$$

The combinatorial interpretation of the coefficients  ${}^{\text{exc}}A_n(s, q)$  was found by Shareshian and Wachs [20] in the form

$${}^{\text{exc}}A_n(s, q) = \sum_{\sigma \in \mathfrak{S}_n} s^{\text{exc } \sigma} q^{\text{maj } \sigma} \quad (n \geq 0).$$

A further step can be made by calculating the exponential generating function for the polynomials  $A_n(s, t, q) := \sum_{\sigma \in \mathfrak{S}_n} s^{\text{exc } \sigma} t^{\text{des } \sigma} q^{\text{maj } \sigma}$  ( $n \geq 0$ ), as was done in [7]:

$$\sum_{n \geq 0} A_n(s, t, q) \frac{u^n}{(t; q)_{n+1}} = \sum_{r \geq 0} t^r \frac{(1 - sq)(usq; q)_r}{((u; q)_r - sq(usq; q)_r)(1 - uq^r)}.$$

Identities (4.2) that relate the evaluations of Eulerian polynomials at  $t = -1$  to tangent numbers can also be carried over to a  $q$ -environment. This gives rise to a new family of  $q$ -analogs of tangent numbers using the combinatorial model of *doubloons* (see [8, 9].)

Following Reiner [17, 18], Eulerian polynomials attached to other groups than the symmetric group have been defined and calculated, in particular for Weyl groups. What is needed is the concept of *descent*, which naturally occurs as soon as the notions of *length* and *positive roots* can be introduced. The Eulerian polynomials for Coxeter groups of spherical type have been explicitly calculated by Cohen [4], who gave the full answer to a question raised by Hirzebruch [14], who on the other hand, pointed out the relevance of Euler’s memoir [6] to contemporary Algebraic Geometry.

## References

- [1] Bernoulli, Jacques. *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicè scripta de ludo pilae reticularis*. Basileae: impensis Thurnisiorum, fratrum, 1713.
- [2] Carlitz, Leonard.  $q$ -Bernoulli and Eulerian numbers, *Trans. Am. Math. Soc.*, vol. **76**, 1954, p. 332–350.
- [3] Carlitz, Leonard. A combinatorial property of  $q$ -Eulerian numbers, *Am. Math. Mon.*, vol. **82**, 1975, p. 51–54.
- [4] Cohen, Arjeh M. Eulerian polynomials of spherical type, *Münster J. Math.*, vol. **1**, 2008, p. 1–7.
- [5] Euler, Leonhard. *Institutiones calculi differentialis cum eius usu in analysi finitorum ac Doctrina serierum*, Academiae Imperialis Scientiarum Petropolitanae, St. Petersburg, 1755, chap. VII (“Methodus summandi superior ulterius promot”).
- [6] Euler, Leonhard. Remarques sur un beau rapport entre les séries des puissances tant directes que réciproques, *Mémoires de l’Académie des Sciences de Berlin*, vol. **27**, 1768, p. 83–106. Also in *Opera Omnia*, Ser. I, *Commentationes analyticae ad theoriam serierum infinitarum pertinentes*, II, vol. **15**, p. 70–90, Teubner, Leipzig, 1927.
- [7] Foata, Dominique; Han, Guo-Niu. Fix-Mahonian Calculus, III: A quadruple distribution, *Monatsh. Math.*, vol. **154**, 2008, 177–197.
- [8] Foata, Dominique; Han, Guo-Niu. Doubleloons and new  $q$ -tangent numbers, 16 p., *Quarterly J. Math.* (in press)
- [9] Foata, Dominique; Han, Guo-Niu. The doubleloon polynomial triangle, 20 p., *Ramanujan J.*, (The Andrews Festschrift) **23** (2010, in press)
- [10] Foata, Dominique; Schützenberger, Marcel-Paul. *Théorie géométrique des polynômes eulériens*. Lecture Notes in Mathematics, **138**, Berlin, Springer, 1970. (<http://igd.univ-lyon1.fr/~slc/books/index.html>).
- [11] Genocchi, Angelo. Intorno all’espressione generale de’ numeri Bernoulliani. *Ann. Sci. Mat. Fis.*, vol. **3**, 1852, p. 395–405
- [12] Gasper, George; Rahman, Mizan. *Basic hypergeometric series*. Encyclopedia of Math. and its Appl. **35**, Cambridge Univ. Press, Cambridge, 1990.
- [13] Heine, Heinrich Eduard. Über die Reihe..., *J. Reine Angew. Math.*, vol. **34**, 1847, p. 210–212.
- [14] Hirzebruch, Friedrich. Eulerian Polynomials, *Münster J. Math.* vol. **1**, 2008, p. 9–14.
- [15] MacMahon, Percy Alexander., *Combinatory Analysis*, vol. 1 and 2. Cambridge, Cambridge Univ. Press, 1915, (Reprinted by Chelsea, New York, 1955).
- [16] Nielsen, Niels. *Traité élémentaire des nombres de Bernoulli*. Gauthier-Villars, Paris, 1923
- [17] Reiner, Vic. The distribution of descents and length in a Coxeter group, *Electron. J. Combinator.*, vol. **2**, 1995, R25.
- [18] Reiner, Vic. Descents and one-dimensional characters for classical Weyl groups, *Discrete Math.*, vol. **140** 1995, p. 129–140.
- [19] Riordan, John. *An Introduction to Combinatorial Analysis*. New York, Wiley, 1958.
- [20] Shreshian, John; Wachs, Michelle L.  $q$ -Eulerian Polynomials: Excedance Number and Major Index. *Electron. Res. Announc. Am. Math. Soc.*, vol. **13** 2007, p. 33–45.
- [21] Stanley, Richard P. Binomial posets, Möbius inversion, and permutation enumeration, *J. Combin. Theory Ser. A*, vol. **20**, 1976, 336–356.

# Crystal Symmetry Viewed as Zeta Symmetry II

Shigeru Kanemitsu and Haruo Tsukada

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** In this paper, we continue our previous investigations on applications of the Epstein zeta-functions. We shall mostly state the results for the lattice zeta-functions, which can be immediately translated into those for the corresponding Epstein zeta-functions. We shall take up the generalized Chowla–Selberg (integral) formula and state many concrete special cases of this formula.

**Mathematics Subject Classification (2000)** 11F66, 11M26, 11M41

**Key words and phrases** Lattice zeta-function · Generalized Chowla–Selberg formula · Zeta-function of a crystal

## 1 Introduction

In our previous papers [17–19], we have developed the theory of the Epstein zeta-functions with emphasis on its ample applications to crystal symmetries in two aspects.

The first is the Madelung constants associated to ionic crystals as in [18], where we first gave their precise definition (before this, the constants had been treated empirically as given a priori). The main contributions are [18, Theorem 1] (Mellin–Barnes type formula) and [18, Theorem 2] (generalized Chowla–Selberg formula) from which we may deduce most of the preceding results for the Madelung constants. However, in [18] we gave only a small portion of these consequences, and it is the purpose of this paper to assemble more substantial amount of concrete examples scattered in literature (or generalizations thereof).

---

S. Kanemitsu and H. Tsukada

Department of Information and Computer Sciences, School of Humanity-Oriented Science and Engineering, University of Kinki, Iizuka, Fukuoka 820-8555, Japan  
e-mail: [kanemitsu@fuk.kindai.ac.jp](mailto:kanemitsu@fuk.kindai.ac.jp); [tsukada@fuk.kindai.ac.jp](mailto:tsukada@fuk.kindai.ac.jp)

We recall that as is described on [18, pp. 114–115], the (generalized) Chowla–Selberg formula being a consequence of [18, Theorem 1], rests, in the long run, on the Mellin–Barnes integral formula

$$(1 + x)^{-s} = \frac{1}{2\pi i} \int_{(c)} \frac{\Gamma(s - z) \Gamma(z)}{\Gamma(s)} x^{-z} dz \tag{1.1}$$

for  $x > 0, 0 < c < \sigma = \operatorname{Re} s$ .

This seems to have been used effectively by Hardy [13] for the first time (which was elucidated fully in [17]), then by Berndt [1, 2], the latter being one of the main contributions to the theory of the Epstein zeta-functions.

It is to be noted that the Mellin–Barnes integrals have been extensively used in another context related to the mean square of Dirichlet  $L$ -functions notably by Katsurada and Matsumoto (cf. e.g., [20, 22]). Then finally, Terras [25] has taken up the method to treat the case of general lattices. We refer to [23] for a general theory of Mellin–Barnes integrals. Terras also mentions the Madelung constants and in book form they also appear in [12, 12, 16].

The second aspect is the incomplete gamma function expansion for the perturbed Epstein zeta-function, known as the Ewald expansion in other disciplines [10, 25] in the spirit of (Kuz'min-Linnik-) Lavrik [21], which was successively applied to the elucidation of the screened Coulomb potential first studied by Hautot [14] and then extensively by Chaba–Pathria [4, 5] among others. In [18, Sect. 3], we have expounded this second aspect rather fully.

However, we have not stated the results on the lattice zeta-functions themselves, which correspond to those on the Epstein zeta-functions (cf. [18, p. 106, ll.4–5 from below]). In view of this, we shall state in this paper those results on lattice zeta-functions, which have their counterparts for the Epstein zeta-functions. The main result we shall use is the generalized Chowla–Selberg formula (Theorem 2 below), which we state as a counterpart of Theorem 1 below on the  $K$ -Bessel expansion for the decomposition of a lattice into sublattices. We remark that our Theorem 2 itself is a generalizaion of the corresponding previous results (cf. e.g., [25]).

*Notation.*  $\mathbb{Z}, \mathbb{R},$  and  $\mathbb{C}$  signify the rational integers, the real numbers, and the complex numbers, respectively. For a lattice  $L, L \otimes \mathbb{R}$  means an extension of the coefficient ring (from  $\mathbb{Z}$  to  $\mathbb{R}$ ) and  $\operatorname{Hom}(L, \mathbb{R})$  is the space of all homomorphisms from  $L$  to  $\mathbb{R}$ .

## 2 Lattice Zeta-Function

Let  $L$  be a lattice, i.e., a free Abelian group of finite rank ( $n$ , say) with biadditive form  $(\ , \ )_L$ . Let  $L'$  denote the dual lattice of  $L: L' = \operatorname{Hom}(L, \mathbb{Z})$ . Then for lattice elements  $p, q$  with real coefficients,  $p \in L \otimes \mathbb{R}, q \in L' \otimes \mathbb{R}$ , we introduce



the completed lattice zeta-function  $\Lambda(L, p, q, s)$  by the Dirichlet series absolutely convergent for  $\sigma > \frac{n}{2}$ :

$$\Lambda(L, p, q, s) = \frac{\Gamma(s)}{\pi^s} \sum_{\substack{x \in L \\ x+p \neq 0}} \frac{e^{2\pi i q(x)}}{(x+p, x+p)_{L \otimes \mathbb{R}}^s}, \tag{2.1}$$

where we understand the meaning of  $q(x)$  through isomorphisms

$$L' \otimes \mathbb{R} \cong \text{Hom}(L, \mathbb{R}) \cong \text{Hom}_{\mathbb{R}}(L \otimes \mathbb{R}, \mathbb{R}).$$

Let  $\mathbf{e}_1, \dots, \mathbf{e}_n$  be a basis of  $L$ ,  $L = \mathbb{Z}\mathbf{e}_1 \oplus \dots \oplus \mathbb{Z}\mathbf{e}_n$ . Let  $\phi$  denote the extension to  $\mathbb{R}^n$  of the canonical isomorphism

$$\phi_0 : \mathbb{Z}^n \longrightarrow L, \quad x = \phi_0(\mathbf{a}),$$

for  $\mathbf{a} = (a_1, \dots, a_n)$  i.e.

$$\phi : \mathbb{R}^n \longrightarrow L \otimes \mathbb{R}, \quad x = \phi(\mathbf{a}) = a_1\mathbf{e}_1 + \dots + a_n\mathbf{e}_n,$$

for  $\mathbf{a} = (a_1, \dots, a_n) \in \mathbb{R}^n$ .

The associated Gram matrix is then defined by

$$Y = \begin{pmatrix} (\mathbf{e}_1, \mathbf{e}_1)_L & \cdots & (\mathbf{e}_1, \mathbf{e}_n)_L \\ \vdots & \ddots & \vdots \\ (\mathbf{e}_n, \mathbf{e}_1)_L & \cdots & (\mathbf{e}_n, \mathbf{e}_n)_L \end{pmatrix}.$$

Then we have

$$(\phi(\mathbf{a}), \phi(\mathbf{a}))_{L \otimes \mathbb{R}} = Y[\mathbf{a}].$$

Let  $M = (\mathbf{e}_1, \dots, \mathbf{e}_n)$ . Then we have

$$\phi(\mathbf{a}) = M\mathbf{a}, \quad Y = {}^t M M.$$

If we define  $\mathbf{g}, \mathbf{h} \in \mathbb{Z}^n$  by  $p = \phi(\mathbf{g})$  and  $q(x) = q \circ \phi(\mathbf{a}) = \mathbf{h} \cdot \mathbf{a}$  ( $\mathbf{a} \in \mathbb{R}^n$ ), then

$$\begin{aligned} \Lambda(L, p, q, s) &= \frac{\Gamma(s)}{\pi^s} \sum_{\substack{\mathbf{a} \in \mathbb{Z}^n \\ \mathbf{a} + \mathbf{g} \neq \mathbf{0}}} \frac{e^{2\pi i q \circ \phi(\mathbf{a})}}{(\phi(\mathbf{a} + \mathbf{g}), \phi(\mathbf{a} + \mathbf{g}))_{L \otimes \mathbb{R}}^s} \\ &= \frac{\Gamma(s)}{\pi^s} \sum_{\substack{\mathbf{a} \in \mathbb{Z}^n \\ \mathbf{a} + \mathbf{g} \neq \mathbf{0}}} \frac{e^{2\pi i \mathbf{h} \cdot \mathbf{a}}}{Y[\mathbf{a} + \mathbf{g}]^s}, \end{aligned}$$

whence

**Proposition 1.** *Under the above notation, we have*

$$\Lambda(L, p, q, s) = \Lambda(Y, \mathbf{g}, \mathbf{h}, s).$$

Hence, whenever we speak about a lattice zeta-function, we may do well with the corresponding Epstein zeta-function with the Gram matrix.

*Example 1.* (i) The simple cubic (s.c.) structure, NaCl (Sodium Chloride). In this case we have  $\mathbb{Z}^3 = \mathbb{Z} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$  with Gram matrix  $I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ .

The zeta-function is

$$Z(\mathbb{Z}^3, 0, 0, s) = Z(I, \mathbf{o}, \mathbf{o}, s) = \sum_{\substack{\mathbf{a} \in \mathbb{Z}^3 \\ \mathbf{a} \neq \mathbf{o}}} \frac{1}{|\mathbf{a}|^{2s}}. \tag{2.2}$$

(ii) The body centered cubic (b.c.c) structure, CsCl (Caesium Chloride). The face centered cubic (f.c.c.) structure, aka A3. ZnS (Zincblende) structure (diamond). CaF<sub>2</sub> (Fluorite) structure. For details on these and their diagrams, cf. [18].

### 3 Results on Lattice Zeta-Functions

Invoking the functional equation [18, (1.10)]

$$\Lambda(Y, \mathbf{g}, \mathbf{h}, s) = \frac{e^{-2\pi i \mathbf{g} \cdot \mathbf{h}}}{\sqrt{|Y|}} \Lambda\left(Y^{-1}, \mathbf{h}, -\mathbf{g}, \frac{n}{2} - s\right), \tag{3.1}$$

for the RHS of the formula in Proposition 1, we immediately deduce

**Proposition 2.** *(Functional equation for the zeta function of a lattice)*

*For  $p \in L \otimes \mathbb{R}, q \in L' \otimes \mathbb{R}$ , we have*

$$\Lambda(L, p, q, s) = \frac{e^{-2\pi i q(p)}}{\text{Vol}(L \otimes \mathbb{R}/L)} \Lambda\left(L', q, -p, \frac{n}{2} - s\right),$$

(where  $L' = \text{Hom}(L, \mathbb{Z}), L' \otimes \mathbb{R} \cong \text{Hom}(L, \mathbb{R}) \cong \text{Hom}_{\mathbb{R}}(L \otimes \mathbb{R}, \mathbb{R})$ ).

In [18], we deduced a Bessel series expansion ([18, Theorem 1]) for the Epstein zeta-function  $\Lambda(Y, \mathbf{g}, \mathbf{h}, s)$  from the functional equation thereof on appealing to the Mellin–Barnes integral [18, (2.35)]. In the same way, we may deduce a Bessel series expansion for the (perturbed) lattice zeta-function  $\Lambda(Y, p, q, s)$ , which, however, from the point of view of the modular relation principle, is a natural manifestation of the functional equation ([18, Theorem 1]):

**Proposition 3.**

$$\begin{aligned}
 & \frac{\Gamma(s)}{\pi^s} \sum_{x \in L} \frac{e^{2\pi i q(x)}}{((x+p, x+p)_{L \otimes \mathbb{R}} + b)^s} \\
 &= \frac{2}{\text{Vol}(L \otimes \mathbb{R}/L)} \sum_{\substack{y \in L' \\ y+q \neq 0}} e^{-2\pi i (y+q)(p)} \sqrt{\frac{(y+q, y+q)_{L' \otimes \mathbb{R}}}{b}}^{s-\frac{n}{2}} \\
 & \quad \times K_{s-\frac{n}{2}}\left(2\sqrt{(y+q, y+q)_{L' \otimes \mathbb{R}} b} \pi\right) \\
 & \quad + \delta(q) \frac{1}{\text{Vol}(L \otimes \mathbb{R}/L)} \frac{\Gamma(s-\frac{n}{2})}{\pi^{s-\frac{n}{2}}} \frac{1}{b^{s-\frac{n}{2}}}
 \end{aligned} \tag{3.2}$$

for  $\text{Re } s > \frac{n}{2}$ , where

$$\delta(q) = \begin{cases} 1 & q \in L' \\ 0 & q \notin L' \end{cases} \quad (q \in L' \otimes \mathbb{R})$$

and

$$K_s(z) = \frac{1}{2} \int_0^\infty e^{-\frac{1}{2}z(t+\frac{1}{t})} t^{s-1} dt, \quad \text{Re } s > -\frac{1}{2}, |\arg z| < \frac{\pi}{4}$$

signifies the modified Bessel function of the second kind.

Corresponding to a block decomposition ([18, p. 116]) of the matrix  $Y$  associated to the lattice  $L$ , we have a decomposition of  $L$ :

$$L = L_1 \oplus L_2$$

( $L_1, L_2$  are sublattices of  $L$ ), where the decomposition is not necessarily orthogonal. Therefore, we take the orthogonal complement  $(L_1 \otimes \mathbb{R})^\perp$  of  $L_1 \otimes \mathbb{R}$  in  $L \otimes \mathbb{R}$  and introduce the projections

$$\pi_1^\parallel : L_2 \otimes \mathbb{R} \rightarrow L_1 \otimes \mathbb{R}$$

the orthogonal projection to  $L_1 \otimes \mathbb{R}$  and

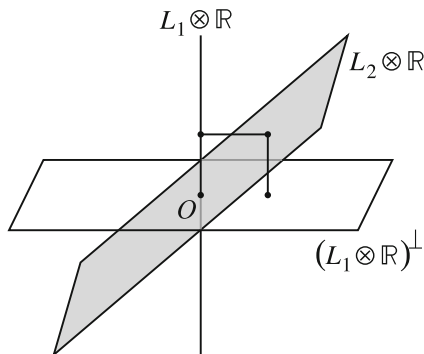
$$\pi_1^\perp : L_2 \otimes \mathbb{R} \rightarrow (L_1 \otimes \mathbb{R})^\perp$$

the orthogonal projection to  $(L_1 \otimes \mathbb{R})^\perp$ .

Under this setting, we have a counterpart of [18, Theorem 2] (which we restate as Theorem 2 below).

**Theorem 1.** (The generalized Chowla–Selberg formula)

$$\begin{aligned}
 & \Lambda(L, (p_1, p_2), (q_1, q_2), s) \\
 &= \delta(p_2) e^{-2\pi i q_2(p_2)} \Lambda(L_1, p_1, q_1, s) \\
 &+ \delta(q_1) \frac{1}{\text{Vol}(L_1 \otimes \mathbb{R}/L_1)} \Lambda(\pi_1^\perp L_2, \pi_1^\perp(p_2), q_2 \circ \pi_1^{\perp-1}, s - \frac{n}{2}) \\
 &+ \frac{2 e^{-2\pi i q_1(p_1)}}{\text{Vol}(L_1 \otimes \mathbb{R}/L_1)} \sum_{\substack{y \in L'_1 \\ y+q_1 \neq 0}} \sum_{\substack{x \in L_2 \\ x+p_2 \neq 0}} e^{2\pi i(-y(p_1)+q_2(x))} e^{-2\pi i(y+q_1) \circ \pi_1^\parallel(x+p_2)} \\
 &\times \sqrt{\frac{(y+q_1, y+q_1)_{L'_1 \otimes \mathbb{R}}}{(\pi_1^\perp(x+p_2), \pi_1^\perp(x+p_2))_{\pi_1^\perp L_2 \otimes \mathbb{R}}}}^{s-\frac{n}{2}} \\
 &\times K_{s-\frac{n}{2}} \left( 2 \sqrt{(y+q_1, y+q_1)_{L'_1 \otimes \mathbb{R}} (\pi_1^\perp(x+p_2), \pi_1^\perp(x+p_2))_{\pi_1^\perp L_2 \otimes \mathbb{R}}} \pi \right) \\
 &= \delta(p_2) e^{-2\pi i q(p)} \frac{1}{\text{Vol}(L_1 \otimes \mathbb{R}/L_1)} \Lambda(L'_1, q_1, -p_1, \frac{n}{2} - s) \\
 &+ \delta(q_1) \frac{1}{\text{Vol}(L_1 \otimes \mathbb{R}/L_1)} \Lambda(\pi_1^\perp L_2, \pi_1^\perp(p_2), q_2 \circ \pi_1^{\perp-1}, s - \frac{n}{2}) \\
 &+ \frac{2 e^{-2\pi i q_2(p_2)}}{\text{Vol}(L_1 \otimes \mathbb{R}/L_1)} \sum_{\substack{y \in L'_1+q_1 \\ y \neq 0}} \sum_{\substack{x \in L_2+p_2 \\ x \neq 0}} e^{2\pi i(-y(p_1)+q_2(x))} e^{-2\pi i y \circ \pi_1^\parallel(x)} \\
 &\times \sqrt{\frac{(y, y)_{L'_1 \otimes \mathbb{R}}}{(\pi_1^\perp(x), \pi_1^\perp(x))_{\pi_1^\perp L_2 \otimes \mathbb{R}}}}^{s-\frac{n}{2}} \\
 &\times K_{s-\frac{n}{2}} \left( 2 \sqrt{(y, y)_{L'_1 \otimes \mathbb{R}} (\pi_1^\perp(x), \pi_1^\perp(x))_{\pi_1^\perp L_2 \otimes \mathbb{R}}} \pi \right). \tag{3.3}
 \end{aligned}$$



**Theorem 2.** (The generalized Chowla–Selberg type formula for matrices cf. [18, Theorem 2], [25, Example 4, p. 208]) Let  $Y = \begin{pmatrix} A & B \\ \iota B & C \end{pmatrix}$  be a block

decomposition with  $A$  an  $n \times n$  matrix and  $B$  an  $n \times m$  matrix and let  $\mathbf{g} = \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{pmatrix}$ ,  $\mathbf{h} = \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix}$ ,  $\mathbf{g}_1, \mathbf{h}_1 \in \mathbb{Z}^n$ ,  $\mathbf{g}_2, \mathbf{h}_2 \in \mathbb{Z}^m$  be the corresponding block decompositions of vectors. Set

$$D = C - {}^t B A^{-1} B.$$

Then under the above notation, we have

$$\begin{aligned} & \Lambda(Y, \mathbf{g}, \mathbf{h}, s) \\ &= \delta(\mathbf{g}_2) e^{-2\pi i \mathbf{g}_2 \cdot \mathbf{h}_2} \Lambda(A, \mathbf{g}_1, \mathbf{h}_1, s) + \delta(\mathbf{h}_1) \frac{1}{\sqrt{|A|}} \Lambda\left(D, \mathbf{g}_2, \mathbf{h}_2, s - \frac{n}{2}\right) \\ &+ \frac{2e^{-2\pi i \mathbf{g}_1 \cdot \mathbf{h}_1}}{\sqrt{|A|}} \sum_{\substack{\mathbf{a} \in \mathbb{Z}^n \\ \mathbf{a} + \mathbf{h}_1 \neq \mathbf{0}}} \sum_{\substack{\mathbf{b} \in \mathbb{Z}^m \\ \mathbf{b} + \mathbf{g}_2 \neq \mathbf{0}}} e^{2\pi i(-\mathbf{g}_1 \cdot \mathbf{a} + \mathbf{h}_2 \cdot \mathbf{b})} e^{-2\pi i A^{-1} B(\mathbf{b} + \mathbf{g}_2) \cdot (\mathbf{a} + \mathbf{h}_1)} \\ &\times \sqrt{\frac{A^{-1}[\mathbf{a} + \mathbf{h}_1]}{D[\mathbf{b} + \mathbf{g}_2]}}^{s - \frac{n}{2}} K_{s - \frac{n}{2}}\left(2\sqrt{A^{-1}[\mathbf{a} + \mathbf{h}_1] D[\mathbf{b} + \mathbf{g}_2]} \pi\right), \end{aligned} \tag{3.4}$$

where

$$\delta(\mathbf{g}) = \begin{cases} 1 & \mathbf{g} \in \mathbb{Z}^n \\ 0 & \mathbf{g} \notin \mathbb{Z}^n \end{cases} \quad (\mathbf{g} \in \mathbb{R}^n).$$

We shall use only the following two special cases of Theorem 2, Corollary 1 being Theorem 2 with  $n = 1, s = 1 (s - \frac{n}{2} = \frac{1}{2})$ .

**Corollary 1.**

$$\begin{aligned} & \Lambda(Y, \mathbf{g}, \mathbf{h}, 1) \\ &= \delta(\mathbf{g}_2) e^{-2\pi i \mathbf{g}_2 \cdot \mathbf{h}_2} \Lambda(A, \mathbf{g}_1, h_1, 1) + \delta(h_1) \frac{1}{\sqrt{A}} \Lambda\left(D, \mathbf{g}_2, \mathbf{h}_2, \frac{1}{2}\right) \\ &+ \frac{e^{-2\pi i \mathbf{g}_1 h_1}}{\sqrt{A}} \sum_{\substack{\mathbf{a} \in \mathbb{Z} \\ \mathbf{a} + h_1 \neq 0}} \sum_{\substack{\mathbf{b} \in \mathbb{Z}^m \\ \mathbf{b} + \mathbf{g}_2 \neq \mathbf{0}}} e^{2\pi i(-\mathbf{g}_1 a + \mathbf{h}_2 \cdot \mathbf{b})} e^{-2\pi i \frac{1}{A} B(\mathbf{b} + \mathbf{g}_2)(a + h_1)} \\ &\times \frac{1}{\sqrt{D[\mathbf{b} + \mathbf{g}_2]}} \exp\left(-\frac{2}{\sqrt{A}} |a + h_1| \sqrt{D[\mathbf{b} + \mathbf{g}_2]} \pi\right). \end{aligned} \tag{3.5}$$

**Corollary 2.** ( $n = 2, m = 1, s = \frac{1}{2}$ )

$$\begin{aligned} & \Lambda\left(Y, \mathbf{g}, \mathbf{h}, \frac{1}{2}\right) \\ &= \delta(g_2) e^{-2\pi i g_2 h_2} \Lambda\left(A, \mathbf{g}_1, \mathbf{h}_1, \frac{1}{2}\right) + \delta(\mathbf{h}_1) \frac{1}{\sqrt{|A|}} \Lambda\left(\frac{|Y|}{|A|}, g_2, h_2, -\frac{1}{2}\right) \\ &+ \frac{e^{-2\pi i \mathbf{g}_1 \cdot \mathbf{h}_1}}{\sqrt{|A|}} \sum_{\substack{\mathbf{a} \in \mathbb{Z}^2 \\ \mathbf{a} + \mathbf{h}_1 \neq \mathbf{0}}} \sum_{\substack{b \in \mathbb{Z} \\ b + g_2 \neq 0}} e^{2\pi i(-\mathbf{g}_1 \cdot \mathbf{a} + h_2 b)} e^{-2\pi i A^{-1} B(b + g_2) \cdot (\mathbf{a} + \mathbf{h}_1)} \\ &\times \frac{1}{\sqrt{A^{-1}[\mathbf{a} + \mathbf{h}_1]}} \exp\left(-\frac{2\sqrt{|Y|}}{\sqrt{|A|}} \sqrt{A^{-1}[\mathbf{a} + \mathbf{h}_1]} |b + g_2| \pi\right), \end{aligned} \tag{3.6}$$

where the first term may also be written as

$$\delta(g_2) e^{-2\pi i g_2 h_2} \frac{e^{-2\pi i \mathbf{g}_1 \cdot \mathbf{h}_1}}{\sqrt{|A|}} \Lambda\left(A^{-1}, \mathbf{h}_1, -\mathbf{g}_1, \frac{1}{2}\right).$$

*Proof.* Theorem 2 with  $n = 2, m = 1, s = \frac{1}{2}, (s - \frac{n}{2} = -\frac{1}{2})$  reads

$$\begin{aligned} & \Lambda\left(Y, \mathbf{g}, \mathbf{h}, \frac{1}{2}\right) \\ &= \delta(g_2) e^{-2\pi i g_2 h_2} \Lambda\left(A, \mathbf{g}_1, \mathbf{h}_1, \frac{1}{2}\right) + \delta(\mathbf{h}_1) \frac{1}{\sqrt{|A|}} \Lambda\left(D, g_2, h_2, -\frac{1}{2}\right) \\ &+ \frac{2e^{-2\pi i \mathbf{g}_1 \cdot \mathbf{h}_1}}{\sqrt{|A|}} \sum_{\substack{\mathbf{a} \in \mathbb{Z}^2 \\ \mathbf{a} + \mathbf{h}_1 \neq \mathbf{0}}} \sum_{\substack{b \in \mathbb{Z} \\ b + g_2 \neq 0}} e^{2\pi i(-\mathbf{g}_1 \cdot \mathbf{a} + h_2 b)} e^{-2\pi i A^{-1} B(b + g_2) \cdot (\mathbf{a} + \mathbf{h}_1)} \\ &\times \sqrt{\frac{D[b + g_2]}{A^{-1}[\mathbf{a} + \mathbf{h}_1]}}^{\frac{1}{2}} K_{-\frac{1}{2}}\left(2\sqrt{A^{-1}[\mathbf{a} + \mathbf{h}_1]} D[b + g_2] \pi\right). \end{aligned}$$

Substituting  $D = \frac{|Y|}{|A|}$  and appealing to the formula

$$K_{\frac{1}{2}}(z) = K_{-\frac{1}{2}}(z) = \sqrt{\frac{\pi}{2z}} e^{-z}, \tag{3.7}$$

this leads to the assertion. The last passage is a consequence of the functional equation (3.1), thereby completing the proof.  $\square$

At this point, we shall assign an exact meaning of the electrostatic energy of a crystal X, thereby giving a precise definition of the Madelung constants. Cf. for more details, [18, p. 103]. Adopting the empirical formula for the crystal structure X,

$$X = C_{n+} A_{n-}$$

where C signifies a cation of electric charge  $+N_+e$  and A an anion of electric charge  $-N_-e$  with  $e$  designating the elementary electric charge, and making a usual convention  $n_+N_+ = n_-N_-$ . We adopt a coordinate system such that the shortest distance  $r$  between the ions is equal to 1. We denote by  $S_{++}$ , respectively by  $S_{+-}$ , the coordinates of cations, respectively those of anions with respect to a coordinate system with a cation at the origin. Dually, let  $S_{-+}$ , respectively,  $S_{--}$  denote the coordinates of cations, respectively, those of anions, with respect to a coordinate system with an anion at the origin.

Then the electrostatic energy of the crystal X may be expressed as

$$U_X = \frac{1}{2} (n_+U_+ + n_-U_-)$$

where

$$U_+ = \frac{1}{4\pi\epsilon_0} \sum_{\substack{x \in S_{++} \\ x \neq 0}} \frac{N_+^2 e^2}{\sqrt{x_1^2 + x_2^2 + x_3^2} r} - \frac{1}{4\pi\epsilon_0} \sum_{x \in S_{+-}} \frac{N_+ N_- e^2}{\sqrt{x_1^2 + x_2^2 + x_3^2} r},$$

and similarly for  $U_-$ . But in ordinary sense, the series that appear here are divergent, and so we adopt zeta-regularization.

For  $S = S_{++}, S_{+-}, S_{-+}, S_{--}$ , we introduce the zeta-function of  $S$  by

$$Z_S(s) = \sum_{\substack{x \in S \\ x \neq 0}} \frac{1}{(x_1^2 + x_2^2 + x_3^2)^s}$$

for  $\sigma$  large enough. Defining the zeta function of the crystal X by

$$Z_X(s) = \frac{n_+}{2} Z_{S_{+-}}(s) - \frac{n_-}{2} Z_{S_{++}}(s) + \frac{n_-}{2} Z_{S_{-+}}(s) - \frac{n_+}{2} Z_{S_{--}}(s),$$

we define the associated Madelung constant by

$$M_X = Z_X\left(\frac{1}{2}\right),$$

whereby we define the electrostatic energy of the crystal X by

$$U_X = -\frac{1}{4\pi\epsilon_0} \frac{N_+ N_- e^2}{r} M_X,$$

where  $\frac{1}{4\pi\epsilon_0} = c^2 \times 10^{-7}$  with  $c$  designating the speed of light.

### 4 Applications

As applications of Theorem 2, we give the following, in addition to [18, Corollary 2].

**Corollary 3.** (An analogue of Hautot’s formula 1) Let  $I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$  and

$$\mathbf{c}_0 = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}. \text{ Then}$$

$$\begin{aligned} Z\left(I, \mathbf{o}, \mathbf{c}_0, \frac{1}{2}\right) &= 3 \Lambda(B_1, \mathbf{o}, \mathbf{c}_0, 1) \\ &= -\frac{\pi}{2} + 6\sqrt{2} \sum_{b_1=1}^{\infty} \sum_{b_2=0}^{\infty} \frac{(-1)^{b_1+b_2}}{\sqrt{b_1^2 + b_2^2}} \operatorname{csch}\left(\sqrt{2b_1^2 + 2b_2^2} \pi\right), \end{aligned}$$

where  $B_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$  is the Gram matrix associated to the lattice

$$L_1 = \mathbb{Z} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

*Proof.* We introduce two lattices accompanying  $L_1$ :

$$L_2 = \mathbb{Z} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

and

$$L_3 = \mathbb{Z} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

with Gram matrices  $B_2 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$  and  $B_3 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ , respectively.

We also recall the body-centered cubic (b. c. c.) lattice

$$L_b = \mathbb{Z} \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$$



with its Gram matrix  $B = \begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix}$  ([18, p. 108]).

We note that

$$L_1 \cup L_2 \cup L_3 = \mathbb{Z}^3, \quad L_1 \cap L_2 = L_2 \cap L_3 = L_1 \cap L_3 = L_b. \quad (4.1)$$

Since

$$\begin{aligned} \Lambda(B_j, \mathbf{o}, \mathbf{c}_0, s) &= \frac{\Gamma(s)}{\pi^s} \sum_{\substack{\mathbf{a} \in \mathbb{Z}^3 \\ \mathbf{a} \neq \mathbf{0}}} \frac{(-1)^{a_1+a_2+a_3}}{B_j[\mathbf{a}]^s} = \frac{\Gamma(s)}{\pi^s} \sum_{\mathbf{a} \in L_b} \frac{1}{I[\mathbf{a}]^s} + \frac{\Gamma(s)}{\pi^s} \sum_{\mathbf{a} \in L_j - L_b} \frac{-1}{I[\mathbf{a}]^s}, \\ &= \Lambda(B, \mathbf{o}, \mathbf{o}, s) - \frac{\Gamma(s)}{\pi^s} \sum_{\mathbf{a} \in L_j - L_b} \frac{1}{I[\mathbf{a}]^s}, \quad (j = 1, 2, 3), \end{aligned}$$

we infer that

$$\begin{aligned} 3 \Lambda(B_1, \mathbf{o}, \mathbf{c}_0, s) &= \Lambda(B_1, \mathbf{o}, \mathbf{c}_0, s) + \Lambda(B_2, \mathbf{o}, \mathbf{c}_0, s) + \Lambda(B_3, \mathbf{o}, \mathbf{c}_0, s) \\ &= 3 \Lambda(B, \mathbf{o}, \mathbf{o}, s) \\ &\quad - \frac{\Gamma(s)}{\pi^s} \sum_{\mathbf{a} \in L_1 - L_b} \frac{1}{I[\mathbf{a}]^s} - \frac{\Gamma(s)}{\pi^s} \sum_{\mathbf{a} \in L_2 - L_b} \frac{1}{I[\mathbf{a}]^s} - \frac{\Gamma(s)}{\pi^s} \sum_{\mathbf{a} \in L_3 - L_b} \frac{1}{I[\mathbf{a}]^s} \\ &= 3 \Lambda(B, \mathbf{o}, \mathbf{o}, s) - \frac{\Gamma(s)}{\pi^s} \sum_{\mathbf{a} \in \mathbb{Z}^3 - L_b} \frac{1}{I[\mathbf{a}]^s} \\ &= 4 \Lambda(B, \mathbf{o}, \mathbf{o}, s) - \Lambda(I, \mathbf{o}, \mathbf{o}, s) \end{aligned} \quad (4.2)$$

we have

$$\Lambda\left(I, \mathbf{o}, \mathbf{c}_0, \frac{1}{2}\right) = 4 \Lambda(B, \mathbf{o}, \mathbf{o}, 1) - \Lambda(I, \mathbf{o}, \mathbf{o}, 1) = 3 \Lambda(B_1, \mathbf{o}, \mathbf{c}_0, 1). \quad (4.3)$$

Now we apply Theorem 2 (or Corollary 2) to  $\Lambda(B_1, \mathbf{o}, \mathbf{c}_0, 1)$  for the decomposition  $B_1 = \left( \begin{array}{c|cc} 1 & 0 & 0 \\ \hline 0 & 2 & 0 \\ 0 & 0 & 2 \end{array} \right)$  to obtain

$$\Lambda(B_1, \mathbf{o}, \mathbf{c}_0, 1) = \Lambda\left(1, 0, \frac{1}{2}, 1\right) + \sum_{\substack{\mathbf{b} \in \mathbb{Z}^2 \\ \mathbf{b} \neq \mathbf{o}}} \sum_{a \in \mathbb{Z}} \frac{(-1)^{b_1+b_2}}{\sqrt{2b_1^2 + 2b_2^2}} \exp\left(-2 \left| a + \frac{1}{2} \right| \sqrt{2b_1^2 + 2b_2^2} \pi\right). \tag{4.4}$$

The sum over  $a$  can be seen to be  $\operatorname{csch}\left(\sqrt{2b_1^2 + 2b_2^2} \pi\right)$  (cf. [18, p. 116]), while the first term on the right is the value at  $s = 1$  of  $2(2^{1-2s} - 1)\zeta(2s)$ , which is  $-\frac{\pi^2}{6}$ . Altogether we may rewrite (4.4) as

$$\Lambda(B_1, \mathbf{o}, \mathbf{c}_0, 1) = -\frac{\pi}{6} + 2\sqrt{2} \sum_{b_1=1}^{\infty} \sum_{b_2=0}^{\infty} \frac{(-1)^{b_1+b_2}}{\sqrt{b_1^2 + b_2^2}} \operatorname{csch}\left(\sqrt{2b_1^2 + 2b_2^2} \pi\right) \tag{4.5}$$

Substituting (4.5) into (4.3) concludes the assertion. □

**Corollary 4.** (An analogue of Hautot’s formula 2) In the previous notation, we have

$$Z(I, \mathbf{o}, \mathbf{c}_0, 1) = -\pi + 6 \sum_{\substack{\mathbf{b} \in \mathbb{Z}^2 \\ \mathbf{b} \neq \mathbf{o}}} \frac{1}{\sqrt{b_1^2 + 2b_2^2}} \operatorname{csch}\left(\sqrt{b_1^2 + 2b_2^2} \pi\right)$$

*Proof.* We shall first prove that

$$\Lambda\left(I, \mathbf{o}, \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix}, s\right) = \Lambda\left(\frac{1}{2}C, \mathbf{o}, \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \end{pmatrix}, s\right), \tag{4.6}$$

where  $\frac{1}{2}C = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$  is the Gram matrix associated to the lattice  $L_4 =$

$\mathbb{Z} \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ 0 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ . The lattice  $L_3 = \mathbb{Z} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \oplus \mathbb{Z} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$

introduced in the proof of Corollary 3 is the kernel of the homomorphism  $f : \mathbb{Z}^3 \rightarrow \{-1, 1\}$ ,  $f(\mathbf{a}) = (-1)^{a_1+a_2}$ . Hence

$$\Lambda\left(I, \mathbf{o}, \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix}, s\right) = \frac{\Gamma(s)}{\pi^s} \sum_{\substack{\mathbf{a} \in \mathbb{Z}^3 \\ \mathbf{a} \neq \mathbf{o}}} \frac{f(\mathbf{a})}{I[\mathbf{a}]^s} = \frac{\Gamma(s)}{\pi^s} \sum_{\substack{\mathbf{a} \in L_3 \\ \mathbf{a} \neq \mathbf{o}}} \frac{1}{I[\mathbf{a}]^s} + \frac{\Gamma(s)}{\pi^s} \sum_{\mathbf{a} \in \mathbb{Z}^3 - L_3} \frac{-1}{I[\mathbf{a}]^s}. \tag{4.7}$$

On the other hand, if we define  $g : L_4 \rightarrow \mathbb{Z}$  by

$$g \left( b_1 \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix} + b_2 \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ 0 \end{pmatrix} + b_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right) = (-1)^{b_1} + (-1)^{b_2},$$

then  $\text{Im}(g) = \{-2, 0, 2\}$  and

$$\begin{aligned} & \{x \in L_4 \mid g(x) = 2\} \\ &= \left\{ b_1 \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix} + b_2 \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ 0 \end{pmatrix} + b_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \mid b_1, b_2 \in 2\mathbb{Z} \right\} = L_3, \\ & \{x \in L_4 \mid g(x) = -2\} \\ &= \left\{ b_1 \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix} + b_2 \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ 0 \end{pmatrix} + b_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \mid b_1, b_2 \in 2\mathbb{Z} + 1 \right\} = \mathbb{Z}^3 - L_3, \end{aligned}$$

whence

$$\begin{aligned} & \Lambda \left( \frac{1}{2}C, \mathbf{o}, \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \end{pmatrix}, s \right) + \Lambda \left( \frac{1}{2}C, \mathbf{o}, \begin{pmatrix} 0 \\ \frac{1}{2} \\ 0 \end{pmatrix}, s \right) \\ &= \frac{\Gamma(s)}{\pi^s} \sum_{\substack{\mathbf{b} \in \mathbb{Z}^3 \\ \mathbf{b} \neq \mathbf{o}}} \frac{(-1)^{b_1} + (-1)^{b_2}}{\left(\frac{1}{2}C[\mathbf{b}]\right)^s} = \frac{\Gamma(s)}{\pi^s} \sum_{x \in L_4} \frac{g(x)}{I[x]^s} \\ &= \frac{\Gamma(s)}{\pi^s} \sum_{\substack{\mathbf{a} \in L_3 \\ \mathbf{a} \neq \mathbf{o}}} \frac{2}{I[\mathbf{a}]^s} + \frac{\Gamma(s)}{\pi^s} \sum_{\mathbf{a} \in \mathbb{Z}^3 - L_3} \frac{-2}{I[\mathbf{a}]^s}. \end{aligned} \tag{4.8}$$

Comparing (4.7) and (4.8) yields

$$\Lambda \left( I, \mathbf{o}, \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix}, s \right) = \frac{1}{2} \Lambda \left( \frac{1}{2}C, \mathbf{o}, \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \end{pmatrix}, s \right) + \frac{1}{2} \Lambda \left( \frac{1}{2}C, \mathbf{o}, \begin{pmatrix} 0 \\ \frac{1}{2} \\ 0 \end{pmatrix}, s \right),$$

but since

$$\Lambda \left( \frac{1}{2}C, \mathbf{o}, \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \end{pmatrix}, s \right) = \Lambda \left( \frac{1}{2}C, \mathbf{o}, \begin{pmatrix} 0 \\ \frac{1}{2} \\ 0 \end{pmatrix}, s \right),$$

we conclude (4.6).

Hence by (4.6), (4.7), and the scaling property

$$\Lambda(cY, \mathbf{g}, \mathbf{h}, s) = c^{-s} \Lambda(Y, \mathbf{g}, \mathbf{h}, s), \quad c > 0, \tag{4.9}$$

we find that

$$\Lambda\left(I, \mathbf{o}, \mathbf{c}_0, \frac{1}{2}\right) = 3\Lambda\left(\frac{1}{2}C, \mathbf{o}, \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \end{pmatrix}, 1\right) = 6\Lambda\left(C, \mathbf{o}, \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \end{pmatrix}, 1\right). \tag{4.10}$$

Applying Theorem 2 (Corollary 1) for the decomposition  $C = \begin{pmatrix} 1|0 & 0 \\ 0|1 & 0 \\ 0|0 & 2 \end{pmatrix}$  as in the proof of (4.4), we conclude that

$$\Lambda\left(C, \mathbf{o}, \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \end{pmatrix}, 1\right) = -\frac{\pi}{6} + \sum_{\substack{\mathbf{b} \in \mathbb{Z}^2 \\ \mathbf{b} \neq \mathbf{o}}} \frac{1}{\sqrt{b_1^2 + 2b_2^2}} \operatorname{csch}\left(\sqrt{b_1^2 + 2b_2^2} \pi\right).$$

Substituting this in (4.10) completes the proof. □

**Corollary 5.** (An analogue of the Benson–Mackenzie formula)

$$\begin{aligned} Z\left(I, \mathbf{o}, \mathbf{o}, \frac{1}{2}\right) &= \frac{\pi}{3} + 4\zeta\left(\frac{1}{2}\right)\beta\left(\frac{1}{2}\right) + 8 \sum_{a_1=1}^{\infty} \sum_{a_2=0}^{\infty} \frac{1}{\sqrt{a_1^2 + a_2^2}} \frac{1}{\exp\left(2\sqrt{a_1^2 + a_2^2} \pi\right) - 1} \\ &= -\pi + 12\pi \sum_{a_1=1}^{\infty} \sum_{a_2=0}^{\infty} \left(\operatorname{csch}\left(\sqrt{a_1^2 + a_2^2} \pi\right)\right)^2 \\ &= -2.83729747948 \dots \end{aligned}$$

*Proof.* As in [18, p. 115] we may write

$$\begin{aligned} Z(I, \mathbf{o}, \mathbf{o}, s) &= \sum_{\substack{\mathbf{a} \in \mathbb{Z}^3 \\ \mathbf{a} \neq \mathbf{o}}} \frac{1}{I[\mathbf{a}]^s} = \sum_{\substack{\mathbf{a} \in \mathbb{Z}^3 \\ \mathbf{a} \neq \mathbf{o}}} \frac{a_1^2 + a_2^2 + a_3^2}{(a_1^2 + a_2^2 + a_3^2)^{s+1}} \\ &= 3 \sum_{\substack{b \in \mathbb{Z} \\ b \neq 0}} \left( b^2 \sum_{\mathbf{a} \in \mathbb{Z}^2} \frac{1}{(I_2[\mathbf{a}] + b^2)^{s+1}} \right) \end{aligned}$$

for  $\sigma > \frac{3}{2}$ . Then Theorem 2 gives

$$\begin{aligned}
 & Z(I, \mathbf{o}, \mathbf{o}, s) \\
 &= 3 \sum_{\substack{b \in \mathbb{Z} \\ b \neq 0}} \left( b^2 \frac{2 \pi^{s+1}}{\Gamma(s+1)} \sum_{\substack{\mathbf{a} \in \mathbb{Z}^2 \\ \mathbf{a} \neq \mathbf{o}}} \sqrt{\frac{I_2[\mathbf{a}]}{b^2}}^s K_s(2 \sqrt{I_2[\mathbf{a}] b^2} \pi) \right. \\
 &\quad \left. + b^2 \frac{\pi^{s+1}}{\Gamma(s+1)} \frac{\Gamma(s)}{\pi^s} \frac{1}{b^{2s}} \right) \\
 &= \frac{6 \pi^{s+1}}{\Gamma(s+1)} \sum_{\mathbf{a} \in \mathbb{Z}^2} \sum_{\substack{b \in \mathbb{Z} \\ b \neq 0}} b^2 \sqrt{\frac{I_2[\mathbf{a}]}{b^2}}^s K_s(2 \sqrt{I_2[\mathbf{a}] b^2} \pi) + \frac{3\pi}{s} \sum_{\substack{b \in \mathbb{Z} \\ b \neq 0}} \frac{1}{b^{2s-2}},
 \end{aligned}$$

the last term being  $\frac{6\pi}{s} \zeta(2s - 2)$ . Hence, in particular,

$$\begin{aligned}
 & Z\left(I, \mathbf{o}, \mathbf{o}, \frac{1}{2}\right) \\
 &= \frac{6 \pi^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} \sum_{\mathbf{a} \in \mathbb{Z}^2} \sum_{\substack{b \in \mathbb{Z} \\ b \neq 0}} b^2 \sqrt{\frac{I_2[\mathbf{a}]}{b^2}}^{\frac{1}{2}} K_{\frac{1}{2}}(2 \sqrt{I_2[\mathbf{a}] b^2} \pi) + 12\pi \zeta(-1) \\
 &= 6\pi \sum_{\mathbf{a} \in \mathbb{Z}^2} \sum_{\substack{b \in \mathbb{Z} \\ b \neq 0}} |b| \exp(-2 \sqrt{I_2[\mathbf{a}] |b|} \pi) - \pi
 \end{aligned}$$

on using  $\zeta(-1) = -\frac{B_2}{2} = -\frac{1}{12}$  and (3.7).

In view of

$$\sum_{n=1}^{\infty} n r^n = \frac{r}{(1-r)^2}$$

for  $|r| < 1$ , the first series sums to

$$12\pi \sum_{\mathbf{a} \in \mathbb{Z}^2} \frac{\exp(-2 \sqrt{I_2[\mathbf{a}]} \pi)}{(1 - \exp(-2 \sqrt{I_2[\mathbf{a}]} \pi))^2},$$

which is immediately seen to be the second term on the right of desired identity.  $\square$

*Example 2.* We have yet another identity in contrast to the one in Corollary 5.

$$\begin{aligned}
 & Z\left(I, \mathbf{o}, \mathbf{o}, \frac{1}{2}\right) \\
 &= \frac{\pi}{3} + 4\zeta\left(\frac{1}{2}\right)\beta\left(\frac{1}{2}\right) + 8 \sum_{a_1=1}^{\infty} \sum_{a_2=0}^{\infty} \frac{1}{\sqrt{a_1^2 + a_2^2}} \frac{1}{\exp\left(2\sqrt{a_1^2 + a_2^2} \pi\right) - 1},
 \end{aligned}$$

where  $\beta(s) = L(s, \chi_4)$ ,  $\chi_4$  meaning the primitive Dirichlet character modulo 4.

*Proof.* By Corollary 2,

$$\begin{aligned}
 & \Lambda\left(Y, \mathbf{o}, \mathbf{o}, \frac{1}{2}\right) \\
 &= \Lambda\left(A, \mathbf{o}, \mathbf{o}, \frac{1}{2}\right) + \frac{1}{\sqrt{|A|}} \Lambda\left(\frac{|Y|}{|A|}, 0, 0, -\frac{1}{2}\right) \\
 &+ \frac{1}{\sqrt{|A|}} \sum_{\substack{\mathbf{a} \in \mathbb{Z}^2 \\ \mathbf{a} \neq \mathbf{o}}} \sum_{\substack{b \in \mathbb{Z} \\ b \neq 0}} \frac{\cos(2A^{-1}B \cdot \mathbf{a} b \pi)}{\sqrt{A^{-1}[\mathbf{a}]}} \exp\left(-\frac{2\sqrt{|Y|}}{\sqrt{|A|}} \sqrt{A^{-1}[\mathbf{a}]} |b| \pi\right),
 \end{aligned}$$

which further becomes by the scaling property and the last line in Corollary 2,

$$\begin{aligned}
 & \frac{1}{\sqrt{|A|}} \Lambda\left(A^{-1}, \mathbf{o}, \mathbf{o}, \frac{1}{2}\right) + \frac{\sqrt{|Y|}}{3|A|} \pi \\
 &+ \frac{1}{\sqrt{|A|}} \sum_{\substack{\mathbf{a} \in \mathbb{Z}^2 \\ \mathbf{a} \neq \mathbf{o}}} \sum_{\substack{b \in \mathbb{Z} \\ b \neq 0}} \frac{\cos(2A^{-1}B \cdot \mathbf{a} b \pi)}{\sqrt{A^{-1}[\mathbf{a}]}} \exp\left(-\frac{2\sqrt{|Y|}}{\sqrt{|A|}} \sqrt{A^{-1}[\mathbf{a}]} |b| \pi\right),
 \end{aligned}$$

where we substituted the value  $\Lambda(1, 0, 0, \frac{1}{2}) = -\frac{1}{2}\Gamma(\frac{1}{2})\zeta(-1) = \frac{\pi}{3}$ .

Specializing  $Y$  to be  $I$ , we get

$$\begin{aligned}
 & \Lambda\left(I, \mathbf{o}, \mathbf{o}, \frac{1}{2}\right) \\
 &= \Lambda\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{o}, \mathbf{o}, \frac{1}{2}\right) + \frac{\pi}{3} + \sum_{\substack{\mathbf{a} \in \mathbb{Z}^2 \\ \mathbf{a} \neq \mathbf{o}}} \sum_{\substack{b \in \mathbb{Z} \\ b \neq 0}} \frac{1}{\sqrt{I_2[\mathbf{a}]}} \exp\left(-2\sqrt{I_2[\mathbf{a}]} |b| \pi\right).
 \end{aligned}$$

The inner series of the last term

$$\sum_{\substack{\mathbf{a} \in \mathbb{Z}^2 \\ \mathbf{a} \neq \mathbf{o}}} \sum_{b=1}^{\infty} \frac{2}{\sqrt{I_2[\mathbf{a}]}} \exp\left(-2\sqrt{I_2[\mathbf{a}]} \pi\right)^b$$

is a geometric series, and so it amounts to the last series in our corollary. Finally it

suffices to note the identity  $\Lambda\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \mathbf{o}, \mathbf{o}, s\right) = 4\zeta(s)\beta(s)$ . □

## References

- [1] B. C. Berndt, Identities involving the coefficients of a class of Dirichlet series IV, *Trans. Am. Math. Soc.* **149** (1970), 179–185.
- [2] B. C. Berndt, Identities involving the coefficients of a class of Dirichlet series VI, *ibid.*, **160** (1971), 157–167.
- [12] J. M. Borwein and P. B. Borwein, *Pi and the AGM: A study in analytic number theory and computational complexity*, Wiley, New York, (1987).
- [4] A. N. Chaba and R. K. Pathria, Evaluation of a class of lattice sums in arbitrary dimensions, *J. Math. Phys.* **16** (1975), 1457–1460.
- [5] A. N. Chaba and R. K. Pathria, Evaluation of a class of lattice sums using Poisson's summation formula. II, *J. Phys. A: Math. Gen.* **9** (1976), 1411–1423.
- [6] S. Chowla and A. Selberg, On Epstein's zeta-function (I), *Proc. Nat. Acad. Sci. USA* **35** (1949), 371–374; *Collected Papers of Atle Selberg I*, Springer Verlag, (1989), 367–370. *The Collected Papers of Sarvadaman Chowla II*, CRM, (1999), 719–722.
- [7] A. Selberg and S. Chowla, On Epstein's zeta-function, *J. Reine Angew. Math.* **227** (1967), 86–110; *Collected Papers of Atle Selberg I*, Springer Verlag, (1989), 521–545; *The Collected Papers of Sarvadaman Chowla II*, CRM, (1999), 1101–1125.
- [8] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups* (2nd. ed.), Springer, New York, (1993).
- [9] R. E. Crandall, New representations for the Madelung constant, *Exp. Math.* **8** (1999), 367–379.
- [10] P. Ewald, Zur Theorie allgemeiner Zetafunktionen II, *Ann. Phys.* **63** (1921), 205–216.
- [22] M. L. Glasser, The evaluation of lattice sums I: Analytic procedures, *J. Math. Phys.* **14** (1973), 409–413; Comments by A. Hautot, *ibid.* **15** (1984), 268.
- [12] M. L. Glasser and I. J. Zucker, Lattice sums., *Theoretical Chemistry: Advances and Perspectives*, Vol. 5, ed. by D. Henderson, Academic, New York (1980), 67–139.
- [13] G. H. Hardy, Some multiple integrals, *Quart. J. Math. (Oxford)*(2) **5** (1908), 357–375; *Collected Papers. Vol. V* (1972), 434–452, Comments 453.
- [14] A. Hautot, A new method for the evaluation of slowly convergent series, *J. Math. Phys.* **15** (1974), 1722–1727.
- [15] A. Hautot, New applications of Poisson's summation formula, *J. Phys. A Math. Gen.* **8** (1975), 853–862.
- [16] S. Kanemitsu and H. Tsukada, *Vistas of special functions*, World Scientific, Singapore (2007), pp. 215.
- [17] S. Kanemitsu, Y. Tanigawa, H. Tsukada and M. Yoshimoto, On Bessel series expressions for some lattice sums II, *J. Phys. A Math. Gen.* **37** (2004), 719–734.
- [18] S. Kanemitsu, Y. Tanigawa and H. Tsukada, Crystal symmetry viewed as zeta symmetry, Proc. Intern. Sympos. Zeta-functions, Topology and Quantum Physics, Kluwer Academic, Dordrecht (2005), 91–129.
- [19] S. Kanemitsu, Y. Tanigawa and W.-P. Zhang, On Bessel series expressions for some lattice sums, *Chebyshevskii Sb.* **5** (2004), 128–137.
- [20] M. Katsurada, An application of Mellin–Barnes type of integrals to the mean square of  $L$ -functions, *Liet. Matem. Rink.* **38** (1998), 98–112.
- [21] A. F. Lavrik, An approximate functional equation for the Dirichlet  $L$ -function, *Trudy Moskov. Math. Obsč* **18** (1968), 91–104 = *Trans. Moskow Math. Soc.* **18** (1968), 101–115.
- [22] K. Matsumoto, Recent developments in the mean square theory of the Riemann zeta and other zeta-functions, in *Number Theory* ed. by R. P. Bambah et al., Hindustan Books Agency, (2000) 241–286.
- [23] R. B. Paris and D. Kaminski, *Asymptotics and Mellin-Barnes Integrals*, Cambridge University Press, Cambridge, (2001).
- [24] A. Terras, Bessel series expansions of the Epstein zeta function and the functional equation, *Trans. Am. Math. Soc.*, **183**, (1973) 477–486.

- [25] A. Terras, *Harmonic Analysis on Symmetric Spaces and Applications I*, Springer, New York, (1985).
- [22] G. N. Watson, *A treatise on the theory of Bessel function*, second edition, CUP, Cambridge, (1966).
- [27] I. J. Zucker, Exact results for some lattice sums in 2, 4, 6 and 8 dimensions, *J. Phys. A Math. Nucl. Gen.* **7** (1974), 1568–1575.
- [28] I. J. Zucker, Functional equation for poly-dimensional zeta functions and the evaluation of Madelung constants, *J. Phys. A Math. Gen.* **9** (1976), 499–505.



# Positive Homogeneous Minima for a System of Linear Forms

Srinivasacharya Raghavan

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** An upper bound for the “positive homogeneous minimum” for  $n$  linearly independent linear forms in  $n$  real variables in terms of their determinant is obtained.

**Mathematics Subject Classification (2000)** 11H46, 11H50, 11J20

**Key words and phrases** Homogeneous minima · Systems of linear forms · Inhomogeneous minima · Wood’s conjecture · Minkowski’s conjecture

For real irrational  $\omega$  with bounded even partial quotients in its simple continued fraction expansion and unbounded odd partial quotients, the inequality  $0 < xy - \omega y^2 < \varepsilon$  is solvable in integers  $x, y$  for any given  $\varepsilon > 0$ , while the same is not true of the inequality  $-\varepsilon < xy - \omega y^2 < 0$ . Consider two real linear forms  $L_1(x, y) := x - \omega y$  and  $L_2(x, y) := y$ , in the standard notation [6], their *homogeneous minimum* denoted by  $M_H$  is just the infimum of the absolute value  $abs L_1(x, y) L_2(x, y)$  of the product  $L_1(x, y) L_2(x, y)$  taken over pairs of integers  $(x, y)$  different from  $(0, 0)$ . The *positive minimum*  $M_P((0, 0))$  is defined as the infimum of  $L_1(x, y) L_2(x, y)$  taken over pairs of integers  $(x, y)$  for which  $L_1(x, y)$  and  $L_2(x, y)$  are both  $> 0$ ; likewise, the *negative minimum*  $M_N((0, 0))$  is the infimum of  $L_1(x, y) L_2(x, y)$  over all pairs of integers  $(x, y)$  such that  $L_1(x, y) L_2(x, y) < 0$ . For the two homogeneous linear forms  $L_1$  and  $L_2$ , we note that “ $M_H = 0$ ” and  $M_P((0, 0)) = 0$  but  $M_N((0, 0))$  is not 0! Hence positive or negative minima seem to be of interest on their own!

More generally, for  $n := r + 2s$  in  $\mathbf{N}$ , with non-negative integral  $r$  and  $s$ , let  $L_1, L_2, \dots, L_r, L_{r+1}, L_{r+2}, \dots, L_{r+s}, L_{r+s+1}, L_{r+s+2}, \dots, L_{r+2s}$  be  $n$  linearly independent linear forms in  $n$  real variables  $x_1, x_2, \dots, x_n$ , having  $D > 0$  as the absolute value of their determinant; let us assume further that  $L_1, L_2, \dots, L_r$  have real

---

S. Raghavan  
26, C.I.T. Colony II Main Road, Mylapore, Chennai 600004, India  
e-mail: [navgraha@bsnl.in](mailto:navgraha@bsnl.in)

coefficients and  $L_{r+1}, L_{r+3}, \dots, L_{r+2s-1}$  have complex coefficients while the  $s$  forms  $L_{r+2}, L_{r+4}, \dots, L_{r+2s}$  are just the complex conjugates respectively of  $L_{r+1}, L_{r+3}, \dots, L_{r+2s-1}$ . One defines their *homogeneous minimum*  $M_H$  as the infimum of the absolute value of the product  $L_1(\mathbf{x})L_2(\mathbf{x})\dots L_n(\mathbf{x})$  taken over all integral  $n$ -tuples  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  different from the zero  $n$ -tuple  $\mathbf{0} := (0, 0, \dots, 0)$ . For a given  $n$ -tuple  $\mathbf{a} := (a_1, a_2, \dots, a_n)$  with real  $a_1, a_2, \dots, a_r$  and complex  $a_{r+1}, a_{r+3}, \dots, a_{r+2s+1}$  having precisely  $a_{r+2}, a_{r+4}, \dots, a_{r+2s}$  as their respective complex conjugates, the *inhomogeneous minimum*  $M_I(\mathbf{a})$  is defined as the infimum of the absolute value of the product  $(L_1(\mathbf{x})+a_1)(L_2(\mathbf{x})+a_2)\dots(L_n(\mathbf{x})+a_n)$  taken over all integral  $n$ -tuples  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  not equal to  $\mathbf{0}$  and finally the *inhomogeneous minimum*  $M_I$  as the supremum of  $M_I(\mathbf{a})$  taken over all  $n$ -tuples  $\mathbf{a}$  as described above. Analogously, the *inhomogeneous positive minima*  $M_P(\mathbf{a})$  and  $M_P$  are defined (cf. [2, 6]) imposing the *additional* conditions  $L_1(\mathbf{x}) + a_1 > 0, L_2(\mathbf{x}) + a_2 > 0, \dots, L_r(\mathbf{x}) + a_r > 0$ . It is well-known that  $D/M_H$  is bounded below by a constant ( $>1$ ) depending only on  $n$ . For  $s = 0$ , Minkowski's celebrated conjecture states that  $M_I$  is bounded above by  $D/2^n$  towards which Chebotarev's (in 1934) was the first contribution with  $2^{n/2}$  in lieu of  $2^n$ ; as of now, the validity of the conjecture stands upheld for all  $n < 7$ , thanks to a galaxy of eminent mathematicians including Landau, Mordell, Davenport, Birch, Swinnerton-Dyer, Dyson, Skubenko, Bambah and Woods, and finally McMullen (whose result in 2005 leading to the confirmation of Minkowski's conjecture for  $n = 6$  was kindly highlighted by the referee in the interest of the present author). Recently, the work on Woods' Conjecture for  $n = 7$  by R.J. Hans-Gill, M. Raka, and R. Sehmi taken with McMullen's 2005 result has led to the confirmation of Minkowski's conjecture for  $n = 7$ .

The focus of this note is, however, on lower bounds (if any) for  $D/M_P(\mathbf{0})$ . The case of  $M_P$  or even of  $M_P(\mathbf{a})$  for any individual  $n$ -tuple  $\mathbf{a}$  in the form prescribed above (say e.g.,  $\mathbf{a}_o := (1/2, 1/2, \dots, 1/2)$ ) is ruled out in view of the following interesting example that we were privileged to learn in person from Professor J. W. S. Cassels (cf. [4, 5]) (given in the notation above): If  $L_i(\mathbf{x}) = x_i (1 \leq i \leq r)$ ,  $L_{r+2j-1}(\mathbf{x}) = x_{r+2j-1} + (\sqrt{-1}) \varepsilon x_{r+2j} (1 \leq j \leq s)$  with  $\varepsilon > 0$  and  $n := r + 2s$  so that  $D = (2\varepsilon)^s$ , one finds that  $M_I(\mathbf{a}_o)/D > 1/(2^{n+s}\varepsilon^s)$  and is thus unbounded as  $\varepsilon$  tends to 0 (the same being true of  $M_P(\mathbf{a}_o)/D$  clearly and noting too that  $M_H = 0$ ). Following Davenport's result " $M_I < c_0(n)D$ " (for  $s = 0$ , with  $c_0(n) := (n2^{n-1}(n!)^{(n-1)/2}\Gamma(1+n/2)/(\Gamma(1/2))^n)^n$ , we have Barnes' estimate " $M_I M_H^{s/(n-s)} < c_1(n)D^{n/(n-s)}$ " for a constant  $c_1(n)$  depending only on  $n$ , Davenport's refinements and further work of Rieger [6] (cf. a complete bibliography in the excellent article [1]) invariably requiring that  $M_H$  be non-zero; one asks if there exists any upper bound for other relevant minima purely in terms of  $D$  without such a condition on  $M_H$ . Pursuing ideas of Siegel and Davenport (also of Barnes), Rieger [6] *inter alia* recovered Chalk's upper bound [2] for  $M_P/D$  (viz. when  $s = 0$ ). Baffled somewhat by exclusion of the case " $s > 0$ " and driven by an urge to stay away from a condition such as " $M_H$  is not 0", it may be of some

interest to seek, in the case “ $s > 0$ ”, an upper bound (in view of the specific example above, neither for  $M_P$  nor for  $M_P(\mathbf{a})$  with general  $\mathbf{a}$  but only) for  $M_P(\mathbf{0})/D$  as in the following main result of this note.

**Theorem.** Given  $n$  linearly independent linear forms  $L_1, L_2, \dots, L_r, L_{r+1}, L_{r+2}, \dots, L_{r+s}, L_{r+s+1}, L_{r+s+2}, \dots, L_{r+2s}$  as above in  $x_1, x_2, \dots, x_n$ , with  $D > 0$  as the absolute value of the determinant, we have for the positive inhomogeneous minimum  $M_P(\mathbf{0})$  the upper bound  $c D$  where  $c := (2^n \Gamma(1 + n/2) / (\Gamma(1/2))^n)^{n+1} (n(n!)^{(n-1)/2})^n$ .

*Proof.* If  $M_P(\mathbf{0}) = 0$ , there is nothing to prove. Following a method devised by Siegel (for  $M_I$  in the case  $s = 0$ ), “modified” then by Davenport [3] (and adapted in [6] by Rieger) to obtain other upper bounds for  $M_I$ , let us take the positive definite quadratic form  $Q(\mathbf{x}) := L_1(\mathbf{x})^2 + L_2(\mathbf{x})^2 + \dots + L_r(\mathbf{x})^2 + (abs L_{r+1}(\mathbf{x}))^2 + \dots + (abs L_n(\mathbf{x}))^2 = L_1(\mathbf{x})^2 + L_2(\mathbf{x})^2 + \dots + L_r(\mathbf{x})^2 + 2(abs L_{r+1}(\mathbf{x}))^2 + \dots + 2(abs L_{r+2s-1}(\mathbf{x}))^2 = K_1(\mathbf{x})^2 + K_2(\mathbf{x})^2 + \dots + K_r(\mathbf{x})^2 + K_{r+1}(\mathbf{x})^2 + K_{r+2}(\mathbf{x})^2 + \dots + K_{r+2s-1}(\mathbf{x})^2 + K_n(\mathbf{x})^2$  where  $abs L_k(\mathbf{x})$  for real  $n$ -tuples  $\mathbf{x}$  denotes (its) absolute value and where we have rewritten  $L_i(\mathbf{x})$  as  $K_i(\mathbf{x})$  for  $i = 1, 2, \dots, r$  and defined  $K_{r+j}(\mathbf{x}) := \sqrt{2} \Re L_{r+j}(\mathbf{x})$  and  $K_{r+j+1}(\mathbf{x}) := \sqrt{2} \Im L_{r+j}(\mathbf{x})$  for  $j = 1, 3, \dots, 2s-1$ . Linking the Minkowski successive minima  $t_1 \leq t_2 \leq \dots \leq t_n$  with the determinant  $\Delta := D^2$  of  $Q$ , we have the well-known inequalities

$$D \leq \sqrt{t_1} \sqrt{t_2} \dots \sqrt{t_n} \leq c_n D \quad \text{where} \quad c_n := 2^n \Gamma(1 + n/2) / (\Gamma(1/2))^n. \tag{1}$$

For non-zero real  $\mathbf{x}$  with  $Q(\mathbf{x}) < t_n$ , there exist real numbers  $a_1, a_2, \dots, a_n$  not all 0 such that  $a_1 K_1(\mathbf{x}) + a_2 K_2(\mathbf{x}) + \dots + a_n K_n(\mathbf{x}) = 0$ . With the notation  $abs$  as above, we can assume (after suitably permuting  $1, 2, \dots, n$  if necessary) that  $0 < abs a_n \geq abs a_i$  for  $1 \leq i \leq n$ . We have then  $K_n(\mathbf{x})^2 = ((a_1 K_1(\mathbf{x}) + \dots + a_{n-1} K_{n-1}(\mathbf{x})) / a_n)^2 \leq (n-1) (K_1(\mathbf{x})^2 + \dots + K_{n-1}(\mathbf{x})^2)$  in view of the Cauchy-Schwarz inequality; therefore,  $(K_1(\mathbf{x})^2 + \dots + K_{n-1}(\mathbf{x})^2) \geq Q(\mathbf{x})/n$ . If, in addition,  $\mathbf{x}$  satisfies the condition  $Q(\mathbf{x}) < t_{n-1}$ , one has a similar linear relation  $b_1 K_1(\mathbf{x}) + b_2 K_2(\mathbf{x}) + \dots + b_{n-1} K_{n-1}(\mathbf{x}) = 0$  with  $0 < abs b_{n-1} \geq abs b_i$  for  $i = 1, 2, \dots, n-1$  (after a suitable permutation of the indices) leading to the inequality  $K_1(\mathbf{x})^2 + \dots + K_{n-2}(\mathbf{x})^2 \geq Q(\mathbf{x}) / (n(n-1))$ . Proceeding in this manner, given any non-zero integral  $\mathbf{x}$ , there exists  $h$  such that  $K_1(\mathbf{x})^2 + \dots + K_h(\mathbf{x})^2 \geq t_h / (n(n-1) \dots (h+1))$ . In particular, for any non-zero integral  $\mathbf{x}$ , abbreviating  $K_i(\mathbf{x})^2$  as  $K_i^2$  for all  $i$ , we have

$$K_1^2 / t_1 + \dots + K_n^2 / t_n \geq 1/n!. \tag{2}$$

In other words, for a suitable permutation  $\sigma(1), \sigma(2), \dots, \sigma(n)$  of  $1, 2, \dots, n$ , such that for all integral  $\mathbf{x} \neq \mathbf{0}$ , we have

$$R(\mathbf{x}) := K_1(\mathbf{x})^2 / t_{\sigma(1)} + \dots + K_n(\mathbf{x})^2 / t_{\sigma(n)} \geq 1/n!. \tag{3}$$

If the successive minima of  $R(\mathbf{x})$  are  $s_1 \leq s_2 \leq \dots \leq s_n$ , then by (2),  $s_1 \geq 1/n!$  and so, in view of (1), we have

$$\sqrt{s_1} + \sqrt{s_2} + \dots + \sqrt{s_n} \leq n \sqrt{s_n} \leq n (n!)^{(n-1)/2} \sqrt{s_1} \sqrt{s_2} \dots \sqrt{s_n} \leq n (n!)^{(n-1)/2} c_n. \tag{4}$$

Let  $\mathbf{x}^{(j)}$  for  $j = 1, 2, \dots, n$  be an integral vector (chosen once for all) at which  $R$  assumes the value  $s_j$ ; writing  $L_*$  briefly for  $L_*(\mathbf{x}^{(j)})$  and for  $1 \leq l \leq s$ , defining  $u_{\sigma(r+l)} := \min(t_{\sigma(r+2l-1)}, t_{\sigma(r+2l)})$ , we clearly have  $(1/n!) \leq s_j = R(\mathbf{x}^{(j)}) = L_1^2/t_{\sigma(1)} + \dots + L_r^2/t_{\sigma(r)} + 2 \text{ abs } L_{r+1}^2/u_{\sigma(r+1)} + 2 \text{ abs } L_{r+3}^2/u_{\sigma(r+2)} + \dots + 2 \text{ abs } L_{r+2s-1}^2/u_{\sigma(r+s)}$ . For  $1 \leq j \leq n$ ,  $1 \leq i \leq r$  and for  $1 \leq k, l \leq s$ , we thus obtain the estimates

$$\begin{aligned} (L_i(\mathbf{x}^{(j)}))^2 &\leq s_j t_{\sigma(i)}, \quad (\text{abs } L_{r+2k-1}(\mathbf{x}^{(j)}))^2 \leq s_j u_{\sigma(r+k)}, \\ (L_{r+2l}(\mathbf{x}^{(j)}))^4 &\leq s_j^2 t_{\sigma(r+2l-1)} t_{\sigma(r+2l)}. \end{aligned} \tag{5}$$

We note that, for any fixed given linear form  $L_i$ , at least one of the  $n$  numbers  $L_i(\mathbf{x}^{(1)}), \dots, L_i(\mathbf{x}^{(n)})$  must be non-zero. Consequently, for  $1 \leq l \leq n$ ,  $\lambda_i := \text{abs } L_i(\mathbf{x}^{(1)}) + \text{abs } L_i(\mathbf{x}^{(2)}) + \dots + \text{abs } L_i(\mathbf{x}^{(n)})$  is always positive. Indeed, since the  $(n, n)$  matrix  $(L_i(\mathbf{x}^{(j)}))$  has non-zero determinant, there exists a permutation  $\rho$  of  $\{1, 2, \dots, n\}$  such that the product  $L_1(\mathbf{x}^{(\rho(1))})L_2(\mathbf{x}^{(\rho(2))}) \dots L_n(\mathbf{x}^{(\rho(n))})$  is non-zero. Therefore, for  $1 \leq i \leq n$ ,  $\lambda_i := +\text{abs } L_i(\mathbf{x}^{(1)}) + \text{abs } L_i(\mathbf{x}^{(2)}) + \dots + \text{abs } L_i(\mathbf{x}^{(n)})$  is always positive. Now there do exist real numbers  $\eta_1, \eta_2, \dots, \eta_n$  satisfying the system of  $n$  linear equations  $L_i(\mathbf{x}^{(1)})\eta_1 + L_i(\mathbf{x}^{(2)})\eta_2 + \dots + L_i(\mathbf{x}^{(n)})\eta_n = \lambda_i/2$  for  $i = 1, 2, \dots, n$ . We take  $y_1, y_2, \dots, y_n$  in  $\mathbf{Z}$  such that  $-1/2 \leq \eta_j - y_j < 1/2$  and define  $\mathbf{x} = y_1 \mathbf{x}^{(1)} + y_2 \mathbf{x}^{(2)} + \dots + y_n \mathbf{x}^{(n)}$ . Note first that for  $i = 1, 2, \dots, n$ ,  $n \text{ abs } L_i(\mathbf{x}^{(1)})\eta_1 + L_i(\mathbf{x}^{(2)})\eta_2 + \dots + L_i(\mathbf{x}^{(n)})\eta_n$  necessarily equals 0; further  $L_i(\mathbf{x}) = \lambda_i/2 - (L_i(\mathbf{x}^{(1)})\eta_1 - y_1)L_i(\mathbf{x}^{(2)})\eta_2 - y_2 + \dots + L_i(\mathbf{x}^{(n)})\eta_n - y_n$ ) which (on duly omitting indices  $k$  with  $L_i(\mathbf{x}^k) = 0!$ ) can be rewritten as  $\text{abs } L_i(\mathbf{x}^{(1)})(1/2 - \text{sgn } L_i(\mathbf{x}^{(1)})(\eta_1 - y_1)) + \dots + \text{abs } L_i(\mathbf{x}^{(n)})(1/2 - ((L_i(\mathbf{x}^{(n)})/\text{abs } L_i(\mathbf{x}^{(n)}))\eta_n - y_n))$ . Since in addition we have also  $(1/2 - (\eta_k - y_k)) > 0, (1/2 + (\eta_k - y_k)) \geq 0$  under the given circumstances, we see that for  $i = 1, 2, \dots, r, L_i(\mathbf{x}) > 0$ . Moreover,  $\text{abs } L_i(\mathbf{x}) \leq \lambda_i \leq \sqrt{t_{\sigma(i)}} (\sqrt{s_1} + \sqrt{s_2} + \dots + \sqrt{s_n})$ . In view of (1), (4), and (5), we have

$$\begin{aligned} \text{abs}(L_1(\mathbf{x})L_2(\mathbf{x}) \dots L_n(\mathbf{x})) &\leq (t_{\sigma(1)}t_{\sigma(2)} \dots t_{\sigma(n)})^{1/2} (\sqrt{s_1} + \sqrt{s_2} + \dots + \sqrt{s_n})^n \\ &\leq c_n^{n+1} (n(n!)^{(n-1)/2})^n D =: cD \end{aligned}$$

and the theorem is finally proved. □

*Remark.* One may note that the bound (involving just  $D$  and no higher power thereof) given by Theorem above is in accord with the (*best possible*) bound (for the case  $s = 0$ ) due to Chalk [2]. Let us keep in view (for non-zero  $s$ ), Cassels' example above. Taking then both  $s$  and  $M_H$  to be non-zero (or simply  $M_H$  to be 1), it is natural for one, on the other hand, to regard the Theorem (albeit with explicit

constants) as a simple special case of Rieger's results (5), (6) on page 127 of [6] but that way one ends up only with a bound involving  $D$  raised to a power higher than 1, in general. We may also point out that with such a result not available earlier, we were forced to require the algebraic number field  $K$  (featuring in Propositions 2–4 and the (main) Theorem of the paper: Values of Quadratic Forms", *Comm. Pure Appl. Math.* **30**(1977), 273–281) to be totally real.

## References

- [1] R. P. Bambah, V. C. Dumir and R. J. Hans-Gill: Non-homogeneous Problems: Conjectures of Minkowski and Watson in *Number Theory*, Indian National Science Academy, 2000, 15–41.
- [2] J. H. H. Chalk: On the positive values of linear forms, *Quarterly J. Math.* **18**(1947), 215–227.
- [3] H. Davenport: Note on a result of Siegel, *Acta Arithmetica* **2**(1937), 262–265.
- [4] H. Davenport: On the products of  $n$  linear forms, *Proc. Camb. Phil. Soc.* **49**(1953), 190–193.
- [5] P. Gruber and C. G. Lekkerkerker: *Geometry of Numbers*, Second Edition, North-Holland Mathematical Library, **37**(1987).
- [6] G. J. Rieger: Einige Bemerkungen ueber inhomogene Linearformen, *J. Reine Angew. Math.* Bd. **203**(1960), 126–129.

# The Divisor Matrix, Dirichlet Series, and $SL(2, \mathbf{Z})$

Peter Sin and John G. Thompson

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** A representation of  $SL(2, \mathbf{Z})$  by integer matrices acting on the space of analytic ordinary Dirichlet series is constructed, in which the standard unipotent element acts as multiplication by the Riemann zeta function. It is then shown that the Dirichlet series in the orbit of the zeta function are related to it by algebraic equations.

**Mathematics Subject Classification (2000)** 11M06, 20C12

**Key words and phrases** Zeta function · Dirichlet series · Group representation · Divisor matrix

## 1 Introduction

This paper is concerned with group actions on the space of analytic Dirichlet series. A *formal* Dirichlet series is a series of the form  $\sum_{n=1}^{\infty} a_n n^{-s}$ , where  $\{a_n\}_{n=1}^{\infty}$  is a sequence of complex numbers and  $s$  is formal variable. Such series form an algebra  $\mathcal{D}[[s]]$  under the operations of termwise addition and scalar multiplication and multiplication defined by Dirichlet convolution:

$$\left( \sum_{n=1}^{\infty} a_n n^{-s} \right) \left( \sum_{n=1}^{\infty} b_n n^{-s} \right) = \sum_{n=1}^{\infty} \left( \sum_{ij=n} a_i b_j \right) n^{-s}. \quad (1)$$

---

P. Sin and J.G. Thompson  
Department of Mathematics, University of Florida, PO Box 118105, Gainesville,  
FL 32611-8105, USA  
e-mail: [sin@ufl.edu](mailto:sin@ufl.edu); [johngriggst@aol.com](mailto:johngriggst@aol.com)

It is well-known that this algebra, sometimes called the *algebra of arithmetic functions*, is isomorphic with the algebra of formal power series in a countably infinite set of variables. The *analytic* Dirichlet series, those which converge for some complex value of the variable  $s$ , form a subalgebra  $\mathcal{D}\{s\}$ , shown in [2] to be a local, non-noetherian unique factorization domain. As a vector space, we can identify  $\mathcal{D}[[s]]$  with the space of sequences and  $\mathcal{D}\{s\}$  with the subspace of sequences satisfying a certain polynomial growth condition. It is important for our purposes that the space of sequences is the dual  $E^*$  of a space  $E$  of countable dimension, since the linear operators on  $\mathcal{D}\{s\}$  of interest to us are induced from operators on  $E$ . We take  $E$  to be the space of columns, indexed by positive integers, which have only finitely many nonzero entries. In the standard basis of  $E$ , endomorphisms acting on the left are represented by column-finite matrices with rows and columns indexed by the positive integers. They act on  $E^*$  by right multiplication. The endomorphisms of  $E$  which preserve  $\mathcal{D}\{s\}$  in their right action form a subalgebra  $\mathcal{DR}$ . There is a natural embedding of  $\mathcal{D}\{s\}$  into  $\mathcal{DR}$  mapping a series to the *multiplication operator* defined by convolution with the series. The Riemann zeta function  $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ ,  $\text{Re}(s) > 1$ , is mapped to the *divisor matrix*  $D = (d_{i,j})_{i,j \in \mathbf{N}}$ , defined by

$$d_{i,j} = \begin{cases} 1, & \text{if } i \text{ divides } j, \\ 0 & \text{otherwise.} \end{cases}$$

It is also of interest to study noncommutative subalgebras of  $\mathcal{DR}$  or nonabelian subgroups of  $\mathcal{DR}^\times$  which contain  $D$ . In [8], it was shown that  $\langle D \rangle$  could be embedded as the cyclic subgroup of index 2 in an infinite dihedral subgroup of  $\mathcal{DR}^\times$ . Given a group  $G$ , the problem of finding a subgroup of  $\mathcal{DR}^\times$  isomorphic with  $G$  and containing  $D$  is equivalent to the problem of finding a matrix representation of  $G$  into  $\mathcal{DR}^\times$  in which some group element is represented by  $D$ .

As a reduction step for this general problem, it is desirable to transform the divisor matrix into a Jordan canonical form. Since  $\mathcal{DR}$  is neither closed under matrix inversion nor similarity, such a reduction is useful only if the transition matrices belong to  $\mathcal{DR}^\times$ . We show explicitly (Lemma 4.5 and Theorem 4.8) that  $D$  can be transformed to a Jordan canonical form by matrices in  $\mathcal{DR}^\times$  with integer entries.

The remainder of the paper is devoted to the group  $G = \text{SL}(2, \mathbf{Z})$ . We consider the problem of constructing a representation  $\rho : G \rightarrow \mathcal{DR}^\times$  such that the standard unipotent element  $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$  is represented by  $D$  and such that an element of order 3 acts without fixed points. The precise statement of the solution of this problem is Theorem 3.1 below. Roughly speaking, the result on the Jordan canonical form reduces the problem to one of constructing a representation of  $G$  in which  $T$  is represented by a standard infinite Jordan block. In the simplified problem, the polynomial growth condition on the matrices representing group elements becomes a certain exponential growth condition and the fixed-point-free condition is unchanged. The construction is the content of Theorem 5.1. Since the group  $G$  has few relations, it is relatively easy to define matrix representations satisfying the growth and fixed-point-free conditions and with  $T$  acting indecomposably. However, the growth condition is not in general preserved under similarity and it is a

more delicate matter to find such a representation for which one can prove that the matrix representing  $T$  can be put into Jordan form by transformations preserving the growth condition.

As we have indicated, the construction of the representation  $\rho$  involves making careful choices and we do not know of an abstract characterization of  $\rho$  as a matrix representation. The  $CG$ -module affording  $\rho$  can, however, be characterized as the direct sum of isomorphic indecomposable modules where the isomorphism type of the summands is uniquely determined up to  $CG$ -isomorphism by the indecomposable action of  $T$  and the existence of a filtration by standard 2-dimensional modules (Theorem 8.1). Although  $\rho$  is just one among many matrix representations of  $G$  into  $\mathcal{DR}^\times$  which satisfy our conditions, we proceed to examine the orbit of  $\zeta(s)$  under  $\rho(G)$ . We find (Theorem 10.3) that one series  $\varphi(s)$  in this orbit is related to  $\zeta(s)$  by the cubic equation

$$(\zeta(s) - 1)\varphi(s)^2 + \zeta(s)\varphi(s) - \zeta(s)(\zeta(s) - 1) = 0. \quad (2)$$

We also show (Theorem 9.1) that, as a consequence of relations in the image of the group algebra, the other series in the orbit belong to  $\mathbf{C}(\zeta(s), \varphi(s))$ .

The cubic equation may be rewritten as:

$$-\varphi(s) = (\zeta(s) - 1)(\varphi(s)^2 + \varphi(s) - \zeta(s)). \quad (3)$$

The second factor on the right is a unit in  $\mathcal{D}\{s\}$ , so  $\varphi$  and  $\zeta(s) - 1$  are associate irreducible elements in the factorial ring  $\mathcal{D}\{s\}$ . The fact that there is a cubic equation relating  $\zeta(s)$  and an associate of  $\zeta(s) - 1$  should be contrasted with the classical theorem of Ostrowski [5], which states that  $\zeta(s)$  does not satisfy any algebraic differential-difference equation.

Matrices resembling finite truncations of the divisor matrix were studied by Redheffer in [6]. For each natural number  $n$ , he considered the matrix obtained from the upper left  $n \times n$  submatrix of  $D$  by setting each entry in the first column equal to 1. Research on Redheffer's matrices has been motivated by the fact that their determinants are the values of Mertens' function, which links them directly to the Riemann Hypothesis. (See [1, 11], and [10].)

## 2 Basic Definitions and Notation

Let  $\mathbf{N}$  denote the natural numbers  $\{1, 2, \dots\}$  and  $\mathbf{C}$  the complex numbers. Let  $E$  be the free  $\mathbf{C}$ -module with basis  $\{e_n\}_{n \in \mathbf{N}}$ . With respect to this basis, the endomorphism ring  $\text{End}_{\mathbf{C}}(E)$  acting on the left of  $E$  becomes identified with the ring  $\mathcal{A}$  of matrices  $A = (a_{i,j})_{i,j \in \mathbf{N}}$ , with complex entries, such that each column has only finitely many nonzero entries. The dual space  $E^*$  becomes identified with the space  $\mathbf{C}^{\mathbf{N}}$  of sequences of complex numbers, with  $f \in E^*$  corresponding to the sequence  $(f(e_n))_{n \in \mathbf{N}}$ . We will write  $(f(e_n))_{n \in \mathbf{N}}$  as  $f$  and  $f(e_n)$  as  $f(n)$  for short.



In this notation, the natural right action of  $\text{End}_{\mathbb{C}}(E)$  on  $E^*$  is expressed as a right action of  $\mathcal{A}$  on  $\mathbb{C}^{\mathbb{N}}$  by

$$(fA)(n) = \sum_{m \in \mathbb{N}} a_{m,n} f(m), \quad f \in \mathbb{C}^{\mathbb{N}}, A \in \mathcal{A}.$$

(The sum has only finitely many nonzero terms.)

Let  $\mathcal{DS}$  be the subspace of  $\mathbb{C}^{\mathbb{N}}$  consisting of the sequences  $f$  for which there exist positive constants  $C$  and  $c$  such that for all  $n$ ,  $|f(n)| \leq Cn^c$ . A sequence  $f$  lies in  $\mathcal{DS}$  if and only if the Dirichlet series  $\sum_n f(n)n^{-s}$  converges for some complex number  $s$ , which gives a canonical bijection between  $\mathcal{DS}$  and the space  $\mathcal{D}\{s\}$  of analytic Dirichlet series. Let  $\mathcal{DR}$  be the subalgebra of  $\mathcal{A}$  consisting of all elements which leave  $\mathcal{DS}$  invariant.

A sufficient condition for membership in  $\mathcal{DR}$  is provided by the following lemma, whose proof is straightforward.

**Lemma 2.1.** *Let  $A = (a_{i,j})_{i,j \in \mathbb{N}} \in \mathcal{A}$ . Suppose that there exist positive constants  $C$  and  $c$  such that the following hold.*

- (i)  $a_{i,j} = 0$  whenever  $i > Cj^c$ .
- (ii) For all  $i$  and  $j$  we have  $|a_{i,j}| \leq Cj^c$ .

*Then  $A \in \mathcal{DR}$ . Furthermore, the set of all elements of  $\mathcal{A}$  which satisfy these conditions, where the constants may depend on the matrix, is a subring of  $\mathcal{DR}$ .*

We let  $\mathcal{DR}_0$  denote the subring of  $\mathcal{DR}$  defined by the lemma. If  $\sum_{n \in \mathbb{N}} f(n)n^{-s} \in \mathcal{D}\{s\}$ , then its multiplication operator has matrix with  $(i, ni)$  entries equal to  $f(n)$  for all  $i$  and  $n$  and all other entries zero, so the multiplication operators form a commutative subalgebra of  $\mathcal{DR}_0$ .

*Remarks 2.2.* There exist elements of  $\mathcal{DR}$  which do not satisfy the hypotheses of Lemma 2.1. An example is the matrix  $(a_{i,j})_{i,j \in \mathbb{N}}$  defined by

$$a_{i,j} = \begin{cases} \frac{1}{j^{j^2}}, & \text{if } i = j^j, \\ 0, & \text{otherwise.} \end{cases}$$

There are invertible matrices in  $\mathcal{DR}_0$  whose inverses are not in  $\mathcal{DR}$ . For example, the matrix  $(b_{i,j})_{i,j \in \mathbb{N}}$ , given by

$$b_{i,j} = \begin{cases} 0, & \text{if } i > j, \\ 1, & \text{if } i = j, \\ -1 & \text{if } i < j \end{cases}$$

is obviously in  $\mathcal{DR}_0$ , while its inverse, given by

$$b'_{i,j} = \begin{cases} 0, & \text{if } i > j, \\ 1, & \text{if } i = j, \\ 2^{j-i-1} & \text{if } i < j \end{cases}$$

is not in  $\mathcal{DR}$ .

### 3 An Action of $SL(2, \mathbf{Z})$ on Dirichlet Series

The group  $G = SL(2, \mathbf{Z})$  is generated by the matrices

$$S = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad \text{and} \quad T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}. \tag{4}$$

These matrices satisfy the relations

$$S^4 = (ST)^6 = 1, \quad S^2 = (ST)^3 \tag{5}$$

which, as is well known, form a set of defining relations for  $SL(2, \mathbf{Z})$  as an abstract group.

With the above definitions, we can state one of our principal results.

**Theorem 3.1.** *There exists a representation  $\rho : SL(2, \mathbf{Z}) \rightarrow \mathcal{A}^\times$  with the following properties.*

(a) *The underlying  $\mathbf{CSL}(2, \mathbf{Z})$ -module  $E$  has an ascending filtration*

$$0 = E_0 \subset E_1 \subset E_2 \subset \dots$$

*of  $\mathbf{CSL}(2, \mathbf{Z})$ -submodules such that for each  $i \in \mathbf{N}$ , the quotient module  $E_i/E_{i-1}$  is isomorphic to the standard 2-dimensional  $\mathbf{CSL}(2, \mathbf{Z})$ -module.*

(b)  $\rho(T) = D$ .

(c)  $\rho(Y)$  is an integer matrix for every  $Y \in SL(2, \mathbf{Z})$ .

(d)  $\rho(SL(2, \mathbf{Z})) \subseteq \mathcal{DR}_0$ .

The facts needed for the proof of Theorem 3.1 are established in the following sections and the proof is completed in Sect. 6.

### 4 A Jordan Form of the Divisor Matrix

For  $m, k \in \mathbf{N}$ , let

$$A_k(m) = \{(m_1, m_2, \dots, m_k) \in (\mathbf{N} \setminus \{1\})^k \mid m_1 m_2 \cdots m_k = m\}$$

and let  $\alpha_k(m) = |A_k(m)|$ .

The following properties of these numbers follow from the definitions.

**Lemma 4.1.** (a)  $\alpha_k(1) = 0$ .

(b)  $\alpha_k(m) = 0$  if  $m < 2^k$  and  $\alpha_k(2^k) = 1$ .

(c)

$$(\zeta(s) - 1)^k = \sum_{m=2^k}^{\infty} \frac{\alpha_k(m)}{m^s}.$$

By considering the first  $k - 1$  entries of elements of  $A_k(m)$ , we see that for  $k > 1$ , we have

$$\alpha_k(m) = \left( \sum_{d|m} \alpha_{k-1}(d) \right) - \alpha_{k-1}(m). \tag{6}$$

Induction yields the following formula.

**Lemma 4.2.**

$$\sum_{i=1}^{k-1} (-1)^{k-1-i} \sum_{d|m} \alpha_i(d) = \alpha_k(m) + (-1)^k \alpha_1(m). \tag{7}$$

**Lemma 4.3.** *There exists a constant  $c$  such that  $\alpha_k(m) \leq m^c$  for all  $k$  and  $m$ .*

*Proof.* We choose  $c$  with  $\zeta(c)=2$ . We proceed by induction on  $k$ . Since  $\alpha_1(m) \leq 1$ , the result is true when  $k = 1$ . Suppose for some  $k$  we have that for all  $m$ ,

$$\alpha_k(m) \leq m^c.$$

Then by (6) we have

$$\alpha_{k+1}(m) \leq \sum_{\substack{d|m \\ 1 < d < m}} \alpha_k(m/d) \leq m^c \sum_{\substack{d|m \\ 1 < d < m}} d^{-c} \leq m^c (\zeta(c) - 1) = m^c,$$

which completes the inductive proof. □

*Remark 4.4.* Since  $\zeta(2) = \pi^2/6$ , the constant  $c$  can be chosen from the real interval  $(1, 2)$ .

Let  $J = (J_{i,j})_{i,j \in \mathbb{N}}$  be the matrix defined by

$$J_{i,j} = \begin{cases} 1, & \text{if } j \in \{i, 2i\}, \\ 0 & \text{otherwise.} \end{cases}$$

Let  $Z = (\alpha(i, j))_{i,j \in \mathbb{N}}$  be the matrix described in the following way. The odd rows have a single nonzero entry, equal to 1 on the diagonal. Let  $i = 2^k d$  with  $d$  odd. Then the  $i^{\text{th}}$  row of  $Z$  is equal to the  $d^{\text{th}}$  row of  $(D - I)^k$ .

**Lemma 4.5.** *The matrix  $Z$  has the following properties:*

- (a)  $\alpha(i, j) = \delta_{i,j}$ , if  $i$  is odd.
- (b) If  $i = d2^k$ , where  $d$  is odd and  $k \geq 1$ , then

$$\alpha(i, j) = \begin{cases} \alpha_k(j/d) & \text{if } d \mid j, \\ 0 & \text{otherwise.} \end{cases}$$

- (c)  $\alpha(im, jm) = \alpha(i, j)$  whenever  $m$  is odd.
- (d)  $Z$  is upper unitriangular.
- (e)  $ZDZ^{-1} = J$ .
- (f)  $Z \in \mathcal{DR}_0$ .

Moreover,  $Z$  is the unique matrix satisfying (a) and (e).

*Proof.* Part (a) is by definition. Part (b) follows from Lemma 4.1(c) upon multiplying by the Dirichlet series with one term  $d^{-s}$ . Part (c) is a special case of (b). Part (d) also follows from (b). Part (f) is then immediate from Lemma 4.3. In the equation

$$Z(D - I) = (J - I)Z \tag{8}$$

the  $n^{\text{th}}$  row of each side is equal to the  $2n^{\text{th}}$  row of  $Z$ . This proves (e) since  $Z$  is invertible by (d). To prove the last statement we see that if (e) holds then by (8) we have for all  $i$  and  $k \in \mathbb{N}$ ,

$$\sum_{\substack{j|k \\ j < k}} \alpha(i, j) = \alpha(2i, k),$$

which determines  $Z$  uniquely since the rows with odd index are specified by (a). □

Our aim is to determine  $Z^{-1}$  explicitly.

For each prime  $p$  and each integer  $m$ , let  $v_p(m)$  denote the exponent of the highest power of  $p$  which divides  $m$  and let  $v(m) = \sum_p v_p(m)$ .

**Lemma 4.6.** *We have for all  $m \in \mathbb{N}$ ,*

$$\sum_{k=1}^{v(m)} (-1)^k \alpha_k(m) = \begin{cases} (-1)^{v(m)}, & \text{if } m \text{ is squarefree and } m > 1, \\ 0 & \text{otherwise.} \end{cases}$$

*Proof.* The case  $m = 1$  is trivial. Suppose  $m = p_1^{\lambda_1} p_2^{\lambda_2} \cdots p_r^{\lambda_r}$ , with  $\lambda_1, \lambda_2, \dots, \lambda_r \geq 1$  and  $r \geq 1$ . In the ring of formal power series  $\mathbb{C}[[t_1, \dots, t_r]]$  in  $r$  indeterminates, we set

$$\begin{aligned} y &= \frac{1}{(1-t_1)(1-t_2)\cdots(1-t_r)} - 1 \\ &= \sum_{(n_1, \dots, n_r) \in (\mathbb{N} \cup \{0\})^r} \alpha_1(p_1^{n_1} \cdots p_r^{n_r}) t_1^{n_1} \cdots t_r^{n_r}. \end{aligned} \tag{9}$$

Then for  $k \geq 1$

$$y^k = \sum_{(n_1, \dots, n_r) \in (\mathbb{N} \cup \{0\})^r} \alpha_k(p_1^{n_1} \cdots p_r^{n_r}) t_1^{n_1} \cdots t_r^{n_r}. \tag{10}$$

Then we have

$$\begin{aligned} \prod_{i=1}^r (1 - t_i) - 1 &= \frac{-y}{1 + y} \\ &= \sum_{k \in \mathbb{N}} (-1)^k y^k \\ &= \sum_{(n_1, \dots, n_r) \in (\mathbb{N} \cup \{0\})^r} \left[ \sum_{k \in \mathbb{N}} (-1)^k \alpha_k(p_1^{n_1} \cdots p_r^{n_r}) \right] t_1^{n_1} \cdots t_r^{n_r}. \end{aligned} \tag{11}$$

The lemma follows by equating the coefficients of monomials. □

*Remark 4.7.* The above proof is similar to the argument in [4], p. 21, used to show that  $\sum_k (-1)^k \alpha_k(m)/k$  is equal to  $1/h$  if  $m$  is the  $h$ -th power of a prime, and zero otherwise. The lemma has also the following enumerative proof, based on another combinatorial interpretation of the sets  $A_k(m)$ . From the above factorization of  $m$ , let  $\lambda$  be the partition of  $n$  defined by the  $\lambda_i$ . Let  $N = \{1, \dots, n\}$  and let  $F_\lambda$  be the set of functions  $h : N \rightarrow \{p_1, \dots, p_r\}$  such that  $|h^{-1}(p_i)| = \lambda_i$  for  $i = 1, \dots, r$ . The symmetric group  $S_n$  acts transitively on the right of  $F_\lambda$  by the rule  $(h\sigma)(y) = h(\sigma(y))$ ,  $y \in N$ ,  $\sigma \in S_n$ . The stabilizer  $S_\lambda$  of the function mapping the first  $\lambda_1$  elements to  $p_1$ , the next  $\lambda_2$  elements to  $p_2$ , etc. is isomorphic to  $S_{\lambda_1} \times S_{\lambda_2} \times \cdots \times S_{\lambda_r}$ . A  $k$ -decomposition of  $n$  is a  $k$ -tuple  $(n_1, \dots, n_k)$  of integers  $n_i \geq 1$  such that  $n_1 + n_2 + \cdots + n_k = n$ . Let  $\Pi = \{\sigma_1, \dots, \sigma_{n-1}\}$  be the set of fundamental reflections, with  $\sigma_i = (i, i + 1)$ . The subgroup  $W_K$  of  $S_n$  generated by a subset  $K$  of  $\Pi$  is called a standard parabolic subgroup of rank  $|K|$ . Given a  $k$ -decomposition  $(n_1, \dots, n_k)$  of  $n$ , we have a set decomposition of  $N$  into subsets  $N_1 = \{1, \dots, n_1\}$ ,  $N_2 = \{n_1 + 1, \dots, n_1 + n_2\}$ , ...,  $N_k = \{n_1 + \cdots + n_{k-1} + 1, \dots, n\}$ . The stabilizer of this decomposition is a standard parabolic subgroup of rank  $n - k$  and this correspondence is a bijection between  $k$ -decompositions and standard parabolic subgroups of rank  $n - k$ .

For each pair  $((n_1, \dots, n_k), h)$  consisting of a  $k$ -decomposition and a function  $h \in F_\lambda$ , we obtain an element  $(m_1, \dots, m_k) \in A_k(m)$  by setting  $m_i = \prod_{j \in N_i} h(j)$ . Every element of  $A_k(m)$  arises in this way and two pairs define the same element of  $A_k(m)$  if and only if the  $k$ -decompositions are equal and the corresponding functions are in the same orbit under the action of the parabolic subgroup of the  $k$ -decomposition.

Thus, we have

$$\alpha_k(m) = |A_k(m)| = \sum_{\substack{K \subseteq \Pi \\ |K|=n-k}} |\{W_K\text{-orbits on } F_\lambda\}|.$$

The number of  $W_K$ -orbits on  $F_\lambda$  can be expressed as the inner product of permutation characters, so

$$\alpha_k(m) = \sum_{\substack{K \subseteq \Pi \\ |K|=n-k}} \langle 1_{W_K}^{S_n}, 1_{S_\lambda}^{S_n} \rangle.$$

Now, it is a well-known fact [7] that

$$\sum_{K \subseteq \Pi} (-1)^{|K|} 1_{W_K}^{S_n} = \epsilon,$$

where  $\epsilon$  is the sign character. Hence,

$$\begin{aligned} \sum_{k=1}^n (-1)^k \alpha_k(m) &= (-1)^n \left\langle \sum_{K \subseteq \Pi} (-1)^{|K|} 1_{W_K}^{S_n}, 1_{S_\lambda}^{S_n} \right\rangle \\ &= (-1)^n \langle \epsilon, 1_{S_\lambda}^{S_n} \rangle \\ &= (-1)^n \langle \epsilon, 1 \rangle_{S_\lambda} \\ &= \begin{cases} (-1)^n, & \text{if } \lambda = 1^n, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Let  $X$  be the diagonal matrix with  $(i, i)$  entry equal to  $(-1)^{v_2(i)}$ , for  $i \in \mathbb{N}$ .

**Theorem 4.8.**  $Z^{-1} = XZX$ . In particular,  $Z^{-1} \in \mathcal{DR}_0$ .

*Proof.* If  $i$  is odd then the  $i^{\text{th}}$  row of  $Z$  is zero except for 1 in the  $i^{\text{th}}$  column, so the same holds for  $XZX$ . By the last assertion of Lemma 4.5 it is sufficient to show that

$$D(XZX) = (XZX)J$$

or, equivalently,

$$(XDX)Z = Z(XJX).$$

The matrices  $XDX = (d'_{i,j})_{i,j \in \mathbb{N}}$  and  $XJX = (c'_{i,j})_{i,j \in \mathbb{N}}$  are given by

$$d'_{i,j} = \begin{cases} (-1)^{v_2(i)+v_2(j)}, & \text{if } i \mid j, \\ 0 & \text{otherwise.} \end{cases}, \quad c'_{i,j} = \begin{cases} 1, & \text{if } j = i, \\ -1, & \text{if } j = 2i, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, we must show that

$$\sum_{m \geq 1} (-1)^{v_2(m)} \alpha(im, j) = \begin{cases} \alpha(i, j), & \text{if } j \text{ is odd,} \\ \alpha(i, j) - \alpha(i, j/2), & \text{if } j \text{ is even.} \end{cases} \tag{12}$$

It is sufficient to consider the case  $i = 2^k$ , for  $k \geq 1$ , by Lemma 4.5(c). In this case, the left hand side of (12) can be rewritten as

$$\sum_{\substack{d|j \\ d \text{ odd}}} \sum_{e=0}^{v(j)-k-v(d)} (-1)^e \alpha(2^{k+e}, j/d) = (-1)^k \sum_{\substack{d|j \\ d \text{ odd}}} \sum_{r=k}^{v(j/d)} (-1)^r \alpha(2^r, j/d). \tag{13}$$

Suppose that we can prove for all  $j$ , that

$$(-1)^k \sum_{d|j} \sum_{r=k}^{v(j/d)} (-1)^r \alpha(2^r, j/d) = \alpha(2^k, j). \tag{14}$$

Then we will have proved (12) if  $j$  is odd. If  $j$  is even, we note that  $d$  is a divisor of  $j/2$  if and only if  $2d$  is an even divisor of  $j$  so that (14) implies

$$\begin{aligned} \alpha(2^k, j/2) &= (-1)^k \sum_{d|(j/2)} \sum_{r=k}^{v((j/2)/d)} (-1)^r \alpha(2^r, (j/2)/d) \\ &= (-1)^k \sum_{\substack{d'|j \\ d' \text{ even}}} \sum_{r=k}^{v(j/d')} (-1)^r \alpha(2^r, j/d'). \end{aligned}$$

Thus, from (13) we see that (12) also follows from (14) when  $j$  is even. It remains to prove (14). We can assume  $j > 1$ , by Lemma 4.1(a). Lemma 4.6, applied to the left hand side of (14), yields

$$(-1)^{k-1} + (-1)^{k-1} \sum_{d|j} \left( \sum_{r=1}^{k-1} (-1)^r \alpha_r(j/d) \right) \tag{15}$$

because the total contribution from the squarefree case of Lemma 4.6 is

$$(-1)^k \sum_{\substack{d|j \\ j/d \text{ squarefree} \\ j/d > 1}} (-1)^{v(j/d)} = (-1)^{k-1}.$$

We can rewrite (15) as

$$(-1)^{k-1} + \sum_{r=1}^{k-1} (-1)^{k-1-r} \sum_{d|j} \alpha_r(d),$$

which, by Lemma 4.2 is equal to  $\alpha_k(j)$ . This proves (14). □

## 5 Construction of Representations

Let  $J_\infty$  be the “infinite Jordan block”, indexed by  $\mathbf{N} \times \mathbf{N}$ , defined by

$$(J_\infty)_{i,j} = \begin{cases} 1, & \text{if } j = i \text{ or } j = i + 1, \\ 0 & \text{otherwise.} \end{cases}$$

In the following theorem,  $T$  and  $S$  are the generators of  $SL(2, \mathbf{Z})$  defined in (4).

**Theorem 5.1.** *There exists a representation  $\tau : SL(2, \mathbf{Z}) \rightarrow \mathcal{A}^\times$  with the following properties.*

(a) *Let  $E_i$  be the subspace of  $E$  spanned by  $\{e_1, \dots, e_{2i}\}$ ,  $i \in \mathbf{N}$ . Then*

$$0 = E_0 \subset E_1 \subset E_2 \subset \dots$$

*is a filtration of  $CSL(2, \mathbf{Z})$ -modules and for each  $i \in \mathbf{N}$  the quotient module  $E_i/E_{i-1}$  is isomorphic to the standard 2-dimensional  $CSL(2, \mathbf{Z})$ -module.*

(b)  $\tau(T) = J_\infty$ .

(c)  $\tau(Y)$  is an integer matrix for every  $Y \in SL(2, \mathbf{Z})$ .

(d) *There is a constant  $C$  such that for all  $i$  and  $j$  we have  $|\tau(S)_{i,j}| \leq 2^{Cj}$ .*

Later we will show (Theorem 8.1) that there is a unique  $CSL(2, \mathbf{Z})$ -module with a filtration by standard modules and such that  $T$  acts indecomposably and unipotently on every  $T$ -invariant subspace.

We define a sequence of integers  $\{b_n\}_{n \geq 0}$  recursively by<sup>1</sup>

$$b_0 = b_1 = 1, \quad b_n + \sum_{\substack{i,j \geq 1 \\ i+j=n}} b_i b_j = 0 \quad \text{for all } n \geq 2. \quad (16)$$

Let  $\mathbf{C}[[t]]$  denote the ring of formal power series over  $\mathbf{C}$  and let  $g(t) \in \mathbf{C}[[t]]$  be defined by

$$1 + g(t) = \sum_{k=0}^{\infty} b_k t^k.$$

Then the recurrence relations satisfied by the  $b_i$  can be stated as the equation

$$g(t)^2 + g(t) = t. \quad (17)$$

Thus,

$$g(t) = \frac{-1 + \sqrt{1 + 4t}}{2},$$

---

<sup>1</sup> As J-P. Serre has pointed out to us, this is the sequence of Catalan numbers, up to signs.



where the positive square root is taken since  $g(t)$  has no constant term. By Taylor expansion we obtain

$$b_m = \frac{(-1)^{m-1}}{m} \binom{2m-2}{m-1}, \quad (m \geq 2), \quad b_1 = b_0 = 1. \tag{18}$$

Let

$$B_0 = T, \quad B_1 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad B_i = \begin{bmatrix} 0 & b_i \\ b_i & 0 \end{bmatrix}, \quad (i \geq 2)$$

and define

$$\tilde{J} = \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & \dots \\ 0 & B_0 & B_1 & B_2 & \dots \\ 0 & 0 & B_0 & B_1 & \dots \\ 0 & 0 & 0 & B_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$\tilde{S} = \text{diag}(S, S, \dots),$$

$$\tilde{R} = -\tilde{S}\tilde{J}.$$

For any ring  $R$ , let  $M_n(R)$  denote the ring of  $n \times n$  matrices over  $R$ . Let  $\mathcal{U}$  denote the ring of matrices of the form

$$U = \begin{bmatrix} X^{(0)} & X^{(1)} & X^{(2)} & X^{(3)} & \dots \\ 0 & X^{(0)} & X^{(1)} & X^{(2)} & \dots \\ 0 & 0 & X^{(0)} & X^{(1)} & \dots \\ 0 & 0 & 0 & X^{(0)} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{19}$$

where, for all  $n \geq 0$ ,

$$X^{(n)} = \begin{bmatrix} x_{1,1}^{(n)} & x_{1,2}^{(n)} \\ x_{2,1}^{(n)} & x_{2,2}^{(n)} \end{bmatrix} \in M_2(\mathbf{C})$$

and the  $X^{(n)}$  are repeated down the diagonals. For example,  $\tilde{S}$ ,  $\tilde{J}$ , and  $\tilde{R}$  all belong to  $\mathcal{U}$ . The center  $Z(\mathcal{U})$  consists of those matrices in which the submatrices  $X^{(n)}$  are all scalar matrices. The map  $\mathbf{C}[[t]] \rightarrow Z(\mathcal{U})$  sending  $\sum_{n \geq 0} a_n t^n$  to the matrix with  $X^{(n)} = a_n I$ , for all  $n \geq 0$ , is a  $\mathbf{C}$ -algebra isomorphism, and extends to a  $\mathbf{C}[[t]]$ -algebra isomorphism

$$\gamma : \mathcal{U} \rightarrow M_2(\mathbf{C}[[t]]), \quad U \mapsto \begin{bmatrix} x_{1,1}(t) & x_{1,2}(t) \\ x_{2,1}(t) & x_{2,2}(t) \end{bmatrix}, \tag{20}$$

where

$$x_{i,j}(t) = \sum_{n=0}^{\infty} x_{i,j}^{(n)} t^n, \quad i, j \in \{1, 2\}.$$

We have:

$$\gamma(\tilde{S}) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad \gamma(\tilde{J}) = \begin{bmatrix} 1 & 1 + g(t) \\ g(t) & 1 + t \end{bmatrix}, \quad \gamma(\tilde{R}) = \begin{bmatrix} g(t) & 1 + t \\ -1 & -1 - g(t) \end{bmatrix}. \tag{21}$$

**Lemma 5.2.** (a)  $\tilde{S}^2 = -I$ .

(b)  $\tilde{R}^2 + \tilde{R} + I = 0$ .

(c) *There exists a representation  $\tau_1$  of  $G$  such that  $\tau_1(S) = \tilde{S}$  and  $\tau_1(T) = \tilde{J}$ .*

*Proof.* Part (a) is obvious and (b) is easy to check by direct computation using (21) and (17). By (a) and (b), the elements  $\tilde{S}$  and  $\tilde{J}$  satisfy the defining relations (5) for  $SL(2, \mathbb{Z})$ , so (c) holds.  $\square$

The representation  $\tau_1$  satisfies all the conditions of Theorem 5.1 except for (b). To complete the proof of Theorem 5.1, we shall conjugate this representation by an upper unitriangular integer matrix  $P$  such that  $P\tilde{J}P^{-1} = J_\infty$ . In order to check that  $P\tilde{S}P^{-1}$  satisfies condition (d) of Theorem 5.1, we will need to compute  $P$  and its inverse explicitly.

### 5.1 Transforming $\tilde{J}$ into Jordan Form

A matrix  $P$  such that  $P\tilde{J}P^{-1} = J_\infty$  can be found by following the usual method for computing Jordan blocks. Thus, for  $n \in \mathbb{N}$ , we define the  $n^{\text{th}}$  row of  $P$  to be the first row of  $(\tilde{J} - I)^{n-1}$ , setting  $(\tilde{J} - I)^0 = I$ . Then  $P$  is upper unitriangular, hence invertible, and from its definition,  $P$  satisfies the equivalent equation

$$P(\tilde{J} - I) = (J_\infty - I)P.$$

We now compute the entries of  $P$  explicitly. In order to do this, we use the isomorphism  $\gamma$  of (20). Let  $\tilde{H} = \gamma(\tilde{J} - I)$ . Then

$$\tilde{H} = \begin{bmatrix} 0 & 1 + g(t) \\ g(t) & t \end{bmatrix}.$$

It is easy to compute the powers of  $\tilde{J} - I$  by diagonalizing  $\tilde{H}$ . Since  $g(t)^2 + g(t) = t$ , the characteristic polynomial of  $H$  is  $\chi(x) = x^2 - tx - t$ . Let  $\lambda_1$  and  $\lambda_2$  be the roots of this polynomial in some extension field and for  $n \geq 0$ , let  $h_n = (\lambda_1^{n+1} - \lambda_2^{n+1})/(\lambda_1 - \lambda_2)$  be the complete symmetric polynomial of degree

$n$  in two variables, evaluated at  $(\lambda_1, \lambda_2)$ . Then  $h_n$  is a polynomial in the coefficients of  $\chi(x)$ , so it is a polynomial in  $t$ . We have  $h_0 = 1$  and  $h_1 = t$ . A straightforward computation shows that, for  $n \geq 2$ ,

$$\tilde{H}^n = \begin{bmatrix} th_{n-2} & (1 + g(t))h_{n-1} \\ g(t)h_{n-1} & h_n \end{bmatrix}. \tag{22}$$

It follows from (22) and the equation  $g(t)^2 + g(t) = t$  that the polynomials  $h_n$  satisfy the recurrence

$$h_n = th_{n-1} + th_{n-2}, \quad (n \geq 2), \quad h_0 = 1, \quad h_1 = t.$$

By inspection, the solution is

$$h_n = \sum_{r=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n-r}{r} t^{n-r}.$$

Thus, we can compute the entries of  $P$  as coefficients of the powers of  $t$  in the top rows of the  $\tilde{H}^n$ . For  $\ell \geq 3$  and  $s \geq 0$ , we have

$$\begin{aligned} p_{\ell,2s+1} &= \text{coefficient of } t^s \text{ in } th_{\ell-3} \\ &= \binom{s-1}{\ell-s-2}, \\ p_{\ell,2s+2} &= \text{coefficient of } t^s \text{ in } (1 + g(t))h_{\ell-2} \\ &= \sum_{k=0}^{\lfloor s+1-\frac{\ell}{2} \rfloor} b_k \binom{s-k}{\ell+k-s-2}. \end{aligned} \tag{23}$$

*Remark 5.3.* Here and elsewhere, we employ the convention for binomial coefficients that  $\binom{a}{b} = 0$  unless  $a \geq b \geq 0$ .

We now turn to the computation of  $P^{-1}$ . Suppose a matrix  $Q = (q_{i,j})_{i,j \in \mathbb{N}}$  satisfies the two conditions

$$\tilde{J}Q = QJ_\infty \quad \text{and} \quad q_{1,j} = \delta_{1,j}. \tag{24}$$

The first condition implies that  $PQ$  commutes with  $J_\infty$  and the second that  $(PQ)_{1,j} = \delta_{1,j}$ , from which it follows that  $PQ = I$  and  $Q = P^{-1}$ . We find a matrix  $Q$  satisfying (24) by first finding a matrix  $A$  such that

$$(\tilde{J} - I)A = A(J_\infty - I) \tag{25}$$

and then modifying it. To compute  $A$ , we must first enlarge the ring  $\mathcal{U}$ . Let  $\widehat{\mathcal{U}}$  denote the set of matrices of the form

$$W = \begin{bmatrix} \dots & X^{(m)} & X^{(m+1)} & X^{(m+2)} & X^{(m+3)} & \dots \\ \dots & X^{(m-1)} & X^{(m)} & X^{(m+1)} & X^{(m+2)} & \dots \\ \dots & X^{(m-2)} & X^{(m-1)} & X^{(m)} & X^{(m+1)} & \dots \\ \dots & X^{(m-3)} & X^{(m-2)} & X^{(m-1)} & X^{(m)} & \dots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix},$$

where the blocks  $X^{(m)} \in M_2(\mathbb{C})$ , for  $m \in \mathbb{Z}$ , are repeated down the diagonals and have the property that for some  $m_0 \in \mathbb{Z}$ , which may depend on  $W$ ,  $X^{(m)} = 0$  whenever  $m < m_0$ . We shall refer to the two columns of  $W$  headed by  $X^{(m)}$  as the  $[m, 1]$  column and the  $[m, 2]$  column, respectively. For  $W \in \widehat{\mathcal{U}}$  and  $m \in \mathbb{Z}$ , we denote by  $W(m)$  the submatrix of  $W$  whose first column is the  $[m, 1]$  column of  $W$ . To be concise, we can write  $W = (X^{(m)})_{m \in \mathbb{Z}}$ , since the top row determines the whole matrix. A product is defined as follows. Let  $W' = (Y^{(m)})_{m \in \mathbb{Z}} \in \widehat{\mathcal{U}}$ . Then  $WW' = (Z^{(m)})_{m \in \mathbb{Z}}$ , where

$$Z^{(m)} = \sum_{i+j=m} X^{(i)}Y^{(j)}.$$

This product can be computed as an ordinary matrix product as follows. Let  $m_0$  and  $n_0$  be chosen such that  $X^{(m)} = 0$  for all  $m < m_0$  and  $Y^{(n)} = 0$  for all  $n < n_0$ . Then  $WW'$  is obtained from the ordinary matrix product  $W(m_0)W'(n_0)$  by adjoining columns of zeros to the left and declaring the first column of  $W(m_0)W'(n_0)$  to be the  $[m_0 + n_0, 1]$  column of the new matrix. The answer is independent of the choice of  $m_0$  and  $n_0$ , due to the diagonal pattern of elements of  $\widehat{\mathcal{U}}$ . Together with the usual vector space structure on matrices, the above product makes  $\widehat{\mathcal{U}}$  into a  $\mathbb{C}$ -algebra. The subset of elements  $W \in \widehat{\mathcal{U}}$  such that  $X^{(m)} = 0$  for all  $m < 0$  forms a subalgebra isomorphic to the algebra  $\mathcal{U}$  defined in (19). Let  $\mathbb{C}((t))$  denote the field of formal Laurent series, the field of fractions of  $\mathbb{C}[[t]]$ . The center  $Z(\widehat{\mathcal{U}})$  consists of the elements in which all the submatrices  $X^{(m)}$  are scalar. The map sending the Laurent series  $\sum_n a_n t^n$  to the element  $(X^{(m)})_{m \in \mathbb{Z}}$  such that  $X^{(m)} = a_m I$  for all  $m$ , is an isomorphism of  $\mathbb{C}((t))$  with  $Z(\widehat{\mathcal{U}})$ . This extends to an isomorphism of  $\mathbb{C}((t))$ -algebras

$$\widehat{\gamma} : \widehat{\mathcal{U}} \rightarrow M_2(\mathbb{C}((t))),$$

which is the unique extension of the isomorphism (20).

Now, the element  $\tilde{J} - I$  is invertible in  $\widehat{\mathcal{U}}$ , since  $\tilde{H} = \widehat{\gamma}(\tilde{J} - I)$  has determinant  $-t$ . We define  $A = (a_{i,j})_{i,j \in \mathbb{N}}$  by columns. For  $n \in \mathbb{N}$ , we set the  $n^{\text{th}}$  column of  $A$  equal to the  $[0, 1]$  column of  $(\tilde{J} - I)^{-(n-1)}$ . Then  $A$  satisfies (25), by construction. To compute the entries of  $A$ , we invert  $\tilde{H}$  and its powers (22) to obtain

$$(\tilde{H})^{-1} = -t^{-1} \begin{bmatrix} t & -(1 + g(t)) \\ -g(t) & 0 \end{bmatrix}$$

and

$$(\tilde{H})^{-n} = (-1)^n t^{-n} \begin{bmatrix} h_n & -(1 + g(t))h_{n-1} \\ -g(t)h_{n-1} & th_{n-2} \end{bmatrix}, \quad n \geq 2.$$

Then we read off the coefficients of the appropriate powers of  $t$  in the first columns. The first two columns of  $A$  are given by

$$\begin{aligned} a_{i,1} &= \delta_{i,1}, & i \in \mathbf{N}, \\ a_{1,2} &= -1, & a_{2,2} = 1, & a_{i,2} = 0, & i \geq 3. \end{aligned} \tag{26}$$

For  $m \geq 3$  and  $s \geq 0$ , we have

$$\begin{aligned} a_{2s+1,m} &= \text{coefficient of } t^{-s} \text{ in } (-1)^{m-1} t^{-(m-1)} h_{m-1} \\ &= (-1)^{m-1} \binom{m-s-1}{s} \\ a_{2s+2,m} &= \text{coefficient of } t^{-s} \text{ in } (-1)^m t^{-(m-1)} g(t) h_{m-2} \\ &= (-1)^m \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor - s} b_k \binom{m-s-k-1}{s+k-1}. \end{aligned} \tag{27}$$

Let  $Q = AJ_\infty = (q_{i,j})_{i,j \in \mathbf{N}}$ . We check that  $Q$  has the properties (24). Since  $\tilde{J}A = AJ_\infty$ , it is clear that  $\tilde{J}Q = QJ_\infty$ . We have

$$q_{i,j} = \begin{cases} a_{i,j}, & \text{if } j = 1, \\ a_{i,j} + a_{i,j-1}, & \text{if } j \geq 2. \end{cases} \tag{28}$$

Since  $a_{1,m} = (-1)^{m-1}$ , it follows that  $q_{1,j} = \delta_{1,j}$ . Thus,  $Q = P^{-1}$ .

Finally, the entries of  $Q$  are obtained by applying (28) to (26) and (27). Thus,  $q_{i,1} = \delta_{i,1}$  and  $q_{i,2} = \delta_{i,2}$ , for  $i \in \mathbf{N}$ . For  $m \geq 3$  and  $s \geq 0$ , we have

$$\begin{aligned} q_{2s+1,m} &= (-1)^{m-1} \binom{m-s-2}{s-1} \\ q_{2s+2,m} &= (-1)^m \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor - s} b_k \binom{m-s-k-2}{s+k-2}. \end{aligned} \tag{29}$$

**Lemma 5.4.** For all  $i$  and  $j$ , we have  $|q_{i,j}| \leq 2^{3j}$  and  $|p_{i,j}| \leq 2^{2j}$ .

*Proof.* The bound  $|b_k| \leq 2^{2k-2}$  for  $k \geq 1$  follows from (18). It is then elementary to verify the bounds of the lemma from the formulae (23) and (29).  $\square$

**Proof of Theorem 5.1** We define

$$\tau(Y) = P\tau_1(Y)P^{-1}, \quad Y \in SL(2, \mathbb{Z}).$$

From its construction,  $\tau$  satisfies conditions (a), (b), and (c) of Theorem 5.1. It follows from Lemma 5.4 that  $\tau(S) = P\tilde{S}P^{-1}$  satisfies (d).  $\square$

## 6 Proof of Theorem 3.1

The matrix  $Z^{-1}$  studied in Sect. 4 is the transition matrix from the basis  $\{e_n\}_{n \in \mathbb{N}}$  of  $E$  to a new basis  $\{e'_n\}_{n \in \mathbb{N}}$ . The linear transformation represented by the divisor matrix  $D$  in the basis  $\{e_n\}_{n \in \mathbb{N}}$  is represented by  $J$  in the basis  $\{e'_n\}_{n \in \mathbb{N}}$ .

Since  $\mathbb{N} = \bigcup_{d \text{ odd}} \{d2^{k-1} \mid k \in \mathbb{N}\}$  we have a decomposition

$$E = \bigoplus_{d \text{ odd}} E(d),$$

where  $E(d)$  is the subspace of  $E$  spanned by the elements  $e'_{d2^{k-1}}, k \in \mathbb{N}$ .

We consider the isomorphisms

$$\phi_d : E \rightarrow E(d), \quad e_k \mapsto e'_{d2^{k-1}}.$$

For each odd number  $d$ , let  $\mathcal{A}(d)$  be the subring of  $\mathcal{A}$  consisting of matrices whose entries  $a_{i,j}$  are zero unless  $i$  and  $j$  both belong to the set  $\{d2^{k-1} \mid k \in \mathbb{N}\}$ .

The above isomorphisms induce isomorphisms

$$\psi_d : \mathcal{A} \rightarrow \mathcal{A}(d).$$

and a homomorphism

$$\psi : \mathcal{A} \rightarrow \prod_{d \text{ odd}} \mathcal{A}(d) \subseteq \mathcal{A}, \quad \psi(A) = (\psi_d(A))_{d \text{ odd}}$$

We have

$$\psi(J_\infty) = J.$$

Now for  $A \in \mathcal{A}$ , we have  $\psi(A)_{i,j} = 0$  unless there exists an odd number  $d$  and  $k, \ell \in \mathbb{N}$  with  $(i, j) = (d2^{k-1}, d2^{\ell-1})$ , in which case  $\psi(A)_{i,j} = A_{k,\ell}$ .

Let  $\tau$  be the representation given by Theorem 5.1 and let  $\tau(S) = (s_{k,\ell})_{k,\ell \in \mathbb{N}}$ . By Theorem 5.1(a),  $s_{k,\ell} = 0$  if  $k > \ell + 1$ . This means  $\psi(\tau(S))_{i,j} = 0$  if  $i > 2j$ . By Theorem 5.1(d), there exists a constant  $C$  such that  $|s_{k,\ell}| \leq 2^{C\ell}$ , for all  $k$  and  $\ell$ , which implies that  $|\psi(\tau(S))_{i,j}| \leq 2^C j^C$ , for all  $i$  and  $j$ . We conclude that

$\psi(\tau(S)) \in \mathcal{DR}_0$ . Since  $\psi(\tau(T)) = J$ , it follows that  $\psi(\tau(\text{SL}(2, \mathbf{Z}))) \subseteq \mathcal{DR}_0$ . Finally, the representation

$$\rho : \text{SL}(2, \mathbf{Z}) \rightarrow \mathcal{A}^\times, \quad Y \mapsto Z^{-1}\psi(\tau(Y))Z$$

satisfies all of the conditions of Theorem 3.1. The proof of Theorem 3.1 is now complete.  $\square$

*Remarks 6.1.* By a closer examination of the proof, we can strengthen the conclusions of Theorem 3.1 in the following ways. First, we have actually constructed the subgroup of  $\mathcal{DR}_0^\times$  isomorphic to the direct product of copies of  $\text{SL}(2, \mathbf{Z})$  (indexed by the odd numbers) with the representation  $\rho$  conjugate to the diagonal embedding. Also part (d) can be sharpened to state that for  $Y \in \text{SL}(2, \mathbf{Z})$  we have  $\rho(Y)_{i,j} = 0$  whenever  $i > 2j$ .

### 7 Extending Representations to $\text{GL}(2, \mathbf{Z})$

Let

$$W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

We have

$$W^2 = 1, \quad WSW = S^{-1}, \quad WRW = R^{-1}. \tag{30}$$

The relations (30) and (5) together form a set of defining relations for  $\text{GL}(2, \mathbf{Z}) = \langle \text{SL}(2, \mathbf{Z}), W \rangle$ .

In the following lemma, the isomorphism  $\gamma$  is defined in (20) and the matrices  $\gamma(\tilde{S})$  and  $\gamma(\tilde{R})$  are from (21).

**Lemma 7.1.** *Let*

$$W(t) = \frac{1}{\sqrt{t^2 + 4t + 1}} \begin{bmatrix} -t & 2g(t) + 1 \\ 2g(t) + 1 & t \end{bmatrix}.$$

Then  $W(t)$  is, up to a sign, the unique element of  $\text{GL}(2, \mathbf{C}[[t]])$  such that

- (i)  $W(t)^2 = 1$ .
- (ii)  $W(t)\gamma(\tilde{S})W(t) = \gamma(\tilde{S})^{-1}$
- (iii)  $W(t)\gamma(\tilde{R})W(t) = \gamma(\tilde{R})^{-1}$
- (iv)  $W(0) = W$ .

*Proof.* The proof is straightforward, by matrix calculations in  $M_2(\mathbf{C}[[t]])$ , using the relation (17).

**Lemma 7.2.** *For  $i, j \in \{1, 2\}$ , let  $w_{i,j}(t) = \sum_{n=0}^\infty r_n t^n$ . Then there exists a constant  $C$ , such that  $|r_n| \leq 2C^n$ .*

*Proof.* We consider those power series  $\sum_{n=0}^{\infty} s_n t^n$  with real coefficients for which there exists a constant  $D$ , which may depend on the series, such that  $|s_n| \leq 2^{Dn}$ . We observe that the product of two such series has the same property. Since  $g(t)$  has this property and since  $t^2 + 4t + 1 = (t + (2 + \sqrt{3}))(t + (2 - \sqrt{3}))$ , we are reduced to proving the bound for the Taylor series, centered at 0, of  $f(t) = (t+a)^{-\frac{1}{2}}$ , where  $a > 0$ . We have

$$f^{(n)}(t) = (-1)^n \frac{1 \cdot 3 \cdot 5 \cdots (2n - 1)}{2^n} (t + a)^{-\frac{(2n+1)}{2}},$$

Hence,

$$\left| \frac{f^{(n)}(0)}{n!} \right| = \frac{1}{2^{2n}} \binom{2n}{n} a^{-\frac{(2n+1)}{2}} \leq \frac{1}{\sqrt{a}} \left( \frac{1}{a} \right)^n.$$

□

**Proposition 7.3.** *The representation  $\rho : SL(2, \mathbf{Z}) \rightarrow \mathcal{DR}_0^\times$  can be extended to  $GL(2, \mathbf{Z})$ .*

*Proof.* Set  $\tilde{W} = \gamma^{-1}(W(t))$ . Then by Lemma 7.1, the group generated by  $\tilde{S}$ ,  $\tilde{R}$ , and  $\tilde{W}$  is isomorphic to  $GL(2, \mathbf{Z})$  and we can extend the representation  $\tau_1$  from  $SL(2, \mathbf{Z})$  to  $GL(2, \mathbf{Z})$  by setting  $\tau_1(W) = \tilde{W}$ . Hence we can also extend the representations  $\tau$  and  $\rho$  by setting  $\tau(W) = P \tau_1(W) P^{-1}$  and  $\rho(W) = Z^{-1} \psi(\tau(W)) Z$ . Then Lemma 7.2 and Lemma 5.4 imply that  $\rho(GL(2, \mathbf{Z})) \subseteq \mathcal{DR}_0$ . □

*Remark 7.4.* Note that  $\tau_1(W)$ ,  $\tau(W)$ , and  $\rho(W)$  are not integral matrices.

## 8 Uniqueness of $M_\infty$

Let  $S$  and  $T$  be the generators of  $G = SL(2, \mathbf{Z})$  as given in (4). Let  $V$  denote the standard 2-dimensional  $CG$ -module.

We shall call a  $CG$ -module *T-indecomposable module* if  $T$  acts indecomposably and unipotently on every  $T$ -invariant subspace. One example is the  $CG$ -module, which we shall denote by  $M_\infty$ , defined by the representation  $\tau$  of Theorem 5.1.

**Theorem 8.1.**  *$M_\infty$  is the unique  $T$ -indecomposable  $CG$ -module which has an ascending filtration  $\{M_n\}_{n \in \mathbf{N}}$  in which every quotient  $M_n/M_{n-1}$  is isomorphic to  $V$ .*

Some lemmas are needed for the proof of Theorem 8.1.



**Lemma 8.2.**  $\text{Ext}_{\text{CG}}^1(V, V) \cong \mathbf{C}$ .

*Proof.* Suppose we have a module extension  $M$  of  $V$  by itself and let  $\mu : G \rightarrow \text{GL}(M)$  denote the representation. Since the cyclic group  $\langle ST \rangle$  of order 6 acts semisimply, we may choose a basis of  $M$  such that

$$\mu(ST) = \begin{bmatrix} ST & 0 \\ 0 & ST \end{bmatrix} \quad \text{and} \quad \mu(S) = \begin{bmatrix} S & z(S) \\ 0 & S \end{bmatrix},$$

for some  $2 \times 2$  matrix  $z(S)$ . Since  $\mu(S)^2 = -I$ , we have  $z(S)S + Sz(S) = 0$ , so

$$z(S) = \begin{bmatrix} a & b \\ b & -a \end{bmatrix}$$

for some  $a, b \in \mathbf{C}$ .

By a further change of basis, we can reduce to

$$\mu(S) = \begin{bmatrix} 0 & -1 & a & 0 \\ 1 & 0 & 0 & -a \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

while leaving  $\mu(ST)$  unchanged. Thus,  $\dim \text{Ext}_{\text{CG}}^1(V, V) \leq 1$ . Lastly, if  $a \neq 0$  then  $\mu(T) = -\mu(S)\mu(ST)$  acts indecomposably.

**Lemma 8.3.** *For each natural number  $n$  there is, up to isomorphism, a unique  $T$ -indecomposable CG-module  $M(n)$  of length  $n$  and having all composition factors isomorphic to  $V$ .*

*Proof.* We already have existence of such a module, as a submodule of  $M_\infty$ . We prove by induction that  $\text{Ext}_{\text{CG}}^1(V, M(k)) \cong \mathbf{C}$ . The case  $k = 1$  is Lemma 8.2. We apply  $\text{Hom}_{\text{CG}}(V, -)$  to the short exact sequence

$$0 \rightarrow M(k - 1) \rightarrow M(k) \rightarrow V \rightarrow 0.$$

The long exact sequence of cohomology is:

$$\begin{aligned} 0 \rightarrow \text{Hom}_{\text{CG}}(V, M(k - 1)) \rightarrow \text{Hom}_{\text{CG}}(V, M(k)) \rightarrow \text{Hom}_{\text{CG}}(V, V) \\ \rightarrow \text{Ext}_{\text{CG}}^1(V, M(k - 1)) \rightarrow \text{Ext}_{\text{CG}}^1(V, M(k)) \rightarrow \text{Ext}_{\text{CG}}^1(V, V) \rightarrow \end{aligned}$$

The desired conclusion follows by induction and Lemma 8.2. □

**Lemma 8.4.** *Let  $M, M'$  be isomorphic to  $M(n)$  and let  $N, N'$  be their maximal CG-submodules. Then any CG-isomorphism from  $N$  to  $N'$  can be extended to an isomorphism from  $M$  to  $M'$ .*

*Proof.* We argue by induction on  $n$ , the case  $n = 1$  being trivial. We assume  $n > 1$ . By Lemma 8.3,  $N'$  has, for each  $k \leq n - 1$ , a unique submodule  $N'(k) \cong M(k)$  of length  $k$  and these are all the submodules of  $N'$ . Let  $\psi : N \rightarrow N'$  be a given isomorphism. Choose any isomorphism  $\phi : M \rightarrow M'$ . Replacing  $\phi$  by a scalar multiple, we can assume that  $\alpha := \phi|_N - \psi \in \text{Hom}_{\mathbf{C}G}(N, N')$  is not an isomorphism, so it has a nonzero kernel  $K$ . Hence  $\alpha$  induces an isomorphism  $N/K \rightarrow N'(k)$  for some  $k < n - 1$ . By induction, this isomorphism may be extended to an isomorphism  $\bar{\beta} : M/K \rightarrow N'(k + 1)$ . The induced map  $\beta : M \rightarrow N'(k + 1)$  is an extension of  $\alpha$ . Thus,  $\psi$  extends to  $\phi - \beta$ , which is an isomorphism, since  $N'(k + 1) \subsetneq M'$ .  $\square$

**Proof of Theorem 8.1.** Let  $M_\infty$  and  $M'_\infty$  be modules satisfying the conditions of Theorem 8.1. Then the submodules  $M_n$  and  $M'_n$  in their respective filtrations are isomorphic with  $M(n)$ . By Lemma 8.4, we can define isomorphisms  $\phi_n : M_n \rightarrow M'_n$  recursively for  $n \in \mathbf{N}$  so that  $\phi_{n+1}$  extends  $\phi_n$ . We can therefore define  $\phi : M_\infty \rightarrow M'_\infty$  as follows. Each  $m \in M_\infty$  belongs to  $M_n$  for some  $n$ . By the extension property,  $\phi_n(m)$  does not depend on  $n$ , so we can define a map  $\phi$  by  $\phi(m) = \phi_n(m)$ , which is easily seen to be an isomorphism.  $\square$

## 9 Dirichlet Series in the $SL(2, \mathbf{Z})$ -Orbit of $\zeta(s)$

We may identify  $\mathcal{DS}$  with  $\mathcal{D}\{s\}$  and consider the action of  $SL(2, \mathbf{Z})$  on analytic Dirichlet series via  $\rho$ . We denote the Dirichlet series with one term  $1^{-s}$  simply by 1. We have  $1.\rho(T) = \zeta(s)$ . We set  $\varphi(s) := 1.\rho(-S)$  and write

$$\varphi(s) := \sum_{n=1}^{\infty} a_n n^{-s},$$

where  $a_n = \rho(-S)_{1,n}$ . We denote the abscissae of conditional and absolute convergence of  $\varphi(s)$  by  $\sigma_c$  and  $\sigma_a$ , respectively.

Let  $\mathbf{C}(\zeta(s), \varphi(s))$  be the subfield of the field of meromorphic functions of the half-plane  $\text{Re}(s) > \max(1, \sigma_c)$  generated by the functions  $\zeta(s)$  and  $\varphi(s)$ . It will be shown below that the Dirichlet series in the orbit  $1.\rho(SL(2, \mathbf{Z}))$  all converge in this half-plane and that the analytic functions they define belong to  $\mathbf{C}(\zeta(s), \varphi(s))$ . Let  $\mathbf{Z}G$  denote the integral group ring. The representation  $\rho$  extends uniquely to a ring homomorphism from  $\mathbf{Z}G$  to  $\mathcal{A}$ , which we will denote by  $\rho$  also. The kernel of this homomorphism contains the 2-sided ideal  $Q$  generated by the elements  $S + S^{-1}$  and  $R + R^{-1} - 1$ . Since  $R = ST$ , we have the relation

$$TS = 1 + ST^{-1}$$

in  $\mathbf{Z}G/Q$ . It follows that  $\mathbf{Z}G/Q$  and hence  $\rho(\mathbf{Z}G)$  is generated as an abelian group by the images of the elements  $T^m$  and  $ST^m$ ,  $m \in \mathbf{Z}$ .

**Theorem 9.1.** *The Dirichlet series in the common  $SL(2, \mathbf{Z})$ -orbit of 1,  $\zeta(s)$ , and  $\varphi(s)$  all converge for  $\text{Re}(s) > \max(1, \sigma_c)$ , and belong to the additive subgroup of  $\mathbf{C}(\zeta(s), \varphi(s))$  generated by the elements  $\zeta(s)^m$  and  $\varphi(s)\zeta(s)^m$ ,  $m \in \mathbf{Z}$ .*

*Proof.* We first note that  $\zeta(s)$  has no zeros in the half-plane  $\text{Re}(s) > \max(1, \sigma_c)$  and that  $\frac{1}{\zeta(s)} = \sum_{n=1}^{\infty} \mu(n)n^{-s}$ , converges absolutely there. Here,  $\mu$  is the Möbius function. In this half-plane, we have  $1.\rho(T^m) = 1.D^m = \zeta(s)^m$  and  $1.\rho(ST^m) = 1.\rho(S)\rho(T^m) = -\varphi(s)\zeta(s)^m$ , for every  $m \in \mathbf{Z}$ . The theorem now follows from the discussion preceding it.  $\square$

### 10 The Cubic Equation Relating $\zeta(s)$ and $\varphi(s)$

Let  $\mathbf{N}_0 = \mathbf{N} \cup \{0\}$  be the set of nonnegative integers.

**Lemma 10.1.** *We have*

$$a_n = \rho(-S)_{1,n} = \alpha_1(n) + \sum_{\ell \geq 4} (-1)^\ell \alpha_{\ell-1}(n) \sum_{k=2}^{\lfloor \frac{\ell}{2} \rfloor} b_k \binom{\ell - k - 2}{k - 2}.$$

(See Sect. 4 and formula (16) for the definitions of  $\alpha_k(n)$  and  $b_k$ .)

*Proof.* This is computed directly from the general formula for  $\rho$ :

$$\rho(-S) = Z^{-1} \psi(P \tau_1(-S) P^{-1}) Z.$$

We recall the following information.

- (a) The matrix  $\tau_1(-S)$  is the block-diagonal matrix with the  $2 \times 2$  block  $-S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$  repeated along the main diagonal. (Lemma 5.2)
- (b) The first rows of  $Z^{-1}$  and  $P$  are equal to the first row of the identity matrix. (Lemma 4.5 and Sect. 5.1.)
- (c) For  $A = (a_{i,j})_{i,j \in \mathbf{N}}$ , we have  $\psi(A)_{i,j} = 0$  unless there exist  $k, \ell \in \mathbf{N}$  and an odd number  $d$  such that  $(i, j) = (2^{k-1}d, 2^{\ell-1}d)$ , in which case  $\psi(A)_{i,j} = a_{k,\ell}$ . (Sect. 6.)
- (d) From formula (29), the entries in the second row of  $P^{-1} = (q_{k,\ell})$  are given by  $q_{2,1} = 0, q_{2,2} = 1$  and for  $\ell \geq 3$ ,

$$q_{2,\ell} = (-1)^\ell \sum_{k=2}^{\lfloor \frac{\ell}{2} \rfloor} b_k \binom{\ell - k - 2}{k - 2}. \tag{31}$$

- (e) The entries of the matrix  $Z = (\alpha(i, j))_{i,j \in \mathbf{N}}$  satisfy the equation  $\alpha(2^r, j) = \alpha_r(j)$ , for  $r \in \mathbf{N}$ . (Lemma 4.5.)

By (b), the first row of  $\rho(-S)$  is obtained by multiplying the first row of  $\psi(P \tau_1(-S) P^{-1})$  with  $Z$ . By (c), the only nonzero entries in the first row of  $\psi(P \tau_1(-S) P^{-1})$  are the entries  $\psi(P \tau_1(-S) P^{-1})_{1,2^{\ell-1}} = (P \tau_1(-S) P^{-1})_{1,\ell}$  for  $\ell \in \mathbf{N}$ . Then by (b) and (a),

$$(P \tau_1(-S) P^{-1})_{1,\ell} = (\tau_1(-S) P^{-1})_{1,\ell} = q_{2,\ell}. \tag{32}$$

Hence, by (e) and (d),

$$\begin{aligned} a_n &= \sum_{\ell \in \mathbb{N}} q_{2,\ell} \alpha(2^{\ell-1}, n) \\ &= \alpha_1(n) + \sum_{\ell \geq 3} q_{2,\ell} \alpha_{\ell-1}(n), \end{aligned} \tag{33}$$

and the lemma follows since  $q_{2,3} = 0$ , by (d). □

Let  $\Omega = \{p_1, \dots, p_r\}$  be a finite set primes and let  $t_1, \dots, t_r$  be indeterminates. We will be interested in the formal power series

$$F_{\Omega} = \sum_{(n_1, \dots, n_r) \in \mathbb{N}_0^r} a_{p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}} t_1^{n_1} \dots t_r^{n_r}.$$

Let

$$y = \frac{1}{(1-t_1)(1-t_2)\dots(1-t_r)} - 1 = \sum_{(n_1, \dots, n_r) \in \mathbb{N}_0^r} \alpha_1(p_1^{n_1} \dots p_r^{n_r}) t_1^{n_1} \dots t_r^{n_r}.$$

Then for  $\ell \geq 1$

$$y^{\ell} = \sum_{(n_1, \dots, n_r) \in \mathbb{N}_0^r} \alpha_{\ell}(p_1^{n_1} \dots p_r^{n_r}) t_1^{n_1} \dots t_r^{n_r}.$$

Then we have

$$\begin{aligned} \frac{-y}{1+y} &= \sum_{\ell \in \mathbb{N}} (-1)^{\ell} y^{\ell} \\ &= \sum_{(n_1, \dots, n_r) \in \mathbb{N}_0^r} \left[ \sum_{\ell \in \mathbb{N}} (-1)^{\ell} \alpha_{\ell}(p_1^{n_1} \dots p_r^{n_r}) \right] t_1^{n_1} \dots t_r^{n_r}. \end{aligned} \tag{34}$$

Set

$$\begin{aligned} f_{\Omega} &= \sum_{(n_1, \dots, n_r) \in \mathbb{N}_0^r} \sum_{\ell \in \mathbb{N}} (-1)^{\ell} \alpha_{\ell-1}(p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}) \sum_{k=2}^{\lfloor \frac{\ell}{2} \rfloor} b_k \binom{\ell-k-2}{k-2} t_1^{n_1} \dots t_r^{n_r} \\ &= \sum_{\ell \in \mathbb{N}} \sum_{k=2}^{\lfloor \frac{\ell}{2} \rfloor} b_k (-1)^{\ell} \binom{\ell-k-2}{k-2} y^{\ell-1} \\ &= \sum_{k \geq 2} \left[ \sum_{\ell \geq 2k} (-1)^{\ell} \binom{\ell-k-2}{k-2} y^{\ell-1} \right] b_k. \end{aligned}$$

By Lemma 10.1,

$$F_{\Omega} = y + f_{\Omega}.$$

For  $k \geq 2$  we set

$$C_k = \sum_{\ell \geq 2k} (-1)^{\ell} \binom{\ell - k - 2}{k - 2} y^{\ell - 1}$$

so that

$$f_{\Omega} = \sum_{k \geq 2} b_k C_k.$$

Next we consider, for  $k \in \mathbf{N} \setminus \{1\}$ , the generalized binomial coefficients

$$p_k(x) = \frac{(x - k - 2)(x - k - 3) \cdots (x - 2k + 1)}{(k - 2)!}$$

as polynomials in  $x$  of degree  $k - 2$ .

*Remark 10.2.* Note that  $p_k(\ell) = \binom{\ell - k - 2}{k - 2}$  for  $\ell$  an integer  $\geq 2k$  but, for example, when  $\ell - k - 2$  is a negative integer, the value  $p_k(\ell)$  may be nonzero, while our convention concerning binomial coefficients (Remark 5.3) would say that  $\binom{\ell - k - 2}{k - 2} = 0$ .

In order to find  $C_2$  and  $C_3$ , we shall evaluate

$$\widehat{C}_k = \sum_{\ell \in \mathbf{N}} (-1)^{\ell} p_k(\ell) y^{\ell}.$$

For  $k = 2$ , we have  $p_2(\ell) = 1$ , so  $\widehat{C}_2 = \frac{-y}{1+y}$  by (34). Hence

$$C_2 = \frac{-1}{1+y} - \sum_{\ell=1}^3 (-1)^{\ell} y^{\ell-1} = \frac{-1}{1+y} + 1 - y + y^2 = \frac{y^3}{1+y}. \tag{35}$$

For  $k = 3$ , we have  $p_3(\ell) = \ell - 5$ , so

$$\begin{aligned} \widehat{C}_3 &= \sum_{\ell \in \mathbf{N}} (-1)^{\ell} \ell y^{\ell} - 5 \sum_{\ell \in \mathbf{N}} (-1)^{\ell} y^{\ell} \\ &= \left( \frac{-y}{1+y} + \frac{y^2}{(1+y)^2} \right) + 5 \frac{y}{1+y} \\ &= \frac{4y}{1+y} + \frac{y^2}{(1+y)^2}, \end{aligned} \tag{36}$$

where the second equality is obtained by applying the operator  $y \frac{d}{dy}$  to the first and second members of (34). Therefore,

$$\begin{aligned} C_3 &= \frac{4}{1+y} + \frac{y}{(1+y)^2} - [(-1)p_3(1) + p_3(2)y - p_3(3)y^2 + p_3(4)y^3] \\ &= \frac{4}{1+y} + \frac{y}{(1+y)^2} - 4 + 3y - 2y^2 + y^3 \\ &= \frac{y^5}{(1+y)^2}. \end{aligned} \tag{37}$$

Suppose  $k \geq 3$ . We have

$$\begin{aligned} C_k &= y^{2k-1} + \sum_{\ell \geq 2k+1} (-1)^\ell \binom{\ell-k-2}{k-2} y^{\ell-1} \\ &= y^{2k-1} + \sum_{\ell \geq 2k+1} (-1)^\ell \binom{\ell-1-k-2}{k-2} y^{\ell-1} + \sum_{\ell \geq 2k+1} (-1)^\ell \binom{\ell-1-k-2}{k-3} y^{\ell-1}. \end{aligned}$$

Set

$$A = \sum_{\ell \geq 2k+1} (-1)^\ell \binom{\ell-1-k-2}{k-2} y^{\ell-1}, \quad B = \sum_{\ell \geq 2k+1} (-1)^\ell \binom{\ell-1-k-2}{k-3} y^{\ell-1}.$$

In  $A$ , set  $\ell' = \ell - 1$  and in  $B$ , set  $k' = k - 1$ . Then

$$A = -yC_k, \quad B = y^2C_{k-1} - y^{2k-1}$$

Thus,

$$C_k = y^{2k-1} + A + B = y^{2k-1} - yC_k + y^2C_{k-1} - y^{2k-1} = -yC_k + y^2C_{k-1}.$$

Therefore,

$$C_k = \frac{y^2}{1+y} C_{k-1}, \quad \text{with } C_2 = \frac{y^3}{1+y}$$

so

$$C_k = \frac{y^{2k-1}}{(1+y)^{k-1}}.$$

Hence

$$\frac{C_k C_{k'}}{C_{k+k'}} = \frac{y^{2(k+k')-2}}{(1+y)^{k+k'-2}} \cdot \frac{(1+y)^{k+k'-1}}{y^{2(k+k')-1}} = \frac{1+y}{y}.$$

$$f_\Omega^2 = \left( \sum_{k \geq 2} b_k C_k \right)^2 = \sum_{k, k' \geq 2} b_k b_{k'} C_{k+k'} \frac{(1+y)}{y}$$

Then, from the definition (16) of the  $b_k$ ,

$$\begin{aligned} \frac{y}{1+y} f_\Omega^2 &= \sum_{K \geq 4} \left( \sum_{k=2}^{K-2} b_k b_{K-k} \right) C_K \\ &= \sum_{K \geq 4} (-b_K - 2b_{K-1}) C_K \\ &= - \sum_{K \geq 2} b_K C_K + b_2 C_2 + b_3 C_3 - 2 \frac{y^2}{1+y} \sum_{L \geq 3} b_L C_L \\ &= -f_\Omega - C_2 + 2C_3 - \frac{2y^2}{1+y} f_\Omega - \frac{2y^2}{1+y} C_2 \\ &= - \left( 1 + \frac{2y^2}{1+y} \right) f_\Omega - \frac{y^3}{1+y} + \frac{2y^5}{(1+y)^2} - \frac{2y^2}{1+y} \cdot \frac{y^3}{1+y}. \end{aligned}$$

Therefore, we have

$$y f_\Omega^2 + (1+y+2y^2) f_\Omega + y^3 = 0. \tag{38}$$

Since  $F_\Omega = f_\Omega + y$ , this yields

$$y F_\Omega^2 + (1+y) F_\Omega - y(1+y) = 0. \tag{39}$$

Set

$$P(z, w) = zw^2 + (1+z)w - z(1+z) \tag{40}$$

The discriminant  $\Delta(z)$  is equal to  $(1+z)^2 + 4z^2(1+z)$ . Set  $c = \min\{|e| \mid e \in \mathbf{C} \text{ and } \Delta(e) = 0\}$ . Then there is a formal power series  $u = \sum_{n=0}^\infty \gamma_n z^n$  such that  $P(z, u) = 0$  and  $u$  defines an analytic function in  $\{z \in \mathbf{C} \mid |z| < c\}$ . Now the roots of  $\Delta(z)$  are  $-1$  and  $e, \bar{e} = \frac{-1 \pm \sqrt{-15}}{8}$ . Since  $|e| = \frac{1}{2}$ , it follows that  $u$  converges for  $|z| < \frac{1}{2}$ . Applied to (39), we see that if  $t_i$  take complex values with  $|\prod_{i=1}^r \frac{1}{1-t_i} - 1| < \frac{1}{2}$ , the power series  $F_\Omega$  converges. In particular for  $s \in \mathbf{C}$  with sufficiently large real part, we have convergence when we set the  $t_i = p_i^{-s}$ . If we denote by  $\mathbf{N}_\Omega$  the set of natural numbers for which every prime factor belongs to  $\Omega$ , and define

$$\varphi_\Omega(s) = \sum_{n \in \mathbf{N}_\Omega} a_n n^{-s}, \quad \text{and} \quad \zeta_\Omega(s) = \sum_{n \in \mathbf{N}_\Omega} n^{-s}, \tag{41}$$

we obtain the equation

$$(\zeta_\Omega(s) - 1)\varphi_\Omega(s)^2 + \zeta_\Omega(s)\varphi_\Omega(s) - \zeta_\Omega(s)(\zeta_\Omega(s) - 1) = 0. \tag{42}$$

Initially, we know that this equation holds for  $s$  with sufficiently large real part. The Dirichlet series  $\zeta_\Omega(s)$  and  $\varphi_\Omega(s)$  converge absolutely in the half-plane  $\text{Re}(s) > \max(1, \sigma_a)$ , where both  $\zeta(s)$  and  $\varphi(s)$  converge absolutely. It is then a general property of Dirichlet series that they converge uniformly on compact subsets of this half-plane, defining analytic functions there. Then, by the principle of analytic continuation, the equation (42) holds in this half-plane. If we take  $\Omega$  to be the set of the first  $r$  primes and allow  $r$  to increase, the resulting sequences of analytic functions  $\zeta_\Omega(s)$  and  $\varphi_\Omega(s)$  defined in the above half-plane converge to  $\zeta(s)$  and  $\varphi(s)$ , respectively.

**Theorem 10.3.** *In the half plane  $\text{Re}(s) > \max(1, \sigma_c)$ , we have*

$$(\zeta(s) - 1)\varphi(s)^2 + \zeta(s)\varphi(s) - \zeta(s)(\zeta(s) - 1) = 0. \tag{43}$$

*Proof.* The validity of this algebraic relation for  $\text{Re}(s) > \max(1, \sigma_a)$  is immediate from the foregoing discussion. Since  $\zeta(s)$  and  $\varphi(s)$  represent analytic functions throughout the half-plane  $\text{Re}(s) > \max(1, \sigma_c)$ , the relation is valid on this larger region by the principle of analytic continuation.  $\square$

*Remark 10.4.* Since  $\phi(s)$  defines an analytic function in  $\text{Re}(s) > \sigma_c$ , it follows that  $\zeta(s) - 1$  cannot be equal to any root of  $\Delta(z)$  for  $s$  in this half-plane. By Theorem 11.6 (C) of [9],  $\zeta(s)$  takes on every nonzero value in  $\text{Re}(s) > 1$ . Therefore,  $\sigma_c > 1$ . A sharper bound follows from [3], which proves the existence of a constant  $C \approx 1.764$  such that the closure  $M(\sigma)$  of the set of values of  $-\log \zeta(\sigma + it)$ ,  $t \in \mathbf{R}$ , is bounded by a convex curve when  $\sigma < C$ , and a ring-shaped domain between two convex curves when  $\sigma > C$ . From this, it follows by computation that  $\zeta(s) = \frac{7 \pm \sqrt{-15}}{8}$  for some  $s$  with  $\text{Re}(s)$  arbitrarily close to 1.8, so  $\sigma_c \geq 1.8$ . We also know from the results of [9], p. 300, that  $\zeta(s)$  never takes the value  $\frac{-7 \pm \sqrt{-15}}{8}$  when  $\text{Re}(s) > 1.92$ .

*Remark 10.5.* A slight modification of the discussion above shows that (42) holds for an arbitrary set  $\Omega$  of primes, again for  $\text{Re}(s) > \sigma_a$ .

The function  $\zeta(s)$  can be extended to a meromorphic function in the whole complex plane, whose only singularity is a simple pole at  $s = 1$ . Then (43) defines analytic continuations of  $\varphi(s)$  along arcs in the plane which do not pass through  $s = 1$  or the branch points  $\left\{s \mid \zeta(s) = 0 \text{ or } \frac{7 \pm \sqrt{-15}}{8}\right\}$ , with the exception that one of the two branches at each point  $s$  with  $\zeta(s) = 1$  has a simple pole there. By [3], we know that there is a constant  $C \approx 1.764$  such that  $\zeta(s) \neq 1$  for all  $s$  with  $\text{Re}(s) > C$ .

### 10.1 Some Generalizations

In the discussion following (40), we could equally well have substituted  $t_i = M(p_i)p_i^{-s}$ , where  $M$  is any bounded, completely multiplicative complex



function of the natural numbers, such as a Dirichlet character. In that case, if we set

$$\zeta_M(s) = \sum_{n=1}^{\infty} M(n)n^{-s}, \quad \varphi_M(s) = \sum_{n=1}^{\infty} a_n M(n)n^{-s}, \tag{44}$$

the same reasoning shows that  $\zeta_M(s)$  and  $\varphi_M(s)$  are related by (43), just as  $\zeta(s)$  and  $\varphi(s)$  are, in a suitable half-plane.

We can also extend our discussion to number fields. For this purpose, a necessary remark is that, by Lemma 10.1, the coefficient  $a_n$  in  $\varphi(s)$  depends only on the partition  $\lambda : e_1 \geq e_2 \geq \dots e_r \geq 1$  defined by the exponents  $e_i$  which occur in the prime factorization of  $n$ , in that if  $n$  and  $n'$  define the same partition then  $a_n = a_{n'}$ . We write  $a_\lambda$  for this common value.

Let  $K$  be a number field. The factorization of an ideal  $\mathfrak{g}$  of its ring of integers  $\mathfrak{o}$  into prime ideals determines a partition  $\lambda$ , so we may set  $a_{\mathfrak{g}} = a_\lambda$ . With these notations, our previous discussion up to (40) remains valid if the set  $\Omega$  is taken to be a finite set  $\{\mathfrak{P}_1, \mathfrak{P}_2, \dots, \mathfrak{P}_r\}$  of prime ideals in  $\mathfrak{o}$ , instead of rational primes. Then, in the paragraph following (40), if we substitute  $t_i = N(\mathfrak{P}_i)^{-s}$ , we deduce, as before, that the Dedekind zeta function of  $K$ ,

$$\zeta_K(s) = \sum_{\mathfrak{g}} N(\mathfrak{g})^{-s} \tag{45}$$

is related to the Dirichlet series

$$\varphi_K(s) = \sum_{\mathfrak{g}} a_{\mathfrak{g}} N(\mathfrak{g})^{-s} \tag{46}$$

by the cubic relation (43), in the appropriate half-plane.

### 11 A Functional Equation for $\varphi(s)$

The classical functional equation for  $\zeta(s)$  can be written as

$$\zeta(1-s) = a(s)\zeta(s), \tag{47}$$

where  $a(s) = \frac{\Gamma(s/2)\pi^{-s/2}}{\Gamma((1-s)/2)\pi^{-(1-s)/2}}$ .

If we apply this to (43) with  $s$  replaced by  $(1-s)$  and then eliminate  $\zeta(s)$  from the resulting equation, using (43), a functional equation relating  $\varphi(s)$  and  $\varphi(1-s)$  is obtained. Let

$$\begin{aligned} G(a, x, y) = & a^4 x^4 - a^3 x^2(x^2 + x + 1)(y^2 + y + 1) \\ & + a^2[x^2(y^2 + y + 1)^2 + y^2(x^2 + x + 1)^2 - 2x^2 y^2] \\ & - ay^2(x^2 + x + 1)(y^2 + y + 1) + y^4. \end{aligned} \tag{48}$$

Then  $G(a, x, y)$  is irreducible in  $\mathbf{C}[a, x, y]$  and  $G(a(s), \varphi(s), \varphi(1-s)) = 0$ .

**Acknowledgments** We thank Peter Sarnak for some helpful discussions and for bringing [4] to our attention.

## References

- [1] Barrett, Wayne W. ; Jarvis, Tyler J. Spectral properties of a matrix of Redheffer. Directions in matrix theory (Auburn, AL, 1990). *Linear Algebra Appl.* 162/164 (1992), 673–683.
- [2] F. Bayart, A. Mouze, Factorialité de l’anneau des séries de Dirichlet analytiques, *C. R. Acad. Sci. Paris, Ser. I* 336 (2003).
- [3] H. Bohr and B. Jessen, On the Distribution of the Values of the Riemann Zeta Function, *Am. J. Math.* 58, (1936), No. 1, 35–44.
- [4] Ju. V. Linnik, “The Dispersion Method in Binary Additive Problems”, *Translations of Mathematical Monographs*, Vol. 4, American Mathematical Society, Providence, Rhode Island, (1963).
- [5] A. Ostrowski, Über Dirichletsche Reihen und algebraische Differentialgleichungen, *Math Z.* 8 (1920), 241–298.
- [6] R. Redheffer, Eine explizit lösbare Optimierungsaufgabe. (German) *Numerische Methoden bei Optimierungsaufgaben*, Band 3 (Tagung, Math. Forschungsinst., Oberwolfach, 1976), pp. 213–216. *Internat. Ser. Numer. Math.*, Vol. 36, Birkhäuser, Basel, (1977).
- [7] L. Solomon, The orders of the finite Chevalley groups, *J. Algebra* 3 (1966) 376–393.
- [8] J. G. Thompson, *The Divisor Matrix and  $SL(2, \mathbf{Z})$* , Preprint, University of Florida, (2007).
- [9] E. C. Titchmarsh, “The theory of the Riemann zeta-function”, Second edition. Edited and with a preface by D. R. Heath-Brown. The Clarendon, Oxford University Press, New York, (1986).
- [10] Vaughan, R. C. On the eigenvalues of Redheffer’s matrix. I. Number theory with an emphasis on the Markoff spectrum (Provo, UT, 1991), 283–296, *Lecture Notes in Pure and Appl. Math.*, 147, Dekker, New York, (1993).
- [11] Vaughan, R. C. On the eigenvalues of Redheffer’s matrix. II. *J. Austral. Math. Soc. Ser. A* 60 (1996), no. 2, 260–273.

# Proof of a Conjecture of Alladi Ramakrishnan on Circulants

Michel Waldschmidt

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** In the course of studying a higher dimensional generalization of the Pythagorean equation and its connections to the Lorentz transformation, Alladi Ramakrishnan made a conjecture on a determinant of a certain circulant matrix and published it in his paper *Pythagoras to Lorentz via Fermat*. This conjecture was proved by the author in a letter to Alladi Ramakrishnan. That letter is reproduced here with a note by the Editor explaining the background.

**Subject Classification (2000)** Primary 15A15; Secondary 83A05

**Key words and phrases** Pythagorean equation · Higher dimensional generalization · Lorentz transformation · Circulants · Determinants · Fermat equation

## Introductory Note by Editor

Professor Alladi Ramakrishnan was intellectually active until the very end. Indeed, even in his retirement, he often came back to the Lorentz transformation in Special Relativity and provided new derivations and interpretations. This note concerns a conjecture he made on the value of a determinant of a certain circulant matrix in his paper *Pythagoras to Lorentz via Fermat*. Although Alladi Ramakrishnan made these observations on such circulants prior to 2000, the note [3] was published only in 2003 as part of a collection of his papers on relativity in his book on that subject.

---

M. Waldschmidt  
Université P et M. Curie (Paris VI) Institut de Mathématiques, 175 rue du Chevaleret, F-75013  
Paris, France  
e-mail: [miw@math.jussieu.fr](mailto:miw@math.jussieu.fr)

Ramakrishnan wrote the celebrated Pythagorean equation in the form

$$a^2 - b^2 = c^2 \tag{1}$$

and viewed the left hand side of (1) as the determinant of the  $2 \times 2$  matrix

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix} \tag{2}$$

He thought of the famous Fermat equation as

$$a^n - b^n = c^n.$$

However, as a physicist, he was more interested in equations that had integer solutions, and thus looked at ways to generalize (1) to an equation in  $n$ -dimensions that had integer solutions.

Alladi Ramakrishnan's view of the left hand side of (1) as the determinant of the matrix in (2) led him to a new proof of the Pythagorean theorem on right triangles, and a new interpretation of the Lorentz transformation (see [1]). This also motivated him to consider the  $3 \times 3$  circulant matrix

$$\begin{pmatrix} a & b & c \\ b & c & a \\ c & a & b \end{pmatrix} \tag{3}$$

and its determinant

$$a^3 + b^3 + c^3 - 3abc$$

and led him to a cubic extension of (1), namely

$$a^3 + b^3 + c^3 - 3abc = d^3, \tag{4}$$

and its link with a cubic analogue of the Lorentz transformation he proposed. The generalization of (1) that Alladi Ramakrishnan considered was the  $n \times n$  circulant matrix

$$C_n = \begin{pmatrix} a_1 & a_2 & \cdots & a_{n-1} & a_n \\ a_n & a_1 & \cdots & a_{n-2} & a_{n-1} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ a_2 & a_3 & \cdots & a_n & a_1 \end{pmatrix} \tag{5}$$

and its determinant replacing the left hand side of (1). On the one hand, Ramakrishnan studied its properties in terms of continuous variables  $a_i$ , and on the other hand, he determined [3] integer solutions of

$$|C_n| = b^n, \tag{6}$$

where  $|A|$  denotes the determinant of a matrix  $A$ . In this regard, he conjectured that if  $n = 2m + 1$  is odd, and the elements of  $C_n$  are

$$a_1 = N + m, \quad a_2 = N + m - 1, \quad \dots, \quad a_n = N - m, \tag{7}$$

then

$$|C_{2m+1}| = N(2m + 1)^{2m}. \tag{8}$$

He noted in [1] that when  $n = 2m$ , the evaluation of  $C_n$  is not as elegant, but conjectured that with certain suitable choices one can get

$$|C_{2m}| = N^{2m}. \tag{9}$$

In the letter dated June 8, 2000, Michel Waldschmidt proves these conjectures of Alladi Ramakrishnan and this letter containing the proof is attached.

In 1999, Ramakrishnan [2] showed that circulant matrices with determinant unity transform  $n$  variables in such a way that the determinant of the circulant formed by the variables is invariant, and proposed this as the generalization of the two variable Lorentz invariant to  $n$  variables. In [3], Ramakrishnan observed that his conjectures are equivalent to stating that there are an infinite number of circulants of any dimension with rational elements having determinant unity. Ramakrishnan noted in [3] that the determinant of a circulant is the product of its eigenvalues each of which is a linear combination of  $1, \omega, \omega^2, \dots, \omega^{n-1}$ , where  $\omega$  is a primitive  $n$ -th root of unity. In his fundamental work on  $L$ -Matrices, Ramakrishnan repeatedly used matrices with entries as the  $n$ -th roots of unity. Thus, his familiarity with the properties of such matrices led him to consider (6) as an interesting generalization of the Pythagorean equation.

*Krishnaswami Alladi*

## References

- [1] Alladi Ramakrishnan, “A reflection principle”, *Phys. Educ.*, **30** (1995), 204–205.
- [2] Alladi Ramakrishnan, “Cubic and general extensions of the Lorentz transformation”, *J. Math. Anal. Appl.*, **229** (1999), 88–92.
- [3] Alladi Ramakrishnan, “Pythagoras to Lorentz via Fermat – spanning the interval with light and delight”, in *Special Relativity*, East–West Books, Madras (2003), 90–97.

Paris, June 8, 2000

Dear Professor Alladi Ramakrishnan,

I am pleased to tell you that the conjectures you stated in your paper “Pythagoras to Lorentz” are true.

More precisely, for  $k$  a positive integer, denote by  $C_k(z_1, \dots, z_k)$  the circulant matrix

$$\begin{pmatrix} z_1 & z_2 & \cdots & z_{k-1} & z_k \\ z_k & z_1 & \cdots & z_{k-2} & z_{k-1} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ z_2 & z_3 & \cdots & z_k & z_1 \end{pmatrix}$$

and by  $P_k(z)$  the polynomial

$$\det C_k(z, z - 1, \dots, z - k + 1).$$

Then

$$P_k(z) = k^{k-1} \left( z - \frac{k-1}{2} \right). \tag{1}$$

In particular, if  $k = 2m + 1$  is odd then

$$P_{2m+1}(m + n) = (2m + 1)^{2m} n.$$

Further, for  $k = 2m$  even,

$$\det C_{2m}(n + m, n + m - 1, \dots, n + 1, n - 1, \dots, n - m) = c(m)n, \tag{2}$$

where  $c(m)$  depends only on  $m$ .

Here are the proofs.

The first remark is that if  $A = (a_{ij})_{1 \leq i, j \leq n}$  is a  $n \times n$  square matrix, the polynomial

$$P(z) = \det(z + a_{ij})_{1 \leq i, j \leq n}$$

can be written as,

$$P(z) = cz + \det(A) \tag{3}$$

with a constant  $c$ . This is easily checked by replacing each row but the first one by its difference with the first one, and then expanding with minors on the first row.<sup>1</sup>

Next for  $k = 2m$ , consider the circulant

$$C_{2m}(m, m - 1, \dots, 1, -1, \dots, -m + 1, -m).$$

---

<sup>1</sup> As pointed out by C. Levesque in March 2009, subtracting each column (starting from the second one) from the first one yields the coefficient  $k^{k-1}$ .

The sum of all rows (as well as the sum of all columns) is 0. Hence

$$\det C_{2m}(m, m - 1, \dots, 1, -1, \dots, -m + 1, -m) = 0. \tag{4}$$

It is plain that (3) and (4) imply (2). They also imply

$$P_k(z) = c_k \left( z - \frac{k - 1}{2} \right), \tag{5}$$

with some constant  $c_k$  depending only on  $k$ , but we are going to reprove this result (and compute  $c_k$ ) by another way.

It is well-known (and easy to prove) that

$$\det C_k(z_1, \dots, z_k) = \prod_{\zeta} (z_1 + \zeta z_2 + \dots + \zeta^{k-1} z_k) = \prod_{\zeta} \sum_{i=0}^{k-1} \zeta^i z_{i+1},$$

where  $\zeta$  ranges over the  $k$ -th roots of unity. Hence

$$P_k(z) = \prod_{\zeta} \sum_{i=0}^{k-1} \zeta^i (z - i).$$

Now

$$\sum_{i=0}^{k-1} \zeta^i = \begin{cases} k & \text{for } \zeta = 1, \\ 0 & \text{for } \zeta \neq 1, \end{cases}$$

and we derive (5) with

$$c_k = k \prod_{\zeta \neq 1} \sum_{i=0}^{k-1} (-i) \zeta^i = (-1)^{k-1} k \prod_{\zeta \neq 1} \sum_{i=0}^{k-1} i \zeta^i.$$

The sum

$$\sum_{i=0}^{k-1} i \zeta^i = \zeta + 2\zeta^2 + \dots + (k - 1)\zeta^{k-1}$$

is the value at the point  $\zeta$  of  $zf'(z)$ , where  $f'$  is the derivative of the polynomial

$$f(z) = 1 + z + \dots + z^{k-1} = \frac{z^k - 1}{z - 1}.$$

Since

$$f'(z) = \frac{kz^{k-1}}{z - 1} - \frac{z^k - 1}{(z - 1)^2},$$

for  $\zeta$  satisfying  $\zeta^k = 1$  and  $\zeta \neq 1$  we have

$$\zeta f'(\zeta) = \frac{k}{\zeta - 1}.$$

Now

$$\prod_{\zeta \neq 1} (\zeta - 1)$$

is nothing else than the resultant of the two polynomials  $z - 1$  and  $f(z)$ , hence

$$\prod_{\zeta \neq 1} (\zeta - 1) = (-1)^{k-1} f(1) = (-1)^{k-1} k.$$

Therefore,

$$\prod_{\zeta \neq 1} \sum_{i=0}^{k-1} i \zeta^i = \prod_{\zeta \neq 1} \frac{k}{\zeta - 1} = \frac{k^{k-1}}{(-1)^{k-1} f(1)} = (-1)^{k-1} k^{k-2} \quad \text{and} \quad c_k = k^{k-1}.$$

This completes the proof of (1).

*Michel Waldschmidt*



**Part III**  
**Probability and Statistics**

# Branching Random Walks

**K.B. Athreya**

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** A branching random walk is a branching tree such that with each line of descent a random walk is associated. This paper provides some results on the asymptotics of the point processes generated by the positions of the  $n$ th generation individuals. An application to the photon–electron energy cascade is also given.

**Mathematics Subject Classification (2000)** 60J80

**Key words and phrases** Branching processes · Random walks

## 1 Introduction

The author would like to thank Professor Alladi Krishnaswami for the invitation to contribute a paper to this volume dedicated to the memory of his father Professor Alladi Ramakrishnan. Professor Ramakrishnan was one of the pioneers with Professors Bhabha and Heitler to work on nuclear cascades. The stochastic model they used is an example of a branching random walk, the subject of the present paper. On a personal note, I learnt a few years ago from Professor Ramakrishnan that he was a student of my father Shri T. A. Balasundaram Iyer at the well-known P. S. High School in Mylapore, Madras (Chennai, now), India during the thirties. Also, Professor Ramakrishnan and I used to meet often at the Madras Music Festival at the Music Academy in Chennai during the second half of December. We shared a deep interest in carnatic music, the classical music of south India.

---

K.B. Athreya

Department of Mathematics, Iowa State University, Ames, IA 50011, USA

Indian Institute of Science, Bangalore 560012, India

e-mail: [kba@iastate.edu](mailto:kba@iastate.edu); [kbathreya@gmail.com](mailto:kbathreya@gmail.com)

## 2 Branching Random Walks

Let  $\{p_j\}_{j \geq 0}$  be a probability distribution, i.e.,  $p_j \geq 0, \forall j \geq 0$  and  $\sum_{j=0}^{\infty} p_j = 1$ .

Let  $\{\xi_{ni} : 1 \leq i < \infty, 0 \leq n < \infty\}$  be a family of independent and identically distributed random variables (i.i.d. r.v.s) with distribution  $P(\xi_{11} = j) = p_j, j \geq 0$ . Let  $Z_0$  be a negative integer valued random variable. Let

$$Z_{n+1} = \sum_{i=1}^{Z_n} \xi_{ni}, \quad n \geq 0. \tag{2.1}$$

This sequence  $\{Z_n\}_{n \geq 0}$  is called a *Galton–Watson branching process with offspring distribution*  $\{p_j\}_{j \geq 0}$ . Since the family  $\{\xi_{ni} : 1 \leq i < \infty, 0 \leq n < \infty\}$  is i.i.d.,  $\{Z_n\}_{n \geq 0}$  is a Markov chain. Its state space is  $\mathbb{N}^+ \equiv \{0, 1, 2, \dots\}$  and transition function is

$$p_{ij} = P\left(\sum_{r=1}^i \xi_{1r} = j\right) \quad \text{for all } i \geq 0, j \geq 0.$$

The recurrence relation (2.1) is to be interpreted as follows. All the  $Z_n$  individuals in the  $n$ th generation produce offspring independently of each other, the  $i$ th one producing  $\xi_{ni}$  offspring,  $1 \leq i \leq Z_n$ , and their total is  $Z_{n+1}$ , the size of the  $(n + 1)$ th generation. If  $Z_0 = 1$ , then there is a unique probability measure on the family tree initiated by this ancestor. If  $Z_0 = k > 1$ , then each of these  $k$  ancestors initiates a family tree and these are i.i.d. random trees.

Now, on this family tree impose the following movement structure. If an individual is located at  $x$  in  $\mathbb{R}$ , the real line and produces  $k$  children, then these  $k$  children move to  $x + X_{kj}, 1 \leq j \leq k$  where for each  $k, (X_{k1}, X_{k2}, \dots, X_{kk})$  is a random vector with a joint distribution  $\pi_k$  on  $\mathbb{R}^k$ . The random vector  $(X_{k1}, X_{k2}, \dots, X_{kk})$  is stochastically independent of the history up to that generation as well as the movement of the offspring of other individuals.

Let  $\zeta_n \equiv \{x_{ni} : 1 \leq i \leq Z_n\}$  be the positions of the  $Z_n$  individuals of the  $n$ th generation. For each  $n \geq 0, \zeta_n$  is a collection of random numbers of random points on  $\mathbb{R}$ , i.e., a *point process*. The sequence of pairs of  $\{Z_n, \zeta_n\}_{n \geq 0}$  is called *branching random walk*. The probability distribution of this process is completely specified by

- (1) The offspring distribution  $\{p_j\}_{j \geq 0}$
- (2) The family of probability measures  $\{\pi_k\}_{k \geq 1}$
- (3) The initial population size  $Z_0$ , and
- (4) The locations  $\zeta_0 \equiv \{x_{0i}, 1 \leq i \leq Z_0\}$  of the initial ancestors

It is clear that  $\{\zeta_n\}_{n \geq 0}$  is also a Markov chain whose state space is the set of all finite subsets of  $\mathbb{R}$ .

The problem of interest is what happens to the point process  $\zeta_n$  as  $n \rightarrow \infty$ . In particular, we could ask what happens to the spatial distribution of points of  $\zeta_n$ . That

is, if  $Z_n(x)$  is the number of points in  $\zeta_n$  that are less than or equal to  $x$ , how does  $Z_n(x)$  behave as  $n \rightarrow \infty$ ? Does there exist  $x_n$  such that the proportion  $\frac{Z_n(x_n)}{Z_n}$  has a nontrivial limit as  $n \rightarrow \infty$ ?

The sequence  $\{Z_n\}_{n \geq 0}$  has been well-studied. (See Harris [6], Athreya and Ney [1], Jagers [7, 8], Mode [9].) A summary of the main results is given in Sect. 3.

It is clear that the movement along any one line of descent is that of a classical random walk. Thus, if  $X_{ki}$  are identically distributed with mean  $\mu$  and finite variance  $\sigma^2$ , then the location of an individual of the  $n$ th generation should be approximately Gaussian (by the central limit theorem. See Feller [5], Athreya and Lahiri [4].) with mean  $n\mu$  and variance  $n\sigma^2$ . This suggests that if  $Z_n \rightarrow \infty$  as  $n \rightarrow \infty$  and if  $x_n = \sigma\sqrt{n}x + n\mu$ , then  $\frac{Z_n(x_n)}{Z_n}$  could have  $\Phi(x)$ , the standard  $N(0, 1)$  c.d.f., as its limit. This turns out to be true and is the main result of this paper. There are extensions to the case where  $X_{11}$  has infinite variance. The proof of our main result needs the following result from branching processes. It says that if two individuals are chosen at random from the  $n$ th generation and if  $\tau_n$  denotes the generation number of their last common ancestor, then in a growing population  $\tau_n$  converges to a proper random variable in distribution. That is, the last common ancestor should have been born way at the beginning of the tree.

In the next section, we review some basic results from branching processes. The fourth section has the main result and the result on branching processes mentioned above. The fifth section treats the photon–electron energy cascade studied by Bhabha, Heitler, and Ramakrishnan. The final section outlines some open questions.

### 3 Results on Branching Processes

For the sequence  $\{Z_n\}_{n \geq 0}$ , the following two results are well-known. (See Athreya and Ney [1].)

**Theorem 3.1.** *Let  $0 < m = \sum_{j=1}^{\infty} jp_j < \infty$ . Let  $P(Z_0 < \infty) = 1$ . Then*

(i)  $m \leq 1 \Rightarrow P(Z_n \rightarrow 0 \text{ as } n \rightarrow \infty) = 1$

(ii)  $m > 1 \Rightarrow P(Z_n \rightarrow 0 \text{ as } n \rightarrow \infty | Z_0 = 1) = q < 1$ , where  $q$  is the unique root of  $s = f(s)$ ,  $0 \leq s < 1$ ,  $f(s) = \sum_{j=1}^{\infty} p_j s^j$ , and

$$P(Z_n \rightarrow \infty \text{ as } n \rightarrow \infty | Z_0 = 1) = 1 - q,$$

and for any  $k > 1$ ,

$$P(Z_n \rightarrow 0 \text{ as } n \rightarrow \infty | Z_0 = k) = q^k.$$

*Remark 1.* The branching process  $\{Z_n\}_{n \geq 0}$  is called *subcritical*, *critical*, or *supercritical* according as  $m <$ ,  $=$  or  $> 1$ .

**Theorem 3.2.** (i) Let  $0 < m < 1$ . Let  $P(Z_0 < \infty) = 1$ . Then, as  $n \rightarrow \infty, \forall j \geq 1$ ,

$$P(Z_n = j | Z_n > 0) \rightarrow b_j, \quad \sum_{j=1}^{\infty} b_j = 1$$

and  $B(s) = \sum_{j=1}^{\infty} b_j s^j, 0 \leq s \leq 1$ , is the unique solution of the functional equation

$$B(s) = mB(s) + (1 - m), \quad 0 \leq s \leq 1$$

such that  $B(0) = 0$  and  $B(1) = 1$ .

(ii) Let  $m = 1, \sum_{j=1}^{\infty} j^2 p_j < \infty$ . Then, as  $n \rightarrow \infty, \forall 0 < x < \infty$ ,

$$P\left(0 < \frac{Z_n}{n} < x \mid Z_n > 0\right) \rightarrow 1 - e^{-\frac{2x}{\sigma^2}},$$

where  $\sigma^2 = \sum_{j=1}^{\infty} j^2 p_j - 1$ .

(iii) Let  $m > 1$ . Then  $\left\{W_n \equiv \frac{Z_n}{m^n}\right\}_{n \geq 0}$  is a nonnegative martingale and  $\lim_{n \rightarrow \infty} W_n = W$  exists w.p.1. Further,

$$\sum_{j=1}^{\infty} j \log j p_j < \infty \Rightarrow E(W | Z_0 = 1) = 1 \text{ and } P(W = 0 | Z_0 = 1) = q$$

$$\sum_{j=1}^{\infty} j \log j p_j = \infty \Rightarrow P(W = 0 | Z_0 = 1) = 1.$$

**Theorem 3.3.** Let  $1 < m \equiv \sum_{j=1}^{\infty} j p_j < \infty, p_0 = 0$  and  $\sum_{j=1}^{\infty} j(\log j) p_j < \infty$ . Pick two individuals from the  $n$ th generation. Let  $\tau_n$  be the generation number of their last common ancestor. Then, for  $j \geq 0$ ,

$$\lim_{n \rightarrow \infty} P(\tau_n = j) = b_j \quad \text{exists}$$

and

$$\sum_{j=1}^{\infty} b_j = 1.$$

That is,  $\tau_n$  converges in distribution to a proper random variable  $\tau$ .

*Proof.* A more general result is available in Athreya [2]. For completeness, a proof is given here: For  $1 \leq j < \infty$ ,

$$P(\tau_n < j) = E \left( \frac{\sum_{i_1 \neq i_2} Z_{n-j,i_1}^{(j)} Z_{n-j,i_2}^{(j)}}{Z_n(Z_n - 1)} \right), \tag{3.1}$$

where  $\{Z_{k,i}^{(j)} : k \geq 0\}$  is the branching process initiated by the  $i$ th individual in the  $j$ th generation and the summation is over  $1 \leq i_1, i_2 \leq Z_j$ . By Theorem 3.2, (iii), there exists i.i.d. r.v.s  $\{W_i\}_{i \geq 1}$  such that

$$\lim_{n \rightarrow \infty} \frac{Z_{n-j,i}^{(j)}}{m^{n-j}} \equiv W_i \quad \text{exists w.p.1}$$

and

$$\lim_{n \rightarrow \infty} \frac{Z_n}{m^n} = \frac{1}{m^j} \sum_{i=1}^{Z_j} W_i.$$

Thus, the sum on the right side of (3.1) converges w.p.1 to

$$\frac{\sum_{i_1 \neq i_2, 1 \leq i_1 < i_2 \leq Z_j} W_{i_1} W_{i_2}}{\left( \sum_{i=1}^{Z_j} W_i \right)^2}.$$

By the bounded convergence theorem, it follows that

$$\lim_{n \rightarrow \infty} P(\tau_n < j) \equiv E\phi(Z_j)$$

exists for each  $j \geq 1$  where

$$\phi(k) \equiv E \left( 1 - \frac{\sum_{i=1}^k W_i^2}{\left( \sum_{i=1}^k W_i \right)^2} \right),$$

where  $\{W_i\}_{i \geq 1}$  are i.i.d. distributed as

$$W \equiv \lim_{n \rightarrow \infty} \frac{Z_n}{m^n} \quad \text{with } Z_0 = 1.$$

Now we show that  $\lim_{k \rightarrow \infty} \phi(k) = 0$ .

By Theorem 3.2 (iii), under the hypothesis  $\sum_{j=1}^{\infty} j(\log j)p_j < \infty$ ,  $EW < \infty$ . This implies

$$\lim_{x \rightarrow \infty} xP(W > x) = 0.$$

Now, if

$$M_n = \max_{1 \leq i \leq n} W_i,$$

then

$$\begin{aligned} P(M_n \leq n\epsilon) &= (P(W_1 \leq n\epsilon))^n \\ &= (1 - P(W_1 > n\epsilon))^n \\ &= \left(1 - \frac{nP(W_1 > n\epsilon)}{n}\right)^n. \end{aligned}$$

Since  $\forall \epsilon > 0$ ,  $nP(W_1 > n\epsilon) \rightarrow 0$  as  $n \rightarrow \infty$ ,

$$P(M_n \leq n\epsilon) \rightarrow 1.$$

Also by the strong law of large numbers

$$\frac{\sum_{i=1}^n W_i}{n} \rightarrow 1 \quad \text{w.p.1.}$$

Now,

$$\frac{\sum_{i=1}^k W_i^2}{\left(\sum_{i=1}^k W_i\right)^2} \leq \frac{M_k}{S_k} \rightarrow 0 \quad \text{in probability}$$

yielding  $\lim_{k \rightarrow \infty} \phi(k) = 0$ . □

## 4 Branching Random Walks

Turning now to the branching random walk sequence  $\{\zeta_n\}_{n \geq 0}$ , it is clear from the above two results that in the critical and subcritical cases, the population dies out and hence  $\zeta_n$  becomes the empty set for large  $n$  with probability one. So the case of interest is primarily the supercritical case although we could consider the conditional distribution of  $\zeta_n$  conditioned on the event  $\{Z_n > 0\}$ .

In the supercritical case, if we assume  $p_0 = 0$  then  $q = 0$  and so  $Z_n \rightarrow \infty$  w.p.1. Now if we look at any one line of descent and the position of the individuals in that line of descent, that forms an ordinary random walk with jump distribution  $F$ . This suggests that if  $F$  has mean zero and finite variance, then the position of an individual in the  $n$ th generation should be approximately Gaussian. A result confirming to this is the following.

**Theorem 4.1.** For each  $x \in \mathbb{R}$ , let  $Z_n(x) \equiv \#\{i : x_{ni} \leq x\}$  when  $\zeta_n = \{x_{ni} : 1 \leq i \leq Z_n\}$ . Assume

- (i)  $p_0 = 0$
- (ii)  $1 < m = \sum_{j=1}^{\infty} jp_j < \infty$
- (iii)  $\mu \equiv \int_{\mathbb{R}} x dF(x) = 0$ , and
- (iv)  $\sigma^2 = \int_{\mathbb{R}} x^2 dF(x) < \infty$

Let  $Y_n$  be the position of an individual chosen at random (by simple random sampling) from the  $n$ th generation. Then, for any  $y \in \mathbb{R}$ , as  $n \rightarrow \infty$ ,

(i)

$$\frac{Z_n(\sqrt{n}\sigma y)}{Z_n} \rightarrow \Phi(y) \tag{4.1}$$

in mean square where  $\Phi(\cdot)$  is the standard normal distribution function given by

$$\Phi(y) \equiv \int_{-\infty}^y \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx, \quad -\infty < y < \infty.$$

(ii)

$$P(Y_n \leq \sqrt{n}\sigma y) \rightarrow \Phi(y) \quad \text{as } n \rightarrow \infty. \tag{4.2}$$

*Proof.* Let  $\zeta_n \equiv \{x_{ni} : 1 \leq i \leq Z_n\}$  be the position of the  $Z_n$  individuals of the  $n$ th generation. Fix  $-\infty < y < \infty$ . Let

$$\delta_{ni} = \begin{cases} 1, & \text{if } x_{ni} \leq y\sigma\sqrt{n}, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$Z_n(y\sigma\sqrt{n}) = \sum_{i=1}^{Z_n} \delta_{ni}$$



and

$$E\left(\frac{Z_n(y\sigma\sqrt{n})}{Z_n}\right) = E(\delta_{n1}) = P(x_{n1} \leq y\sigma\sqrt{n}) = P(S_n \leq y\sigma\sqrt{n} - x_{01}),$$

where  $S_n = \sum_{i=1}^n \eta_i$ ,  $\{\eta_i\}_{i \geq 1}$  are i.i.d. with distribution  $\pi$ , and  $x_{01}$  is the location of the initial ancestor of the  $n$ th generation individual located at  $x_{ni}$ . By the central limit theorem, for  $-\infty < y < \infty$ ,

$$P(S_n \leq y\sigma\sqrt{n} - x_{01}) \rightarrow \Phi(y) \quad \text{as } n \rightarrow \infty.$$

Next,

$$E\left(\frac{Z_n(y\sigma\sqrt{n})}{Z_n}\right)^2 = E\left(\frac{1}{Z_n^2} \sum_{i,j=1, i \neq j}^{Z_n} \delta_{ni}\delta_{nj}\right) + E\left(\sum_{i=1}^{Z_n} \delta_{ni} \frac{1}{Z_n}\right).$$

Since  $0 \leq \frac{1}{Z_n^2} \sum_{i=1}^{Z_n} \delta_{ni} \leq \frac{1}{Z_n}$  and  $Z_n \rightarrow \infty$  w.p.1, then second term above goes to zero. By symmetry considerations conditioned on the branching tree (but not the random walk),

$$E\left(\frac{1}{Z_n^2} \sum_{i,j=1, i \neq j}^{Z_n} \delta_{ni}\delta_{nj}\right) = E\left(\frac{Z_n(Z_n - 1)}{Z_n^2} \delta_{n1}\delta_{n2}\right).$$

Let  $\tau_n$  be the generation number of the last common ancestor of the individuals  $I_{n1}$  and  $I_{n2}$  corresponding to  $x_{n1}$  and  $x_{n2}$ , respectively.

Let  $x_{\tau_n}$  be the location of this common ancestor in the  $\tau_n$ th generation. Then we can write

$$x_{ni} = x_{\tau_n} + Y_{ni}, \quad i = 1, 2,$$

where  $Y_{ni}$  is the net displacement of the individual  $I_{ni}$  from generation  $\tau_n$  to  $n$ .

Clearly,  $Y_{n1}$  and  $Y_{n2}$  are independent. Thus,

$$\begin{aligned} E(\delta_{n1}\delta_{n2} | \tau_n, x_{\tau_n}) &= E(I(Y_{n1} \leq y\sigma\sqrt{n} - x_{\tau_n})I(Y_{n2} \leq y\sigma\sqrt{n} - x_{\tau_n}) | \tau_n, x_{\tau_n}) \\ &= E(P(S_{n-\tau_n} \leq y\sigma\sqrt{n} - x_{\tau_n} | \tau_n, x_{\tau_n}))^2, \end{aligned}$$

where  $\{S_j\}_{j \geq 0}$  is a random walk independent of  $\tau_n$  and  $x_{\tau_n}$  with step distribution  $\pi_1$ .

By Theorem 3.3,  $\tau_n$  converges in distribution to a proper random variable and hence so does  $x_{\tau_n}$ .

By the central limit theorem

$$P(S_{n-\tau_n} \leq y\sigma\sqrt{n} - x_{\tau_n} | \tau_n, x_{\tau_n}) \rightarrow \Phi(y) \quad \text{as } n \rightarrow \infty.$$

Now by the bounded convergence theorem

$$E(E(\delta_{n1}\delta_{n2} | \tau_n, x_{\tau_n})) \rightarrow (\Phi(y))^2.$$

Since  $Z_n \rightarrow \infty$ ,  $\frac{Z_n^2}{Z_n(Z_n - 1)} \rightarrow 1$  w.p.1. Thus,

$$E\left(\frac{1}{Z_n^2} \sum_{i,j=1, i \neq j}^{Z_n} \delta_{ni}\delta_{nj}\right) \rightarrow (\Phi(y))^2 \quad \text{as } n \rightarrow \infty.$$

This in turn yields

$$E\left(\frac{Z_n(x_n)}{Z_n}\right)^2 \rightarrow (\Phi(y))^2 \quad \text{as } n \rightarrow \infty.$$

Since  $E\left(\frac{Z_n(x_n)}{Z_n}\right) \rightarrow \Phi(y)$  as  $n \rightarrow \infty$ , we may conclude that

$$E\left(\frac{Z_n(x_n)}{Z_n} - \Phi(y)\right)^2 \rightarrow 0$$

proving assertion (4.1).

Next, since

$$P(Y_n \leq y\sigma\sqrt{n}) = E\left(\frac{Z_n(x_n)}{Z_n}\right),$$

assertion (4.2) follows. □

## 5 Energy Cascades

Consider an elementary particle undergoing fission. Suppose we start with one particle with energy  $E_0$ . Suppose

- (i) It splits into a random number  $\xi$  of new particles with probability distribution  $P(\xi = j) = p_j, j \geq 0$
- (ii) If  $\xi = k, k \geq 1$ , the energy  $E_0$  of the parent particle is distributed among the  $k$  offspring particles as  $E_0 Y_{k1}, E_0 Y_{k2}, \dots, E_0 Y_{kk}$ , where  $(Y_{k1}, Y_{k2}, \dots, Y_{kk})$  has

a probability distribution  $\pi_k$  over the simplex  $\left\{ \vec{p} = (p_1, p_2, \dots, p_k) : p_i \geq 0, \right.$

$\left. \sum_{i=1}^k p_i = 1 \right\}$  such that  $\pi_k$  is unchanged under permutation

(iii)  $\forall k \geq 1, x_{k1}$  has a distribution independent of  $k$ , and

(iv) If  $\xi = 0$ , the fission stops

After  $n$  generations, let the energies of the  $Z_n$  particles in the  $n$ th generation be  $e_{n1}, e_{n2}, \dots, e_{nZ_n}$ , respectively. It is clear from (ii) that if  $x_{01} = \log E_0$  and  $x_{ni} = \log e_{ni}, i \geq 1, n \geq 0$  and  $\zeta_n = (x_{n1}, x_{n2}, \dots, x_{nZ_n})$ , then  $(Z_n, \zeta_n)_{n \geq 0}$  is a branching random walk as described in Sect. 4. If

$$Z_n(x) = \#\{i : e_{ni} \leq x, 1 \leq i \leq Z_n\}$$

is the number of particles with energy less than or equal to  $x$  after  $n$  splits have occurred, then by Theorem 4.1 the following holds.

**Theorem 5.1.** *Let the family of distributions  $\{\pi_k\}_{k \geq 1}$  satisfies*

- (i)  $\forall k \geq 1, \pi_k$  is unchanged under permutation, i.e.,  $P((Y_{k1}, Y_{k2}, \dots, Y_{kk}) \in A_1 \times A_2 \times \dots \times A_k) = P((Y_{ki_1}, Y_{ki_2}, \dots, Y_{ki_k}) \in A_1 \times A_2 \times \dots \times A_k)$  for every permutation  $(i_1, i_2, \dots, i_k)$  of  $(1, 2, \dots, k)$
- (ii)  $\pi_k(Y_{k1} \in A) \equiv \mu(A), A \in \mathfrak{B}(\mathbb{R})$ , is independent of  $k$
- (iii)  $E \log Y_{11} = \mu, V(\log Y_{11}) = \sigma^2 < \infty$ , and
- (iv)  $p_0 = 0, 1 < m = \sum_{j=1}^{\infty} j p_j < \infty$  and  $\sum_{j=1}^{\infty} j(\log j) p_j < \infty$

Let  $Z_n(x_n) = \#\{i : x_{ni} \leq x_n\}$  be the number of elementary particles of the  $n$ th generation with energy  $e_{ni}$  such that  $x_{ni} = \log e_{ni} \leq n\mu + \sqrt{n}\sigma x$ . Then, for  $\forall -\infty < x < \infty$ ,

$$E \left( \frac{Z_n(x_n)}{Z_n} - \Phi(x) \right)^2 \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{u^2}{2}} du, \quad -\infty < x < \infty$$

is the standard  $N(0, 1)$  c.d.f. and hence

$$\frac{Z_n(x_n)}{Z_n} \rightarrow \Phi(x) \quad \text{in probability as } n \rightarrow \infty$$

and

$$\frac{Z_n(x_n)}{m^n} \rightarrow \Phi(x)W \quad \text{in probability} \quad \text{as } n \rightarrow \infty,$$

where  $W = \lim_{n \rightarrow \infty} \frac{Z_n}{m^n}$  as in Theorem 3.2.

## 6 Extensions and Open Problems

The model considered in Sect. 2 can be extended in many directions. We describe a few of them below.

### 6.1 Non-Gaussian Limits

Suppose the displacement random variables  $X_{11}$  has no finite second moment. Assume  $X_{11}$  is in the domain of attraction of a stable law of order  $\alpha$  with  $0 < \alpha \leq 2$ . Then there exist  $a_n$  and  $b_n$  such that

$$P(S_n \leq a_n + b_n x) \rightarrow F_\alpha(x), \quad -\infty < x < \infty$$

as  $n \rightarrow \infty$ , where  $F_\alpha(\cdot)$  is the c.d.f. of a stable law of order  $\alpha$ . Now we can adapt the argument of Theorem 4.1 to conclude that

(i)

$$\frac{Z_n(a_n + b_n x)}{Z_n} \rightarrow F_\alpha(x), \quad -\infty < x < \infty$$

in mean square and hence in probability.

(ii)

$$\frac{Y_n - a_n}{b_n} \xrightarrow{d} G_\alpha, \quad \text{as } n \rightarrow \infty,$$

where  $Y_n$  is the position of a randomly chosen individual in the  $n$ th generation.

### 6.2 Continuous Time

Suppose now that each individual lives a random length of time with distribution  $G(\cdot)$ . Let  $Z(x, t)$  be the number of individuals alive at time  $t$  with positions less than or equal to  $x$ .

It is known (Athreya [3]) that the generation number  $N_t$  of a randomly chosen individual grows like  $t$  and

$$\frac{N_t}{t} \rightarrow \frac{1}{\mu_\alpha},$$

where  $\mu_\alpha = m \int_0^\infty x e^{-\alpha x} dG(x)$  and  $0 < \alpha < \infty$  is the Malthusian parameter defined by  $m \int_0^\infty e^{-\alpha x} dG(x) = 1$ .

A randomly chosen individual's position is distributed as  $S_{N_t}$  where  $\{S_n\}_{n \geq 0}$  is a random walk with jump distribution and  $N_t$  is a random variable independent of  $\{S_n\}_{n \geq 0}$ .

An interesting conjecture is that if the displacement has finite second moment then for some constant  $\sigma$

$$\frac{Z_t \left( \frac{t}{\mu_\alpha} \mu + \sqrt{\frac{t}{\mu_\alpha}} x \sigma \right)}{Z_t}$$

converges in mean square to  $\Phi(x)$ .

### 6.3 Critical Case

If  $m = 1$ , then the population dies out in finite time w.p.1. That is, w.p.1  $Z_n = 0$  for some large  $n$ . (Theorem 3.2 (ii)) But under finite second moments conditioned on nonextinction, i.e.,  $\{Z_n > 0\}$ ,  $Z_n$  is of the order  $n$  and hence on this event  $\frac{Z_n(x_n)}{Z_n}$  should converge in probability to some limit if  $x_n \rightarrow \infty$  at an appropriate rate. This needs to be established. A difficulty in this case is that the random time  $\tau_n$ , the generation number of the last common ancestor of two randomly chosen individuals from those alive in the  $n$ th generation may grow like  $n$ . An extension of this to the continuous case is also an interesting problem.

### 6.4 Multitype Case

The extension of Theorem 4.1 to the case when the underlying branching process is a multitype one with displacement distribution dependent on the type is also an interesting question. Here the supercritical and critical cases, discrete and continuous time cases are all open.

## References

- [1] Athreya, K. B. and Ney, P. (2000) *Branching Processes*, Dover, Minolta, NY.
- [2] Athreya, K. B. (2006) On the last common ancestor problem in branching processes, unpublished.
- [3] Athreya, K. B. (2009) Applications of size biasing in branching processes, *Int. Symp. on Probability Theory & Stochastic processes*, Cochin University, India, Feb 2009.
- [4] Athreya, K. B. and Lahiri, B. P. (2006) *Measure Theory and Probability Theory*, Springer, NY.
- [5] Feller, W. (1968) *An Introduction of Probability Theory and Its Applications*, Wiley, NY.
- [6] Harris, T. E. (1963) *The Theory of Branching Processes*, Prentice-Hall, NJ.
- [7] Athreya, K. B. and Jagers, P. (1997) *Classical and Modern Branching Processes*, Springer, NY.
- [8] Jagers, P. (1975) *Branching Processes with Biological Applications*, Wiley, London.
- [9] Mode, C. (1971) *Multi-type Branching Processes – Theory and Applications*, Elsevier, NY.

# A Commentary on the Logistic Distribution

Malay Ghosh, Kwok Pui Choi, and Jialiang Li

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** The paper provides a series representation of the logistic probability density function in terms of differently scaled double exponential distributions with terms of the series alternating in signs. This representation is used to calculate moments, moment generating function, and characteristic function of a logistic distribution. The same representation is also used to derive the logistic distribution as the scale mixture of a normal distribution.

**Mathematics Subject Classification (2000)** 62E15

**Key words and phrases** Double exponential · Generating functions · Moments · Normal · Scale mixture · Series representation

## 1 Introduction

The logistic distribution occupies a prominent role in the theory and practice of statistics. One of the earliest applications was in biology to describe how species populations grow in competition [18]. This distribution is also used in epidemiology [19] to describe the spreading of epidemics, in psychology to describe learning [19], and in technology to describe how new technologies diffuse and substitute each other [10]. The distribution is pivotal in item response theory, for example, in the very basic Rasch model [17, 21, 22], in categorical data analysis [3] and in

---

M. Ghosh

Department of Statistics, University of Florida, P.O. Box 118545, Gainesville,  
FL 32611-8545, USA

e-mail: [ghoshm@stat.ufl.edu](mailto:ghoshm@stat.ufl.edu)

K.P. Choi and J. Li

Department of Statistics and Applied Probability, National University of Singapore,  
6 Science Drive 2, Singapore 117546

e-mail: [stackp@nus.edu.sg](mailto:stackp@nus.edu.sg); [stalj@nus.edu.sg](mailto:stalj@nus.edu.sg)

case-control studies [7], a topic which has become the cornerstone in epidemiologic research. Johnson et al. [14] devoted a whole chapter on logistic distributions, while a large body of articles covering many facets of this distribution appeared in the edited volume [6] entitled “Handbook of the Logistic Distribution.” In particular, the logistic distribution may be used to model a latent variable for the binary outcome data [5] and give rise to the well-known logistic regression model. The idea of introducing the latent logistic variable is especially helpful for multilevel mixed-effects logistic regression models. One of the several key definitions of variance partitioning coefficients (VPC) for such sophisticated models depends explicitly on the logistic distribution and uses the variance of a logistic variable as the level one variance [12]. For example, the VPC for a two-level logistic regression model  $\log \frac{p_{ij}}{1-p_{ij}} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i$ , where  $i$  indicates clusters and  $j$  indicates observations with a cluster, is given by

$$\text{VPC} = \frac{\text{var}(b_i)}{\text{var}(b_i) + \pi^2/3},$$

where  $\text{var}(b_i)$  is the variance of the random intercept and  $\pi^2/3$  is the variance of the standard logistic distribution. Unlike the unbounded variance parameter  $\text{var}(b_i)$ , the VPC parameter simply lies within  $(0, 1)$  and is scale-free. Furthermore, the computed VPC value provides an insight on the proportion of variation explained at the cluster level (level two) and is therefore frequently reported in longitudinal data analysis [11]. A tri-level version of the above VPC parameter has also been introduced recently in the literature based on a similar construction [8, 20]. The logistic distribution is indispensable for interpreting the variance parameters in these problems. Understanding the properties for the logistic distribution is thus meaningful and necessary.

Despite the popularity of the logistic distribution among both theoretical and applied researchers, many aspects of this distribution are still unfamiliar to most statisticians. As an example, we may cite the very important result of Andrews and Mallows [2] that the logistic distribution is a scale mixture of a normal distribution. Even the derivation of some of the basic results such as the variance, kurtosis, and other parameters of interest have still not entered most statistical textbooks.

The purpose of this note is to give a series representation of the standard logistic probability density function (pdf) in terms of several differently scaled double exponential distributions centered at zero with terms of the series alternating in signs. This particular representation was somewhat implicit in Johnson et al. [14, p. 117] while calculating the moments of a logistic distribution, but was never made explicit. With this series representation, we will be able to provide fairly simple expressions for moments, moment generating functions (mgf), and characteristic functions (cf) from the corresponding results for the double exponential distribution. One of the highlights of our approach is that contour integration is avoided altogether, especially for deriving the characteristic functions. Similar results for the normal, double exponential, and  $t$ -family of distributions are available in [9].

Andrews and Mallows [2] showed how a standard logistic distribution could be obtained as the scale mixture of a standard normal distribution, and provided also



an explicit expression of the mixing distribution. They found the latter by inverting a Laplace transform. In this note, we derive this distribution directly from the same series representation by writing the double exponential distribution also as a scale mixture of a normal distribution.

The main results are given in Sect. 2. Section 3 contains a summary of our results and suggests some potential future research.

## 2 The Main Results

The standard logistic distribution has a pdf of the form

$$f(x) = \frac{\exp(-x)}{(1 + \exp(-x))^2} = \frac{\exp(-|x|)}{(1 + \exp(-|x|))^2}, \tag{1}$$

due to the symmetry of  $f$  around zero. With the expansion,  $(1 + \exp(-|x|))^{-2} = \sum_{k=1}^{\infty} (-1)^{k-1} k \exp[-(k - 1)|x|]$  for  $x \neq 0$ , it is possible to rewrite (1) as

$$\begin{aligned} f(x) &= \sum_{k=1}^{\infty} (-1)^{k-1} k \exp(-k|x|) \\ &= 2 \sum_{k=1}^{\infty} (-1)^{k-1} (k/2) \exp(-k|x|), \quad x \neq 0. \end{aligned} \tag{2}$$

Noting that  $g_k(x) = (k/2) \exp(-k|x|)$  is the pdf of a double exponential distribution with zero mean and scale parameter  $k^{-1}$ , one gets the exact series representation. Direct application of ratio test shows that this series converges absolutely for  $x \neq 0$ . Indeed, this series converges uniformly for  $|x| > \epsilon$  for any  $\epsilon > 0$ .

The moments of the logistic distribution are easily obtained from (2). Due to symmetry of  $f$  around zero, all the odd moments of  $f$  are zero. Write  $f = h_1 - h_2$  where  $h_1(x) := 2 \sum_{k=1}^{\infty} 2^{-1} (2k - 1) \exp[-(2k - 1)|x|]$  and  $h_2(x) := 2 \sum_{k=1}^{\infty} k \exp(-2k|x|)$ . An even moment is easily calculated as

$$E(X^{2m}) = 2 \int_0^{\infty} x^{2m} f(x) dx = 2 \int_0^{\infty} x^{2m} h_1(x) dx - 2 \int_0^{\infty} x^{2m} h_2(x) dx.$$

Note that, for  $m \geq 1$ ,

$$\begin{aligned} 2 \int_0^{\infty} x^{2m} h_1(x) dx &= 2 \sum_{k=1}^{\infty} (2k - 1) \int_0^{\infty} x^{2m} \exp[-(2k - 1)x] dx \\ &= 2(2m)! \sum_{k=1}^{\infty} (2k - 1)^{-2m} \end{aligned}$$

and similarly

$$2 \int_0^\infty x^{2m} h_2(x) dx = 2(2m)! \sum_{k=1}^\infty (2k)^{-2m}.$$

Therefore,

$$\begin{aligned} E(X^{2m}) &= 2(2m)! \sum_{k=1}^\infty (2k-1)^{-2m} - 2(2m)! \sum_{k=1}^\infty (2k)^{-2m} \\ &= 2(2m)! \zeta(2m) - 4(2m)! \sum_{k=1}^\infty (2k)^{-2m} \\ &= 2(2m)! [1 - 2^{-(2m-1)}] \zeta(2m), \end{aligned} \tag{3}$$

where  $\zeta(s) = \sum_{n=1}^\infty n^{-s}$  is the well-known Riemann zeta function [1, p. 256].

In particular,  $\text{var}(X) = E(X^2) = 2\zeta(2) = \pi^2/3$ . Johnson et al. [14, p. 117] also provided the expression in the last line of (3).

The mgf and cf can be deduced directly from (3). For  $|t| < 1$ , we have

$$\begin{aligned} E \exp(tX) &= E[\exp(tX) + \exp(-tX)]/2 \\ &= E \left\{ 1 + \sum_{n=1}^\infty \frac{t^{2n} X^{2n}}{(2n)!} \right\} \\ &= 1 + \sum_{n=1}^\infty \frac{E X^{2n}}{(2n)!} t^{2n} \\ &= 1 + \sum_{n=1}^\infty \frac{[2^{2n-1} - 1] \zeta(2n)}{2^{2(n-1)}} t^{2n} \\ &= 1 + \sum_{n=1}^\infty \frac{(-1)^{n-1} 2[2^{2n-1} - 1] B_{2n} \pi^{2n}}{(2n)!} t^{2n} \\ &= \frac{\pi t}{\sin(\pi t)}, \end{aligned}$$

where we used a known identity of Riemann zeta function and Bernoulli numbers in the penultimate equality [6, p. 34], and Taylor expansion of  $\frac{\pi t}{\sin(\pi t)}$  in the last equality [13, p. 35]. Recall that Bernoulli numbers  $B_n$  are given by the series expansion:

$$\frac{t}{e^t - 1} = \sum_{n=0}^\infty \frac{B_n}{n!} t^n.$$

For example,  $B_0 = 1, B_1 = -1/2, B_2 = 1/6, B_4 = -1/30, \dots$  Similarly,

$$\begin{aligned} E \exp(itX) &= E[\exp(itX) + \exp(-itX)]/2 \\ &= 1 + \sum_{n=1}^\infty (-1)^n \frac{E X^{2n}}{(2n)!} t^{2n} \end{aligned}$$

$$\begin{aligned}
 &= 1 + \sum_{n=1}^{\infty} (-1)^n \frac{[2^{2n-1} - 1]\zeta(2n)}{2^{2(n-1)}} t^{2n} \\
 &= 1 - \sum_{n=1}^{\infty} \frac{2[2^{2n-1} - 1]B_{2n}\pi^{2n}}{(2n)!} t^{2n} \\
 &= \frac{\pi t}{\sinh(\pi t)} \quad [13, \text{p. 35}].
 \end{aligned}$$

We illustrate another use of the series representation, (2), of the logistic pdf to derive a result of Andrews and Mallows [2]. To this end, first rewrite

$$(k/2) \exp(-k|x|) = \int_0^{\infty} \frac{kr^{1/2}}{\sqrt{2\pi}} \exp\left(-\frac{rk^2x^2}{2}\right) \frac{\exp\left(-\frac{1}{2r}\right)}{2r^2} dr \tag{4}$$

$$= \int_0^{\infty} \left\{ \frac{v}{\sqrt{2\pi}} \exp\left(-\frac{v^2x^2}{2}\right) \right\} \frac{k^2}{v^3} \exp\left(-\frac{k^2}{2v^2}\right) dv. \tag{5}$$

To see (4), rewrite the right hand side of (4) as

$$\begin{aligned}
 &(k/2) \int_0^{\infty} (2\pi r^3)^{-1/2} \exp\left[-\frac{1}{2r}(r^2k^2x^2 + 1)\right] dr \\
 &= (k/2) \exp(-k|x|) \int_0^{\infty} (2\pi r^3)^{-1/2} \exp\left[-\frac{1}{2r}(rk|x| - 1)^2\right] dr. \tag{6}
 \end{aligned}$$

Recognizing the above integrand as the pdf of an inverse Gaussian distribution with mean  $(k|x|)^{-1}$  and scale parameter 1, (6) simplifies to  $(k/2) \exp(-k|x|)$  which is the left hand side of (4). Using the substitution  $k\sqrt{r} = v$ , (5) follows from (4).

In view of (5), we have

$$\begin{aligned}
 h_1(x) &= 2 \sum_{k=1}^{\infty} 2^{-1}(2k - 1) \exp[-(2k - 1)|x|] \\
 &= \int_0^{\infty} \left\{ \frac{v}{\sqrt{2\pi}} \exp\left(-\frac{v^2x^2}{2}\right) \right\} \left[ 2 \sum_{k=1}^{\infty} \frac{(2k - 1)^2}{v^3} \exp\left(-\frac{(2k - 1)^2}{2v^2}\right) \right] dv
 \end{aligned}$$

and

$$h_2(x) = \int_0^{\infty} \left\{ \frac{v}{\sqrt{2\pi}} \exp\left(-\frac{v^2x^2}{2}\right) \right\} \left[ 2 \sum_{k=1}^{\infty} \frac{(2k)^2}{v^3} \exp\left(-\frac{(2k)^2}{2v^2}\right) \right] dv.$$

Hence

$$f(x) = h_1(x) - h_2(x) = \int_0^{\infty} \left\{ \frac{v}{\sqrt{2\pi}} \exp\left(-\frac{v^2x^2}{2}\right) \right\} g(v)dv,$$

where

$$g(v) = 2 \sum_{k=1}^{\infty} (-1)^{k-1} (k^2/v^3) \exp(-k^2/(2v^2)). \quad (7)$$

Consequently, we see that  $X|V = v \sim N(0, v^{-2})$  and the mixing pdf  $g(v)$  of  $V$ . This provides an alternative proof of a result in Andrews and Mallows [2] who obtained it by inverting a Laplace transform. As also noted in Andrews and Mallows [2],  $W = 1/(2V)$  has the Kolmogorov distribution with pdf [16, p.480]

$$h(w) = 8w \sum_{k=1}^{\infty} (-1)^{k-1} k^2 \exp(-2k^2w^2).$$

### 3 Summary

The paper provides a series representation of the logistic pdf, the terms of the series being differently scaled double exponential pdf and also alternating in signs. It will be interesting to see whether a similar representation is available for the multivariate logistic pdf.

The logistic distribution, indeed, appears in various statistical problems other than the logistic regression mentioned earlier. In survival analysis, this distribution is considered as a common parametric error distribution in an accelerated failure time model [15, p. 37]. In practice, the instructive results we derive in this paper may also benefit the analytic studies of lifetime data when the logistic distribution approximates the data closely.

**Acknowledgments** This research took place when the first author was visiting the Department of Statistics and Applied Probability, National University of Singapore. He acknowledges gratefully the research opportunities made available to him during this period.

### References

- [1] Abramowitz, M., and Stegun, I. A. (1972). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Dover, New York.
- [2] Andrews, D. F., and Mallows, C. F. (1974). Scale mixtures of a normal distributions. *Journal of the Royal Statistical Society, Series B*, **36**, 99–102.
- [3] Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.
- [4] Balakrishnan, N. (1992). *Handbook of the Logistic Distribution*. Marcel Dekker, New York.
- [5] Bartholomew, D. J., and Knott, M. (1999). *Latent Variable Models and Factor Analysis*. Edward Arnold, London.
- [6] Bateman, H. (1953). *Higher Transcendental Functions, Vol. I*. McGraw Hill, New York.
- [7] Breslow, N. E., and Day, N. E. (1980). *Statistical Methods in Cancer Research I. The Analysis of case–Control Studies*. IARC, Lyon.

- [8] Browne, W. J., Subramanian, S. V., Jones, K., and Goldstein, H. (2005). Variance partitioning in multilevel logistic models that exhibit overdispersion. *Journal of the Royal Statistical Society, Series A*, **168**, 599–613.
- [9] Datta, G. S., and Ghosh, M. (2007). Characteristic functions without contour integration. *The American Statistician*, **61**, 67–70.
- [10] Fisher, J. C., and Pry, R. H. (1971). A simple substitutional model for technological change. *Technological Forecasting and Social Change*, **3**, 75–88.
- [11] Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2003). *Applied Longitudinal Analysis*. Wiley, New York.
- [12] Goldstein, H., Browne, W. J., and Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, **1**, 223–232.
- [13] Gradshteyn, I. S., and Ryzhik, I. M. (1980). *Tables, Integrals, Series, and Products*. Academic, New York.
- [14] Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distribution, Vol. 2*. Wiley, New York.
- [15] Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York.
- [16] Kendall, S. M., and Stuart, A. (1979). *The Advanced Theory of Statistics, Vol. 2*. Fourth Edition. Charles Griffin & Company Limited, London.
- [17] Lord, F. M. (1965). A note on the normal ogive or logistic curve in item analysis. *Psychometrika*, **30**, 371–372.
- [18] Lotka, A. J. (1925). *Elements of Physical Biology*. Williams and Wilkins, Baltimore.
- [19] Modis, T. (1992). *Predictions: Society's Telltale Signature Reveals Past & Forecasts the Future*. Simon and Schuster, New York, pp 97–105.
- [20] Rabe-Hesketh, S., and Skrondal, A. (2005). *Multilevel and Longitudinal Modeling Using Stata*. StatCorp LP, College Station, Texas.
- [21] Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics*. University of California, Berkeley, pp 321–334.
- [22] Sanathanan, L. (1974). Some properties of the logistic distribution for dichotomous response. *Journal of the American Statistical Association*, **69**, 744–749.

# Entropy and Cross Entropy: Characterizations and Applications

C.R. Rao

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** The paper provides an axiomatic setup for an entropy function as a measure of diversity. A general definition of cross entropy is given and its use in solving a variety of stochastic and nonstochastic optimization problems is mentioned. A method of deriving a cross entropy function associated with a given entropy function is given.

**Mathematics Subject Classification (1991)** 94A17

**Key words and phrases** Cross entropy · Diversity · Decomposition of diversity · Kullback–Leibler divergence · Maximum entropy principle · Shannon entropy

## 1 Introduction

Let  $\mathcal{P}$  be a set of probability distributions defined on a measurable space  $(\mathcal{X}, \mathcal{B})$ . Entropy of  $p \in \mathcal{P}$ , denoted by  $H(p)$ , was originally designed as a measure of uncertainty of the outcomes of a probability distribution  $p$ , or how close  $p$  is to uniform distribution. The most popular choice of  $H$ , known as Shannon entropy, is

$$H(p) = \sum_1^k p_i \log p_i \quad (1.1)$$

---

C.R. Rao

Advanced Institute of Mathematics, Statistics, and Computer Science, University of Hyderabad campus, Central University PO, Hyderabad 500046, Andhra Pradesh, India  
e-mail: [crr1@psu.edu](mailto:crr1@psu.edu)

in the case of a multinomial distribution with  $k$  classes and cell probabilities  $p_1, \dots, p_k$ , and

$$H(p) = - \int p(x) \log p(x) dv \quad (1.2)$$

in the continuous case, where  $dv$  is the volume element in  $\mathcal{X}$ .

The functions defined in (1.1) and (1.2) have been used in building models of probability distributions for elementary particles in physics and in solving some problems in communication theory. The same functions have been used as measures of diversity by ecologists in discussing problems of differences in frequencies of different species of animals inhabiting a locality. Some key references are [10, 12, 13, 15–19, 24].

While  $H(\cdot)$  is defined on  $\mathcal{P}$ , there is another function  $C(p|q)$  defined on  $\mathcal{P} \times \mathcal{P}$ , called cross entropy, not necessarily symmetric, designed to examine how close a probability distribution  $q$  is to a given distribution  $p$ . A well-known cross entropy function is Kullback–Leibler [9] divergence

$$\begin{aligned} C(p|q) &= \sum p_i \log \frac{p_i}{q_i}, \text{ in the discrete case,} \\ &= \int p(x) \log \frac{p(x)}{q(x)} dv, \text{ in the continuous case.} \end{aligned} \quad (1.3)$$

Cross entropy received numerous applications in solving complicated optimization problems as described in [2, 8]. An interesting application of the cross-entropy method is in estimating rare-events probability as discussed in [5].

In this paper, a general discussion of entropy and cross-entropy measures and their characterizations, and possible applications are given.

This paper is dedicated to the memory of Professor Alladi Ramakrishnan who not only made fundamental contributions to frontier areas of stochastic processes, elementary particle physics, special theory of relativity, and matrix algebra but also created a monument for himself by establishing the Institute of Mathematical Sciences to promote basic research in key areas of science.

## 2 Entropy Functional

A strict definition of  $H(\cdot)$  would depend on how uncertainty in prediction is defined and loss in making predictions of outcomes as in [3, 23]. However, we can state some general postulates.

$$\begin{aligned} A_1 : & H(p) \geq 0 \forall p \in \mathcal{P} \text{ and } = 0 \text{ if } p \text{ is degenerate.} \\ A_2 : & H(\lambda p + \mu q) - \lambda H(p) - \mu H(q) = J(p, q : \lambda, \mu) \geq 0 \\ & \text{for all } p \text{ and } q \text{ and } \lambda > 0, \mu > 0, \lambda + \mu = 1, \text{ and } = 0 \text{ if } p = q. \end{aligned}$$

While  $A_1$  requires  $H(\cdot)$  to be a nonnegative function,  $A_2$  implies that uncertainty in a mixture of distributions is strictly greater than the average of the uncertainties in each of the components if they are different, i.e.,  $H(\cdot)$  is a strongly concave function on  $\mathcal{P}$ .

Dalton and Pielou postulated the following condition on  $H(\cdot)$ , calling it a diversity measure, when  $p$  is multinomial in  $k$  classes:

$$H(p_1, \dots, p_i, \dots, p_j, \dots, p_k) \leq H(p_1, \dots, p_i + \delta, \dots, p_j - \delta, \dots, p_k)$$

$$\text{for } p_i < p_i + \delta < p_j - \delta < p_j$$

i.e.,  $H(\cdot)$  increases if some part is transferred from a large  $p_j$  to a smaller  $p_i$ .

This condition is implied by the postulate  $A_2$  if  $H(\cdot)$  is a symmetric function of  $p_1, \dots, p_k$  as shown below. Consider two probability vectors

$$p = (p_1, \dots, p_i, \dots, p_j, \dots, p_k), \tag{2.1}$$

$$q = (p_1, \dots, p_j, \dots, p_i, \dots, p_k). \tag{2.2}$$

Then the  $i$ th and  $j$ th values in  $\lambda p + \mu q$  are

$$\lambda p_i + \mu p_j = p_i + \delta, \lambda p_j + \mu p_i = p_j - \delta, \tag{2.3}$$

which give

$$\lambda = 1 - \delta / (p_j - p_i), \mu = \delta / (p_j - p_i). \tag{2.4}$$

Using  $\lambda$  and  $\mu$  as in (2.4), the postulate  $A_2$  and the symmetry  $H(p) = H(q)$ ,

$$H(P) = \lambda H(p) + \mu H(q) \leq H(\lambda p + \mu q) = H(p_1, \dots, p_i + \delta, \dots, p_j - \delta, \dots, p_k).$$

This also demonstrates that a symmetric  $H(\cdot)$ , under postulates  $A_1$  and  $A_2$ , attains the maximum value when all  $p_i$  are equal.

Some examples of entropy functions in the case of multinomial distributions, which have been used in various applications are as follows.

- (1)  $-\sum p_i \log p_i$ , [24].
- (2)  $\sum \sum d_{ij} p_i p_j, d_{11} = \dots = d_{kk}$ , and the matrix  $(d_{ik} + d_{jk} - d_{ij} - d_{kk}), i, j = 1, \dots, k - 1$  is nonnegative definite, Rao's [18] quadratic entropy.
- (3)  $1 - \sum p_i^2$ , Gini-Simpson [25].  
(special case of 2 with  $d_{ii} = 0 \forall i$  and  $d_{ij} = 1$  for all  $i \neq j$ ).
- (4)  $(1 - \alpha)^{-1} \log \sum p_i^\alpha, \alpha > 0$ , [21].
- (5)  $(\alpha - 1)^{-1} (1 - \sum p_i^\alpha)$ , [4].
- (6)  $\left[ 1 - \left( \sum p_i^{1/\gamma} \right)^\gamma \right] / (1 - 2^{\gamma-1}), \gamma > 0, \neq 1$ , ( $\gamma$  - entropy).
- (7)  $-\sum p_i \log p_i - \sum (1 - p_i) \log (1 - p_i)$ , paired entropy.



All these entropy functions satisfy the postulates  $A_1$  and  $A_2$  and all except 2 attain the maximum value when all  $p_i$  are equal. It may be noted that the quadratic entropy has a great potential for applications in statistics and ecology as shown in [16, 17, 26].

In the continuous case, some entropy functions are as follows:

- (1)  $-\int p(x) \log(x) dv$ , Shannon.
- (2)  $\int K(x, y) p(x) p(y) dv_x dv_y$ , Rao's quadratic entropy where  $K$  is a conditionally negative definite kernel, i.e.,

$$\sum_1^n \sum_1^n K(x_i x_j) a_i a_j \leq 0$$

for any  $n$  and  $x_1, \dots, x_n$  such that  $\sum a_i = 0$ .

- (3)  $\frac{1}{1-\alpha} \log \int p^\alpha dv$ , Renyi.

## 2.1 Maximum Entropy Principle

Physicists used the principle of maximizing entropy in generating models for distribution of particles such as molecules subject to certain kinematic restrictions. For instance, a particle occupying a certain position in a phase space will have an energy  $E$  whose average may be known giving the equation such as

$$\int E(x) p(x) dv = c. \quad (2.5)$$

If we choose Shannon entropy, the problem is to find  $p$  such that

$$-\int p(x) \log p(x) dv \quad (2.6)$$

is a maximum subject to the condition (2.5). The solution [14, p. 173] is obtained as

$$p = \alpha \exp(\lambda E), \quad (2.7)$$

which is known as Maxwell–Boltzmann distribution. For the use of (2.7) in building models such as Helly's law for the equilibrium of sedimentation, Maxwell distribution of velocities and angular distribution of areas of elementary magnets in a magnetic field, the reader is referred to Joos [7], Jaynes [6], and Rao [14, pp. 172–175].

If instead of Shannon entropy, we choose to maximize Renyi's entropy

$$\frac{1}{1-\alpha} \log \int p^\alpha dv,$$

we obtain the distribution [14, p. 175] with maximum entropy subject to (2.5) as

$$p(x) = (\lambda E(x) + \mu)^{1/(\alpha-1)}, \tag{2.8}$$

which provides a family of models to explain various physical phenomena as an alternative to (2.7). Some comparison of the models (2.7) and (2.8) may be made in different situations using observational data. It would also be of interest to build models using other entropy functions.

### 3 Cross Entropy

#### 3.1 Characterization

There are situations in statistical theory and optimization problems where the true probability distribution  $p$  is not known but we use a surrogate distribution  $q$  for analysis of observations drawn from  $p$ , or  $p$  is known but it is easy to generate observations from  $q$  to estimate some quantities, such as probabilities of large deviations in  $p$ , by a technique known in statistics as importance sampling. In such problems, to make a choice of  $q$  we need a measure of how close  $q$  is to  $p$ . Such a measure is known as cross entropy and is indicated by  $C(p|q)$ . We suggest a few postulates for the choice of  $C(p|q)$ .

$$B_1 : C(p|q) \geq 0 \forall p, q \in \mathcal{P}, = 0 \text{ only if } p = q,$$

$$B_2 : C(\lambda p + \mu q|q) \leq C(p|q), \lambda > 0, \mu > 0, \lambda + \mu = 1.$$

The postulate  $B_2$  is a natural requirement as the mixture  $\lambda p + \mu q$  has some component of  $q$  which would make  $q$  closer to  $\lambda p + \mu q$ . Rao and Nayak [20] provide a general choice of  $C(p|q)$  based on a given entropy function  $H(p)$  with the smooth differentiability property

$$H(\lambda p + \mu q) - H(q) = \lambda f(q, p - q) + o(\lambda), \tag{3.1}$$

where  $f(q, p - q), p, q \in \mathcal{P}$ , is such that  $f(q, 0) = 0$  and  $f(q, \alpha(p - q)) = \alpha f(q, p - q)$ .

We may compute  $f(q, p - q)$  as

$$\lim_{\lambda \rightarrow 0} \frac{H(q + \lambda(p - q)) - H(q)}{\lambda}. \tag{3.2}$$

We then define

$$C_H(p|q) = f(q, p - q) + H(q) - H(p) \tag{3.3}$$

as the cross-entropy of  $q$  with respect to  $p$  based on a given entropy function  $H(\cdot)$ .

Consider for instance Shannon entropy,  $-\sum p_i \log p_i$ . Then

$$f(q, p - q) = -\sum (p_i - q_i) \log q_i, \quad (3.4)$$

$$\begin{aligned} C_H(p, q) &= f(q, p - q) + H(q) - H(p) \\ &= \sum p_i \log \frac{p_i}{q_i}. \end{aligned} \quad (3.5)$$

Let us verify whether  $C_H(p|q)$  as defined in (3.3) satisfies the postulates  $B_1$  and  $B_2$ . Using concavity of  $H(\cdot)$

$$\begin{aligned} H(q + \lambda(p - q)) &> \lambda H(p) + \mu H(q) \\ \frac{H(q + \lambda(p - q)) - H(q)}{\lambda} &> (H(p) - H(q)). \end{aligned} \quad (3.6)$$

Taking the limit as  $\lambda \rightarrow 0$

$$\begin{aligned} f(q, p - q) &> H(p) - H(q), \\ C_H(p|q) &= f(q, p - q) - H(p) + H(q) > 0, \end{aligned} \quad (3.7)$$

which proves  $B_1$ . To prove  $B_2$ , consider

$$\begin{aligned} C_H(p|q) - C_H(\lambda p + \mu q|q) &= \mu (f(q, p - q) + H(q) - H(p)) \\ &\quad + H(\lambda p + \mu q) - \lambda H(p) - \mu H(q) \geq 0. \end{aligned} \quad (3.8)$$

Let us examine whether Kullback–Leibler divergence measure

$$C_H(p|q) = \sum p_i \log \frac{p_i}{q_i} \quad (3.9)$$

satisfies the postulate  $B_2$ .

$$\begin{aligned} C_H(p|q) - C_H(\lambda p + \mu q|q) &= \sum \left[ p_i \log p_i - p_i \log q_i + (\lambda p_i + \mu q_i) \log q_i \right. \\ &\quad \left. - (\lambda p_i + \mu q_i) \log(\lambda p_i + \mu q_i) \right] \\ &= \sum \left[ \mu p_i \log \frac{p_i}{q_i} + \lambda p_i \log p_i + \mu q_i \log q_i \right. \\ &\quad \left. - (\lambda p_i + \mu q_i) \log(\lambda p_i + \mu q_i) \right] \geq 0. \end{aligned} \quad (3.10)$$

### 3.2 Decomposition of $H(\cdot)$

Let  $\bar{P} = \lambda_1 P_1 + \cdots + \lambda_m P_m$ , where  $P_1, \dots, P_m$  are probability measures and  $\lambda_i \geq 0 \forall i$  and  $\sum \lambda_i = 1$ . Then

$$H(\bar{P}) = \sum \lambda_i H(P_i) + \sum \lambda_i C_H(P_i|\bar{P}). \quad (3.11)$$

From (3.3)

$$\begin{aligned}
 C_H(P_i|\bar{P}) &= f(\bar{P}, P_i - \bar{P}) + H(\bar{P}) - H(P_i) \\
 \sum \lambda_i C_H(P_i|\bar{P}) &= \sum \lambda_i f(\bar{P}, P_i - \bar{P}) + H(\bar{P}) - \sum \lambda_i H(P_i) \\
 &= f(\bar{P}, \sum \lambda_i (P_i - \bar{P})) + H(\bar{P}) - \sum \lambda_i H(P_i) \\
 &= H(\bar{P}) - \sum \lambda_i H(P_i),
 \end{aligned}$$

which proves (3.11).

### 3.3 Some Applications of Cross Entropy

A comprehensive account of the use of cross entropy in solving complicated optimization problems, and estimation of probabilities of rare events is given in [2, 5, 8, 22]. Some of the applications in statistics are the construction of classification regions using Support Vector Machines as in [11] and iterative methods of cluster analysis as in [8].

An example of how cross entropy is used is as follows. Suppose the problem is that of estimating  $\gamma = E_p[\Phi(x)]$ , the expectation of a function  $\Phi(x)$  with respect to a given probability distribution  $p(x)$ . For instance, if we want to find the probability of  $x \geq a$ , we can express it as the expectation of the function  $I_{x \geq a}$ , where  $I$  is the indicator function.

A general Monte Carlo technique of estimating  $\gamma$  is to draw a sample  $x_1, \dots, x_n$  from  $p(x)$  and estimate  $\gamma$  by

$$\hat{\gamma} = n^{-1} \sum \Phi(x_i). \tag{3.12}$$

Observing that

$$\int \Phi(x)p(x) dx = \int \Phi(x) \frac{p(x)}{q(x)} q(x) dx \tag{3.13}$$

and

$$\gamma = E_q \left[ \Phi(x) \frac{p(x)}{q(x)} \right], \tag{3.14}$$

we may draw a sample  $(x'_1, \dots, x'_n)$  from  $q$  and estimate  $\gamma$  by

$$\hat{\gamma} = n^{-1} \sum \Phi(x'_i)p(x'_i)/q(x'_i). \tag{3.15}$$

The best choice of  $q$  which reduces the variance of  $\hat{\gamma}$  to zero is

$$q^*(x) = \frac{\Phi(x)p(x)}{\gamma}. \tag{3.16}$$

However, the solution depends on the unknown  $\gamma$ . An alternative is to choose a family of probability distributions indexed by a number of parameters

$$q(x, \theta), \theta = (\theta_1, \dots, \theta_s) \quad (3.17)$$

and estimate  $\theta$  by minimizing the cross entropy

$$C_H(q^*(x)|q(x, \theta)), \quad (3.18)$$

where  $q^*$  is as determined in (3.16). If we are using KL divergence measure, the problem reduces to

$$\max_{\theta} \int q^*(x) \log q(x, \theta) dv. \quad (3.19)$$

There are a number of ways of solving (3.19), analytically or algorithmically, depending on the functional forms of  $q^*(x)$  and  $q(x, \theta)$ . Reference may be made to [1, 22].

Up-to-date, cross entropy method has been used to solve a variety of optimization problems arising in statistics, operations research, engineering, and finance. The superiority of the CE method over other computational methods has been demonstrated in the papers referred to above.

## References

- [1] Celesti, F., Dambreville, F., and Le Ladre, J.P. (2006). Optimal path planning using cross-entropy method, ([fusion.carthel.com/technical\\_program/abstracts/303.htm](http://fusion.carthel.com/technical_program/abstracts/303.htm)).
- [2] de-Boer, P.T., Kroese, D.P., Mannor, S., and Rubinstein, R.Y. (2005). A tutorial on cross-entropy method, *Annals of Operations Research*, **134**, 19–67.
- [3] Habermann, S.J. (1982). Analysis of dispersion of multinomial probabilities, *J. Amer. Statist. Ass.*, **77**, 568–580.
- [4] Havrada, J. and Charvát, F. (1967). Quantification method in classification processes: Concept of structural  $\alpha$ -entropy, *Kybernetika*, **30**, 30–35.
- [5] Homem-de-Mello, T. (2007). A study of the cross-entropy method for rare-event probability estimation, *INFORMS J. on Computing*, **19**, 381–394.
- [6] Jaynes, E.T. (1957). Information theory and statistical mechanics, *Physical Review*, **106**, 620–630.
- [7] Joos, G. (1951). *Theoretical Physics*, Haffner, New York.
- [8] Kroese, D.T., Rubinstein, R.Y., and Taimre, T. (2007). Application of cross-entropy method to clustering, *J. Glob Optim.*, **37**, 137–157.
- [9] Kullback, S. and Leibler, R.A. (1951). On information and sufficiency, *Ann. Math. Statist.*, **22**, 79–86.
- [10] Lewontin, R.C. (1972). The apportionment of human diversity, *Evolutionary Biology*, **6**, 381–398.
- [11] Mannor, S., Peleg, D., and Rubinstein, R. (2005). *Proc. 22nd International Conference on Machine Learning*.
- [12] Patil, G.P. and Taillie, C. (1982). Diversity as a concept and its measurement, *J. Amer. Statist. Ass.*, **77**, 548–567.
- [13] Pielou, E.C. (1975). *Ecological Diversity*, Wiley, New York.

- [14] Rao, C.R. (1973). *Linear Statistical Inference and its Applications* (Second edition), Wiley, New York.
- [15] Rao, C.R. (1982a). Diversity and dissimilarity coefficients: A unified approach, *Theor. Popul. Bio*, **21**, 24–43.
- [16] Rao, C.R. (1982b). Diversity, its measurement, decomposition, apportionment and analysis, *Sankhya*, **44**, 1–21.
- [17] Rao, C.R. (1982c). Gini-Simpson index of diversity: A characterization, generalization and applications, *Utilitas Mathematica*, **21**, 273–282.
- [18] Rao, C.R. (1984). Convexity properties of entropy functions and analysis of diversity, In *Inequalities in Statistics and Probability*, Ed. Y.L. Tong, IMS Lecture Notes, **5**, 68–77.
- [19] Rao, C.R. (1986). Rao's axiomatization of diversity measures, In *Encyclopedia of Statistical Sciences*, Vol 7, Wiley, New York, 614–617.
- [20] Rao, C.R. and Nayak, T. (1985). Cross-entropy, dissimilarity measures and quadratic entropy, *IEEE Transactions of Information Theory*, **31**, 589–593.
- [21] Renyi, A. (1961). On measures of information and entropy, *Proc. 4-th Berkeley Symposium on Math. Stat. and Prob.*, 547–561.
- [22] Rubinstein, R. and Kroese, D.P. (2004). The cross entropy method: A unified approach to combinational optimization, Monte Carlo simulation and machine learning, *Information Science & Statistics*, Springer.
- [23] Savage, L.J. (1971). Elicitation of personal probabilities and expectations, *J. Amer. Statist. Ass.*, **66**, 783–801.
- [24] Shannon, C.E. (1948). A mathematical theory of communication, *Bell System Technical Journal*, **27**, 379–423, 623–656.
- [25] Simpson, E.H. (1949). Measurement of diversity, *Nature*, **163**, 688.
- [26] Zoltan, B.-D. (2008). Rao's quadratic entropy as a measure of functional diversity based on multiple traits, *J. Vegetation Science*, **16**, 533–540.

# Optimal Weights for a Class of Rank Tests for Censored Bivariate Data

Samuel S. Wu, P.V. Rao, and Aparna Raychaudhuri

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** The problem of testing equality of survival distributions on the basis of paired censored survival data has received considerable attention in literature. Some of the important statistics used for such purposes can be expressed as linear combinations of two statistics, one based on uncensored pairs and the other based on the censored pairs. Raychaudhuri and Rao (Nonparametric Statistics, 1996, 6, 1–11) investigated properties of two classes of such statistics and derived expressions for the optimal coefficients (weights) for the linear combination that will maximize efficacy within each class. As the optimal weights depend upon the form of the underlying survival and censoring distributions, statistics with optimal weights can only be used with estimated weights. This article presents a method of estimating optimal weights on the basis of an assumed model that specifies the distribution of the difference between the observed survival times conditional on the censoring pattern. The model, in addition to dispensing with the usual assumption that the survival and censoring variables are independent, also permits a graphical check of its lack of fit on the basis of observed data. The performance of statistics with the estimated weights is evaluated by using two simulation studies – one with data generated under the assumed model and the other assuming independence of the survival and censoring times. Simulation results show that the optimal statistics with estimated weights have good power properties in all cases considered, and that they compare well with other commonly used tests for paired censored survival data. An advantage of the tests with optimal weights is that, unlike their competitors, these tests have demonstrated performance characteristics in some cases where the assumption of independent censoring may not be justified.

**Mathematics Subject Classification (2000)** Primary 62N03, Secondary 62G10

**Key words and phrases** Paired censored survival data · Efficacy · Optimal weights

---

S.S. Wu, P.V. Rao, and A. Raychaudhuri  
University of Florida, University of Florida and Statistical Consultant  
e-mail: [samwu@biostat.ufl.edu](mailto:samwu@biostat.ufl.edu); [pejavervrao@gmail.com](mailto:pejavervrao@gmail.com); [aparna\\_raychaudhuri@yahoo.com](mailto:aparna_raychaudhuri@yahoo.com)

## 1 Introduction

In statistical literature survival times refer to the times to occurrences of an event in a given population of individuals. In biomedical applications, the event of interest could be the death from a disease or relapse of a symptom in a population of treated patients. In engineering applications, survival times may represent the life lengths of a particular type of aircraft engines. Survival times are encountered in a wide variety of other applications such as sociology (duration of first employment) and insurance (amounts of disability insurance claims). For obvious reasons survival times are also referred to as lifetimes or failure times.

A special feature of survival times is that they are amenable to censoring. A survival time is said to be right censored if the event of interest has not occurred before the end of the observation period. For example, in a study of the relapse-free time of treated patients, complete information about relapse time is not available for subjects who do not relapse during the study period. All we know about such subjects is that their relapse times are longer than their corresponding censoring times—the lengths of time the subjects were under observation. Hence the two pieces of information available about the survival time of an individual whose survival and censoring times are  $X$  and  $C$ , respectively, are  $Y = \min(X, C)$  and  $\delta = I(X \leq C)$ , where  $I(A)$  is the indicator function of  $A$ . Here  $Y$  is the observed survival time and  $\delta$  is its censoring status. If  $\delta = 1$ , then the observed survival time is the actual survival time. Otherwise, the actual survival time is longer than the observed survival time.

Let  $(X_{1i}, X_{2i})$ ,  $i = 1, \dots, n$  be independent bivariate survival times each distributed as a bivariate random variable  $(X_1, X_2)$  with continuous density  $\psi(x_1, x_2)$ . Let  $\{C_i : i = 1, \dots, n\}$  be an independent random sample of censoring times from a continuous population. Suppose that the survival times are not observable because of right censoring and the observed data consist of (1) the observed times  $(Y_{1i}, Y_{2i})$ , where  $Y_{ki} = \min(X_{ki}, C_i)$ , and (2) the censoring pattern  $(\delta_{1i}, \delta_{2i})$ , where  $\delta_{ki} = I(X_{ki} \leq C_i)$ , ( $i = 1, \dots, n; k = 1, 2$ ).

Testing the null hypothesis,  $H_0 : \psi(x_1, x_2) = \psi(x_2, x_1)$ , based on observed values of  $(Y_{ki}, \delta_{ki})$  is an important problem in biomedical research. For example, when times to responses to two drugs are measured on experimental units that are matched on the basis of shared characteristics, censored paired responses of the type  $Y_{ki}$  and  $\delta_{ki}$  result if there is a possibility that the response times  $X_1$  and  $X_2$  within a pair may be censored by a common censoring time. The null hypothesis that the response times have a common distribution can be tested by testing  $H_0$ .

The Florida Geriatric Research Program (FGRP) – a longitudinal study of the elderly begun in 1975 in Dunedin, Florida provides an example of the need for testing  $H_0$  on the basis of paired survival data. Over 6,500 ambulatory people at least 65 years of age have enrolled in FGRP and over 2,000 return each year for annual screenings. The screenings include detailed assessment of symptoms and diseases, as well as a SMAC-23 and blood pressure, heart rate, height and weight assessments. If we define  $X_1$  and  $X_2$  as the ages at which a subject's PCL13 (albumin) and HGB (hemoglobin) values reached "abnormal" levels for the first time during the observation period then the null hypothesis  $H_0$  can be interpreted as the hypothesis



that, at any given age of the patient, the likelihood of first abnormal PCL13 is the same as the likelihood of first abnormal HGB. In other words, testing  $H_0$  will help answer the question “Which, if any, of the two abnormalities is likely to occur first?” Let  $C$  denote the age at the most recent followup for a subject. A subject who does not present with one or both of the abnormalities at age  $C$  has a censored value for one or both of the  $X_i$ . Thus, if the age at the last observation is considered as the censoring time for each subject then the data for testing  $H_0$  can be regarded as a sample of paired survival data with a common censoring time.

There exist a number of procedures for testing  $H_0$ . Among these are procedures suggested by Woolson and Lachenbruch [1] who developed a family,  $\mathcal{C}_1$ , of score statistics for testing  $H_0$ . Popovich and Rao [2] proposed an alternative family of statistics,  $\mathcal{C}_2$ , for the same problem.

Statistics in  $\mathcal{C}_1$  or  $\mathcal{C}_2$  can be represented as linear combinations

$$T = L_u T_u + L_c T_c \tag{1}$$

where  $L_u$  and  $L_c$  are scalar coefficients (possibly random) and  $T_u$  and  $T_c$  are appropriately chosen statistics based on the uncensored and censored pairs, respectively. Dabrowska [3] used a counting process representation to derive the asymptotic relative efficiencies (AREs) of the tests in  $\mathcal{C}_1$ . Raychaudhuri and Rao [5] assumed a log-linear model and used Dabrowska’s approach to compare the efficacies of selected statistics in each of these two classes to the efficacies of the corresponding optimal statistics – statistics that maximize the efficacy within a class. On the basis of a simulation study, Raychaudhuri and Rao [4] concluded that the optimal statistics can have high efficiencies that depend on the heaviness of censoring and correlation between pairs.

Unfortunately, unlike the scores in the Woolson-Lachenbruch class of statistics, the coefficients in the linear combination defining the optimal statistic depend on the joint distribution of the survival and censoring times. Consequently, optimal coefficients can be specified only if one is willing to assume a form for this joint distribution.

In this article, we describe a method for determining the coefficients in the optimal statistics on the basis of assumptions that do not require specification of the form of the joint distribution of  $C_i$  and  $(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})$ . Let  $Z_i = Y_{2i} - Y_{1i}$ . Our approach assumes a model for the conditional distribution of the observed absolute difference conditional on the censoring pattern of  $X_i$ . That is, we assume a model for the conditional distribution of  $|Z_i|$  given the values of  $(\delta_{1i}, \delta_{2i})$ . As we shall see, the appropriateness of the assumed model can be assessed on the basis of observed data.

A brief description of the statistics in  $\mathcal{C}_1$  and  $\mathcal{C}_2$  with expressions of optimal statistics is presented in Sect. 2. Estimators of the optimal coefficients are proposed in Sect. 3, and Sect. 4 contains the results of a simulation study of the performance of statistics with estimated optimal weights. A real data example is presented in Sect. 5.

## 2 Efficacies of Statistics in $\mathcal{C}_1$ and $\mathcal{C}_2$

The pair  $(Y_{1i}, Y_{2i})$  will be said to be doubly censored, singly censored, or uncensored according as both in the pair are censored, exactly one in the pair is censored or none in the pair is censored. Following Dabrowska [3], let

$$\begin{aligned} B_1 &= \{i : Z_i > 0, (Y_{1i}, Y_{2i}) \text{ is an uncensored pair}\}, \\ B_2 &= \{i : Z_i < 0, (Y_{1i}, Y_{2i}) \text{ is an uncensored pair}\}, \\ B_3 &= \{i : Z_i > 0, (Y_{1i}, Y_{2i}) \text{ is a singly censored pair}\}, \\ B_4 &= \{i : Z_i < 0, (Y_{1i}, Y_{2i}) \text{ is a singly censored pair}\}, \\ B_5 &= \{i : (Y_{1i}, Y_{2i}) \text{ is a doubly censored pair}\}, \end{aligned}$$

and define the counting processes:

$$N_j(t) = \sum_{i=1}^n I\{|Z_i| \leq t, i \in B_j\} \quad j = 1, 2, 3, 4.$$

Clearly, the  $N_j(t)$  count the occurrences of uncensored and singly censored absolute differences in the interval  $(0, t]$ . For  $i = 1, 2$ , let  $J_{iu}(\cdot)$  and  $J_{ic}(\cdot)$  be given score functions defined on  $(0, 1)$ . Then the classes of statistics  $\mathcal{C}_1$  and  $\mathcal{C}_2$  introduced by Woolson and Lachenbruch [1] and Popovich and Rao [2] can be represented in the form of (1):

$$T_i = L_{iu}T_{iu} + L_{ic}T_{ic}, \quad i = 1, 2, \tag{2}$$

where

$$T_{iu} = \int J_{iu}(1 - \hat{S}_i)d(N_1 - N_2), \quad \text{and} \quad T_{ic} = \int J_{ic}(1 - \hat{S}_i)d(N_3 - N_4),$$

and  $\hat{S}_1(t)$  and  $\hat{S}_2(t)$  are the Kaplan-Meier [5] estimator of  $S(t) = P(|X_1 - X_2| > t)$  based on  $\{|Z_i| : i \in \cup_{j=1}^4 B_j\}$  and  $\{|Z_i| : i \in \cup_{j=1}^2 B_j\}$ , respectively. Special cases of (2) are:

1.  $L_{1u} = L_{1c} = 1$  and  $J_{1u}(v) = J_{1c}(v) = 1$ , in which case,  $T_1$  reduces to the Woolson-Lachenbruch sign statistic (WL sign statistic). In this case, the symbols  $T_{1su}$  and  $T_{1sc}$  will be used to denote  $T_{1u}$  and  $T_{1c}$ , respectively.
2.  $L_{1u} = L_{1c} = 1$  and  $J_{1u}(v) = v, J_{1c}(v) = 1/2(1 + v)$ , in which case  $T_1$  reduces to the Woolson-Lachenbruch Wilcoxon signed-rank statistic (WL Wilcoxon signed rank statistic). In this case, the symbols  $T_{1wu}$  and  $T_{1wc}$  will denote  $T_{1u}$  and  $T_{1c}$ , respectively.
3. When  $J_{2u}(v) = v, T_{2u}$  reduces to the Wilcoxon signed rank statistic calculated from the uncensored pairs. Hence, the symbol  $T_{2wu}$  will be used to denote  $T_{2u}$  in this case. We will use the symbol  $T_{2wc}$  to denote  $T_{2c}$ . Note that  $T_{2c}$  is the same as  $T_{1sc}$  in (2). Thus,  $T_{1sc}$  and  $T_{2wc}$  denote the same statistic.

Efficacies of the statistics in  $\mathcal{C}_1$  and  $\mathcal{C}_2$  are derived in Dabrowska [3] and Raychaudhuri and Rao [4]. For  $i = 1, 2$ , let  $\{T_{iu}^*, T_{ic}^*\}$  denote the standardized (under  $H_0$ ) versions of  $T_{iu}$  and  $T_{ic}$ , respectively. Define  $r_i = \frac{e_{iu}}{e_{ic}}$ , where  $e_{iu}$  and  $e_{ic}$  are the efficacies of  $T_{iu}^*$  and  $T_{ic}^*$ , respectively. Raychaudhuri and Rao [4] showed that the linear combination

$$T_i^* = L_{iu}^* T_{iu}^* + L_{ic}^* T_{ic}^*, \tag{3}$$

where

$$L_{iu}^* = \sqrt{\frac{r_i}{1+r_i}} \quad \text{and} \quad L_{ic}^* = \sqrt{\frac{1}{1+r_i}},$$

attains maximum efficacy in  $\mathcal{C}_i$ .

Assuming a log-linear model for survival times, Raychaudhuri and Rao [4] derived expressions for  $e_{iu}$  and  $e_{ic}$  and used them to compare the asymptotic relative efficiencies (AREs) of nine selected statistics. Three linear combinations of the two statistics in each of the three pairs  $(T_{1su}, T_{1sc})$ ,  $(T_{1wu}, T_{1wc})$ , and  $(T_{2wu}, T_{2wc})$  were studied. The coefficients in the linear combinations were:

$$\begin{aligned} L_{iu}^* &= \sqrt{\frac{r_i}{1+r_i}} & L_{ic}^* &= \sqrt{\frac{1}{1+r_i}}, & \text{(optimal weights)} \\ L_{iu}^* &= \frac{1}{\sqrt{2}}, & L_{ic}^* &= \frac{1}{\sqrt{2}}, & \text{(equal weights)} \\ L_{iu}^* &= \sqrt{\frac{\sigma_{iu}^2}{\sigma_{iu}^2 + \sigma_{ic}^2}}, & L_{ic}^* &= \sqrt{\frac{\sigma_{ic}^2}{\sigma_{iu}^2 + \sigma_{ic}^2}}, & \text{(proportional weights)} \end{aligned}$$

where  $\sigma_{iu}^2$  and  $\sigma_{ic}^2$  are the null variances of  $T_{iu}$  and  $T_{ic}$ , respectively. In the sequel, we shall use the symbols  $T_{1s-op}$ ,  $T_{1s-eq}$ ,  $T_{1s-pr}$ ,  $T_{1w-op}$ ,  $T_{1w-eq}$ ,  $T_{1w-pr}$ ,  $T_{2w-op}$ ,  $T_{2w-eq}$ ,  $T_{2w-pr}$  to denote these nine statistics.

Raychaudhuri and Rao [4] observed that the statistics in  $\mathcal{C}_2$  perform as well as those in  $\mathcal{C}_1$  and that (1) the efficacy of the optimal statistic in  $\mathcal{C}_2$  is higher than that of the optimal statistic in  $\mathcal{C}_1$ , and (2) the statistics in  $\mathcal{C}_2$  have the attractive property that for small sample sizes, they can be used to perform conditional distribution-free exact tests of  $H_0$ .

### 3 Estimating Optimal Weights

The optimal weights can be estimated on the basis of the following model for the conditional distributions of  $|Z_i|$ . Let  $h(t)$  and  $H(t)$  be the density and survival functions of a continuous random variable symmetrically distributed with median 0 and assume that

$$\Pr(|Z_i| \geq t \mid i \in B_j) = \frac{H(tv_j + \theta_j)}{H(\theta_j)}, \quad j = 1, \dots, 4, \tag{4}$$

where, for some  $-\infty < \theta < \infty$  and  $\nu > 0$ ,  $\theta_j = (-1)^j \theta$  for  $j = 1, \dots, 4$ ,  $\nu_j = 1$  for  $j = 1, 2$  and  $\nu_j = \nu$  for  $j = 3, 4$ . Let

$$F_j(t : \theta_j, \nu_j) = \Pr(|Z_i| \geq t, i \in B_j), \quad j = 1, \dots, 4.$$

Then  $F_j(0 : \theta_j, \nu_j) = \Pr(i \in B_j)$  and (4) can be expressed as:

$$F_j(t : \theta_j, \nu_j) = F_j(0 : \theta_j, \nu_j) \frac{H(t\nu_j + \theta_j)}{H(\theta_j)}, \quad j = 1, \dots, 4. \tag{5}$$

Equation (5) specifies a two parameter model for the conditional distribution of  $|Z_i|$  conditional on  $i \in B_j$ ,  $j = 1, \dots, 4$ . The parameter  $\theta$  can be interpreted as a measure of the degree of location shift in the survival functions of  $X_1$  and  $X_2$ . The null hypothesis  $H_0$  implies  $F_1(0 : \theta_1, \nu_1) = F_2(0 : \theta_2, \nu_2)$  and  $F_3(0 : \theta_3, \nu_3) = F_4(0 : \theta_4, \nu_4)$ . In addition, if one of the commonly used symmetric densities:

$$\begin{aligned} h(t) &= \frac{1}{2}e^{-|t|}, && \text{(double exponential),} \\ h(t) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}, && \text{(normal)} \\ h(t) &= \frac{e^{-t}}{(1 + e^{-t})^2}, && \text{(logistic)} \end{aligned} \tag{6}$$

is used in (5) then the null hypothesis  $H_0$  also implies  $\theta = 0$ . The parameter  $\nu$  describes how censoring affects the distributions of  $Z$ . If  $\nu = 1$ , the conditional distributions of censored and uncensored differences are the same.

In view of the wide variety of possible choices for  $H$ , the model in (5) provides a simple structure that is flexible enough to represent a variety of joint distributions of survival and censoring times. Indeed, the model (5) does not require the assumption of independent censoring – that  $(X_1, X_2)$  and  $C$  are independent. Thus, optimal properties of tests under model (5) will hold as long as the conditional distributions have the required form. Furthermore, a practical advantage of estimating optimal weights using (5) is that the observed data can be used, as described in Appendix 1, for a graphical check of the appropriateness of a selected  $H$ .

Let  $N_j$  be the number of elements in  $B_j$ ,  $N = \sum_{j=1}^5 N_j$ ,  $\hat{\alpha}_{jr}$  be as defined in Appendix 1, and  $|Z|_{(jr)}$ ,  $r = 1, \dots, N_j$ , denote the ordered absolute  $Z$ s in  $B_j$ . In Appendix 1 it is shown that if the model (5) is appropriate for the data then the plots of  $\{(H^{-1}(\hat{\alpha}_{jr}), |Z|_{(jr)}) : r = 1, \dots, N_j\}$ ,  $j = 1, 2, 3, 4$ , should approximate two sets of parallel straight lines. The plot of points in  $B_1$  and  $B_2$  should be parallel with intercepts  $\theta$  and  $-\theta$  and a common slope of 1. The plots of the points in  $B_3$  and  $B_4$  should be parallel with intercepts  $\theta$  and  $-\theta$  and a common slope equal to  $\nu$ .

Furthermore, let  $g(\theta)$  be a function with derivative  $g'(\theta)$  at  $\theta = 0$ . It can be shown (see Appendix 2) that if the assumption

$$\frac{F_1(0 : \theta, \nu)}{F_2(0 : \theta, \nu)} = \exp \{g(\theta) - g(0)\}, \quad \frac{F_3(0 : \theta, \nu)}{F_4(0 : \theta, \nu)} = \exp \{\eta + g(\theta) - g(0)\}, \tag{7}$$

is added to the assumptions in (5), then the efficacy ratios to determine optimal weights can be completely specified in terms of  $F_j = F_j(0 : 0, \nu_j)$ ,  $\nu$ ,  $h(\cdot)$ , and  $g'(\theta)$ . Before proceeding further, it should be noted that  $\frac{F_1(0:\theta, \nu)}{F_2(0:\theta, \nu)}$  and  $\frac{F_3(0:\theta, \nu)}{F_4(0:\theta, \nu)}$  are the odds of observing a positive  $Z$  with respect to a negative  $Z$  given that  $Z$  corresponds to an uncensored pair and a singly censored pair, respectively. Thus, the conditions expressed in (7) is an assumption about the forms of the odds of observing a positive difference in the uncensored and censored pairs.

We propose estimating optimal weights as follows.

1. Estimate  $F_j$  with its unbiased estimator  $N_j/N$ .
2. Let  $\bar{Z}_j$  be the mean of the  $Z_i$  in  $B_j$ . In Appendix 3, we use (5) to show that

$$E(\bar{Z}_1) = \nu E(\bar{Z}_3) \text{ and } E(\bar{Z}_2) = \nu E(\bar{Z}_4). \tag{8}$$

From (8), an intuitively reasonable estimator for  $\nu$  is

$$\hat{\nu} = \begin{cases} \frac{\bar{Z}_1}{\bar{Z}_3} & \text{if } N_2 + N_4 = 0 \text{ and } \min\{N_1, N_3\} > 0, \\ \frac{\bar{Z}_2}{\bar{Z}_4} & \text{if } N_1 + N_3 = 0 \text{ and } \min\{N_2, N_4\} > 0, \\ \frac{1}{2} \left( \frac{\bar{Z}_1}{\bar{Z}_3} + \frac{\bar{Z}_2}{\bar{Z}_4} \right) & \text{if } \min\{N_1, N_2, N_3, N_4\} > 0, \\ 1 & \text{otherwise.} \end{cases} \tag{9}$$

3. Select an appropriate form of  $h(\cdot)$ . Any symmetric density on  $(-\infty, \infty)$  is a possible choice for  $h(\cdot)$ . Three such densities are described in (6). Sometimes, a nonstandard version of  $h(\cdot)$  may be more suitable for describing the conditional distributions of  $|Z_i|$ . The appropriateness of a selected  $h(\cdot)$  can be checked using the graphical procedure described earlier.
4. Specify a value of  $g'(\theta)$ . As the optimal weights depend only on the derivative of  $g(\theta)$  at  $\theta = 0$ , the estimated optimal statistic remains the same for all  $g(\cdot)$  with a given value of  $g'(\theta)$ . From (7), it follows that specifying  $g'(\theta)$  is equivalent to specifying the rate at which the odds of a positive  $Z$  in the censored and the uncensored pairs change in the neighborhood of the null hypothesis  $H_0 : \theta = 0$ . Higher values of  $g'(\theta)$  correspond to more severe departures from  $H_0$ .
5. Use the computing formulas in Appendix 2 to estimate optimal weights.

## 4 Simulations

Two simulation studies were performed to evaluate the powers of linear rank tests with optimal weights estimated as described in Sect. 3. In the first study, we compared nine statistics,  $T_{1s-j}$ ,  $T_{1w-j}$ ,  $T_{2w-j}$ ,  $j = eq, pr, op$ , resulting from using equal weights, proportional weights or optimal weights in the linear combination (3). We assumed that the  $Z_i$  satisfy conditions (5) and (7).

The nine statistics in the first study utilize information in within pair differences, whereas the Akritas [7, 8] (AK) and paired Prentice-Wilcoxon [6, 8] (PPW) statistics utilize scores assigned to the pooled sample  $\{X_{1i}, X_{2i} : i = 1, \dots, n\}$ . In the second study, we compared the powers of the statistics evaluated in the first study with the powers of AK and PPW statistics. For the purpose of comparison with existing results, the second study was performed using conditions similar to those in Woolson and O’Gorman [8], where the tests based on AK and PPW statistics were compared to several tests for censored paired data.

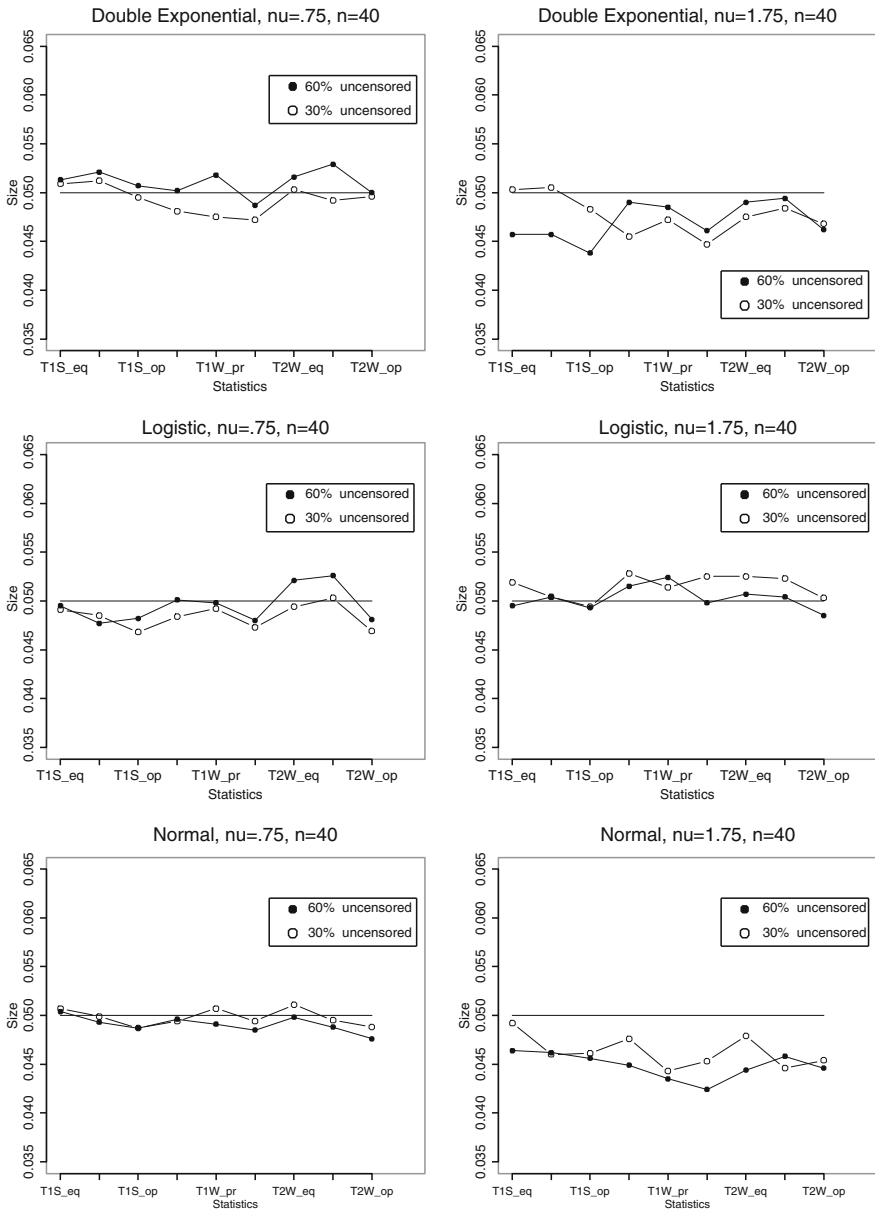
In the first simulation, we estimated the size and power of the nine tests at .05 level based on 10,000 samples of sizes  $n = 40$  and  $n = 100$ . The samples were generated under different situations determined by the density  $h(\cdot)$ , the parameters  $\theta$  and  $\nu$ , and the censoring pattern. The three densities listed in (6) were used. For the odds ratio in (7), we selected  $g(\theta) = .6\theta$  ( $g'(0) = .6$ ), and estimated the size at  $\theta = 0$  and powers at  $\theta = .5, 1, 1.5, 2$ . Three values,  $\nu = .75, 1, 1.75$ , were included in the study. As  $\nu < 1$  implies that the mean and standard deviation of the null distribution of a censored  $Z$  is larger than that of the corresponding uncensored  $Z$ , the value  $\nu = .75$  was selected to represent this case. The value  $\nu = 1.75$  was selected to represent the case where a censored  $Z$  has smaller mean and standard deviation than that of an uncensored  $Z$ . Two censoring patterns were investigated: (a) 30% uncensored with  $\Pr(i \in B_1 \cup B_2) = .3$ ,  $\Pr(i \in B_3 \cup B_4) = .6$ ; and (b) 60% uncensored with  $\Pr(i \in B_1 \cup B_2) = .6$ ,  $\Pr(i \in B_3 \cup B_4) = .3$ . The following is a summary of the conclusions from the results of the first study.

All nine statistics held their levels fairly well under all conditions investigated in the study. Figure 1 shows the estimated levels under six different conditions.

Figure 2 provides a comparison of the powers under three distributions and two censoring patterns when  $\theta = 1$ ,  $\nu = 1.75$ , and  $n = 100$ . The differences between powers exhibited similar patterns in all cases considered.

The tests based on sign statistics had substantially lower powers than the powers of other tests. In every instance, the use of optimal weights increased power and the test based on  $T_{1w-op}$  had maximum power with the test based on  $T_{2W-op}$  not too far behind. Furthermore, all tests had maximum power when  $h(t)$  was normal, with the power of the corresponding test when  $h(t)$  is logistic following close behind. Also, an increase in the expected proportion of uncensored observations resulted in an increase of the power of every test considered.

The effect of the value of  $\nu$  (an indicator of the difference between the distributions of the censored and uncensored  $Z$ 's) can be seen in Figure 3, where we show the estimated powers for  $\nu = .75, 1, 1.75$  at two alternative hypotheses:  $\theta = .5$  and



**Figure 1** Estimated levels of .05-level tests (10,000 Simulations)

$\theta = 1$ , when the expected proportion of uncensored pairs is 30%, the density is logistic, and  $n = 100$ . It is clear that  $\nu$  has very little effect on the power of these tests, implying that precise specification of  $\nu$  is not a critical issue in determining optimal coefficients.

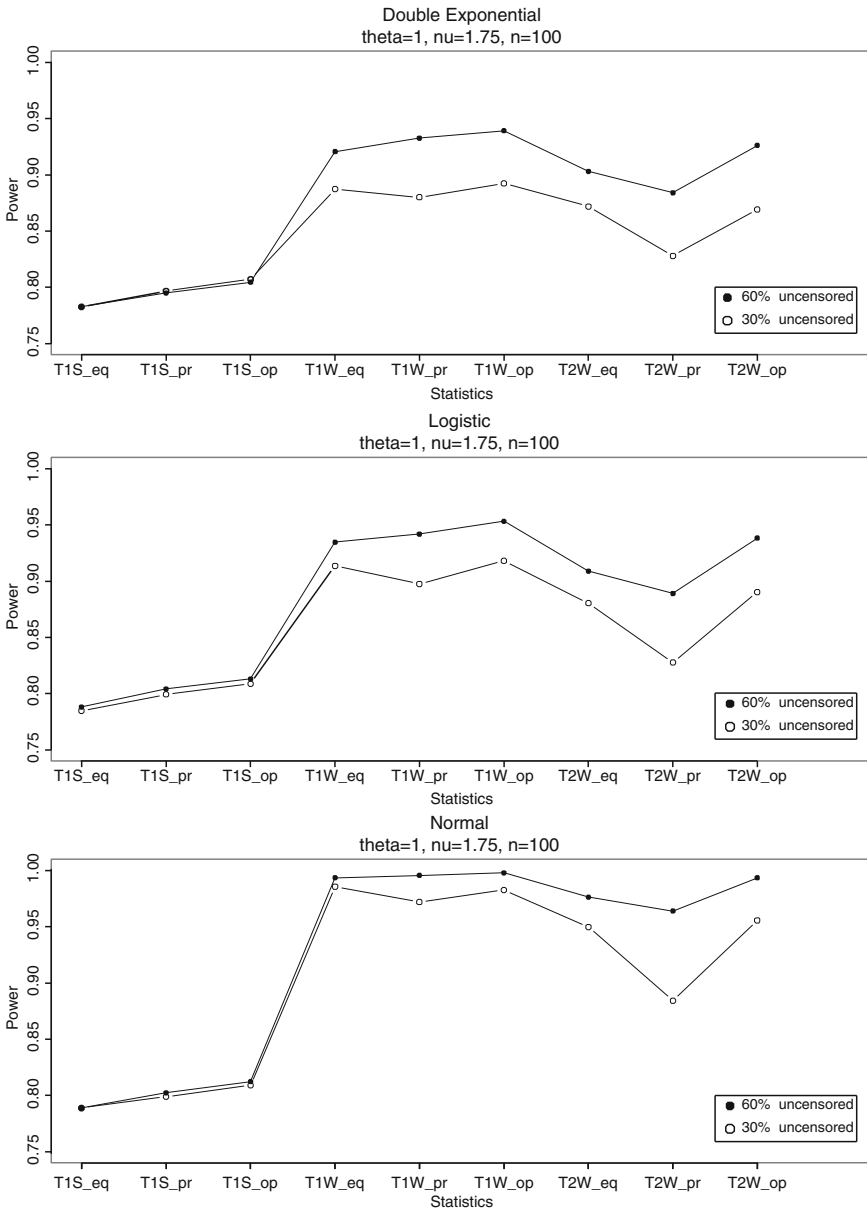


Figure 2 Estimated powers for different censoring rates and  $h(\cdot)$  (10,000 simulations)

Figure 4 plots the powers of the nine tests under three distributions and two sample sizes when  $\theta = 1, \nu = 1$  and the expected proportion of uncensored pairs equals 30%. The plotting symbols are the positions of the statistics in the horizontal axes of the panels in Figures 2 or 3.



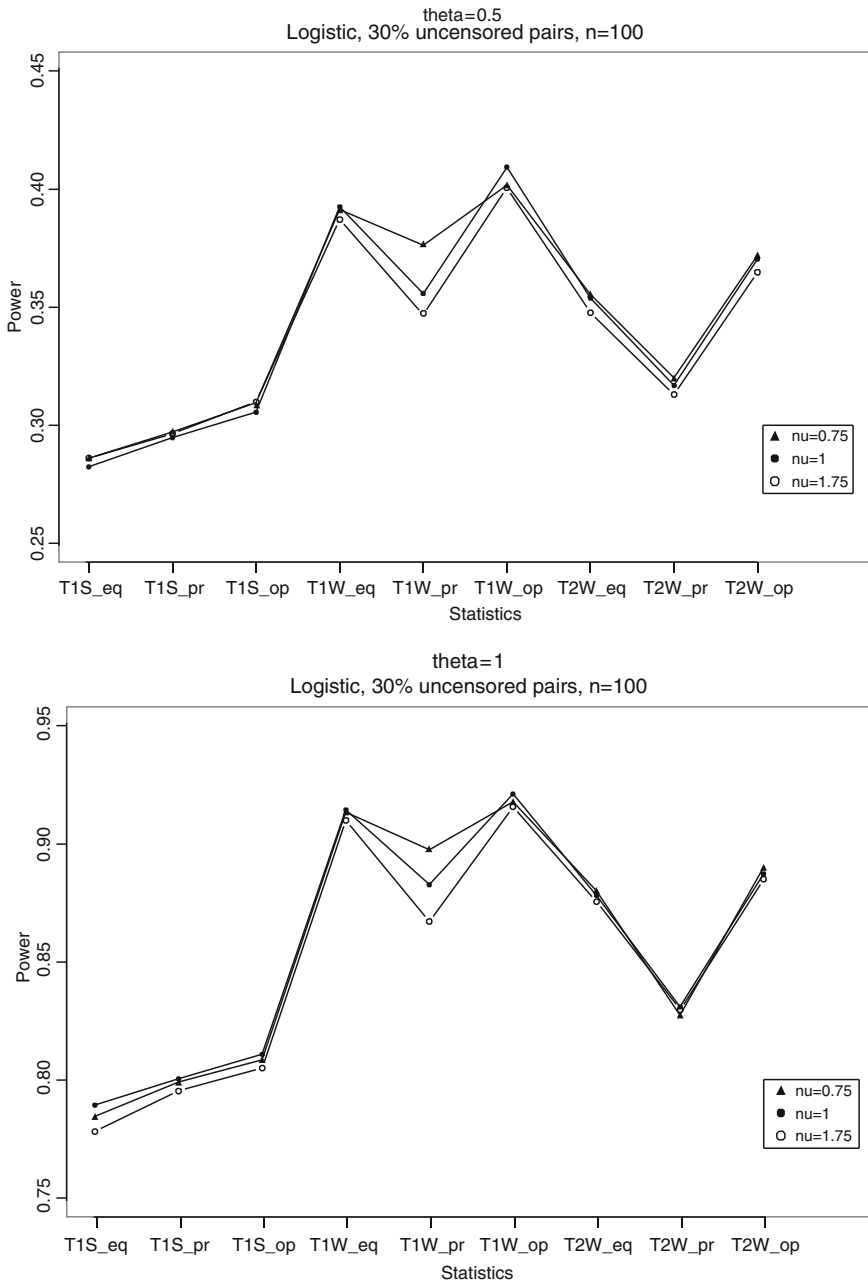
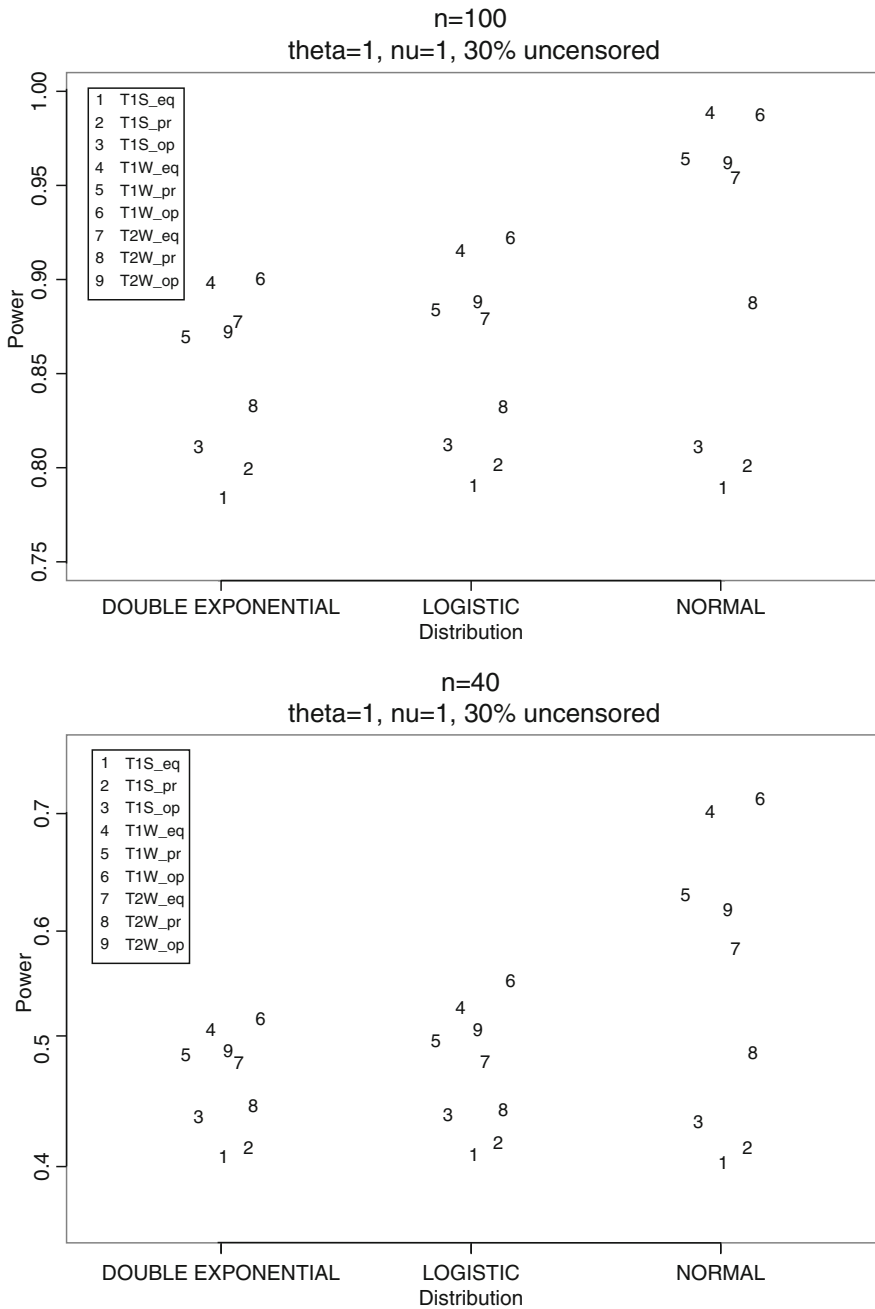


Figure 3 Estimated powers for different  $\nu$  and  $h(\cdot)$  (10,000 simulations)



**Figure 4** Estimated powers for different sample sizes and  $h(\cdot)$  (10,000 simulations)

As one would expect, Figure 4 shows that the sample size has a big impact on power. In addition, the heaviness of the tail of  $h(\cdot)$  also influences power. The normal distribution, which has the lightest tail among the three distributions considered, corresponds to highest power. The lowest powers are typically associated with double exponential, which has the heaviest tail among the three distributions studied.

In the second simulation, we compared the powers of AK and PPW tests under three situations. For the first situation, the survival and censoring times were generated using the additive model:  $X_{1,i} = U_i + U_{200+i}$ ,  $X_{2,i} = \theta + U_{100+i} + U_{200+i}$ ,  $C_i = c(\theta)U_{300+i}$ ;  $i = 1, 2, \dots, 100$ . The  $U$ 's are independent identically distributed exponential variates with mean one, and the constant  $c(\theta)$  was chosen so that we can expect 70% uncensored pairs. The  $U_{200+i}$  terms provide correlation among the paired observations. For this model,  $X_{2,i} - X_{1,i}$  has a shifted double exponential distribution with shift  $\theta$ .

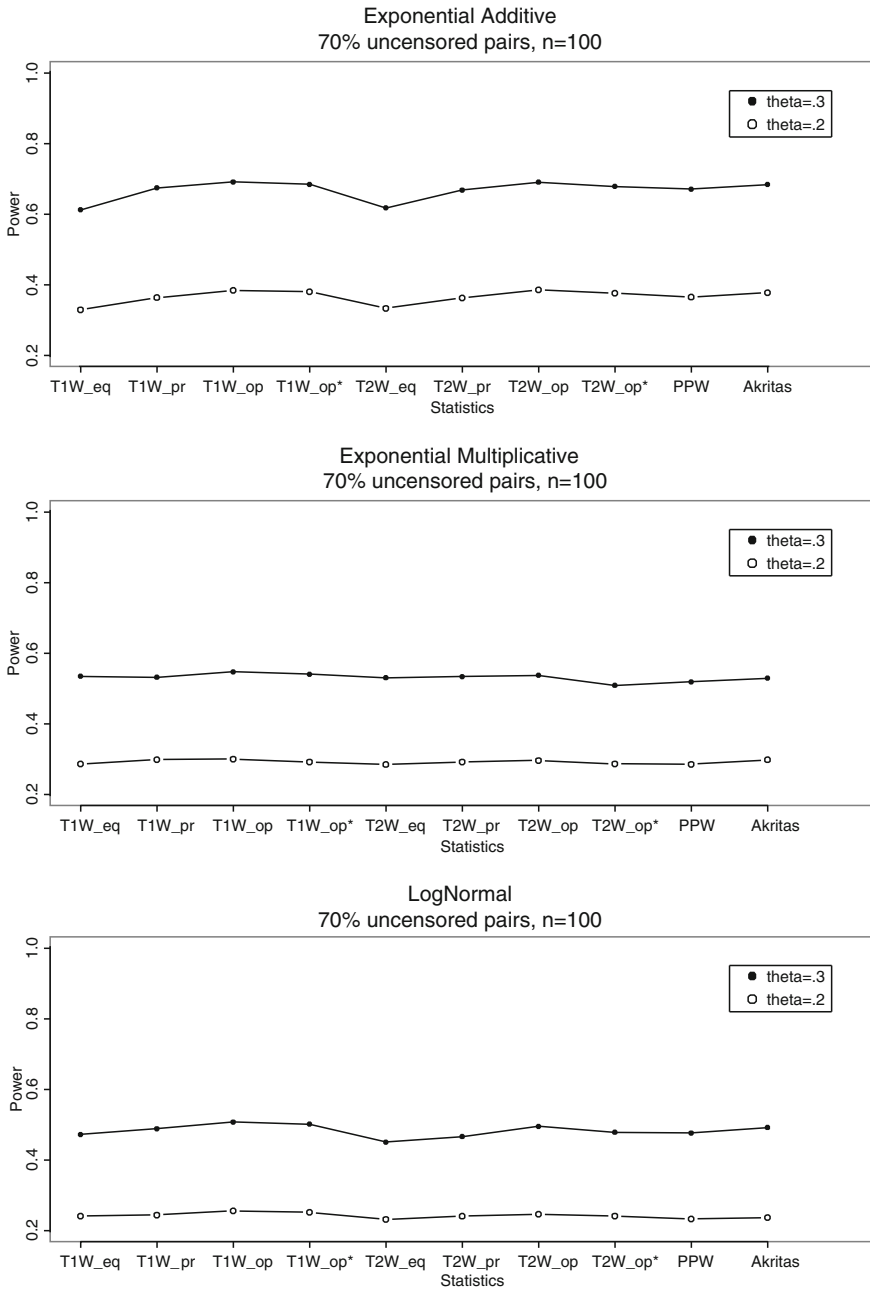
The second situation utilized a multiplicative model, e.g.,  $X_{1,i} = [U_i U_{200+i}]^s$ ,  $X_{2,i} = \exp(\theta)[U_{100+i} U_{200+i}]^s$ ,  $C_i = [c(\theta)U_{300+i}]^s$ , where the  $U$ 's are independent identically distributed exponential variates with mean one. We used  $s = \sqrt{2/\text{var}(\text{logistic})} = .780$ , such that  $\log X_{2,i} - \log X_{1,i}$  has a logistic distribution with location parameter  $\theta$  and variance 2.

For the third situation, we generated the survival and censoring times under the model:  $X_{1,i} = U_i U_{200+i}$ ,  $X_{2,i} = U_{100+i} U_{200+i}$ ,  $C_i = U_{300+i}$ , where  $U_i, U_{200+i} \sim \text{Lognormal}(0, 1)$ ,  $U_{100+i} \sim \text{Lognormal}(\theta, 1)$ ,  $U_{300+i} \sim \text{Lognormal}(c(\theta), 2)$ , and the constant  $c(\theta)$  was chosen so that the expected proportion of uncensored pairs equals 70%. In this situation  $\log X_{2,i} - \log X_{1,i}$  has a normal distribution with location parameter  $\theta$  and variance 2.

Figure 5 shows the powers (estimated on the basis of 2,000 simulations) of ten tests in each of the three situations described earlier.

In addition to the AK and PPW tests, the figure shows the powers of six tests based on the statistics  $\{T_{i_w-j} : i = 1, 2; j = eq, pr, op\}$ . Because of their poor performance in the first study, the second study did not include the three tests based on sign statistics. Instead, two additional tests based on statistics with optimal weights calculated using an incorrect  $h(\cdot)$  were evaluated. In Figure 5, these statistics are denoted as  $T_{i_w-op*}$ . The weights for these statistics were estimated using logistic density in situation 1 and situation 3, and normal density in situation 2.

In every situation considered, there was very little difference between the powers of the tests based on  $T_{i_w-op}$  and  $T_{i_w-op*}$ . Thus, it appears that the performance of an optimal test does not depend heavily on the choice of a correct  $h(\cdot)$ . Also, the optimal tests performed as well as the AK and PPW tests. The fact that the performance of the optimal tests when the data were generated using model (5) is similar to their performance when data were generated assuming independent censoring model, indicates that these tests may be preferable over AK and PPW tests, particularly if model (5) fits the data well and there is reason to doubt the independent censoring assumption.



**Figure 5** Estimated powers of statistics with optimal coefficients compared with AK and PPW statistics (2,000 simulations)

### 5 Example

In this section, we will illustrate the tests discussed in this paper with the FGRP data on two variables – albumin level (PCL13) and hemoglobin level (HGB) – taken from the SMAC-23 data. As in the introductory section, we define  $X_1$  and  $X_2$  as the ages at which a subject’s PCL13 and HGB values reached “abnormal” levels for the first time during the observation period. The age at the last observation will be treated

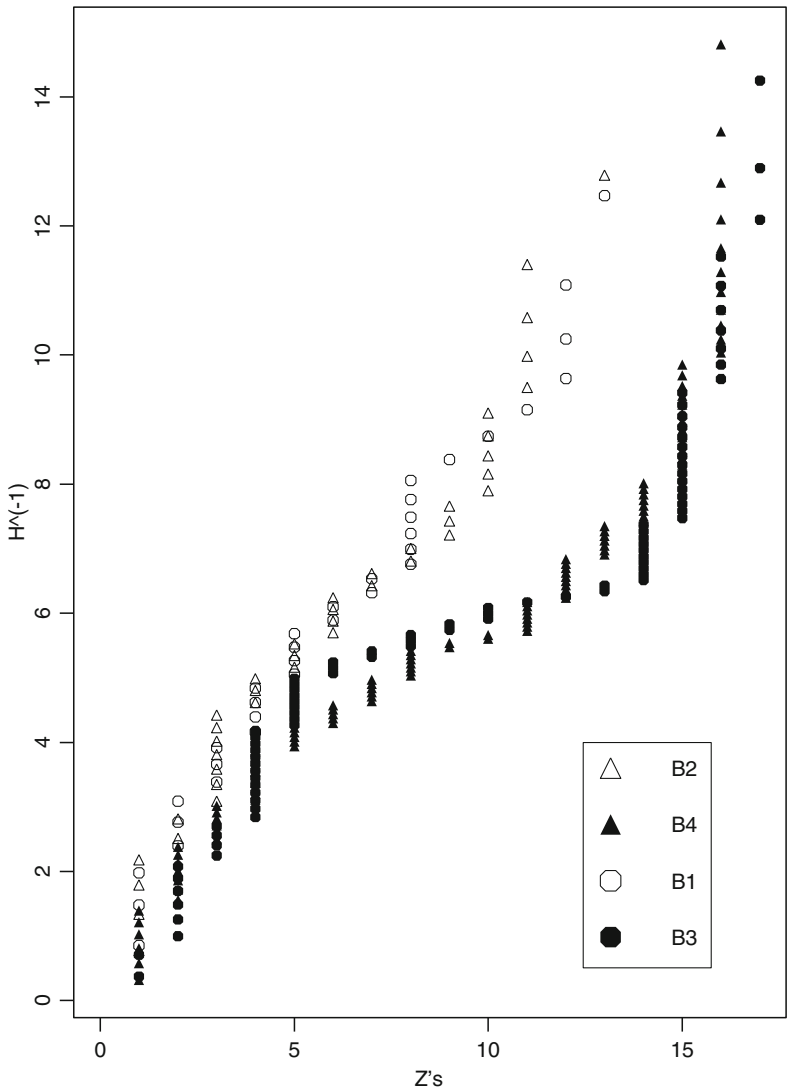


Figure 6 Q-Q plot for logistic density

as the censoring time for each subject. We will assume the model in (5) and test the null hypothesis that the probability distributions of the ages at which PCL13 and HGB reached “abnormal” levels for the first time are identical against a two sided alternative that the distributions are not the same. This will help answer the question “Which, if any, of the two abnormalities is likely to occur first?”

In our dataset, there were 34, 40, 86, and 115 pairs in the sets  $B_1, \dots, B_4$ , respectively, so we estimate  $F_1$  and  $F_3$  by .12 and .31. Secondly, we estimate  $\nu$  by  $\frac{1}{2} \left( \frac{\bar{Z}_1}{\bar{Z}_3} + \frac{\bar{Z}_2}{\bar{Z}_4} \right) = .65$ . To calculate an optimal statistic, one needs to select a density  $h(\cdot)$ . Figure 6 shows the Q–Q plot of a logistic distribution with location parameter 5.6 and scale parameter 1.9. The s-shaped pattern for censored pairs indicate a lack of fit of the model for the censored differences. However, since the performances of the optimal tests are not very sensitive to correct choice of  $h(\cdot)$ , we decided that the selected logistic distribution is adequate for purpose of illustrating the use of optimal tests. The calculated values of the optimal statistics are  $T_{1W} = 2.18$ , and  $T_{2W} = 2.17$ , both correspond to a two-sided  $p$ -value of 0.03. Hence, there was statistically significant evidence that the probability distributions of the ages at which PCL13 and HGB reached “abnormal” levels for the first time are not the same.

## 6 Conclusions

A model for conditional distributions of the absolute differences between randomly right censored paired survival times is proposed for estimation of optimal coefficients for two classes of statistics considered by Raychaudhuri and Rao [4]. The model has a simple structure and is flexible enough to represent a wide variety of conditional distributions of the observed differences between censored pairs. Simulation studies indicate that no single test is the best test in all situations considered, but the tests with estimated weights compare well with such other recommended tests as AK and PPW tests for paired censored survival data. An advantage of the tests with estimated optimal weights is that the tests can be used in cases where the assumption of independent censoring is questionable but the conditional model seems reasonable.

**Acknowledgment** The authors wish to thank Dr. William E. Hale, former Director of the FGRP and the Morton Plant Health System, and Dr. Ron G. Marks, formerly of the Division of Biostatistics of the Department of Statistics, University of Florida for their support of this study.

## References

- [1] Woolson, R. F., and Lachenbruch, P. A. (1980), “Rank Tests for Censored Matched Pairs”, *Biometrika*, 67, 597–606.
- [2] Popovich, E. A., and Rao, P. V. (1985), “Conditional Tests for Censored Matched Pairs”, *Communications in Statistics*, 14(9), 2041–2056.

[3] Dabrowska. D. M. (1990), “Signed-Rank Tests for Censored Matched Pairs”, *Journal of the American Statistical Association*, 85, 478–485.  
 [4] Raychaudhuri, A. and Rao, P. V. (1996), “Efficacies of Some Rank Tests for Censored Bivariate Data”, *Nonparametric Statistics*, 6, 1–11.  
 [5] Kalbfleisch, J. D., and Prentic, R. L. (1980), “The Statistical Analysis of Failure Time Data”, *John Wiley and Sons*, NY.  
 [6] O’Brien, P. C., and Fleming, T. R. (1987), “A paired Prentice-Wilcoxon Test For Censored Paired Data”, *Biometrics*, 43, 169–180.  
 [7] Akritas, M. G. (1992), “Rank Transform Statistics with Censored Data”, *Statistics and Probability Letters*, 13, 209–221.  
 [8] Woolson, R. F., and O’Gorman, T. W. (1992), “A Comparison of Several Tests for Censored Paired Data”, *Statistics in Medicine*, 11, 193–208.

## Appendix 1

### Graphical Plots for Selecting $H$

Let  $|Z|_{(jr)}, r = 1, \dots, N_j$ , denote the ordered  $|Z|$ 's in  $B_j$ . Using the approximations,

$$\Pr(|Z_i| \geq |Z|_{(jr)} \mid i \in B_j) \doteq \frac{N_j - r + 1}{N_j + 1},$$

we have the relationships:

$$\begin{aligned} \frac{N_1 - r + 1}{N_1 + 1} &\doteq \frac{H(|Z|_{(1r)} - \theta)}{H(-\theta)} & r = 1, \dots, N_1, \\ \frac{N_2 - r + 1}{N_2 + 1} &\doteq \frac{H(|Z|_{(2r)} + \theta)}{H(\theta)} & r = 1, \dots, N_2, \\ \frac{N_3 - r + 1}{N_3 + 1} &\doteq \frac{H(v|Z|_{(3r)} - \theta)}{H(-\theta)} & r = 1, \dots, N_3, \\ \frac{N_4 - r + 1}{N_4 + 1} &\doteq \frac{H(v|Z|_{(4r)} + \theta)}{H(\theta)} & r = 1, \dots, N_4. \end{aligned} \tag{10}$$

Let

$$\begin{aligned} \alpha_{1r} &= \frac{N_1 - r + 1}{N_1 + 1} H(-\theta), \\ \alpha_{2r} &= \frac{N_2 - r + 1}{N_2 + 1} H(\theta), \\ \alpha_{3r} &= \frac{N_3 - r + 1}{N_3 + 1} H(-\theta), \\ \alpha_{4r} &= \frac{N_4 - r + 1}{N_4 + 1} H(\theta). \end{aligned} \tag{11}$$

Then

$$\begin{aligned}
 H^{-1}(\alpha_{1r}) &\doteq |Z|_{(1r)} - \theta & r = 1, \dots, N_1, \\
 H^{-1}(\alpha_{2r}) &\doteq |Z|_{(2r)} + \theta & r = 1, \dots, N_2, \\
 H^{-1}(\alpha_{3r}) &\doteq v|Z|_{(3r)} - \theta & r = 1, \dots, N_3, \\
 H^{-1}(\alpha_{4r}) &\doteq v|Z|_{(4r)} + \theta & r = 1, \dots, N_4.
 \end{aligned}
 \tag{12}$$

Thus, if we replace  $\alpha_{jr}$  in (12) with an appropriate estimator  $\hat{\alpha}_{jr}$ , the plots of  $\{(H^{-1}(\hat{\alpha}_{jr}), |Z|_{(jr)}) : r = 1, \dots, N_j\}, j = 1, 2, 3, 4$ , can be used to check the adequacy of the selected  $H$ .

The  $\alpha_{jr}$ 's can be estimated by replacing  $\theta$  in (11) with an estimator  $\hat{\theta}$ . If the density  $h(\cdot)$  has a monotone hazard function, that is if  $h(t)/H(t)$  is monotone, then it is easy to show that  $H(t + \theta)/H(\theta)$  is a monotone function of  $\theta$  and it has a uniquely defined inverse. Thus, each equation in (10) can be solved to obtain a corresponding estimator  $\hat{\theta}_{jr}$ . Any measure of the location of  $\hat{\theta}_{jr}$ 's (e.g. mean, median, etc.) provides a reasonable estimator of  $\theta$ .

## Appendix 2

### Efficacy Ratios

This section provides the efficacy ratios needed for determining the optimal weights.

Let  $F_j(t) = F_j(t : 0, v_j)$  and  $f_j(t) = \frac{d}{dt} F_j(t)$ . Thus  $F_j(t)$  and  $f_j(t)$  denote, respectively, the null ( $\theta = 0$ ) sub-survival and density functions of the  $|Z_i|$  in  $B_j$ . Also, let  $F_j = F_j(0)$ ,  $F(t) = \sum_{j=1}^4 F_j(t)$ , and  $k(t) = g'(0) - 2h(0)/H(0) - 2h'(t)/h(t)$ . Then, for the model in (5), equations (4) and (6) in Raychaudhuri and Rao [4] can be used to express the efficacy ratios as:

$$r_{1s} = \frac{I_1}{I_2} = \frac{F_1}{F_3} \qquad r_{1w} = \frac{I_4^2 I_7}{I_6^2 I_5} \qquad r_{2w} = \frac{3I_3^2 F_3}{I_2^2 F_1}$$

where

$$\begin{aligned}
 I_1 &= \int_0^\infty k(t) f_1(t) dt, & I_2 &= \int_0^\infty k(tv) f_3(t) dt, \\
 I_3 &= \int_0^\infty k(t) \left[ 1 - \frac{F_1(t)}{F_1} \right] f_1(t) dt, & I_4 &= \int_0^\infty k(t) [1 - F(t)] f_1(t) dt, \\
 I_5 &= 2 \int_0^\infty [1 - F(t)]^2 f_1(t) dt, & I_6 &= \int_0^\infty k(tv) [1 - \frac{1}{2} F(t)] f_3(t) dt, \\
 I_7 &= 2 \int_0^\infty [1 - \frac{1}{2} F(t)]^2 f_3(t) dt.
 \end{aligned}$$



The expression for  $k(t)$  is derived based on Sect. 2.2 of Debrowska [3], which implies that  $k(t) = \frac{d}{d\theta} \left\{ \frac{f_{1\theta}(t)}{f_{2\theta}(t)} \right\} |_{\theta=0}$ , where  $f_{j\theta}(t) = \frac{d}{dt} F_j(t : \theta, v_j)$ .

Therefore, under conditions (5) and (7), the integrals  $I_1 \dots I_7$  can be expressed as follows.

$$I_1 = g'(0)F_1,$$

$$I_2 = g'(0)F_3,$$

$$I_3 = I_1 - F_1[\alpha/2 + 2r_2(1)],$$

$$I_4 = I_1 - 2F_1\alpha[F_1r_1(1) + F_3r_1(v)] + 4F_1[F_1r_2(1) + F_3r_2(v)],$$

$$I_5 = 2F_1 + 8F_1[F_1^2r_3(1) + F_3^2r_3(v)] - 8F_1[F_1r_1(1) + F_3r_1(v)] + 16F_1^2F_3r_4(v),$$

$$I_6 = I_2 - F_3\alpha[F_1r_1(1/v) + F_3r_1(1)] + 2F_3[F_1r_2(1/v) + F_3r_2(1)],$$

$$I_7 = 2F_3 + 2F_3[F_1^2r_3(1/v) + F_3^2r_3(1)] - 4F_3[F_1r_1(1/v) + F_3r_1(1)] + 4F_1F_3^2r_4(1/v),$$

where  $\alpha = g'(0) - 2h(0)/H(0)$ ,  $\phi(t) = h'(t)/h(t)$ , and

$$r_1(v) = \int_0^\infty H(tv)h(t)dt/H(0)^2,$$

$$r_2(v) = \int_0^\infty \phi(t)H(tv)h(t)dt/H(0)^2,$$

$$r_3(v) = \int_0^\infty H(tv)^2h(t)dt/H(0)^3,$$

$$r_4(v) = \int_0^\infty H(t)H(tv)h(t)dt/H(0)^3.$$

Thus, the integrals  $I_1 \dots I_7$  are easily evaluated once we determine  $r_j(v)$  for  $j = 1, 2, 3, 4$ . Since the  $r_j(v)$  have the general form:

$$\int_0^\infty A(t)h(t)dt = - \int_0^\infty A[H^{-1}\{H(t)\}]dH(t) = \int_0^{H(0)} A[H^{-1}(y)]dy,$$

a grid-search procedure can be used for rapid calculation of  $\{r_j(v) : j = 1, 2, 3, 4\}$ .

### Appendix 3

#### *Proof of (8)*

We will establish the relationship:  $E(\bar{Z}_1) = \nu E(\bar{Z}_3)$ . The proof of  $E(\bar{Z}_2) = \nu E(\bar{Z}_4)$  is similar.

$$\begin{aligned}
 E(\bar{Z}_1) &= E(Z_i \mid i \in B_1) = \frac{1}{F_1(0; \theta_1, \nu_1)} \int_0^\infty t dF_1(t; \theta_1, \nu_1) \\
 &= \frac{1}{H(-\theta)} \int_0^\infty th(t - \theta) dt \\
 &= \frac{\nu^2}{H(-\theta)} \int_0^\infty uh(u\nu - \theta) du \\
 &= \nu E(\bar{Z}_3).
 \end{aligned}$$

# Connections Between Bernoulli Strings and Random Permutations

Jayaram Sethuraman<sup>†</sup> and Sunder Sethuraman<sup>\*</sup>

*Dedicated to the memory of Professor Alladi Ramakrishnan.*

*We dedicate this paper to the memory of Professor Alladi Ramakrishnan, the founder of the world renown Mathematical Sciences Institute of Chennai. We fondly continue to cherish our memories of several contacts at a personal level with Professor Ramakrishnan.*

**Summary** A sequence of random variables, each taking only two values “0” or “1,” is called a Bernoulli sequence. Consider the counts of occurrences of strings of the form  $\{11\}, \{101\}, \{1001\}, \dots$  in Bernoulli sequences. Counts of such Bernoulli strings arise in the study of the cycle structure of random permutations, Bayesian nonparametrics, record values etc.

The joint distribution of such counts is a problem worked on by several researchers. In this paper, we summarize the recent technique of using conditional marked Poisson processes which allows to treat all cases studied previously. We also give some related open problems.

**Mathematics Subject Classification (2000)** Primary 60C05; Secondary 60K99

**Key words and phrases** Bernoulli · Cycles · Strings · Spacings · Nonhomogeneous · Poisson processes · Random permutations · Records

---

Research partially supported by ARO-W911NF-09-1-0338<sup>†</sup> and NSF-DMS 0906713<sup>\*</sup>. Approved for public release, distribution unlimited.

J. Sethuraman

Department of Statistics, Florida State University, Tallahassee, FL 32306, USA

e-mail: [sethu@stat.fsu.edu](mailto:sethu@stat.fsu.edu)

S. Sethuraman

Department of Mathematics, 396 Carver Hall, Iowa State University, Ames, IA 50011, USA

e-mail: [sethuram@iastate.edu](mailto:sethuram@iastate.edu)

## 1 Introduction

Consider a sequence of independent Bernoulli random variables  $Y_k$  where  $P(Y_k = 1) = 1 - P(Y_k = 0) = 1/k$  for  $k \geq 1$ . We will call such a sequence as a  $\text{Bern}(1, 0)$  sequence. Such a sequence notes the outcomes, “1” for a success and “0” for a failure, in an experiment conducted over times  $k = 1, 2, \dots$ . Notice that there is an infinite number of successes in the sequence, that is  $\sum_{k \geq 1} Y_k = \infty$  a.s. since  $\sum_{k \geq 1} E[Y_k] = \infty$ . However, the number,  $Z_1$ , of consecutive pairs of successes, or strings  $\{11\}$ , is a.s. finite since

$$E(Z_1) = \sum_{k \geq 1} E[Y_k Y_{k+1}] = \sum_{k \geq 1} \frac{1}{k(k+1)} = 1.$$

As an illustration of counting strings of the form  $\{11\}$ , we see that there are five strings of the form  $\{11\}$  in the truncated sequence  $\{01011101111\}$ . What can one say about the distribution of  $Z_1$ ?

Persi Diaconis, around 1996, surprisingly recognized that  $Z_1$  is distributed as a Poisson random variable with mean 1! Several studies of  $Z_1$  and related counts of other strings followed from this observation, which became the subject of friendly mathematical conversation. In fact, we learned of the problem from Krishna Athreya, who heard it during the course of a dinner at a conference.

This topic can be generalized. For  $m \geq 2$ , let  $Z_m = \sum_{k \geq 1} X_k [\prod_{l=1}^{m-1} (1 - X_{k+l})] X_{k+m}$ , be the count of strings where a success is followed by exactly  $m - 1$  failures before the next success, that is the number of strings of the form  $\{1 \underbrace{0 \dots 0}_{m-1} 1\}$ .

Analogous to  $Z_1$ , all the counts  $Z_m$  for  $m \geq 2$  are finite a.s. Intriguingly, the counts  $\mathbf{Z} = \{Z_m\}_{m \geq 1}$  turn out to be independent random variables, and the distribution of  $Z_m$  is Poisson with mean  $1/m$  for  $m \geq 1$ .

How to explain this phenomena, and how robust and relevant is it? Consider the situation where the success probabilities are “perturbed” in certain ways, that is when  $X_1, X_2, \dots$  are independent with Bernoulli distributions satisfying  $P(X_k = 1) = a/(a + b + k - 1)$  for  $a > 0$ ,  $b \geq 0$ , and  $k \geq 1$ . We will call such a Bernoulli sequence as a  $\text{Bern}(a, b)$  sequence. In this case also, it turns out the joint distribution of the counts  $\mathbf{Z}$  can be described in terms of a mixture of Poisson variables. Interestingly,  $\text{Bern}(a, b)$  sequences have been found to arise naturally in the study of random permutations, record values, Bayesian nonparametrics, and species allocation models.

However, for strings which are not of the form  $\{10 \dots 01\}$ , it seems that “nice” distributional expressions for their counts may not be available even with respect to sequence  $\text{Bern}(1, 0)$ . For instance, although the generating function for the count  $W_3 = \sum_{k \geq 1} X_k X_{k+1} X_{k+2}$ , of three consecutive successes, i.e., of the string  $\{111\}$ , can be found, its distribution is not known in a “closed form.” See [15] for more details.

As another independent Bernoulli sequence, consider  $Y_1, Y_2, \dots$  where  $Y_1 \equiv 1$ ,  $P(Y_k = 1) = a/(a + b + k - 2)$  for  $k \geq 2$  for  $a > 0, b \geq 0$ , which we call  $\text{Bern}_1(a, b)$ . This sequence appends a 1 to a  $\text{Bern}(a, b)$  sequence, thereby picking up an additional  $k$ -string corresponding to any leading 0's in the  $\text{Bern}(a, b)$  sequence. Another interpretation of  $\text{Bern}_1(a, b)$  arises from the following observation. The conditional distribution of the tail segment  $(Y_n, Y_{n+1}, \dots)$  in a  $\text{Bern}(a, b)$  sequence given  $Y_n = 1$  is  $\text{Bern}_1(a, b + n + 2)$ .

It can be proved that the joint distribution of  $\mathbf{Z}$  is sensitive to the value of  $b$  in a  $\text{Bern}_1(a, b)$  sequence. Namely, when  $b \geq 1$ , the joint distribution is again a mixture of Poisson variables, but is not when  $0 \leq b < 1$ .

By now there are several different ways to find the joint distribution of  $\mathbf{Z} = \{Z_m\}_{m \geq 1}$ , for instance by using combinatorial techniques [1–4], generating functions of moments [12, 17], Polya and Hoppe urns [7], and Poisson process embedding [8–10]. The purpose of this note is to summarize existing results, and to describe the last method in [10], the technique of using conditional marked Poisson process models, through which the joint distribution of  $\mathbf{Z}$  can be found for a large class of Bernoulli sequences including all sequences studied before, in particular  $\text{Bern}(a, b)$ ,  $\text{Bern}_1(a, b)$ , and dependent sequences.

The plan of the article is to give motivating examples in Sect. 2, and to detail the technique of conditional marked Poisson processes in Sect. 3. In Sects. 4, 5, and 6, this method is applied to find the joint distribution of  $\mathbf{Z}$  when  $\mathbf{Y} = \text{Bern}(a, b)$ , when  $\mathbf{Y} = \text{Bern}_1(a, b)$ , and also when  $\mathbf{Y}$  are some types of dependent Bernoulli sequences. In the following, we rely on the exposition in [10, 17].

## 2 Examples

Bernoulli sequences arise naturally in several situations. We give four examples below with respect to random permutations, Bayesian nonparametric statistics, production failures, and record values.

*Example 2.1.* This example will show that the Bernoulli sequence  $\text{Bern}(1, 0)$  arises in the limit in the study of cycles in random permutations. Let  $\mathbb{S}_n = \{1, 2, \dots, n\}$ , and consider the Feller algorithm to generate a permutation  $\pi : \mathbb{S}_n \rightarrow \mathbb{S}_n$  uniformly among the  $n!$  choices (cf. [5]):

1. Draw an element uniformly from  $\mathbb{S}_n$ , and call it  $\pi(1)$ . If  $\pi(1) = 1$ , a 1-cycle is completed. If  $\pi(1) \neq 1$ , make another draw uniformly from  $\mathbb{S}_n \setminus \{\pi(1)\}$ , and call it  $\pi(\pi(1))$ . If  $\pi(\pi(1)) = 1$ , a 2-cycle is completed. If  $\pi(\pi(1)) \neq 1$ , continue drawing from  $\mathbb{S}_n \setminus \{\pi(1), \pi(\pi(1))\}, \dots$  naming them  $\pi(\pi(\pi(1)))$ , and so on, until a cycle (of some length) is finished.
2. From the elements left in  $\mathbb{S}_n \setminus \{\pi(1), \pi(\pi(1)), \dots, 1\}$  after the first cycle is completed, follow the process in step 1 with the smallest remaining number taking the role of “1” to finish a second cycle. Repeat until all elements of  $\mathbb{S}_n$  are exhausted.

Let  $I_k^{(n)}$  be the indicator that a cycle is completed at the  $k$ th Feller draw from  $\mathbb{S}_n$ . A moment's thought convinces us that  $\{I_k^{(n)}\}_{k=1}^n$  are independent Bernoulli random variables with  $P(I_k^{(n)} = 1) = 1/(n - k + 1)$  since, at time  $k$  and independent of the past, exactly one choice from the remaining  $n - k + 1$  members left in  $\mathbb{S}_n$  completes the cycle. Denote  $C_k^{(n)}$  as the number of  $k$ -cycles in  $\pi$ ,

$$C_k^{(n)} = \begin{cases} I_1^{(n)} + \sum_{i=1}^{n-1} I_i^{(n)} I_{i+1}^{(n)} & \text{for } k = 1 \\ \prod_{l=1}^{k-1} (1 - I_l^{(n)}) I_k^{(n)} + \sum_{i=1}^{n-k} I_i^{(n)} \prod_{l=i+1}^{i+k-1} (1 - I_l^{(n)}) I_{i+k}^{(n)} & \text{for } 2 \leq k \leq n. \end{cases}$$

Now let  $\mathbf{Y}$  be the sequence  $\text{Bern}(1, 0)$  where  $P(Y_k = 1) = 1/k$  for  $k \geq 1$  so that  $Y_k \stackrel{d}{=} I_{n-k+1}^{(n)}$  in distribution, for  $1 \leq k \leq n$ . Since  $Y_n$ , and  $Y_{n-k+1} \prod_{l=n-k+2}^n (1 - Y_l)$  for  $2 \leq k \leq n$  all vanish in probability as  $n \uparrow \infty$ , we can conclude, for each  $k \geq 1$ , that  $\lim_{n \rightarrow \infty} C_k^{(n)} \stackrel{d}{=} Z_k$  in distribution.

Finally, as is well-known, the asymptotic cycle counts  $\{\lim_n C_k^{(n)}\}_{k \geq 1}$  are distributed as independent Poisson random variables with respective means  $1/k$  for  $k \geq 1$  (cf. [13]). Hence,  $\mathbf{Z} \stackrel{d}{=} \prod_{k \geq 1} \text{Po}(1/k)$ . See also [1, 2] for more discussion with Ewens sampling formula.

*Example 2.2.* Consider the standard nonparametric inference problem of estimating the unknown distribution function  $F$  from data  $X_1, X_2, \dots$  which are independently and identically distributed as  $F$ . In Bayesian inference, one would place a Dirichlet prior  $\mathcal{D}(\alpha)$  on  $F$ . Here  $\alpha$  is a finite measure on  $\mathbb{R}_1$  with  $a = \alpha(\mathbb{R}_1) > 0$ . Under these circumstances, one can show that there will be repetitions among  $X_1, X_2, \dots$ . Let  $\beta_1 = 1, \beta_n = I(X_n \notin \{X_1, \dots, X_{(n-1)}\})$  for  $n = 2, 3, \dots$ . Thus,  $\beta = 1$  if  $X_n$  is different from  $X_1, \dots, X_{(n-1)}$  and zero other wise. It is well-known that  $\beta_1, \beta_2, \dots$  are independent and  $P(\beta_n = 1) = a/(a + n - 1)$  for  $n = 1, 2, \dots$  and thus form a  $\text{Bern}(a, 0)$  sequence. For details, see [6, 14]. This example is also relevant in counting species among animals that are captured, and is part of the definition of species allocation models.

*Example 2.3.* Suppose items are produced and examined routinely over time. Alternatively, the item can be a long “chip” with successive spatial components. The data consist of a Bernoulli sequence  $\{Y_1, Y_2, \dots\}$ , where  $Y_n = 1$  means that there is a flaw (and  $Y_n = 0$  means that there is no flaw) at time  $n$  or at the  $n$ th spatial component. In practice, given improvements in production scheme or other attributes,  $P(Y_n = 1)$  will go to 0 as  $n$  gets large. Isolated flaws do not signify failures. However, successive flaws like  $\{11\}, \{101\}, \dots$  signify failures of say of type  $1, 2, \dots$ . One would like to know the distribution of the number of failures of type  $1, 2, \dots$ , e.g., the distribution of the joint distribution  $\mathbf{Z}$ .

*Example 2.4.* The following is another way to generate a  $\text{Bern}(1, 0)$  sequence from record values. Let  $\{\beta_i\}_{i \geq 1}$  be independent, identically distributed (iid)  $\text{Uniform}[0, 1]$  random variables, and define  $Y_1 = 1$  and  $Y_n = I(\beta_n \text{ is a record}) = I(\beta_n > \max(\beta_1, \dots, \beta_{(n-1)}))$ ,  $n \geq 2$ . R enyi's theorem shows that  $\{Y_n\}_{n \geq 1}$  are independent and  $P(Y_n = 1) = 1/n$  for  $n \geq 1$ , that is  $\mathbf{Y} = \text{Bern}(1, 0)$ .

### 3 Conditional Marked Poisson Process (CMPP)

To introduce the technique of conditional marked Poisson processes, let us further examine Example 2.4 of Sect. 2, and derive in its context the joint distribution of the count vector  $\mathbf{Z}$  associated with sequence  $\mathbf{Y} = \text{Bern}(1, 0)$ . With the same notations, define  $\tau_1 = 1, X_1 = \beta_1, \tau_n = \inf\{m : m > \tau_{n-1}, \beta_m > X_{\tau_{n-1}}\}, X_n = \beta_{\tau_n}$  for  $n \geq 2$ . Then,  $\{X_i\}_{i \geq 1}$  are the record values among  $\{\beta_i\}_{i \geq 1}$  and  $\{\tau_n\}_{i \geq 1}$  are the record times. Notice that the point process  $N$  on  $[0, 1]$  defined by  $N(A) = \sum_{i \geq 1} \delta_{X_i}(A)$  is a nonhomogeneous Poisson process on  $[0, 1]$  with intensity  $1/(1 - x)$  (cf. [16]).

For each record value  $X_i$ , we can associate a Geometric( $1 - X_i$ ) variable  $L_i$ , a mark, corresponding to the number of uniform random variables in  $\{\beta_i\}_{i \geq 1}$  to the next record. Then, by thinning decompositions,  $Z_k = \sum_{i \geq 1} I(L_i = k) = \sum_{i \geq 1} \delta_{X_i}([0, 1])I(L_i = k)$  for  $k \geq 1$  are independent Poisson variables with respective means  $\int_0^1 (1 - x)^{-1} x^{k-1} (1 - x) dx = 1/k$  for  $k \geq 1$ .

The idea now is to reverse the discussion above, and starting from what we call a conditional marked Poisson process (CMPP), which is slightly more general than a marked Poisson process, we determine a Bernoulli sequence  $\mathbf{Y}$  and compute the corresponding joint distribution of  $\mathbf{Z}$  through Poisson thinning decompositions.

**Conditional Marked Poisson Process** Consider a sequence of random variables  $(\mathbf{X}, \mathbf{L}) = \{(X_i, L_i)\}_{i \geq 0}$  on  $\mathbb{R} \times \mathbb{N}$  where  $\mathbb{N} = \{1, 2, \dots\}$ , and the point process  $N$  on  $\mathbb{R}$  given by  $N(A) = \sum_{i \geq 1} \delta_{X_i}(A)$ . Let also  $g : \mathbb{R} \rightarrow [0, \infty)$  be a probability density function (pdf), and for each  $x \in \mathbb{R} r(x, \cdot), q(x, \cdot) : \mathbb{N} \rightarrow [0, 1]$  be probability mass functions, and  $\lambda_x(\cdot) : \mathbb{R} \rightarrow [0, \infty)$  be an intensity function.

Then, we say that  $(\mathbf{X}, \mathbf{L})$  forms a CMPP  $\mathcal{M}(g, r, \lambda, q)$  if the following hold:

1.  $X_0$  has pdf  $g$ ,
2. Conditional on  $X_0 = x_0, N$  is a nonhomogeneous Poisson process with intensity function  $\lambda_{x_0}(\cdot)$ ,
3.  $P(L_0 = k | \mathbf{X}) = r(X_0, k)$  for  $k \geq 1$ , and
4.  $P(L_n = k | \mathbf{X}, L_0, L_1, \dots, L_{n-1}) = q(X_n, k)$  for  $k, n \geq 1$ .

Let  $L_0^* = L_0$ , and  $L_r^* = L_{r-1}^* + L_r$  for  $r \geq 1$ . We now define a Bernoulli sequence  $\mathbf{Y}$  based on  $(\mathbf{X}, \mathbf{L})$  as follows:  $Y_n = 1$  if  $n$  is of the form  $L_r^*$  for some  $r \geq 0$ , and  $Y_n = 0$  otherwise. A different way to say this is

$$Y_n = \begin{cases} 0, & \text{when } n < L_0^*, \text{ or } L_r^* < n < L_{r+1}^* \text{ for } r \geq 0 \\ 1, & \text{when } n = L_r^* \text{ for some } r \geq 0. \end{cases} \tag{3.1}$$

In the Bernoulli sequence  $\mathbf{Y}$ , there is a 1 : 1 correspondence between  $k$ -strings and marks  $L_n = k$ , which signify a “1” followed by  $(k - 1)$  “0”s and then succeeded by a “1.” Thus, the count vector  $\mathbf{Z}$  associated with  $\mathbf{Y}$  is given by

$$Z_k = \sum_{n \geq 1} I(L_n = k), \quad \text{for } k \geq 1. \tag{3.2}$$

We note the zeroth mark  $L_0$  is not included in the above summation since any  $Y_i$  with  $i < L_0$  is part of an initial segment of zeros of the sequence not preceded by a “1,” and so does not contribute to any  $k$ -string, for  $k \geq 1$ .

**Theorem 3.1.** *Suppose  $\int \lambda_w(x)q(x, k)dx < \infty$  for all  $w \in \mathbb{R}$  and  $k \geq 1$ . Then, the count vector  $\mathbf{Z}$  associated with sequence  $\mathbf{Y}$ , defined through CMPP  $(\mathbf{X}, \mathbf{L}) = \mathcal{M}(g, r, \lambda, q)$ , is distributed as follows. Given the value  $X_0 = x_0$ ,*

$$\mathbf{Z} \stackrel{d}{=} \prod_{k \geq 1} \text{Po} \left( \int \lambda_{x_0}(x)q(x, k)dx \right).$$

*Remark 3.2.* The distribution of  $\mathbf{Z}$  does not depend on the transition function  $r$ , consistent with the discussion of  $L_0$  before the theorem.

*Proof of Theorem 3.1.* Recall the count vector representation (3.2). Conditional on  $X_0 = x_0$ , the point process  $M$  on  $\mathbb{R} \times \mathbb{N}$  given by  $M(A \times \{k\}) = \sum_{i \geq 1} \delta_{X_i}(A)I(L_i = k)$  is a Poisson process on  $\mathbb{R} \times \mathbb{N}$  with intensity function  $\lambda_{x_0}(x)q(x, k)$  (cf. Proposition 4.10.1 (b) [16]). Hence, it follows that, given  $X_0 = x_0$ , the variables  $M(\mathbb{R} \times \{k\}) = \sum_{n \geq 1} I(L_n = k) = Z_k$  are independent Poisson variables with respective means  $\int \lambda_{x_0}(x)q(x, k)dx$ , for  $k \geq 1$ .  $\square$

### 4 The Sequence Bern( $a, b$ )

We now give a CMPP model which produces a Bern( $a, b$ ) sequence. Recall that a sequence  $\mathbf{Y}$  is a Bern( $a, b$ ) sequence if  $Y_1, Y_2, \dots$  are independent and  $P(Y_k = 1) = a/(a + b + k - 1)$  for  $k = 1, 2, \dots$ . Denote, as usual, for  $\alpha, \beta > 0$ , the Beta function

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}. \tag{4.1}$$

Let

1.  $\bar{g}(x) = x^{b-1}(1-x)^{a-1}/B(b, a)$  on  $0 < x < 1$ , the Beta( $b, a$ ) pdf,
2.  $\bar{f}(x, k) = x^{k-1}(1-x)$  for  $k \geq 1$ ,
3.  $\bar{\lambda}_w(x) = [a/(1-x)]I(w < x < 1)$ , and
4.  $\bar{q}(x, k) = x^{k-1}(1-x)$  for  $k \geq 1$ .

We note the Poisson process in the above CMPP model with intensity  $\bar{\lambda}_w(\cdot)$  can be generated in the following way. First, the point process formed by the record values from an iid sequence of Beta( $1, a$ ) random variables is a Poisson process



with intensity  $a/(1-x)$ , the Beta(1,  $a$ ) failure rate (cf. [16] Proposition 4.11.1 (b)). Next, we thin this process as follows. Let  $X_0 \stackrel{d}{=} \text{Beta}(b, a)$ , and  $\{X_i\}_{i \geq 1}$  be the record values from an iid sequence of Beta(1,  $a$ ) random variables, subject to  $X_i > X_0$  for  $i \geq 1$ . Then, conditional on  $X_0 = w$ , the point process  $\bar{N}$  defined by  $\bar{N}(A) = \sum_{i \geq 1} \delta_{X_i}(A)$  is the desired Poisson process with intensity function  $\bar{\lambda}_w(x) = [a/(1-x)]I(w < x < 1)$ .

**Proposition 4.1.** *The model  $(\mathbf{X}, \mathbf{L}) = \mathcal{M}(\bar{g}, \bar{r}, \bar{\lambda}, \bar{q})$  produces an independent Bernoulli sequence  $\mathbf{Y} \stackrel{d}{=} \text{Bern}(a, b)$  for  $a > 0$  and  $b > 0$  whose count vector  $\mathbf{Z}$ , conditional on the value  $x_0$  of a Beta( $b, a$ ) random variable, is distributed as  $\prod_{k \geq 1} \text{Po}(a(1-x_0^k)/k)$ .*

*Remark 4.2.* As a corollary, by taking  $b \downarrow 0$ , we recover the count vector distribution for Bern( $a, 0$ ) as simply  $\mathbf{Z} \stackrel{d}{=} \prod_{k \geq 1} \text{Po}(a/k)$ . Note that  $(X_0, L_0) \rightarrow (0, 1)$  in distribution as  $b \downarrow 0$ .

*Proof of Proposition 4.1.* The second part on the count vector distribution follows from Theorem 3.1, noting for  $k \geq 1$ , that

$$\int_0^1 \bar{\lambda}_{x_0}(x) \bar{q}(x, k) dx = \int_{x_0}^1 ax^{k-1} dx = \frac{a(1-x_0^k)}{k}. \tag{4.2}$$

The first part is proved by showing that the finite dimensional distributions of the Bernoulli sequence  $\mathbf{Y}$  agree with those of a Bern( $a, b$ ). Observe that the distribution of  $\{Y_i\}_{i \geq 1}$  given through (3.1) is uniquely determined by the probabilities of cylinder sets of the form  $E = E(k_0, \dots, k_n)$ ,

$$\begin{aligned} E &= (L_0 = k_0, L_1 = k_1, \dots, L_n = k_n) \\ &= \left( Y_t = 1 \text{ for } t \in \{K_0, K_1, \dots, K_n\}, \text{ and } Y_t = 0 \text{ otherwise for } 1 \leq t \leq K_n \right), \end{aligned} \tag{4.3}$$

where  $k_0, k_1, \dots, k_n$  are positive integers and  $K_0 = k_0, K_1 = K_0 + k_1, \dots, K_n = K_{n-1} + k_n$  are their partial sums. The random variables  $\{Y_n\}$  will form a Bern( $a, b$ ) sequence if

$$P(E(k_0, \dots, k_n)) = \prod_{i=1}^{K_n} \frac{b+i-1}{a+b+i-1} \prod_{r=0}^n \frac{a}{b+K_r-1}. \tag{4.4}$$

Let  $A_n = \{0 < x_0 < x_1 < \dots < x_n < 1\}$ . We now use the Beta variables representation given just above Proposition 4.1. Observe

$$P(E) = \int_{A_n} \bar{g}(x_0) \bar{r}(x_0, k_0) \prod_{i=1}^n \left[ P(X_i \in dx_i | X_i > x_{i-1}) \bar{q}(x_i, k_i) \right] dx_0.$$

Since  $P(X_i \in dx_i | X_i > x_{i-1}) = a(1 - x_i)^{a-1} / (1 - x_{i-1})^a dx_i$  for  $1 \leq i \leq n$ , we have further that the last line equals

$$\begin{aligned} & \frac{a^n}{B(b, a)} \int_{A_n} x_0^{b+k_0-2} \prod_{i=1}^n x_i^{k_i-1} (1 - x_n)^a dx_0 \dots dx_n \\ &= \frac{B(b + K_n - 1, a + 1)}{B(b, a)} \cdot \frac{a^n}{\prod_{s=0}^{n-1} (b + K_s - 1)} \\ &= \frac{a \prod_{r=0}^{K_n-2} (b + r)}{\prod_{r=0}^{K_n-1} (a + b + r)} \cdot \frac{a^n}{\prod_{s=0}^{n-1} (b + K_s - 1)}, \end{aligned}$$

which is equal to the probability in (4.4). □

We note following ideas based on Theorem 2.2 in [9] (Theorem 3.1 in this note). Holst [8] shows that an alternate CMPP model based on iid exponential random variables can also give rise to a  $\text{Bern}(a, b)$  and yield the same results for  $\mathbf{Z}$ .

### 5 The Sequence $\text{Bern}_1(a, b)$

Recall  $\text{Bern}_1(a, b)$  is the independent Bernoulli sequence  $\mathbf{Y}$  where  $P(Y_1 = 1) = 1$  and  $P(Y_k = 1) = a / (a + b + k - 2), k = 2, 3, \dots$  We now construct a CMPP model corresponding to  $\text{Bern}_1(a, b)$  sequence when  $a > 0, b > 1$ . Thus, the joint distribution of strings  $\mathbf{Z}$  in a  $\text{Bern}_1(a, b)$  sequence when  $a > 0, b > 1$  can be written as a certain mixture of Poissons.

Let  $a > 0$  and  $b > 1$ . Define

1.  $g^*(x) = x^{b-2}(1-x)^a / B(b-1, a+1)$  on  $0 < x < 1$ , the  $\text{Beta}(b - 1, a + 1)$  pdf,
2.  $r^*(x, 1) = 1$ ,
3.  $\lambda_w^*(x) = [a / (1 - x)]I(w < x < 1)$ , and
4.  $q^*(x, k) = x^{k-1}(1 - x)$  for  $k \geq 1$ .

Note that the Poisson process in the above CMPP model with intensity  $\lambda^*$  can be generated, as in Proposition 4.1, by taking  $X_0 \stackrel{d}{=} \text{Beta}(b - 1, a + 1)$ , and  $\{X_i\}_{i \geq 1}$  as the sequence of records from an iid sequence of  $\text{Beta}(1, a)$  random variables, subject to the condition  $X_1 > X_0$ .

**Proposition 5.1.** *The CMPP model  $(\mathbf{X}, \mathbf{L}) = \mathcal{M}(g^*, r^*, \lambda^*, q^*)$  produces an independent Bernoulli sequence  $\mathbf{Y} \stackrel{d}{=} \text{Bern}_1(a, b)$  for  $a > 0$  and  $b > 1$ , and, conditional on a  $\text{Beta}(b - 1, a + 1)$  variable  $X_0 = x_0$ , the distribution of its count vector  $\mathbf{Z}$  is  $\prod_{k \geq 1} \text{Po}(a(1 - x_0^k) / k)$ .*

*Remark 5.2.* As a corollary, by taking  $b \downarrow 1$ , we find the count vector distribution for  $\text{Bern}_1(a, 1)$  to be simply  $\mathbf{Z} \stackrel{d}{=} \prod_{k \geq 1} \text{Po}(a/k)$ . [In fact,  $\text{Bern}_1(a, 1)$  coincides with the sequence  $\text{Bern}(a, 0)$  mentioned earlier in Remark 4.2.]

*Proof of Proposition 5.1.* That the Bernoulli sequence  $\mathbf{Y}$  defined from  $\mathbf{X}, \mathbf{L}$  is  $\text{Bern}_1(a, b)$ , and the associated counts  $\mathbf{Z}$  are the desired mixture of Poissons follows from the same method as in Proposition 4.1. See [10] for details.  $\square$

Although a CMPP model does not lead to a  $\text{Bern}_1(a, b)$  sequence for  $a > 0, 0 \leq b < 1$ , the distributions of the associated count vector  $\mathbf{Z}$  can still be described with direct calculations in terms of a recurrence relation. However, it can be shown the distribution of  $\mathbf{Z}$  is not a mixture of Poissons. For more specifics, see [10].

## 6 Dependent Bernoulli Sequences

The CMPP model given in Sect. 3 can also produce dependent Bernoulli sequences. In all these cases, as a consequence of Theorem 3.1, the joint distribution of the count vector  $\mathbf{Z}$  are fully described as a mixture of Poisson variables.

We describe briefly two such examples.

*Example 6.1.* For  $a > 0$  and  $b > 0$ , denote  $P_{a,b}$  as the probability distribution of the CMPP  $\mathcal{M}(\bar{g}, \bar{r}, \bar{\lambda}, \bar{q})$  described in Proposition 4.1 which gives rise to the Bernoulli sequence  $\text{Bern}(a, b)$ . Let now  $r^+(x, k) = kx^{k-1}(1-x)^2$  for  $k \geq 1$ . Consider the associated CMPP model  $\mathcal{M}(\bar{g}, r^+, \bar{\lambda}, \bar{q})$  with  $\bar{g}, \bar{\lambda}, \bar{q}$  the same as in Proposition 4.1. Denote the probability measure under this model as  $P^+ = P_{a,b}^+$ .

Note that  $r^+(x, k) = k[\bar{r}(x, k) - \bar{r}(x, k + 1)]$  where  $\bar{r}(x, k) = x^{k-1}(1-x)$ . Recall the cylinder set  $E \stackrel{\text{def}}{=} E(k_0, \dots, k_n)$  from (4.3) where  $k_0, k_1, \dots, k_n$  are positive integers, and  $K_0, K_1, \dots, K_n$  their partial sums. It is easy to see that

$$P^+(E) = k_0 \left[ P_{a,b} \left( E(k_0, \dots, k_n) \right) - P_{a,b} \left( E(k_0 + 1, k_1, \dots, k_n) \right) \right].$$

From this expression, the distribution of  $\mathbf{Y}$  can be recovered, and shown with a few calculations not to be an independent sequence, e.g.,  $P^+(Y_1 = Y_2 = 1) \neq P^+(Y_1 = 1)P^+(Y_2 = 1)$ . For details see [10].

However, by noting Remark 3.2, the count vectors under  $P_{a,b}$  and  $P^+$  have the same distribution  $\prod_{k \geq 1} \text{Po}(a(1-x_0^k)/k)$ .

*Example 6.2.* Let  $\mathbf{Y}$  be the Bernoulli sequence  $\text{Bern}(1, 0)$  generated by the CMPP model based on  $(\mathbf{X}, \mathbf{L})$  discussed in Example 2.4 and Remark 4.2. Note that the count vector  $\mathbf{Z}$  does not change if one interchanges  $(X_1, L_1)$  and  $(X_2, L_2)$ . More precisely, let  $X_0^* = X_0, L_0^* = L_0, X_1^* = X_2, L_1^* = L_2, X_2^* = X_1, L_2^* = L_1$ , and  $X_n^* = X_n, L_n^* = L_n$  for  $n = 3, 4, \dots$ . Then, as the counts are invariant under such a switch,  $\mathbf{Z}^* = \mathbf{Z}$  still has distribution  $\prod_{k=1}^{\infty} \text{Po}(1/k)$ . However, the underlying Bernoulli sequence  $\mathbf{Y}^*$  generated by  $(\mathbf{X}^*, \mathbf{L}^*)$  is no longer independent. Again, one can show  $P(Y_1^* = Y_2^* = 1) \neq P(Y_1^* = 1)P(Y_2^* = 1)$ . Details can be found in [10].

## 7 Some Open Problems

We indicate two intriguing questions, although certainly many more can be envisioned.

1. As indicated in the introduction, the generating function of  $W_3$ , the count of strings of the form  $\{111\}$  in  $\text{Bern}(a, b)$  has been identified in the nice paper [15]. However, we do not have a good specification of the exact distribution. We know even less about counts of strings of the form  $\{1111\}$ ,  $\{11111\}$ , etc. although some recursions are given in [15]. Can one say something more about these counts?
2. In an interesting paper [11], the following question is raised. Consider the sequence  $\text{Bern}(a, 0)$ . We know that the count  $Z_1$  of strings of the form  $\{11\}$  is finite. Let  $N_1$  be the last  $n$  such that  $Y_{n-1}Y_n = 1$ . Is there a stopping time  $\tau$  on  $\mathbf{Y}$  such that  $P(\tau = N_1)$  is maximized among all stopping times? Hsiao [11] constructs such a  $\tau$  and shows that it is of a threshold type, that is there is a  $t \in \mathbb{N}$  such that  $\tau = \min\{n : n \geq t, Y_{n-1}Y_n = 1\}$ . It will be interesting to answer this question for Bernoulli sequences  $\text{Bern}(a, b)$  and for other counts  $Z_2, Z_3, \dots$

## References

- [1] Arratia, R., Barbour, A.D. and Tavaré, S. (1992) Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.* **2** 519–535.
- [2] Arratia, R., Barbour, A.D. and Tavaré, S. (2003) *Logarithmic Combinatorial Structures: A Probabilistic Approach*. European Mathematical Society, Zürich.
- [3] Arratia, R., and Tavaré, S. (1992) The cycle structure of random permutations. *Ann. Probab.* **20** 1567–1591.
- [4] Chern, H.-H., Hwang, H.-K. and Yeh, Y.-N. (2000) Distribution of the number of consecutive records. *Random Struct. Algor.* **17** 169–196.
- [5] Feller, W. (1945) The fundamental limit theorems in probability. *Bull. Amer. Math. Soc.* **51** 800–832.
- [6] Ghosh, J.K., and Ramamoorthi, R.V. (2003) *Bayesian Nonparametrics*. Springer, New York.
- [7] Holst, L. (2007) Counts of failure strings in certain Bernoulli sequences. *J. Appl. Probab.* **44** 824–830.
- [8] Holst, L. (2008) A note on embedding certain Bernoulli sequences in marked Poisson processes *J. Appl. Probab.* **45** 1181–1185.
- [9] Huffer, F., Sethuraman, J. and Sethuraman, S. (2008) A study of counts of Bernoulli strings via conditional Poisson processes. Available at arXiv 0801.2115v1.pdf.
- [10] Huffer, F., Sethuraman, J. and Sethuraman, S. (2009) A study of counts of Bernoulli strings via conditional Poisson processes. *Proc. Amer. Math. Soc.* **137** 2125–2134.
- [11] Hsiao S. (2009) Selecting the last consecutive records in a record process. *Technical Report*.
- [12] Joffe, A., Marchand, E., Perron, F. and Popadiuk, P. (2004) On sums of products of Bernoulli variables and random permutations. *J. Theoret. Probab.* **17** 285–292.
- [13] Kolchin, V.F. (1971) A problem of the allocation of particles in cells and cycles of random permutations. *Theory Probab. Appl.* **16** 74–90.
- [14] Korwar, R.M., and Hollander, M. (1973) Contributions to the theory of Dirichlet processes. *Ann. Probab.* **1** 705–711.

- [15] Móri, T.F. (2001) On the distribution of sums of overlapping products. *Acta Scientiarum Mathematica (Szeged)* **67** 833–841.
- [16] Resnick, S.I. (1994) *Adventures in Stochastic Processes*. Second Ed. Birkhäuser, Boston.
- [17] Sethuraman, J., and Sethuraman, S. (2004) On counts of Bernoulli strings and connections to rank orders and random permutations. In *A festschrift for Herman Rubin. IMS Lecture Notes Monograph Series* **45** 140–152.

# Storage Models for a Class of Master Equations with Separable Kernels

P. R. Vittal\*, S. Jayasankar, and V. Muralidhar

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** We discuss a number of storage problems for a class of one-dimensional master equations with separable kernels. For this class of problems, the integral equation for the first overflow or first emptiness can be transformed exactly into ordinary differential equations. Analysis is done with a generalised separable kernel. Using imbedding method, closed form solutions are obtained for the first overflow without or with emptiness in a given time. The first passage time for emptiness without or with overflow in a given time is also obtained. The imbedding technique is also used to study the expected amount of overflow in a given time. Diffusion approximation for this model is also obtained using suitable statistical conditions.

**Mathematics Subject Classification (2000)** 60k25, 11465

**Key words and phrases** Master equation · Imbedding method · Finite storage system · Overflow · Emptiness · Absorbing and reflecting barriers · Diffusion approximation

## 1 Introduction

There is a lot of literature connected with the studies of storage theory, queueing theory, dam theory, risk theory, neuronal spike discharge activities, communication theory, etc. The problems met with in these fields can be analyzed by identical

---

\* This work was completed when one of the authors (P.R. Vittal) was in ISI (Bangalore) in January 2009.

P.R. Vittal and S. Jayasankar  
Ramakrishna Mission, Vivekananda College, Chennai - 600 004, India  
e-mail: [vittal\\_ramaseshan@yahoo.com](mailto:vittal_ramaseshan@yahoo.com); [ksksjayjay@gmail.com](mailto:ksksjayjay@gmail.com)

V. Muralidhar  
Mohamed Sathak College, Chennai, India  
e-mail: [mbvadari@yahoo.co.in](mailto:mbvadari@yahoo.co.in)

techniques. Moran [16] has initiated the studies of storage systems by working out solutions for discrete time models. Further contributions in this field have been made by Gani [9], Prabhu [22], and others. The content of an infinite dam with Poisson inputs has been studied in considerable detail, and a survey of this interesting field has been studied by Prabhu [23]. However, a storage model with finite boundaries poses difficult problems. For Poisson inputs, Tackas [30] adopted the combinatorial techniques to study the fluctuations of the content of a finite dam. The extension to continuous time version of Moran's discrete model was carried out by Maron [17] and Downton [7] using a limiting method.

A systematic version of continuous time model has been formulated by Kendall [15]. He has obtained an elegant result for the wet period of the dam. Cochen [5, 10] has made use of the Pollaczek [21] integral equation for various models of general storage theory. A number of time-dependent results for some of these storage models have been given by Saaty [27], Yeo [35], Chover and Yeo [3], Gover and Muller [11], and many others. In all these cases, the input is a Poisson or renewal process and the amount of input is governed by an independent and identically distributed random variable. The concept of first passage densities for a compound Poisson process and ideas of renewal theory and product densities are elaborated in an excellent review in the *Handbuch der Physik* (Ramakrishnan [25]). The epochs of inputs are assumed to constitute a stationary renewal point process as has been used by Srinivasan [27] and Phatarfod [20]. The method of using backward integral equation described by Bellman and Harris [1] has been used by Srinivasan [28]. Phatarfod has used the Wald [34] identity for studying the wet period of a finite dam.

Regarding the release policy, different models have been thought of. The release has been considered as a deterministic process with a constant rate (Srinivasan [28]). The general type of deterministic release of problems has been studied by Cinlar and Pinsky [4]. The exponential release rule has been considered by Yeo [35] and Vasudevan and Vittal [30, 31] for a finite dam model and Keilson and Mermin [14] for studying short noise problems. First exit times for compound Poisson process for certain type of positive and negative jumps have been studied by Pery et al. [18]. In another paper [19], they also derived results for the expected total discounted cost of switching and maintaining extra capacity as well as the total expected discounted loss of discharged services in queues. This type of modeling also has been applied for the reception of light on the retina of eye when the light is switched on in a dark room. Impulses are received in a Poisson manner and between impulses exponential loss of light occurs while one can 'see' when the level of light on the retina reaches a constant threshold  $k$ .

Karlin and Fabens [13] used the renewal process to permit certain interdependence between successive inputs for discrete time models in the theory of stationary inventory models. In problems relating to the warehouse model, the demand for the storage occurs in a Poisson manner by outputs governed by independent and identically distributed random variables. When the storage falls below a certain specified reorder level, its orders are received. They are not refused but kept on record and filled in later. These have been described by Prabhu [23]. In insurance problems, claims (outputs) are taken as random and the inputs as deterministic to study the survival time (Cramer Herald [6]). In neurobiology, sequences of neuronal

firings referred to as spike trains arise from the so-called spontaneous activity of response of the neurons to external stimuli. Spike trains and the corresponding internal histograms of spontaneous activity of a single neuron have been recorded experimentally using electrodes (Redman and Lambard [26]). Many mathematical and statistical models have been proposed to reproduce the neuronal activity to fit the data. Details of theoretical models of single neurons are given in Feinberg [8], Holden [12], Srinivasan and Sampath [29], and many others.

All these problems pose the same type of questions. The quantities of interest that which are studied in these problems are similar to the computation of First Passage Time Density for overflows or emptiness for a finite storage system. A search of the literature reveals that there is enough scope for studying continuous time models with random inputs and random outputs with or without deterministic release. The case of infinite depth dam with Poisson inputs and Poisson release has been considered by Puri and Senthuria [24]. A finite dam model with Poisson inputs and Poisson outputs and a deterministic release policy has been studied by Vasudevan et al. [32].

In this contribution, we will be concerned with a finite storage system with Poisson inputs and Poisson outputs. The amounts of inputs and outputs are governed by independent and identically distributed random variables with general distribution. In Sect. 2, using the imbedding method [2] we derive the differential equation for the first passage time density (FPTD) for overflow and the closed form solution in terms of Laplace transform (LT) for FPTD. Here, we treat  $X = 0$  and  $X = k$  as the barriers of the finite storage system. Also, we treat  $X = 0$  and  $X = k$  as absorbing barriers.

In Sect. 3, we derive the result for FPTD for overflow treating  $X = 0$  as a reflecting barrier. This means arbitrary emptiness is allowed before time  $t$ . In Sect. 4, we study the expected amount of overflow in time  $t$  and without emptiness in this period. In Sect. 5, we derived the result for the expected amount of overflow in time  $t$  allowing arbitrary emptiness. In Sect. 6, we obtain the diffusion approximation for the model.

## 2 First Passage Time for Overflow Without Emptiness

Consider a storage model with Poisson inputs and Poisson outputs. The input and output sizes form a transition density function  $k(x, y)$ .

The model describing the process is

$$X(t) = x + \sum_{n=1}^{N(t)} Z_n \quad (2.1)$$

where  $X(0) = x$  is the initial level of the warehouse and  $N(t)$  is the number of inputs and outputs which occur in a Poisson process with intensity  $\gamma$ . Here,  $Z_n$  is a sequence of independent and identical random variables with transition density  $k[x, y]$ .



$X(t)$  is the level of the warehouse at time  $t$ . The storage is of finite capacity  $X = k$ . This means the process  $X(t)$  has two barriers  $X = 0$  and  $X = k$ . We are interested in finding the first passage time density for overflow before emptiness.

Define  $f(x, k, t)$  as the probability that the overflow occurs for the first time between time  $t$  and  $t + dt$  without emptiness occurring in the interval  $(0, t)$ .

Consider the dynamics of the process for  $f(x, k, t)$  in the initial interval of time  $dt$ . The following mutually exclusive pairwise events may occur.

- (a) There is no random input or output.
- (b) There is a random input or output but the level of the store reaches  $X < k$ .
- (c) There is a random output so that the level of the store is  $X > k$ .

Considering the above possibilities for the initial interval of time  $dt$  the equation of motion for the process is

$$f(x, k, t + dt) = (1 - \gamma dt) f(x, k, t) + \gamma dt \int_x^k k(x, y) f(y, k, t) dy + \gamma dt \int_0^x k(x, y) f(y, k, t) dy + \delta(t) \gamma dt \int_k^\infty k(x, y) dy. \quad (2.2)$$

In the first and third integral on the right,  $y \geq x$  and in the second integral  $x \geq y$ .

Also,  $\delta(t)$  occurring with the last integral on the right is the Dirac delta function and the third integral on the right corresponds to the event of crossing the level  $X = k$  in the initial interval of time  $dt$ .

On proceeding to the limit as  $dt \rightarrow 0$  in (2.2), we get results in the integral equation

$$\frac{\partial f}{\partial t} + \gamma f = \gamma \int_x^k k(x, y) f(y, k, t) dy + \gamma \delta(t) \int_k^\infty k(x, y) dy + \gamma \int_0^x k(x, y) f(y, k, t) dy. \quad (2.3)$$

Define the Laplace transform (LT) of  $f(x, k, t)$  as

$$\bar{f}(x, k, l) = \int_0^\infty e^{-lt} f(x, k, t) dt.$$

Taking the LT with respect to  $t$ , (2.3) becomes

$$\begin{aligned}
 (\ell + \gamma)\bar{f} &= \gamma \int_x^k k(x, y) \bar{f}(y, k, \ell) dy \\
 &+ \gamma \int_k^\infty k(x, y) dy + \gamma \int_0^x k(x, y) \bar{f}(y, k, \ell) dy. \tag{2.4}
 \end{aligned}$$

In the theory of Fredholm integral equations, separable kernels play a special role in converting the integral equations to algebraic form. We now introduce a class of kernels for which the master equation can be reduced to much simpler form and for which the FPTD in one dimension is readily solvable.

Hence, we consider the kernel  $k(x, y)$  as

$$k(x, y) = \begin{cases} a(x) b(y) \rho(x), & y \geq x \\ b(x) a(y) \rho(x), & x \geq y. \end{cases} \tag{2.5}$$

This form of kernel allows positive and negative jumps following asymmetric or symmetric random walks.

The above form of  $k(x, y)$  satisfies the balance equation

$$k(x, y) \rho(y) = k(y, x) \rho(x). \tag{2.6}$$

The function  $\rho(x)$  that appears in (2.5) is the normalizing function so that

$$\int_{-\infty}^\infty k(x, y) dx = 1. \tag{2.7}$$

With this choice of  $k(x, y)$  in (2.5), the integral equation (2.4) can be written as

$$\begin{aligned}
 (\ell + \gamma)\bar{f} &= \gamma a(x) \int_x^k b(y) \rho(y) \bar{f}(y, k, \ell) dy \\
 &+ \gamma b(x) \int_0^x a(y) \rho(y) \bar{f}(y, k, \ell) dy \\
 &+ \gamma a(x) \int_k^\infty b(y) \rho(y) dy \tag{2.8}
 \end{aligned}$$

That is  $(\ell + \gamma)\bar{f} = aU + bV + aU_1$  (2.9)

where

$$U = \gamma \int_0^x b(y) \rho(y) \bar{f}(y, k, \ell) dy$$

$$V = \gamma \int_k^\infty a(y) \rho(y) \bar{f}(y, k, \ell) dy$$

$$U_1 = \gamma \int_k^\infty b(y) \rho(y) dy.$$

Differentiating twice with respect to  $x$ , (2.9) becomes

$$\begin{aligned} (\ell + \gamma)\bar{f}' &= a'U + b'V + a'U_1 \\ &= a'(U + U_1) + b'V \end{aligned} \tag{2.10}$$

$$(\ell + \gamma)\bar{f}'' = a''(U + U_1) + b''V - a'b\rho + b'a\rho. \tag{2.11}$$

Eliminate  $U + U_1$  and  $V$  from (2.10) and (2.11)

$$b'(\ell + \gamma)\bar{f}' = ab'(U + U_1) + bb'V \tag{2.12}$$

$$b(\ell + \gamma)\bar{f}' = a'b(V + U_1) + bb'V. \tag{2.13}$$

Subtracting (2.13) from (2.12), we get

$$U + U_1 = \frac{(\ell + \gamma)(b\bar{f}' - b'\bar{f})}{a'b - ab'}. \tag{2.14}$$

Similarly,

$$V = \frac{-(\ell + \gamma)(a\bar{f}' - a'f)}{a'b - ab'}. \tag{2.15}$$

Using (2.14) and (2.15) in (2.11), we arrive at

$$\begin{aligned} (\ell + \gamma)f'' - \frac{(\ell + \gamma)(a''b' - ab'')}{a'b - ab'}\bar{f}' \\ + \left[ \frac{(\ell + \gamma)(a''b' - b''a')}{a'b - ab'} - (ab' - a'b)\rho \right] \bar{f} = 0. \end{aligned} \tag{2.16}$$

In order to have a closed form solution for  $\bar{f}(x, k, \ell)$ , we choose

$$a(x) = e^{\alpha x}$$

$$b(x) = e^{-\beta x}$$

and

$$\rho(x) = e^{(\beta-\alpha)x}. \quad (2.17)$$

The differential equation (2.16) reduces to

$$(\ell + \gamma)\bar{f}'' - (\ell + \gamma)(\alpha - \beta)\bar{f}' - \ell\alpha\beta f = 0. \quad (2.18)$$

The characteristic equation of (2.18) is

$$(\ell + \gamma)m^2 - (\ell + \gamma)(\alpha - \beta)m - \ell\alpha\beta = 0. \quad (2.19)$$

The solution of the differential equation (2.18) with constant coefficients is

$$\bar{f}(x, k, \ell) = A(k, \ell)e^{m_1x} + B(k, \ell)e^{m_2x}, \quad (2.20)$$

where

$$m_1, m_2 = \frac{(\ell + \gamma)(\alpha - \beta) \pm \sqrt{(\ell + \gamma)^2(\alpha - \beta)^2 + 4\ell\alpha\beta(\ell + \gamma)}}{2(\ell + \gamma)}. \quad (2.21)$$

To determine  $A$  and  $B$ , we write down the imbedding equation (2.4) in the special case of  $k(x, y)$  given by (2.17) and substitute (2.20) in (2.5)

$$\begin{aligned} (\ell + \gamma)\bar{f} &= \frac{\gamma\alpha\beta}{\alpha + \beta} e^{\alpha x} \int_x^k e^{-\alpha y} (Ae^{m_1y} + \beta e^{m_2y}) dy \\ &+ \frac{\gamma\alpha\beta}{\alpha + \beta} e^{-\beta x} \int_0^x e^{\beta y} (Ae^{m_1y} + Be^{m_2y}) dy + \frac{\gamma\alpha\beta e^{\alpha x}}{\alpha + \beta} \int_k^\infty e^{-\alpha y} dy \end{aligned} \quad (2.22)$$

One can easily see that the coefficients of  $Ae^{m_1x}$  and  $Be^{m_2x}$  vanish separately.

As the solution for  $\bar{f}(x, k, \ell)$  is true for all values of  $u$  in  $0 \leq x \leq k$ , we arrive at two conditions connecting  $A$  and  $B$ , namely

$$\frac{Ae^{m_1k}}{m_1 - \alpha} + \frac{Be^{m_2k}}{m_2 - \alpha} - \frac{1}{\alpha} = 0 \quad (2.23)$$

$$\frac{Ae^{m_1k}}{m_1 + \beta} + \frac{Be^{m_2k}}{m_2 + \beta} = 0 \quad (2.24)$$

Equation (2.20) together with (2.23) and (2.24) gives the closed form analytical solution for the LT of  $f(u, k, t)$ .

**Special case:** In the case of symmetric random walk,  $\alpha = \beta = \lambda$  (say), (2.19) reduces to

$$(\ell + \gamma)\bar{f}'' - \lambda^2\ell = 0 \tag{2.25}$$

where

$$m_1 = -m_2 = \lambda\sqrt{\frac{\ell}{\ell + \gamma}}. \tag{2.26}$$

**Mean Passage time**

The  $n$ th moment for the FPTD is given by

$$T_n(x) = (-1)^n \left. \frac{d^n}{d\ell^n} \bar{f}(u, \ell) \right|_{\ell=0}. \tag{2.27}$$

The differential equation for the  $n$ th moment is given by

$$\gamma T_n''(x) - nT_{n-1}''(x) + \lambda^2 nT_{n-1}(x) = 0. \tag{2.28}$$

This is obtained by using the Leibnitz theorem for the  $n$ th derivative of the product of two functions with (2.24).

Noting that  $T_0(x) = 1$ , for the case  $n = 1$ ,

$$T_1'' = \frac{-\lambda^2}{\gamma} \tag{2.29}$$

which is the differential equation derived by Vittal et al. [33]. The mean passage time for overflow can easily be determined as in [33].

**3 First Passage Time for Overflow with Arbitrary Number of Emptiness**

Here, our interest is in obtaining the first passage time density for overflow with arbitrary number of emptiness. This means we are treating  $X = 0$  as a reflecting barrier. The analysis for this model is exactly the same as in the previous model but for a change in the boundary condition to be incorporated in the imbedding equation (2.4). Here, once the level of the store crosses  $X = 0$  downward, it resets at  $X = 0$  and starts the dynamics of the process from  $X = 0$ . Defining  $f_1(x, k, t)$  as the

FPTD for overflow, treating  $X = 0$  as a reflecting barrier and  $\bar{f}_1(u, k, \ell)$  as its LT, we have

$$\begin{aligned}
 (\ell + \gamma)\bar{f}_1 &= \gamma \int_x^k k(x, y) \bar{f}_1(y, k, \ell) dy + \gamma \int_k^\infty k(x, y) dy \\
 &\quad + \gamma \int_0^x k(x, y) \bar{f}_1(y, k, \ell) dy + \gamma \int_{-\infty}^0 k(x, y) dy. \tag{3.1}
 \end{aligned}$$

In the first two integrals  $y \geq x$  and with last two integrals  $x \geq y$ .

We arrive at the same differential equation (2.18) with  $\bar{f}$  being replaced  $\bar{f}_1$ . For the choice of  $k(x, y)$  given in (2.5), the general solution for  $\bar{f}_1(x, k, \ell)$  is

$$\bar{f}_1(x, k, \ell) = A_1(k, \ell)e^{m_1x} + B_1(k, \ell)e^{m_2x}. \tag{3.2}$$

The equations for  $A_1$  and  $B_1$  are

$$\frac{A_1 e^{m_1k}}{m_1 - \alpha} + \frac{B_1 e^{m_2k}}{m_2 - \alpha} - \frac{1}{\alpha} = 0 \tag{3.3}$$

and

$$\frac{A_1 e^{m_1k}}{m_1 + \beta} + \frac{B_1 e^{m_2k}}{m_2 + \beta} + \frac{1}{\beta} = 0. \tag{3.4}$$

Thus,  $\bar{f}_1(x, k, \ell)$  is completely determined by (3.2) together with (3.3) and (3.4).

It will be a straightforward procedure to determine mean passage time for overflow using (2.26).

### 4 Expected Amount of Overflow in a Given Time

Define  $S(x, k, t)$  as the expected amount of overflow in time  $t$  assuming that there is no emptiness of the store in time  $t$ .

The imbedding equation for  $S(x, k, t)$  is

$$\begin{aligned}
 \frac{\partial S}{\partial t} + \gamma S &= \gamma \int_x^k k(x, y) S(y, k, t) dy + \gamma \delta(t) \int_k^\infty k(x, y) (y - k) dy \\
 &\quad + \gamma \int_k^\infty k(x, y) dy \cdot S(k, k, t) + \gamma \int_0^x k(x, y) S(y, k, t) dy. \tag{4.1}
 \end{aligned}$$

Here, we have to observe that the second integral on the RHS corresponds to the excess overflow in the initial interval of time  $dt$  and the third integral takes care of the process repeating from the restart level  $X = k$ .

Define

$$\bar{S}(x, k, \ell) = \int_0^\infty e^{-\ell t} S(x, k, t) dt. \tag{4.2}$$

Taking the LT with respect to  $t$ , (4.1) gets converted to

$$\begin{aligned} (\ell + \gamma)\bar{S} &= \gamma \int_x^k k(x, y) \bar{S}(y, k, \ell) dy + \gamma \int_k^\infty k(x, y) (y - k) dy \\ &+ \gamma \int_k^\infty k(x, y) dy \bar{S}(k, k, \ell) + \gamma \int_0^x k(x, y) \bar{S}(y, k, \ell) dy. \end{aligned} \tag{4.3}$$

Taking  $k(x, y)$  as given in (2.5), we write the (4.3) as

$$\begin{aligned} (\ell + \gamma)\bar{S} &= \gamma a(x) \int_x^\infty b(y) \rho(y) \bar{S}(y, k, \ell) dy \\ &+ \gamma a(x) \int_k^\infty b(y) \rho(y) (y - k) dy \\ &+ \gamma a(x) \int_k^\infty b(y) \rho(y) dy \cdot \bar{S}(k, k, \ell) \\ &+ \gamma b(x) \int_0^x a(y) \rho(y) \bar{S}(y, k, \ell) dy. \end{aligned} \tag{4.4}$$

The differential equation for  $\bar{S}(y, k, \ell)$  is the same as (2.20) except  $\bar{f}$  is being replaced by  $\bar{S}$ . For the choice of  $a(x)$ ,  $b(x)$  and  $\rho(x)$  as in (2.17), the differential equation for  $\bar{S}(x, k, \ell)$  is

$$(\ell + \gamma) \frac{\partial^2 \bar{S}}{\partial x^2} - (\ell + \gamma)(\alpha - \beta) \frac{\partial \bar{S}}{\partial x} - \ell \alpha \beta \bar{S} = 0 \tag{4.5}$$

whose solution is

$$\bar{S}(x, k, \ell) = A_2(k, \ell)e^{m_1 x} + B_2(k, \ell)e^{m_2 x} \tag{4.6}$$

where  $m_1$  and  $m_2$  are the roots of the equation

$$(\ell + \gamma)m^2 - (\ell + \gamma)(\alpha - \beta)m - \ell\alpha\beta = 0. \tag{4.7}$$

As done in earlier sections to determine  $A_2$  and  $B_2$ , substituting (4.6) in (4.4) and using the choices of  $a(x)$ ,  $b(x)$ , and  $\rho(x)$  as taken earlier, we arrive at the equation

$$\begin{aligned} (\ell + \gamma)\bar{S} &= \frac{\gamma\alpha\beta}{\alpha + \beta} e^{\alpha x} \int_x^k [A_2 e^{(m_1 - \alpha)y} + B_2 e^{(m_2 - \alpha)y}] dy \\ &+ \frac{\gamma\alpha\beta}{\alpha + \beta} e^{\alpha x} \int_k^\infty e^{-\alpha y} (y - k) dy \\ &+ \frac{\gamma\alpha\beta}{\alpha + \beta} e^{\alpha x} \int_k^\infty e^{-\alpha y} dy \cdot (A_2 e^{m_1 k} + B_2 e^{m_2 k}) \\ &+ \frac{\gamma\alpha\beta}{\alpha + \beta} e^{-\beta x} \int_0^x [A_2 e^{(m_1 + \alpha)y} + B_2 e^{(m_2 + \alpha)y}] dy. \end{aligned} \tag{4.8}$$

This equation results in two independent conditions connecting  $A_2$  and  $B_2$  as

$$\frac{A_2 m_1 e^{m_1 k}}{m_1 - a} + \frac{B_2 m_2 e^{m_2 k}}{m_2 - a} + \frac{1}{\alpha} = 0 \tag{4.9}$$

$$\frac{A_2 e^{m_1 k}}{m_1 + \beta} + \frac{B_2 e^{m_2 k}}{m_2 + \beta} = 0. \tag{4.10}$$

Equations (4.6), (4.9), and (4.10) give the complete closed form solution for  $\bar{S}(x, k, \ell)$ , the LT of the function giving the expected amount of overflow in time  $t$ .

### 5 Expected Amount of Overflow Allowing Arbitrary Number of Emptiness

Here, we treat  $X = 0$  as a reflecting barrier. we already treated  $X = k$  as a reflecting barrier. Also here the procedure is exactly the same as in the last section but for the additional term  $\gamma \int_0^k k(x, y) dy \cdot \bar{S}_1(0, k, \ell)$  figuring in the (4.4). This only means that after the process  $X(t)$  crosses the level  $X = 0$  in the initial interval of time  $dt$



and it restarts again from  $X = 0$  to proceed to overflow between time  $t$  and  $t + dt$ .  $\bar{S}_1(x, k, \ell)$  is taken as the LT of  $S_1(x, k, t)$ , the expected amount of overflow in time  $t$  allowing arbitrary number of emptiness.

The solution for  $\bar{S}_1(x, k, \ell)$  is

$$\bar{S}_1(x, k, \ell) = A_3(k, \ell)e^{m_1x} + B_3(k, \ell) e^{m_2x} \tag{5.1}$$

subject to the conditions

$$\frac{A_3m_1}{m_1 - \alpha} e^{m_1k} + \frac{B_3m_2}{m_2 - \alpha} e^{m_2k} + \frac{1}{\alpha} = 0 \tag{5.2}$$

and

$$\frac{Ae^{m_1k}}{m_1 + \beta} + \frac{Be^{m_2k}}{m_2 + \beta} + \frac{1}{\beta} = 0. \tag{5.3}$$

## 6 Diffusion Approximation

In all the earlier sections we considered, jumps in the level of warehouse are in both the directions. If we impose certain statistical conditions on the jump size and jump frequency, we obtain a diffusion equation. In our model, we take

$$a(x) = e^{-\lambda x}, b(x) = e^{\lambda x}, \text{ and } \rho(x) = \frac{\lambda}{2}.$$

Then the statistical conditions to be imposed for the diffusion limit are

1. The average jump size is very small ( $\lambda \rightarrow \infty$ ).
2. The number of occurrence of Poisson jumps is very large ( $\gamma \rightarrow \infty$ ) such that  $\lim_{\gamma \rightarrow \infty} \frac{\gamma}{\lambda^2} = D$  where  $D$  has the dimensions of a diffusion constant.

Let us define  $P(x, t)$  as the probability for the position of the warehouse at time  $t$  given that  $X(0) = x$ . For a change, we arrive at the differential equation for the free motion of the level of the process  $X(t)$  by a different approach.

We split  $P(x, t)$  as  $P^+(x, t)$  and  $P^-(x, t)$  such that

$$P(x, t) = P^+(x, t) + P^-(x, t). \tag{6.1}$$

Here,  $P^+(x, t)$  is the probability for the level of the warehouse at any height with an initial jump in the positive direction whenever it occurs, and  $P^-(x, t)$  the

corresponding probability with the initial jump in the negative direction whenever it occurs. The integro-differential equation for  $P^+(x, t)$  is

$$e^{-\lambda x} \left( \frac{\partial P^+}{\partial t} + \gamma P^+ \right) = \frac{\gamma \lambda}{2} \int_x^\infty e^{-\lambda y} P(y, t) dy. \quad (6.2)$$

Differentiating this with respect to  $x$ , we get

$$\frac{\partial^2 P^+}{\partial t \partial x} + \gamma \frac{\partial P^+}{\partial x} - \lambda \left( \frac{\partial P^+}{\partial t} + \gamma P^+ \right) = -\frac{\gamma \lambda}{2} P. \quad (6.3)$$

The similar equation for  $P^-(x, t)$  is

$$\frac{\partial^2 P^-}{\partial t \partial x} + \gamma \frac{\partial P^-}{\partial x} + \lambda \left( \frac{\partial P^-}{\partial t} + \gamma P^- \right) = \frac{\gamma \lambda}{2} P. \quad (6.4)$$

Adding (6.3) and (6.4), we get

$$\frac{\partial^2 P}{\partial t \partial x} + \gamma \frac{\partial P}{\partial x} - \lambda \left( \frac{\partial P^+}{\partial t} - \frac{\partial P^-}{\partial t} \right) - \gamma \lambda (P^+ - P^-) = 0. \quad (6.5)$$

Subtracting (6.4) from (6.3), we get

$$\left( \frac{\partial^2 P^+}{\partial t \partial x} - \frac{\partial^2 P^-}{\partial t \partial x} \right) + \gamma \left( \frac{\partial P^+}{\partial x} - \frac{\partial P^-}{\partial x} \right) - \lambda \left( \frac{\partial P}{\partial t} + \gamma P \right) = -\gamma \lambda P. \quad (6.6)$$

Differentiating (6.5) with respect to  $x$ ,

$$\frac{\partial^3 P}{\partial t \partial x^2} + \gamma \frac{\partial^2 P}{\partial x^2} - \lambda \left( \frac{\partial^2 P^+}{\partial t \partial x} - \frac{\partial^2 P^-}{\partial t \partial x} \right) - \gamma \lambda \left( \frac{\partial P^+}{\partial x} - \frac{\partial P^-}{\partial x} \right) = 0. \quad (6.7)$$

Multiplying (6.6) by  $\lambda$  and adding it to (6.7), we arrive at

$$\frac{\partial^3 P}{\partial t \partial x^2} + \gamma \frac{\partial^2 P}{\partial x^2} - \lambda^2 \left( \frac{\partial P}{\partial t} + \gamma P \right) - \gamma \lambda^2 P = 0. \quad (6.8)$$

Dividing by  $\lambda^2$  and proceeding to the limit as  $\lambda \rightarrow \infty$ ,  $\gamma \rightarrow 0$  such that  $\frac{\gamma}{\lambda^2} \rightarrow D$ , we get the diffusion equation

$$D \frac{\partial^2 P}{\partial x^2} = \frac{\partial P}{\partial t}. \quad (6.9)$$

Adopting to the same limiting procedure in the FPTD, we arrive at the FPTD to reach the barrier  $X = k$  for the diffusion process.

**Conclusion:** In conclusion, we point out that the central idea of this paper is to show the imbedding approach of Richard Bellman works out elegantly to derive a variety of results for a two-way jump Markov process with two barriers and with separable kernels. We treated the cases of both the barriers as absorbing or both reflecting and one barrier absorbing and the other reflecting. This technique is also carried out to calculate the expected amount of overflow in the storage model. One can easily note that the choice of the kernel decides the closed form analytical solution for the FPTD and the expected amount of overflow. The FPTD for emptiness is not exclusively treated in this work as the procedure will be totally similar to the case of first overflow. If the kernel is chosen in the form

$$k(x, y) = \frac{\lambda^{n+1}}{2n!} |x - y|^n e^{-\lambda|x-y|}$$

we will arrive at a differential equation of order  $2(n + 1)$  in the cases we discussed and in particular for the case  $n = 1$ , the differential equation is of order 4. This type of kernel will be studied in a later paper.

**Acknowledgements** The authors dedicate this work to Professor Ramakrishnan, the Founder Director of the Institute of Mathematical Sciences, Chennai (India). The first author acknowledges his initiation to research in general and in particular to Stochastic Processes and Application by Professor Ramakrishnan. The first author also acknowledges the support given by ISI (Bangalore Centre).

## References

- [1] Bellman, R.E. and Harris, T.E. (1978): On the theory of age dependent Stochastic Branching Processes. Proc. Nat. Acad. Sci. (USA).
- [2] Bellman, R.E. and Wing, M.C. (1976): An introduction to invariant imbedding, Academic, New York.
- [3] Chover, J. and Yeo, G.F. (1965): Solutions of some two sided boundary problems for some variables with alternating distributions, J. Appl. Prob. **2** 377–395.
- [4] Cinlar, A. and Pinsky, K. (1972): On dams with additive inputs and a general release rule, J. Appl. Prob. **9** 422–429.
- [5] Cochen, J.W. (1969): The single server queue, North Holland, Amsterdam.
- [6] Cramer Herlad (1955): Collective Risk theory, Jubilee volume of the Skandia Insurance Co., Stockholm.
- [7] Downton (1957): A note on Moran’s theory of dams, Quart. J. Math. **8** 282–286.
- [8] Feinberg, S.E. (1974): Stochastic Models for single neron firing trains – A survey, J. Biom. **30** 399–427.
- [9] Joe Gani (1955): Problems in probability theory of storage system, J.R. Stat. Soc. **B 19** 181–206.
- [10] Cochen, J. (1975): The Wiener-Hopt techniques in Applied Probability in “Perspectives in Probability and Statistics”, ed. J. Gani. Academic, London.
- [11] Gaver, D. and Muller, R.G. (1962): Limiting distributions for some storage problems, studies in Applied Probability and Management Sciences, ed. K.J. Arrow et al. Stanford University Press.
- [12] Holden, A.V. (1976): Models of the stochastic activities of neurons, lecture notes in Biomathematics, Vol. 12, Springer, New York.

- [13] Karlin, S. and Fabens, A.H. (1962): Generalized renewal functions stationary inventory models, *J. Math. Anal. Appl.* **5** 461–487.
- [14] Karlin, S. and Mermin (1959): The second order distribution of integrated shot noise. *IRE Trans. Inform. Theor.* 75–77.
- [15] Kendall, D.G. (1957): Some problems in the theory of dams, *J.R. Stat. Soc. B* **19** 207–212.
- [16] Moran, A.P. (1954): A probability theory of dams and storage systems, *Aust. J. Appl. Sci.* **15** 116–124.
- [17] Moran, A.P. (1959): *The theory of storage*, Methuen, London.
- [18] Peers, D. Stadje, W. and Zacks, S. (2002): First-Exit times for Compound Poisson processes for some types of positive and negative jumps, *Stoch. Models* **18** 139–157.
- [19] Perry, D., Stadje, W. and Zacks, S. (2007): Hysteratic capacity switching queues, *stoch. models* **23** 277–305.
- [20] Phatarfod, R.M. (1971): Some approximate results in renewal and dam theories, *J. Aust. Math. Soc. XII Part 4* 425–432.
- [21] Pollaczek, P. (1957): Problems stochastic poses parle phenomene, *Memorail des sciences mathematiques* 136.
- [22] Prabhu, N.U. (1964): Time dependent results in storage theory, *J. Appl. Prob.* **1** 1–46.
- [23] Prabhu, N.U. (1965): *Stochastic Processes*, Macmillan & Company, New York.
- [24] Puri, S. and Senthuria, J. (1975): An infinite dam with Poisson inputs and Poisson release, *Scand. Actur. J.* 193–202.
- [25] Ramakrishnan, A (1959): *Probability and Stochastic Processes in Handbuch Der Physik*, **4**, Springer, Berlin.
- [26] Redman, S.J. and Lamberd, D.G. (1968a): *J. Neurophysical* **31** 485–491.
- [27] Saaty, T.M. (1961): *Elements of queueing theory*, McGraw Hill, New York.
- [28] Srinivasan, S.K. (1974): Analytic solution of a finite dam governed by general input, *J. App. Prob.* **11** 133–144.
- [29] Srinivasan, S.K. and Sampath, G. (1977): *Stochastic models for space trains of single neurons*, lecture notes in Biomathematics, 16 Springer, New York.
- [30] Tackas, L. (1967): *Combinatorial models in the theory of Stochastic Processes*, Wiley, New York.
- [31] Vasudevan, R. and Vittal, P.R. (1985): Storage problems in continuous time with random inputs, random outputs and deterministic release, *Appl. Math. Comput.* **16** 309–326.
- [32] Vasudevan, R., Vittal, P.R. and Vijaykumar, A. (1979): On a class of two barrier problems, *Matscience* **97** 135–136.
- [33] Vittal, P.R., Jagadesan, T. and Muralidhar, V. (2008): Stochastic dynamics and passage times to diffusion approximations, *Differ. Equ. Dyn. Syst.* **16** 145–163.
- [34] Wald, A. (1947): *Sequential Analysis*, Wiley, New York.
- [35] Yeo, G.F. (1961): The time dependent solutions for an infinite dam with discrete additive inputs, *J.R. Stat. Soc. B* **13** 173–196.

**Part IV**  
**Theoretical Physics and Applied**  
**Mathematics**

# Inverse Consistent Deformable Image Registration

Yunmei Chen and Xiaojing Ye

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** This paper presents a novel variational model for inverse consistent deformable image registration. The proposed model deforms both source and target images simultaneously, and aligns the deformed images in the way that the forward and backward transformations are inverse consistent. To avoid the direct computation of the inverse transformation fields, our model estimates two more vector fields by minimizing their invertibility error using the deformation fields. Moreover, to improve the robustness of the model to the choice of parameters, the dissimilarity measure in the energy functional is derived using the likelihood estimation. The experimental results on clinical data indicate the efficiency of the proposed method with improved robustness, accuracy, and inverse consistency.

**Mathematics Subject Classification (1991)** Primary 62H35; Secondary 65K10

**Key words and phrases** Image registration · Inverse consistent · Variational method · Optimization

## 1 Introduction

Image registration is a very important subject that has been widely applied in medical research and clinical applications. The task of image registration is to find a transformation field that relates points in the source image to their corresponding points in the target image. Deformable image registration allows localized transformations, and is able to account for internal organ deformations. Therefore, it has been increasingly used in health care to assist diagnosis and treatments.

---

Y. Chen and X. Ye  
Department of Mathematics, University of Florida, Gainesville, FL 32611, USA  
e-mail: [yun@ufl.edu](mailto:yun@ufl.edu); [xye@ufl.edu](mailto:xye@ufl.edu)

In particular, deformable image registration has become a critical technique for image guided radiation therapy. It allows more precise tumor targeting and normal tissue preservation. A comprehensive review of image registration in radiation therapy can be found in [Kes06].

A deformable image registration is called inverse consistent, if the correspondence between two images is invariant to the order of the choice of source and target. More precisely, let  $S$  and  $T$  be the source and target images, and  $h$  and  $g$  be the forward and backward transformations, respectively, i.e.,

$$S \circ h = T \quad \text{and} \quad T \circ g = S,$$

then an inverse consistent registration satisfies  $h \circ g = id$  and  $g \circ h = id$ , where  $id$  is the identity map. This can be illustrated by the following diagram with constraints  $g = h^{-1}$ ,  $h = g^{-1}$ :

$$\boxed{S} \begin{array}{c} \xleftarrow{h} \\ \xrightarrow{g} \end{array} \boxed{T}, \quad (1)$$

where each of the two squares in (1) represents the domain on which the labeled image is defined. By applying an inverse consistent registration, measurements, or segmentations on one image can be precisely transferred to the other. In imaging guided radiation therapy, the inverse consistent deformable registration technique provides the voxel-to-voxel mapping between the reference phase and the test phase in four-dimensional (4D) radiotherapy [LOC<sup>+</sup>06]. This technique is referred to “automatic recontouring.”

Inverse consistent deformable image registration has been an active subject of study in the literature. There has been a group of work developed in the context of large deformation by diffeomorphic metric mapping, e.g. [HC03, JDJG04, AGG06, BK07]. The main idea of this method is modeling the forward and backward transformations as a one-parameter diffeomorphism group. Then, a geodesic path connecting two images is obtained by minimizing an energy functional symmetric to the forward and backward transformations. This type of models produce a very good registration results. However, it take long time to compute, because strong regularization of the mappings are required.

Variational method is one of the popular approaches for inverse consistent deformable image registration. This method minimizes an energy functional(s) symmetric to the forward and backward transformations, and in general, consists of three parts: regularization of deformation fields, dissimilarity measure of the target and deformed source images, and penalty of inverse inconsistency [CJ01, ADPS02, RK06, ZJT06]. In [CJ01], Christensen and Johnson proposed to minimize the following coupled energy functionals with respect to  $h$  and  $g$  alternately:

$$\begin{cases} E(h) = \lambda E_s(S \circ h, T) + E_r(u) + \rho \|h - g^{-1}\|_{L^2(\Omega)}^2 \\ E(g) = \lambda E_s(T \circ g, S) + E_r(v) + \rho \|g - h^{-1}\|_{L^2(\Omega)}^2 \end{cases}, \quad (2)$$

where  $u$  and  $v$  are forward and backward deformation fields corresponding to  $h$  and  $g$ , respectively, i.e.,  $h(x) = x + u(x)$  and  $g(x) = x + v(x)$ . The dissimilarity measure  $E_s$  and the regularization of the deformation field  $E_r$  are defined by:

$$E_s(S \circ h, T) = \|S \circ h - T\|_{L^2(\Omega)}^2, \quad E_r(u) = \|a\Delta u + b\nabla(\operatorname{div} u) - cu\|_{L^2(\Omega)}^2$$

with positive constants  $a, b, c > 0$ . The last term in both energy functionals enforces the inverse consistency of  $h$  and  $g$ . The solution  $(u, v)$  to (2) is obtained by iteratively solving a system of two evolution equations associated with their Euler-Lagrange (EL) equations. This model gives considerably good results with parameters chosen carefully. However, it needs to compute the inverse mappings  $g^{-1}$  and  $h^{-1}$  explicitly in each iteration, which is computationally intensive can cause cumulated numerical errors in the estimation of inverse mappings.

The variational models developed in [ADPS02] and [ZJT06] have the same framework as in [CJ01], but with different representations of  $E_s$ ,  $E_r$ , and inverse consistent constraints. In [ADPS02] and [ZJT06] the terms  $\|h \circ g(x) - x\|_{L^2(\Omega)}^2$  and  $\|g \circ h(x) - x\|_{L^2(\Omega)}^2$  are used in the energy functional to enforce the inverse consistency. By using these terms the explicit computation of the inverse transforms of  $h$  and  $g$  can be avoided during the process of finding optimal forward and backward transformations. The similarity measure in [ZJT06] is mutual information for multimodal image registration. The  $E_s(S \circ h, T)$  in [ADPS02] is  $\|S \circ h - T\|_{L^2(\Omega)}^2 / \max |DT|$ . The regularization term  $E_r(u)$  in [ZJT06] is a function of  $Du$ , and that in [ADPS02] is a tensor based smoothing which is designed to prevent the transformation fields from being smoothed across the boundaries of features. In [YS05, YTS<sup>+</sup>08] the proposed models incorporated stochastic errors in the inverse consistent constraints for both forward and backward transformations.

In [LHG<sup>+</sup>05], Leow et al. proposed a nonvariational approach that updates the forward and backward transformations simultaneously by a force that reduces the first two terms in  $E(h)$  and  $E(g)$  in (2) and preserves the inverse consistency. However, in order to simplify the computation this algorithm only takes linear order terms in the Taylor expression to approximate the inverse consistent conditions for updated transformation fields. As a consequence, the truncating errors can be accumulated and exaggerated during iterations. This can lead to large inverse consistent error, despite that it can produce a good matching quickly [ZC08].

In this paper we propose a novel variational model to improve the accuracy, robustness and efficiency of inverse consistent deformable registration. As an alternate to the current framework of variational methods which finds the forward and backward transformations that deform a source image  $S$  to match a target image  $T$  and vice versa, we propose to deform  $S$  and  $T$  simultaneously, and let the registration align the deformed source and deformed target images. It is clear that the disparity between deformed  $S$  and deformed  $T$  is smaller than that between deformed  $S$  and fixed  $T$  or deformed  $T$  and fixed  $S$ . Therefore, the deformation by the bidirectional simultaneous deformations is in general smaller than the deformation by unidirectional deformation that deforms  $S$  full way to  $T$  or  $T$  full way to  $S$ . As shown in



Sect. 5, deforming  $S$  and  $T$  simultaneously leads to a faster and better alignment than deforming  $S$  to the fixed  $T$  or vice versa. Let  $\phi$  and  $\tilde{\phi}$  represent the transformation fields such that  $S \circ \phi$  matches  $T \circ \tilde{\phi}$ . It is not difficult to verify that if  $\phi$  and  $\tilde{\phi}$  are invertible, then the registrations from  $S$  to  $T$ , and  $T$  to  $S$  are inverse consistent. To avoid the direct computation of the inverse transformations of  $\phi$  and  $\tilde{\phi}$ , our model seeks for two additional deformation fields  $\psi$ ,  $\tilde{\psi}$  such that  $\phi$  and  $\psi$  are inverse to each other, and the same for  $\tilde{\phi}$  and  $\tilde{\psi}$ . Moreover, the registration process enforces certain regularization of these four deformation fields, and aligns the deformed  $S$  and deformed  $T$ . Then, the optimal inverse consistent transformations from  $S$  to  $T$ , and  $T$  to  $S$  can be obtained simply by appropriate compositions of these four transformations.

The idea of deforming  $S$  and  $T$  simultaneously has been adopted in the models where the forward or backward transformation is modeled as a one-parameter diffeomorphism group [AGG06]. However, our model finds regularized invertible deformation fields by minimizing the  $L^2$  norms of the deformation fields and inverse consistent errors rather than a one-parameter diffeomorphism group, whose computational cost is very expensive and hence hinders its application in clinical use. Moreover, our model allows parallel computations for all the deformation fields to significantly reduce the computational time.

Furthermore, to improve the robustness of the model to noises and the choice of the parameter  $\lambda$  that balances the goodness of matching and smoothness of the deformation fields (see the  $\lambda$  in  $E(h)$  and  $E(g)$  of (2)), we adopt the maximum likelihood estimate (MLE) that is able to accommodate certain degree of variability in matching to improve the robustness and accuracy of the registration. By using MLE, the ratio of weighting parameters on the sum of squared distance (SSD) of the residue image  $S \circ \phi - T \circ \tilde{\phi}$  and the regularization term is not a fixed  $\lambda$ , but  $\lambda/\sigma^2$  (see (18) later). This results in a self-adjustable weighting factor that makes the choice of  $\lambda$  more flexible, and also speeds up the convergence to the optimal deformation field.

The rest of the paper is organized as follows. In Sect. 2, we present a detailed description of the proposed model. The existence of solutions to the proposed model is shown in Sect. 3. The calculus of variation and an outline of a fast algorithm for solving the proposed model numerically are provided in Sect. 4. In Sect. 5, we present the experimental results on clinical data, and the application in auto recontouring. The last section concludes the paper.

## 2 Proposed Method

Let  $S$  and  $T$  be the source and target images defined on  $\Omega_S$  and  $\Omega_T$  in  $\mathbb{R}^d$ , respectively. Note that, in real applications,  $\Omega_S$  and  $\Omega_T$  are usually fully overlapped. For simplicity we assume that images  $S$  and  $T$  are real-valued functions with continuous derivatives. Let  $|\cdot|$  denote the absolute value (length) of a scalar (vector) in Euclidean spaces, and  $\|\cdot\|$  denote  $\|\cdot\|_{L^2(\Omega)}$  henceforth. We also extend this notation

to vector-valued functions whose components are in  $L^2$  or  $H^1$ :  $u = (u_1, \dots, u_d)^\top$  with each component  $u_j \in H^1(\Omega)$ ,  $j = 1, \dots, d$ , there is

$$\|u\|_{H^1(\Omega)} \triangleq (\|u\|^2 + \|Du\|^2)^{1/2}$$

and

$$\|u\| \triangleq \left( \sum_{j=1}^d \|u_j\|^2 \right)^{1/2}, \quad \|Du\| \triangleq \left( \sum_{j=1}^d \|Du_j\|^2 \right)^{1/2},$$

where

$$\|u_j\| = \left( \int_{\Omega} |u_j(x)|^2 dx \right)^{1/2} \quad \text{and} \quad \|Du_j\| = \left( \int_{\Omega} |Du_j(x)|^2 dx \right)^{1/2},$$

for  $j = 1, \dots, d$ .

## 2.1 Motivation and Ideas of Proposed Method

In this paper, we propose a novel variational model for inverse consistent deformable registration to improve its efficiency and robustness. Our idea differs from the current framework which deforms source image  $S$  to target image  $T$ , or vice versa: as an alternate, we propose to deform  $S$  and  $T$  simultaneously, and match both deformed images. This means that ideally we pursue for a pair of half-way transforms  $\phi : \Omega_S \rightarrow \Omega_M$  and  $\tilde{\phi} : \Omega_T \rightarrow \Omega_M$  such that  $S \circ \phi = T \circ \tilde{\phi}$ , where  $\Omega_M$  is the region where  $S \circ \phi$  and  $T \circ \tilde{\phi}$  have overlap. To ensure the transformations from  $S$  to  $T$  and  $T$  to  $S$  are inverse consistent, the transforms  $\phi$  and  $\tilde{\phi}$  are required to be invertible (but not necessarily to be inverse to each other). Hence, our purpose is to find the transformations  $\phi$  and  $\tilde{\phi}$  such that

$$S \circ \phi = T \circ \tilde{\phi}, \quad \phi, \tilde{\phi} \text{ invertible.} \quad (3)$$

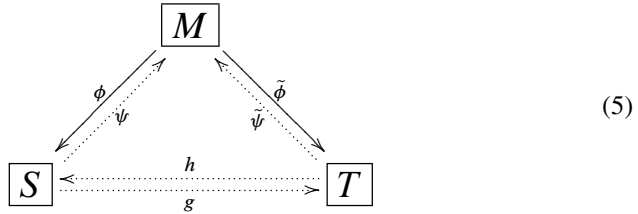
To avoid direct computation of inverses of  $\phi$  and  $\tilde{\phi}$  during iterations, we enforce the invertibility of  $\phi$  and  $\tilde{\phi}$  by finding another two transformations  $\psi : \Omega_M \rightarrow \Omega_S$  and  $\tilde{\psi} : \Omega_M \rightarrow \Omega_T$  such that

$$\begin{aligned} \psi \circ \phi &= id, & \phi \circ \psi &= id, \\ \tilde{\psi} \circ \tilde{\phi} &= id, & \tilde{\phi} \circ \tilde{\psi} &= id. \end{aligned} \quad (4)$$

Once we obtained such  $\psi$  and  $\tilde{\psi}$ , we can construct the objective full-way transformations  $h$  and  $g$  as follows,

$$h = \phi \circ \tilde{\psi}, \quad g = \tilde{\phi} \circ \psi.$$

It is easy to see that  $h$  and  $g$  satisfy the inverse consistent constraints  $h \circ g = g \circ h = id$ . This idea is illustrated by the following diagram, where  $M$  is an intermediate image.



As by deforming  $S$  and  $T$  simultaneously the difference between deformed  $S$  and deformed  $T$  at each iteration, in general, is smaller than that between deformed  $S$  and fixed  $T$ , or deformed  $T$  and fixed  $S$ , the computational cost of deforming both  $S$  and  $T$  is much less than the conventional one that deform  $S$  all the way to  $T$  and  $T$  to  $S$ . In particular, if the underlying deformations of  $h$  and  $g$  are large, deforming both  $S$  and  $T$  can make the each deformation of  $\phi$  and  $\tilde{\phi}$  in the proposed model almost half smaller than that of  $h$  and  $g$ , and achieve a faster convergence for the computation of  $\phi$  and  $\tilde{\phi}$ . Also, seeking  $\psi$  and  $\tilde{\psi}$  along with  $\phi$  and  $\tilde{\phi}$  avoids direct computation of inverse transformations in each iteration as that in (4), which usually causes cumulated errors during iterations if using approximations of the inverses.

Moreover, regularizing the deformation fields is very important to obtain physically meaningful and accurate registrations. Also, if the energy functional consists of only dissimilarity measures and invertible constraints, it is ill-posed in general. Therefore, we propose the following framework for deformable inverse consistent registration:

$$\min_{\phi, \tilde{\phi}, \psi, \tilde{\psi}} R(\phi, \tilde{\phi}, \psi, \tilde{\psi}) + \text{dis}(S \circ \phi, T \circ \tilde{\phi}), \quad \text{s.t. condition (4) holds} \quad (6)$$

where  $R$  is a regularization operator of its arguments,  $\text{dis}(S \circ \phi, T \circ \tilde{\phi})$  measures the dissimilarity between  $S \circ \phi$  and  $T \circ \tilde{\phi}$ .

### 2.2 Alternative Formulation of (4) Using Deformation Fields

Let the functions  $u, \tilde{u}, v$  and  $\tilde{v}$  represent the corresponding deformation fields of the transformations  $\phi, \tilde{\phi}, \psi$  and  $\tilde{\psi}$ , respectively. That is,

$$\begin{aligned} \phi(x) &= x + u(x), & \tilde{\phi}(x) &= x + \tilde{u}(x), \\ \psi(x) &= x + v(x), & \tilde{\psi}(x) &= x + \tilde{v}(x). \end{aligned} \quad (7)$$

Then, the constraints in (4) can be rewritten as:

$$\begin{aligned} u + v(x + u) &= v + u(x + v) = 0, \\ \tilde{u} + \tilde{v}(x + \tilde{u}) &= \tilde{v} + \tilde{u}(x + \tilde{v}) = 0. \end{aligned} \quad (8)$$

### 2.3 MLE Based Derivation for $\text{dis}(S \circ \phi, T \circ \tilde{\phi})$

To improve the robustness of the algorithm for deformable image registration, we use the negative log-likelihood of the residue image as a measure of mismatching. Consider voxel intensities of the residue image defined by:

$$W(x) \triangleq S \circ \phi(x) - T \circ \tilde{\phi}(x), \quad x \in \Omega_M,$$

as independent samples drawn from a Gaussian distribution of mean zero and variance  $\sigma^2$  to be optimized (see remark later for the reason of this assumption), whose probability density function (pdf) is denoted by  $P(\cdot|\sigma)$ . Then the likelihood of the residual image  $W(x)$  can be computed as:

$$\mathcal{L}(\sigma|\{W(x), x \in \Omega\}) = \prod_{x \in \Omega} P(W(x)|\sigma) = \prod_{x \in \Omega} \left( \frac{1}{\sqrt{2\pi}\sigma} e^{-|S \circ \phi - T \circ \tilde{\phi}|^2/2\sigma^2} \right). \quad (9)$$

Then, by writing the summation over all  $x \in \Omega$  as an integral over  $\Omega$  the negative log-likelihood function is given as follows:

$$\|S \circ \phi - T \circ \tilde{\phi}\|^2/2\sigma^2 + |\Omega| \log \sqrt{2\pi}\sigma.$$

Omitting the constant  $|\Omega| \log \sqrt{2\pi}$ , we define the dissimilarity term as:

$$\text{dis}(S \circ \phi, T \circ \tilde{\phi}) \triangleq \|S \circ \phi - T \circ \tilde{\phi}\|^2/2\sigma^2 + |\Omega| \log \sigma. \quad (10)$$

which can be rewritten as our MLE fitting term  $F$  by using corresponding deformation fields  $u$  and  $\tilde{u}$ :

$$\begin{aligned} F(u, \tilde{u}, \sigma) &\triangleq \text{dis}(S(x + u), T(x + \tilde{u})) \\ &= \|S(x + u) - T(x + \tilde{u})\|^2/2\sigma^2 + |\Omega| \log \sigma. \end{aligned} \quad (11)$$

*Remark 2.1.* Let  $\hat{P}$  be the estimation of the pdf for the random variable  $X \triangleq W(x)$ ,  $x \in \Omega$ . We show later why it is reasonable to assume  $\hat{P}$  to be a Gaussian distribution of zero mean and variance  $\sigma^2$ .

In fact,  $\hat{P}$  is a function in  $C_0(\mathbb{R})$ , the space of all the continuous functions on real line vanishing at infinity with the supreme norm. Let  $H_0(\mathbb{R})$  be the Hilbert

space consisting of all linear combinations of  $\kappa(x_l, x)$  for finite many of  $x_l \in \mathbb{R}$ , where

$$\kappa(x_l, x) = (2\pi\sigma^2)^{-1/2} e^{-(x_l-x)^2/2\sigma^2}, \quad \forall x \in \mathbb{R}. \tag{12}$$

Define an inner product on  $H_0(\mathbb{R})$  by:

$$\left\langle \sum_{i=1}^m a_i \kappa(x_i, \cdot), \sum_{j=1}^n b_j \kappa(y_j, \cdot) \right\rangle_{H_0(\mathbb{R})} = \sum_{i=1}^m \sum_{j=1}^n a_i b_j \kappa(x_i, y_j).$$

We claim that

$$H_0(\mathbb{R}) \text{ is dense in } C_0(\mathbb{R}). \tag{13}$$

In fact, if the claim (13) is not true, by Hahn-Banach theorem there exists a bounded signed measure  $m$  in the dual space of  $C_0(\mathbb{R})$ , such that

$$\int_{\mathbb{R}} \hat{P} dm \neq 0, \tag{14}$$

but  $\int_{\mathbb{R}} f dm = 0$ , for all  $f \in H_0(\mathbb{R})$ . In particular, for any  $x \in \mathbb{R}$ ,

$$\int_{\mathbb{R}} \kappa(x, y) dm_y = 0,$$

where  $\kappa(\cdot, \cdot)$  is as in (12), and hence,

$$\int_{\mathbb{R} \times \mathbb{R}} \kappa(x, y) dm_x dm_y = 0.$$

This implies  $m = 0$ , which contradicts (14). Therefore, the claim holds.

By this claim it is easy to see that

$$\hat{P}(z) \approx \sum_{l=1}^k \alpha_l \kappa(x_l, z) = (2\pi\sigma^2)^{-1/2} \sum_{l=1}^k \alpha_l e^{-(x_l-z)^2/2\sigma^2} \tag{15}$$

for some  $\{x_l; \alpha_l\}_{l=1}^k$ . Since a good registration requires the intensities of the residue image  $W(x)$  close to zero. Hence, in (15) the only dominate term in the sum should be the one corresponding to  $x_l = 0$ , and other terms are negligible. This means that  $\hat{P}$  is approximately  $\mathcal{N}(0, \sigma^2)$ , the Gaussian distribution with mean 0 and variance  $\sigma^2$ .

### 2.4 Proposed Model

Based on the discussion earlier, we are ready to present the proposed model. We define the regularization term  $R(\phi, \tilde{\phi}, \psi, \tilde{\psi})$  in (6) using their corresponding deformation fields as

$$R(\phi, \tilde{\phi}, \psi, \tilde{\psi}) = R(u, \tilde{u}, v, \tilde{v}) \triangleq \|Du\|^2 + \|D\tilde{u}\|^2 + \|Dv\|^2 + \|D\tilde{v}\|^2. \tag{16}$$

By plugging (16) and (11) into (6), and replacing the constraint in (6) by (8), the proposed model can be written as:

$$\min_{u, \tilde{u}, v, \tilde{v}, \sigma} R(u, \tilde{u}, v, \tilde{v}) + \lambda F(u, \tilde{u}, \sigma), \quad \text{s.t. condition (8) holds,} \quad (17)$$

where  $R(u, \tilde{u}, v, \tilde{v})$  and  $F(u, \tilde{u}, \sigma)$  are defined in (16) and (11), respectively.

To solve problem (17), we relax the equality constraints of inverse consistency, and penalize their violation using quadratic functions, then write it as an unconstrained energy minimization problem

$$\min_{u, \tilde{u}, v, \tilde{v}, \sigma} R(u, \tilde{u}, v, \tilde{v}) + \lambda F(u, \tilde{u}, \sigma) + \mu (\mathcal{I}(u, v) + \mathcal{I}(\tilde{u}, \tilde{v})), \quad (18)$$

where and  $\mathcal{I}(u, v)$  is the cost of inverse inconsistency of  $u$  and  $v$ :

$$\mathcal{I}(u, v) = \mathcal{I}_v(u) + \mathcal{I}_u(v), \quad (19)$$

with

$$\mathcal{I}_v(u) = \|u + v(x + u)\|^2 \quad \text{and} \quad \mathcal{I}_u(v) = \|v + u(x + v)\|^2. \quad (20)$$

Similarly, we have  $\mathcal{I}(\tilde{u}, \tilde{v})$ . With sufficiently large  $\mu$ , solving (18) gives an approximation to the solution of (17).

The term  $F(u, \tilde{u}, \sigma)$  is from the negative log-likelihood of the residue image (11). Minimizing this term forces the mean of the residue image to be zero, but allows it to have a variance to accommodate certain variability. This makes the model more robust to noise and artifacts, and less sensitive to the choice of the parameter  $\lambda$  than the model using the SSD, i.e., the squared  $L^2$ -norm, of the residue image as a dissimilarity measure as in (2). The parameter  $\lambda$  balances the smoothness of deformation fields and goodness of alignments, and affects the registration result significantly. In the proposed model, the ratio of the SSD of the residue image over the smoothing terms is  $\lambda/\sigma^2$  rather than a prescribed  $\lambda$ . Since  $\sigma$  is to be optimized, and from its EL equation  $\sigma$  is the standard deviation of the residue image. Therefore, in the proposed model the weight on the matching term updates during iterations. When the alignment gets better,  $\sigma$  the standard deviation of the residue as shown in (35) decreases, and hence the weight on the matching term automatically increases. This self-adjustable feature of the weight not only enhances the accuracy of alignment but also makes the choice of  $\lambda$  flexible, and results in a fast convergence.

As shown earlier, the final forward and backward transforms  $h$  and  $g$  can be obtained by:

$$h = \phi \circ \tilde{\psi} = x + \tilde{v} + u(x + \tilde{v}) \quad \text{and} \quad g = \tilde{\phi} \circ \psi = x + \tilde{u} + v(x + \tilde{u}).$$

Thus, the corresponding final full-way forward and backward deformation fields  $\tilde{u}$  and  $\tilde{v}$  are given as:

$$\tilde{u} = \tilde{v} + u(x + \tilde{v}) \quad \text{and} \quad \tilde{v} = \tilde{u} + v(x + \tilde{u}), \quad (21)$$

respectively. Then the inverse consistent constraints (4) can be represented using  $\bar{u}, \bar{v}$  as follows:

$$\bar{u} + \bar{v}(x + \bar{u}) = \bar{v} + \bar{u}(x + \bar{v}) = 0. \tag{22}$$

### 3 Existence of Solutions

In this section we prove the existence of solutions  $(u, \tilde{u}, v, \tilde{v}, \sigma)$  to the proposed model (18). For simplicity, we assume that both  $S$  and  $T$  defined on the same domain  $\Omega$ , which is simply connected, closed and bounded in  $\mathbb{R}^d$  with Lipschitz boundary  $\partial\Omega$ . Also  $S, T \in C^1(\Omega)$ . As in reality, deformation field cannot be unbounded, we restrict  $u, \tilde{u}, v, \tilde{v}$  to be in a closed subset of  $L^\infty(\Omega)$ :

$$\mathcal{B} \triangleq \{u \in L^\infty(\Omega) : \|u\|_{L^\infty(\Omega)} \leq B, B \in \mathbb{R}_+ \text{ only depends on } \Omega\}$$

Then, we seek solutions  $(u, \tilde{u}, v, \tilde{v}, \sigma)$  to the problem (18) in the spaces  $u, \tilde{u}, v, \tilde{v} \in H^1(\Omega) \cap \mathcal{B}$  and  $\sigma \in \mathbb{R}_+$ . For short notations, we let  $w$  denote the quaternion  $(u, \tilde{u}, v, \tilde{v})$ . Then, we show the existence of solutions to the following minimization problem:

$$\min_{(w, \sigma) \in (H^1 \cap \mathcal{B}) \times \mathbb{R}_+} E(w, \sigma) \tag{23}$$

where

$$E(w, \sigma) = \|Dw\|^2 + \lambda F(w, \sigma) + \mu \mathcal{I}(w)$$

and  $F$  and  $\mathcal{I}$  are defined correspondingly in (18) using the simplified notation of  $w$ , i.e.,

$$\begin{aligned} \|Dw\|^2 &= \|Du\|^2 + \|D\tilde{u}\|^2 + \|Dv\|^2 + \|D\tilde{v}\|^2, \\ F(w, \sigma) &= \|S(x + u) - T(x + \tilde{u})\|^2 / \sigma^2 + |\Omega| \log \sigma, \\ \mathcal{I}(w) &= \mathcal{I}_v(u) + \mathcal{I}_u(v) + \mathcal{I}_{\tilde{v}}(\tilde{u}) + \mathcal{I}_{\tilde{u}}(\tilde{v}). \end{aligned}$$

and the terms on the right side of  $\mathcal{I}(w)$  are defined as in (20). The  $\lambda$  and  $\mu$  are prescribed positive constants.

**Theorem 3.1.** *The minimization problem (23) admits solutions  $(w, \sigma) \in (H^1 \cap \mathcal{B}) \times \mathbb{R}_+$ .*

*Proof.* For  $(w, \sigma) \in (H^1 \cap \mathcal{B}) \times \mathbb{R}_+$ ,  $E(w, \sigma)$  is bounded below. Hence, there exists a minimizing sequence  $\{(w_k, \sigma_k)\}_{k=1}^\infty \subset (H^1 \cap \mathcal{B}) \times \mathbb{R}_+$  such that

$$\lim_{k \rightarrow \infty} E(w_k, \sigma_k) = \inf_{(H^1 \cap \mathcal{B}) \times \mathbb{R}_+} E(w, \sigma).$$

Therefore,  $\{\|Dw_k\|\}_{k=1}^\infty$  are uniformly bounded. Along with  $w_k \in \mathcal{B}$  we know that  $\{w_k\}_{k=1}^\infty$  is a bounded sequence in  $H^1$ . By the weak compactness of  $H^1$  and the

fact that  $H^1$  is precompact in  $L^2$ , there exists a convergent subsequence, which is still denoted by  $\{w_k\}_{k=1}^\infty$ , and a function  $\hat{w} \in H^1$ , such that

$$w_k \rightharpoonup \hat{w} \text{ weakly in } H^1, \tag{24}$$

$$w_k \rightarrow \hat{w} \text{ strongly in } L^2, \text{ and a.e. in } \Omega. \tag{25}$$

Moreover, since  $E(w_k, \sigma_k) \rightarrow \infty$  if  $\sigma_k \rightarrow 0$  or  $\infty$ , there is a constant  $C > 0$  such that  $\{\sigma_k\}_{k=1}^\infty$  are bounded below and above by  $1/C$  and  $C$ , respectively. Hence, there is a subsequence of  $\{\sigma_k\}_{k=1}^\infty$  and a scalar  $\hat{\sigma} \in \mathbb{R}_+$ , without changing the notation for the subsequence we have

$$\sigma_k \rightarrow \hat{\sigma} \in \mathbb{R}_+. \tag{26}$$

From the weak lower semicontinuity of norms and (24), we know

$$\|D\hat{w}\|^2 \leq \liminf_{k \rightarrow \infty} \|Dw_k\|^2. \tag{27}$$

Also, as  $\mathcal{I}(w) \leq 8B$  for any  $w \in H^1 \cap \mathcal{B}$  and  $w_k \rightarrow \hat{w}$  a.e. in  $\Omega$ , we get, by dominant convergence theorem, that

$$\lim_{k \rightarrow \infty} \mathcal{I}(w_k) = \mathcal{I}(\hat{w}). \tag{28}$$

By the same argument with the smoothness of  $S$  and  $T$ , the convergence of  $\{\sigma_k\}_{k=1}^\infty$ , and the fact that  $w_k \rightarrow \hat{w}$  a.e. in  $\Omega$ , we can also have

$$\lim_{k \rightarrow \infty} F(w_k, \sigma_k) = F(\hat{w}, \hat{\sigma}) \tag{29}$$

Combining (27), (28) with (29), we obtain that

$$E(\hat{w}, \hat{\sigma}) \leq \liminf_{k \rightarrow \infty} E(w_k, \sigma_k) = \inf_{(H^1 \cap \mathcal{B}) \times \mathbb{R}_+} E(w, \sigma).$$

Furthermore, because  $\{w_k\}_{k=1}^\infty \subset \mathcal{B} \subset L^\infty(\Omega)$ , we know

$$w_k \rightharpoonup^{W*} \hat{w} \text{ weakly* in } L^\infty$$

and hence  $\hat{w} \in \mathcal{B}$ . Therefore,  $(\hat{w}, \hat{\sigma}) \in (H^1 \cap \mathcal{B}) \times \mathbb{R}_+$ . Hence

$$E(\hat{w}, \hat{\sigma}) = \inf_{(H^1 \cap \mathcal{B}) \times \mathbb{R}_+} E(w, \sigma).$$

which implies that  $(\hat{w}, \hat{\sigma})$  is a solution to the minimization problem (23). □



### 4 Numerical Scheme

In this section, we provide the numerical scheme for solving (18). As the compositions in the inverse consistency constraints  $\mathcal{I}_u(v)$  and  $\mathcal{I}_v(u)$  bring a difficulty in getting an explicit form of the EL equations for the deformation fields and their inverses, in our computation, instead of directly solving (18), we solve the following two coupled minimization problems alternately:

$$\begin{cases} \min_{u, \tilde{u}} E_{v, \tilde{v}, \sigma}(u, \tilde{u}) \\ \min_{v, \tilde{v}} E_{u, \tilde{u}}(v, \tilde{v}) \end{cases} \tag{30}$$

where

$$E_{v, \tilde{v}, \sigma}(u, \tilde{u}, \sigma) = \|Du\|^2 + \|D\tilde{u}\|^2 + \lambda F(u, \tilde{u}, \sigma) + \mu (\mathcal{I}_v(u) + \mathcal{I}_{\tilde{v}}(\tilde{u})) \tag{31}$$

and

$$E_{u, \tilde{u}}(v, \tilde{v}) = \|Dv\|^2 + \|D\tilde{v}\|^2 + \mu (\mathcal{I}_u(v) + \mathcal{I}_{\tilde{u}}(\tilde{v})). \tag{32}$$

By taking first variation with respect to  $u, \tilde{u}, v, \tilde{v}$ , we get the EL equations:

$$\begin{cases} -\Delta u + \frac{\lambda}{\sigma^2} W_{u, \tilde{u}} DS(x + u) + \mu \langle I + Dv(x + u), u + v(x + u) \rangle = 0 \\ -\Delta v + \mu \langle I + Du(x + v), v + u(x + v) \rangle = 0 \\ -\Delta \tilde{u} - \frac{\lambda}{\sigma^2} W_{u, \tilde{u}} DT(x + \tilde{u}) + \mu \langle I + D\tilde{v}(x + \tilde{u}), \tilde{u} + \tilde{v}(x + \tilde{u}) \rangle = 0 \\ -\Delta \tilde{v} + \mu \langle I + D\tilde{u}(x + \tilde{v}), \tilde{v} + \tilde{u}(x + \tilde{v}) \rangle = 0 \end{cases}, \tag{33}$$

in  $\Omega$ , with free Neumann boundary conditions for each of them on  $\partial\Omega$ :

$$\langle Du, n \rangle = \langle D\tilde{u}, n \rangle = \langle Dv, n \rangle = \langle D\tilde{v}, n \rangle = 0, \quad \text{on } \partial\Omega, \tag{34}$$

where  $W_{u, \tilde{u}} \triangleq S(x + u) - T(x + \tilde{v})$ ,  $I$  is the identity matrix of size  $d$ , and  $n$  is the outer normal of  $\partial\Omega$ . Also, the first variation of  $\sigma$  gives

$$\sigma = \|S(x + u) - T(x + \tilde{u})\|/|\Omega|^{1/2}. \tag{35}$$

The solution to the EL equations (33) can be obtained by finding the stationary solution to the evolution equations associated with the EL equations. In numerical implementation, we use semi-implicit discrete form of the evolution equations. The additive operator splitting (AOS) scheme was applied to solve the problem faster [WtHRV98]. An alternative way of AOS to solve the semi-implicit discrete evolution equation in this case can be obtained by applying discrete cosine transforms (DCT) to diagonalize the Laplace operator with the assumption that the deformation fields have symmetric boundary condition, which is compatible with (34).

In two-dimensional (2D) case, the semi-implicit discrete form of (33) with fixed step sizes  $\tau_u, \tau_v$  for the evolution equations of  $u^{(k+1)}$  as:

$$\frac{u_{i,j}^{(k+1)} - u_{i,j}^{(k)}}{\tau_u} = \Delta_{i,j} u^{(k+1)} - D_{i,j} \left( \lambda F \left( u^{(k)}, \tilde{u}^{(k)}, \sigma^{(k)} \right) + \mu \mathcal{I}_{v^{(k)}} \left( u^{(k)} \right) \right), \quad (36)$$

and  $v^{(k+1)}$  as

$$\frac{v_{i,j}^{(k+1)} - v_{i,j}^{(k)}}{\tau_v} = \Delta_{i,j} v^{(k+1)} - \mu D_{i,j} \mathcal{I}_{u^{(k)}} \left( v^{(k)} \right), \quad (37)$$

where  $\Delta_{i,j}$  and  $D_{i,j}$  represent the discrete Laplacian and gradient operators at the pixel indexed by  $(i, j)$ , respectively. The 3D case is a simple analog with one more subscript in indices. Similarly, we have the discrete evolution equation for  $\tilde{u}$  and  $\tilde{v}$  with the two components within each of the three pairs  $(u, \tilde{u})$ ,  $(v, \tilde{v})$  and  $(S, T)$  switched in (36) and (37). With AOS scheme being applied, the computation for each update of  $u$  involves of solving  $d$  tridiagonal systems whose computational costs are linear in  $N$ , where  $N$  is the total number of pixels in  $S$  (or  $T$ ). Also, in each iteration of updating  $u$  and  $v$ , there needs  $2(d+1)$  interpolations with size  $N$ . It is important to point out that, in each iteration, the computations of  $u, \tilde{u}, v, \tilde{v}$  can be carried out in parallel. We summarize **icDIR** in Algorithm 1, where the maximum inverse consistency error (ICE)  $\delta_c$  is defined by:

$$\delta_c = \max_x \{ |\bar{u} + \bar{v}(x + \bar{u})|, |\bar{v} + \bar{u}(x + \bar{v})| \}, \quad (38)$$

and  $\bar{u}$  and  $\bar{v}$  are the final full-way deformation fields shown in (21). That is, it measures the maximum ICE of deformations obtained by quaternion  $(u, \tilde{u}, v, \tilde{v})$ . The parameter  $\mu$  in (18) may increase during iterations to ensure smaller ICE. In each inner loop with fixed  $\mu$ , the computation is terminated when the mean of

---

### Algorithm 1 Inverse Consistent Deformable Image Registration (**icDIR**)

---

Input  $S, T$ , and  $\tau_u, \tau_v, \lambda, \mu > 0, \epsilon = 0.5, \delta_c = 1$ . Initialize  $(u^{(0)}, \tilde{u}^{(0)}, v^{(0)}, \tilde{v}^{(0)}) = 0, k = 0$ .  
**while**  $\delta_c \geq \epsilon$  **do**  
  **repeat**  
    {All terms in  $(u^{(k+1)}, \tilde{u}^{(k+1)}, v^{(k+1)}, \tilde{v}^{(k+1)})$  can be calculated in parallel}  
    Calculate  $(u^{(k+1)}, v^{(k+1)})$  using (36) and (37).  
    Calculate  $(\tilde{u}^{(k+1)}, \tilde{v}^{(k+1)})$  using (36) and (37) with  $(u^{(k+1)}, v^{(k+1)})$  replaced by  $(\tilde{u}^{(k+1)}, \tilde{v}^{(k+1)})$ .  
    update  $\sigma^{(k+1)}$  by (35).  
     $k \leftarrow k + 1$   
  **until** convergence  
  **return**  $(u, \tilde{u}, v, \tilde{v})^\mu$   
   $(u, \tilde{u}, v, \tilde{v}) \leftarrow (u, \tilde{u}, v, \tilde{v})^\mu, \mu \leftarrow 2\mu$ .  
  Compute  $\bar{u}$  and  $\bar{v}$  using (21) and then  $\delta_c$  using (38).  
**end while**

---

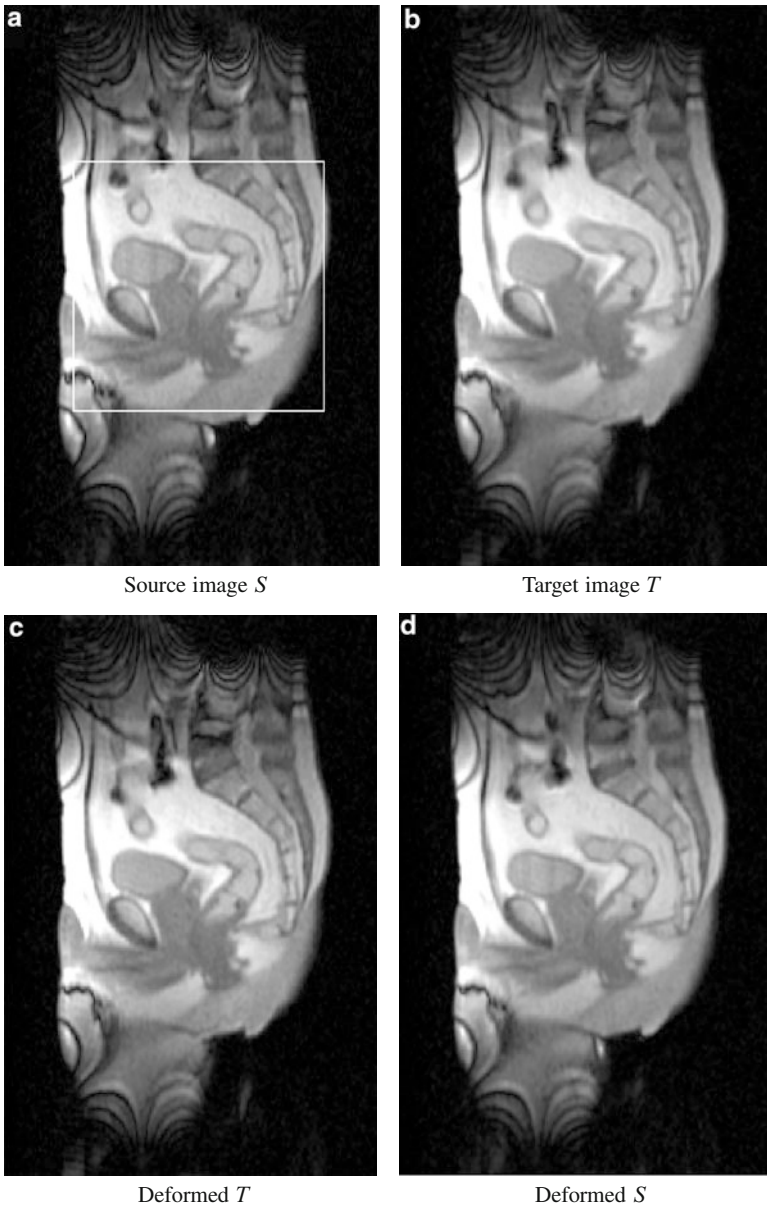
$CC(S(x + \bar{u}), T)$  and  $CC(T(x + \bar{v}), S)$  converges. We set a stopping tolerance  $\epsilon = 0.5$  and terminate the whole computation once  $\delta_c$  is lower than  $\epsilon$ , in which case the maximum ICE is less than half of the grid size between two concatenate pixels/voxels and hence the inverse consistency is exactly satisfied with respect to the original resolution of the images.

## 5 Experimental Results

In this section, we present the experimental results of proposed model using algorithm 1 (**icDIR**). All implementations involved in the experiments were coded in Matlab v7.3 (R2006b), except the Thomas tridiagonal solver, which was coded in C++. We used build-in functions `interp2/interp3` of Matlab with default settings for interpolations. All Computations were performed on a Linux (version 2.6.16) workstation with Intel Core 2 CPU at 1.86GHz and 2GB memory.

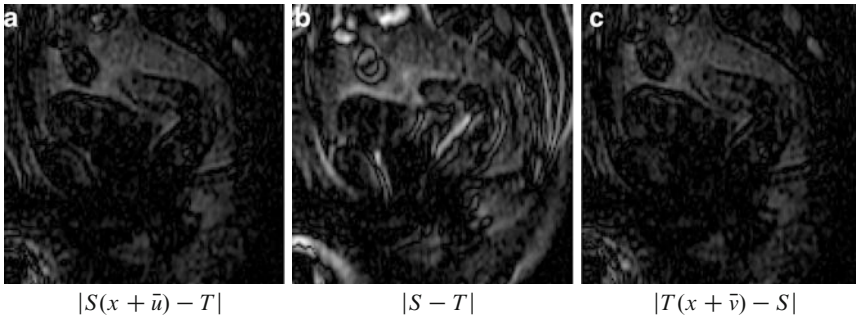
We first test the accuracy of registration and auto recontouring of the proposed algorithm on a clinical data set of 100 2D-prostate MR images. Each image, called a phase, is a 2D image of dimension  $288 \times 192$  that focuses on the prostate area. The first phase is used as a source image  $S$ , as shown in Figure 1(a). The boundaries of the regions of interests (ROI) in  $S$  were delineated by contours and superimposed by medical experts, as enlarged and shown in Figure 4(a). The rest 99 phases were considered as targets. In this experiment we applied the proposed model (18) with parameters  $(\lambda, \mu, \tau)$  set to be (0.05, 0.2, 0.05) to  $S$  and  $T$ s. For demonstration, we only showed the result using the 21st phase as  $T$ , as depicted in Figure 1(b). The deformed  $T$  and deformed  $S$ , i.e.,  $T(x + \bar{v})$  and  $S(x + \bar{u})$ , are shown in the Figure 1(c) and 1(d), respectively, where  $\bar{u}$  and  $\bar{v}$  are defined in (21) using the optimal  $(u, \bar{u}, v, \bar{v})$  obtained by model (18). The errors of the alignments,  $|T(x + \bar{v}) - S|$  and  $|S(x + \bar{u}) - T|$ , on the squared area (shown in Figure 1(a)) are displayed in Figure 2(a) and 2(c), respectively. With comparison to the original error  $|S - T|$  shown in Figure 2(b), we can see the errors of alignments are significantly reduced. This indicates that the proposed registration model (18) has high accuracy in matching two images.

The final optimal forward and backward deformation fields  $\bar{u}$  and  $\bar{v}$  are displayed by applying them to a domain of regular grids, shown in Figure 3(a) and 3(c), respectively. Furthermore, to validate the accurate inverse consistency obtained by our model (18), we applied  $\bar{u} + \bar{v}(x + \bar{u})$  on a domain with regular grids, and plotted the resulting grids in Figure 3(b). The resulting grids by  $\bar{v} + \bar{u}(x + \bar{v})$  had the same pattern so we omitted it here. From Figure 3(b), we can see that the resulting grids are the same as the original regular grids. This indicates that the inverse consistent constraints  $\bar{u} + \bar{v}(x + \bar{u}) = \bar{v} + \bar{u}(x + \bar{v}) = 0$  are well preserved. We also computed the maximum ICE  $\delta_c$  using  $\bar{u}, \bar{v}$  and (38) and the result was 0.46. The mean ICE  $(\|\bar{u} + \bar{v}(x + \bar{u})\| + \|\bar{v} + \bar{u}(x + \bar{v})\|) / 2|\Omega|$  versus the number of iterations is plotted in Figure 5, which shows the inverse consistency is preserved during the registration. These imply that the proposed algorithm provides an accurate inverse consistent registration.

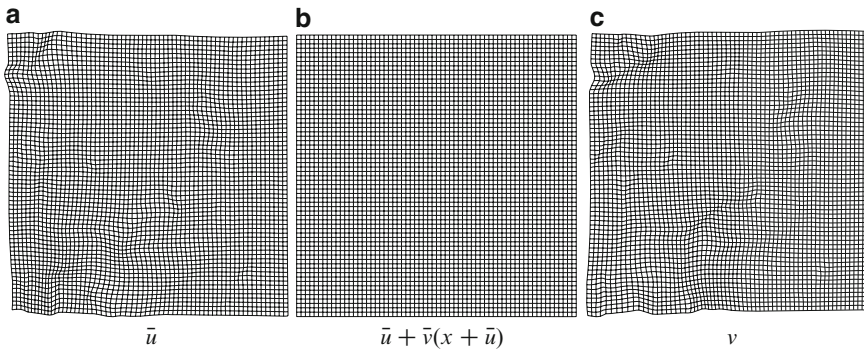


**Figure 1** Inverse consistent registration result by proposed model (18). (a) source image  $S$ . (b) target image  $T$ . (c) deformed  $T$ , i.e.,  $T(x + \bar{v})$ . (d) deformed  $S$ , i.e.,  $S(x + \bar{u})$

An accurate inverse consistent registration can transform segmentations from one image to another accurately. One of the applications is auto-recontouring, that deforms the expert's contours from a planning image to new images during the course of radiation therapy. In this experiment, we had expert's contours superimposed on



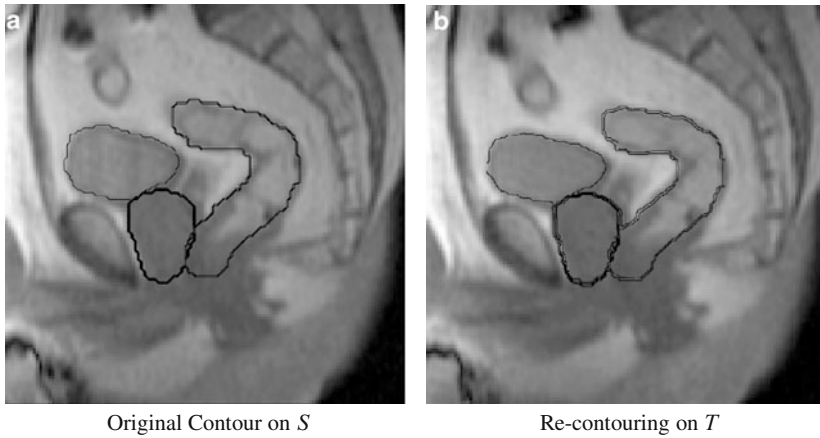
**Figure 2** Residue image (in the square area shown in Figure 1(a)) obtained by proposed model (18). (a)  $|S(x + \bar{u}) - T|$ . (b) initial  $|S - T|$ . (c)  $|T(x + \bar{v}) - S|$



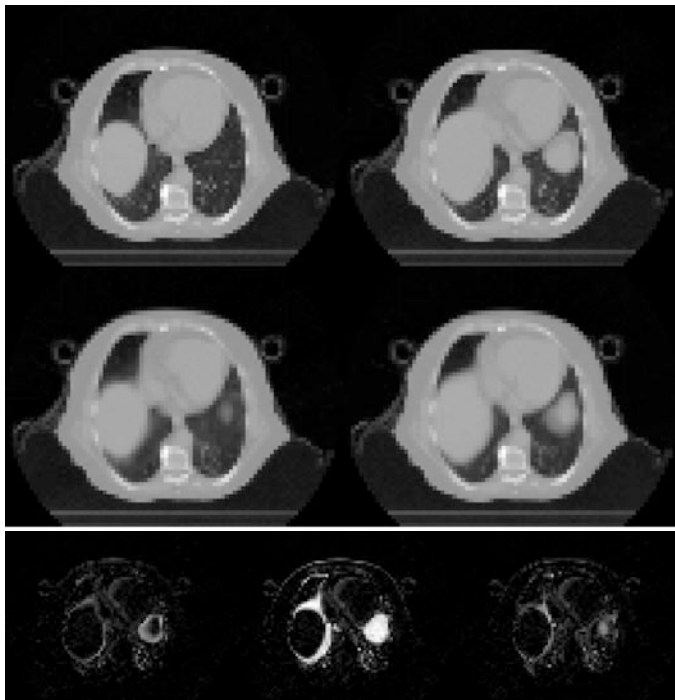
**Figure 3** Deformation fields obtained by proposed model (18) in the zoomed-in area applied on regular grid with half of original resolution of images. (a)  $\bar{u}$ . (b)  $\bar{u} + \bar{v}(x + \bar{u})$ , which demonstrates the inverse consistency is well preserved. (c)  $\bar{v}$

the source image  $S$  as shown in Figure 4(a). Then by applying the deformation field  $\bar{u}$  on this contours we get the deformed contours on the target image  $T$  as shown in Figure 4(b). The accuracy in auto recontouring is evident.

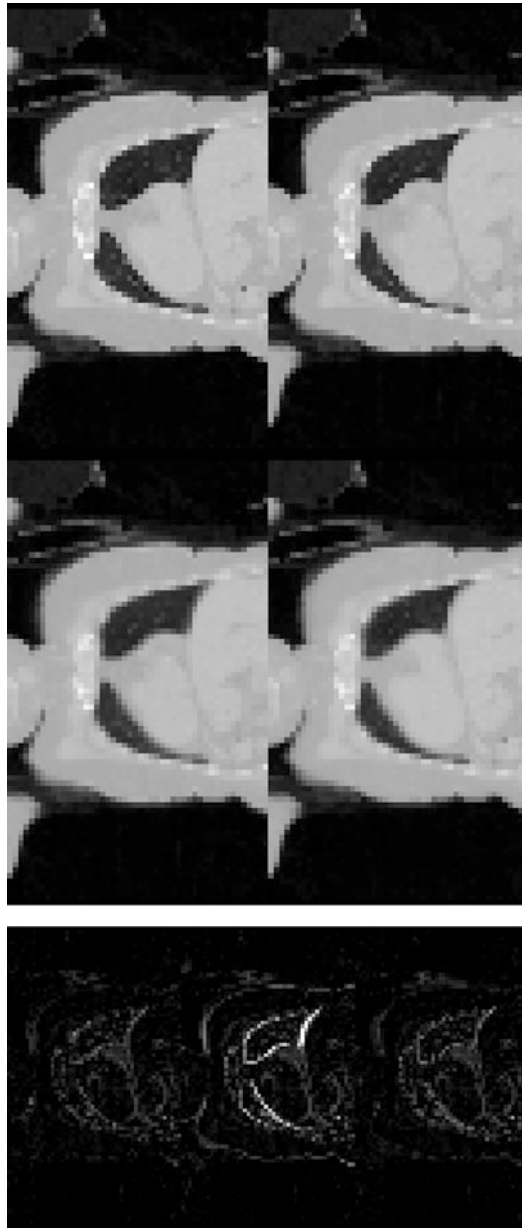
The second experiment was aimed to test the efficiency of the proposed model (18) in registering 3D images. We applied (18) to a pair of 3D chest CT images of dimension  $64 \times 83 \times 48$  taken from the same subject but at different periods. The parameters  $(\lambda, \mu, \tau)$  were set to be  $(.05, .1, .004)$ . The registration was performed in 3D, but for demonstration, we only show the corresponding axial ( $xy$  plane with  $z = 33$ ), sagittal ( $yz$  plane with  $x = 25$ ), and coronal ( $zx$  plane with  $y = 48$ ) slices. The registration results are plotted in Figures 5, 6 and 7, respectively. In each figure, the images in the upper row are  $S$  and  $T$ , respectively, and the images in the middle row are deformed  $T$  and  $S$ , i.e.,  $T(x + \bar{v})$  and  $S(x + \bar{u})$ , respectively. The bottom row shows the residual images  $|S(x + \bar{u}) - T|$ ,  $|S - T|$  and  $|T(x + \bar{v}) - S|$ . The mean of  $CC(S(x + \bar{u}), T)$  and  $CC(T(x + \bar{v}), S)$  reached 0.998 after 50 iterations, and the mean of inverse consistency errors was 0.015. The results show the high accuracy of proposed model (18) and the well preserved inverse consistency.



**Figure 4** Auto recontouring result using the deformation field  $\bar{u}$  obtained by proposed model (18). Images are zoomed-ins of the square area shown in Figure 1(a). (a)  $S$  with initial contours on ROIs (drawn by medical expert). (b)  $T$  with recontouring on ROIs by applying deformation field  $\bar{u}$  to the initial contours

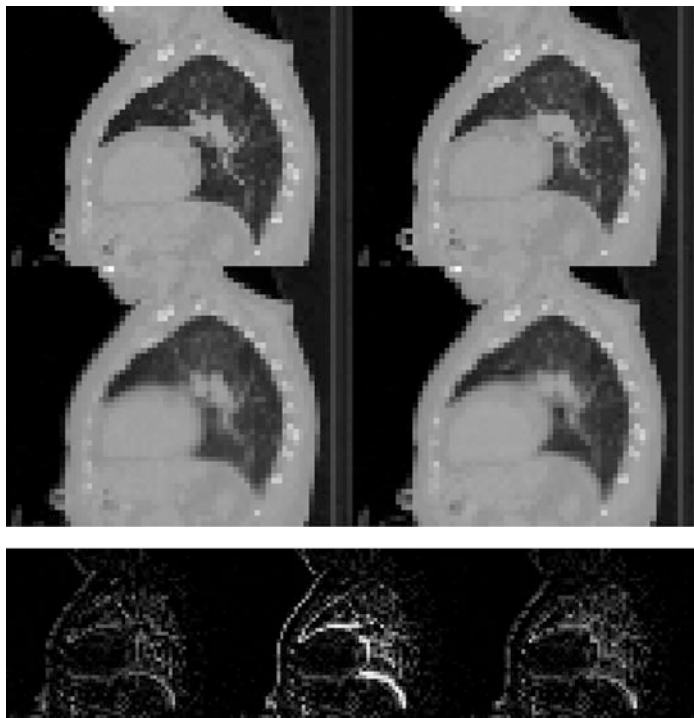


**Figure 5** Registration result of proposed model (18) applied to 3D chest CT image. This figure shows the  $z = 33$  slice at axial direction. Upper left:  $S$ . Upper right:  $T$ , Middle left: deformed  $T$ , i.e.,  $T(x + \bar{v})$ . Middle right: deformed  $S$ , i.e.,  $S(x + \bar{u})$ . Bottom left: residue image  $|S(x + \bar{u}) - T|$ . Bottom middle: initial residue image  $|S - T|$ . Bottom right: residue image  $|T(x + \bar{v}) - S|$



**Figure 6** Registration result of proposed model (18) applied to 3D chest CT image. This figure shows the  $x = 25$  slice at sagittal direction. Upper left:  $S$ . Upper right:  $T$ , Middle left: deformed  $T$ , i.e.,  $T(x + \bar{v})$ . Middle right: deformed  $S$ , i.e.,  $S(x + \bar{u})$ . Bottom left: residue image  $|S(x + \bar{u}) - T|$ . Bottom middle: initial residue image  $|S - T|$ . Bottom right: residue image  $|T(x + \bar{v}) - S|$





**Figure 7** Registration result of proposed model (18) applied to 3D chest CT image. This figure shows the  $y = 48$  slice at coronary direction. Upper left:  $S$ . Upper right:  $T$ , Middle left: deformed  $T$ , i.e.,  $T(x + \bar{v})$ . Middle right: deformed  $S$ , i.e.,  $S(x + \bar{u})$ . Bottom left: residue image  $|S(x + \bar{u}) - T|$ . Bottom middle: initial residue image  $|S - T|$ . Bottom right: residue image  $|T(x + \bar{v}) - S|$

The third experiment was aimed to compare the effectiveness of model (18) with the following conventional full-way inverse consistent deformable registration model:

$$\min_{u, v, \sigma_u, \sigma_v} \|Du\|^2 + \|Dv\|^2 + \lambda J(u, v, \sigma_u, \sigma_v) + \mu (\mathcal{I}_v(u) + \mathcal{I}_u(v)), \quad (39)$$

where  $u$  and  $v$  are forward and backward deformation fields, respectively, and the term  $J$  is defined by:

$$J(u, v, \sigma_u, \sigma_v) = \|S(x + u) - T\|^2 / 2\sigma_u^2 + \|T(x + v) - S\|^2 / 2\sigma_v^2 + |\Omega| \log \sigma_u \sigma_v.$$

The comparison is made on the efficiency and accuracy of matching, as well as the preservation of inverse consistency. The accuracy of matching is measured by correlation coefficients ( $CC$ ) between the target image and deformed source image with the optimal forward and backward deformations obtained by model (39) and proposed model (18), respectively. Recall that for any two images  $S$  and  $T$  both with  $N$  pixels, the  $CC$  of  $S$  and  $T$  is defined by:



$$CC(S, T) = \frac{\sum_{i=1}^N (S_i - \bar{S})(T_i - \bar{T})}{\sqrt{\sum_{i=1}^N (S_i - \bar{S})^2 \sum_{i=1}^N (T_i - \bar{T})^2}},$$

where  $S_i$  and  $T_i$  are the intensities at the  $i$ th pixels of  $S$  and  $T$ , respectively,  $\bar{S}$  and  $\bar{T}$  are the mean intensities of  $S$  and  $T$ , respectively. The maximum value of  $CC$  is 1, in which case  $S$  and  $T$  are (positively) linearly related. In this experiment we applied models (39) and (18) to the images in the first experiment shown in Figure 1 with the same parameters ( $\lambda, \mu, \tau$ ) to be (.05, .2, .05). In Figure 5, we plotted the  $CC$  obtained by model (39) and proposed model (18) at each iteration. One can observe that the  $CC$  obtained by model (18) is higher and increases faster than model (39). This demonstrates that proposed model (18) is more efficient than the conventional full-way model. The reason is that the disparity between deformed  $S$  and deformed  $T$  is smaller than that between deformed  $S$  and fixed  $T$  or deformed  $T$  and fixed  $S$ . When  $S$  and  $T$  are deformed simultaneously, the two directional deformation fields are not necessarily to be large even if the underlying deformation field is large, which usually makes it difficult for the full-way based registration model to reach a satisfactory alignment in short time.

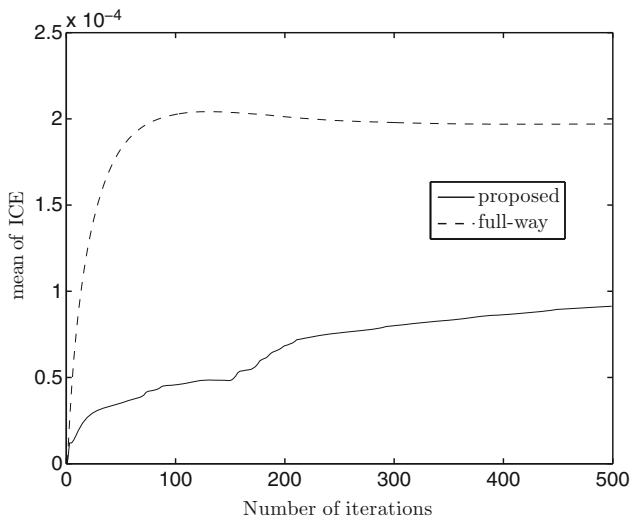
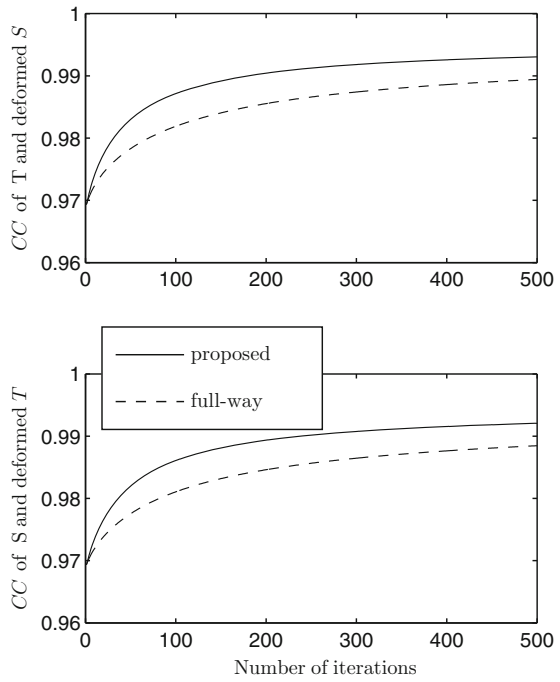
The last experiment is aim to test the robustness of the model to noises and the choice of the parameter  $\lambda$  with the use of MLE based approach (11) for measuring the goodness of matching. The images  $S$  and  $T$  in Figure 1 with additive Gaussian noises (standard deviation is 3% of largest intensity value of  $S$ ) were used in this experiment. The  $CC$  between  $S$  and  $T$  before registration is  $CC(S, T) = 0.901$ . We applied model (18) with  $\sigma$  to be updated/optimized by its EL equation (35), and  $\sigma$  to be set  $\sigma = 1$ , that is the same as using SSD as similarity measure, respectively, to the noise data mentioned earlier. We proceeded the registration with various values of  $\lambda$ , but kept other parameters fixed. Then the numbers of iterations (Iter) for convergence and the final  $CC$  were recorded and shown in Table 1. One can see that while  $\lambda$  decreases, the accuracy of model (18) using fixed  $\sigma$  reduces as the final  $CC$  become much smaller, and it also takes much longer time for the algorithm to converge. On the other hand, with  $\sigma$  being updated (whose computational cost is extremely cheap) model (18) can obtain good matching in much less iterations for a large range of  $\lambda$ . This shows that model with MLE fitting is much less sensitive to noise and the choice of  $\lambda$ , and can achieve fast and accurate results compared with the model using SSD to measure mismatching (Figures 8 and 9).

**Table 1** Number of iterations used for convergence and the final  $CC$  obtained by proposed model with  $\sigma$  updated/fixed

| $\lambda$ | Update $\sigma$ |      | Fix $\sigma$ |      |
|-----------|-----------------|------|--------------|------|
|           | $CC$            | Iter | $CC$         | Iter |
| 1e2       | 0.962           | 48   | 0.955        | 89   |
| 1e1       | 0.962           | 97   | 0.946        | 420  |
| 1e0       | 0.960           | 356  | 0.933        | 1762 |

For a large range of  $\lambda$ , updating  $\sigma$  in each iteration consistently leads to faster convergence and higher accuracy

**Figure 8** *CC* in each iteration obtained by full-way model (39) and proposed model (18). Proposed model (18) gives quick matching with better accuracy, as *CC* by model (18) increase much faster and can reach higher limits than that by full-way model (39)



**Figure 9** Mean of inverse consistent errors (ICE) of the final deformation fields obtained by using full-way model (39) and proposed model (18). The value is much smaller than the size of grid between concatenate pixels, which shows that the inverse consistency is preserved

## References

- [ADPS02] L. Alvarez, R. Deriche, T. Papadopoulos, and J. Sanchez, *Symmetrical dense optical flow estimation with occlusions detection*, Proceedings of European Conference on Computer Vision (2002), 721–735.
- [AGG06] B. B. Avants, M. Grossman, and J. C. Gee, *Symmetric diffeomorphic image registration: Evaluating automated labeling of elderly and neurodegenerative cortex and frontal lobe*, Proceedings of Biomedical Image Registration **4057** (2006), 50–57.
- [BK07] Mirza Faisal Beg and Ali Khan, *Symmetric data attachment terms for large deformation image registration*, IEEE Transactions on Medical Imaging **26** (2007), no. 9, 1179–1189.
- [CJ01] G. E. Christensen and H. J. Johnson, *Consistent image registration*, IEEE Transactions on Medical Imaging **20** (2001), no. 7, 721–735.
- [HC03] J. C. He and G. E. Christensen, *Large deformation inverse consistent elastic image registration*, Proceedings of Information Processing in Medical Imaging **2732** (2003), 438–449.
- [JDJG04] S. Joshiand, B. Davis, M. Jomier, and G. Gerig, *Unbiased diffeomorphic atlas construction for computational anatomy*, Neuroimage, Supplement **23** (2004), no. 1, 151–160.
- [Kes06] M. L. Kessler, *Image registration and data fusion in radiation therapy*, British Journal on Radiology **79** (2006), 99–108.
- [LHG<sup>+</sup>05] A.D. Leow, S.C. Huang, A. Geng, J. Becker, S. Davis, A. Toga, and P. Thompson, *Inverse consistent mapping in 3d deformable image registration: its construction and statistical properties*, Proceedings of Information Processing in Medical Imaging (2005), 493–503.
- [LOC<sup>+</sup>06] W. Lu, G. H. Olivera, Q. Chen, M. Chen, and K. Ruchala, *Automatic re-contouring in 4d radiotherapy*, Physics in Medicine and Biology **51** (2006), 1077–1099.
- [RK06] P. Rogelj and S. Kovacic, *Symmetric image registration*, Medical Image Analysis **10** (2006), no. 3, 484–494.
- [WtHRV98] Joachim Weickert, Bart M. ter Haar Romeny, and Max A. Viergever, *Efficient and reliable schemes for nonlinear diffusion filtering*, IEEE Transactions on Image Processing **7** (1998), no. 3, 398–410.
- [YS05] S.K. Yeung and P.C. Shi, *Stochastic inverse consistency in medical image registration*, International Conference on Medical Image Computing and Computer-Assisted Intervention **8** (2005), no. 2, 188–196.
- [YTS<sup>+</sup>08] Sai-Kit Yeung, Chi-Keung Tang, Pengcheng Shi, Josien P.W. Pluim, Max A. Viergever, Albert C.S. Chung, and Helen C. Shen, *Enforcing stochastic inverse consistency in non-rigid image registration and matching*, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2008), 1–8.
- [ZC08] Q. Zeng and Y. Chen, *Accurate inverse consistent non-rigid image registration and its application on automatic re-contouring*, Proceedings of International Symposium on Bioinformatics Research and Applications (2008), 293–304.
- [ZJT06] Z. Zhang, Y. Jiang, and H. Tsui, *Consistent multi-modal non-rigid registration based on a variational approach*, Pattern Recognition Letters (2006), 715–725.

# A Statistical Model for the Quark Structure of the Nucleon

V. Devanathan and S. Karthiyayini

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** The deep inelastic scattering experiments reveal that the nucleon is a composite object consisting of quarks and gluons. Treating them as Fermi and Bose gases, statistical distribution functions are used to describe their momentum distributions in the rest frame. When transformed to the infinite momentum frame, they yield quark and gluon distribution functions. A thermodynamical bag model is proposed to obtain realistic distribution functions that yield correctly the nucleon structure functions. By including the spin degree of freedom in the Fermi statistical distribution functions, the quark spin distribution functions and the polarized nucleon structure functions are obtained.

**Mathematics Subject Classification (2000)** 62P35, 81U35, 81V05, 81V25, 81V35

**Key words and phrases** Fermi and Bose statistical distribution functions · Deep inelastic scattering · Quarks and gluons · Quark distribution functions · Nucleon structure functions

## 1 Introduction

The deep inelastic scattering of leptons (electrons or muons) on nucleon clearly indicates that the nucleon is a composite object consisting of point particles known as partons. These partons are quarks which interact among themselves by exchange of bosons known as gluons.

---

V. Devanathan  
Tamil Nadu Academy of Sciences, Department of Nuclear Physics, University of Madras,  
Guindy Campus, Chennai 600 025, India  
e-mail: [vdevanathan@hotmail.com](mailto:vdevanathan@hotmail.com)

S. Karthiyayini  
Physics Division, BITS-Pilani, Dubai, UAE  
e-mail: [rajkar6761@gmail.com](mailto:rajkar6761@gmail.com)

The constituent quark model completely explains the static properties of the nucleon such as its mass, spin and magnetic moment. The dynamical properties of the nucleon as observed in Deep Inelastic Scattering (DIS) of leptons are not well understood, especially when there is a large energy transfer. It is the purpose of this article to discuss a statistical model for the nucleon that can explain both the static and dynamical properties of the nucleon in a coherent manner.

The deep inelastic cross section is expressed in terms of the nucleon structure functions. In the parton model of the nucleon, the DIS cross section can be expressed as the incoherent sum of elastic lepton–quark cross sections, and hence it depends on the quark distribution functions. Thus the deep inelastic scattering experiments give valuable information on the quark distribution functions. The statistical model that is proposed for the nucleon yields the quark distribution functions that compare favourably well with the experimental results obtained from deep inelastic scattering experiments.

At present, there is no rigorous theory to deduce these parton distribution functions. Only certain parametrized forms [1–9] are available in literature for the parton distribution functions. As the nucleon is found to consist of quarks and gluons, one can use the statistical distribution functions to find the momentum distribution of these quarks and gluons. Using these statistical distribution functions, several statistical models have been proposed by various authors [10–14] but the one that is discussed here is that of Devanathan et al. [15–19]. This model has an inbuilt mechanism by which sea quarks and gluons are produced copiously in the small  $x$  region. This is a distinguishing feature of this model whereas the other models do not give any physical picture for the production of sea quarks and gluons in the small  $x$  region.

First, let us briefly discuss the DIS of leptons and the kinematic variables that describe such a scattering. If we choose the Lorentz invariant quantities as kinematic variables, then we can study the event in any frame of reference – laboratory frame in which the nucleon is at rest or the infinite momentum frame (IMF) in which the nucleon momentum is extremely large. In IMF, one should expect the momentum of the composite object (nucleon) should be equal to the sum of the momenta of the constituents (quarks and gluons). The Bjorken variable  $x$  which is considered as the inelasticity parameter is now interpreted as the fraction of the momentum carried by quarks in IMF. The DIS experiments indicate that only 45% of the nucleon momentum is carried by the quarks and the missing momentum (55%) is attributed to the gluons which do not participate in the electromagnetic or weak interaction.

The DIS experiments with polarized leptons and polarized target indicate that only 30% of the nucleon spin is accounted by the quark spins. This has come to be known as the proton spin puzzle. Now, it is fairly accepted that the orbital angular momentum of the quarks and gluons account for the balance of the proton spin.

The statistical model that is proposed here is an attempt to answer some of the questions that came up during the study of DIS.

1. The constituent quark model completely describes the static properties of the nucleon but fails miserably to explain the dynamic properties such as DIS nucleon structure functions.

*Is it possible to develop a model which explains both the static and dynamic properties?*

2. The Bjorken variable  $x$  which has been initially introduced to describe a specific kinematic configuration of DIS, acquires two nice physical interpretations:
  - (a) The Bjorken variable  $x$  can be considered as the inelasticity parameter.

$x = 1$  : Elastic scattering

$x < 1$  : Inelastic scattering

$x = 0$  : Extreme limit of inelasticity

(b) In IMF,  $x$  is interpreted as the fraction of the momentum carried by the quark. *Since  $x$  is a Lorentz invariant quantity, is there any interplay between these two physical interpretations of  $x$ ?*

3. At  $x = 1$ , only three valence quarks are present and at small  $x$ , a large number of sea quarks and gluons are seen besides the valence quarks. If we consider the nucleon as a MIT bag consisting of quark–gluon gas, then only valence quarks are seen at temperature  $T = 0$  but a large number of sea quarks and gluons are produced at large  $T$ .

*Since  $x$  and  $T$  show similar features, is it possible to establish a connection between  $x$  and  $T$ ?*

4. The invariant mass  $W$  of the final hadronic state depends upon  $x$  and it is a measure of the energy transfer to the nucleon in DIS.

*Can the invariant mass  $W$  of the final hadronic state be used to obtain the temperature of the MIT bag that is used to represent the nucleon and thus establish a connection between  $x$  and  $T$ ?*

5. The nucleon structure function  $F_2(x)$  vanishes as  $x \rightarrow 1$  and shows a steep rise as  $x \rightarrow 0$ .

*Is it possible to explain this asymptotic behaviour of  $F_2(x)$  by postulating a thermodynamical bag model for the nucleon with temperature  $T \rightarrow 0$  as  $x \rightarrow 1$  and  $T$  steeply increasing as  $x \rightarrow 0$ ?*

As it is the deep inelastic scattering of leptons that reveals much of the quark structure of the nucleon, let us start with a brief review of DIS. Then, following Devanathan et al. [15–19], we shall treat the nucleon as a MIT bag consisting of quarks and gluons. Treating them as Fermi gas and Bose gas and using their relevant statistical distribution functions, equations of state for the nucleon are obtained and solved self-consistently. Transformation of the Fermi and Bose statistical distribution functions to the infinite momentum frame yields the quark and gluon distribution functions. The nucleon structure functions are calculated using the quark distribution functions and compared with the results obtained from DIS experiments. If the spin degree of freedom is included in the Fermi statistical distribution function then one can obtain the quark spin distribution function and obtain the polarized nucleon structure functions.

## 2 Deep Inelastic Scattering of Leptons

Let us briefly discuss the DIS of leptons (electrons or muons) and the kinematic variables that describe such a scattering [20]. In DIS, only the scattered lepton is observed and not the hadronic final state (Figure 1). So, the DIS is characterized by the kinematic variables  $E'$ , the energy of the scattered lepton and  $\theta$ , the angle of scattering. Instead of  $E'$  and  $\theta$ , it is more convenient to describe DIS by the variables  $Q^2 (= -q^2)$  and  $x$ , where  $q^2$  is the square of the four-momentum transfer and  $x$  is the Bjorken variable.

At high energies, the square of the four-momentum transfer is:

$$\begin{aligned} q^2 &= k^2 + k'^2 - 2k \cdot k' = m_l^2 + m_l^2 - 2(E E' - \mathbf{k} \cdot \mathbf{k}') \\ &= -2E E' (1 - \cos \theta) = -4E E' \sin^2(\theta/2), \end{aligned} \quad (1)$$

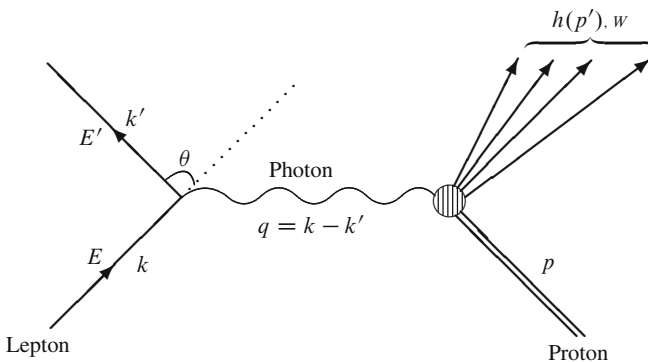
neglecting the mass  $m_l$  of the lepton which is negligible when compared to its energy. Thus,

$$Q^2 = -q^2 = 4E E' \sin^2(\theta/2) \quad (2)$$

$$x = \frac{Q^2}{2p \cdot q} = \frac{Q^2}{2M\nu}, \quad (3)$$

where  $M$  denotes the mass of the nucleon and  $\nu = E' - E$  the energy transfer to the nucleon in the laboratory system. Thus, we choose the Lorentz invariant quantities  $Q^2$  and  $x$  as the kinematic variables. The invariant mass  $W$  of the final hadronic state is given by:

$$\begin{aligned} W^2 &= (p + q)^2 = p^2 + 2p \cdot q + q^2 = M^2 + 2M\nu - Q^2 \\ &= M^2 + Q^2 \left( \frac{1}{x} - 1 \right). \end{aligned} \quad (4)$$



**Figure 1** Deep inelastic scattering of lepton on proton.  $k, k', p, p'$  are four-momenta and  $q$  denotes the four-momentum transfer.  $E$  and  $E'$  denote the energies of the incident and scattered lepton in the laboratory. The final hadronic state  $h(p')$  is not observed and the invariant mass of the final hadronic state is denoted by  $W$

It can be seen from (4), that  $W = M$  if  $x = 1$  and  $W$  increases rapidly with the decrease of  $x$ . Thus, the Bjorken variable  $x$  can be considered as the inelasticity parameter. The Bjorken variable  $x = 1$  denotes the elastic scattering and  $x < 1$  denotes the inelastic scattering. The study of small  $x$  region is of great interest and offers a big challenge to the understanding of the experimental results in this region.

The DIS cross section of electrons or muons can be expressed in terms of two structure functions  $F_1(x, Q^2)$  and  $F_2(x, Q^2)$ . These structure functions are almost found to be independent of  $Q^2$  and this is known as Bjorken scaling. Besides, there exists a relation between these two structure functions in the limit  $Q^2, \nu \rightarrow \infty$

$$F_2(x, Q^2) = 2xF_1(x, Q^2) \tag{5}$$

and this is known as Callon–Gross relation. Using the scaling approximation and the Callon–Gross relation, the DIS cross section can be written as:

$$\frac{d^2\sigma_{IN}}{dx dy} = \frac{ME}{2\pi} \left(\frac{e^2}{q^2}\right)^2 (1 - y + \frac{1}{2}y^2) F_2(x), \tag{6}$$

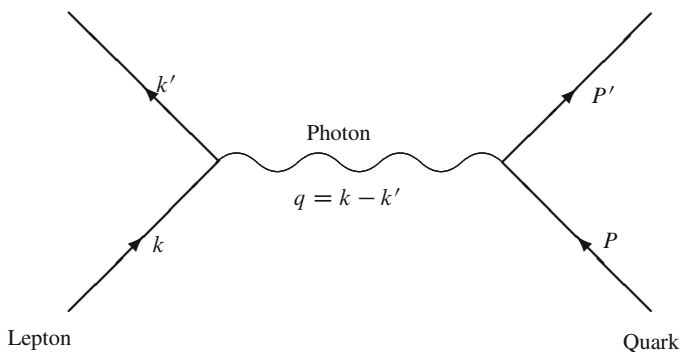
where

$$y = \frac{q \cdot p}{k \cdot p} = \frac{M\nu}{ME} = \frac{\nu}{E}. \tag{7}$$

### The Parton Model

In the parton model, the lepton–nucleon cross section can be expressed as the incoherent sum of elastic lepton–quark cross sections. Figure 2 depicts the elastic scattering of lepton on quark. To gain an insight into the quark structure of the nucleon, let us consider the kinematics of the lepton–quark scattering.

$$P^2 = P'^2 = m_q^2, \tag{8}$$



**Figure 2** Elastic scattering of lepton by quark.  $k, k', P, P'$  are four-momenta and  $q$  denotes the four-momentum transfer.  $P^2 = P'^2 = m_q^2$



where  $m_q$  is the quark mass. The four-momentum transfer is given by:

$$P' - P = k - k' = q, \quad (9)$$

from which it follows

$$\begin{aligned} P'^2 + P^2 - 2P' \cdot P &= q^2, \\ 2P^2 - 2(P + q) \cdot P &= q^2, \\ q^2 + 2q \cdot P &= 0. \end{aligned} \quad (10)$$

From (3) and (10), the Bjorken variable  $x$  is given by:

$$x = -\frac{q^2}{2q \cdot p} = \frac{q \cdot P}{q \cdot p}. \quad (11)$$

If one chooses a frame in which the nucleon has infinite momentum, then the four momenta  $P$  and  $p$  become parallel since any transverse three-momentum of the quark can be neglected. Then from (11), we obtain

$$P_\mu = x p_\mu. \quad (12)$$

The Bjorken variable  $x$  has been initially introduced to define a specific kinematical configuration of DIS. It has been interpreted later as the inelasticity parameter and now, in the infinite momentum frame (IMF), it acquires a new physical meaning as the fraction of the nucleon momentum carried by the quark. The interplay between these two definitions of  $x$  leads to a new interpretation of the nucleon structure function.

### Nucleon Structure Functions

Let us now express the lepton–nucleon cross section as an incoherent sum of elastic lepton–quark cross sections. If  $f(x)$  is the quark distribution function of a given flavour, then

$$d^2\sigma_{lN} = \sum_f f(x) dx d\sigma_{lf}, \quad (13)$$

where  $d\sigma_{lf}$  denotes the lepton–quark cross section.

$$\frac{d^2\sigma_{lf}}{dy} = \frac{ME}{2\pi} \left( \frac{e_f e}{q^2} \right)^2 x (1 - y + \frac{1}{2}y^2), \quad (14)$$

where  $e_f$  denotes the charge of the quark of flavour  $f$  and  $e$  is the unit charge.

From (6), (13) and (14), we deduce an expression for the nucleon structure function in terms of the quark distribution functions  $f(x)$ .

$$F_2(x) = \sum_f \left(\frac{e_f}{e}\right)^2 x f(x). \tag{15}$$

### Quark Distribution Functions

Restricting our considerations to light quarks  $u, d$  and  $s$ , we need to consider only the following quark distribution functions:

$$u(x), \bar{u}(x), d(x), \bar{d}(x), s(x), \bar{s}(x).$$

In the infinite momentum frame (IMF),  $u(x)dx$  is interpreted as the probability of finding the  $u$  quark carrying a fraction of momentum lying between  $x$  and  $x + dx$  of the total nucleon momentum. So, the proton can be described by the following set of equations:

$$\int_0^1 [u(x) - \bar{u}(x)]dx = 2 \tag{16}$$

$$\int_0^1 [d(x) - \bar{d}(x)]dx = 1 \tag{17}$$

$$\int_0^1 [s(x) - \bar{s}(x)]dx = 0 \tag{18}$$

The range of  $x$  integration is from 0 to 1. A similar description can be given for the neutron. Because of the isospin invariance, the distribution functions for neutron are related to those of proton. The following nomenclature is used.

- $u(x)$  =  $x$  distribution of  $u$  quark in a proton
- =  $x$  distribution of  $d$  quark in a neutron
- $d(x)$  =  $x$  distribution of  $d$  quark in a proton
- =  $x$  distribution of  $u$  quark in a neutron
- $s(x)$  =  $x$  distribution of  $s$  quark in a proton or neutron

Similar definitions are used for  $\bar{u}(x), \bar{d}(x)$  and  $\bar{s}(x)$ .

Using (15), we can now write down the nucleon structure functions in terms of the quark distribution functions.

$$\text{proton : } F_2^p(x) = x \left[ \frac{4}{9}(u(x) + \bar{u}(x)) + \frac{1}{9}(d(x) + \bar{d}(x) + s(x) + \bar{s}(x)) \right]. \tag{19}$$

$$\text{neutron : } F_2^n(x) = x \left[ \frac{4}{9}(d(x) + \bar{d}(x)) + \frac{1}{9}(u(x) + \bar{u}(x) + s(x) + \bar{s}(x)) \right]. \tag{20}$$

The Gottfried Sum Rule (GSR) can be written in terms of the quark distribution functions.

$$\begin{aligned}
 \text{GSR} &= \int_0^1 \frac{1}{x} (F_2^p(x) - F_2^n(x)) dx \\
 &= \frac{1}{3} \int_0^1 (u(x) + \bar{u}(x) - d(x) - \bar{d}(x)) dx \\
 &= \frac{1}{3} + \frac{2}{3} \int_0^1 (\bar{u}(x) - \bar{d}(x)) dx. \tag{21}
 \end{aligned}$$

If  $\bar{u}(x) = \bar{d}(x)$ , then  $\text{GSR} = 1/3$ . The new muon collaboration experiment yields a value

$$\text{GSR} = 0.240 \pm 0.016.$$

This implies that the contribution from  $\bar{d}(x)$  is more than the contribution from  $\bar{u}(x)$ . This theory predicts qualitatively this feature and it is evident from the perusal of Table 1 and Figure 5.

### 3 The Statistical Model of the Nucleon

The deep inelastic scattering of leptons on nucleons indicates that the nucleon consists of three valence quarks, sea quarks and gluons, confined within a small volume.

Proton:  $u u d$  + quark–antiquark pairs + gluons  
 (Valence quarks) (Sea quarks)

Neutron:  $u d d$  + quark–antiquark pairs + gluons  
 (Valence quarks) (Sea quarks)

Based on this observation, let us develop a statistical model for the nucleon. It is a MIT bag consisting of quark–gluon gas, for which the Fermi distribution function is used for describing the quarks and the Bose distribution function for describing the gluons. Treating the quarks as particles of zero rest mass, the number density of  $u$ -quarks with momentum lying between  $p$  and  $p + dp$  at temperature  $T$  is given by the Fermi distribution function [15–20].

$$n_u(p) = \frac{g}{(2\pi)^3} \frac{1}{e^{(\epsilon - \mu_u)/T} + 1}, \tag{22}$$

where  $\epsilon$  is the energy and  $\mu_u$  the chemical potential of the  $u$ -quark. The degeneracy factor  $g$  is 6 which is the number of degrees of freedom (3 due to colour and 2 due to spin) available for each flavour of quarks. Similar equations can be written for the  $d$ -quarks and the antiquarks. The chemical potential for the  $d$ -quark  $\mu_d$  is, in general, different from that of  $u$ -quark. The chemical potential for the antiquark is

of opposite sign to the chemical potential of the quark. With this observation, we can write down the distributions functions for  $d$ -quarks and antiquarks.

$$n_d(p) = \frac{6}{(2\pi)^3} \frac{1}{e^{(\epsilon - \mu_d)/T} + 1}, \quad (23)$$

$$n_{\bar{u}}(p) = \frac{6}{(2\pi)^3} \frac{1}{e^{(\epsilon + \mu_u)/T} + 1}, \quad (24)$$

$$n_{\bar{d}}(p) = \frac{6}{(2\pi)^3} \frac{1}{e^{(\epsilon + \mu_d)/T} + 1}, \quad (25)$$

Given the distribution functions, we can obtain the number density (number per unit volume) of each flavour of quarks by integration over the momentum.

$$n_u = \int n_u(p) d^3 p, \quad n_d = \int n_d(p) d^3 p, \quad (26)$$

$$n_{\bar{u}} = \int n_{\bar{u}}(p) d^3 p, \quad n_{\bar{d}} = \int n_{\bar{d}}(p) d^3 p, \quad (27)$$

For the proton, the number of  $u$  valence quarks is 2 and the number of  $d$  valence quarks is 1. If  $V$  is the volume of proton, then

$$(n_u - n_{\bar{u}})V = 2; \quad (n_d - n_{\bar{d}})V = 1. \quad (28)$$

For the gluons, there is no number conservation and hence the chemical potential for the gluon is zero. The number density of the gluons is given by the Bose distribution function.

$$n_g(p) = \frac{16}{(2\pi)^3} \frac{1}{e^{\epsilon/T} - 1}. \quad (29)$$

The degeneracy factor for the gluons is 16, of which 8 is due to the colour degree of freedom and 2 due to the transverse components of spin.

In a similar way, one can find the energy density  $\varepsilon_q$  of each flavour of quarks and antiquarks ( $u$ ,  $d$ ,  $\bar{u}$  and  $\bar{d}$ ) and calculate their contributions to the total energy density. As we have assumed zero rest mass for the quarks, the energy of the quark  $\varepsilon$  is numerically equal to its momentum  $p$  in the natural units ( $\hbar = c = 1$ ).

$$\varepsilon_q = \int \frac{6}{(2\pi)^3} \frac{p}{e^{(p - \mu_q)/T} + 1} d^3 p, \quad \varepsilon_{\bar{q}} = \int \frac{6}{(2\pi)^3} \frac{p}{e^{(p + \mu_q)/T} + 1} d^3 p, \quad q = u, d. \quad (30)$$

For the gluons, the energy density is

$$\varepsilon_g = \int \frac{16}{(2\pi)^3} \frac{p}{e^{p/T} - 1} d^3 p, \quad (31)$$

The energy density due to all the quarks and gluons is the sum.

$$\varepsilon = \varepsilon_u + \varepsilon_d + \varepsilon_{\bar{u}} + \varepsilon_{\bar{d}} + \varepsilon_g. \quad (32)$$

Now, we are in a position to write down the equations of state for the proton.

$$\varepsilon(T)V + BV = W, \quad (33)$$

$$n_u - n_{\bar{u}} = 2/V = \mu_u T^2 + \mu_u^3/\pi^2, \quad (34)$$

$$n_d - n_{\bar{d}} = 1/V = \mu_d T^2 + \mu_d^3/\pi^2, \quad (35)$$

$$P = (1/3)\varepsilon(T) - B = 0. \quad (36)$$

The energy density  $\varepsilon(T)$  is given by<sup>1</sup>

$$\begin{aligned} \varepsilon(T) &= \varepsilon_u + \varepsilon_{\bar{u}} + \varepsilon_d + \varepsilon_{\bar{d}} + \varepsilon_g \\ &= \frac{3}{4\pi^2}(\mu_u^4 + \mu_d^4) + \frac{3}{2}T^2(\mu_u^2 + \mu_d^2) + \frac{37}{30}\pi^2 T^4, \end{aligned} \quad (37)$$

and it is a function of temperature. So, this bag model describes the nucleon not only in the ground state ( $T = 0$ ) but also in the excited states at higher temperature. So, it can be truly described as the thermodynamical bag model of the nucleon. The bag constant is denoted by  $B$  and the volume of the bag by  $V$ . The mass of the nucleon  $M$  in the thermodynamical bag model corresponds to  $T = 0$  and  $W$  denotes the mass of the excited nucleon at some finite temperature  $T$ . Equation (36) arises from the pressure balance condition or the energy minimization condition with respect to the bag volume.

Let us consider the ground state of the nucleon which corresponds to  $T = 0$ . Given the mass of the nucleon ( $M = 938.4$  MeV), we can determine all the other four quantities by solving the four equations (33) – (36), using any numerical method such as the Newton-Raphson method.

$$\mu_u = 335.9 \text{ MeV}, \quad \mu_d = 266.6 \text{ MeV}, \quad B^{1/4} = 145.68 \text{ MeV}, \quad R = 0.985 \text{ fm}. \quad (38)$$

It is remarkable that this naive approach yields correctly the nucleon radius  $R$ . Assuming the value of the bag constant<sup>2</sup>  $B$ , one can determine  $W, \mu_u, \mu_d, V$  at any

<sup>1</sup> For the derivation of (34), (35) and (37), the reader is referred to: V. Devanathan, *Ch. 14, Nuclear Physics*, Narosa Publishing House, New Delhi, India and Alpha Science International, Oxford, UK. (2006)

<sup>2</sup> The bag constant is known to decrease as the temperature increases

$$B = B_0[1 - (T/T_c)^4],$$

where  $B_0$  is a bag constant corresponding to  $T = 0$  and  $T_c$  is the critical temperature determined from the pressure balance equation (36) by imposing the condition that the chemical potential vanishes at the critical temperature:

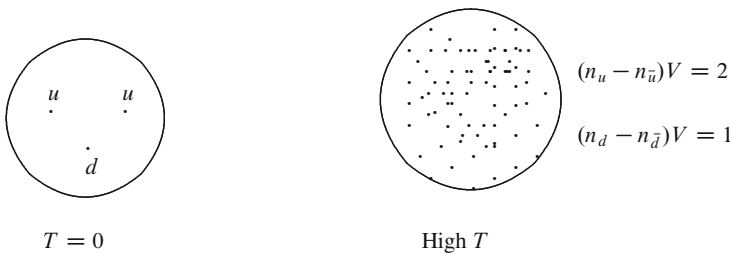
$$T_c = (90B_0/37\pi^2)^{1/4} \approx 102.6 \text{ MeV}.$$

higher temperature by solving the equations of state. As it is possible to extend the study to higher temperatures by this method, this is known as the thermodynamical bag model (TBM).

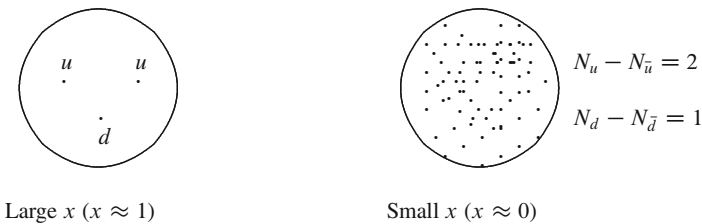
At  $T = 0$ , only three valance quarks are observed and as the temperature is increased, more and more sea quarks (quark–antiquark pairs) are produced. This results in the increase of the mass of the nucleon as well as its volume (*vide* Figure 3).

This picture has a remarkable similarity with the features noticed in DIS. At  $x = 1$ , only three valance quarks are observed and as  $x \rightarrow 0$ , more and more sea quarks are detected in the nucleon (*vide* Figure 4).

The thermodynamical bag model has been used extensively by Devanathan and his collaborators [15–19] to investigate the quark distribution functions and the nucleon structure functions observed in deep inelastic scattering of leptons on nucleons. They have also used the thermodynamical bag model to study the nucleon–nucleon potential in terms of the quarks [21] and the static properties of hadrons [22]. It is observed that the thermodynamical bag model offers a clear insight into the transition of static properties of the nucleon into its dynamical properties, as observed in deep inelastic scattering of leptons.



**Figure 3** Proton: At  $T = 0$ , proton consists of only three valance quarks but at high temperature, a large number of sea quarks are observed along with valance quarks



**Figure 4** Quark structure of proton: At large  $x$ , only three valance quarks are observed but at small  $x$ , a large number of sea quarks are observed along with valance quarks

### Statistical Distribution Functions in IMF

The statistical distribution functions (22)–(25) and (29) when transformed to the infinite momentum frame yields the quark distribution functions. This has been done by Cleymans and Thews, Mac and Ugaz [10, 11] and Devanathan et al. [15–19]. We give here only the final results.

$$u(x) = 2A \ln[1 + \exp\{(\mu_u - \frac{1}{2}xM)/T\}], \quad (39)$$

$$d(x) = 2A \ln[1 + \exp\{(\mu_d - \frac{1}{2}xM)/T\}], \quad (40)$$

$$\bar{u}(x) = 2A \ln[1 + \exp\{(-\mu_u - \frac{1}{2}xM)/T\}], \quad (41)$$

$$\bar{d}(x) = 2A \ln[1 + \exp\{(-\mu_d - \frac{1}{2}xM)/T\}], \quad (42)$$

$$g(x) = -\frac{16}{3}A \ln[1 - \exp\{-\frac{1}{2}(xM/T)\}], \quad (43)$$

with

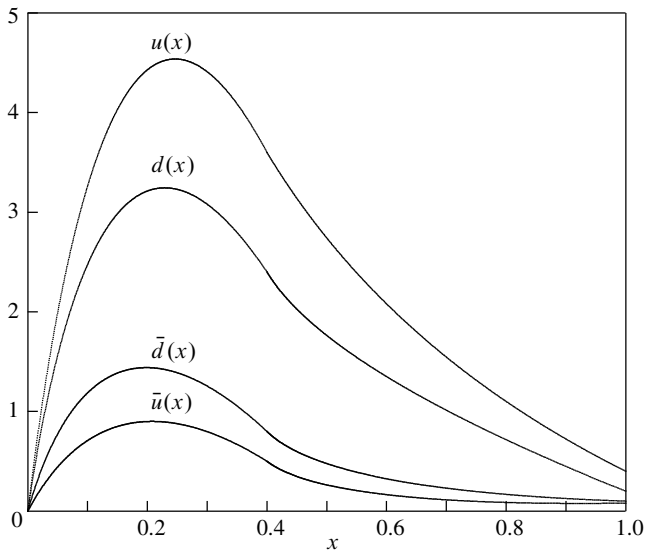
$$A = \frac{3M^2 V x T}{4\pi^2}. \quad (44)$$

The distribution functions given by (39)–(43) depend on the parameters  $T$ ,  $V$ ,  $\mu_u$  and  $\mu_d$ . These are not free parameters. The invariant mass  $W$  of the final hadronic state is determined for a given value of  $x$  using (4) and taking it as the mass of the excited nucleon, all the parameters are obtained by solving the equations of state (33)–(36) of the nucleon as given in Table 1.

The distribution functions for the  $u$  and  $d$  quarks are depicted in Figure 5 for the proton using a set of parameters, so obtained. It is seen that  $u$  quarks dominate over the  $d$  quarks whereas for the antiquarks, the opposite feature is observed. The dominance of the  $\bar{d}$  quarks over the  $\bar{u}$  quarks – a feature so essential for the explanation of the Gottfried Sum Rule (GSR) – comes out naturally in this formalism.

**Table 1** Table showing the dependence of temperature  $T$ , bag radius  $R$  and chemical potentials  $\mu_u$  and  $\mu_d$  on the Bjorken variable  $x$  along with the quark distribution functions for the proton in DIS ( $Q^2 = 4 \text{ GeV}^2$ )

| $x$  | $W$ (MeV) | $T$ (MeV) | $R$ (fm) | $\mu_u$ (MeV) | $\mu_d$ (MeV) | $xu(x)$ | $xd(x)$ | $x\bar{u}(x)$ | $x\bar{d}(x)$ |
|------|-----------|-----------|----------|---------------|---------------|---------|---------|---------------|---------------|
| 0.15 | 4854      | 85.7      | 2.1275   | 50.1          | 25.7          | 0.442   | 0.306   | 0.085         | 0.130         |
| 0.20 | 4110      | 85.5      | 2.0057   | 59.3          | 30.7          | 0.530   | 0.337   | 0.070         | 0.117         |
| 0.30 | 3197      | 84.9      | 1.8297   | 76.7          | 40.6          | 0.559   | 0.301   | 0.035         | 0.069         |
| 0.40 | 2623      | 84.2      | 1.6958   | 94.2          | 51.2          | 0.468   | 0.216   | 0.014         | 0.032         |
| 0.50 | 2209      | 83.1      | 1.5805   | 113.2         | 63.5          | 0.345   | 0.136   | 0.005         | 0.012         |
| 0.60 | 1883      | 81.6      | 1.4722   | 135.2         | 78.9          | 0.235   | 0.080   | 0.001         | 0.004         |
| 0.70 | 1611      | 79.0      | 1.3622   | 162.6         | 100.0         | 0.153   | 0.043   | 0.000         | 0.001         |
| 0.80 | 1371      | 74.0      | 1.2414   | 200.6         | 132.2         | 0.093   | 0.021   | 0.000         | 0.000         |
| 0.90 | 1151      | 61.1      | 1.1025   | 259.4         | 187.6         | 0.042   | 0.006   | 0.000         | 0.000         |
| 0.95 | 1045      | 45.9      | 1.0347   | 298.1         | 226.6         | 0.012   | 0.001   | 0.000         | 0.000         |
| 1.00 | 938       | 0.0       | 0.9849   | 335.9         | 266.6         | 0.000   | 0.000   | 0.000         | 0.000         |



**Figure 5** The quark distribution functions  $u(x), d(x), \bar{u}(x), \bar{d}(x)$  as a function of  $x$ , obtained using the parameters given in Set 1

This feature is observed because the chemical potential  $\mu_u$  is greater than  $\mu_d$  as a consequence of two valence  $u$  quarks and one valence  $d$  quark for the proton.

Since the quark distribution functions have been obtained by transforming the Fermi distribution functions from the rest frame to the infinite momentum frame, it is expected that their integrals should yield the total number of quarks of a particular flavour. For instance

$$N_u = \int n_u(p)d^3 p = \int_0^1 u(x)dx, \tag{45}$$

where  $N_u$  denotes the total number of  $u$  quarks. Similar expressions can be written for  $N_d, N_{\bar{u}}$  and  $N_{\bar{d}}$ . These relations have been verified by numerical integration for a given set of parameters  $T, V, \mu_u$  and  $\mu_d$ . Calculations done with the following two sets of parameters are presented in Table 2.

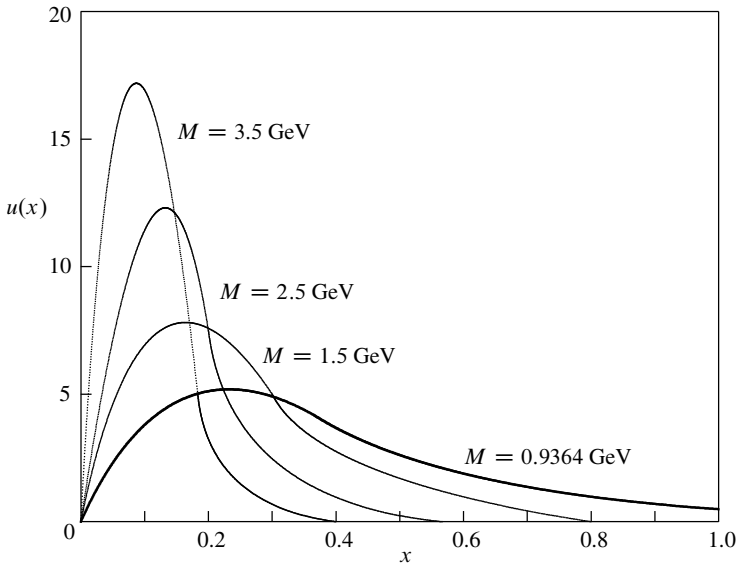
- Set 1:  $T = 84.9 \text{ MeV}, V = 25.69 \text{ fm}^3, \mu_u = 76.7 \text{ MeV}, \mu_d = 40.6 \text{ MeV}.$
- Set 2:  $T = 85.9 \text{ MeV}, V = 51.02 \text{ fm}^3, \mu_u = 39.9 \text{ MeV}, \mu_d = 20.3 \text{ MeV}.$

In Table 2, it is shown that the number of quarks of a particular flavour obtained by performing the integration in the rest frame or in the IMF is essentially the same for the proton yielding the value 2 for the  $u$  valence quarks ( $N_u^V$ ) and 1 for the  $d$  valence quark ( $N_d^V$ ). In IMF, the values are slightly less since the distribution functions extend to a small extent beyond the physical region  $x = 1$  due to the neglect of second-order terms in the transformation of the Fermi statistical distribution function from the rest frame to the IMF. If the upper limit of the integration is increased slightly beyond  $x = 1$ , we obtain once again the correct number of valence quarks.



**Table 2** Number of quarks (expectation values) of a particular flavour obtained by performing the integration in the rest frame and in the IMF for two different sets of parameters

| Number of quarks            | Set 1      |       | Set 2      |       |
|-----------------------------|------------|-------|------------|-------|
|                             | Rest frame | IMF   | Rest frame | IMF   |
| $N_u$                       | 2.482      | 2.398 | 3.502      | 3.384 |
| $N_{\bar{u}}$               | 0.482      | 0.468 | 1.502      | 1.455 |
| $N_u^V = N_u - N_{\bar{u}}$ | 2.000      | 1.930 | 2.000      | 1.929 |
| $N_d$                       | 1.721      | 1.666 | 2.859      | 2.765 |
| $N_{\bar{d}}$               | 0.721      | 0.700 | 1.859      | 1.800 |
| $N_d^V = N_d - N_{\bar{d}}$ | 1.000      | 0.966 | 1.000      | 0.965 |



**Figure 6** The quark distribution function  $u(x)$  as a function of  $x$  for different values of nucleon mass  $M$

This ensures the correctness of the quark distribution function obtained by transformation of the Fermi statistical distribution function from the rest frame to the infinite momentum frame.

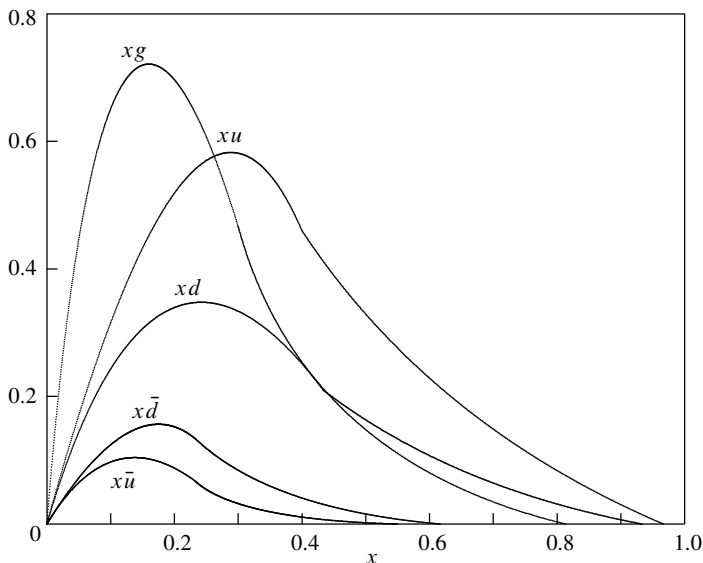
The quark distribution functions (39)–(42) in IMF involve the mass  $M$  whereas the Fermi statistical distribution functions (22)–(25) do not depend explicitly on  $M$ . The effect of changing  $M$  in the quark distribution function is investigated and presented in Figure 6. It is found that as the value of  $M$  is increased, the curve shifts towards the lower values of  $x$  but the areas enclosed by the curve remains a constant and is equal to the number of quarks of that flavour. This is a very important observation and this has suggested to us the thermodynamical bag model to obtain the parton distribution functions of the correct asymptotic behaviour.

### 4 The Thermodynamical Bag Model

The parton distribution functions (39)–(43) involve the parameters  $T, V, \mu_u$  and  $\mu_d$  which are obtained by solving the equations of state of the nucleon bag for a given mass  $W$  of the excited nucleon which depends on the Bjorken variable  $x$ . The quark distribution functions calculated with a fixed set of parameters do not exhibit the correct asymptotic behaviour and a new approach has been developed by Devanathan et al. [15–19] by treating the parameters as functions of the Bjorken variable  $x$  and normalizing the distribution function so obtained by using one or two free parameters. A single parameter is found to be sufficient to obtain quark distribution functions which compare favourably well with the experimental results for values of  $x > 0.15$ . The resulting parton distribution functions are shown in Figure 7. They exhibit the correct asymptotic behaviour. They vanish as  $x \rightarrow 1$  because  $T \rightarrow 0$ . For the study of smaller  $x$  region, the mass of the nucleon  $M$  in (39)–(43) is to be replaced by  $W$  and additional parameters have to be used to normalize the quark distribution functions so as to yield  $N_u - N_{\bar{u}} = 2$  and  $N_d - N_{\bar{d}} = 1$ .

Such a procedure is justified because the differential cross section for the DIS can be written as the scalar product of leptonic tensor  $L^{\mu\nu}$  and hadronic tensor  $W_{\mu\nu}$ .

$$\frac{d^2\sigma(E, E', \theta)}{d\Omega dE'} \sim L^{\mu\nu} W_{\mu\nu}. \tag{46}$$

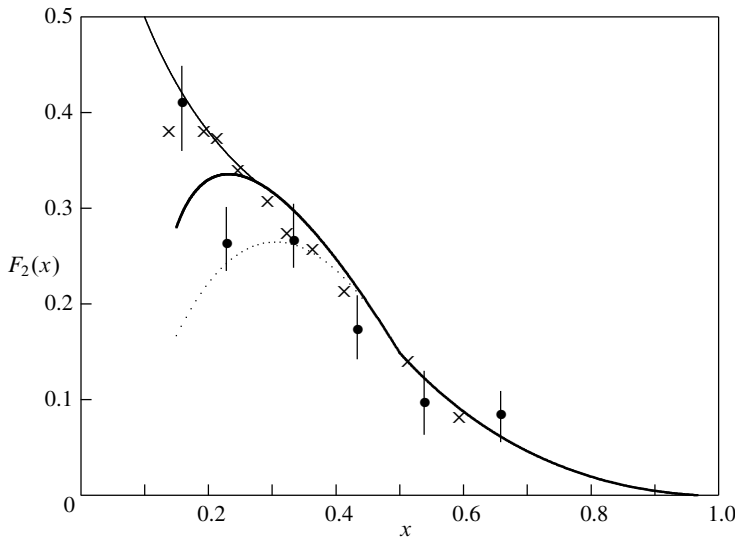


**Figure 7** Weighted quark and gluon distribution functions  $xu(x), xd(x), x\bar{u}(x), x\bar{d}(x)$  and  $xg(x)$  obtained using the thermodynamical bag model for  $Q^2 = 4 \text{ GeV}^2$

The hadronic tensor arises from the hadronic current–current interaction.

$$W_{\mu\nu} \sim \sum_h \langle p | J_\mu^\dagger | h \rangle \langle | J_\nu | p \rangle, \quad (47)$$

where the summation is over all the possible final hadronic states  $h$  of invariant mass  $W$ . The hadronic tensor, in turn, is expressed in terms of the nucleon structure functions, which, in turn, is expressed in terms of the quark distribution functions. If one attempts to deduce the quark distribution function from the Fermi distribution function, in some way the excitation of the target nucleon to the final hadronic state and its subsequent deexcitation should be incorporated in the phenomenological model. This is what is done in the thermodynamical bag model by identifying  $W$  with the mass of the excited nucleon and normalizing the distribution functions so obtained to yield the experimentally observed value 0.45 for the fraction of the momentum carried by the quarks by using a single parameter  $\eta(Q^2)$ . The distribution functions so obtained are shown in Figure 7. The proton structure function  $F_2(x)$  calculated using the quark distribution functions is given in Figure 8. The dotted line curve gives the contribution of valence quarks alone to  $F_2(x)$ . The thin line curve indicates a steep rise in the small  $x$  region due to increase in the proton mass as envisaged in TBM.



**Figure 8** Proton structure function  $F_2(x)$  as a function of  $x$  along with the experimental data [23–25]. The dotted curve depicts the contribution of valence quarks alone. The thin line curve indicates the steep rise in the small  $x$  region due to increase in proton mass  $M$  as envisaged in TBM. ( $Q^2 = 4 \text{ GeV}^2$ )

## 5 The Nucleon Spin

The deep inelastic scattering experiments with polarized leptons on polarized proton yield valuable information on the polarized nucleon structure functions and the quark spin distribution functions. The experimental data implied that only a small fraction of the proton spin is carried by the quarks. This startling result has come to be known as *the proton spin puzzle* and referred to as *the spin crisis*.

In the parton model, the nucleon spin structure function can be expressed in terms of the quark spin distribution functions.

$$g_1(x) = \frac{1}{2} \sum_i e_i^2 [(q_i^\uparrow(x) + \bar{q}_i^\uparrow(x)) - (q_i^\downarrow(x) + \bar{q}_i^\downarrow(x))], \quad (48)$$

where  $e_i$  is the charge of the quark (antiquark) of flavour  $i$  and  $q_i^\uparrow(x)(\bar{q}_i^\uparrow(x))$  is the quark (antiquark) distribution function of momentum fraction  $x$ , having the helicity parallel to that of the nucleon and  $q_i^\downarrow(x)(\bar{q}_i^\downarrow(x))$  is that with helicity antiparallel to that of the nucleon. Restricting our considerations to  $u$  and  $d$  quarks only, let us write down explicitly the proton and neutron spin structure functions

$$g_1^p(x) = \frac{1}{2} \left[ \frac{4}{9} \{ (u^\uparrow(x) + \bar{u}^\uparrow(x)) - (u^\downarrow(x) + \bar{u}^\downarrow(x)) \} + \frac{1}{9} \{ (d^\uparrow(x) + \bar{d}^\uparrow(x)) - (d^\downarrow(x) + \bar{d}^\downarrow(x)) \} \right], \quad (49)$$

$$g_1^n(x) = \frac{1}{2} \left[ \frac{1}{9} \{ (u^\uparrow(x) + \bar{u}^\uparrow(x)) - (u^\downarrow(x) + \bar{u}^\downarrow(x)) \} + \frac{4}{9} \{ (d^\uparrow(x) + \bar{d}^\uparrow(x)) - (d^\downarrow(x) + \bar{d}^\downarrow(x)) \} \right], \quad (50)$$

with the usual notation

$$d^n(x) = u^p(x) = u(x); \quad u^n(x) = d^p(x) = d(x). \quad (51)$$

The integrals of the proton and neutron spin functions  $\Gamma_1^p$  and  $\Gamma_1^n$  have special significance since their difference is the Bjorken Sum Rule (BSR), obtained from general considerations and hence considered sacrosanct.

$$\Gamma_1^p = \int_0^1 g_1^p(x) dx; \quad \Gamma_1^n = \int_0^1 g_1^n(x) dx. \quad (52)$$

$$\text{BSR} = \Gamma_1^p - \Gamma_1^n = \frac{1}{6} \frac{g_A}{g_V} \left( 1 - \frac{\alpha(Q^2)}{\pi} \right), \quad (53)$$

where  $g_A$  and  $g_V$  denote the weak interaction axial vector and vector coupling constants and the multiplicative factor  $\left( 1 - \frac{\alpha(Q^2)}{\pi} \right)$  is the perturbative QCD correction factor.

Following exactly the procedure used for obtaining the unpolarized nucleon structure functions, we can now obtain the polarized nucleon structure functions. Including the spin degree of freedom in the Fermi distribution functions, the number densities of  $u$  and  $\bar{u}$  quarks with spin up and spin down can be written as

$$n_u^\uparrow = \frac{g}{8\pi^2} \int \frac{d^3 p}{\exp[(p - \mu_u - \frac{1}{2}\gamma_u)/T] + 1}. \quad (54)$$

$$n_u^\downarrow = \frac{g}{8\pi^2} \int \frac{d^3 p}{\exp[(p - \mu_u + \frac{1}{2}\gamma_u)/T] + 1}. \quad (55)$$

$$n_{\bar{u}}^\uparrow = \frac{g}{8\pi^2} \int \frac{d^3 p}{\exp[(p + \mu_u - \frac{1}{2}\gamma_u)/T] + 1}. \quad (56)$$

$$n_{\bar{u}}^\downarrow = \frac{g}{8\pi^2} \int \frac{d^3 p}{\exp[(p + \mu_u + \frac{1}{2}\gamma_u)/T] + 1}. \quad (57)$$

The multiplicative factor  $g$  denotes the colour degeneracy ( $g = 3$ ) and the additional factor  $\gamma_u$  is the spin parameter. A similar set of equations can be written for  $d$  quarks for which the chemical potential is  $\mu_d$  and the spin parameter is  $\gamma_d$ . The energy densities can also be written in a similar way assuming the quarks to be of zero rest mass.

The Lagrangian multipliers  $T, \mu_u, \mu_d, \gamma_u$  and  $\gamma_d$  are determined from the constraints on energy, particle number and the spin of the system. For the proton, the following equations of state determine the Lagrangian multipliers.

$$\varepsilon(T)V + BV = W, \quad (58)$$

$$V[(n_u^\uparrow + n_u^\downarrow) - (n_{\bar{u}}^\uparrow + n_{\bar{u}}^\downarrow)] = 2, \quad (59)$$

$$V[(n_d^\uparrow + n_d^\downarrow) - (n_{\bar{d}}^\uparrow + n_{\bar{d}}^\downarrow)] = 1, \quad (60)$$

$$V[(n_u^\uparrow + n_{\bar{u}}^\downarrow) - (n_u^\downarrow + n_{\bar{u}}^\uparrow)] = a, \quad (61)$$

$$V[(n_d^\uparrow + n_{\bar{d}}^\uparrow) - (n_d^\downarrow + n_{\bar{d}}^\downarrow)] = b, \quad (62)$$

$$P = (1/3)\varepsilon(T) - B = 0. \quad (63)$$

Using the Fermi distribution functions, (59)–(62) can be written in terms of  $\mu_u, \mu_d, \gamma_u, \gamma_d$ .

$$\frac{V}{2\pi^2} [2\pi^2\mu_u T^2 + 2\mu_u^3 + \frac{3}{2}\mu_u\gamma_u^2] = 2 \quad (64)$$

$$\frac{V}{2\pi^2} [2\pi^2\mu_d T^2 + 2\mu_d^3 + \frac{3}{2}\mu_d\gamma_d^2] = 1 \quad (65)$$

$$\frac{V}{2\pi^2} [\pi^2\gamma_u T^2 + 3\gamma_u\mu_u^2 + \gamma_u^3/4] = a \quad (66)$$

$$\frac{V}{2\pi^2} [\pi^2\gamma_d T^2 + 3\gamma_d\mu_d^2 + \gamma_d^3/4] = b \quad (67)$$

The total energy density  $\varepsilon(T)$  is given by:

$$\begin{aligned} \varepsilon(T) &= (\varepsilon_u^\uparrow + \varepsilon_u^\downarrow + \varepsilon_{\bar{u}}^\uparrow + \varepsilon_{\bar{u}}^\downarrow) + (\varepsilon_d^\uparrow + \varepsilon_d^\downarrow + \varepsilon_{\bar{d}}^\uparrow + \varepsilon_{\bar{d}}^\downarrow) + \varepsilon_g. \\ &= \frac{37}{30}\pi^2 T^4 + \frac{3}{2}T^2(\mu_u^2 + \mu_d^2 + \frac{1}{4}\gamma_u^2 + \frac{1}{4}\gamma_d^2) \\ &\quad + \frac{3}{4\pi^2}(\mu_u^4 + \mu_d^4 + \frac{1}{16}\gamma_u^4 + \frac{1}{16}\gamma_d^4 + \frac{3}{2}\mu_u^2\gamma_u^2 + \frac{3}{2}\mu_d^2\gamma_d^2) \end{aligned} \quad (68)$$

Given  $W$  and  $B$ , the other six quantities  $T, V, \mu_u, \mu_d, \gamma_u, \gamma_d$  can be determined uniquely by solving the earlier equations. The quantities  $a$  and  $b$  denote the separate spin contributions from the  $u$  and  $d$  quarks. If  $a + b = 1$ , then the entire nucleon spin will be accounted by the quarks. Close [26] argues that for the proton, the spin contribution from  $u$  quarks is  $\frac{4}{3}$  ( $a = \frac{4}{3}$ ) and from  $d$  quarks is  $-\frac{1}{3}$  ( $b = -\frac{1}{3}$ ), such that  $a + b = 1$ . The nucleon spin structure functions  $g_1^p$  and  $g_1^n$ , obtained from DIS experiments with polarized leptons on polarized targets, agree remarkably well with this theoretical study for the values of the spin parameters  $a = 1.1$  and  $b = -0.7$ . That means that only 40% of the nucleon spin is accounted for by the quark spins, since  $a + b = 0.4$ .

By boosting the Fermi distribution functions (54)–(57) to the IMF, we obtain the quark spin distribution functions  $u^\uparrow(x), u^\downarrow(x), \bar{u}^\uparrow(x), \bar{u}^\downarrow(x)$ .

$$u^\uparrow(x) = A \ln[1 + \exp[(\mu_u + \frac{1}{2}\gamma_u - \frac{1}{2}xM)/T]], \quad (69)$$

$$u^\downarrow(x) = A \ln[1 + \exp[(\mu_u - \frac{1}{2}\gamma_u - \frac{1}{2}xM)/T]], \quad (70)$$

$$\bar{u}^\uparrow(x) = A \ln[1 + \exp[(-\mu_u + \frac{1}{2}\gamma_u - \frac{1}{2}xM)/T]], \quad (71)$$

$$\bar{u}^\downarrow(x) = A \ln[1 + \exp[(-\mu_u - \frac{1}{2}\gamma_u - \frac{1}{2}xM)/T]], \quad (72)$$

with

$$A = \frac{3M^2 V x T}{4\pi^2}.$$

Similar expressions are obtained for the quark spin distribution functions  $d^\uparrow(x), d^\downarrow(x), \bar{d}^\uparrow(x), \bar{d}^\downarrow(x)$ , by replacing the chemical potential by  $\mu_d$  and the spin parameter by  $\gamma_d$ .

### A Consistency Check

Since the quark spin distribution functions (69)–(72) are obtained by transforming the Fermi distribution functions (54)–(57) to the IMF, it is expected that their integrals should yield the total number of quarks of a particular flavour and helicity.

$$N_u^\uparrow = \int n_u^\uparrow(p) d^3 p = \int_0^1 u^\uparrow(x) dx. \quad (73)$$

This has been verified by numerical integration. It is also observed that the Fermi distribution function  $n_u(p)$  in the rest frame does not involve the nucleon mass  $M$  but the quark distribution function  $u^\uparrow(x)$  in IMF depends on the mass  $M$ . The quark distribution function  $u^\uparrow(x)$  shifts towards the smaller values of  $x$  for higher values of  $M$  as observed in Figure 6. But the integral

$$\int_0^1 u^\uparrow(x) dx$$

is independent of  $M$ . This is an important observation since the nucleon mass  $M$  has to be replaced by  $W$  for small values of  $x$

Further, it is observed that the exponential function in the quark spin distribution functions (69)–(72) is much larger than 1 and hence the quark distribution functions obey a simple additive law.

$$u(x) = u^\uparrow(x) + u^\downarrow(x). \quad (74)$$

$$d(x) = d^\uparrow(x) + d^\downarrow(x). \quad (75)$$

$$\bar{u}(x) = \bar{u}^\uparrow(x) + \bar{u}^\downarrow(x). \quad (76)$$

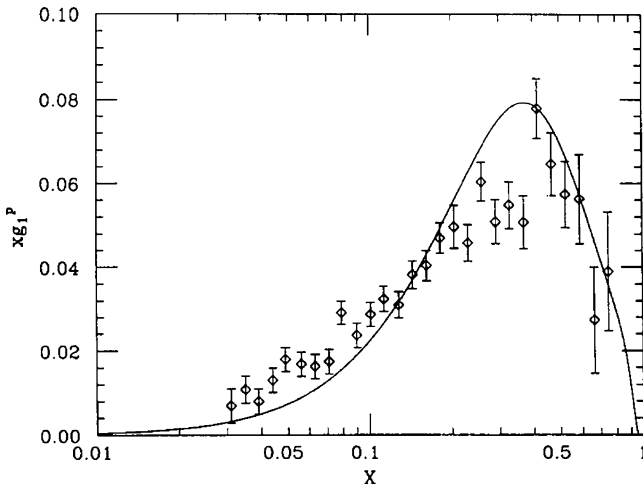
$$\bar{d}(x) = \bar{d}^\uparrow(x) + \bar{d}^\downarrow(x). \quad (77)$$

### Nucleon Spin Structure Functions

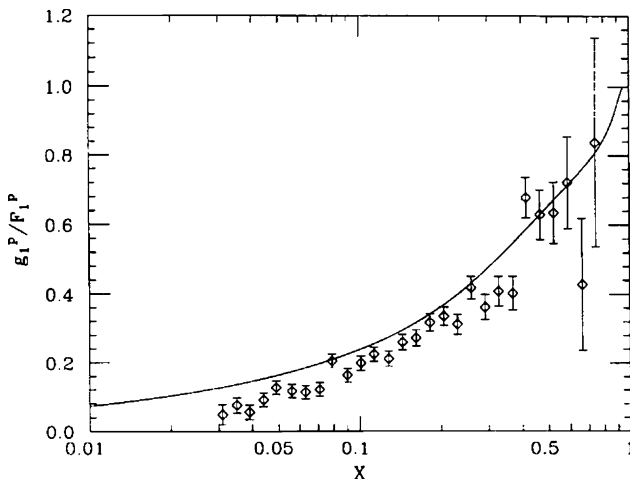
The nucleon spin structure functions  $g_1^p(x)$  and  $g_1^n(x)$  are obtained from (49) and (50). The parameters  $T, V, \mu_u, \mu_d, \gamma_u, \gamma_d$  are obtained by solving the equations of state (58)–(63). Since the quark distribution functions calculated with a fixed set of parameters do not satisfy the correct asymptotic behaviour, a new approach known as the thermodynamical bag model has been developed by Devanathan et al. [15–19]. In this model, the parameters are treated as functions of the Bjorken variable  $x$  and the distribution functions so obtained have to be renormalized to yield the number of valence quarks. This renormalization procedure is to include the structure of the hadronic current in the quark distribution functions. A remarkable agreement is obtained with the experimental data [27, 28] by Devanathan and McCarthy (DM) [18, 19] and they are presented in Table 3. The nucleon spin structure functions cal-

**Table 3** Comparison of the theoretical calculations ( $Q^2 = 3 \text{ GeV}^2$ ) of Devanathan and McCarthy with the experimental data on spin observables

|              | DM (Theoretical calculation) | Experimental data            |
|--------------|------------------------------|------------------------------|
| $\Gamma_1^p$ | 0.134                        | $0.129 \pm 0.004 \pm 0.010$  |
| $\Gamma_1^n$ | -0.033                       | $-0.033 \pm 0.008 \pm 0.013$ |
| BSR          | 0.168                        | $0.162 \pm 0.024$            |
| $\Delta u$   | 0.685                        | $0.821 \pm 0.034$            |
| $\Delta d$   | -0.320                       | $-0.437 \pm 0.035$           |
| $\Delta s$   | ...                          | $-0.098 \pm 0.037$           |
| $\Delta q$   | 0.365                        | $0.287 \pm 0.104$            |



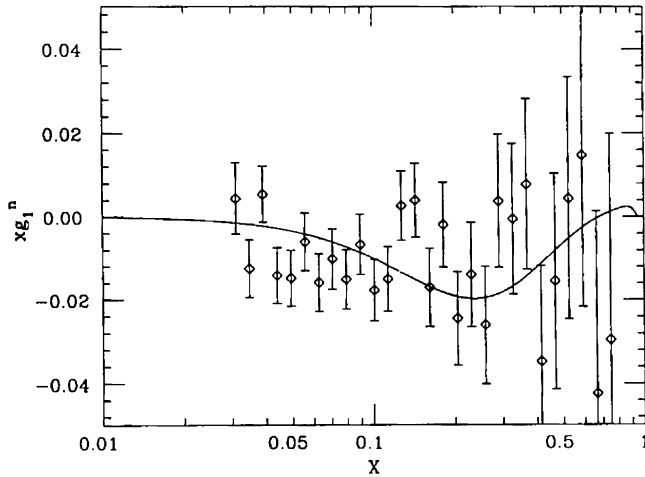
**Figure 9** The graph  $xg_1^p(x)$  vs  $x$  at  $Q^2 = 3 \text{ GeV}^2$ . The solid curve represents the theoretical calculation and the experimental data are from the E143 collaboration



**Figure 10** The graph  $g_1^p(x)/F_1^p(x)$  vs  $x$  at  $Q^2 = 3 \text{ GeV}^2$ . The solid curve represents the theoretical calculation [DM] and the experimental data are from the E143 collaboration

culated in this model are shown in Figures 9–11 along with the experimental data [27, 28]. This model calculation yields the Bjorken sum rule correctly. The strange quarks have not been included in this calculation, the inclusion of which will only increase the number of equations of state defining the nucleon bag and will involve a greater numerical effort to assess their contribution.





**Figure 11** The graph  $xg_1^n(x)$  vs  $x$  at  $Q^2 = 3 \text{ GeV}^2$ . The *solid curve* represents the theoretical calculation [DM] and the experimental data are from the E142 collaboration

## 6 Conclusion

The statistical model that is discussed in this article does not involve any free parameters. All the parameters have been determined by solving the equations of state of the nucleon considered as a MIT bag consisting of quarks and gluons. By boosting the statistical distribution functions, so obtained, to the infinite momentum frame (IMF), we obtain the quark distribution functions. These quark distribution functions are not realistic and they do not fit in with the experimental data. Taking explicitly the dependence of the quark distribution functions on the Bjorken variable  $x$ , renormalized quark distribution functions are obtained using a parameter  $\eta(x)$  to obtain realistic quark distribution functions for  $x > 1.5$ . To obtain fit with experimental data in small  $x$  region, the nucleon mass  $M$  is to be replaced by the invariant mass  $W$  of the final hadronic state and renormalized by using one more parameter. However this parameter is not absolutely free, as it is constrained by the condition that the realistic quark distribution function so obtained should yield, when integrated, the number of valence quarks in the nucleon. Both the unpolarized and polarized quark distribution functions and the unpolarized and polarized nucleon structure functions are obtained and they compare admirably well with the available experimental data. The small  $x$  region has not been studied in detail in this model but the steep rise in the structure function, observed experimentally, in the small  $x$  region, is adequately explained by this model. In conclusion, it is reiterated that the statistical model, discussed here, is a QCD inspired model and it gives a clear physical insight into the dynamical properties of the nucleon that are observed in DIS.

**Acknowledgements** This article is dedicated to the memory of Prof. Alladi Ramakrishnan who has inspired us to take to research and teaching. He can be truly called the Father of Theoretical Physics in South India and he has been a source of inspiration to successive generations of students in this part of the country. Much of the work reported in this review has been done in collaboration with K. Ganesamurthy and J. S. McCarthy. The authors thank Professor Krishnaswami Alladi for inviting us to contribute to this memorial volume.

## References

- [1] D. K. Dukes and J. F. Owens, *Phys. Rev.* **D30**, 49 (1984)
- [2] J. F. Owens, *Phys. Lett.* **B266**, 126 (1991)
- [3] V. Berger and R. J. N. Phillips, *Nuclear Phys.* **B73**, 269 (1974)
- [4] A. D. Martin, R. Roberts and W. J. Stirling, *Phys. Rev.* **D37**, 1161; *Phys. Lett.* **B206**, 327 (1989); *Mod. Phys. Lett.* **A4**, 1135 (1989)
- [5] M. Gluck, E. Reya and A. Vogt, *Z. Phys.* **C48**, 471 (1990); **C53**, 127 (1992)
- [6] A. D. Martin, W. J. Stirling and R. G. Roberts, *Phys. Lett.* **B306**, 145 (1993); *Phys. Rev.* **D47**, 867 (1993); *Phys. Rev.* **D50**, 6734 (1994)
- [7] M. Gluck, E. Reya and A. Vogt, *Phys. Lett.* **B306**, 391 (1993)
- [8] J. Botts et al., *Phys. Lett.* **B304**, 159 (1993)
- [9] CTEQ Collaboration, H. L. Lai et al., *Phys. Rev.* **D51**, 4763 (1995)
- [10] J. Cleymans and R. L. Thews, *Z. Phys.* **C37**, 315 (1988)
- [11] E. Mac and E. Ugaz, *Z. Phys.*, **C43**, 655 (1989)
- [12] R. P. Bickerstaff and J. T. Londergan, *Phys. Rev.* **C42**, 3621 (1990)
- [13] R. S. Bhalerao, *Phys. Lett.* **B380**, 1 (1996)
- [14] C. Bourrely, F. Buccella and J. Soffer, *Eur. Phys. J.* **C23**, 487 (2002); *Mod. Phys. Lett.* **A18**, 771 (2003); *Eur. Phys. J.* **C41**, 327 (2005); *Mod. Phys. Lett.* **A21**, 143 (2006); *Phys. Lett.* **B648**, 39 (2007)
- [15] V. Devanathan, S. Karthiyayini and K. Ganesamurthy, *Mod. Phys. Lett.* **A9**, 3455 (1994)
- [16] V. Devanathan, S. Karthiyayini and K. Ganesamurthy, *Parton Distributions and Nucleon Structure Functions*, in Perspectives in Theoretical Nuclear Physics (p.119), Eds. K. Srinivasa Rao and L. Satpathy, Wiley Eastern, New York (1994)
- [17] V. Devanathan, "Deep inelastic scattering and nucleon structure functions", in *New Perspectives in Classical and Quantum Physics*, Eds. P.P. Delsanto and A. W. Sáenz, Gordon and Breach, London (1995)
- [18] V. Devanathan, *Parton spin distribution functions*, in Selected Topics in Mathematical Physics: Professor R. Vasudevan Memorial Volume, Eds. R. Sridhar, K. Srinivasa Rao and V. Lakshminarayanan, Allied Publishers Ltd. (1995)
- [19] V. Devanathan and J. S. McCarthy, *Mod. Phys. Lett.* **A11**, 147 (1996)
- [20] V. Devanathan, *Ch. 14, Nuclear Physics*, Narosa Publishing House, New Delhi, India and Alpha Science International, Oxford, UK. (2006)
- [21] M. Rajasekaran, K. Ganesamurthy and V. Devanathan, *Mod. Phys. Lett.* **A5**, 473 (1990)
- [22] M. Rajasekaran, N. Meenakumari and V. Devanathan, *Mod. Phys. Lett.* **A5**, 2537 (1990)
- [23] G. Miller et al., *Phys. Rev.* **D5**, 528 (1972)
- [24] M. Derric et al., *Z. Phys.* **C65**, 379 (1995)
- [25] T. Ahmed et al. *Nuclear Phys.* **B439**, 471 (1995)
- [26] F. E. Close, *An Introduction to Quarks and Partons*, Academic Press, New York (1978)
- [27] SLAC, The E143 Collaboration, K. Abe et al., *Phys. Rev. Lett.* **74**, 346 (1995); **75**, 25 (1995); *Phys. Rev.* **D58**, 25 (1996); 112003 (1998)
- [28] SLAC, The E142 Collaboration, P. L. Anthony et al., *Phys. Rev. Lett.* **71**, 959 (1996); *Phys. Rev.* **D54**, 6620 (1996)

# On Generalized Clifford Algebras and their Physical Applications

Ramaswamy Jagannathan

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** Generalized Clifford algebras (GCAs) and their physical applications were extensively studied for about a decade from 1967 by Alladi Ramakrishnan and his collaborators under the name of  $L$ -matrix theory. Some aspects of GCAs and their physical applications are outlined here. The topics dealt with include: GCAs and projective representations of finite abelian groups, Alladi Ramakrishnan's  $\sigma$ -operation approach to the representation theory of Clifford algebra and GCAs, Dirac's positive energy relativistic wave equation, Weyl-Schwinger unitary basis for matrix algebra and Alladi Ramakrishnan's matrix decomposition theorem, finite-dimensional Wigner function, finite-dimensional canonical transformations, magnetic Bloch functions, finite-dimensional quantum mechanics, and the relation between GCAs and quantum groups.

**Mathematics Subject Classification (2010)** 15A66, 20C25, 20C35, 81R05

**Key words and phrases** Clifford algebra · Generalized Clifford algebras · Projective representations of finite abelian groups ·  $L$ -matrix theory · Dirac equation · Dirac's positive-energy relativistic wave equation · Dark matter · Heisenberg-Weyl commutation relation · Finite-dimensional Wigner function · Finite-dimensional canonical transformations · Finite-dimensional quantum mechanics · Kinematic confinement of quarks · Magnetic Bloch functions · Quantum groups

---

R. Jagannathan  
Chennai Mathematical Institute, Plot H1, SIPCOT IT Park, Padur P.O., Siruseri 603103,  
Tamilnadu, India  
Formerly of MATSCIENCE, The Institute of Mathematical Sciences, Chennai  
e-mail: [jagan@cmi.ac.in](mailto:jagan@cmi.ac.in); [jagan@imsc.res.in](mailto:jagan@imsc.res.in)

# 1 Introduction

Extensive studies on Clifford algebra, its generalizations, and their physical applications were made for about a decade starting from 1967, under the name of *L*-Matrix Theory, by Alladi Ramakrishnan and his collaborators at The Institute of Mathematical Sciences (MATSCIENCE) including me, his Ph.D student during 1971–1976. When I joined MATSCIENCE in August 1971, as a student, the book [1], containing all the results of their papers on the subject, up to then, was getting ready to be released; I had participated in the final stage of proof reading of the book. Chandrasekaran had just completed his Ph.D. thesis on the topic [2]. Subsequently, I started my thesis work on the same topic under the guidance of Alladi Ramakrishnan. I had also the guidance of Ranganathan, Santhanam, and Vasudevan, senior faculty members of the institute, who had also started their scientific careers under the guidance of Alladi Ramakrishnan and had contributed largely to the development of *L*-matrix theory. In my Ph.D. thesis [3] I had studied certain group theoretical aspects of generalized Clifford algebras (GCAs) and their physical applications. After my Ph.D. work also, I have applied the elements of GCAs in studies of certain problems in quantum mechanics and quantum groups. Here, I would like to outline some aspects of GCAs and their applications essentially based on my work.

A generalized Clifford algebra (GCA) can be presented, in general, as an algebra having a basis with generators  $\{e_j | j = 1, 2, \dots, n\}$  satisfying the relations:

$$\begin{aligned}
 e_j e_k &= \omega_{jk} e_k e_j, & \omega_{jk} e_l &= e_l \omega_{jk}, & \omega_{jk} \omega_{lm} &= \omega_{lm} \omega_{jk}, \\
 e_j^{N_j} &= 1, & \omega_{jk}^{N_j} &= \omega_{jk}^{N_k} = 1, & \forall j, k, l, m &= 1, 2, \dots, n.
 \end{aligned}
 \tag{1.1}$$

In any irreducible matrix representation, relevant for physical applications, one will have

$$\begin{aligned}
 \omega_{jk} &= \omega_{kj}^{-1} = e^{2\pi i v_{jk} / N_{jk}}, & N_{jk} &= \text{g.c.d}(N_j, N_k), \\
 & & j, k &= 1, 2, \dots, n
 \end{aligned}
 \tag{1.2}$$

where  $v_{jk}$ s are integers. Consequently, one can write

$$\begin{aligned}
 \omega_{jk} &= e^{2\pi i t_{jk} / \hat{N}}, & t_{kj} &= -t_{jk}, & \hat{N} &= \text{l.c.m}[N_{jk}], \\
 & & j, k &= 1, 2, \dots, n.
 \end{aligned}
 \tag{1.3}$$

Thus, any GCA can be characterized by an integer  $\hat{N}$  and an antisymmetric integer matrix

$$T = \begin{pmatrix} 0 & t_{12} & t_{13} & \dots & t_{1n} \\ -t_{12} & 0 & t_{23} & \dots & t_{2n} \\ -t_{13} & -t_{23} & 0 & \dots & t_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -t_{1n} & -t_{2n} & -t_{3n} & \dots & 0 \end{pmatrix}.
 \tag{1.4}$$

In the following we shall study the representation theory of GCAs and physical applications of some special cases of these algebras.

## 2 Projective Representations of Finite Abelian Groups and GCAs

GCAs arise in the study of projective, or ray, representations of finite abelian groups. Let us consider the finite abelian group  $G \cong \mathbb{Z}_{N_1} \otimes \mathbb{Z}_{N_2} \otimes \dots \otimes \mathbb{Z}_{N_n}$  with  $\{c_1^{m_1} c_2^{m_2} \dots c_n^{m_n}\}$  as its generic element where the generators  $\{c_j\}$  satisfy the relations

$$c_j c_k = c_k c_j, \quad c_j^{N_j} = 1, \quad j = 1, 2, \dots, n. \quad (2.1)$$

A projective representation  $D(G)$  of a group  $G$  is defined as

$$D(g_j)D(g_k) = \varphi(g_j, g_k)D(g_j g_k), \quad \varphi(g_j, g_k) \in \mathbb{C}, \quad \forall g_j, g_k \in G, \quad (2.2)$$

where the given factor set  $\{\varphi(g_j, g_k)\}$  is such that

$$\varphi(g_j, g_k)\varphi(g_j g_k, g_l) = \varphi(g_j, g_k g_l)\varphi(g_k, g_l), \quad \forall g_j, g_k, g_l \in G, \quad (2.3)$$

and

$$\varphi(E, g_j) = \varphi(g_j, E) = 1, \quad \forall g_j \in G, \quad (2.4)$$

with  $E$  as the identity element of  $G$ . For an abelian group, (2.2) implies:

$$\begin{aligned} D(g_j)D(g_k) &= \varphi(g_j, g_k)D(g_j g_k) = \varphi(g_j, g_k)D(g_k g_j) \\ &= \frac{\varphi(g_j, g_k)}{\varphi(g_k, g_j)}D(g_k)D(g_j), \quad \forall g_j, g_k \in G, \end{aligned} \quad (2.5)$$

or,

$$D(g_j)D(g_k) = \Omega_\varphi(g_j, g_k)D(g_k)D(g_j), \quad (2.6)$$

with

$$\Omega_\varphi(g_j, g_k) = \frac{\varphi(g_j, g_k)}{\varphi(g_k, g_j)}, \quad \forall g_j, g_k \in G. \quad (2.7)$$

Using (2.2) it is easy to see that we can write

$$D\left(\prod_{j=1}^n c_j^{m_j}\right) = \phi\left(\prod_{j=1}^n c_j^{m_j}\right) \left\{\prod_{j=1}^n D(c_j)^{m_j}\right\}, \quad (2.8)$$

with

$$\phi \left( \prod_{j=1}^n c_j^{m_j} \right) = \prod_{j=1}^n \prod_{p_j=1}^{m_j} \varphi \left( c_j, \prod_{l=0}^{n-j} c_j^{N_j-p_j} c_{j+l}^{m_{j+l}} \right)^{-1}. \quad (2.9)$$

From this it follows that

$$D \left( c_j^{N_j} \right) = \phi \left( c_j^{N_j} \right) D(c_j)^{N_j} = I, \quad \forall j = 1, 2, \dots, n., \quad (2.10)$$

where

$$\phi \left( c_j^{N_j} \right) = \prod_{p_j=1}^{N_j} \varphi \left( c_j, c_j^{N_j-p_j} \right)^{-1}. \quad (2.11)$$

Let us now define

$$e_j = \phi \left( c_j^{N_j} \right)^{1/N_j} D(c_j), \quad \forall j = 1, 2, \dots, n. \quad (2.12)$$

Then, it is found that the required representations satisfying (2.2–2.7), for the given factor set, are immediately obtained from (2.8–2.9) once the ordinary representations of  $\{e_j | j = 1, 2, \dots, n\}$  are found such that

$$\begin{aligned} e_j e_k &= \omega_{jk}^{(\varphi)} e_k e_j, \quad \omega_{jk}^{(\varphi)} e_l = e_l \omega_{jk}^{(\varphi)}, \quad \omega_{jk}^{(\varphi)} \omega_{lm}^{(\varphi)} = \omega_{lm}^{(\varphi)} \omega_{jk}^{(\varphi)}, \\ e_j^{N_j} &= 1, \quad \left( \omega_{jk}^{(\varphi)} \right)^{N_j} = \left( \omega_{jk}^{(\varphi)} \right)^{N_k} = 1, \quad \text{with } \omega_{jk}^{(\varphi)} = \Omega_{\varphi}(c_j, c_k), \\ &\forall j, k, l, m = 1, 2, \dots, n. \end{aligned} \quad (2.13)$$

Comparing (2.13) with (1.1) it is clear that the problem of finding the projective representations of any finite abelian group for any given factor set reduces to the problem of finding the ordinary representations of a generalized Clifford algebra defined by (1.1).

### 3 Representations of GCAs

Let us now consider a GCA associated with a specific antisymmetric integer matrix  $T$  as in (1.4) and an integer  $\hat{N}$ . The  $T$ -matrix can be related to its skew-normal form  $\mathcal{T}$  by a transformation as follows:

$$\begin{aligned} \mathcal{T} &= \begin{pmatrix} 0 & t_1 \\ -t_1 & 0 \end{pmatrix} \oplus \dots \oplus \begin{pmatrix} 0 & t_s \\ -t_s & 0 \end{pmatrix} \oplus O_{n-2s}, \\ T &= U \mathcal{T} \tilde{U} \ (\pm \text{mod. } \hat{N}), \end{aligned} \quad (3.1)$$

where  $O_{n-2s}$  is an  $(n-2s) \times (n-2s)$  null matrix,  $U = [u_{jk}]$  is a unimodular integer matrix with  $|u_{jk}| \leq \hat{N}$ , and  $\tilde{U}$  is the transpose of  $U$ . For any given antisymmetric integer matrix  $T$  it is possible to get the skew normal form  $\mathcal{T}$  and the corresponding  $U$ -matrix explicitly by a systematic procedure (see, e.g., [4]). Now, let  $\{\epsilon_j | j = 1, 2, \dots, n\}$  be a set of elements satisfying the commutation relations

$$\begin{aligned} \epsilon_{2j-1}\epsilon_{2j} &= e^{2\pi i t_j / \hat{N}} \epsilon_{2j}\epsilon_{2j-1}, & j &= 1, 2, \dots, s, \\ \epsilon_k\epsilon_l &= \epsilon_l\epsilon_k & \text{otherwise.} \end{aligned} \tag{3.2}$$

It is clear that this set of relations generate a GCA corresponding to  $\mathcal{T}$  as its  $T$ -matrix. It is straightforward to verify that if we construct  $\{e_j | j = 1, 2, \dots, n\}$  from  $\{\epsilon_j | j = 1, 2, \dots, n\}$  through a product transformation [3, 5]

$$e_j = \mu_j \epsilon_1^{u_{j1}} \epsilon_2^{u_{j2}} \dots \epsilon_n^{u_{jn}}, \quad \forall j = 1, 2, \dots, n, \tag{3.3}$$

where  $[u_{jk}] = U$  and  $\{\mu_j | j = 1, 2, \dots, n\}$  are complex numbers, then, in view of (3.1),

$$e_j e_k = e^{2\pi i t_{jk} / \hat{N}} e_k e_j, \quad \forall j, k = 1, 2, \dots, n., \tag{3.4}$$

as required in (1.1)–(1.3); the complex numbers  $\{\mu_j\}$  are normalization factors which are to be chosen such that

$$e_j^{N_j} = 1, \quad \forall j = 1, 2, \dots, n. \tag{3.5}$$

Now, let the matrix representations of  $\{\epsilon_j | j = 1, 2, \dots, 2s\}$  be given by:

$$\begin{aligned} \epsilon_1 &= I \otimes I \otimes I \otimes \dots \otimes I \otimes A_1, \\ \epsilon_2 &= I \otimes I \otimes I \otimes \dots \otimes I \otimes B_1, \\ \epsilon_3 &= I \otimes I \otimes I \otimes \dots \otimes A_2 \otimes I, \\ \epsilon_4 &= I \otimes I \otimes I \otimes \dots \otimes B_2 \otimes I, \\ &\vdots \\ \epsilon_{2s-1} &= A_s \otimes I \otimes I \otimes \dots \otimes I \otimes I, \\ \epsilon_{2s} &= B_s \otimes I \otimes I \otimes \dots \otimes I \otimes I, \end{aligned} \tag{3.6}$$

where

$$\begin{aligned} A_j B_j &= \omega_j^{\tau_j} B_j A_j, \\ \text{with } \omega_j &= e^{2\pi i / N_j}, \quad N_j = \hat{N} / (\text{g.c.d.}(t_j, \hat{N})), \\ \tau_j &= t_j / (\text{g.c.d.}(t_j, \hat{N})), \quad j = 1, 2, \dots, s, \end{aligned} \tag{3.7}$$

and  $I_s$  are identity matrices of appropriate dimensions. As  $\{\epsilon_k | k = 2s + 1, 2s + 2, \dots, n\}$  commute among themselves and also with all other  $\{e_j | j = 1, 2, \dots, 2s\}$  they are represented by unimodular complex numbers which can be absorbed in the normalization factors  $\{\mu_j\}$  in (3.3). This shows that if the matrix representations of all  $A_s$  and  $B_s$  satisfying (3.7) are known, then the problem of representation of the given GCA is solved. Explicitly, one has, apart from multiplicative normalizing phase factors,

$$e_j \sim A_s^{\mu_j(2s-1)} B_s^{\mu_j(2s)} \otimes A_{s-1}^{\mu_j(2s-3)} B_{s-1}^{\mu_j(2s-2)} \otimes \dots \otimes A_1^{\mu_j 1} B_1^{\mu_j 2},$$

$$\forall j = 1, 2, \dots, n. \tag{3.8}$$

Note that  $\omega_j^{\tau_j}$  s in (3.7) are primitive roots of unity. Thus, the representation theory of any GCA depends essentially on the central relation

$$AB = \omega BA, \tag{3.9}$$

where  $\omega$  is a nontrivial primitive root of unity. If  $\omega$  is a primitive  $N$ th root of unity then the normalization relations for  $A$  and  $B$  can be

$$A^{jN} = I, \quad B^{kN} = I, \quad \text{where } j, k = 1, 2, \dots \tag{3.10}$$

The central relation (3.9) determines the representation of  $A$  and  $B$  uniquely up to multiplicative phase factors and the normalization relation (3.10) fixes these phase factors. For more details on projective representations of finite abelian groups and their relation to GCAs, and other different approaches to GCAs, see [6]–[13].

### 4 The Clifford Algebra

Hamilton’s quaternion, generalizing the complex number, is given by:

$$q = q_0 1 + q_1 i + q_2 j + q_3 k, \tag{4.1}$$

where  $\{q_0, q_1, q_2, q_3\}$  are real numbers, 1 is the identity unit, and  $\{i, j, k\}$  are imaginary units such that

$$ij = -ji, \quad jk = -kj, \quad ki = -ik,$$

$$i^2 = j^2 = k^2 = -1, \tag{4.2}$$

and

$$ij = k, \quad jk = i, \quad ki = j. \tag{4.3}$$



It should be noted that the relations in (4.3) are not independent of the commutation and normalization relations (4.2); to see this, observe that  $ijk$  commutes with each one of the imaginary units  $\{i, j, k\}$  and hence  $ijk \sim 1$ . The ‘geometric algebra’ of Clifford [14] has the generating relations

$$\begin{aligned} \iota_j \iota_k &= -\iota_k \iota_j, & \text{for } j \neq k \\ \iota_j^2 &= -1, & \forall j, k = 1, 2, \dots, n, \end{aligned} \tag{4.4}$$

obtained by generalizing (4.2). This is what has become the Clifford algebra defined by the generating relations

$$\begin{aligned} e_j e_k &= -e_k e_j, & \text{for } j \neq k, \\ e_j^2 &= 1, & \forall j, k = 1, 2, \dots, n, \end{aligned} \tag{4.5}$$

which differ from (4.4) only in the normalization conditions, and evolved into the GCA (1.1). Thus, the Clifford algebra (4.5) corresponds to (1.1) with the choice

$$\omega_{jk} = -1, \quad N_j = 2, \quad \forall j, k = 1, 2, \dots, n, \tag{4.6}$$

associated with the  $T$ -matrix

$$T = \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ -1 & 0 & 1 & \dots & 1 \\ -1 & -1 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & \dots & 0 \end{pmatrix}. \tag{4.7}$$

and

$$\hat{N} = 2. \tag{4.8}$$

The corresponding skew normal form is

$$\mathcal{T} = \underbrace{\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus \dots \oplus \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}}_{m \text{ times}}, \tag{4.9}$$

when  $n = 2m$ . When  $n = 2m + 1$ ,

$$\mathcal{T} = \underbrace{\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus \dots \oplus \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}}_{m \text{ times}} \oplus 0. \tag{4.10}$$

In this case the  $U$ -matrices are

$$U = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & -1 & 1 & \dots & -1 & 1 & -1 & 1 \\ 0 & 1 & -1 & 1 & \dots & -1 & 1 & -1 & 1 \end{pmatrix}, \tag{4.11}$$

for  $n = 2m$ ,

and

$$U = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -1 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & -1 & 1 & \dots & -1 & 1 & -1 & 1 & 0 \\ 0 & 1 & -1 & 1 & \dots & -1 & 1 & -1 & 1 & 0 \\ -1 & 1 & -1 & 1 & \dots & -1 & 1 & -1 & 1 & 1 \end{pmatrix}, \tag{4.12}$$

for  $n = 2m + 1$ ,

such that

$$T = UT\tilde{U} \pmod{2}. \tag{4.13}$$

Now, equation (3.2) becomes in this case, for both  $n = 2m$  and  $n = 2m + 1$ ,

$$\begin{aligned} \epsilon_{2j-1}\epsilon_{2j} &= -\epsilon_{2j}\epsilon_{2j-1}, & j = 1, 2, \dots, m, \\ \epsilon_k\epsilon_l &= \epsilon_l\epsilon_k, & \text{otherwise,} \end{aligned} \tag{4.14}$$

with the matrix representations

$$\begin{aligned} \epsilon_1 &= I \otimes I \otimes I \otimes \dots \otimes I \otimes A_1, \\ \epsilon_2 &= I \otimes I \otimes I \otimes \dots \otimes I \otimes B_1, \\ \epsilon_3 &= I \otimes I \otimes I \otimes \dots \otimes A_2 \otimes I, \\ \epsilon_4 &= I \otimes I \otimes I \otimes \dots \otimes B_2 \otimes I, \\ &\vdots \\ \epsilon_{2m-1} &= A_m \otimes I \otimes I \otimes \dots \otimes I \otimes I, \\ \epsilon_{2m} &= B_m \otimes I \otimes I \otimes \dots \otimes I \otimes I, \end{aligned} \tag{4.15}$$

where

$$A_j = \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B_j = \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

$$\forall j = 1, 2, \dots, m. \tag{4.16}$$

In the case of  $n = 2m + 1$ , as  $\epsilon_{2m+1}$  commutes with all other  $\epsilon_j$ s it can be just taken to be 1. The matrices  $\sigma_1$  and  $\sigma_3$  are the well known first and the third Pauli matrices, respectively, and the second Pauli matrix is given by:

$$\sigma_2 = i\sigma_1\sigma_3 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}. \tag{4.17}$$

Then, in view of (3.8) and (4.11, 4.12), the required representations of (4.5) are given in terms of the Pauli matrices by:

$$\begin{aligned} e_1 &= \sigma_1 \otimes I \otimes \dots \otimes I \otimes I, \\ e_2 &= \sigma_3 \otimes I \otimes \dots \otimes I \otimes I, \\ e_3 &= \sigma_2 \otimes \sigma_1 \otimes I \otimes \dots \otimes I \otimes I, \\ e_4 &= \sigma_2 \otimes \sigma_3 \otimes I \otimes \dots \otimes I \otimes I \\ &\vdots \\ e_{2m-1} &= \sigma_2 \otimes \sigma_2 \otimes \dots \otimes \sigma_2 \otimes \sigma_1, \\ e_{2m} &= \sigma_2 \otimes \sigma_2 \otimes \dots \otimes \sigma_2 \otimes \sigma_3, \\ e_{2m+1} &= \sigma_2 \otimes \sigma_2 \otimes \dots \otimes \sigma_2 \otimes \sigma_2. \end{aligned} \tag{4.18}$$

Note that this representation is Hermitian and unitary. One can show that this is an irreducible representation. Also it should be noted that the earlier representation matrices are defined only upto multiplication by  $\pm 1$  because  $e_j^2 = 1$  for all  $j$ .

Let us now write down the generators of the first four Clifford algebras:

$$\begin{aligned} C^{(2)} : e_1^{(2)} &= \sigma_1, \quad e_2^{(2)} = \sigma_3, \\ C^{(3)} : e_1^{(3)} &= \sigma_1, \quad e_2^{(3)} = \sigma_3, \quad e_3^{(3)} = \sigma_2, \\ C^{(4)} : e_1^{(4)} &= \sigma_1 \otimes I, \quad e_2^{(4)} = \sigma_3 \otimes I, \\ &e_3^{(4)} = \sigma_2 \otimes \sigma_1, \quad e_4^{(4)} = \sigma_2 \otimes \sigma_3, \\ C^{(5)} : e_1^{(5)} &= \sigma_1 \otimes I, \quad e_2^{(5)} = \sigma_3 \otimes I, \\ &e_3^{(5)} = \sigma_2 \otimes \sigma_1, \quad e_4^{(5)} = \sigma_2 \otimes \sigma_3, \quad e_5^{(5)} = \sigma_2 \otimes \sigma_2, \end{aligned} \tag{4.19}$$

where the superscript indicates the number of generators in the corresponding algebra. The dimension of the irreducible representation of the Clifford algebra with  $2m$ , or  $2m + 1$ , generators is  $2^m$ . One can show that for the algebra with an even number of generators there is only one unique irreducible representation up to

equivalence. In the case of the algebra with an odd number of generators there are two inequivalent irreducible representations where the other representation is given by multiplying all the matrices of the first representation by  $-1$ . These statements form Pauli's theorem on Clifford algebra.

An obvious irreducible representation of the identity and the three imaginary units of Hamilton's quaternion algebra (4.2, 4.3) is given by:

$$1 = I, \quad i = -i\sigma_1, \quad j = -i\sigma_3, \quad k = i\sigma_2. \tag{4.20}$$

From the above it is clear that, as Clifford remarked [14], the geometric algebra, or the Clifford algebra, is a compound of quaternion algebras the units of which are commuting with one another. Actually, (3.2) and (3.3) correspond precisely to Clifford's original construction of geometric algebra starting with commuting quaternion algebras; matrix representations and realization of commuting quaternion algebras in terms of direct products did not exist at that time. Later, obviously unaware of Clifford's work, Dirac [15] used the same procedure to construct his four matrices  $\{\alpha_x, \alpha_y, \alpha_z, \beta\}$ , building blocks of his relativistic theory of electron and other spin-1/2 particles, starting with the three Pauli matrices  $\{\sigma_1, \sigma_2, \sigma_3\}$ . The Dirac matrices are given by:

$$\alpha_x = \sigma_1 \otimes \sigma_1, \quad \alpha_y = \sigma_1 \otimes \sigma_2, \quad \alpha_z = \sigma_1 \otimes \sigma_3, \quad \beta = \sigma_3 \otimes I, \tag{4.21}$$

which can be shown to be equivalent to the representation of  $C^{(4)}$  given earlier; as already mentioned,  $C^{(4)}$  has only one inequivalent irreducible representation. Clifford algebra is basic to the theory of spinors, theory of fermion fields, Onsager's solution of the two dimensional Ising model, etc. For detailed accounts of Clifford algebra and its various physical applications see, e.g., [16]–[18].

### 5 Alladi Ramakrishnan's $L$ -Matrix Theory and $\sigma$ -Operation

Representation theory of Clifford algebra has been expressed by Alladi Ramakrishnan [1] in a very nice framework called the  $L$ -matrix theory. Let

$$L^{(2m+1)}(\underline{\lambda}) = \sum_{j=1}^{2m+1} \lambda_j e_j^{(2m+1)}, \tag{5.1}$$

called an  $L$ -matrix, be associated with a  $(2m + 1)$ -dimensional vector  $\underline{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_{2m+1})$ . It follows that

$$\left( L^{(2m+1)}(\underline{\lambda}) \right)^2 = \left( \sum_{j=1}^{2m+1} \lambda_j^2 \right) I = \|\underline{\lambda}\|^2 I, \tag{5.2}$$

where  $I$  is the  $2^m \times 2^m$  identity matrix. Thus,  $L^2$  represents the square of the norm, or the length, of the vector  $\underline{\lambda}$ . In other words,  $L$  is a square root of  $\sum \lambda_j^2$  linear in  $\{\lambda_j\}$ .

From (4.19) observe that

$$\begin{aligned} e_1^{(5)} &= e_1^{(3)} \otimes I, & e_2^{(5)} &= e_2^{(3)} \otimes I, \\ e_3^{(5)} &= e_3^{(3)} \otimes e_1^{(3)}, & e_4^{(5)} &= e_3^{(3)} \otimes e_2^{(3)}, & e_5^{(5)} &= e_3^{(3)} \otimes e_3^{(3)}. \end{aligned} \tag{5.3}$$

Thus, one can write

$$\begin{aligned} L^{(5)}(\underline{\lambda}) &= \sum_{j=1}^5 \lambda_j e_j^{(5)} \\ &= e_1^{(3)} \otimes \lambda_1 I + e_2^{(3)} \otimes \lambda_2 I + e_3^{(3)} \otimes (\lambda_3 e_1^{(3)} + \lambda_4 e_2^{(3)} + \lambda_5 e_3^{(3)}), \end{aligned} \tag{5.4}$$

i.e.,  $L^{(5)}$  can be obtained from  $L^{(3)}$  by replacing  $\lambda_1, \lambda_2,$  and  $\lambda_3$  by  $\lambda_1 I, \lambda_2 I,$  and  $L^{(3)}(\lambda_3, \lambda_4, \lambda_5)$ , respectively. From (4.18) it is straightforward to see that this procedure generalizes: an  $L^{(2m+3)}$  can be obtained from an  $L^{(2m+1)}$  by replacing  $(\lambda_1, \lambda_2, \dots, \lambda_{2m})$ , respectively, by  $(\lambda_1 I, \lambda_2 I, \dots, \lambda_{2m} I)$ , and  $\lambda_{2m+1}$  by  $L^{(3)}(\lambda_{2m+1}, \lambda_{2m+2}, \lambda_{2m+3})$ . This procedure is called  $\sigma$ -operation by Alladi Ramakrishnan. It can be shown that the induced representation technique of group theory takes this form in the context of Clifford algebra [19]. Actually, in this procedure any one of the parameters of  $L^{(2m+1)}$  can be replaced by an  $L^{(3)}$  and the remaining parameters  $\{\lambda_j\}$  can be replaced, respectively, by  $\{\lambda_j I\}$  with suitable relabelling. As we shall see later this  $\sigma$ -operation generalizes to the case of GCAs with ordered  $\omega$ -commutation relations.

Another interesting result of Alladi Ramakrishnan is on the diagonalization of an  $L$ -matrix. An  $L^{(2m+1)}$ -matrix of dimension  $2^m$  obeys

$$\left( L^{(2m+1)} \right)^2 = \sum_{j=1}^{2m+1} \lambda_j^2 I = \Lambda^2 I, \tag{5.5}$$

and hence has  $(\Lambda, -\Lambda)$  as its eigenvalues each being  $2^{m-1}$ -fold degenerate. In general, let us call the matrix  $e_2^{(2m+1)}$ , or  $e_2^{(2m)}$ , as  $\beta$ :

$$\beta = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}, \tag{5.6}$$

where  $I$  is the  $2^{m-1}$ -dimensional identity matrix. Thus, the diagonal form of  $L$  is  $\Lambda\beta$ . Then, from the relation

$$L(L + \Lambda\beta) = \Lambda^2 I + L\Lambda\beta = (L + \Lambda\beta)\Lambda\beta, \tag{5.7}$$

it follows that  $(L + \Lambda\beta)$  is the matrix diagonalizing  $L$  and the columns of  $(L + \Lambda\beta)$  are the eigenvectors of  $L$ . Note that an  $L^{(2^m)}$ -matrix, of dimension  $2^m$ , can be treated as an  $L^{(2^{m+1})}$ -matrix with one of the  $\lambda$ s as zero.

Let us now take a Hermitian  $L(\underline{\lambda})$ -matrix where all the  $\lambda$ -parameters are real. As  $\beta = e_2$  anticommutes with all the other  $e_j$ s we get

$$(L + \Lambda\beta)^2 = 2\Lambda^2 I + \Lambda(L\beta + \beta L) = 2\Lambda(\Lambda + \lambda_2)I. \quad (5.8)$$

Hence

$$U = \frac{L + \Lambda\beta}{\sqrt{2\Lambda(\Lambda + \lambda_2)}} \quad (5.9)$$

is Hermitian and unitary ( $U = U^\dagger = U^{-1}$ ) and is such that

$$U^{-1}LU = \Lambda\beta. \quad (5.10)$$

Thus, the columns of  $U$  are normalized eigenvectors of the Hermitian  $L$ . This result has been applied [1] to solve in a very simple manner Dirac's relativistic wave equation [15],

$$i\hbar \frac{\partial \psi(\vec{r}, t)}{\partial t} = \left[ -i\hbar c \left( \alpha_x \frac{\partial}{\partial x} + \alpha_y \frac{\partial}{\partial y} + \alpha_z \frac{\partial}{\partial z} \right) + mc^2 \beta \right] \psi(\vec{r}, t), \quad (5.11)$$

where  $\psi(\vec{r}, t)$  is the 4-component spinor associated with the free spin-1/2 particle.

## 6 Dirac's Positive-Energy Relativistic Wave Equation

Students of Alladi Ramakrishnan got excellent training as professional scientists. He emphasized that the students should master any topic of research by studying the works of the leaders in the field and should communicate with their peers whenever necessary. In this connection, I would like to recall proudly an incident.

Following a suggestion of Santhanam, my fellow junior student Dutt and I started studying a paper of Dirac [20] in which he had proposed a positive-energy relativistic wave equation:

$$\begin{aligned} & i\hbar \frac{\partial [\hat{q}\psi(\vec{r}, t; q_1, q_2)]}{\partial t} \\ & = \left[ -i\hbar c \left( \alpha'_x \frac{\partial}{\partial x} + \alpha'_y \frac{\partial}{\partial y} + \alpha'_z \frac{\partial}{\partial z} \right) + mc^2 \beta' \right] [\hat{q}\psi(\vec{r}, t; q_1, q_2)], \end{aligned} \quad (6.1)$$

$[\hat{q}\psi]$  being a 4-component column matrix with elements  $(\hat{q}_1\psi, \hat{q}_2\psi, \hat{q}_3\psi, \hat{q}_4\psi)$  where

$$[\hat{q}_j, \hat{q}_k] = -\beta'_{jk}, \quad j, k = 1, 2, 3, 4, \quad (6.2)$$

and

$$\beta' = \sigma_2 \otimes I, \quad \alpha'_x = -\sigma_1 \otimes \sigma_3, \quad \alpha'_y = \sigma_1 \otimes \sigma_1, \quad \alpha'_z = \sigma_3 \otimes I. \quad (6.3)$$

Unlike the standard relativistic wave equation for the electron (5.11) which has both positive and negative (antiparticle) energy solutions, the new Dirac equation (6.1) has only positive energy solutions. Further, more interestingly, this positive-energy particle would not interact with an electromagnetic field. Around November 1974, Dutt and I stumbled upon an equation which had only negative-energy solutions. Our negative-energy relativistic wave equation was exactly the same as Dirac's positive-energy equation (6.1) except only for a slight change in the commutation relations of the internal variables  $(\hat{q}_1, \hat{q}_2, \hat{q}_3, \hat{q}_4)$  in the equation; instead of (6.2), we took

$$[\hat{q}_j, \hat{q}_k] = \beta'_{jk}, \quad j, k = 1, 2, 3, 4. \quad (6.4)$$

When I told Alladi Ramakrishnan about this he told us that we could not meddle with Dirac's work and keep quiet. He suggested that I should write to Dirac and get his opinion on our work. I wrote to Dirac who was in The Florida State University at that time. I received a letter from him within a month! His reply was: "*Dear Jagannathan, The equation you propose would correctly describe a particle with only negative-energy states. It would be the correct counterpart of the positive-energy equation, but of course it would not have any physical application. Yours sincerely, P. A. M. Dirac.*" Immediately, Alladi Ramakrishnan forwarded our paper for rapid publication [21].

So far, no one has found any application for Dirac's positive-energy equation. Attempts to modify it so that these positive-energy particles could interact with electromagnetic field have not succeeded. May be, these positive-energy Dirac particles and their negative-energy antiparticles constitute the dark matter of our universe.

## 7 GCAs with Ordered $\omega$ -Commutation Relations

We shall now consider a GCA (1.1) with ordered  $\omega$ -commutation relations, i.e.,

$$\begin{aligned} e_j e_k &= \omega e_k e_j, \quad \omega = e^{2\pi i/N}, \quad \forall j < k, \\ e_j^N &= 1, \quad j, k = 1, 2, \dots, n. \end{aligned} \quad (7.1)$$

The associated  $T$ -matrix has elements

$$t_{jk} = \begin{cases} 1, & \text{for } j < k, \\ 0, & \text{for } j = k, \\ -1, & \text{for } j > k, \end{cases} \tag{7.2}$$

and  $\hat{N} = N$ . This is exactly same as for the Clifford algebra except for the value of  $\hat{N}$ . So, the treatment of representation theory of this GCA is along the same lines as for the Clifford algebra:  $T$  matrix is the same as in (4.7) for any  $n$  and  $\mathcal{T}$  and  $U$  matrices are the same as in (4.9) and (4.11) for  $n = 2m$  and (4.10) and (4.12) for  $n = 2m + 1$ , respectively. The only difference is that in the case of Clifford algebra  $A_j^{-1} = A_j$  and  $B_j^{-1} = B_j$  for any  $j$ , where as now  $A_j^{-1} = A_j^{N-1}$  and  $B_j^{-1} = B_j^{N-1}$  for any  $j$ . Thus, in view of (3.8) and (4.11, 4.12), the required representations of (7.1) are given by:

$$\begin{aligned} e_1 &= A \otimes I \otimes \cdots \otimes I \otimes I, \\ e_2 &= B \otimes I \otimes \cdots \otimes I \otimes I, \\ e_3 &= \mu A^{-1} B \otimes A \otimes I \otimes \cdots \otimes I \otimes I, \\ e_4 &= \mu A^{-1} B \otimes B \otimes I \otimes \cdots \otimes I \otimes I, \\ e_5 &= \mu^2 A^{-1} B \otimes A^{-1} B \otimes A \otimes I \otimes \cdots \otimes I, \\ e_6 &= \mu^2 A^{-1} B \otimes A^{-1} B \otimes B \otimes I \otimes \cdots \otimes I, \\ &\vdots \\ e_{2m-1} &= \mu^{m-1} A^{-1} B \otimes A^{-1} B \otimes \cdots \otimes A^{-1} B \otimes A, \\ e_{2m} &= \mu^{m-1} A^{-1} B \otimes A^{-1} B \otimes \cdots \otimes A^{-1} B \otimes B, \\ e_{2m+1} &= \mu^m A^{-1} B \otimes A^{-1} B \otimes \cdots \otimes A^{-1} B \otimes A^{-1} B, \end{aligned} \tag{7.3}$$

where  $\mu = \omega^{(N+1)/2}$  and

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix}, \tag{7.4}$$

$$B = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & \omega & 0 & \dots & 0 \\ 0 & 0 & \omega^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \omega^{N-1} \end{pmatrix}, \tag{7.5}$$



$N \times N$  unitary matrices, obeying

$$AB = \omega BA, \quad A^N = B^N = I. \tag{7.6}$$

The matrices  $A$  and  $B$  in (7.4) and (7.5), respectively, provide the only irreducible representation for the relation (7.6) [22]. It can also be shown that the GCA  $\mathcal{C}_N^{(n)}$  defined by (7.1) has only one  $N^m$ -dimensional irreducible representation, as given by (7.3) without  $e_{2m+1}$ , when  $n = 2m$  and there are  $N$  inequivalent irreducible representations of dimension  $N^m$  (differing from (7.3) only by multiplications by powers of  $\omega$ ) when  $n = 2m + 1$  (see, e.g., [23, 24]). This is the generalization of Pauli's theorem for the GCA (7.1). When  $N = 2$  it is seen that  $A = \sigma_1, B = \sigma_3$ , and the representation (7.3) becomes the representation (4.18) of the Clifford algebra.

From the structure of the representation (7.3) it is clear that the  $\sigma$ -operation procedure should work in this case also. Let the  $n$ -dimensional vector  $\underline{\lambda} = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$  be associated with an  $\mathcal{L}$ -matrix defined by:

$$\mathcal{L}^{(n)} = \sum_{j=1}^n \lambda_j e_j^{(n)} \tag{7.7}$$

Then, from the commutation relations (7.1) it follows that

$$\left(\mathcal{L}^{(n)}\right)^N = \left(\sum_{j=1}^n \lambda_j^N\right) I. \tag{7.8}$$

Thus, the  $N$ -th root of  $\sum_{j=1}^n \lambda_j^N$  is given by  $\mathcal{L}^{(n)}$  which is linear in  $\lambda_j$ s. This fact helps linearize certain  $N$ -th order partial differential operators using the GCA [6] similar to the way Clifford algebra helps linearize certain second order partial differential operators (e.g., Dirac's linearization of  $\hat{H}^2 = -\hbar^2 c^2 \nabla^2 + m^2 c^4$  to get his relativistic Hamiltonian  $\hat{H} = -i\hbar c(\alpha_x \partial/\partial x + \alpha_y \partial/\partial y + \alpha_z \partial/\partial z) + mc^2 \beta$ ). Now, it can be easily seen [1] that  $\mathcal{L}^{(2m+3)}$  is obtained from  $\mathcal{L}^{(2m+1)}$  by the  $\sigma$ -operation: replace  $(\lambda_1, \lambda_2, \dots, \lambda_{2m})$  in  $\mathcal{L}^{(2m+1)}$  by  $(\lambda_1 I, \lambda_2 I, \dots, \lambda_{2m} I)$ , respectively, where  $I$  is the  $N$ -dimensional identity matrix, and  $\lambda_{2m+1}$  by  $\mathcal{L}^{(3)}(\lambda_{2m+1}, \lambda_{2m+2}, \lambda_{2m+3})$ .

From the above it is clear that the matrices  $A$  and  $B$  in (7.4) and (7.5), respectively, obeying the relation (7.6), play a central role in the study of GCAs. If we want to have two matrices  $A_j$  and  $B_j$  obeying

$$A_j B_j = e^{2\pi i j/N} B_j A_j, \quad \text{g.c.d}(j, N) = 1, \tag{7.9}$$

then,  $A_j$  is same as  $A$  in (7.4) and  $B_j$  is given by  $B$  in (7.5) with  $\omega$  replaced by  $\omega^j$ , upto multiplicative factors which are to be determined by the required normalization relations like (3.10). In the following we shall outline some of the physical applications of the matrices  $A$  and  $B$ .

One approach to study the representation theory of the GCA with ordered  $\omega$ -commutation relations (7.1) is to study the vector, or the ordinary, representations of the group

$$\mathcal{G} : \left\{ \omega^{j_0} e_1^{j_1} e_2^{j_2} \dots e_n^{j_n} \mid j_0, j_1, j_2, \dots, j_n = 0, 1, 2, \dots, N - 1 \right\}. \quad (7.10)$$

This group has been called a generalized Clifford group (GCG) and the study of its representation theory involves interesting number theoretical aspects ([23, 24]). Particularly, by studying the representations of the lowest order GCG generated by  $A$ ,  $B$ , and  $\omega$  one can show that  $A$  and  $B$  have only one irreducible representation as given by (7.4) and (7.5). Study of spin systems defined on a GCG also involves very interesting number theoretical problems [25]. Alladi Ramakrishnan and collaborators used the  $L$ -matrix theory for studying several topics like idempotent matrices, special unitary groups arising in particle physics, algebras derived from polynomial conditions, Duffin-Kemmer-Petiau algebra, and para-Fermi algebra (for details see [1]). They studied essentially the GCA with ordered  $\omega$ -commutation relations (7.1). The more general GCAs (1.1) were studied later in ([3, 5, 13, 23, 24]). In gauge field theories Wilson operators and 't Hooft operators satisfy commutation relations of the form in (7.6) and the corresponding algebra is often called the 't Hooft-Weyl algebra (see, e.g., [26]). For the various other physical applications of GCAs see, e.g., ([27, 28]).

### 8 Weyl-Schwinger Unitary Basis for Matrix Algebra and Alladi Ramakrishnan's Matrix Decomposition Theorem

Heisenberg's canonical commutation relation between position and momentum operators of a particle, the basis of quantum mechanics, is

$$[\hat{q}, \hat{p}] = i\hbar. \quad (8.1)$$

Weyl [22] wrote it in exponential form as:

$$e^{i\eta \hat{p}/\hbar} e^{i\xi \hat{q}/\hbar} = e^{i\xi \eta/\hbar} e^{i\xi \hat{q}/\hbar} e^{i\eta \hat{p}/\hbar}, \quad (8.2)$$

where the parameters  $\xi$  and  $\eta$  are real numbers, and studied its representation as the large  $N$  limit of the relation :

$$AB = \omega BA, \quad \omega = e^{2\pi i/N}. \quad (8.3)$$

Note that the Heisenberg-Weyl commutation relation (8.2) takes the form (8.3) when  $\xi\eta/\hbar = 2\pi/N$ . Weyl established that the relation (8.3), subject to the normalization condition

$$A^N = B^N = I, \quad (8.4)$$

has only one irreducible representation as given in (7.4) and (7.5). Analysing the large  $N$  limits of  $A$  and  $B$ , he showed that the the relation (8.2), or equivalently the Heisenberg commutation relation (8.1), has the unique (upto equivalence) irreducible representation given by the Schrödinger representation

$$\hat{q}\psi(q) = q\psi(q), \quad \hat{p}\psi(q) = -i\hbar \frac{d}{dq}\psi(q), \quad \text{for any } \psi(q). \quad (8.5)$$

This result, or the Stone-von Neumann theorem obtained later by a more rigorous approach, is of fundamental importance for physics because it establishes the uniqueness of quantum mechanics. Thus, Weyl viewed quantum kinematics as an irreducible Abelian group of unitary ray rotations in system space.

Following the earlier approach to quantum kinematics Weyl gave his correspondence rule for obtaining the quantum operator  $\hat{f}(\hat{q}, \hat{p})$  for a classical observable  $f(q, p)$ :

$$\begin{aligned} \hat{f}(\hat{q}, \hat{p}) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) e^{i(\xi\hat{q} + \eta\hat{p})}, \\ g(\xi, \eta) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} dq dp f(q, p) e^{-i(\xi q + \eta p)}. \end{aligned} \quad (8.6)$$

The fact that the set of  $N^2$  linearly independent unitary matrices  $\{A^k B^l | k, l = 0, 1, 2, \dots, (N - 1)\}$  forms a basis for the  $N \times N$ -matrix algebra is implicit in this suggestion that any quantum operator corresponding to a classical observable can be written as a linear combination of the unitary operators  $\{e^{i(\xi\hat{q} + \eta\hat{p})}\}$ .

Schwinger [29] studied in detail the role of the matrices  $A$  and  $B$  in quantum mechanics and hence the set  $\{A^k B^l | k, l = 0, 1, 2, \dots, (N - 1)\}$  is often called Schwinger's unitary basis for matrix algebra. Let us write an  $N \times N$  matrix  $M$  as:

$$M = \sum_{k,l=0}^{N-1} \mu_{kl} A^k B^l. \quad (8.7)$$

From the structure of the matrices  $A$  and  $B$  it is easily found that

$$\text{Tr} \left[ (A^k B^l)^\dagger (A^m B^n) \right] = N \delta_{km} \delta_{ln}. \quad (8.8)$$

Hence,

$$\mu_{kl} = \frac{1}{N} \text{Tr} \left[ (A^k B^l)^\dagger M \right] = \text{Tr} \left[ B^{-l} A^{-k} M \right]. \quad (8.9)$$

Alladi Ramakrishnan wrote (8.7) equivalently as:

$$M = \sum_{k,l=0}^{N-1} c_{kl} B^k A^l, \quad (8.10)$$

and expressed the coefficients  $\{c_{kl}\}$  in a very nice form [1] :

$$C = S^{-1}R, \tag{8.11}$$

$$C = \begin{pmatrix} c_{00} & c_{01} & c_{02} & \dots & c_{0,N-1} \\ c_{10} & c_{11} & c_{12} & \dots & c_{1,N-1} \\ c_{20} & c_{21} & c_{22} & \dots & c_{2,N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{N-2,0} & c_{N-2,1} & c_{N-2,2} & \dots & c_{N-2,N-1} \\ c_{N-1,0} & c_{N-1,1} & c_{N-1,2} & \dots & c_{N-1,N-1} \end{pmatrix}, \tag{8.12}$$

$$S^{-1} = \frac{1}{N} \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \omega^{-1} & \omega^{-2} & \dots & \omega^{-(N-1)} \\ 1 & \omega^{-2} & \omega^{-4} & \dots & \omega^{-2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{-(N-2)} & \omega^{-2(N-2)} & \dots & \omega^{-(N-2)(N-1)} \\ 1 & \omega^{-(N-1)} & \omega^{-2(N-1)} & \dots & \omega^{-(N-1)(N-1)} \end{pmatrix}, \tag{8.13}$$

$$R = \begin{pmatrix} M_{00} & M_{01} & M_{02} & \dots & M_{0,N-1} \\ M_{11} & M_{12} & M_{13} & \dots & M_{10} \\ M_{22} & M_{23} & M_{24} & \dots & M_{21} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ M_{N-2,N-2} & M_{N-2,N-1} & M_{N-2,0} & \dots & M_{N-2,N-3} \\ M_{N-1,N-1} & M_{N-1,0} & M_{N-1,1} & \dots & M_{N-1,N-2} \end{pmatrix}. \tag{8.14}$$

Note that  $S^{-1}$  is the inverse of the Sylvester, or the finite Fourier transform, matrix. He called (8.10)–(8.14) as a matrix decomposition theorem. Comparing (8.7) and (8.10) it is clear that  $\mu_{kl} = \omega^{-kl} c_{lk}$ .

### 9 Finite-Dimensional Wigner Function

Let  $N = 2\nu + 1$  and choose

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix}, \tag{9.1}$$

and

$$B = \begin{pmatrix} \omega^{-\nu} & 0 & 0 & \dots & 0 & 0 \\ 0 & \omega^{-\nu+1} & 0 & \dots & 0 & 0 \\ 0 & 0 & \omega^{-\nu+2} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \omega^{\nu-1} & 0 \\ 0 & 0 & 0 & \dots & 0 & \omega^{\nu} \end{pmatrix}, \tag{9.2}$$

where  $\omega = e^{2\pi i/(2\nu+1)}$ . Note that  $AB = \omega BA$  and  $A^{2\nu+1} = B^{2\nu+1} = I$ . Let us now write a  $2\nu + 1$ -dimensional matrix  $M$  as:

$$M = \sum_{k,l=-\nu}^{\nu} v_{kl} \omega^{kl/2} B^k A^l, \tag{9.3}$$

where

$$v_{kl} = \frac{1}{2\nu + 1} \text{Tr} \left[ \omega^{-kl/2} A^{-l} B^{-k} M \right]. \tag{9.4}$$

If the matrix  $M$  is to be Hermitian, i.e.,  $M^\dagger = M$ , then the condition to be satisfied is that  $v_{kl}^* = v_{-k,-l}$ .

Let  $W = (w_{kl})$ , with  $k, l = -\nu, -\nu + 1, \dots, \nu - 1, \nu$ , be a real matrix and define the finite-dimensional Fourier transform

$$v_{\xi\eta} = \frac{1}{2\nu + 1} \sum_{k,l=-\nu}^{\nu} w_{kl} \omega^{-\xi k - \eta l}. \tag{9.5}$$

We have

$$v_{\xi\eta}^* = v_{-\xi,-\eta}. \tag{9.6}$$

Hence, the matrix

$$\begin{aligned} H &= \sum_{\xi,\eta=-\nu}^{\nu} v_{\xi,\eta} \omega^{\xi\eta/2} B^\xi A^\eta \\ &= \frac{1}{2\nu + 1} \sum_{\xi,\eta=-\nu}^{\nu} \sum_{k,l=-\nu}^{\nu} w_{kl} \omega^{-\xi k - \eta l + (\xi\eta/2)} B^\xi A^\eta \end{aligned} \tag{9.7}$$

is Hermitian. This property, that to every real matrix  $W$  there is associated a unique Hermitian matrix  $H$ , is the basis of the Weyl correspondence (8.6). For a given Hermitian matrix  $H$  the associated real matrix  $W$  is obtained from (9.7) as:

$$w_{kl} = \text{Tr} \left[ \omega^{\xi k + \eta l - (\xi\eta/2)} A^{-\eta} B^{-\xi} H \right]. \tag{9.8}$$

In the large  $\nu$  limit this provides the converse of the Weyl rule (8.6) for obtaining the classical observable corresponding to a quantum operator or the Wigner transform of a quantum operator; in particular, the Wigner phase-space quasiprobability distribution function can be obtained as the limiting case of (9.8) corresponding to the choice of  $H$  as the quantum density operator [3]. Thus, the formula (9.8) can be viewed as an expression of the finite-dimensional Wigner function corresponding to the case when  $H$  is a finite-dimensional density matrix. For more details on finite-dimensional, or discrete, Wigner functions, which are of current interest in quantum information theory, see, e.g., [30].

## 10 Finite-Dimensional Quantum Canonical Transformations

As seen earlier, the relation (7.6) has a unique representation for  $A$  and  $B$  as given by (7.4) and (7.5). Let us take  $N$  to be even and make a transformation

$$A \longrightarrow A' = \omega^{-kl/2} A^k B^l, \quad B \longrightarrow B' = \omega^{-mn/2} A^m B^n, \quad (10.1)$$

where  $(k, l, m, n)$  can be in general taken to be nonnegative integers in  $[0, N - 1]$ , and require

$$A' B' = \omega B' A', \quad A'^N = B'^N = I. \quad (10.2)$$

This implies that we should have

$$kn - lm = 1 \pmod{N}, \quad (10.3)$$

and the factors  $\omega^{-kl/2}$  and  $\omega^{-mn/2}$  ensure that  $A'^N = B'^N = I$ . The uniqueness of the representation requires that there should be a definite solution to the equivalence relation

$$SA = A'S, \quad SB = B'S. \quad (10.4)$$

Substituting the explicit matrices for  $A$  and  $B$  from (7.4) and (7.5) it is straightforward to solve for  $S$ . We get

$$S_{xy} = \omega^{-(nx^2 - 2xy + ky^2)/2m}, \quad x, y = 0, 1, 2, \dots, N - 1. \quad (10.5)$$

From the association, following Weyl,

$$A \longrightarrow e^{i\nu\hat{p}/\hbar}, \quad B \longrightarrow e^{i\xi\hat{q}/\hbar}, \quad (10.6)$$

it follows that in the limit of  $N \longrightarrow \infty$  the finite-dimensional transformation (10.1) becomes the linear canonical transformation of the pair  $(\hat{q}, \hat{p})$ ,

$$\hat{q}' = n\hat{q} + m\hat{p}, \quad \hat{p}' = l\hat{q} + k\hat{p}. \quad (10.7)$$

By taking the corresponding limit of the matrix  $S$  in (10.5) one gets the unitary transformation corresponding to the quantum linear canonical transformation (10.7) ([3, 31]) (for details of the quantum canonical transformations see [32]).

## 11 Magnetic Bloch Functions

For an electron of charge  $-e$  and mass  $m$  moving in a crystal lattice under the influence of an external constant homogeneous magnetic field the stationary state wavefunction corresponding to the energy eigenvalue  $E$  satisfies the Schrödinger equation

$$\begin{aligned} \hat{\mathcal{H}}\psi(\vec{r}) &= E\psi(\vec{r}), \\ \hat{\mathcal{H}} &= \frac{1}{2m} \left( \vec{\hat{p}} + e\vec{A} \right)^2 + V(\vec{r}), \end{aligned} \tag{11.1}$$

where  $\vec{\hat{p}}$  is the momentum operator  $-i\hbar\vec{\nabla}$ ,  $V(\vec{r})$  is the periodic crystal potential, and  $\vec{A} = \frac{1}{2}(\vec{B} \times \vec{r})$  is the vector potential of the magnetic field  $\vec{B}$ . In the absence of the magnetic field the Hamiltonian is invariant under the group of lattice translations and as a consequence the corresponding wavefunction takes the form of a Bloch function:

$$\psi_{\vec{B}=0}(\vec{r}) = \sum_{\vec{R}} e^{-i\vec{K}\cdot\vec{R}} u(\vec{r} + \vec{R}), \tag{11.2}$$

where  $\{\vec{R}\}$  is the set of all lattice vectors and  $\vec{K}$  is a reciprocal lattice vector within a Brillouin zone. This is the basis of the band theory of solids. In the presence of a magnetic field the Hamiltonian  $\hat{\mathcal{H}}$  is not invariant under the lattice translation group. Now, the invariance group is the so-called magnetic translation group with its generators given by, apart from some phase factors,  $\{\tau_j = e^{i\vec{a}_j \cdot (\vec{\hat{p}} - e\vec{A})} | j = 1, 2, 3\}$  where  $\vec{a}_j$ s are the primitive lattice vectors. These generators obey the algebra:

$$\tau_j \tau_k = e^{-ie\vec{B} \cdot \vec{a}_j \times \vec{a}_k / \hbar} \tau_k \tau_j, \quad j, k = 1, 2, 3, \tag{11.3}$$

a GCA! We can obtain the irreducible representations of this algebra in terms of  $A$  and  $B$  matrices. Once the inequivalent irreducible representations of the magnetic translation group are known, using the standard group theoretical techniques we can construct the symmetry-adapted basis functions for the Schrödinger equation (11.1). This leads to a generalization of the Bloch function (11.2), the magnetic Bloch function, given by:

$$\psi(\vec{r}) = \sum_{\vec{R}} e^{-i\left[\left(\vec{K} + \frac{e}{2\hbar}\vec{B} \times \vec{r}\right) \cdot \vec{R} + \phi(\vec{R})\right]} u(\vec{r} + \vec{R}), \tag{11.4}$$

where

$$\phi(n_1\vec{a}_1+n_2\vec{a}_2+n_3\vec{a}_3) = \frac{e}{2\hbar}\vec{B}\cdot(n_1n_2\vec{a}_1\times\vec{a}_2+n_1n_3\vec{a}_1\times\vec{a}_3+n_2n_3\vec{a}_2\times\vec{a}_3). \quad (11.5)$$

If the term  $\phi(\vec{R})$  is dropped from this expression then it reduces to the well known form proposed by Peierls (for more details see ([31, 33, 34]) and references therein). Understanding the dynamics of a Bloch electron in a magnetic field is an important problem of condensed matter physics with various practical applications.

## 12 Finite-Dimensional Quantum Mechanics

Following are the prophetic words of Weyl [22]: *The kinematical structure of a physical system is expressed by an irreducible Abelian group of unitary ray rotations in system space. .... If the group is continuous this procedure automatically leads to Heisenberg's formulation. .... Our general principle allows for the possibility that the Abelian rotation group is entirely discontinuous, or that it may even be a finite group. .... But the field of discrete groups offers many possibilities which we have not yet been able to realize in Nature; perhaps, these holes will be filled by applications to nuclear physics.*

Keeping in mind the earlier statement of Weyl and the later work of Schwinger [29], a finite-dimensional quantum mechanics was developed by Santhanam and collaborators. Following Weyl, let us make the association

$$A \longrightarrow e^{i\eta\hat{p}/\hbar}, \quad B \longrightarrow e^{i\xi\hat{q}/\hbar}. \quad (12.1)$$

Now if we interpret the finite dimensional matrices  $A$  and  $B$  as corresponding to finite-dimensional momentum and position operators, say,  $P$  and  $Q$ , respectively, with finite discrete spectra, then, the corresponding system will have confinement purely as a result of its kinematical structure. The matrices  $P$  and  $Q$  can be obtained by taking the logarithms of  $A$  and  $B$ . The commutation relation between  $P$  and  $Q$  was first calculated by Santhanam and Tekumalla [35] (Tekumalla was my senior fellow student at our institute). Further work by Santhanam ([36]–[40]) along these lines resulted in the study of the Hermitian phase operator in finite dimensions as a precursor to the currently well known Pegg-Barnett formalism (see, e.g., [41]).

Later, we developed a formalism of finite-dimensional quantum mechanics (FDQM) ([42]–[44]) in which we studied the solutions of the Schrödinger equation with finite-dimensional matrix Hamiltonians obtained by replacing the position and momentum operators by finite-dimensional matrices  $Q$  and  $P$ . In [44] I interpreted quark confinement as a kinematic confinement as a consequence of its Weylian finite-dimensional quantum mechanics. Recently, dynamics of wave packets has been studied within the formalism of FDQM [45].



### 13 GCAs and Quantum Groups

Experience of working on GCAs helped me later in my work on quantum groups. An  $n \times n$  linear transformation matrix  $M$  acting on the noncommutative  $n$ -dimensional Manin vector space and its dual is a member of the quantum group  $GL_q(n)$  if its noncommuting elements  $m_{jk}$  satisfy certain commutation relations. For example, the elements of a  $2 \times 2$  quantum matrix belonging to  $GL_q(2)$ ,

$$M = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}, \quad (13.1)$$

have to satisfy the commutation relations

$$\begin{aligned} m_{11}m_{12} &= q^{-1}m_{12}m_{11}, & m_{11}m_{21} &= q^{-1}m_{21}m_{11}, \\ m_{12}m_{22} &= q^{-1}m_{22}m_{12}, & m_{21}m_{22} &= q^{-1}m_{22}m_{21}, \\ m_{12}m_{21} &= m_{21}m_{12}, & m_{11}m_{22} - m_{22}m_{11} &= (q^{-1} - q)m_{12}m_{21}. \end{aligned} \quad (13.2)$$

Some of these relations are already GCA-like, or Heisenberg-Weyl-like. It was shown in ([46, 47]) that, in general, all the commutation relations of  $GL_q(n)$  can be formulated in a similar form and hence the representations of these elements can be found utilising the representation theory of the Heisenberg-Weyl relations. Extending these ideas further, we developed in [48] a systematic scheme for constructing the finite and infinite dimensional representations of the elements of the quantum matrices of  $GL_q(n)$ , where  $q$  is a primitive root of unity, and discussed the explicit results for  $GL_q(2)$ ,  $GL_q(3)$ , and  $GL_q(4)$ . In this work we essentially used the product transformation technique ([3, 5]) developed in the context of representation theory of GCAs. In [49] we extended this formalism to the two-parameter quantum group  $GL_{p,q}(2)$  and the two-parameter quantum supergroup  $GL_{p,q}(1|1)$ .

### 14 Conclusion

To summarize, I have reviewed here some aspects of GCAs and their physical applications, mostly related to my own work. I learnt about it in the school of Alladi Ramakrishnan and it has been useful to me throughout my academic career so far. I would like to conclude with the following remark on GCAs by Alladi Ramakrishnan [50]:

*The structure is too fundamental to be unnoticed, too consistent to be ignored, and much too pretty to be without consequence.*

**Acknowledgement** I dedicate this article, with gratitude, to the memory of my teacher Professor Alladi Ramakrishnan under whose guidance I started my scientific career at MATSCIENCE, The Institute of Mathematical Sciences, Chennai.

## References

- [1] Alladi Ramakrishnan, *L-Matrix Theory or Grammar of Dirac Matrices*, Tata-McGraw Hill, NY, USA, 1972.
- [2] P. S. Chandrasekaran, *Clifford Algebra, its Generalization, and their Applications to Symmetries and Relativistic Wave Equations*, Ph.D. Thesis, University of Madras, 1971.
- [3] R. Jagannathan, *Studies in Generalized Clifford Algebras, Generalized Clifford Groups, and their Physical Applications*, Ph.D. Thesis, University of Madras, 1976.
- [4] M. Newman, *Integral Matrices*, Academic Press, New York, 1972.
- [5] Alladi Ramakrishnan and R. Jagannathan, *Topics in Numerical Analysis - II*, Ed. J. H. Miller, Academic Press, New York, p. 133, 1976.
- [6] K. Morinaga and T. Nono, *J. Sci. - Hiroshima Univ. A* **16** (1952).
- [7] K. Yamazaki, *J. Fac. Sci. - Univ. Tokyo, Sec.1*, **10** (1964) 147.
- [8] I. Popovici and C. Gheorghe, *C. R. Acad. Sci. (Paris)*, A **262** (1966) 682.
- [9] A. O. Morris, *Quart. J. Math.* **18** (1967) 7.
- [10] A. O. Morris, *J. Lond. Math. Soc. (2)*, **7** (1973) 235.
- [11] Alladi Ramakrishnan (Ed.), *Proc. MATSCIENCE Conf. Clifford algebra, its generalizations, and applications*, MATSCIENCE, 1971.
- [12] N. B. Backhouse and C. J. Bradley, *Proc. Am. Math. Soc.*, **36** (1972) 260.
- [13] R. Jagannathan, *Springer Lecture Notes in Mathematics* **1122** (1984) (Proc. 4th MATSCIENCE Conf. on Number Theory), Ed. K. Alladi, p. 130.
- [14] W. K. Clifford, *Am. J. Math. Pure Appl.*, **1** (1878) 350; see *Mathematical Papers by W. K. Clifford*, Ed. H. Robert Tucker, Chelsea, 1968.
- [15] P. A. M. Dirac, *Proc. Roy. Soc., A* **177** (1928) 610.
- [16] D. Hestenes, *Space-Time Algebra*, Gordon & Breach, London, 1966.
- [17] J. S. R. Chisholm and A. K. Common (Eds.), *Clifford Algebras and their Applications in Mathematical Physics*, D. Reidel, 1986.
- [18] C. Doran and A. Lasenby, *Geometric Algebra for Physicists*, Cambridge University Press, Cambridge, 2003.
- [19] Alladi Ramakrishnan and I. V. V. Raghavacharyulu, *Symposia in Theoretical Physics and Mathematics*, **8**, Plenum Press, 1968, Alladi Ramakrishnan (Ed.), p. 25.
- [20] P. A. M. Dirac, *Proc. Roy. Soc., A* **322** (1971) 435.
- [21] R. Jagannathan and H. N. V. Dutt, *J. Math. Phys. Sci. (IIT-Madras)* **IX** (1975) 301.
- [22] H. Weyl, *The Theory of Groups and Quantum Mechanics*, Dover, NY, USA (1950).
- [23] R. Jagannathan and N. R. Ranganathan, *Rep. Math. Phys.* **5** (1974) 131.
- [24] R. Jagannathan and N. R. Ranganathan, *Rep. Math. Phys.* **7** (1975) 229.
- [25] R. Jagannathan and T. S. Santhanam, *Springer Lecture Notes in Mathematics* **938**, (1982) (Proc. 3rd MATSCIENCE Conf. on Number Theory), Ed. K. Alladi, p. 82.
- [26] V. P. Nair, *Quantum Field Theory - A Modern Perspective*, Springer, Berlin (2005).
- [27] A. K. Kwasniewski, *J. Phys. A: Math. Gen.* **19** (1986) 1469.
- [28] A. K. Kwasniewski, W. Bajguz, and I. Jaroszewski, *Adv. Appl. Clifford Algebras* **8** (1998) 417.
- [29] J. Schwinger, *Quantum Kinematics and Dynamics*, Benjamin, New York (1970).
- [30] S. Chaturvedi, E. Ercolessi, G. Marmo, G. Morandi, N. Mukunda, and R. Simon, *J. Phys. A: Math. Gen.* **39** (2006) 1405.
- [31] R. Jagannathan, *MATSCIENCE Rep.* **87** (1977) 6.
- [32] M. Moshinsky and C. Quesne, *J. Math. Phys.* **12** (1971) 1772.
- [33] N. R. Ranganathan and R. Jagannathan, *Proc. 2nd International Colloquium on Group Theoretical Methods in Physics*, Univ. Nijmegen, (1973) p. B232.
- [34] R. Jagannathan and N. R. Ranganathan, *Phys. Stat. Sol. B* **74** (1976) 74.
- [35] T. S. Santhanam and A. R. Tekumalla, *Found. Phys.* **6** (1976) 583.
- [36] T. S. Santhanam, *Phys. Lett. A* **56** (1976) 345.
- [37] T. S. Santhanam, *Found. Phys.* **7** (1977) 121.
- [38] T. S. Santhanam, *Nuovo Cim. Lett.* **20** (1977) 13.

- [39] T. S. Santhanam, *Uncertainty Principle and Foundations of Quantum Mechanics*, Eds. W. Price and S. S. Chissick, John Wiley, (1977) p. 227.
- [40] T. S. Santhanam and K. B. Sinha, *Aust. J. Phys.* **31** (1978) 233.
- [41] R. Tanas, A. Miranowicz, and Ts. Gantsog, *Progress in Optics XXXV* (1996) 355.
- [42] R. Jagannathan, T. S. Santhanam, and R. Vasudevan, *Int. J. Theor. Phys.* **20** (1981) 755.
- [43] R. Jagannathan and T. S. Santhanam, *Int. J. Theor. Phys.* **21** (1982) 351.
- [44] R. Jagannathan, *Int. J. Theor. Phys.* **22** (1983) 1105.
- [45] J. Y. Bang and M. S. Berger, *Phys. Rev. A* **80** (2009) 022105.
- [46] E. G. Floratos, *Phys. Lett.* **B233** (1989) 395.
- [47] J. Weyers, *Phys. Lett.* **B240** (1990) 396.
- [48] R. Chakrabarti and R. Jagannathan, *J. Phys. A:Math. Gen.* **24** (1991) 1709.
- [49] R. Chakrabarti and R. Jagannathan, *J. Phys. A:Math. Gen.* **24** (1991) 5683.
- [50] Alladi Ramakrishnan, in *The Structure of Matter - Rutherford Centennial Symposium*, 1971, University of Canterbury, Christchurch, New Zealand, Ed. B. G. Wybourne, 1972.

# $(p, q)$ -Rogers-Szegö Polynomial and the $(p, q)$ -Oscillator

Ramaswamy Jagannathan and Raghavendra Sridhar

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** A  $(p, q)$ -analog of the classical Rogers-Szegö polynomial is defined by replacing the  $q$ -binomial coefficient in it by the  $(p, q)$ -binomial coefficient corresponding to the definition of  $(p, q)$ -number as  $[n]_{p,q} = (p^n - q^n)/(p - q)$ . Exactly like the Rogers-Szegö polynomial is associated with the  $q$ -oscillator algebra, the  $(p, q)$ -Rogers-Szegö polynomial is found to be associated with the  $(p, q)$ -oscillator algebra.

**Mathematics Subject Classification (2010)** 33D45, 33D80, 33D90

**Key words and phrases**  $q$ -hypergeometric series ·  $(p, q)$ -hypergeometric series ·  $q$ -special functions ·  $(p, q)$ -special functions ·  $q$ -binomial coefficients · Rogers-Szegö polynomial ·  $(p, q)$ -binomial coefficients ·  $(p, q)$ -Rogers-Szegö polynomial · quantum groups ·  $q$ -oscillator ·  $(p, q)$ -oscillator ·  $(p, q)$ -Steiltjes-Wigert polynomial · continuous  $(p, q)$ -Hermite polynomial

## 1 Introduction

The  $q$ -oscillator algebra plays a central role in the physical applications of quantum groups (for a review of quantum groups and their applications, see, e.g., [1–3]). It was used [4–7] to extend the Jordan-Schwinger realization of the  $sl(2)$  algebra

---

R. Jagannathan

Chennai Mathematical Institute, Plot H1, SIPCOT IT Park, Padur P.O. Siruseri 603103, Tamilnadu, India

Formerly of MATSCIENCE, The Institute of Mathematical Sciences, Chennai

e-mail: [jagan@cmi.ac.in](mailto:jagan@cmi.ac.in); [jagan@imsc.res.in](mailto:jagan@imsc.res.in)

R. Sridhar

30/1, Sundar Enclave, Valmiki Street, Thiruvanniyur Chennai 600041, Tamilnadu, India

Formerly of MATSCIENCE, The Institute of Mathematical Sciences, Chennai

e-mail: [sridhar@imsc.res.in](mailto:sridhar@imsc.res.in)

in terms of harmonic oscillators to the  $q$ -analogue of the universal enveloping algebra of  $sl(2)$ , namely,  $U_q(sl(2))$ . In order to extend this  $q$ -oscillator realization of  $U_q(sl(2))$  to the two-parameter quantum algebra  $U_{p,q}(gl(2))$ , the  $(p, q)$ -oscillator algebra was introduced in [8] (see also [9, 10]).

Heine's  $q$ -number, or the basic number,

$$[n]_q = \frac{1 - q^n}{1 - q}, \quad (1.1)$$

is well known in the mathematics literature. The  $(p, q)$ -oscillator necessitated the introduction of the  $(p, q)$ -number, or the twin-basic number,

$$[n]_{p,q} = \frac{p^n - q^n}{p - q}, \quad (1.2)$$

a natural generalization of the  $q$ -number, such that

$$\lim_{p \rightarrow 1} [n]_{p,q} \longrightarrow [n]_q. \quad (1.3)$$

With the introduction of this  $(p, q)$ -number, the essential elements of the  $(p, q)$ -calculus, namely,  $(p, q)$ -differentiation,  $(p, q)$ -integration, and the  $(p, q)$ -exponential, were also studied in [8]. This led to a more detailed study of  $(p, q)$ -hypergeometric series and  $(p, q)$ -special functions [11–13]. Meanwhile, the  $(p, q)$ -binomial coefficients,  $(p, q)$ -Stirling numbers, and the  $(p, q)$ -binomial theorem for noncommutative operators were studied [14, 15] in the analysis of certain physical problems. Interestingly, in the same year 1991 the  $(p, q)$ -number was introduced in the mathematics literature in connection with set partition statistics [16]. A very general formalism of deformed hypergeometric functions has been developed in [17]. Some applications of  $(p, q)$ -hypergeometric series in the context of two-parameter quantum groups can be found in [18, 19].

It is noted in [4] that the classical Rogers-Szegő polynomials provide a basis for a coordinate representation of the  $q$ -oscillator. Several aspects of this close connection between the  $q$ -oscillator algebra and the Rogers-Szegő polynomials, and the related continuous  $q$ -Hermite polynomials, have been analyzed in detail later (see [20–24]). In this paper, after a brief review of the known connection between the Rogers-Szegő polynomial and the  $q$ -oscillator, we shall define a  $(p, q)$ -Rogers-Szegő polynomial and show that it is connected with the  $(p, q)$ -oscillator.

As explained below in Sect. 4, it is not possible to rewrite a  $(p, q)$ -hypergeometric series, or a  $(p, q)$ -analog of a  $q$ -function, as a regular  $q$ -hypergeometric series or a  $q$ -function routinely by rescaling the independent variable. Particularly, this is not possible in the case of the  $(p, q)$ -Rogers-Szegő polynomial considered here.

## 2 Harmonic Oscillator

The harmonic oscillator is associated with the creation (or raising) operator  $\hat{a}_+$ , the annihilation (or lowering) operator  $\hat{a}_-$ , and the number operator  $\hat{n}$  satisfying the algebra

$$[\hat{n}, \hat{a}_+] = \hat{a}_+, \quad [\hat{n}, \hat{a}_-] = -\hat{a}_-, \quad [\hat{a}_-, \hat{a}_+] = 1, \tag{2.1}$$

where  $[\hat{A}, \hat{B}]$  stands for the commutator  $\hat{A}\hat{B} - \hat{B}\hat{A}$ . Note that

$$\hat{n} = \hat{a}_+\hat{a}_-. \tag{2.2}$$

Let

$$h_n(x) = (1+x)^n, \quad \psi_n(x) = \frac{1}{\sqrt{n!}}h_n(x). \tag{2.3}$$

It follows that

$$\frac{d}{dx}\psi_n(x) = \sqrt{n}\psi_{n-1}(x), \tag{2.4}$$

$$(1+x)\psi_n(x) = \sqrt{n+1}\psi_{n+1}(x), \tag{2.5}$$

$$(1+x)\frac{d}{dx}\psi_n(x) = n\psi_n(x), \tag{2.6}$$

$$\frac{d}{dx}((1+x)\psi_n(x)) = (n+1)\psi_n(x). \tag{2.7}$$

Thus, it is clear that the set  $\{\psi_n(x)|n = 0, 1, 2, \dots\}$  forms a basis for the following Bargmann-Fock realization of the oscillator algebra (2.1):

$$\hat{a}_+ = (1+x), \quad \hat{a}_- = \frac{d}{dx}, \quad \hat{n} = (1+x)\frac{d}{dx}. \tag{2.8}$$

It is to be noted that (2.5) and (2.6) are, respectively, the recurrence relation and the differential equation for  $\psi_n(x)$ .

Let  $\{\hat{a}_-^{(1)}, \hat{a}_+^{(1)}, \hat{n}^{(1)}\}$  and  $\{\hat{a}_-^{(2)}, \hat{a}_+^{(2)}, \hat{n}^{(2)}\}$  be two sets of oscillator operators each satisfying the algebra (2.1) and commuting with each other. Then, for the generators of  $sl(2)$  satisfying the Lie algebra,

$$[x_0, x_+] = x_+, \quad [x_0, x_-] = -x_-, \quad [x_-, x_+] = 2x_0, \tag{2.9}$$

one has the Jordan-Schwinger realization

$$x_+ = \hat{a}_+^{(1)}\hat{a}_-^{(2)}, \quad x_- = \hat{a}_+^{(2)}\hat{a}_-^{(1)}, \quad x_0 = \frac{1}{2}(\hat{n}^{(1)} - \hat{n}^{(2)}). \tag{2.10}$$

### 3 $q$ -Oscillator and the Rogers-Szegö Polynomial

When  $U(sl(2))$ , the universal enveloping algebra of  $sl(2)$ , is  $q$ -deformed, the resulting  $U_q(sl(2))$  has generators  $\{X_-, X_+, X_0\}$  satisfying the algebra

$$\begin{aligned} [X_0, X_+] &= X_+, & [X_0, X_-] &= -X_-, \\ X_- + X_- - q^{-1} X_- X_+ &= \frac{1 - q^{2X_0}}{1 - q} = [2X_0]_q. \end{aligned} \tag{3.1}$$

The  $q$ -oscillator is associated with the annihilation operator  $\hat{A}_-$ , creation operator  $\hat{A}_+$ , and the number operator  $\hat{N}$  satisfying the algebra

$$[\hat{N}, \hat{A}_-] = -\hat{A}_-, \quad [\hat{N}, \hat{A}_+] = \hat{A}_+, \quad \hat{A}_- \hat{A}_+ - q \hat{A}_+ \hat{A}_- = 1. \tag{3.2}$$

It should be noted that in this case  $\hat{N} \neq \hat{A}_+ \hat{A}_-$ . Instead, we have

$$\hat{A}_+ \hat{A}_- = \frac{1 - q^{\hat{N}}}{1 - q} = [\hat{N}]_q, \tag{3.3}$$

and

$$\hat{A}_- \hat{A}_+ = \frac{1 - q^{\hat{N}+1}}{1 - q} = [\hat{N} + 1]_q, \tag{3.4}$$

Now, let  $\{\hat{A}_-^{(1)}, \hat{A}_+^{(1)}, \hat{N}^{(1)}\}$  and  $\{\hat{A}_-^{(2)}, \hat{A}_+^{(2)}, \hat{N}^{(2)}\}$  be two sets of  $q$ -oscillator operators each satisfying the algebra (3.2) and commuting with each other. Then, taking

$$X_+ = \hat{A}_+^{(1)} q^{-\hat{N}^{(2)}/2} \hat{A}_-^{(2)}, \quad X_- = \hat{A}_+^{(2)} q^{-\hat{N}^{(2)}/2} \hat{A}_-^{(1)}, \quad X_0 = \frac{1}{2} (\hat{N}^{(1)} - \hat{N}^{(2)}), \tag{3.5}$$

we get a Jordan-Schwinger-type realization of the  $U_q(sl(2))$  (3.1).

Now, we have to recall some definitions from the theory of  $q$ -series. The  $q$ -shifted factorial is defined as

$$(a; q)_n = \begin{cases} 1, & \text{for } n = 0, \\ \prod_{k=0}^{n-1} (1 - aq^k), & \text{for } n = 1, 2, \dots \end{cases} \tag{3.6}$$

The  $q$ -binomial coefficient is defined by

$$\begin{bmatrix} n \\ k \end{bmatrix}_q = \frac{(q; q)_n}{(q; q)_k (q; q)_{n-k}}. \tag{3.7}$$

For more details on q-series, see [25]. With the definition

$$[0]_q! = 1, \quad [n]_q! = [n]_q[n - 1]_q \dots [2]_q[1]_q, \quad \text{for } n = 1, 2, \dots, \tag{3.8}$$

we have

$$\begin{bmatrix} n \\ k \end{bmatrix}_q = \frac{[n]_q!}{[k]_q![n - k]_q!} \tag{3.9}$$

and

$$\lim_{q \rightarrow 1} \begin{bmatrix} n \\ k \end{bmatrix}_q = \frac{n!}{k!(n - k)!} = \binom{n}{k}. \tag{3.10}$$

The Rogers-Szegő polynomial is defined as

$$H_n(x; q) = \sum_{k=0}^n \begin{bmatrix} n \\ k \end{bmatrix}_q x^k. \tag{3.11}$$

This can be naturally expected to be related to the basis of a realization of the q-oscillator since

$$\lim_{q \rightarrow 1} H_n(x; q) = h_n(x), \tag{3.12}$$

and the q-oscillator becomes the ordinary oscillator in the limit  $q \rightarrow 1$ .

To exhibit the relation between  $H_n(x; q)$  and the q-oscillator, we shall closely follow [22], although our treatment is slightly different. Let us define

$$\psi_n(x; q) = \frac{1}{\sqrt{[n]_q!}} H_n(x; q) = \frac{1}{\sqrt{[n]_q!}} \sum_{k=0}^n \begin{bmatrix} n \\ k \end{bmatrix}_q x^k. \tag{3.13}$$

The Jackson q-difference operator is defined by

$$\hat{D}_q f(x) = \frac{f(x) - f(qx)}{(1 - q)x}. \tag{3.14}$$

It is straightforward to see that

$$\hat{D}_q \psi_n(x; q) = \sqrt{[n]_q} \psi_{n-1}(x; q). \tag{3.15}$$

The q-binomial coefficients obey the recurrence relation

$$\begin{bmatrix} n + 1 \\ k \end{bmatrix}_q = \begin{bmatrix} n \\ k \end{bmatrix}_q + \begin{bmatrix} n \\ k - 1 \end{bmatrix}_q - (1 - q^n) \begin{bmatrix} n - 1 \\ k - 1 \end{bmatrix}_q. \tag{3.16}$$

From this it follows that  $\psi_n(x; q)$  satisfies the recurrence relation

$$\sqrt{[n + 1]_q} \psi_{n+1}(x; q) = (1 + x) \psi_n(x; q) - x(1 - q) \sqrt{[n]_q} \psi_{n-1}(x; q). \tag{3.17}$$



Using (3.15) we can write this relation as

$$\left[ (1+x) - (1-q)x\hat{D}_q \right] \psi_n(x; q) = \sqrt{[n+1]_q} \psi_{n+1}(x; q). \tag{3.18}$$

Thus, it is seen from (3.15) and (3.18) that the set of polynomials  $\{\psi_n(x; q) | n = 0, 1, 2, \dots\}$  provides a basis for a realization of the  $q$ -oscillator algebra (3.2) as follows. Let us define the number operator  $\hat{N}$  formally as

$$\hat{N}\psi_n(x; q) = n\psi_n(x; q). \tag{3.19}$$

Note that

$$\hat{D}_q^{n+1}\psi_n(x; q) = 0, \tag{3.20}$$

and  $\hat{D}_q^m\psi_n(x; q) \neq 0$  for any  $m < n + 1$ . Then,

$$\hat{A}_-\psi_n(x; q) = \hat{D}_q\psi_n(x; q) = \sqrt{[n]_q} \psi_{n-1}(x; q), \tag{3.21}$$

$$\begin{aligned} \hat{A}_+\psi_n(x; q) &= \left[ (1+x) - (1-q)x\hat{D}_q \right] \psi_n(x; q) \\ &= \sqrt{[n+1]_q} \psi_{n+1}(x; q), \end{aligned} \tag{3.22}$$

$$\hat{A}_+\hat{A}_-\psi_n(x; q) = [n]_q\psi_n(x; q) = [\hat{N}]_q\psi_n(x; q), \tag{3.23}$$

$$\hat{A}_-\hat{A}_+\psi_n(x; q) = [n+1]_q\psi_n(x; q) = [\hat{N} + 1]_q\psi_n(x; q). \tag{3.24}$$

From this one can easily verify that the relations in (3.2) are satisfied by  $\{\hat{A}_+, \hat{A}_-, \hat{N}\}$ . It may be noted that these relations (3.21)–(3.24) are the  $q$ -generalizations of the harmonic oscillator relations (2.4)–(2.7) to which they reduce in the limit  $q \rightarrow 1$ . Substituting the explicit expressions for  $\hat{A}_+$  and  $\hat{A}_-$  in (3.23), we get the  $q$ -differential equation for  $\psi_n(x; q)$  (or  $H_n(x; q)$ ; see [22]):

$$\left( (1-q)x\hat{D}_q^2 - (1+x)\hat{D}_q + [n]_q \right) \psi_n(x; q) = 0, \tag{3.25}$$

which reduces to (2.6) in the limit  $q \rightarrow 1$ .

### 4 $(p, q)$ -Oscillator and the $(p, q)$ -Rogers-Szegö Polynomial

A genuine two-parameter quantum deformation exists only for  $U(gl(2))$  and not for  $U(sl(2))$ . The two-parameter deformation of  $U(gl(2))$  leads to  $U_{p,q}(gl(2))$  which is generated by  $\{\hat{\mathcal{X}}_0, \hat{\mathcal{X}}_+, \hat{\mathcal{X}}_-\}$  satisfying the commutation relations

$$\begin{aligned} [\hat{\mathcal{X}}_0, \hat{\mathcal{X}}_+] &= \hat{\mathcal{X}}_+, & [\hat{\mathcal{X}}_0, \hat{\mathcal{X}}_-] &= -\hat{\mathcal{X}}_-, \\ \hat{\mathcal{X}}_+\hat{\mathcal{X}}_- - (pq)^{-1}\hat{\mathcal{X}}_-\hat{\mathcal{X}}_+ &= \frac{p^{2\hat{\mathcal{X}}_0} - q^{2\hat{\mathcal{X}}_0}}{p - q} = [2\hat{\mathcal{X}}_0]_{p,q}, \end{aligned} \tag{4.1}$$

and a central element  $\hat{Z}$ , which we shall ignore for the present purpose. Here,  $[ \ ]_{p,q}$  is as defined in (1.2).

To get an oscillator realization of the algebra (4.1), we need the (p, q)-oscillator algebra defined by

$$[\hat{N}, \hat{A}_+] = \hat{A}_+, \quad [\hat{N}, \hat{A}_-] = -\hat{A}_-, \quad \hat{A}_- \hat{A}_+ - q \hat{A}_+ \hat{A}_- = p^{\hat{N}}, \quad (4.2)$$

where  $\{\hat{A}_+, \hat{A}_-, \hat{N}\}$  are, respectively, the creation, annihilation, and number operators. In this case,

$$\hat{A}_+ \hat{A}_- = \frac{p^{\hat{N}} - q^{\hat{N}}}{p - q} = [\hat{N}]_{p,q}, \quad \hat{A}_- \hat{A}_+ = \frac{p^{\hat{N}+1} - q^{\hat{N}+1}}{p - q} = [\hat{N} + 1]_{p,q}. \quad (4.3)$$

Note the symmetry of this relation under the exchange of p and q. So, the last relation in (4.2) can also be taken, equivalently, as

$$\hat{A}_- \hat{A}_+ - p \hat{A}_+ \hat{A}_- = q^{\hat{N}}. \quad (4.4)$$

The (p, q)-oscillator unifies several special cases of q-oscillators, including the q-fermion oscillator [26]. Now, we shall show that a (p, q)-deformation of the Rogers-Szegö polynomial can be used for a realization of the (p, q)-oscillator algebra exactly in the same way as the classical Rogers-Szegö polynomial is used for a realization of the q-oscillator algebra as seen above. To this end we proceed as follows.

First, let us recall some essential elements of the (p, q)-series (for more details, see [12, 13]). The (p, q)-shifted factorial is defined by

$$(a, b; p, q)_n = \begin{cases} 1, & \text{for } n = 0, \\ \prod_{k=0}^{n-1} (ap^k - bq^k), & \text{for } n = 1, 2, \dots \end{cases} \quad (4.5)$$

Note that

$$(a, b; p, q)_n = a^n p^{n(n-1)/2} (b/a; q/p)_n. \quad (4.6)$$

In view of this, it is not possible to rewrite a (p, q)-hypergeometric series, or a (p, q)-analog of a q-function, routinely as a q/p-hypergeometric series or a q/p-function, with the same or a rescaled independent variable, unless the factors depending on a and p in the numerator and the denominator cancel in each term, or are such that the uncanceled factor in each term is of the same power as the independent variable. The (p, q)-binomial coefficient is defined by

$$\begin{bmatrix} n \\ k \end{bmatrix}_{p,q} = \frac{(p, q; p, q)_n}{(p, q; p, q)_k (p, q; p, q)_{n-k}}. \quad (4.7)$$

With the definition

$$[0]_{p,q}! = 1, \quad [n]_{p,q}! = [n]_{p,q}[n-1]_{p,q} \dots [2]_{p,q}[1]_{p,q}, \quad \text{for } n = 1, 2, \dots, \tag{4.8}$$

we have

$$\begin{bmatrix} n \\ k \end{bmatrix}_{p,q} = \frac{[n]_{p,q}!}{[k]_{p,q}![n-k]_{p,q}!} \tag{4.9}$$

and

$$\lim_{p \rightarrow 1} \begin{bmatrix} n \\ k \end{bmatrix}_{p,q} = \begin{bmatrix} n \\ k \end{bmatrix}_q. \tag{4.10}$$

Let us now define the  $(p, q)$ -Rogers-Szegö polynomial as

$$H_n(x; p, q) = \sum_{k=0}^n \begin{bmatrix} n \\ k \end{bmatrix}_{p,q} x^k, \tag{4.11}$$

and take

$$\psi_n(x; p, q) = \frac{1}{\sqrt{[n]_{p,q}!}} H_n(x; p, q) = \frac{1}{\sqrt{[n]_{p,q}!}} \sum_{k=0}^n \begin{bmatrix} n \\ k \end{bmatrix}_{p,q} x^k. \tag{4.12}$$

Note that

$$\begin{bmatrix} n \\ k \end{bmatrix}_{p,q} = p^{k(n-k)} \begin{bmatrix} n \\ k \end{bmatrix}_{q/p}, \tag{4.13}$$

and the presence of the factor  $p^{-k^2}$  makes it impossible to rescale  $x$  in any way and hence rewrite the  $(p, q)$ -Rogers-Szegö polynomial  $H_n(x; p, q)$  as a regular Rogers-Szegö polynomial (3.11). Recalling the definition of the  $(p, q)$ -difference operator [8],

$$\hat{D}_{p,q} f(x) = \frac{f(px) - f(qx)}{(p-q)x}, \tag{4.14}$$

it is seen that

$$\hat{D}_{p,q} \psi_n(x; p, q) = \sqrt{[n]_{p,q}} \psi_{n-1}(x; p, q). \tag{4.15}$$

The  $(p, q)$ -analog of (3.16) is given by

$$\begin{bmatrix} n+1 \\ k \end{bmatrix}_{p,q} = p^k \begin{bmatrix} n \\ k \end{bmatrix}_{p,q} + p^{n-k+1} \begin{bmatrix} n \\ k-1 \end{bmatrix}_{p,q} - (p^n - q^n) \begin{bmatrix} n-1 \\ k-1 \end{bmatrix}_{p,q}. \tag{4.16}$$

For a detailed study of the  $(p, q)$ -binomial coefficients, see [27]. This identity (4.16) leads to the following recurrence relation for  $\psi_n(x; p, q)$ :

$$\begin{aligned} \sqrt{[n+1]_{p,q}} \psi_{n+1}(x; p, q) &= \psi_n(px; p, q) + xp^n \psi_n(p^{-1}x; p, q) \\ &\quad - x(p-q) \sqrt{[n]_{p,q}} \psi_{n-1}(x; p, q). \end{aligned} \tag{4.17}$$

To obtain a realization of the (p, q)-oscillator algebra in the basis provided by {ψ<sub>n</sub>(x; p, q) | n = 0, 1, 2, ...}, let us proceed as follows. As before, define the number operator N̂ formally as

$$\hat{N}\psi_n(x; p, q) = n\psi_n(x; p, q). \tag{4.18}$$

Note that

$$\hat{D}_{p,q}^{n+1}\psi_n(x; p, q) = 0, \tag{4.19}$$

and  $\hat{D}_{p,q}^m\psi_n(x; p, q) \neq 0$  for any  $m < n + 1$ . Then, with the scaling operator defined by

$$\hat{\eta}_s f(x) = f(sx), \tag{4.20}$$

it readily follows from (4.15) and (4.17) that we can write

$$\hat{A}_-\psi_n(x; p, q) = \hat{D}_{p,q}\psi_n(x; p, q) = \sqrt{[n]_{p,q}}\psi_{n-1}(x; p, q), \tag{4.21}$$

$$\begin{aligned} \hat{A}_+\psi_n(x; p, q) &= \left(\hat{\eta}_p + x\hat{\eta}_{p-1}p^{\hat{N}} - x(p-q)\hat{D}_{p,q}\right)\psi_n(x; p, q) \\ &= \sqrt{[n+1]_{p,q}}\psi_{n+1}(x; p, q), \end{aligned} \tag{4.22}$$

$$\hat{A}_+\hat{A}_-\psi_n(x; p, q) = [n]_{p,q}\psi_n(x; p, q) = [\hat{N}]_{p,q}\psi_n(x; p, q), \tag{4.23}$$

$$\hat{A}_-\hat{A}_+\psi_n(x; p, q) = [n+1]_{p,q}\psi_n(x; p, q) = [\hat{N} + 1]_{p,q}\psi_n(x; p, q), \tag{4.24}$$

which generalize the corresponding results (3.21)–(3.24) for the q-oscillator; when  $p \rightarrow 1$ , (4.21)–(4.24) reduce to (3.21)–(3.24). It is straightforward to verify that the realizations of { $\hat{A}_-$ ,  $\hat{A}_+$ ,  $\hat{N}$ } in (4.18), (4.21), and (4.22) satisfy the required relations of the (p, q)-oscillator algebra (4.2). Using (4.21) and (4.22) in (4.23), we get the (p, q)-differential equation satisfied by  $\psi_n(x; p, q)$  as

$$\left[(p-q)x\hat{D}_{p,q}^2 - (\hat{\eta}_p + p^{n-1}x\hat{\eta}_{p-1})\hat{D}_{p,q} + [n]_{p,q}\right]\psi_n(x; p, q) = 0, \tag{4.25}$$

which reduces to (3.25) in the limit  $p \rightarrow 1$ .

### 5 Conclusion

Let us conclude with a few remarks.

By choosing different values for p and q, one can study different special cases of  $\psi_n(x; p, q)$ . This is particularly important since there exists many versions of q-oscillators that are special cases of the (p, q)-oscillator. For example, the q-oscillator originally used in connection with  $U_q(su(2))$  [4–7], and more popular

in the physics literature, corresponds to the choice  $p = q^{-1}$ . From the above, it is clear that this oscillator can be realized through the polynomials

$$H_n(x; q^{-1}, q) = \sum_{k=0}^n \begin{bmatrix} n \\ k \end{bmatrix}_{q^2} q^{-k(n-k)} x^k. \tag{5.1}$$

It may be emphasized again that although  $H_n(x; q^{-1}, q)$  is a function with only a single  $q$ -parameter, it cannot be rewritten as a regular Rogers-Szegö polynomial.

In [22], the raising and lowering operators for the Steiltjes-Wigert polynomial have been obtained using the fact that this polynomial is just the Rogers-Szegö polynomial with  $q$  replaced by  $q^{-1}$  and it has been shown that these raising and lowering operators of the Steiltjes-Wigert polynomial provide a realization of the single-parameter deformed oscillator with  $q$  replaced by  $q^{-1}$ . Now, it is clear that one can study the  $(p, q)$ -Steiltjes-Wigert polynomials similarly by replacing  $p$  and  $q$ , respectively, by  $p^{-1}$  and  $q^{-1}$  in the above formalism. Thus, the  $(p, q)$ -Steiltjes-Wigert polynomial is given by

$$G_n(x; p, q) = H_n(x; p^{-1}, q^{-1}) = \sum_{k=0}^n \begin{bmatrix} n \\ k \end{bmatrix}_{p,q} (pq)^{-k(n-k)} x^k, \tag{5.2}$$

which becomes the usual Steiltjes-Wigert polynomial in the limit  $p \rightarrow 1$ .

The continuous  $q$ -Hermite polynomial is defined as

$$H_n(\cos \theta | q) = e^{-in\theta} H_n(e^{2i\theta}; q). \tag{5.3}$$

It is clear that one can define the continuous  $(p, q)$ -Hermite polynomial in an analogous way as

$$H_n(\cos \theta | p, q) = e^{-in\theta} H_n(e^{2i\theta}; p, q). \tag{5.4}$$

This has already been suggested in [13] without any further study. It should be worthwhile to study the  $(p, q)$ -Rogers-Szegö polynomial, the  $(p, q)$ -Steiltjes-Wigert polynomial, and the continuous  $(p, q)$ -Hermite polynomial in detail.

**Acknowledgment** We would like to recall, with gratitude and pride, our long association with Professor Alladi Ramakrishnan who pioneered research in theoretical physics in south India, founded MATSCIENCE, The Institute of Mathematical Sciences, Chennai, and directed it for more than two decades. We are his disciples, and our scientific careers were molded in his school.

## References

1. V. Chari and A. Pressley, *A Guide to Quantum Groups*, Cambridge Univ. Press, Cambridge, 1994.
2. L. C. Biedenharn and M. A. Lohe, *Quantum Group Symmetry and  $q$ -Tensor Algebras*, World Scientific, Singapore, 1995.

3. M. Chaichian and A. Demichev, *Introduction to Quantum Groups*, World Scientific, Singapore, 1996.
4. A. J. Macfarlane, *J. Phys. A : Math. Gen.* **22** (1989) 4581.
5. L. C. Biedenharn, *J. Phys. A : Math. Gen.* **22** (1989) L873.
6. C.-P. Sun and H.-C. Fu, *J. Phys. A : Math. Gen.* **22** (1989) L983.
7. T. Hayashi, *Commun. Math. Phys.*, **127** (1990) 129.
8. R. Chakrabarti and R. Jagannathan, *J. Phys. A : Math. Gen.* **24** (1991) L711.
9. A. Jannussis, G. Brodimas and L. Mignani, *J. Phys. A : Math. Gen.* **24** (1991) L775.
10. M. Arik, E. Demircan, E. Turgut, L. Ekinçi and M. Mungan, *Z. Phys. C : Particles and Fields*, **55** (1992) 89.
11. I. M. Burbani and A. U. Klimyk, *Integral Transforms and Special Functions* **2** (1994) 15.
12. R. Jagannathan, in *Proc. Workshop on Special Functions and Differential Equations* (The Institute of Mathematical Sciences, Chennai, January 1997), Eds. K. Srinivasa Rao, R. Jagannathan, G. Vanden Berghe and J. Van der Jeugt (Allied Pub. New Delhi, 1998) p.158; *arXiv:math/9803142*.
13. R. Jagannathan and K. Srinivasa Rao, *arXiv:math/0602613* (presented at the Internat. Conf. on Number Theory and Mathematical Physics, 20-21 Dec. 2005, Srinivasa Ramanujan Centre, Kumbakonam, India).
14. J. Katriel and M. Kibler, *J. Phys. A : Math. Gen.* **25** (1992) 2683.
15. Yu. F. Smirnov and R. F. Wehrhahn, *J. Phys. A : Math. Gen.* **25** (1992) 5563.
16. M. Wachs and D. White, *J. Combin. Theory A* **56** (1991) 27.
17. I. M. Gelfand, M. I. Graev and V. S. Retakh, *Russian Math. Surveys* **53** (1998) 1.
18. M. Nishizawa, *J. Comp. Appl. Maths.* **160** (2003) 233.
19. V. Sahai and S. Srivastava, *J. Comp. Appl. Maths.* **160** (2003) 271.
20. J. Van der Jeugt, *Lett. Math. Phys.* **24** (1992) 267.
21. R. Floreanini, J. LeTourneux and L. Vinet *J. Phys. A : Math. Gen.* **28** (1995) L287.
22. D. Galetti, *Brazilian J. Phys.* **33** (2003) 148.
23. E. I. Jafarov, S. Lievens, S. M. Nagiyev and J. Van der Jeugt, *J. Phys. A : Math. Theor.* **40** (2007) 5427.
24. S. Odake and R. Sasaki, *Phys. Lett. B* **663** (2008) 141.
25. G. Gasper and M. Rahman, *Basic Hypergeometric Series*, Cambridge Univ. Press, Cambridge, 2004.
26. R. Parthasarathy and K. S. Viswanathan, *J. Phys. A : Math. Gen.* **24** (1991) 613.
27. R. B. Corcino, *Integers : Electronic J. Comb. Numb. Theor.* **8** (2008) # A29.

# Rethinking Renormalization

**John R. Klauder**

**Summary** As applied to quantum theories, the program of renormalization is successful for ‘renormalizable models’ but fails for ‘nonrenormalizable models’. After some conceptual discussion and analysis, an enhanced program of renormalization is proposed that is designed to bring the ‘nonrenormalizable models’ under control as well. The new principles are developed by studying several, carefully chosen, soluble examples, and include a recognition of a ‘hard-core’ behavior of the interaction and, in special cases, an extremely elementary procedure to remove the source of all divergences. Our discussion provides the background for a recent proposal for a nontrivial quantization of nonrenormalizable scalar quantum field models, which is briefly summarized as well.

**Dedication** It is a pleasure to dedicate this article to the memory of Prof. Alladi Ramakrishnan who, besides his own important contributions to science, played a crucial role in the development of modern scientific research and education in his native India. Besides a number of recent informative discussions during his yearly visits to the University of Florida, the present author had the pleasure much earlier of hosting Prof. Alladi during his visit and lecture at Bell Telephone Laboratories.

**Mathematics Subject Classification (2000)** 81T25, 81T16, 83C47

**Key words and phrases** General quantum field theory · Renormalization  
· Quantum field theory on a lattice

## Introduction

Renormalization has been a very successful paradigm for dealing with an important class of quantum theories. Its basic principles are easily stated: The parameters of a classical theory are different from those of a quantum theory because of additional

---

J.R. Klauder  
Department of Physics and Department of Mathematics, University of Florida,  
Gainesville, FL 32611-8440  
e-mail: [klauder@phys.ufl.edu](mailto:klauder@phys.ufl.edu)

self interaction that arises in a quantum theory. In practical terms, the interacting system is commonly treated as a perturbation of a free system, and the power series in the nonlinear coupling often displays divergent terms that need to be canceled and counterterms of a suitable nature are introduced to do just this. If a finite number of distinct counterterms can be found so that every term in the power series expansion is rendered finite, then the theory is called renormalizable, and many such theories have had highly successful applications and in several cases have led to astonishingly accurate predictions when compared to experimental measurements. This aspect of the program of renormalization is considered to be a resounding success and deservedly so. It is natural of course that a successful program such as renormalization has also been proposed to study a wider class of theories than its proponents originally intended, and this is indeed the case. A certain family of field theories fall into the class of being “nonrenormalizable”, an attribute that asserts that the procedures usually ascribed to the program of renormalization are unsuccessful in dealing with certain model problems. If such examples were confined to esoteric models with no potential application to the real world, it would be permissible to ignore those models that are classified as nonrenormalizable. But that is not the case. The most famous example corresponds to the Einstein gravitational field for which the general consensus is that quantum gravity is perturbatively nonrenormalizable. Since the standard procedures of renormalization have failed for such an important case, there have been proposed elaborate alternative theories that entail additional fields or degrees of freedom that are designed to produce a theory that is term-by-term finite within a perturbation analysis. Superstring theory is one such program, and  $N = 8$  supergravity is another. In so doing, these alternative theories have introduced additional fields, which, thanks to the differing properties of fermions and bosons can lead to cancellations among the old, divergent contributions of the original theory and well chosen, new, divergent contributions from the carefully selected additional fields. This general approach is sufficiently broad that it would seem to cover all possible situations regarding how interactions and auxiliary counterterms can appear and interact with each other. However, there is one important class of models that is in practice not covered by the preceding characterization. Admittedly, it is not obvious where one should look if such an overlooked class of examples is to be found. A clue to the overlooked class emerges if we recall that the traditional procedures of regularization and renormalization entail the implicit assumption that if the perturbative interaction is reduced in strength, say by the usual device of reducing the value of the associated coupling constant, then, in the limit that the coupling constant vanishes and the effect of the interaction is formally eliminated, the resulting theory in the limit of a vanishing coupling constant is identical to the free theory with which one started. Stated otherwise, and perhaps more directly, this is the implicit assumption that the set of interacting theories defined as the set that is produced for all nonzero (typically positive) values of the coupling constant is such that as the coupling constant goes to zero, the limit of that set of interacting theories is the free theory itself, i.e., the interacting theories are *continuously connected to the free theory*. This highly natural, implicit assumption covers a lot of the important cases but it certainly does not cover all possibilities some of which may



have some ultimate physical relevance. It is an important feature of this paper that we focus on these outlier model theories, which are typically nonrenormalizable models.

### ***Overview of the Present Paper***

The features ascribed to the renormalization program are not limited to quantum field theory but also arise in quantum mechanical analogues. As such, one can gain real insight into the distinction among super renormalizable, strictly renormalizable, and nonrenormalizable models. A common feature of the latter theories is the occurrence of a *hard-core potential*. From a (Euclidean) functional integral viewpoint, the nonlinear interaction acts partially as a hard core projecting out certain paths that would otherwise appear in the free theory. This fact – which we believe is a defining characteristic for a large class of nonrenormalizable interactions – means that an interacting theory is *not* continuously connected to the free theory as the coupling constant is reduced to zero. This property of the quantum theory is also seen in the classical theory itself by the fact that, generally speaking, the set of solutions of the interacting classical theory does not reduce to the set of solutions that characterizes the free solutions. This aspect will be illustrated for particle systems as well as field systems.

The full dynamics of a classical system involves the action functional and its stationary variation to derive the equations of motion. In a (Euclidean) functional integral formalism, the classical action again plays an important role in the quantum dynamics. Regularization is essential in order to give a functional integral meaning, and it is customary to use a lattice approximation for the time for particle mechanics or for spacetime for field models. The lattice action induces a lattice Hamiltonian operator and in turn a lattice ground state for that Hamiltonian. It is natural that a model can be characterized by either the action, the Hamiltonian, or the ground state. It is important to remark that we focus heavily on the ground state in our analysis. When we take up the discussion of field problems, we will present an argument that shows an important role that the ground state plays.

However, before dealing with fields, we wish to illustrate how the issue of renormalization arises in elementary one dimensional examples.

### **One Dimensional Example**

Consider a classical system for a single, phase space, degree of freedom  $(p, q)$  with a classical Hamiltonian given by

$$H_\lambda(p, q) = \frac{1}{2}(p^2 + q^2) + \lambda|q|^{-\alpha}.$$

For any  $\alpha > 0$ , it follows, just from energy considerations, that the motion of the particle can never be such as to reach the origin  $q = 0$  let alone pass through the value  $q = 0$ . This situation holds for all values of the coupling constant  $\lambda > 0$ , and as a consequence, as  $\lambda \rightarrow 0$ , the set of classical solutions of the interacting theory do *not* correspond to the set of classical solutions of the free theory, namely, that of the free harmonic oscillator given by  $q(t) = A \cos(t - a)$ . Specifically, for any choice of the amplitude  $A$  and the phase  $a$  there will be for *every* solution of the free theory a time  $t$  for which the solution vanishes and even crosses the line  $q = 0$ . In contrast, the solutions of the interacting theory for which  $\lambda > 0$ , all pass by continuity to solutions not of the free theory but to those which are rectified in the sense that they are of the form  $q(t) = \pm |A \cos(t - a)|$  and are all strictly different over time from the usual free theories. We give the name *pseudofree* to the name of the theory, different from the free theory, to which the interacting theory is continuously connected as the nonlinear coupling constant goes to zero. Clearly, if one reintroduces the interaction starting from the pseudofree theory, the form of the new solutions is indeed continuously connected to that of the pseudofree theory.

The easiest way to characterize the pseudofree quantum theory is by its Hamiltonian which is the same as that of the free harmonic oscillator augmented by Dirichlet boundary conditions at  $x = 0$ . If one were contemplating a perturbation series representation of the interacting solution, that power series should not be about the free theory (to which the interacting solutions are not continuously connected!) but rather about the pseudofree theory.

Regarding the quantization of such a model, there are some surprises that can arise. For example, when  $0 < \alpha < 1$ , it follows that the interacting quantum solution is in fact continuously connected to the free quantum theory unlike the situation for the classical case. For  $\alpha > 2$ , on the other hand, there is no modification of the theory that can be made to prevent the theory from passing to a pseudofree theory as the parameter  $\lambda \rightarrow 0$ . In other words, for  $\alpha > 2$ , the interacting quantum theory passes to a pseudofree theory with a set of eigenfunctions and eigenvalues that are generally different from those that characterize the free theory. What happens in the interval  $1 \leq \alpha \leq 2$  is quite interesting and to some extent open to different conclusions. With an eye toward maintaining a continuous connection of the interacting theories to the free theory, it is possible to choose a regularized form for the interaction, namely, a set of potentials of the form  $V_\epsilon(q, \lambda)$  that have the property that as  $\epsilon \rightarrow 0$ , the regularized potentials

$$V_\epsilon(q, \lambda) \rightarrow \lambda |q|^{-\alpha}, \quad q \neq 0.$$

These regularized forms of the potential are rather strictly constrained and they involve polynomial contributions in the coupling constant  $\lambda$ . It is not difficult to determine the general form of the regularized potential simply on the basis of dimensional arguments. In particular, the dimensions of the Hamiltonian are those of the first term  $p^2$ , and taking Planck's constant  $\hbar = 1$  for the present time, the

dimensions are that of  $L^{-2}$  where  $L$  denotes the dimension of length. With the regularization parameter  $\epsilon > 0$  entering initially in the interaction as

$$\lambda |q|^{-\alpha} \rightarrow \lambda (|q| + \epsilon)^{-\alpha},$$

it follows that the dimension of  $\epsilon$ , like  $q$ , is  $L$ . In order that the interaction terms have the right dimensions, i.e.,  $L^{-2}$ , it follows that the dimension of  $\lambda$  is that of  $L^{\alpha-2}$ . For regularization terms we restrict ourselves to terms of the form

$$k_j \lambda^j \epsilon^{-p_j} \delta(q),$$

where  $\delta(q)$  is a Dirac delta function. With  $k_j$  chosen as an unknown dimensionless factor, and since  $\delta(q)$  has dimensions  $L^{-1}$ , it follows that the power  $p_j = 1 - (2 - \alpha)j$  in order to ensure that the regularization terms above each have the desired dimension of the Hamiltonian, namely  $L^{-2}$ . Hence the regularized form of the potential is given by

$$V_\epsilon(q, \lambda) = \lambda (|q| + \epsilon)^{-\alpha} - \sum_{j=1}^J k_j \lambda^j \epsilon^{(2-\alpha)j-1} \delta(q).$$

The factor  $J$  denotes the upper limit of the sum which occurs whenever  $(2 - \alpha)^{-1}$  is nonintegral and  $(2 - \alpha)J < 1 < (2 - \alpha)(J + 1)$  for then all further regularization terms vanish as  $\epsilon \rightarrow 0$ . In this case further analysis shows that the factors  $k_j$  are given by  $k_1 = 2/(\alpha - 1)$  and then

$$k_j = -\frac{1}{[1 - j(2 - \alpha)]} \sum_{q=1}^{j-1} k_{j-q} k_q;$$

if instead,  $(2 - \alpha)^{-1} = J$  is an integer, then the last factor  $k_J$  involves a natural logarithm; see [1]. For  $\alpha = 2$ ,  $J = \infty$ , and all  $p_j = 1$ . For all  $\alpha \leq 2$  such a series provides a regularized potential for which the interacting theory is continuously connected to the free theory as  $\lambda \rightarrow 0$ . It is noteworthy that when  $\alpha < 2$  a finite series of counterterms, each with a diminishing divergence (i.e.,  $p_{j+1} < p_j$ ), provides the proper regularized potential, a property similar to that encountered when dealing with super renormalizable quantum field theories. When  $\alpha = 2$  an infinite series of counterterms, all of equal divergence (i.e.,  $p_{j+1} = p_j$ ), leads to a suitable regularized potential, a property similar to that of so-called strictly renormalizable quantum field theories. For  $\alpha > 2$ , on the other hand, there is no regularized potential that leads to an interacting theory that is continuously connected to the free theory. Of course, the proposed regularization terms based simply on dimensionality do not know this fact, and it may be said that they do their best to signal their inability to provide a solution to the problem by the fact that when  $\alpha > 2$ , the term-by-term divergence actually *increases* (i.e.,  $p_{j+1} > p_j$ ), and moreover,  $p_j \rightarrow \infty$  as  $j \rightarrow \infty$ , a property which is reminiscent of the behavior of nonrenormalizable quantum field theories.

## A Brief Summary

We have discussed this simple quantum mechanical model in some detail in order to show what kind of singular behavior is possible even in quantum mechanics. In particular, we observe that for  $\alpha < 1$ , there is no anomalous behavior in the quantum theory although there is anomalous classical behavior. For  $1 \leq \alpha \leq 2$ , it can be arranged that there is no anomalous quantum behavior although there always will be anomalous classical behavior. The price to pay for this good quantum behavior is the introduction of regularized quantum terms that entail a power series in the coupling constant  $\lambda$ . For  $\alpha > 2$ , on the other hand, there is no escaping the anomalous quantum behavior no matter how one tries to regularize the quantum theory.

## Field Theory Analog – A Brief Detour

We claim there is an analog with the above story for quantum mechanics that plays out in quantum field theory as well. For sufficiently weak perturbations, the interaction can be renormalized so that the resultant interacting theory is continuously connected to the free theory as the coupling constant is reduced to zero; this is the situation that applies to super renormalizable and possibly to strictly renormalizable theories. For sufficiently strong perturbations, the interaction cannot be renormalized so that the interacting theory is continuously connected to the free theory. Instead, for such strong perturbations, the interacting theories are connected to an appropriate pseudofree theory. Later, we will bolster the argument that this is the situation which should apply to nonrenormalizable theories. To make this leap of faith from a singular family of classical problems and their associated quantum problems to a wide class of quantum field theories, it will be helpful to develop a primary *principle* that captures the essence of the singular nature of the interaction that leads to either a continuous connection with the original free theory or instead leads to a continuous connection with a pseudofree theory.

## Path Integral Formulation

The principle we adopt to describe the appearance of pseudofree theories is that of a **hard-core interaction**. The concept behind this principle is most simply appreciated in a functional integral representation of the associated quantum system. This analysis works for either a real time or an imaginary time functional integral, and for its better mathematical structure, we shall choose the latter form. For the quantum mechanical problem that we have so far been discussing, the associated imaginary time (Euclidean) functional integral is given by

$$\mathcal{N} \int e^{-\int \frac{1}{2} [\dot{x}^2 + x^2] + \lambda V(x) dt} \mathcal{D}x.$$

Although the Brownian-like paths  $x(t)$  that enter this functional integral have a nowhere defined (i.e., divergent) derivative – a feature that is surely unlike the classical theory – it is noteworthy that the distinction between the behavior for  $\alpha < 2$  and  $\alpha > 2$  can nevertheless be won by simple classical arguments. For classical paths consider the following simple inequality

$$|x(t_2) - x(t_1)| = \left| \int_{t_1}^{t_2} \dot{x}(t) dt \right| \leq |t_2 - t_1|^{1/2} \left[ \int_{t_1}^{t_2} \dot{x}^2(t) dt \right]^{1/2}.$$

Assuming a finite value for the kinetic energy, it follows, for some  $K < \infty$ , that

$$|x(t_2) - x(t_1)|^{-\alpha} \geq K |t_2 - t_1|^{-\alpha/2}.$$

Setting  $x(t_2) = 0$ , the location of the singularity, we see that

$$\int |x(t)|^{-\alpha} dt \geq K \int |t|^{-\alpha/2} dt.$$

This inequality implies that for  $\alpha > 2$  the integral over the interaction term diverges, while for  $\alpha < 2$  that is not necessarily the case. When the integral over the interaction diverges, the contribution of that path is projected out (by the factor  $e^{-\infty}$ ) for any positive value of the coupling constant. And as the coupling constant is reduced to zero, the contribution of that path is never restored leading to the exclusion of that path in the definition of the pseudofree theory. For the quantum mechanical problem previously discussed, this means that whenever  $\alpha > 2$ , the contribution of all paths that reach or cross the axis  $x = 0$  are projected out of the functional integral; that is the meaning of the statement that the interaction acts in part like a hard core. Our simple argument involving the inequality derived from classical paths does not have anything to say about what happens for  $\alpha < 2$ , but that does not diminish its importance for the region  $\alpha > 2$ .

Before proceeding, let us restate some important issues that arose in our analysis of the one dimensional quantum problem as discussed above. The model we studied had a clearly defined free theory (with  $\lambda \equiv 0$ ) which is just the usual harmonic oscillator. The free propagator (in imaginary time for convenience) is readily given by the sum

$$\langle x'', T|x', 0 \rangle = \sum_{n=0}^{\infty} h_n(x'') e^{-(n+1/2)T} h_n(x'),$$

where the set of functions  $\{h_n(x)\}_{n=0}^{\infty}$  are the Hermite functions defined by the generating function

$$\exp(-s^2 + 2sx - \frac{1}{2}x^2) = \pi^{1/4} \sum_{n=0}^{\infty} (n!)^{-1/2} (s\sqrt{2})^n h_n(x).$$

In the present case the pseudofree theory (denoted by a prime ') has a propagator defined by the expression

$$\langle x'', T|x', 0 \rangle' = \theta(x'' x') \sum_{n=0}^{\infty} h_n(x'') e^{-(n+1/2)T} [h_n(x') - h_n(-x')],$$

where the function  $\theta(u) = 1$  if  $u > 0$  and  $\theta(u) = 0$  if  $u < 0$ . It is the latter expression that incorporates the hard core, projecting out all those paths in the free harmonic oscillator propagator that reach or cross the value  $x = 0$ . Note well: It is the pseudofree theory to which the interacting theories are continuously connected as the coupling constant is reduced to zero. It is the pseudofree theory around which a meaningful perturbation theory for the singular perturbation can be constructed. From the point of view of a Euclidean functional integral, if one attempted to expand a partially hard core interaction about the free theory, this would lead to a series composed of ever more divergent expressions. Regularization of that series would serve to render those terms finite but it would also falsely imply that the interacting theory was continuously connected to the free theory because the regularized power series would reduce to the free theory when the coupling constant is reduced to zero. This property of the regularized perturbation series is entirely erroneous and misleading.

Moreover, the seed of the discontinuous nature of the perturbation about the free theory is already evident in the classical theory itself. This situation holds because the classical solutions of the interacting theory already do not reduce to the solutions of the classical free theory as  $\lambda \rightarrow 0$ . Instead they pass to the classical solutions of the pseudofree theory as noted above. This result has the important consequence that an indelible imprint of the fact that one could be dealing with a discontinuous perturbation (of the free theory) can be determined from an analysis of the classical interacting theory itself! The nature of such an analysis is not too difficult; it rests on the determination that the set of solutions of the interacting theory for arbitrarily small coupling constant is not equivalent to the set of solutions of the free theory itself.

The criterion that a classical pseudofree theory be different from the classical free theory is necessary for a quantum pseudofree theory to be different from a quantum free theory. However, the one dimensional example with  $0 < \alpha < 1$  demonstrates that such a criterion is not sufficient to ensure that the quantum theory also involves a pseudofree theory different from the free theory.

### Shifting the Singularity from $x = 0$ to $x = c$

Suppose, instead of the singularity being at  $x = 0$ , we moved it to the point  $x = c$ , where without loss of generality we can assume that  $c > 0$ . This means that our basic potential is  $\lambda |x - c|^{-\alpha}$ . We now briefly summarize the main changes that occur. First, the classical story. In this case, the free solution given by  $q(t) = A \cos(t - a)$

may remain unchanged if the overall classical energy is sufficiently small, which occurs when  $|A| \leq c$ . When  $|A| > c$ , two solutions are possible, one of the form  $q(t) = \max[A \cos(t - a), c]$  with the phase  $a$  adjusted so that the classical path continues to obey the equation of motion. The second path is given by  $q(t) = \min[A \cos(t - a), c]$  with the phase again adjusted so that the classical path solves the equation of motion. The quantum theory for this case is such that the pseudofree theory is defined by the harmonic oscillator Hamiltonian augmented by Dirichlet boundary conditions at  $x = c$ . As a consequence, the eigenfunctions and eigenvalues of the free harmonic oscillator are almost never relevant in the construction of the pseudofree Hamiltonian. The same conclusions would be drawn from an analysis of the Euclidean functional integral formulation of the quantum theory. For  $\alpha \leq 2$ , a regularized potential qualitatively similar to that discussed before, should be suitable to define an interaction that is continuously connected to the free theory. For  $\alpha > 2$ , however, no regularized form of the potential leads to interacting theories that are continuously connected to the free theory as the coupling constant passes to zero. Any perturbation analysis of the interacting theory when  $\alpha > 2$  must take place about the pseudofree theory. It is noteworthy in this example that as  $c \rightarrow \infty$  the classical solutions all tend to those of the free theory. It is also true that as  $c \rightarrow \infty$ , the pseudofree quantum theory passes to the free quantum theory.

### A Remark on Higher Dimensional Examples

Although these facts have been illustrated for a comparatively simple one-dimensional classical/quantum model, it is not difficult to imagine analogous situations in higher dimensional mechanical systems that lead to a corresponding behavior. For example, a two-dimensional configuration space may have a singular potential of the form  $\lambda(x^2 + y^2)^{-\alpha}$ . However, this example does *not* lead to a discontinuous perturbation since, although there are Brownian motion paths that pass through the singular point  $x = y = 0$  and which therefore need to be discarded, the set of such paths is only of measure zero. To achieve a discontinuous perturbation, one would need a singularity of co-dimension one such as offered by the potential  $\lambda|(x^2 + y^2) - 1|^{-\alpha}$ , for example. There is a rich set of examples of this sort, but we shall not dwell on them for we are after still bigger game, namely, those that arise for an infinite number of variables!

## Classical and Quantum Field Theory

Until now, we have seen simple models for which the interacting theory is not continuously connected to the free theory as the coupling constant is reduced to zero. In the classical regime, such a situation can be seen by comparing the set of solutions

allowed by the free classical theory with the set of solutions allowed by the pseudo-free classical theory. In those cases where the set of solutions of the pseudofree classical theory is a proper subset of the set of solutions of the free classical theory, we have a genuine situation where the interacting theory has left an indelible imprint on the classical theory as the coupling constant is reduced to zero. When it comes to an analysis of the associated quantum theories, however, the classical results offer only a partial guide. In certain cases, the interacting quantum theory is continuously connected to the free theory, and thus there is no distinct pseudofree quantum theory, even though the classical pseudofree and free theories differ from one another; for example, this is the case for the one dimensional model when  $0 < \alpha < 1$ . In such a case, it is natural that a quantum perturbation series about the free theory would be the proper choice. However, there is still another option, and this is the one to which we wish to draw attention, namely when the pseudofree quantum theory is distinct from the free quantum theory. It is for such situations that the interacting quantum theory is not continuously connected to the free quantum theory as the coupling constant is reduced toward zero. It is in such cases that a perturbation series of the interaction taken about the free theory would be wrong while a perturbation series about the pseudofree theory would be the proper choice; for example, this is the case for the one dimensional models when  $\alpha > 2$ .

### *Focus on the Ground State*

We aim to carry these concepts from one dimensional systems to field theoretic systems. Functional integral formulations entail regularization such as that offered by a lattice.

Consider the spacetime lattice formulation of a general problem phrased as a scalar field theory. Let  $\phi_k$  denote the field value at the lattice point  $k = (k_0, k_1, k_2, \dots, k_s)$ , where  $k_j \in \{0, \pm 1, \pm 2, \dots\} \equiv \mathbb{Z}$ ,  $k_0$  refers to the (future) temporal direction, and the remaining  $k_j$ ,  $1 \leq j \leq s$ , denote the  $s$  spatial directions; for a quantum mechanical problem,  $s = 0$ . Assume that spacetime is replaced by a periodic, hypercubic lattice with  $L$  points on an edge and  $L^s \equiv N'$  lattice points in a spatial slice.

In this section we first wish to argue that moments of expressions of interest in the full spacetime distribution can be bounded by suitable averages of related quantities in the ground state distribution. In particular, let the full spacetime average on a lattice be given by

$$\langle [\sum_{k_0} F(\phi, a)a]^p \rangle \equiv M \int [\sum_{k_0} F(\phi, a)a]^p e^{-I(\phi, a, \hbar)} \Pi_k d\phi_k,$$

where  $I$  is the lattice action,  $\sum_{k_0}$  denotes a summation over the temporal direction  $k_0$  only, and  $F(\phi, a)$  is an expression that depends only on fields  $\phi_k$  at a fixed value of  $k_0$ . For example, one may consider  $F(\phi, a) = \sum'_k \phi_k^4 a^s$  or



$F(\phi, a) = \sum'_{k,l} \Omega_{k,l} \phi_k \phi_l a^{2s}$ , for some  $c$ -number kernel  $\Omega_{k,l}$ , etc., where the primed sum implies summation over a spatial slice at fixed  $k_0$ . It follows that

$$\langle [\sum_{k_0} F(\phi, a) a]^p \rangle = \sum_{k_0, \dots, k_0} a^p \langle F(\phi_1, a) \cdots F(\phi_p, a) \rangle,$$

where each  $\phi_j$  refers to the fields at Euclidean time “ $k_0 = j$ ”. A straightforward inequality shows that

$$|\langle F(\phi_1, a) \cdots F(\phi_p, a) \rangle| \leq |\langle F(\phi_1, a)^p \rangle \cdots \langle F(\phi_p, a)^p \rangle|^{1/p}.$$

Finally, for sufficiently large  $N'(ba^s)$ , we note that

$$\langle F(\phi, a)^p \rangle = \int F(\phi, a)^p \Psi(\phi)^2 \Pi'_k d\phi_k,$$

namely, an average in the ground state distribution. The argument behind the last equation is as follows. Quite generally,

$$\langle F(\phi, a)^p \rangle = M \sum_l \int \langle \phi | l \rangle e^{-E_l T} \langle l | \phi \rangle F(\phi, a)^p \Pi'_k d\phi_k,$$

where we have used the resolution of unity  $1 = \int |\phi\rangle \langle \phi| \Pi'_k d\phi_k$  for states for which  $\hat{\phi}(x) |\phi\rangle = \phi(x) |\phi\rangle$ , as well as the eigenvectors  $|l\rangle$  and eigenvalues  $E_l$  for which  $\mathcal{H}|l\rangle = E_l |l\rangle$ . For asymptotically large  $T$ , it follows that only the (unique) ground state contributes, and the former expression becomes

$$\langle F(\phi, a)^p \rangle = \int F(\phi, a)^p |\langle \phi | 0 \rangle|^2 \Pi'_k d\phi_k,$$

now with  $M = 1$ , which is just the expression given above.

In summary, for a finite, hypercubic lattice with periodic boundary conditions, we have derived an important result: **If the sharp time average of  $[F(\phi, a)]^p$  is finite, then it follows that the spacetime average of  $[\sum_{k_0} F(\phi, a) a]^p$  is also finite.**

## Ultralocal Scalar Quantum Fields

As we have done before, we want to illustrate the existence of a pseudofree quantum field theory distinct from any free quantum field theory by means of a straightforward and soluble example. The example we have in mind is the so-called *ultralocal scalar quantum field theory*. This model has been rigorously solved previously, and its most complete story can be found in Chap. 10 of [1]. We start with a brief summary of this model based on that rigorous, nonperturbative analysis. Later we show how a simple and natural argument arrives at a completely satisfactory solution

as well. The advantage of having this simple, alternative argument is that it can be generalized to realistic, relativistically covariant model quantum field theories.

The classical Hamiltonian for a scalar ultralocal field theory with a quartic non-linear interaction is given by

$$H = \int \left\{ \frac{1}{2} [\pi(t, x)^2 + m_0^2 \phi(t, x)^2] + g_0 \phi(t, x)^4 \right\} d^s x.$$

Here,  $s$  is the number of spatial dimensions which is one less than the number  $n$  of spacetime dimensions,  $s = n - 1$ . Note well the absence of spatial derivatives in this expression. Clearly this is not a relativistic model; rather it is a mathematical model that will teach us a great deal when it is successfully quantized.

Initially, we note that there are many functions  $\phi(t, x)$  such that

$$\int [\dot{\phi}(t, x)^2 + m_0^2 \phi(t, x)^2] dt d^s x < \infty, \quad \int \phi(t, x)^4 dt d^s x = \infty,$$

a fact which implies that there is a classical pseudofree theory distinct from the classical free theory. This is an important preliminary remark as we try to determine the status of the quantum theory.

However, let us first make a few remarks about the classical properties of such models.

### Classical Features

The classical equations of motion for this model are given by

$$\ddot{\phi}(t, x) + m_0^2 \phi(t, x) + 4g_0 \phi(t, x)^3 = 0.$$

Indeed, the variable  $x$  is strictly a spectator variable in this equation, and we can relegate it to a subsidiary role simply by rewriting the equation of motion as

$$\ddot{\phi}_x(t) + m_0^2 \phi_x(t) + 4g_0 \phi_x(t)^3 = 0,$$

which shows the equation of motion is simply that of an independent anharmonic oscillator at each point of space. Its solution is given by  $\phi(t, x) \equiv \phi_x(t)$ , where the latter function is based on the initial data, e.g.,  $\phi(0, x) \equiv \phi_x(0)$  and  $\dot{\phi}(0, x) \equiv \dot{\phi}_x(0)$ , two functions of  $x$  which may be taken to be continuous in  $x$ , but need not be so.

Indeed, thanks to the independence of the solution for distinct  $x$  values, one may readily discretize this model by replacing the spatial continuum by a hypercubic spatial lattice with a lattice spacing  $a$  and  $L$  sites on each edge, which leads to a spatial volume given by  $V' \equiv (La)^s \equiv N' a^s$ . To begin, we may replace the classical Hamiltonian by a lattice regularized version given by

$$H_{reg} = \sum_k' \left\{ \frac{1}{2} [\pi_k(t)^2 + m_0^2 \phi_k(t)^2] + g_0 \phi_k(t)^4 \right\} a^s,$$

where  $k \in \mathbb{Z}^s$ ; this expression is nothing but a Riemann sum approximation to the integral given above, and it will converge to the former with  $x = \lim ka$ , as the lattice spacing  $a$  converges to zero. This regularized Hamiltonian gives rise to the regularized equations of motion

$$\ddot{\phi}_k(t) + m_0^2 \phi_k(t) + 4g_0 \phi_k(t)^3 = 0,$$

and even this set of discrete equations of motion converge to the continuum form of the equation of motion as  $a \rightarrow 0$  and  $ka \rightarrow x$ .

### Free Ultralocal Field Theory

An important limiting case arises when  $g_0 = 0$  which is the free theory given by the free Hamiltonian

$$H_0 = \frac{1}{2} \int [\pi(t, x)^2 + m_0^2 \phi(t, x)^2] d^s x.$$

The associated free equations of motion are given by

$$\ddot{\phi}(t, x) + m_0^2 \phi(t, x) = 0,$$

with a solution given in terms of the initial data  $\phi(0, x) \equiv \phi_x(0)$  and  $\dot{\phi}(0, x) \equiv \dot{\phi}_x(0)$ , by the relation

$$\phi(t, x) = \phi_x(0) \cos(m_0 t) + m_0^{-1} \dot{\phi}_x(0) \sin(m_0 t),$$

along with  $\pi(t, x) = \dot{\phi}(t, x)$ , or specifically by

$$\pi(t, x) = -m_0 \phi_x(0) \sin(m_0 t) + \dot{\phi}_x(0) \cos(m_0 t).$$

The lattice regulated free Hamiltonian and the associated free solution is also easily given by

$$H_0 = \frac{1}{2} \sum_k [\pi_k(t)^2 + m_0^2 \phi_k(t)^2] a^s,$$

as well as

$$\begin{aligned} \phi_k(t) &= \phi_k(0) \cos(m_0 t) + m_0^{-1} \dot{\phi}_k(0) \sin(m_0 t), \\ \pi_k(t) &= -m_0 \phi_k(0) \sin(m_0 t) + \dot{\phi}_k(0) \cos(m_0 t). \end{aligned}$$

The free model is therefore nothing but an infinite number of identical harmonic oscillators all with the same angular frequency  $m_0$ ! Clearly, as  $a \rightarrow 0$  and  $ka \rightarrow x$ , the regularized solutions  $\phi_k(t)$  and  $\pi_k(t)$  converge to the continuum solutions  $\phi(t, x)$  and  $\pi(t, x)$ .

## Quantum Theory – First Look

We start the discussion of the quantum theory with the free theory. We promote the classical field at time  $t = 0$  (and then suppress the time argument) to an operator field  $\phi(x) \rightarrow \hat{\phi}(x)$  as well as promote the classical momentum  $\pi(x) \rightarrow \hat{\pi}(x)$ , subject to the canonical commutation relation (in units where  $\hbar = 1$ )

$$[\hat{\phi}(x), \hat{\pi}(y)] = i\delta(x - y).$$

The free quantum Hamiltonian  $\mathcal{H}_0$  is then written as

$$\mathcal{H}_0 = \frac{1}{2} \int [ : \hat{\pi}(x)^2 + m_0^2 \hat{\phi}(x)^2 : ] d^s x,$$

where, as usual, the notation  $:(\cdot):$  denotes normal ordering (all creation operators to the left of all annihilation operators). We denote by  $|0_0\rangle$  the nondegenerate ground state of  $\mathcal{H}_0$  for which  $\mathcal{H}_0|0_0\rangle = 0$  holds, thanks to the normal ordering which removes the (infinite) zero-point energy.

An important relation that characterizes the ground state eigenstate is the expectation functional

$$E_0(f) \equiv \langle 0_0 | e^{i \int \hat{\phi}(x) f(x) d^s x} | 0_0 \rangle = e^{-(1/4m_0) \int f(x)^2 d^s x}.$$

Indeed, the structure of this functional as the exponential of a local integral of  $f(x)$  is dictated by the fact that the temporal development of the operators at any point  $x$  is ultralocal, i.e., the temporal development at  $x$  is completely independent of the time development at a different spatial point  $x'$ . This behavior carries over to the case of the interacting ultralocal model as well, and one expects that whatever the full Hamiltonian operator  $\mathcal{H}$  is, and whatever the associated ground state  $|0\rangle$  is, for which  $\mathcal{H}|0\rangle = 0$  holds, the ground state expectation functional has the form

$$E(f) = \langle 0 | e^{i \int \hat{\phi}(x) f(x) d^s x} | 0 \rangle = e^{-\int L[f(x)] d^s x},$$

for some suitable choice of the function  $L[u]$ .

A canonical representation for the function  $L[u]$  is readily determined. We focus on those cases that are even functions  $L[-u] = L[u]$ , which are then real and satisfy  $L[0] = 0$  and otherwise  $L[u] \geq 0$ . Let  $f(x) \equiv p \chi_\Delta(x)$ , where  $\chi_\Delta(x) \equiv 1$  if  $x \in \Delta$  and zero otherwise; moreover, as a modest abuse of notation, we also set  $\int \chi_\Delta(x) d^s x = \Delta$  as well. Thus

$$\langle 0 | e^{i \int \hat{\phi}(x) f(x) d^s x} | 0 \rangle = e^{-\Delta L[p]} \equiv \int \cos(p\lambda) d\mu_\Delta(\lambda),$$

where we have made use of the symmetry of  $L[u]$ , and the fact that for each  $\Delta > 0$  we are dealing with a characteristic function (Fourier transform of a probability measure  $\mu_\Delta$ ). Thus,

$$L[p] = \lim_{\Delta \rightarrow 0} \Delta^{-1} \int [1 - \cos(p\lambda)] d\mu_\Delta(\lambda).$$

Based on this expression, and assuming convergence, it is clear that the most general function  $L[u]$  is given by the relation

$$L[u] = au^2 + \int_{\lambda \neq 0} [1 - \cos(u\lambda)] d\sigma(\lambda),$$

where  $a \geq 0$  and  $\sigma(\lambda)$  is a nonnegative measure such that

$$\int_{\lambda \neq 0} [\lambda^2 / (1 + \lambda^2)] d\sigma(\lambda) < \infty.$$

The free model solution obtained above is one for which  $a = 1/(4m_0)$  and  $\sigma = 0$ . Let us assume hereafter that  $a = 0$  and  $\sigma \neq 0$ . Observe that it is possible that

$$\int_{\lambda \neq 0} d\sigma(\lambda) = \infty,$$

and in fact this will be the case for the solutions of interest to us because we insist that the spectrum of the field operator  $\hat{\phi}(x)$  is absolutely continuous, and thus for any  $\Delta > 0$ , it is necessary that

$$\lim_{p \rightarrow \infty} e^{-\Delta L[p]} = 0.$$

For simplicity in what follows, we assume that the measure  $\sigma(\lambda)$  is absolutely continuous, and we respect that assumption by setting

$$d\sigma(\lambda) = c(\lambda)^2 d\lambda,$$

where  $c(\lambda)$  is known as the “model function”. It has been found that the choice of the model function completely characterizes the ultralocal model under consideration, and, importantly, apart from the free model, all nonlinear ultralocal models are described by the situation where  $a = 0$  and the model function  $c(\lambda) > 0$  [1].

## Model Function

To ensure that the model function  $c(\lambda)$  has a suitable singularity at  $\lambda = 0$ , we focus our attention on model functions of the form

$$c(\lambda) = (b)^{1/2} \frac{e^{-y(\lambda)/2}}{|\lambda|^\gamma},$$

where  $y(0) = 0$ ,  $\gamma = 1/2$ , and  $b$  is a positive constant with dimensions  $L^{-s}$ . **[Remark:** Other  $\gamma$  values in the range  $1/2 < \gamma < 3/2$ , which are discussed in [1],

can be obtained by suitable, invertible, changes of variables from the case where  $\gamma = 1/2$ .] As a consequence, it follows that

$$\begin{aligned} E(p) &\equiv \langle 0 | e^{ip} Q | 0 \rangle \\ &= e^{-(b\Delta) \int_{[1-\cos(p\lambda)]} \frac{e^{-\gamma(\lambda)}}{|\lambda|} d\lambda} \\ &\simeq (b\Delta) \int \cos(p\lambda) \frac{e^{-\gamma(\lambda)}}{|\lambda|^{1-2b\Delta}} d\lambda, \end{aligned}$$

where  $Q \equiv \int \hat{\phi}(x) \chi_{\Delta}(x) d^s x$  and the last relation holds when  $0 < b\Delta \ll 1$ . Observe that the prefactor  $b\Delta$  in the last expression is an approximate normalization factor (and an asymptotically correct one!) for the ground state distribution.

This latter form of the expectation function for a single degree of freedom readily extends to an infinite set of such fields, with  $p = \{p_k\}$  now, such that

$$E_{\Delta}(p) = \prod'_k \left[ (b\Delta) \int \cos(p_k \phi_k) \frac{e^{-\gamma(\phi_k)}}{|\phi_k|^{1-2b\Delta}} d\phi_k \right].$$

Let us consider  $\sum_k p_k \chi_{\Delta}(x - ka)$ , where here we have in mind that  $\chi_{\Delta}(x)$  denotes a small hypercubic cell around the origin of area  $\Delta = a^s$ . As  $\Delta = a^s \rightarrow 0$  and  $\sum_k p_k \chi_{\Delta}(x - ka) \rightarrow f(x)$ , it follows that

$$\begin{aligned} \lim_{\Delta \rightarrow 0} E_{\Delta}(p) &= E(f) = \langle 0 | e^{i \int \hat{\phi}(x) f(x) d^s x} | 0 \rangle \\ &= \exp\{-b \int d^s x \int [1 - \cos(f(x)\lambda)] e^{-\gamma(\lambda)} d\lambda / |\lambda|\}. \end{aligned}$$

This last relation allows us to identify the regularized ground state of a general ultralocal theory as given (with  $\hbar$  temporarily restored) by the expression

$$\Psi(\phi) \equiv \prod'_k (b\Delta)^{1/2} \frac{e^{-\gamma(\phi_k, a, \hbar)/2\hbar}}{|\phi_k|^{1/2-b\Delta}} \equiv \prod'_k \Psi_k(\phi_k).$$

Given that this expression represents the ground state, it then follows that the regularized Hamiltonian is given by

$$\begin{aligned} \mathcal{H}_{\Delta} &= \sum'_k \left[ -\frac{1}{2} \hbar^2 \frac{\partial^2}{\partial \phi_k^2} a^{-s} + \frac{1}{2} \hbar^2 \frac{1}{\Psi_k(\phi_k)} \frac{\partial^2 \Psi_k(\phi_k)}{\partial \phi_k^2} a^{-s} \right] \\ &\equiv -\frac{1}{2} \sum'_k \hbar^2 \frac{\partial^2}{\partial \phi_k^2} a^{-s} + \mathcal{V}(\phi), \end{aligned}$$

where, for the choice of  $\Psi(\phi)$  given above,

$$\begin{aligned} \mathcal{V}(\phi) &\equiv \sum'_k \left[ \frac{1}{8} y'(\phi_k, a, \hbar)^2 - \frac{1}{4} \hbar y''(\phi_k, a, \hbar) + \frac{1}{2} \hbar \gamma_r y'(\phi_k, a, \hbar) \phi_k^{-1} \right. \\ &\quad \left. + \frac{1}{2} \hbar^2 \gamma_r (\gamma_r + 1) \phi_k^{-2} \right]; \end{aligned}$$

here

$$\gamma_r \equiv \frac{1}{2} - b\Delta = \frac{1}{2} - ba^s.$$

Consider the pseudofree ultralocal case for which

$$y(\phi_k, a, \hbar) = m_0 \phi_k^2 a^s.$$

For this choice, it follows that

$$\mathcal{V}_{pf}(\phi) \equiv \frac{1}{2} \sum_k' [m_0^2 \phi_k^2 a^s - \hbar m_0 (1 - 2\gamma_r) + \hbar^2 \gamma_r (\gamma_r + 1) \phi_k^{-2} a^{-s}].$$

Given the Hamiltonian for this case we can immediately determine the lattice action for this pseudofree ultralocal model. In particular, it follows that

$$I_{pf} = \sum_k \left\{ \frac{1}{2} [(\phi_{k^\#} - \phi_k)^2 a^{n-2} + m_0^2 \phi_k^2 a^n + \hbar^2 (\frac{1}{2} - ba^s)(\frac{3}{2} - ba^s) a^{-2s} \phi_k^{-2} a^n] \right\}.$$

In this expression the factor  $k^\#$  signifies the next lattice point advanced by one unit in the time direction, i.e., if  $k = (k_0, k_1, \dots, k_s)$  then  $k^\# = (k_0 + 1, k_1, \dots, k_s)$ . Note well that any constant term (zero point energy) in the Hamiltonian cancels out with a similar term in the normalization factor in the functional integral and need not be included in the lattice action. Observe that the classical limit for which  $\hbar \rightarrow 0$  accompanied by the continuum limit leads to the classical (Euclidean) action for the free ultralocal model.

### Interacting Ultralocal Models

Drawing on the foregoing analysis of the pseudofree ultralocal model, we may give a brief discussion of interacting ultralocal models. The quartic interaction in the lattice action leads to a lattice Hamiltonian of the form

$$\mathcal{H} = -\frac{1}{2} \hbar^2 \sum_k' \frac{\partial^2}{\partial \phi_k^2} + \mathcal{V}(\phi),$$

where

$$\mathcal{V}(\phi) = \sum_k' \left[ \frac{1}{2} m_0^2 \phi_k^2 a^s + \lambda_0 \phi_k^4 a^s + \frac{1}{2} \hbar^2 \gamma_r (1 + \gamma_r) \phi_k^{-2} a^{-s} \right] - E.$$

The constant  $E$  is chosen so that the ground state  $\Psi(\phi)$  fulfills  $\mathcal{H}\Psi(\phi) = 0$ . Unfortunately, the form of the expression  $y(\phi, a, \hbar)$  that is part of the ground state function is unknown, but it surely has the property that as  $\lambda_0 \rightarrow 0$ , then  $y(\phi, a, \hbar) \rightarrow m_0 \phi^2 a^s$  appropriate to the pseudofree model. Stated otherwise, the quartic interacting theory is continuously connected to the pseudofree model as advertised.

Although we can not analytically describe the ground state for the quartic ultralocal model, we can, as another example, choose a nonquadratic form for  $y(\phi, a, \hbar)$  and see to what interacting model it belongs. For example, let us consider

$$y(\phi, a, \hbar) = m_0 \phi^2 a^s + g_0 \phi^4 a^s,$$

which leads to the potential

$$\mathcal{V}(\phi) = \sum'_k \frac{1}{2} \{ m_0^2 \phi_k^2 + 4m_0 g_0 \phi_k^4 + 4g_0^2 \phi_k^6 - \frac{1}{2} \hbar [m_0(1 - 2\gamma_r) + 2g_0(2\gamma_r - 3)\phi_k^2] + \hbar^2 \gamma_r (1 + \gamma_r) \phi_k^{-2} \} a^s.$$

Evidently this choice describes a model with a mixed quadratic, quartic, and sixth order potential. The first three terms – those without  $\hbar$  as a coefficient – survive in the classical limit as  $\hbar \rightarrow 0$ . Again, as the nonlinear coupling  $g_0 \rightarrow 0$ , it follows that this interacting model is continuously connected to the pseudofree model.

### ***Another Route to Quantize Ultralocal Models***

Let us now derive the pseudofree ultralocal model by an alternative argument. First, we recognize the free model and its ground state on a regularizing lattice as given by

$$\Psi_0(\phi) = \sqrt{K} e^{-\frac{1}{2} m_0 \sum'_k \phi_k^2 a^s},$$

which gives rise to the ground state expectation functional

$$\begin{aligned} E_0(f) &= \lim_{\Delta \rightarrow 0} K \int e^{i \sum'_k p_k \phi_k a^s - m_0 \sum'_k \phi_k^2 a^s} \Pi'_k d\phi_k \\ &= e^{-(1/4 m_0) \int f(x)^2 d^s x}. \end{aligned}$$

Perturbations in the mass for example would involve expressions of the form

$$I_p(m_0) \equiv K \int [\sum'_k \phi_k^2 a^s]^p e^{-m_0 \sum'_k \phi_k^2 a^s} \Pi'_k d\phi_k,$$

for which the result is clearly divergent in the continuum limit where the number  $N'$  of spatial lattice points diverges. It is instructive to see just where that  $N'$  factor originates, and to do so we pass to *hyper-spherical coordinates* defined by the expressions

$$\begin{aligned} \phi_k &\equiv \kappa \eta_k, \quad \kappa \geq 0, \quad -1 \leq \eta_k \leq 1, \\ \kappa^2 &\equiv \sum'_k \phi_k^2, \quad 1 = \sum'_k \eta_k^2. \end{aligned}$$

In terms of these variables, it follows that

$$I_p(m_0) = 2K \int [\kappa^2 a^s]^p e^{-m_0 \kappa^2 a^s} \kappa^{N'-1} d\kappa \delta(1 - \sum'_k \eta_k^2) \Pi'_k d\eta_k.$$



For large  $N'$ , this integral may be estimated by steepest descent methods as

$$I_p(m_0) = O((N'/m_0)^p) I_0(m_0).$$

Moreover, in a perturbation calculation of  $I_1(m_0)$  about  $I_1(1)$  (say) it follows that

$$I_1(m_0) = I_1(1) - \delta m_0 I_2(1) + \frac{1}{2} \delta m_0^2 I_3(1) - \dots,$$

where  $\delta m_0 \equiv m_0 - 1$ . Clearly this series is divergent as  $N' \rightarrow \infty$ , i.e., in the continuum limit. Note well that  $N'$  makes an explicit appearance in this series *only* in the factor  $\kappa^{N'-1}$  that arises from the measure  $\prod'_k d\phi_k$  put into hyper-spherical coordinates.

To eliminate those divergences we need to eliminate that appearance of the factor  $N'$ . The only way to eliminate that factor is to change the ground state from that of the free system to that of the pseudofree system that takes account of the hard core. To attack the hard core directly is difficult and has so far not been a productive direction to follow. But, *and here is the main point of this discussion: To eliminate the factor  $N'$  that arises from the field measure it suffices to ensure that the ground state distribution for the pseudofree theory is such that*

$$\Psi_{pf}^2(\phi) \propto \kappa^{-(N'-R)} e^{-m_0 \sum'_k \phi_k^2 a^s}$$

**for some finite parameter  $R$ .**

For the ultralocal model, we shall more explicitly choose a ground state for the pseudofree model of the form

$$\Psi_{pf}(\phi) = K' \prod'_k |\phi_k|^{-(1-R/N')/2} e^{-\frac{1}{2} m_0 \phi_k^2 a^s},$$

which leads to the desired form and respects the ultralocal symmetry of the model. How do we choose  $R$ ? We require that this expression have an acceptable continuum limit, which we study by examining the characteristic function for the ground state distribution, i.e.,

$$\begin{aligned} E_{pf}(f) &= \lim_{a \rightarrow 0} \prod'_k K \int e^{i p_k \phi_k a^s - m_0 \phi_k^2 a^s} |\phi_k|^{-(1-R/N')} \prod'_k d\phi_k \\ &= \lim_{a \rightarrow 0} \prod'_k \{ 1 - K \int [1 - e^{i p_k \phi_k a^s}] e^{-m_0 \phi_k^2 a^s} |\phi_k|^{-(1-R/N')} \prod'_k d\phi_k \}. \end{aligned}$$

The only way to achieve a meaningful continuum limit is, first, (effectively) choose  $m_0 = (b a^s) m$ , where  $b$  is an arbitrary positive parameter with dimensions of  $L^{-s}$ , which, after a change of variables ( $\phi_k \rightarrow a^{-s} \lambda_k$ ), yields to leading order,

$$E_{pf}(f) = \lim_{a \rightarrow 0} \prod'_k \{ 1 - K \int [1 - e^{i p_k \lambda_k}] e^{-b m \lambda_k^2} |\lambda_k|^{-(1-R/N')} \prod'_k d\lambda_k \},$$

and, second, choose  $K = c(ba^s)$  [which fixes  $R$  to be  $R = 2c(ba^s)N'$ ], and thus

$$E_{pf}(f) = e^{-cb} \int d^s x \int_{\{1 - \cos[f(x)\lambda]\}} e^{-bm\lambda^2} d\lambda/|\lambda|.$$

Normally, the dimensionless factor  $c$  has been chosen as  $c = 1$  or  $c = \frac{1}{2}$ , but any positive value is acceptable.

It is of fundamental importance to observe that we have derived a correct version of the pseudofree ultralocal model by the simple act of choosing the pseudofree ground state distribution to cancel the unwanted factor  $N'$ , the very factor that **causes** the divergences in the first place, and then to ensure as meaningful a continuum limit as possible. This simple act ensures that all the moments of interest are now finite and no infinities arise whatsoever. Since this action has the effect of cancelling all divergences, it acts in all necessary ways as would the presumed hard core. In particular, the so-defined, divergence-free interacting theory does not pass continuously to the free theory but instead it passes to an alternative theory, namely, the pseudofree theory. That kind of limiting behavior is the biggest clue to the fact that the interaction acts as a (partial) hard core. Does the simple act of removing the offending factor  $N'$  accurately correspond to including the effects of the hard core? In fact, it really doesn't matter if the elimination of the factor  $N'$  is an accurate realization of the hard core; the putative "hard core" has already rendered an important service by refocussing our attention beyond those counter terms that are suggested by perturbation theory. Additionally, the study of the soluble ultralocal models has helped us clarify the question of whether removing the factor  $N'$  corresponds to accounting for the hard core. Specifically, the solution obtained from a rigorous viewpoint is identical to the one obtained by the supremely simple prescription of choosing a suitable pseudofree model that eliminates the offending factor  $N'$ . In this sense, the removal of the cause of the divergences, i.e., the factor  $N'$ , has rendered the theory finite in all respects, and since the result completely agrees with the rigorously obtained result, we are certainly entitled to assert that the removal of the factor  $N'$  has accounted for the presence of the hard core in the case of ultralocal models.

We shall see that this breathtakingly elementary procedure, coupled with a judicious choice of further details of the pseudofree model, will provide a divergence-free formulation of additional examples of nonrenormalizable models, formulations that would be difficult to arrive at by any other means. It is reasonable that the procedure to eliminate the source of divergences caused by the measure should apply to other models which, in some sense, are "close" to ultralocal models. It is also reasonable to expect that traditional nonrenormalizable models are good candidates on which to try a similar approach to deal with otherwise uncontrollable divergences.

## Relativistic Models

The classical (Euclidean) action for covariant, quartic self interacting scalar fields is given by

$$I = \int \left\{ \frac{1}{2} [(\nabla\phi(x))^2 + m_0^2\phi(x)^2] + \lambda_0\phi(x)^4 \right\} d^n x,$$

for an  $n$ -dimensional spacetime. To discuss the classical side of the pseudofree situation, we recall a classical Sobolev-type inequality (see, e.g., [1]) given by

$$\{\int \phi(x)^4 d^n x\}^{1/2} \leq c \int [(\nabla \phi(x))^2 + m_0^2 \phi(x)^2] d^n x,$$

which for  $n \leq 4$  holds with  $c = 4/3$  and for  $n \geq 5$ , requires that  $c = \infty$ . This result implies that for  $n \geq 5$ , there are fields  $\phi(x)$  for which the free part of the classical action is finite but for which the quartic interaction diverges. These are just the conditions under which a classical pseudofree theory different from the classical free theory exists. Thus it is possible when  $n \geq 5$  that the quantum theory also has a pseudofree theory different from its free theory.

We recall that a lattice regularized form of the Euclidean functional integral with only two free parameters ( $m_0$  and  $\lambda_0$ ) has been shown to pass to a (generalized) free theory in the continuum limit [2]; thus a richer variety of renormalization counterterms is required to avoid triviality. Since, for  $n \geq 5$  the quantum theories are perturbatively nonrenormalizable leading to a perturbation series composed of infinitely many distinct counterterms, such an approach does not resolve the problem. Our goal is to show that an unconventional counterterm suggested by what is needed to remove the source of the divergences can lead to a satisfactory resolution of all problems with the relativistic models. To that end we now turn our attention to a very different sort of lattice regularized functional integral formulation for self-interacting relativistic scalar fields.

In particular, relativistic interacting scalar models admit an analogous treatment to that of the ultralocal models, and in our present discussion we follow reference [3]. In begin with, let us introduce a lattice action defined by the expression

$$I(\phi, a, \hbar) \equiv \frac{1}{2} \sum_k (\phi_{k^*} - \phi_k)^2 a^{n-2} + \frac{1}{2} m_0^2 \sum_k \phi_k^2 a^n + \lambda_0 \sum_k \phi_k^4 a^n + \frac{1}{2} \hbar^2 \sum_k \mathcal{F}_k(\phi) a^n,$$

where there is an implicit summation over all  $n$  nearest neighbors in the positive sense symbolized by the notation  $k^*$ , and where the nonclassical counterterm is

$$\begin{aligned} \mathcal{F}_k(\phi) \equiv & \frac{1}{4} \left( \frac{N' - 1}{N'} \right)^2 a^{-2s} \sum_{r,t} \frac{J_{r,k} J_{t,k} \phi_k^2}{[\sum'_l J_{r,l} \phi_l^2][\sum'_m J_{t,m} \phi_m^2]} \\ & - \frac{1}{2} \left( \frac{N' - 1}{N'} \right) a^{-2s} \sum_t \frac{J_{t,k}}{[\sum'_m J_{t,m} \phi_m^2]} \\ & + \left( \frac{N' - 1}{N'} \right) a^{-2s} \sum_t \frac{J_{t,k}^2 \phi_k^2}{[\sum'_m J_{t,m} \phi_m^2]^2}. \end{aligned}$$

Here,

$$J_{k,l} \equiv \frac{1}{2s + 1} \delta_{k,l \in \{k \cup k_{nn}\}},$$

where  $\delta_{k,l}$  is a Kronecker delta. This latter notation means that an equal weight of  $1/(2s + 1)$  is given to the  $2s + 1$  points in the set composed of  $k$  and its  $2s$  nearest neighbors in the spatial sense only;  $J_{k,l} = 0$  for all other points in that spatial slice. [Specifically, we define  $J_{k,l} = 1/(2s + 1)$  for the points  $l = k = (k_0, k_1, k_2, \dots, k_s)$ ,  $l = (k_0, k_1 \pm 1, k_2, \dots, k_s)$ ,  $l = (k_0, k_1, k_2 \pm 1, \dots, k_s), \dots$ ,  $l = (k_0, k_1, k_2, \dots, k_s \pm 1)$ .] This definition implies that  $\sum'_l J_{k,l} = 1$ .

For the ultralocal model, the analog of the constants  $J_{k,l}$  is the Kronecker delta, i.e.,  $\delta_{k,l}$ . In that case it was important to respect the physics of the ultralocal model with no interaction between fields at distinct (lattice) points. For the relativistic models, on the other hand, there is indeed communication between spatially neighboring points and we can use that fact to provide a lattice-symmetric, regularized form of the denominator factor. Moreover, the lack of integrability at  $\phi_k = 0$ , for each  $k$ , which was critical for the ultralocal models to ensure that the ground state becomes a generalized Poisson distribution in the continuum limit, is exactly what is *not* wanted in the case of the relativistic models. This latter fact is ensured by the factors  $J_{k,l}$  as chosen.

We first focus on our choice of the pseudofree model in the relativistic case, which is chosen somewhat differently than in the ultralocal case. Specifically, we define the generating function for the lattice regularized, covariant pseudofree model by

$$S_{pf}(h) = M_{pf} \int \exp[Z^{-1/2} \sum_k h_k \phi_k a^n / \hbar - \frac{1}{2} \sum_k (\phi_{k^*} - \phi_k)^2 a^{n-2} / \hbar - \frac{1}{2} \hbar \sum_k \mathcal{F}_k(\phi) a^n] \Pi_k d\phi_k ;$$

here,  $Z$  denotes the so-called field strength renormalization constant to be discussed below. Associated with this choice of the pseudofree generating function is the lattice Hamiltonian for the pseudofree model, which (with the zero point energy subtracted) reads

$$\mathcal{H}_{pf} = -\frac{1}{2} \hbar^2 a^{-s} \sum_k \frac{\partial^2}{\partial \phi_k^2} + \frac{1}{2} \sum_k (\phi_{k^*} - \phi_k)^2 a^{s-2} + \frac{1}{2} \hbar^2 \sum_k \mathcal{F}_k(\phi) a^s - E_0.$$

Lastly, we introduce the expression for the pseudofree ground state

$$\Psi_{pf}(\phi) = \sqrt{K} \frac{e^{-\sum'_{k,l} \phi_k A_{k-l} \phi_l a^{2s} / 2\hbar - W(\phi) a^{(s-1)/2} / \hbar^{1/2}} / 2}{\Pi'_k [\sum'_l J_{k,l} \phi_l^2]^{(N'-1)/4N'}}$$

which, in effect, was chosen *first*, and then the lattice Hamiltonian and the lattice action were derived from it. We discuss the (unknown) function  $W$  below; however, we observe here that the other factors in  $\Psi_{pf}(\phi)$  properly account for both the large field and small field behavior of the ground state.

In the next section we discuss the continuum limit, and in doing so we are again guided by the discussion in [3].

### Continuum Limit

Before focusing on the limit  $a \rightarrow 0$  and  $L \rightarrow \infty$ , we note several important facts about ground-state averages of the direction field variables  $\{\eta_k\}$ . First, we assume that such averages have two important symmetries: (i) averages of an odd number of  $\eta_k$  variables vanish, i.e.,

$$\langle \eta_{k_1} \cdots \eta_{k_{2p+1}} \rangle = 0,$$

and (ii) such averages are invariant under any spacetime translation, i.e.,

$$\langle \eta_{k_1} \cdots \eta_{k_{2p}} \rangle = \langle \eta_{k_1+l} \cdots \eta_{k_{2p}+l} \rangle$$

for any  $l \in \mathbb{Z}^n$  due to a similar translational invariance of the lattice Hamiltonian. Second, we note that for any ground-state distribution, it is necessary that  $\langle \eta_k^2 \rangle = 1/N'$  for the simple reason that  $\sum'_k \eta_k^2 = 1$ . Hence,  $|\langle \eta_k \eta_l \rangle| \leq 1/N'$  as follows from the Schwarz inequality. Since  $\langle [\sum'_k \eta_k^2]^2 \rangle = 1$ , it follows that  $\langle \eta_k^2 \eta_l^2 \rangle = O(1/N'^2)$ . Similar arguments show that for any ground-state distribution

$$\langle \eta_{k_1} \cdots \eta_{k_{2p}} \rangle = O(1/N'^p),$$

which will be useful almost immediately.

### Field Strength Renormalization

For  $\{h_k\}$  a suitable spatial test sequence, we insist that expressions such as

$$\int Z^{-p} [\sum'_k h_k \phi_k a^s]^{2p} \Psi_{pf}(\phi)^2 \Pi'_k d\phi_k$$

are finite in the continuum limit. Due to the intermediate field relevance of the factor  $W$  in the pseudofree ground state, an approximate evaluation of the integral will be adequate for our purposes. Thus, we are led to consider

$$\begin{aligned} & K \int Z^{-p} [\sum'_k h_k \phi_k a^s]^{2p} \frac{e^{-\sum'_{k,l} \phi_k A_{k-l} \phi_l a^{2s}/\hbar - W}}{\Pi'_k [\sum'_l J_{k,l} \phi_l^2]^{(N'-1)/2N'}} \Pi'_k d\phi_k \\ & \simeq 2K_0 \int Z^{-p} \kappa^{2p} [\sum'_k h_k \eta_k a^s]^{2p} \\ & \times \frac{e^{-\kappa^2 \sum'_{k,l} \eta_k A_{k-l} \eta_l a^{2s}/\hbar}}{\Pi'_k [\sum'_l J_{k,l} \eta_l^2]^{(N'-1)/2N'}} d\kappa \delta(1 - \sum'_k \eta_k^2) \Pi'_k d\eta_k, \end{aligned}$$

where  $K_0$  is the normalization factor when  $W$  is dropped. Our goal is to use this integral to determine a value for the field strength renormalization constant  $Z$ . To estimate this integral we first replace two factors with  $\eta$  variables by their appropriate averages. In particular, the quadratic expression in the exponent is estimated by

$$\kappa^2 \sum'_{k,l} \eta_k A_{k-l} \eta_l a^{2s} \simeq \kappa^2 \sum'_{k,l} N'^{-1} A_{k-l} a^{2s} \propto \kappa^2 N' a^{2s} a^{-(s+1)},$$

and the expression in the integrand is estimated by

$$[\sum'_k h_k \eta_k a^s]^{2p} \simeq N'^{-p} [\sum'_k h_k a^s]^{2p}.$$

The integral over  $\kappa$  is then estimated by first rescaling the variable  $\kappa^2 \rightarrow \kappa^2 / (N' a^{s-1} / \hbar)$ , which then leads to an overall integral estimate proportional to

$$Z^{-p} [N' a^{s-1}]^{-p} N'^{-p} [\sum'_k h_k a^s]^{2p};$$

at this point, all factors of  $a$  are now outside the integral. For this result to be meaningful in the continuum limit, we are led to choose  $Z = N'^{-2} a^{-(s-1)}$ . However,  $Z$  must be dimensionless, so we introduce a fixed positive quantity  $q$  with dimensions of an inverse length, which allows us to set

$$Z = N'^{-2} (qa)^{-(s-1)}.$$

### Mass and Coupling Constant Renormalization

A power series expansion of the mass and coupling constant terms lead to the expressions  $\langle [m_0^2 \sum_k \phi_k^2 a^n]^p \rangle$  and  $\langle [\lambda_0 \sum_k \phi_k^4 a^n]^p \rangle$  for  $p \geq 1$ , which we treat together as part of the larger family governed by  $\langle [g_{0,r} \sum_k \phi_k^{2r} a^n]^p \rangle$  for integral  $r \geq 1$ . Thus we consider

$$\begin{aligned} K \int [g_{0,r} \sum'_k \phi_k^{2r} a^s]^p & \frac{e^{-\sum'_{k,l} \phi_k A_{k-l} \phi_l a^{2s} / \hbar - W}}{\Pi'_k [\sum'_l J_{k,l} \phi_l^2]^{(N'-1)/2N'}} \Pi'_k d\phi_k \\ & \simeq 2 K_0 \int g_{0,r}^p \kappa^{2rp} [\sum'_k \eta_k^{2r} a^s]^p \\ & \times \frac{e^{-\kappa^2 \sum'_{k,l} \eta_k A_{k-l} \eta_l a^{2s} / \hbar}}{\Pi'_k [\sum'_l J_{k,l} \eta_l^2]^{(N'-1)/2N'}} d\kappa \delta(1 - \sum'_k \eta_k^2) \Pi'_k d\eta_k. \end{aligned}$$

The quadratic exponent is again estimated as

$$\kappa^2 \sum'_{k,l} \eta_k A_{k-l} \eta_l a^{2s} \propto \kappa^2 N' a^{2s} a^{-(s+1)},$$

while the integrand factor

$$[\Sigma'_k \eta_k^{2r}]^p \simeq N'^p N'^{-rp}.$$

The same transformation of variables used above precedes the integral over  $\kappa$ , and the result is an integral, no longer depending on  $a$ , that is proportional to

$$g_{0,r}^p N'^{-(r-1)p} a^{sp} / N'^{rp} a^{(s-1)rp}.$$

To have an acceptable continuum limit, it suffices that

$$g_{0,r} = N'^{(2r-1)} (qa)^{(s-1)r-s} g_r,$$

where  $g_r$  may be called the physical coupling factor. Moreover, it is noteworthy that  $Z^r g_{0,r} = [N'(qa)^s]^{-1} g_r$ , for all values of  $r$ , which for a finite spatial volume  $V' = N' a^s$  leads to a finite nonzero result for  $Z^r g_{0,r}$ . It should not be a surprise that there are no divergences for all such interactions because the source of all divergences has been neutralized!

We may specialize the general result established above to the two cases of interest to us. Namely, when  $r = 1$  this last relation implies that  $m_0^2 = N'(qa)^{-1} m^2$ , while when  $r = 2$ , it follows that  $\lambda_0 = N'^3 (qa)^{s-2} \lambda$ . In these cases it also follows that  $Z m_0^2 = [N'(qa)^s]^{-1} m^2$  and  $Z^2 \lambda_0 = [N'(qa)^s]^{-1} \lambda$ , which for a finite spatial volume  $V' = N' a^s$  leads to finite nonzero results for  $Z m_0^2$  and  $Z^2 \lambda_0$ , respectively.

## Conclusion

For covariant scalar nonrenormalizable quantum field models, we have shown that the choice of a nonconventional counterterm, but one that is still nonclassical, leads to a formulation for which a perturbation analysis of both the mass term and the nonlinear interaction term, expanded about the appropriate pseudofree model, are term-by-term finite.

Coupled with the discussion for the ultralocal models, it is evident that the present analysis would suggest a related formulation for so-called *Diastrophic Quantum Field Theories* introduced by the author in [4]. These models are distinguished by the fact that they can be viewed as fully relativistic models modified so that some (but not all) of the spatial derivatives are dropped; thus these models lie, in a certain sense, between the relativistic and ultralocal models.

It is also hoped that some of these ideas may have relevance in one or more formulations of quantum gravity, such as, for example, in the program of *Affine Quantum Gravity* introduced by the author; see [5].

## References

- [1] J.R. Klauder, *Beyond Conventional Quantization*, (Cambridge University Press, Cambridge, 2000 & 2005).
- [2] M. Aizenman, “Proof of the Triviality of  $\phi_d^4$  Field Theory and Some Mean-Field Features of Ising Models for  $d > 4$ ”, *Phys. Rev. Lett.* **47**, 1-4, E-886 (1981); J. Fröhlich, “On the Triviality of  $\lambda\phi_d^4$  Theories and the Approach to the Critical Point in  $d \geq 4$  Dimensions”, *Nuclear Physics B* **200**, 281–296 (1982).
- [3] J.R. Klauder, “Taming Nonrenormalizability”, *J. Phys. A: Math. Theor.* **41**, 335208 (7pp) (2008).
- [4] J.R. Klauder, “Covariant Diastrophic Quantum Field Theory”, *Phys. Rev. Lett.* **28**, 769–772 (1972).
- [5] J.R. Klauder, “The Affine Quantum Gravity Programme”, *Class. Quant. Grav.* **19**, 817–826 (2002).



# Magnetism, FeS Colloids, and Origins of Life

Gargi Mitra-Delmotte and A.N. Mitra

*Dedicated to the memory of Professor Alladi Ramakrishnan*

**Summary** A number of features of living systems, reversible interactions and weak bonds underlying motor-dynamics; gel-sol transitions; cellular connected fractal organization; asymmetry in interactions and organization; quantum coherent phenomena; to name some, can have a natural accounting via *physical* interactions, which we therefore seek to incorporate by expanding the horizons of “chemistry-only” approaches to the origins of life. It is suggested that the magnetic “face” of the minerals from the inorganic world, recognized to have played a pivotal role in initiating Life, may throw light on some of these issues. A magnetic environment in the form of rocks in the Hadean Ocean could have enabled the accretion and therefore an ordered confinement of super-paramagnetic colloids within a structured phase. A moderate H-field can help magnetic nanoparticles to not only overcome thermal fluctuations but also harness them. Such controlled dynamics brings in the possibility of accessing quantum effects, which together with frustrations in magnetic ordering and hysteresis (a natural mechanism for a primitive memory) could throw light on the birth of biological information which, as Abel argues, requires a combination of order and complexity. This scenario gains strength from observations of scale-free framboidal forms of the greigite mineral, with a magnetic basis of assembly. And greigite’s metabolic potential plays a key role in the mound scenario of Russell and coworkers—an expansion of which is suggested for including magnetism.

**Mathematics Subject Classification (2010)** 82D40, 92-02, 93-02, 94-02

---

G. Mitra-Delmotte

Present Address: 39 Cite de l’Ocean, Montgaillard, St. Denis 97400, REUNION

e-mail: [gargijj@orange.fr](mailto:gargijj@orange.fr)

A.N. Mitra

Professor Emeritus, Department of Physics, Delhi University

244 Tagore Park, Delhi-110009, India

e-mail: [ganmitra@nde.vsnl.net.in](mailto:ganmitra@nde.vsnl.net.in)

**Key words and phrases** Magnetic-reproduction · Brownian noise · Symmetry-breaking · Ferro-fluids · Super-paramagnetic particle · Ligand-effects · Greigite mineral

## 1 Introduction

Life's hierarchical control structure is a sequence of constraints, each limiting the scope of the preceding level for stepwise harnessing of the physico-chemical laws governing its lowest rung. But the limiting "boundary conditions" are themselves extraneous; they cannot be formally derived from these laws. Furthermore, the higher-level operating principles depend on, but are *not* reducible to, those of the lower ones [127]. Next, the origins of purpose permeating across biology [67], as well as information associated with function, are among the most fundamental of questions in biology [80]. Indeed, the structure-function relationship, where rate-dependent equations representing measurement associated with biostructures are linked to rate-independent constraints associated with bioinformation, is viewed as an epistemological complementarity [123]. According to Pattee, "epistemic operations like observation, detection, recognition, measurement, and control as the essential type of function" demarcate living from non-living organizations. The chances of an organism's survival are crucially dependent on its ability to improve its control strategies that in turn depend on its recognition of environmental patterns. Hence, "To qualify as a measuring device it must have a function, and the most primitive concept of function implies improving fitness of an organism." Pattee's famous "semantic closure principle" places a heavy responsibility on the *observer* who should at minimum be an *organization* that can construct the measuring device and use the results of measurement for its very survival [124]. This scenario seems to be a far cry from the objective (observer-independent) physical laws characterized by Universality and Invariance Principles. And it is indeed a tall order to explain from these "classical" premises the emergence of subjective (observer-dependent) biological infrastructure making measurements for survival. But Pattee recognizes that unlike classical theory, Quantum theory is *not* constrained by observer – independence and promptly invokes Wheeler to make his point: "No elementary quantum phenomenon is a phenomenon until it is a recorded phenomenon (i.e., the results of a measurement)." Indeed, the puzzle is really about how the "Cybernetic cut" [1] could have been crossed using mere physiodynamics, leading to the emergence of a nonphysical (not governed by chance or necessity) mind from physicality that established controls over the same. We further ask whether this mystery could somehow be related to the idea of life having originated in an inorganic world—an idea which has met considerable acceptance. The compelling link to iron (FeS) clusters in early evolved enzymes (and across species in a range of crucial roles, e.g., catalytic, electron transfer, structural), with exhalates on the Hadean ocean floor, is based on the close resemblance of these clusters with greigite (Fe<sub>5</sub>NiS<sub>8</sub>) [62, 138]. Not only are these clusters seen as playing a key role in the origins of metabolism, where geochemical gradients were harnessed, but also, for long

mineral crystal surfaces have, and continue to be seen as scaffolds, thanks to their chemical-information storing/transferring potential, leading to the other-replicating – wing of Life [6, 14, 20, 33, 42, 49]. But in these approaches, a number of features: reversible interactions, weak bonds, gel-sol transitions, cellular connected fractal organization, asymmetry in interactions and organization, to name some, and which are difficult to address using chemical interactions alone, are seen as later arrivals, i.e., upon achievement of complexity in the pre-biotic “soups.” Here again, the path, as to how complexity could have been entrained to lead to Life-like features of today, remains far from being understood. Then, in addition to chemistry, could physical properties of inorganic matter have also acted as a scaffold for onward transmission of several common *physical* features (see below) typical of living systems? To that end, we note that dynamically ordered forms of matter, like framboids, regardless of chemical structure, are the result of physical forces, including magnetism (see Sect. 4).

Now, magnetism has myriad manifestations at different scales – quantum to cosmological [153]. (The repeated appearance of fractal themes is compelling – from magnetic critical phenomena to finer length scales where quasiparticle behavior in a magnetic field can be explained by fractional quantum numbers [48, 64]; Farey series elements,  $F_n$ ; Hausdorff dimension  $h$  [28]. And, there are ubiquitous magnetic influences across kingdoms : navigation sensing in bacteria, algae, protists, bees, ants, fishes, dolphins, turtles, and birds [72, 172]; field effects on growth patterns, differentiation, orientation of plants and fungi [47]; ferromagnetic elements in tissues [74], etc. (In fact, magnetite ( $Fe_3O_4$ , a magnetic mineral) – biomineralization, the most ancient matrix-mediated system, is thought to have served as an ancestral template for exaptation [73]. Indeed, new inputs of quantum events underlying biophenomena like magnetoreception [75] reveal the importance of magnetism in biological systems of today. Most importantly, its vital role in *the science of information technology* persuades us to turn to this enveloping science for any mechanisms beyond the limits of physicochemical principles that could have helped bridge the gap from inanimate matter to life.

In this mini survey

- (1) We give a brief summary of the relevance of quantum searches in biology and therefore to the origin-of-life problem (Sect. 2.1). We briefly review spin and magnetic models offering insights into the emergence of life, leading up to our proposal (Sects. 2.2–4).
- (2) We survey various biophenomena with analogies to magnetic ones in general as well as topological similarities with our magnetism-based proposal in particular (Sects. 3.1–9), and ask whether magnetism could have helped to pave the way for a takeoff from non-life to life.
- (3) We briefly review framboids, where conflicting physical forces usher in dynamic order. Here, the mineral greigite’s magnetic properties underlie its framboid-forming capacity (Sect. 4).
- (4) We outline the mound scenario of Russell and coworkers, with rich metabolism potential, where greigite forms in a colloidal environment. A possible scenario for a magnetic reproducer is drawn (Sect. 5).

## 2 Quantum Searches and the Origins of Life

A brief introduction on quantum searches in biology is followed by their implications in the origin of life. A possible physical system enhancing the propensity of such searches is then suggested.

### 2.1 *Quantum Searches and Biology*

Outstanding biological-search examples can be seen in biological evolution itself, with divergences symbolized by tree nodes; the clonal Darwinian-like phase in the adaptive immune system; brain connections and protein folding. The efficiency of quantum searches over classical ones has prompted the idea that they could have been used by Nature who usually is found to take the cleverest among available options, as illustrated by certain Extremum Principles of Classical physics (Hamilton, Fermat, Maupertius). For instance, in a database of dimension  $d$ , a quantum search gives a square root speedup over its classical counterpart – also valid for the respective nested versions [21]. In a typical scenario, challenges interrupting the networking phase are seen as forcing the biosystem to seek help from a co-existing quantum domain, e.g., a search prompted by a “crisis” in the form of a depleted nutrient could lead the adaptive system to a new pathway for succor. Now, quantum coherence in the set of elements on the affected front could help skirt frustrations in local minima as can happen in a classical search. This access to the wave-property enables a superposition of states and allows a “holistic” decision. Thus in the face of crises, halted networked interactions in a subsystem would prompt the formation of a “quantum decision front.” This would be constantly checked or “measured” by the rest of the system. A fruitful interaction with one chosen path would mean a simultaneous collapse of the quantum superposition of alternative paths [96].

Today, clear signatures of quantum processing in biology are coming in [40], aided by femtosecond laser-based 2D spectroscopy and coherent control approaches, showing how phase relationships in nanostructures modulate the course of bioreactions [106]. As to decoherence evading mechanisms, the role of a gel-state; quasicrystalline order; [51, 66]; are among proposed order-maintaining mechanisms in a wet environment, while “screening effect,” or “cocooning” structural mechanisms are seen as providing insulation against interactions with the environment [29, 30, 119] (see also Sect. 3.9). Indeed, it seems that Nature has quietly been using these strategies all along, i.e., leading to creation of biological language itself, as the Grover-Patel search numbers match those used by Nature! Using Grover’s quantum search method for a marked item in an unsorted database, Patel [119] hit upon the base-pairing logic of nucleic acids in transcription and translation as an excellent quantum search algorithm – a directed walk through a superposition of all possibilities – resulting in a twofold increase in sampling efficacy over its classical counterpart (which at best permits a random walk). Prompted by these insights,

Al-Khalili and McFadden [5] point out that a quantum search would have been far more efficient than a random one for picking out the self-replicator from the primordial soup comprising a dynamic combinatorial library of compounds linked together, say by reversible reactions. But what plausible ingredients could have facilitated such a quantum-assisted leap?

## 2.2 *Spin and Magnetic Systems for the Origin of Life*

Hypothesizing a quantum-mediated process for the transition from non-life to life, Davies [31] proposes that information could have its origins in quantum objects such as spins, whose orientations offer a natural discretization mechanism of genetic information, and which in turn may have been embodied by physical structures in some natural system. Although this would initially be copying bits (no associated phase information so initially no issues of decoherence evasion), the possibility of coherence in this inherently quantum system endows it with a potential for conducting a quantum search for the quantum replicator. Furthermore, he points out that in this envisioned scenario, the collapse of the quantum superposition of states of living and non-living ones to the low probability state of “life” cannot be due to the quantum system’s own doing. Instead it must have been the result of an environmental interaction, serving as a measuring device, thus implying a key role for the environment (cf. [178]). Again, an origin-of-life model based on spin-ordering (a variant of the Ising spin glass) was proposed by Anderson [9], which was albeit prompted from another angle – the correspondence between the complexity due to the impact of frustration in magnetically disordered systems and bio-processes, such as protein-folding [61, 155] (see Sect. 3.1). Then again, Breivik [17] demonstrated that self-ordering of ferromagnetic objects ( $\sim 3$  mm) with reproduction of magnetic templates could be manipulated via dynamic interaction with environmental temperature fluctuations, thereby significantly also connecting information encoded in nucleic acids with non-chemically linked aperiodic polymers. This is because a magnetically packed array is naturally aperiodic (see Sect. 3.8), hence *satisfying Schrodinger’s [150] vision of aperiodic surfaces as efficient information-holders*, in contrast to a periodic crystal lattice with strongly correlated elements. This magnetic mechanism for propagating information also agrees with Dyson’s [38] suggestion that “physical reproduction” preceded chemical replication in the origins of life, the latter being identified with a specific chemical copying process. And interestingly, his use of a magnetic analogy for states, obeying the Boltzmann probability distribution, gels with the kinetic aspects of biological reactions [129].

All of these compels us to ask whether magnetism could have empowered the initial conditions for traversing the bridge dividing life from non-life, by providing simultaneously a scaffold for interactions and connections, where physical representations would allow for higher level abstractions, not of the isolated system but rather in the context of its penetrating environment playing an active role in its decision-making. We have recently proposed [103, 104] that an external field in the

form of magnetic rocks could have enabled accretion of newly forming, magnetic nanoparticles on the Hadean Ocean floor, because of field-induced aggregates have been observed in magnetic fluids showing deviations from ideal behavior.

### 2.3 *Ferrofluids; Field-Induced Structures*

Ferrofluids are colloidal single-domain magnetic nanoparticles ( $\sim 10$  nm) in non-magnetic liquids that can be controlled by moderate H-fields ( $\sim$  tens of milliTesla) [110]. The relevance of these dispersions to natural locations has been considered only rarely, e.g., see [171], perhaps due to their synthetic origins; nevertheless, their amazing properties lead to myriad applications, including ratchet behavior [39]. On the one hand dilute dispersions display ideal single-phase behavior due to prohibited (chemical) interparticle contacts, thanks to synthetic coatings. On the other hand, in the present context we look at the interactions between the magnetic particles although the carrier remains in the liquid state. Such deviations from ideal magnetization behavior can show up on increasing particle concentrations that can be understood in terms of H-field-induced inter-particle interactions leading to internal structure formation [22, 135] and manifesting in *dense phases* – a milder phase transition than to the solid-crystalline one. The structure of hydrated, heterogenous aggregates would depend on factors such as the strength of the applied field, the nature of the ferrofluid, etc [110, 176, 177]. Li et al. [85] have pointed out the dissipative nature of the field-induced aggregates [157] that break up in response to thermal effects upon removal of field. In their gas-like compression model, the total magnetic energy of ferrofluids obtained from an applied field:  $W_T = W_M + W_S$ ; where  $W_M = \mu_0 M H V$  and  $W_S = -T \Delta S$  are the magnetized and the structured energies, respectively,  $V$  is the volume of the ferrofluid sample and  $\Delta S$  is the entropic change due to the microstructure transition of the ferrofluid. An assumed equivalence of  $W_T$  (zero interparticle interactions), with the Langevin magnetized energy  $W_L = \mu_0 M H V$  necessitates a correction in the magnetization, in terms of the entropy change. Hence, these colloidal systems are well equipped to analyze the interplay between competing factors -dipolar interactions, thermal motion, screening effects, etc. leading to the emergence of magnetically structured phases [122].

### 2.4 *Structured Magnetic Phases; Life-Like Dynamics*

On analogous lines to ferrofluids, magnetic rocks providing a surface field strength  $\sim$  tens of milli-Tesla would have turned any newly forming magnetic particle suspension into tiny magnets, leading to the emergence of magnetically structured phases (MSPs). We come to a suggested scenario in Sect. 5. Here, the magnetic entropy property of super-paramagnetic particles offer a ready basis for interchange with the Brownian hits from the surroundings for harnessing this energy, analogous

to complex biological soft matter, while the external magnetic environment plays a key role in controlling their dynamics. Furthermore, we suggested [104] that the presence of charge on particles would permit only the tiny sized particles (carrying one/two units of charge) to diffuse through layers of the magnetically accreted charged layers in response to a non-equilibrium source – a gentle gradient of flux lines (assuming a non-homogeneous H-field from rocks). Non-equilibrium energy driven diffusion of tiny particles (ligand-carrying or otherwise) through the magnetically ordered phase in a close-to-equilibrium manner shows the possibility of controlled dynamics in a confined system.

The connections between field-induced structures of magnetic nanoparticles and biophenomena bring out their ramifications for fluctuation-generated order from dissipative structures as envisaged decades ago [109]. Note that a magnetic environment exerts control on spin states and hence on spin-selective chemical reactions (see [18]). The possibility of yet another magnetic control is via magnetically sensitive reactions whose rates are sensitive to orientations of reactants [170]. Separation of complex mixtures forming at the origins of life would have also been facilitated by magnetic mechanisms, acting in an orthogonal non-interfering manner.

### 3 “The Importance of Being Magnetic”

We now look at some general features of biological systems with similarity to magnetic phenomena, also comparing dynamics in biology vis-a-vis our proposal of a nanoscale assembly controllable by a magnetic environment.

#### 3.1 *Confinement, Connectivity, Frustration-Complexity*

Self-ordering phenomena [109] show how spontaneous order can emerge from inanimate matter, leading to connected components (confined). But the high algorithmic compressibility of order and patterns that can be explained in terms of physical laws would simultaneously make it difficult to generate the complexity (high information carrying capacity) underlying biology [2]. In Shannon’s terminology, the information carrying capacity of a 1D-string is at its maximum when there are no correlations between its components, i.e., when it is a random sequence. A combination of the two-order and unpredictability – might be a better way to understand this paradox of biological complexity [2]. Now, frustrations in magnetically connected systems are well known in literature (see also Sects. 2.2, 4.5). Their presence naturally introduce the element of uncertainty in the midst of long-range correlations. We therefore suggest that a confined system due to magnetic connections, as in our proposal, has the combination for addressing such complexity in the origins of life.

### 3.2 *Nested Hierarchy, Cooperative Dynamics*

Biological structures appear as nested organizations based on coherent feedback through a lattice of interacting, spatially oriented units; self and non-self interactions underlie their cooperative dynamics [87]. And as noted by Min et al. [102], the characteristics of dynamically self-assembled nano-structures with bottom-up complexity, formed by dissipating energy, depend on the constituent particle size, shape, hardness, and composition, apart from their sensitivity to (control by) external fields; this approach was used in generating systems with hierarchial complexity via an interplay of magnetic and hydrodynamic interactions [50] (see also Sect. 4). In this connection recall some facets of magnetism in common with those of self-organizing systems: emergence of global order from local interactions, organizational closure, hierarchy, downward causation, distributed control underlying robustness, bifurcations via boundary conditions, non-linearity due to feedback, etc [57]. Their relevance can be gauged from the insights of Bak and Chen [11]: long-range spatiotemporal correlations (via a non-dimensional scale factor) are the hallmark of self-similarity, manifest as self-organized criticality in natural dynamical systems. Again, Selvam [151] proposed a coherence preservation mechanism via self-similar structures with quasicrystalline order as iterative principles – the main tools for handling non-linear dynamics of perturbations for evolving nested order that connect the microscopic and macroscopic realms with scale-free structures arising out of deterministic chaos. This brings us to Tagore’s couplet:

*Amra shobai raja amader ei rajar rajottey, noiley moder rajar shoney milbey ki  
shottey – Tagore*

(We are all kings in our King’s kingdom, else how do we get along with Him.)

### 3.3 *Polar Cell-Organization and Structures*

On higher scales, the directionality of biochemical processes gets derived from the asymmetric structure of biomolecules and their association into consequently polarized assemblies with increasing complexity [52]. The cytoskeleton, at least in eukaryotes, is organized via transmitted internal or external spatial cues, reflecting the polar organization of the cell [35]. We also note that some fundamental biological structures form from asymmetric monomers. For instance, the directionality of nucleic acid polymers stems from the asymmetry of template-based aligning monomers. The cytoskeletal family of proteins provides another outstanding example. The past two decades revealed how analogous functions are carried out by bacterial homologues of eukaryotic cytoskeletal proteins. Actually, the highly conserved FtsZ, barring a few exceptions, is found across all eubacteria and archaea. Despite its low sequence identity to tubulin, its eukaryotic homologue, the two proteins not only share the same fold but also follow similar self-assembly patterns, forming protofilaments. The longitudinal contact of the assembling monomers is in



a head-to-tail fashion. The other crucial eukaryotic cytoskeletal protein – actin – also shows a distinct asymmetry. It forms double-helical thin filaments composed of two strands. Within these, actin assembles in a head-to-tail manner, similar to its bacterial homologues [99]. Indeed, another association between the cytoskeletal network and percolation systems [161] recalls the long-range connectivity of magnetism (e.g., magnetic percolation clusters forming fractal networks [63]. Again, the diamagnetic anisotropy of planar peptide bonds permits their oriented self-assembly in a magnetic field, seen for fibrous biostructures [159, plus ref].

### 3.4 *Reversible Gel-Sol Transitions*

A far cry from organelles floating in sacs, the cytoplasm appears to have rich structure irrespective of species, with increasingly reported associations of mobile proteins with defined, albeit transient, locations [52]. Again, “site-dipoles” have been proposed for resolving the apparent contradiction between the seemingly random molecular movements and the correlated orientations in assemblies. Thus, the co-operativity among water molecules occupying the site-dipole field surrounding a solute in MD simulations, manifested in coherent patterns ( $\sim 14\text{\AA}^{\circ}$ ) that lasted about 300 ps, even as individual molecules randomly moving in and out of the sites, rapidly lost their orientational memory [58]. Indeed, the cell is viewed as a gel; reversible gel-sol phase transitions underlie its dichotomy that can be accessed via subtle environmental variations leading to finite structural changes [162]. Like hydrated cross-linked polymer gels, the cytoplasm thus exhibits excluded volume effects and sizeable electrical potentials. Biomolecules like proteins and ions play a critical role in structuring of intracellular water [23, 24]. This capacity to lie on the border between liquid and gel states underlies life’s ability to make the most of fluidity of the liquid state as well as long range order of the more solid gel phase, enabling self-assembly of softmatter. Now, in the origins of life, unlike a chemically bonded thermally formed gel, a magnetic gel has the potential of reverting to its colloidal components just like *colloid-gel* transitions pointed out in living systems [162].

### 3.5 *Reversible Interactions; Weak Bonds*

The sensitivity of biomolecular machines to thermal noise is a rather intriguing phenomenon. And, they have evidently learnt to harness these, thanks to the continuous nature of the energy landscape connecting different states. Again, the interconvertibility between different states is permitted due to the use of weak interactions (Van der Waal’s, H-bonds, hydrophobic, etc), used for their temporary maintenance. Significantly, the Berry’s phase-like periodic cycles [8] shown by biomolecular motors reveal different trajectories for two half cycles (with different binding capacities in

forward and backward directions) that can be understood in terms of their internal degrees of freedom. How could such complexity of biological macromolecules have arisen from simple matter, e.g., small molecules with a few discrete energy states, present at the dawn of Life? This is because these very features underlie the efficiency of biological machines that are being increasingly viewed as microscopic systems governed by the fluctuation-dissipation theorem (in the linear regime). The variations in total Gaussian-distributed energy of a macroscopic system with  $N$  particles, relative to the average value, are of the order  $N^{-1/2}$ . Thus, fluctuations would be negligible for macroscopic systems, but they would be relevant for microscopic ones, and also when the total energy of the system is  $\sim k_B T$ .

Next, in small systems in equilibrium or non-equilibrium steady states, the behavior remains unchanged in time, although a constant input of energy is required for the latter, operating away from equilibrium. No net heat transfer occurs in the former, with equal probabilities of absorbing/releasing heat from bath. However, the probability ratio differs from one for nonequilibrium steady state systems that dissipate heat on the average. And heat, being an extensive quantity, the probability of its absorption becomes exponentially smaller with increasing system size. On the other hand, for microscopic systems like biomolecular machines driven by rectified thermal fluctuations, this Maxwell-Demon-like probability can be significant [19]. This has been very succinctly phrased in a recent review [54] as follows: "These engines have one foot in the equilibrium camp and another in the world of fluctuations and non-equilibrium." Indeed, Jaryznski [65] showed that the average of the exponential of the energy of a microscopic system, pulled quickly away from equilibrium (instead of the simple average) works out to have the same value as the equilibrium energy change corresponding to a slow version of the same. This prediction was experimentally verified by Bustamante et al. [19], where the result remained unaffected upon changing the applied shearing force. In this proposal, diffusion of tiny particles driven by non-equilibrium energy, via infinitesimal changes in their relative orientations through the magnetically ordered phase in a close-to-equilibrium manner shows the possibility of controlled dynamics analogous to ATP-driven biomolecular motors (see [104]). Here, the source of non-equilibrium energy is none other than the gentle gradient of flux lines, thanks to a rock magnetic field (non-homogeneous).

### ***3.6 Kinetic Barriers; Records of Constraints via Hysteresis***

A major difference in the dynamics of life's processes lies in the shift of the role of thermodynamics from a directing force in regular chemical reactions to one of supporting the kinetics [129]. In fact, biology teems with examples of chemical reactions that are thermodynamically allowed but await help for going across the kinetic barrier—an intermediate state requiring energy of activation ( $E_a$ ), with the reaction rate primarily dictated by the Boltzmann factor ( $\exp(-E_a/kT)$ ). Catalytic enzymes bring down the barrier by enabling the appropriate relative positioning of

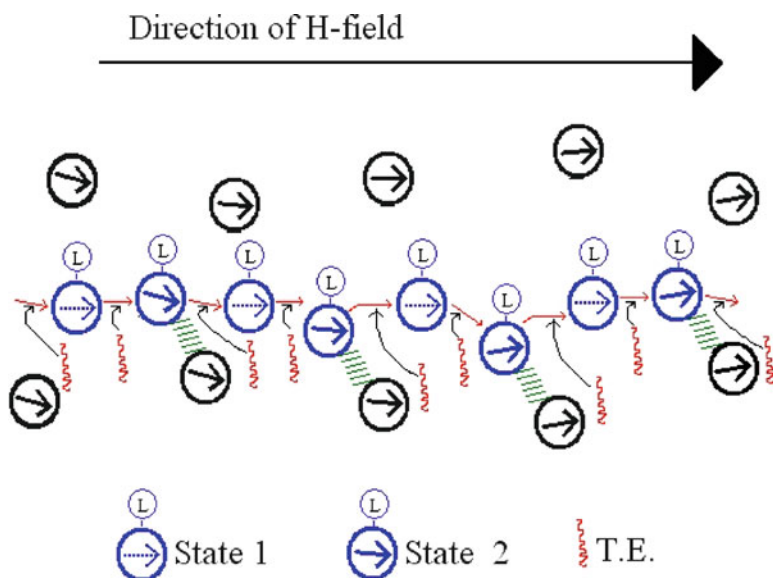
the reactants for reaction to occur. In the Hadean, rigid mineral crystals could have acted likewise, although it is difficult to see how entire metabolic cycles of disparate reactions could have been catalyzed on the same surface [115]. On the other hand, field-energy transfer through a network of magnetic templates within the structured phase [104] offers an alternative scenario for enabling the juxtaposition required for not only one but also an array of reactions, by harnessing thermal fluctuations to orient substrates diffusing into and binding to the templates (cf. [120] oscillator inspired catalytic mechanism for each reaction, see also Sect. 3.9).

Note that Pattee's perception of lifedynamics arising out of an irreducible "whole"-internal interpretation of time-independent symbolic codes (DNA) by their dynamical functional self-expressed constraints (proteins) – neatly subsumes the debate of which branch of life-the metabolic or the replicator-first made its appearance in the origins. Briefly, it may be recalled that constraints create specific conditions for execution of physical laws in the dynamical system they cause their local action, thanks to frozen degrees of freedom in their material structures. Their formation, in turn, depends on records or memory-like preserved constraining configurations, e.g., the dislocation of a growing crystal. And, although these do not form as a consequence of the dynamics of the system in which they function (giving them an elevated "status"), they can govern some dynamical events, by switching on-off in a specific manner. In one scenario of such "entangled" emergence of symbols and metabolism – a "protometabolic" system – where the information specifying the network is distributed in its organization (a membrane-enclosed recursive network of component production) evolves to a self-interpreted genome through a stage dependent on *non-symbolic* records. This is crucially dependent on the latter's ability to act at two levels: as a memory to be expressed and as a way to express this memory [41]. Now, the phenomenon of hysteresis in magnetic materials provides a natural mechanism for the emergence of constraints in a magnetically ordered system. For example, for reactions catalyzed on the magnetic templates [104] as above, the imprint of the bound product in terms of altered orientations of the template particles would itself provide an "observing" mechanism for "recording" (the product of) the reaction.

### ***3.7 Self-Reproduction; Pre-Bio-Molecular Motors***

Not only genetic information but also entire progeny are modeled on the "parent template" that provides the precise spatial information for element organization and patterns, at different levels of the intricately connected hierarchy [52]. Living systems use diverse modes for copying patterns of information : nucleic acids follow a template basis for assembly, membranes grow by extension of existing ones, entire structures can duplicate (e.g., spindle pole body or the dividing cell as a whole), etc. [52]. Now, in contrast to the growth of mineral crystals (in traditional origin-of-life models) restricted to a growing surface, field-assisted alignment of diffusing tiny particles would occur through the "layers" of the accreted assembly, leading to its inflation.

Next, for ratchet-like effects, consider super-paramagnetic ligand-bound particles [100, 101], diffusing through the structured phase in an oriented manner as a consequence of gentle change in flux lines, assuming that magnetic rocks would have provided a non-homogeneous field. Now, two changes are expected upon ligand binding: lowering of both rotational freedom and coercivity [165] on the ligand-bound end. Thus, while unconstrained rotation of ligand-free particles enables alignment and propagation of the “information” in the magnetic dipole-ordered assembly (“reproduction” as above), ligand-binding aids diffusive passage. The constituents of the structured phase- magnetically networked dipoles - are expected to locally perturb the H-field “seen” by the aligning and diffusing particles, moving through its layers-the “templates” (Figure 1). Thus, alignment to consequent template-partners would be alternated by dissociation from the template, in cycles. Infinitesimal steps leading to these altered states would require  $\sim k_B T$ , hence could be facilitated by Brownian hits. This way the main features of today’s biological molecular motors: a non-equilibrium force applied close-to-equilibrium that could reign-in Brownian noise, plus asymmetry (via an H-field gradient), can be recovered [104], since no macroscopic thermal gradient runs these engines. Recall that a “thermal gradient” was proposed by Feynman to circumvent the idea of “biased” Brownian motion (based on structural anisotropy alone) which, despite a right magnitude for driving nano-sized particles [125] is otherwise forbidden by the Second Law of Thermodynamics. The evolution of these motors can perhaps be understood



**Figure 1** Directed interactive diffusion of S-PP through MSP (with parallel correlations). MSP represented in black; State 1/ State 2: lower/higher template-affinity states of the ligand (L)-bound S-PP, in blue; green lines signify alignment in State 2; T.E. or thermal energy from bath; rock H-field direction indicated on top of figure, see text

in terms of non-magnetic “replacements” allowing the exit of such a magnetic system from its geological confines [104]. Indeed, diffusing superparamagnetic units through a viscous medium (due to interparticle magnetic dipolar forces) have a striking parallel to the directed movement of biomolecular motors (in the translational, transcriptional, cytoskeletal assemblies) on aperiodic intracellular surfaces that indicate an invariant topological theme for a ratchet mechanism, namely, movement of a cargo loaded element on a template (representing a varying potential) that harvests thermal fluctuations for dissociating its bound state and spends energy for conformationally controlled directed binding, or an ionic gradient for direction [7].

### 3.8 *Pre-RNA World; Transfer Reactions; Optical Activity*

Both magnetic templates and the particles (free or chemical ligated) diffusing through the phase are part of a magnetically connected network, and therefore seem to have the potential to naturally provide topological correspondences to a variety of biophenomena. For example, in the proposed RNA world, RNA played the roles of *both* DNA and protein – let us call them RNA-sequential and RNA-structural, respectively. Evidently, nature designed DNA for packaging information efficiently, satisfying Shannon’s maximum entropy requirement (no correlations across sequences). This leads to the “chicken-egg” conundrum, as the largely random sequential information encoded in DNA is correlated via RNA with the high degree of stereochemical information in proteins. Now in contrast to hard periodic crystal lattices forged with chemical bonds, confining physical forces in an accreted ensemble gives a natural access to aperiodic surfaces [17], (see Sect. 2.2). We therefore point out that RNA-sequential has obvious parallels with aperiodic layers of a magnetically structured phase hosting directed diffusion of ligand-bound super-paramagnetic particles (above). These very “templates” seem like a primitive translational machinery, where Wächtershäuser’s [167] “bucket brigade-like” transfer reactions carried out by oriented particles play the key adaptor roles *a la* transfer RNAs – the directed diffusion of the particle on an aperiodically packed surface with no correlations (RNA-sequential-like), with the other, ligand-bound to compounds rich in structural information. This “magnetic letters-like” scenario bears a striking resemblance to the tRNA’s bringing the amino acids together for stringing them up on the basis of the sequential information inscribed in the mRNA template. And the maintenance of similar orientation, during diffusive migration (depending upon the gradient of flux lines cutting through the magnetically structured phase, i.e., forward/backward from N to S or S to N; see above) offers a natural mechanism for generating optical activity through symmetry-breaking. This is because the solid-phase-like *arrangement* of ligands, from a racemic mixture (and bound to diffusing-super-paramagnetic particles oriented to the magnetic-rock field) would take place in the *limited* space between densely packed magnetic layers/templates (cf. [95, 166]). And, in the transfer reactions, this directional asymmetry of transport of an oriented dipole due to a non-homogeneous external field has the potential to

push the balance in favor of bond formation between juxtaposed activated units having the same chirality close to the ligand-binding site. This is further aided by the space constraints of such intralayer activity, where the optical activity of the first-bound unit (the symmetry-breaking choice) would set the preferences for those of the subsequently selected ones.

### 3.9 *The Potential for a Quantum-Leap to Life*

These non-trivial correspondences between biological and magnetic phenomena in general and topological correspondences to our proposal in particular, prompt us to push this interface between these apparently unrelated disciplines, to wonder why the functional-approach-based selection of chemical molecules (where changes are largely due to environmental fluctuations) would not have started from a magnetic scaffold defining and dictating these functional/contextual requirements? Indeed, the orientation of each (particle) moment can be viewed as an interpreting gauge of its composite environment-external field (rocks); neighboring particle moments; thermal fluctuations. It offers a “route” for capturing a “stable internal symbolic representation of the environment” to borrow a phrase from Hoffmeyer [60]. So could there have been a possible role of magnetism in endowing a system with constraints, non-creativity, no goals, with the potential to jump to a state with formal processes of controls, learning and instructions, creativity (as in the extended version of Pattee’s work drawn by Abel [1] – life as a bonafide natural programmer), thus empowering the initial conditions for this leap? In this connection, it may be recalled that using the metaphor of an arch of stones, Cairns–Smith had proposed that the scaffold paving the way for “organic takeover” (the “arch”) may well have been provided by clay minerals that were eventually disposed off. Indeed, this idea finds a sort of echo in the suggestion of Patel [121], viz., the choice of carbon with its tetrahedral geometry provide the simplest discretization of the fundamental operations of translation and rotation needed for processing structural information. (Rotations in 3-D are *not* commutative, a fact of crucial importance in representing structural information; in mathematical jargon this goes by the name of the SU(2) group of Pauli matrices/quaternions). Of course “replacements” via quantum searches could well have been biopolymers with capacity for classical searches that would have been more robust against decoherence (cf. the classical wave algorithm proposed by Patel [120]). According to Patel, vibrations and rotations of molecules being harmonic oscillator modes, the catalyst like a mega oscillator can focus the energy of many modes onto the reactant awaiting activation.

This brings us to an important feature accessible via magnetism, viz., a sound entry point for quantum processing. The Matsuno group (2001) has reported the coherent alignment of induced magnetic dipoles in ATP-activated actomyosin complexes that was maintained over the entire filament even in the presence of thermal agitations causing rapid decoherence. The energy of the dipole-dipole interaction per monomeric unit of  $1.1 \times 10^{-22}$  Joule was found to be far below the thermal

energy per degree of freedom at room temperature. This also can be extended to magnetically aligned particles in a natural way. Work is currently in progress regarding the role of a magnetic environment in aiding coherence. This matches with Abel's [2] observation, "... an inanimate environment has no ability to program for a *potential* function that does not yet exist. Yet selection for potential function is exactly what genetic programming requires." Thus, Abel projects life as a bona fide programming system with discretized instructions. Now, the infinitesimal orientational changes of particles (associated moments) diffusing through the layers of the assembly [104] offer yet another occasion for discretization of operations required for processing structural information, e.g., choice of carbon polymers (see above). Indeed, the implications of a ferrofluid network as an analog device can be seen in the recent simulations by the Korenivski group [12, 117]. We therefore suggest that these magnetic nano-particle assemblies could have been the *soft-magnetic-matter* version of Cairns-Smith's mineral scaffold that was replaced by organic matter.

Again, one can find an example of discretization in the biological currency ATP, providing energy for coupling to biochemical reactions. Furthermore, in what is seen as a temperature lowering mechanism enabling molecular motors to act as heat engines, Matsuno and Paton [93] describe the gradual release of energy stored in ATP by actomyosin ATPase, in a sequence of quanta  $E_m$  over time intervals of  $\Delta t_m$ . This underlies the huge order of magnitude discrepancy between the observed time interval of hydrolysis of 1 molecule of ATP  $\sim 10^{-2}$  s, and that calculated by considering energy release of  $E = 5 \times 10^{-3}$  erg (7 kcal/mol) from a singly emitted quantum, or  $\hbar/E \approx 2 \times 10^{-15}$ . The obtained values of  $E_m \sim 2.2 \times 10^{-19}$  erg and  $\Delta t_m \approx 4.5 \times 10^{-9}$  s indicates therefore  $2.2 \times 10^6$  number of coherent energy quanta release during one cycle of energy release from a single ATP molecule. In Kelvin scale, each energy quantum  $E_m$  amounts to  $1.6 \times 10^{-3}$  K associated with the actomyosin complex. Here too we find that a mechanism enabling interchange between the a system's environmental temperature and its own entropy is provided by the (anisotropic) magnetocaloric effect (MCE) [159], which is the property of some magnetic materials to heat up when placed in an H-field and cool down when they are removed (adiabatic). In fact, the heat capacity at the nano-scale turns out to be a few-fold higher than that of bulk systems, thanks to MCE [78]. We have suggested [104] that the exit of the "magnetic ancestor" from the confines of its magnetic environment may have been enabled upon coupling of its envisaged dynamics associated with changes in gradient of flux lines, instead with ATP-the universal biological currency (see Sect. 5.4; also Sects. 2.4, 3.5, 3.7).

In this scenario, biological phenomena with similarity to magnetic ones could be considered as "distant cousins" of their "non-living" counterparts. Thus even quantum processing is viewed as a legacy and not a product of adaptive evolution [34]. Note that magnetic ordering may stem from unpaired p – electron systems [113] (not just 3d, 4f!). The "substitutes," despite increasing complexity, would need to pass on the legacy of multidimensional properties of the Ancestor possessed, especially phase information, e.g., DNA has positional information, with possible phase signatures in its helical structure [79].



## 4 Framboids and the Mineral Greigite

We shall now seek to expand the potential of mineral crystal theories by looking for minerals that can enable magnetic effects, such as those outlined above. This brings us to framboids ([171], see below) as these dynamically ordered terrestrial/extraterrestrial, microcrystal composites formed by structurally *different* materials show the control of packing by *physical* forces.

### 4.1 Framboids; Importance of Physical Properties

In framboids, named after their framboise/raspberry-like patterns, nucleation of clusters is followed by growth of individual nuclei into microcrystals. They have been defined as microscopic spheroidal to sub-spheroidal clusters of equidimensional and equimorphic microcrystals which suggest a homogenous nucleation of the initial microcrystals. Other than the spherical framboids, a highly ordered icosahedral type has been reported where this packing is maintained in its internal structure. The formational environment is evidently critical for the packing in these varied forms. As pointed out by Ohfuji and Akai [111],  $D/d$  ratios of framboids (framboid diameter  $D$  and microcrystal diameter  $d$ ) dominated by irregular or loosely packed cubic-cuboidal microcrystals are low compared to high corresponding values observed for those composed of ordered densely packed octahedral microcrystals. The narrow distribution of sizes and uniform growth of thousands of crystals in framboids within a short time interval was attributed to a regulated balance between rates of nucleation and of crystal growth, as in the La Mer and Dinegar model [82]. Furthermore, the nucleation of a supersaturated solution by the first-formed crystal triggers the separation of many crystals of the same size. This liquid–solid-like phase transition is dependent on packing considerations of hard-sphere-like microcrystals, whose ordering is an outcome of the interplay of close-packing and repulsive forces (see [148]).

As noted by Sawlowicz [148], the framboidal texture is seen in a number of different minerals other than pyrite, i.e., copper and zinc sulfides, greigite, magnetite, magnesioferrite, hematite, goethite, garnet, dolomite, opal, and even in phosphoric derivatives of allophane. This suggests a similar mechanism of formation, despite the structural differences. Studying their presence in sedimentary environments, Sawlowicz [147] found pyrite framboids to be hierarchially structured over three size-scales: microframboids, to framboids, to polyframboids. And since spheroidal microframboids are formed of equant nanocrystals, he suggested (1993, 2000) the formation of nano-framboids, comprising microcluster aggregations ( $\sim 100$  atoms), by analogy with the 3-scale framboidal hierarchy. His observations leading to a proposed formation mechanism center around the key role of the colloid-gel phase leading to the fractal forms. Interestingly, exclusion of organic compounds, were found to lead to simple framboids via an aggregation mechanism while experiments with organic substance stabilized gel-droplets, framboids formed by *particulation*.



This latter route is seen as important for generating the *fractal* complexity. Similar scale free frambooids of greigite that is ferrimagnetic (next), have also been documented [128].

## 4.2 *Framboidal Greigite*

In frambooids reported in sedimentary rocks more than 11,000 years old [134], the central parts of the weakly magnetized frambooids were found to have greigite microcrystals. Sections from these show that the pentagonal arrangement comprise a central pentagonal domain with its sides connected to five rectangular/trapezoid-like regions which are in turn connected via fan-shaped domains. The arrangement pattern of these densely packed octahedral microcrystals linked edge to edge is “lattice-like” (space filled) in the rectangular domains, whereas in the triangular domains the triangles are formed by the (111) faces of the octahedral microcrystals and the voids between them. Thus within these domains the individual faces of the microcrystals do not make any contact. The icosahedral form is seen as generated by stacking twenty tetrahedral sectors packed on three faces out of four, and connected by their apexes at the center. Generally acknowledged as dynamically stable, this form is known to have six 5-fold axes at each apex, and ten 3-fold axes at each face, as can be seen in a number of naturally occurring structures from microclusters like fullerene to some viruses [111]. Furthermore, in an investigation of apparent biologically induced mineralization by symbiotically associating bacterial and archaeal species, frambooidal greigites have been obtained from Black Sea sediments that are ordered clusters of octahedral crystals comprising  $Fe_3S_4$ -spinel (Essentially cubic where sulfur forms a fcc lattice with 32 atoms in the unit cell, and Fe occupies 1/8 of the tetrahedral and 1/2 of the octahedral sites). Their size is restrained by their icosahedral symmetry and under greater pressures at depths of 200 m, the diameters are mostly  $\sim (2.1, 4.2, 6.3 \text{ or } 8.4) \mu$ , with the two intermediate ones predominating. The smallest of these are formed from 20 octahedral crystals ( $0.35 \mu$ ) positioned at the apexes of an icosahedron and surrounding a  $0.5 \mu$  diameter vacancy that give rise to 12 pentagonal depressions on the outside. Nested structures building up from this smallest one lead to the higher sized clusters [128]. Subspheroidal pyrite-frambooids, due to curved polyhedron-like outer facets, probably reflect an internal icosahedral microcrystal organization [111], which are classically forbidden crystallographic symmetries [112].

## 4.3 *Magnetic Interactions*

Magnetic interactions turned out to have an overwhelming influence when Wilkin and Barnes [171] included them in the standard DLVO treatment for interacting colloidal particles that considers attractive van der Waals and double-layer repulsive

interactions, for modeling framboidal pyrite formation. This is based on the alignment of precursor greigite, under the influence of the weak geomagnetic field that would help overcome the thermal energy of particles above a critical size. Ferrimagnetic greigite has a saturation magnetization value  $M_{sat}$  at 298 K ranging between 110 and 130 kA/m. On the basis of microscopic observations by Hoffmann [59] of natural greigite crystals,  $< \mu$  meter – sized greigite can be roughly taken as single-domain particles. Assuming a spherical geometry, the critical grain diameter of constituent crystallites comprising the framboid interior  $d_c = 2a$ , where  $a > 1$ , is given by

$$d_c = (6k_B T / \mu_0 \pi M_{sat} |H|)^{1/3} \quad (1)$$

This result can be obtained from the inequality  $W_{WB} > k_B T$  where we define  $W_{WB} \equiv \mu_0 M_{sat} V H$ . Here  $k_B$  is Boltzmann's constant and  $\mu_0$  the permeability of vacuum. When aligned parallel to weak geomagnetic field ( $\sim 70 \mu\text{T}$ ),  $d_c = 0.1 \mu\text{m}$ . Although framboids can form in varied environments and by other mechanisms (see [112, 148]), this magnetic greigite-precursor mechanism can operate only up to temperatures of 200°C [171], e.g., sediments, in natural waters. Also, as pointed out by Wilkin and Barnes [171], the effect of weak fields leads to spherical structures in ferrofluids (Sect. 2.3) in contrast to aspect ratios approaching infinity in strong fields. They also noted the role of turbulence in facilitating the interplay of opposing interactions.

#### 4.4 Dynamic Ordering; Phyllotaxis; Quasiperiodicity

A characteristic pattern of icosahedral framboids – octahedral microcrystals, large D/d ratio – has been attributed to a high initial nucleation rate and low growth rate of microcrystals [111, 112]. According to Sawlowicz [148], the interplay of surface-minimizing forces with repulsive interactions lead to close-packed framboids, tending to polyhedrons. And, this is a ramification of anastrophic supramolecular organization, with its far-from-equilibrium conditions. Sure enough, the framboid morphology is strongly reminiscent of the ubiquitous phenomena of Phyllotaxis, from subnano to cosmological scales [3, 36, 84]: Repulsive magnetic dipoles, galactic structures, biostructures, from the molecular (proteins, DNA) to macroscopic levels (myriad marine forms), proportions in morphological and branching patterns [36], Benard convection cells, stress-driven self-assembly, bunched crystalline ion beams, atmospheric flows, and flux lattices in layered superconductors. Phyllotactic patterns are produced when the sequential accretion/deposition or appearance/growth of elements is governed by an energy-minimized optimization of the main opposing forces: largest available space vs repulsive interactions. And in magnetically accreted greigite framboids [171] too, a similar interplay of conflicting forces, leads to raspberry-like phyllotactic patterns. This dynamic ordering via accretion of magnetic crystals in the face of short-range repulsive forces does contrast with the build-up of a conventional infinite crystalline lattice, where the nuclear surface

acts as a template for copying a unit cell via local interactions. Rather, it is analogous to a scenario at nano-scales—one associated with the aperiodic, long range order of systems known to form *quasicrystals* whose growth occurs by accretion of pre-formed clusters in the liquid state by the growing nucleus [70]. Now, the relevance to greigite concerns its natural preference for such order as evidenced from observations of nested scale-free icosahedral greigite frambooids [128]. These observations are intriguing in view of the known links between phyllotactic patterns and quasiperiodic phases. For instance, the predominance of edge-to-edge contacts between microcrystals comprising icosahedral greigite frambooids [111, 134] limits possible *conduction* pathways.

#### 4.5 *Magnetic Assemblies in the Laboratory; Long-Range Order?*

Some insights into the above natural assemblies are offered by synthetic ones driven via a different route of evaporation [4, 26, 156], also one under hydrothermal conditions [174]. Apart from external-field control, other physical properties of nano-constituents: crystalline/colloidal state, geometry, susceptibility, coatings, etc, are important criteria for clustering patterns [81, 126]. Next, in soft condensed matter studies, varied and unusual polyhedra have been seen in packing sequences of colloidal polystyrene microspheres, illustrating how certain symmetries, including fivefold rotational symmetry, can arise solely from compression and packing constraints. These can be explained by the use of a minimization principle – that of the second moment of mass distribution wrt the center of mass ( $\sum_i m_i x_i^2$ ), instead of the conventional volume ( $\sim r^3$ ) – optimizes the packing [89, review in 174]. Again, the *route* to formation is another important aspect of assembly; there is no possibility of an internal sphere upon collapse in this evaporation-driven system that starts from spherically packed particles bound to a continuous and smooth (2D) surface, i.e., the droplet interface. This route would not apply to particles compressed via magnetic dipolar forces as in scale-free greigite frambooids, which is more like a problem of packing spheres not only on the surface of a sphere (2d-space), but also rather *into* a finite 3D space, as in some compounds, alloys, quasicrystals that have long range order without periodicity. Recall that frambooidal texture comes via optimized packing of microcrystals (see large  $D/d$  ratios, Sect. 4.1). That structurally different materials form frambooids (Sect. 4.1) also reveal the important role of the colloidal state where physical properties can be accessed, in contrast to the strong influence of chemical properties for packing in (periodic) crystals. An understanding of icosahedral geometry in scale-free greigite frambooids can be had from a study of tessellation of spheres (number  $N$ ; radius  $a$ ) packed on the surface of a large sphere (radius  $R$ ). This shows that energy minimization would lead to buckling into icosahedral forms, considering only small  $R/a$  ratios, as  $N \sim (R/a)^2$  [107]. This in turn could bring in geometrical frustrations but studies on icosahedral magnetic quasicrystals [86] show that geometrical constraints do *not* rule out the possibility of long range magnetic order.

Thus, we find that in the mineral inorganic world too, superimposed physical interactions can dictate assembly organization. Furthermore, it is significant that greigite, which is known to undergo accretion due to magnetic forces [171] and also has a natural propensity for framboid formation [111, nested forms in 128], is also strongly suspected for its “metabolic” potential (next).

## 5 Mound Scenario of Russell et al. and Greigite

In fact, the search for greigite forming on the Hadean Ocean floor led us to the colloidal environment setting of Russell and coworkers where greigite forms across gradients and that leads to a metabolically enriched scenario (next).

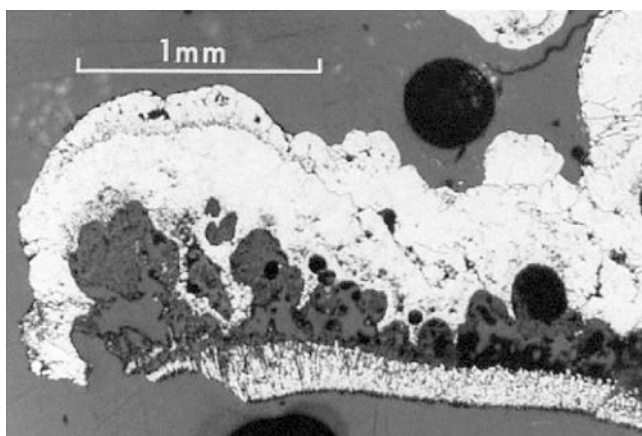
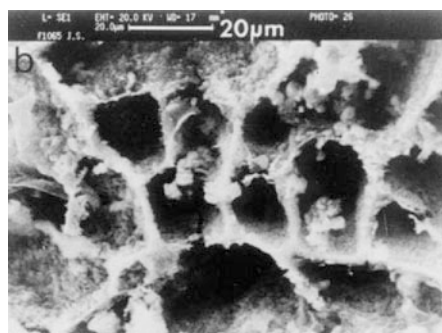
### 5.1 Mound Scenario of Russell et al.

The colloidal environment-based proposal of Russell et al. [138] envisages Life as having emerged in moderate temperature hydrothermal systems, such as mild alkaline seepage springs. Water percolating down through cracks in the hot ocean crusts reacted exothermically with ferrous iron minerals, and returned in convective up-drafts infused with  $\text{H}_2$ ,  $\text{NH}_3$ ,  $\text{HCOO}^-$ ,  $\text{HS}^-$ ,  $\text{CH}_3^-$ ; this fluid ( $\text{pH} \sim 10 \leq 120^\circ\text{C}$ ) exhaled into  $\text{CO}_2$ ,  $\text{Fe}^{2+}$  bearing ocean waters ( $\text{pH} \sim 5.5 \leq 20^\circ\text{C}$ ) [137]. The interface evolved gradually from a colloidal FeS barrier to a single membrane and thence to more precipitating barriers of FeS gel membranes. Since fluids in alkaline hydrothermal environments contain very little hydrogen sulfide, the entry of bisulfide, likely to have been carried in alkaline solution on occasions where the solution met sulfides at depth [142], was controlled. This was perhaps important for a gradual build-up of scale-free clusters leading to the envisaged gel-environment. (As pointed out by Sawlowicz [148] colloids often form more readily in dilute solutions – suspension as a sol – than in concentrated ones where heavy precipitates are likely to form). These barriers controlled the meeting of the two fluids, as they enclosed bubbles entrapping the alkaline exhalate : an aggregate growing by hydrodynamic inflation. The forced entry of buoyant seeps may have led to chimney-like protrusions. Furthermore, theoretical studies by Russell and Hall [141] show the potential of the alkaline hydrothermal solution (expected to flow for at least 30,000 years) for dissolving sulfhydryl ions from sulfides in the ocean crust. The reaction of these with ferrous iron in the acidulous Hadean ocean (derived from very hot springs, [141]) is seen as having drawn a secondary ocean current with the  $\text{Fe}^{2+}$  toward the alkaline spring as a result of entrainment [91]. Hence at the growing front of the mound, the production of daughter bubbles by budding would have been sustained by a constant supply of newly precipitated FeS. Like cells, these mini FeS compartments protected and concentrated the spectrum of energy-rich molecules, borne out by harnessing important gradients across the mound (a true far-from-equilibrium system, driven by energy released from geodynamic sources): redox,

pH, and thermal gradients for electron transfers, primitive metabolism, and directed diffusion, respectively [137]. See also Rickard and Luther [137] for an analysis of the reducing power of FeS for synthesizing organics in this proposed scenario.

Experimental simulations of mound conditions using calculated concentrations of ferrous iron and sulfide (20 mmoles of each) resulted in the formation of a simple membrane. Using solutions with 5 to 20-fold greater concentrations (to make up for their build-up in geological time) generated compartmentalized structures, shown in Figure 2 where the chambers and walls are  $\sim 20$  and  $5 \mu$ , respectively. These have remarkable similarities to porous ones in retrieved Irish orebodies, shown in Figure 3, which had originally inspired the idea that the first compartments involved in the emergence of life were of comparable structure (see [136, 139]). In fact, even submarine mounds seen today are invariably porous [69, 90]. Also, the sulfide comprising what is now pyrite ( $\text{FeS}_2$ ) in the 350 million-year-old submarine Irish deposits (Figure 3) was derived through bacterial sulfate reduction in some-

**Figure 2** FeS compartments. SEM photo of a freeze dried section showing FeS compartments formed on injecting 0.5 M  $\text{Na}_2\text{S}$  solution into 0.5 M of  $\text{FeCl}_2$  [139]. Reproduced with kind permission from M.J. Russell



**Figure 3** FeS botryoids. Polished cross-section of the Tynagh iron sulfide botryoids. Kindly provided by M.J. Russell, see text for details

what alkaline and saline seawater while the iron was contributed by exhaling acidic hydrothermal solutions. On mixing, mackinawite ( $\text{Fe}(\text{Ni})\text{S}$ ) and greigite ( $\text{Fe}_5\text{NiS}_8$ ) would have precipitated to form inorganic membranes at the interface [138, 139].

## 5.2 Greigite Formation from FeS

Figure 2 shows laboratory simulated FeS compartments; the chambers and walls are  $\sim 20$  and  $5 \mu$ , respectively. According to Russell et al. [145], the permeable membranes likely comprise (ferredoxin-like) greigite and mackinawite, and whose metal and sulfide layers work for and against  $e^-$  conduction, respectively. An insight into this calls for a brief outline of iron sulfide transformations under wet and moderate temperature conditions. Amorphous mackinawite ( $\text{FeS}_{(am)}$ ) is the first FeS phase formed from aqueous S(-II) and Fe(II) at ambient temperatures, apparently via two competing pathways governing the relative proportions of the two end-member phase mixture. The long-range ordered phase with bigger crystalline domain size and more compact lattice increases at the cost of sheet-like precipitated aqueous FeS clusters [173].

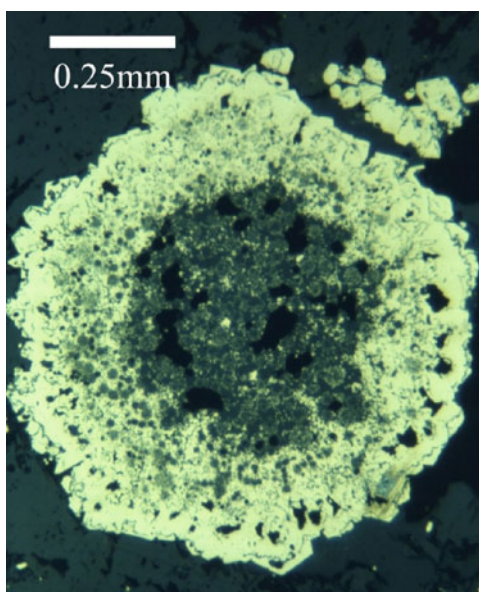
Note that an FeS cluster can display two properties: (1) it can be regarded as a multinuclear complex (where instead of a central atom, as in a complex, a system of bonds connects each atom directly to its neighbors in the polyhedron); and (2) as an embryo since it can develop to form the nucleus of the first condensed phase [132]. The formation of the latter gets initiated by statistical fluctuations in the density of the initial parent phase (e.g., due to supersaturation) and its growth is favored by the difference in chemical potentials between the parent and the new phase. Reviewing aqueous FeS clusters in water environments, Rickard and Morse [132] suggested the enhanced stability of some stoichiometries—stable magic number clusters – from among the apparent continuum of stoichiometries of aqueous FeS clusters. This ranges from  $\text{Fe}_2\text{S}_2$  to  $\text{Fe}_{150}\text{S}_{150}$ , where the first condensed phase ( $\text{FeS}_m$ , mackinawite) appears with a size and volume of 2 nm and  $10 \text{ nm}^3$ , respectively. Although molecular  $\text{Fe}_2\text{S}_2$  is similar in structure to crystalline mackinawite, the Fe–Fe bond lengths and Fe–S–Fe bond angles are seen to approach those of crystalline mackinawite, in tandem with increased size of molecular FeS clusters. The decrease in degree of softness, or water loss, can be gauged from the relative density increase over the smallest  $\text{Fe}_2\text{S}_2$  cluster ( $\geq 10^6$ ), as the structure of hydrated clusters is believed to determine that of the first condensed phase. X-ray diffraction of the first nano-precipitate shows a (lattice expanded) tetragonal mackinawite structure. That the data fit well with other independent estimates is ascribed to the plate-like form of  $\text{FeS}_m$ . The quick transformation of disordered mackinawite to the ordered form is followed by solid state transformation to the more stable but structurally congruent greigite, with a 12% decrease in volume, involving a rearrangement of Fe atoms in a close-packed, cubic array of S atoms. Furthermore, trace amounts of aldehydes are believed to bind to the  $\text{FeS}_{(am)}$  surface, initiating Fe(II) oxidation (S(-II) unaffected); they also prevent the dissolution reaction,  $\text{FeS}_{(am)}$  to  $\text{FeS}_{(aq)}$  (aqueous FeS complex), crucial for pyrite formation (in absence of aldehyde, S(-II) oxidized,



Fe(II) unchanged), thus assisting in greigite formation at the cost of pyrite (perhaps as in bacteria) [130]. Such a solid-state transformation of amorphous mackinawite to greigite can be extended to FeS clusters – Rickard and Luther0 [131] suggest the possibility of organic ligands stabilizing aqueous Fe(III)-bearing sulfide clusters, as seen in similar (greigite-like) cubane forms in FeS proteins. Importantly, FeS membranes formed in the laboratory show a 20 to 40-fold increased durability on adding abiogenic organics. Diffusion controlled reactions would slow down with thickening of aging/hardening of membranes [138].

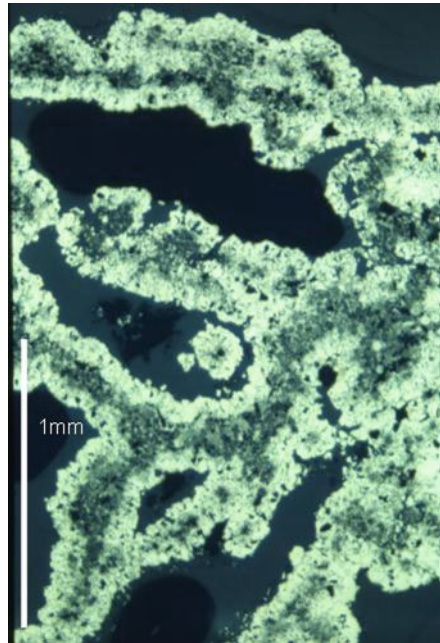
### 5.3 The FeS Gel Environment and Framboids

As noted by Russell et al. [138], citing Kopelman [77], gels lie between liquid and solid states with self-similar clusters, fractal on all scales (permitting diffusion control in heterogenous reactions, ubiquitous in biosystems). They suggested [143, 144] the nucleation of the FeS gel bubbles by iron sulfide: in vitro simulations of iron sulfide chimneys demonstrated formation of macroscopic spherical shells 1–20 mm across, while on a microscopic scale spherical, ordered aggregates of *framboidal pyrite* about 5 micro meter in diameter were found in fossil hydrothermal chimneys (see Figures 4a and 4b; [15, 16, 83]) that seemed to have grown inorganically from the spherical shells of FeS gel. These framboidal sacks of periodic arrays within the extensive reactive surfaces per unit volume of the chimneys could have offered ideal experimental culture chambers and flow reactors well poised for origin-of-life experiments [143]. Indeed, framboids have long been recognized for their fascinating



**Figure 4a** Framboids in chimneys, small vent. Small pyrite vent structure: Reflected ore microscopy of transverse section shows a central area of empty black spaces plus (grey) fine framboidal pyrite, and a fine euhedral authigenic rim surrounded by baryte, with minor pyrite

**Figure 4b** Framboids in chimneys, sheaves. Sheaf system, formed from coalescing rods of anastomosing microcrystalline pyrite. Black areas are empty spaces; central regions are framboidal pyrite with an exterior of crystalline pyrite. (Labeled pictures given by Dr. Adrian Boyce are reproduced with his kind permission; Source: Boyce et al. [15]; Boyce [16]: Exhalation, sedimentation and sulphur isotope geochemistry of the Silvermines Zn+Pb+Ba deposits, County Tipperary, Ireland; Boyce, Unpublished Ph.D. thesis, University of Strathclyde, Glasgow)



features, prompting speculations on their possible role in the origin of life, e.g., Sawlowicz [148] noted the bio-potential of constituent microcrystal surfaces, presence of catalytic metals, fractal structures, to name some.

The above observations of magnetically accreted framboidal greigite (Sect. 4.2) and possibility of framboid assembly in colloidal environment lead us to think that superparamagnetic greigite could have formed magnetic assemblies (in the presence of magnetic rocks) as starting self-reproducing systems, besides being a precursor for nucleic acids, proteins, lipids, etc., that could have been chosen as context-based replacements. This could be significant, as it has long been recognized that much of the path sketched from prebiotic chemistry to the RNA world (a widely accepted hypothesis; see [116]) remains uncharted and for start points (see [152]), there are suggestions of “physically” self-reproducing systems as having preceded “chemically-copying” self-replicators [38]; autocatalytic reactions [68] and self-replicating inorganic [20] or even a combination of organic and inorganic [114] systems.

#### ***5.4 Field Estimate from W-B Model; Motor-Like Dynamics***

We now come to the possibility of magnetic rocks which could further expand the potential of the mound scenario as described above. The associated H-field with rocks, needed for overcoming temperatures  $\sim 50\text{C}$  in the mound, is estimated by



extrapolating the Wilkin and Barnes (W-B) model (1997) for formation of framboidal pyrite via the precursor greigite. When aligned parallel to weak geomagnetic field ( $\sim 70 \mu\text{T}$ ), it gives  $d_c = 0.1 \mu\text{m}$  (see Sect. 4.3). Thus, a rock H-field for accreting 10nm sized particles would have to be 1,000-fold higher. This also is of the same order of magnitude  $\sim 10 \text{ mT}$ , seen for magnetite-based ferrofluids [110]. The saturation magnetization of magnetite ( $M_s = 4.46 \times 10^5 \text{ A/m}$ ) is about 3.5 times greater than that of greigite; from this one expects proportionate values for the fluid susceptibility of a corresponding greigite suspension, building up slowly in the ocean waters (see above). Also, the dipole–dipole interactions between negatively charged greigite particles (as the pH is well above 3 under mound conditions [171]) is likely to be aided by the screening effect due to ionic strength of natural waters [154].

Now, as the geomagnetic field did not even exist at  $\sim 4.1\text{--}4.2 \text{ Ga}$  [55] (whereas life is thought to have initiated at  $\sim 4.2\text{--}4.3 \text{ Ga}$  [140, 141]), we look at local sources for providing a magnetic field  $\sim 50\text{--}100 \text{ mT}$  for enabling accretion of newly forming greigite particles. (For example, the present geomagnetic field strength is too weak to explain the magnetization mechanism of lodestones). To that end, a plausible candidate (cf. [169]) could be isothermal remnant magnetism (acquired by lightning, impact, etc) in say, meteoritic matter on its way to the Ocean floor. In fact, Ostro and Russell have suggested plausible mechanisms for accumulation of reducing meteoritic matter, around the base of the mound. Also, unlike today's conditions, the primitive crust was still extremely reducing when life is thought to have emerged [133, 146], making the presence of ferromagnetic matter a likely event. Further reinforcement of the local H-field would occur through the generation of magnetic minerals like magnetite and awaruite [13, 37, 149] serpentinization of Ocean Floor peridotites (for more details see [104]).

Here, magnetic rocks could have helped not only the accretion of greigite particles, but also gentle changing flux due to non-homogeneous field lines (expected from rocks) could have gently moved incoming particles aligned to the field, i.e., in the same orientation in either the forward or the backward (N–S or S–N) directions, depending upon their position in the structured phase, and using thermal fluctuations to drive ratchet-like effects (see Sect. 3). At the same time, such a magnetic albeit locally confined ancestor, maintained close-to-equilibrium, would also have the potential for coupling with non-equilibrium energy sources (such as pH or redox gradient) - the “metabolic” wing of life-producing energy rich molecules [137]. This capacity of a magnetically controlled system to couple to different gradients, e.g., thermal [10], was also needed to pave the way for complex energy transduction mechanisms. We have suggested [104] that the “innovative evolution” of a bio-ratchet where coupling to non-equilibrium energy (in discrete packets) from energy-rich molecules propelled close-to-equilibrium dynamics (driven so far by a gentle H-field gradient), allowed the exit of the Ancestor from its geological location for seeking out gradient-rich niches elsewhere. This in turn would have led to a progressively decreasing functional dependence on iron sulfide. Nevertheless, the continued presence of magnetic elements (e.g., structural roles) would offer a magnetic basis for the association of its “liberated” replacements as in the multicellular life proposal [32]. The possibility of different “magnetic soups” close to the mound

also converges well with the suggestions of Martin and Russell [92], Koonin and Martin [76], of an initially confined universal ancestor diverging into replicating systems, located separately on a single submarine seepage site (see Sect. 5.1), en route to proto-branches of life. These reproducer-turned replicators could navigate to different openings where survival criteria would induce variations. The transfer of regulatory powers to the genes is likely to have been slow but progressive. In the pre-Mendelian era, there was more plasticity in phenotype – genotype mapping, gradually taking on a one-to-one basis with a decline in morphological plasticity – yet another “robustness” enhancing strategy [108].

### 5.5 *Enzyme Clusters and Natural Violarite Phases*

Note that the composition of iron sulfide clusters found in enzymes,  $Fe_5NiS_8$ , lie between  $FeNi_2S_4$  and  $Fe_3S_4$ . Although a solid solution in this range has not been observed in synthetic dry condition, high temperature experiments, it has been observed in natural violarite (iron-nickel thiospinel) phases [163]. More recently, the supergene oxidation of pentlandite ( $(Fe, Ni)_9S_8$ ) to violarite (includes extensions from  $Fe Ni_2S_4$  toward both  $Fe_3S_4$  and  $Ni_3S_4$ ) was experimentally reproduced under mild hydrothermal conditions [158]. The results show the *feasibility* of high iron/nickel ratios in violarite forming under reducing mound conditions, despite the suggested metastability of these compositions from bonding models. Iron is believed to occur as low spin  $Fe^{2+}$  in  $Fe Ni_2S_4$  that exhibits metallic, Pauli paramagnetic behavior. In contrast, the Mossbauer spectrum of  $Fe_3S_4$  is attributed to high-spin  $Fe^{3+}$  in tetrahedral A and octahedral B sites and its electronic structure from molecular orbital calculations [164] reveal localized 3d electrons with unpaired spins, coupled anti-ferromagnetically at lower temperatures. According to Vaughan and Craig [163], the greater ionic character and larger number of electrons in antibonding orbitals in  $Fe_3S_4$  relative to  $Fe Ni_2S_4$ , could contribute to the instability of intermediate compositions, despite their natural occurrence.

### 5.6 *Coherence: Ferromagnetic–Ferroelectric Effects*

The quest for co-existing (in same or locally different subspaces) ferroelectric effects reinforcing the coherent (“dispersive,” non-dissipative) effects of ferromagnetism arises out of interesting present-day biological observations. Frohlich [44,45] proposed the emergence of a long range coherent state via alignment of dipoles in cell membranes. Ordering of electric dipoles via interactions between structured water and the interior of microtubular cavities brings in a dynamic role of *ferroelectricity* as a frequency-dependent dielectric-constant  $\epsilon(\omega)$ , which gives a big dispersive (non-dissipative) interaction (robust against thermal losses) for small values of  $\omega$  (since the factor  $\epsilon(\omega)$  occurs in the denominator of the correspond-

ing interaction) [94]. Apart from the importance of such coherent electric dipole ordering alignment of actin monomers prior to ATP-activation, Hatori et al. [53] report the coherent alignment of magnetic dipoles induced along the filament, by the flow of protons released from ATP molecules during their hydrolysis (basically a Maxwell displacement current-like dynamical effect). But in contrast to the similar nature of magnetic ordering mechanisms conferring ferromagnetism via exchange interactions of predominantly localized magnetic moments, a variety of ferroelectric ordering mechanisms exist for different types of ferroelectrics, not all of which are well understood. In fact, in materials their co-existence can range from being mutually exclusive, such as due to incompatibility of d-electron criterion for magnetism with off-centering second-order Jahn–Teller effect, all the way to strongly coupled giant magneto-resistance effects (includes non-oxidic ferrimagnetic semiconductor thiospinels  $FeCr_2S_4$  and  $Fe_{0.5}Cu_{0.5}Cr_2S_4$ , that are  $Fe^{2+}$  and  $Fe^{3+}$  end members of solid solution  $Fe_{1-x}C_xCr_2S_4$  ( $0 < x < 0.5$ ) [118]. While lattice distortions with lowered symmetry reduce competing interactions [27, see also 43], an insight into the loss of inversion symmetry comes via the spin-orbit coupling mechanism which gives the electric polarization  $P$  ( $\sim \mathbf{e} \times \mathbf{Q}$ ), where  $\mathbf{e}$  is the spin rotation axis and  $\mathbf{Q}$  is the wave vector of a spiral) induced upon transition to a spiral spin-density-wave state triggered by magnetic frustrations [105]. Apart from the spin-orbit coupling factor, a reduction of crystal symmetry (Fd3m to non-centrosymmetric  $F\bar{4}3m$ ) in several spinel compounds, including  $FeCr_2S_4$  was attributed to a displacement of cations [25, 98]. Similar off-centering was also found in oxide spinels [25], e.g., magnetite  $Fe_3O_4$ . Additionally, a combination of site-centered (extra holes or electrons on metal sublattice, e.g.,  $Fe^{2+}$  and  $Fe^{3+}$ , where anions do not play a role) and bond-centered (the alternation of short and long bonds, in otherwise equivalent sites, lead to a bond-centered charge density wave) charge-ordering was suggested for explaining the multiferroic behavior of  $Fe_3O_4$  below the Verwey transition at 120 K [71]. The co-operative co-existence of ferroelectric and ferro-magnetic properties in these structural relatives of greigite – due to a subtle interplay between charge, spin, orbital, and lattice degrees of freedom [56] – raise the possibility of a similar profile for  $Fe_3S_4$  or close relatives found in enzymes, e.g.,  $Fe_5NiS_8$ , for which no direct evidence is so far available.

## 5.7 Preliminary Experimental Requirements

What is needed first is a robust model system to explore magnetic structure formation together with protocols for monitoring accompanying chemical reactions. Then, the presence of magnetic rocks in the mound, represented by a surface magnetic field strength (say, in the range 0–200 mT) needs to be checked for any magnetic structure formation in different concentrations of newly forming greigite suspension. Here, the dispersity of newly forming greigite clusters whose size range would be expected to closely resemble that of the FeS dispersion ( $Fe_2S_2$  to  $Fe_{150}S_{150}$ )

(see Sect. 5.2) [132]. It could be a reasonable approximation to mimic the build-up, *for fast-forwarding geo-time*, by starting out with known (polydisperse) size ranges, taking into account their initial magnetic susceptibility (along the lines of Wang and Holm [168]). Furthermore, the “team-up” of FeS clusters with organics (see also [131]) may well have deeper roots, as organics play important roles in separate aspects related to proposed magnetic assemblies, viz., (1) stabilize colloidal membranes [138]; (2) facilitate particulation mechanism leading to fractal frambooid formation [147, 148]; (3) enable transformation to greigite in aqueous dispersed FeS, at the cost of pyrite formation [130]; and (4) enable generation of metastable phases intermediate between  $FeNi_2S_4$  and  $Fe_3S_4$  (similar to biological clusters), under mild hydrothermal mound-like conditions [158]. Thus the inclusion/exclusion of organics does need to be closely studied in experimental simulations.

## 6 Conclusions

The adaptive nature of biological systems and their fractal organization cry for a coherent connection between their micro- and macroscopic domains. A physical basis—the quantum mechanical spin – for linking the quantum-classical realms at the very origins of life is suggested in this rudimentary study, rooted in the findings of a spectrum of scientists (see bibliography). This in turn also helps to expand the potential of crystal-based theories, and shows how Life-like dynamics could have been brought about by the magnetic “face” of minerals. We propose that structured phases with a magnetic basis for information-transfer, not too far from the mound (Sect. 5.4), accumulated “metabolites” (mound-synthesized) riding in on diffusing super-paramagnetic greigite particles. The evolution of complexity (biological soft matter with internal degrees of freedom, asymmetry, organization, etc.) where chemistry was trained to replace magnetic effects, plus installation/maintenance of energy transduction mechanisms via energy-rich molecules for using non-equilibrium sources elsewhere, could have led to the release of the Ancestor from its H-field providing location. Now, as “Necessity is the mother of invention” could it be that the “necessity” for independence from an increasingly hostile location brought on the creation of such innovative mechanisms? This possibility seems intriguing in the light of Patel’s findings, where quantum searches seem to be responsible for the creation of biological language itself. Moreover, Russell et al. have argued that life’s hatchery could have been busy by 3.8 Gyr, evolving fast enough for a branch to have reached the ocean surfaces by 3.5 Gyr, as evidenced by photosynthetic signatures. The gestation period of life had to have been less than the umbilical mound’s delivery of the formative hydrothermal solution, i.e., certainly less than 3 million years, and probably less than 30,000 years [46]. Indeed, a magnetic start to Life could provide the ingredients for an intelligent Ancestor, along the lines envisaged by Lloyd [88] for a computing universe. Again, it seems to be a physically feasible embodiment [103, 104] of Paul Davies’s Q-Life proposal (2008), as also acknowledged by him in Merali [97]. A magnetic basis of assembly could also offer robustness to

an “open” system against interference from a decohering environment. On the other hand, as evidence of quantum processing effects in biology trickles in, it appears that Nature is equipped for tackling environmental intrusion. Sure enough, with regard to Brownian noise, Nature seems to know how to not only overcome adversity, but also instead put it to its advantage by harnessing it. At the other-macroscopic-end too, elegant examples can be seen in the seed dispersal strategies that use this very “intrusion” by the environment (wind, water, or even creatures). Thus, the environment apparently provides feedback to the adaptive living system, besides defining “necessity” and acting as a “watch-dog” leading to new nodes in biological evolution [96] (Sect. 2.1). Could it be that the paradigm of environment-decoherence being a big obstacle against quantum processing events in biology, needs to be reviewed since environmental interference itself seems to be an active component of Nature’s search technique?

**Acknowledgements** One of us (ANM) is grateful to Prof Krishnaswami Alladi for this opportunity to be associated with this memorial volume dedicated to (the late) Professor Alladi Ramakrishnan. The theme of the article has been governed by a desire to conform to his versatile interest in an entire gamut of physical science through an appropriate choice of subject. The latter comes from a recent father–daughter (nay daughter–father!) collaboration seeking a *Magnetic Origin of Life*, a subject which represents an ultimate synthesis of physics with biological chemistry through the complex terrain of geological science. We thank Prof. M.J. Russell for inspiration and constant support (data and key references); Prof. Z. Sawlowicz for key references; Dr. A. Boyce for active help with his labeled frambooid-in-chimney pictures; Prof. K. Matsuno for suggesting a closer look at electrostatic effects; Prof. A.K. Pati for bringing “Quantum Aspects of Life” to our notice. This work was entirely financed, with full infrastructural support, by Dr. Jean-Jacques Delmotte; Drs A. Bachhawat and B. Sodermark gave a gentle push; Dr. V. Ghildyal and Mr. Vijay Kumar helped with manuscript processing.

## References

- [1] Abel D (2008) The ‘Cybernetic Cut’: Progressing from Description to Prescription in Systems Theory. *Open Cybernet. Systemat. J.* 2, 234–244
- [2] Abel D (2009) The capabilities of chaos and complexity. *Int. J. Mol. Sci.* 10, 247–291
- [3] Adler I, Barabe D, Jean RV (1997) A history of the study of phyllotaxis. *Annals of Botany* 80, 231–244
- [4] Ahniyaz A, Sakamoto Y, Bergstrom L (2007) Magnetic field-induced assembly of oriented superlattices from maghemite nanocubes. *Proc Natl Acad Sci* 104(45), 17570–17574
- [5] Al-Khalili J, McFadden J (2008) Quantum Coherence and the Search for the First Replicator. In: Abbott D, Davies PCW, Pati A.K. (Ed.s) *Quantum aspects of life*. Imperial College Press, London, 3–18
- [6] Arrhenius GO (2003) Crystals and life. *Helv Chim Acta* 86, 1569–1586
- [7] Astumian RD (1997) Thermodynamics and kinetics of a Brownian motor. *Science* 276 (5314), 917–922
- [8] Astumian R.D. (2007) Adiabatic operation of a molecular machine. *Proc Natn Acad Sci* 104(50), 19715–19718
- [9] Anderson PW (1983) Suggested model for prebiotic evolution: The use of chaos. *Proc Natl Acad Sci USA* 80, 3386–3390
- [10] Baaske P, Weinert F, Duhr S et al. (2007) Extreme accumulation of nucleotides in simulated hydrothermal pore systems. *Proc Natl Acad Sci USA* 104, 9346–51

- [11] Bak P, Chen K (1991) Self-organized criticality. *Sci Amer* 264, 46–53
- [12] Ban S, Korenivski V (2006) Pattern storage and recognition using ferrofluids. *J. Appl. Phys.* 99, 08R907
- [13] Beard JS, Hopkinson L (2000) A fossil, serpentinization-related hydrothermal vent, Ocean Drilling Program Leg 173, Site 1068 (Iberia Abyssal Plain): Some aspects of mineral and fluid chemistry. *J. Geophys. Res.* 105(B7), 16527
- [14] Bernal JD (1949) The physical basis of life. *Proc Royal Soc London* 357A, 537–558
- [15] Boyce AJ, Coleman ML, Russell MJ (1983) Formation of fossil hydrothermal chimneys and mounds from Silvermines, Ireland. *Nature*, 306, 545–550
- [16] Boyce AJ (1990) Exhalation, sedimentation and sulphur isotope geochemistry of the Silvermines Zn + Pb + Ba deposits, County Tipperary, Ireland; (Unpublished) Ph.D. thesis, University of Strathclyde, Glasgow
- [17] Breivik J (2001) Self-Organization of Template-Replicating Polymers and the Spontaneous Rise of Genetic Information, *Entropy* 2001, 3, 273–279
- [18] Buchachenko AL (2000) Recent advances in spin chemistry. *Pure Appl. Chem.*, 72 (12), 2243–2258
- [19] Bustamante C., Liphardt J., Ritort F. (2005) The nonequilibrium thermodynamics of small systems. *Physics Today*, 58, 43–48
- [20] Cairns-Smith AG (1982) Genetic takeover and the mineral origins of life. Cambridge Univ Press, Cambridge
- [21] Cerf NJ, Grover LK, Williams CP (2000) Nested quantum search and NP-complete problems. *Phys Rev A* 61, 032303
- [22] Chantrell RW, Bradbury A, Popplewell J, and Charles SW (1982) Agglomeration formation in magnetic fluid. *J Appl Phys* 53(3), 2742–2744
- [23] Chaplin, M.F., 2004. Does water clustering determine biological structure? Gordon Research Conference (2004), ‘Interfacial Water in Cell Biology’, Mount Holyoke College, USA
- [24] Chaplin, M.F., 2006. Opinion: Do we underestimate the importance of water in cell biology? *Nature Rev. Mol. Cell Biol.* 7, 861–866
- [25] Charnock J, Garner CD, Patrick RAD et al. (1990) An EXAFS study of thiospinel minerals. *Amer Mineral* 75, 247–255
- [26] Cheon J, Park J-I, Choi J-S et al. (2006) Magnetic superlattices and their nanoscale phase transition effects. *Proc Natl Acad Sci* 103(9), 3023–3027
- [27] Chern G-W, Fennie CJ, Tchernyshov O (2006) Broken parity and a chiral ground state in the frustrated magnet  $CdCr_2O_4$ . *Phys Rev B* 74, 060405(R) 4pp
- [28] da Cruz W (2005) A fractal-like structure for the fractional quantum Hall effect. *Chaos Solitons and Fractals* 23, 373–378
- [29] Davies PCW (2003) How bio-friendly is the universe. *Int J Astrobiol* 2, 115
- [30] Davies PCW (2004) Does quantum mechanics play a non-trivial role in life? *Biosystems* 78, 69–79
- [31] Davies PCW (2008) A quantum origin of Life? In: Abbott D, Davies PCW, Pati A.K. (Ed.s) *Quantum aspects of life*. Imperial College Press, London, 3–18
- [32] Davila AF, Winklhofer M, McKay C (2007) Multicellular magnetotactic prokaryote as a target for life search on Mars. 38th Lunar and Planetary Science Conference, (Lunar and Planetary Science XXXVIII), held March 12–16, 2007 in League City, Texas. LPI Contribution No. 1338, p. 1495
- [33] Degens ET, Matheja J, Jackson TA (1970) Template catalysis: Asymmetric polymerization of amino-acids on clay minerals. *Nature* 227, 492–493
- [34] Doll KM, Finke RG (2003) A compelling experimental test of the hypothesis that enzymes have evolved to enhance quantum mechanical tunneling in hydrogen transfer reactions: The  $\beta$ -neopentylcobalamin system combined with prior adocobalamin data. *Inorg Chem* 42 (16), 4849–4856
- [35] Drubin DG (2000) *Cell polarity*. Oxford University Press, Oxford
- [36] Dunlap RA (1997) *The Golden Ratio and Fibonacci Numbers*. World Scientific, New Jersey.

- [37] Dymant J, Arkani-Hamed J, Ghods A (1997) Contribution of serpentized ultramafics to marine magnetic anomalies at slow and intermediate spreading centres: insights from the shape of the anomalies. *Geophys. J. Int.* 129, 691
- [38] Dyson FJ (1999) *Origins of Life*. (2nd ed.) Cambridge Univ Press, Cambridge
- [39] Engel A, Müller HW, Riemann P, Jung A (2003) Ferrofluids as thermal ratchets. *Phys Rev Lett* 91(6), 060602
- [40] Engel GS, Calhoun TR, Read EL et al. (2007) Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. *Nature* 446, 782–786
- [41] Etxeberria A, Moreno A (2001) From complexity to simplicity: Nature and symbols. *Biosystems* 60(1–3), 149–157
- [42] Ferris JP (2005) Mineral catalysis and prebiotic synthesis: Montmorillonite-catalyzed formation of RNA. *Elements* 1, 145–149
- [43] Fritsch V, Hemberger J, Büttgen N, Scheidt EW, et al. (2004) Spin and orbital frustration in  $MnSc_2S_4$  and  $FeSc_2S_4$ . *Phys. Rev. Lett.* 92(11), 116401
- [44] Frohlich H (1968) Longrange coherence and energy storage in biological systems. *Int J Quantum Chem* 2, 6419
- [45] Frohlich H (1975) The extraordinary dielectric properties of biological materials and the action of enzymes. *Proc Natl Acad Sci USA* 72, 42114215
- [46] Früh-Green GL, Kelley DD, Bernasconi SM et al. (2003) 30,000 years of hydrothermal activity at the Lost City vent field. *Science* 301, 495–498
- [47] Galland P, Pazur A (2005) Magnetoreception in plants. *J Plant Res* 118(6), 371–389
- [48] Goerbig MO, Lederer P, Morais Smith C (2004) On the self-similarity in quantum Hall systems. *Europhys Lett* 68, 72–78
- [49] Goldschmidt VM (1952) Geochemical aspects of the origin of complex organic molecules on the Earth, as precursors to organic life. *New Biology* 12, 97–105
- [50] Grzybowski BA, Campbell CJ (2004) Complexity and dynamic self-assembly. *Chem Eng Sci* 59, 1667
- [51] Hagan S, Hameroff S, Tuszyński J (2002) Quantum computation in brain microtubules? Decoherence and biological feasibility. *Phys Rev E* 65, 061901
- [52] Harold FM (2005) Molecules into cells: Specifying spatial architecture. *Microbiol Mol Biol Rev* 69(4), 544–564
- [53] Hatori K, Honda H, Matsuno K (2001) Magnetic Dipoles and Quantum Coherence in Muscle Contraction. arXiv:quant-ph/0104042v1
- [54] Haw M (2007) The industry of Life. *Phys World*, Nov issue.
- [55] Hazen R, Papineau D, Bleeker W, Downs RT, Ferry JM, McCoy TJ, et al. (2008) Mineral evolution. *American Mineralogist* 93, 1693
- [56] Hemberger J, Lunkenheimer P, Ficht R et al. (2006) Multiferroicity and colossal magneto-capacitance in Cr-thiospinels. *Phase Trans* 79, 1065
- [57] Heylighen F (2001) The Science of Self-organization and Adaptivity. In: L. D. Kiel, (ed.) Knowledge Management, Organizational Intelligence and Learning, and Complexity, In: *The Encyclopedia of Life Support Systems*, Oxford
- [58] Higo J, Sasai M, Shirai H et al. (2001) Large vortex-like structures of dipole field in computer models of liquid water and dipole-bridge between biomolecules. *Proc Natl Acad Sci USA* 98, 5961–5964
- [59] Hoffmann V (1992) Greigite ( $Fe_3S_4$ ): magnetic properties and first domain observations. *Phys. Earth Planet. Interiors* 70, 288–301
- [60] Hoffmeyer J (2001) Life and Reference. *Biosystems* 60, 123–130
- [61] Hollander WThF den, Toninelli FL (2004) Spin glasses: A mystery about to be solved. *Nieuw Archief voor Wiskunde*, 5/5(4), 274–278
- [62] Huber C, Wächtershäuser G (1997) Activated acetic acid by carbon fixation on (Fe,Ni)S under primordial conditions. *Science* 276, 245–247
- [63] Itoh S, Kajimoto R, Iwasa K et al. (2006) Fractal structure and critical scattering in the three-dimensional percolating antiferromagnet,  $RbMn_0.31Mg_0.69F_3$ . *Physica B* 385–386(1), 441–443



- [64] Jain JK (2007) Composite fermions. Cambridge Univ Press, Cambridge
- [65] Jaryznski C (1997) Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* 78, 2690
- [66] Jibu M, Hagan S, Hameroff SR et al. (1994) Quantum optical coherence in cytoskeletal microtubules: implications for brain function. *Biosystems* 32(3), 195–209
- [67] Kant I (1790) *Kritik der Teleologischen Urteilskraft*, Berlin and Libau (Collected works Vol. X Ed. W. Weischedel, Frankfurt, 1968)
- [68] Kauffman S (1993) *The origins of order: Self-organization and Selection in Evolution*. Oxford Univ Press, New York
- [69] Kelley D S, Karson J A, Früh-Green GL et al. (2005) A serpentinite-hosted ecosystem: the Lost City hydrothermal field. *Science* 307, 1428–1434
- [70] Keys AS, Glotzer SC (2007) How do Quasicrystals Grow? *Phys Rev Lett* 99, 235503
- [71] Khomskii DI (2004) Multiferroics: different ways to combine magnetism and ferroelectricity. *J Magn Magn Mat* 306, 1
- [72] Kirschvink JL, Gould JL (1981) Biogenic magnetite as a basis for magnetic field detection in animals. *Biosystems* 13(3), 181–201
- [73] Kirschvink JL, Hagadorn JW (2000) A grand unified theory of biomineralization. In: E. Bäuerlein (Ed.) *The Biomineralization of Nano- and Micro- structures* (Wiley VCH, Weinheim, Germany) 139
- [74] Kirschvink JL, Kobayashi-Kirschvink A, Diaz-Ricci JC et al. (1992) Magnetite in human tissues: a mechanism for the biological effects of weak ELF magnetic fields. *Bioelectromagnetics Suppl* 1, 101–113
- [75] Kominis IK (2008) Quantum Zeno Effect Underpinning the Radical-Ion-Pair Mechanism of Avian Magnetoreception. arXiv:0804.2646
- [76] Koonin EV, Martin W (2005) On the origin of genomes and cells within inorganic compartments. *Trends Genet* 21, 647–654
- [77] Kopelman R (1989) Diffusion-controlled reaction kinetics. In: Avnir D (ed) *The fractal approach to heterogenous chemistry*. John Wiley, New York, pp 295–309
- [78] Korolev VV, Arefyev IM, Ramazanova A G (2008) The magnetocaloric effect of superfine magnets. *J Thermal Anal Cal* 92(3), 691–695
- [79] Kwon Y (2007) Quantum mechanism of biological search. *Chaos, Solitons and Fractals* 34 (4), 1037–1038
- [80] Küppers B-O (1990) *Information and the origin of life*. Cambridge, MIT Press.
- [81] Lalatonne Y, Richardi J, Pileni MP (2004) Van der Waals versus dipolar forces controlling mesoscopic organizations of magnetic nanocrystals. *Nat Mater* 3(2), 121–125
- [82] La Mer VK, Dinegar RH (1950) Theory, production and mechanism of formation of monodispersed hydrosols. *J Amer Chem Soc* 72, 4847–4854
- [83] Larter RCL, Boyce AJ, Russell MJ (1981) Hydrothermal pyrite chimneys from the Ballynoe baryte deposit, Silvermines, County Tipperary, Ireland. *Mineralium Deposita* 16, 309–317
- [84] Levitov LS (1991) Fibonacci numbers in botany and physics: Phyllotaxis. *Pis'ma Zh Eksp Teor Fiz* 54(9), 542–545
- [85] Li J, Huang Y, Liu X et al. (2007) Effect of aggregates on the magnetization property of ferrofluids: A model of gaslike compression. *Sci Tech Adv Mater* 8, 448–454
- [86] Lifshitz R (1998) Symmetry of magnetically ordered quasicrystals. *Phys Rev Letts* 80(12), 2717–2720
- [87] Ling GN (2001) *Life at the cell and below-cell level: The hidden history of a functional revolution in Biology*. Pacific Press, New York
- [88] Lloyd S (2006) *Programming the universe: a quantum computer scientist takes on the cosmos*. New York: Knopf
- [89] Manoharan VN, Elsesser MT, Pine DJ (2003) Dense Packing and Symmetry in Small Clusters of Microspheres. *Science*, 301(5632), 483–487
- [90] Marteinson VTh, Kristjánsson JK, Kristmannsdóttir H et al. (2001) Discovery of giant submarine smectite cones on the seafloor in Eyjafjörður, Northern Iceland, and a novel thermal microbial habitat. *Appl Environ Microbiol* 67, 827–833



- [91] Martin W, Baross J, Kelley D et al. (2008) Hydrothermal vents and the origin of life. *Nature Rev Microbiol*, doi:10.1038/nrmicro1991
- [92] Martin W, Russell MJ (2003) On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philos Trans R Soc Lond* 358(1429), 59–83, discussion 83–85
- [93] Matsuno K, Paton RC (2000) Is there a biology of quantum information? *Biosystems* 55(1–3), 39–46
- [94] Mavromatos NE, Nanopoulos DV, Samaras I et al. (1998) Ferroelectrics and their possible involvement in biology. *Adv Struct Biol* 5, 127–134
- [95] McBride JM, Tully JC (2008) Did life grind to a start? *Nature* 452, 161–162
- [96] McFadden J, Al-Khalili J (1999) A quantum mechanical model of adaptive mutations. *Biosystems* 50, 203–211
- [97] Merali Z (2007) Was life forged in a quantum crucible? *New Scientist* 196(2633), 8 Dec, 6–7
- [98] Mertinat M, Tsurkan V, Samusi D et al. (2005) Low-temperature structural transition in  $FeCr_2S_4$ . *Phys Rev B* 71, 100408(R) (4 pages)
- [99] Michie, K.A., Lowe, J., 2006. Dynamic filaments of the bacterial cytoskeleton. *Annu.Rev.Biochem.* 75, 467–92
- [100] Milner-White EJ, Russell MJ (2005) Nests as sites for phosphates and iron-sulfur thiolates in the first membranes: 3 to 6 residue anion-binding motifs. *Origins Life Evol Biosphere* 35, 19–27
- [101] Milner-White EJ, Russell MJ (2008) Predicting the conformations of peptides and proteins in early evolution. *Biol Direct*, doi:10.1186/1745-6150-3-3
- [102] Min Y, Akbulut M, Kristiansen et al. (2008) The role of interparticle and external forces in nanoparticle assembly. *Nature Materials* 7, 527–538
- [103] Mitra-Delmotte G, Mitra AN (2007) Can magnetism-assisted quasiperiodic structures in Russell-FeS ‘bubbles’ offer a quantum coherent origin of life? arXiv:0710.0220v1 [cond-mat.soft]
- [104] Mitra-Delmotte and Mitra (2009) Magnetism, entropy, and the first nano-machines. *Cent. Eur. J. Phys.* DOI 10.2478/s11534-009-0143-4
- [105] Mostovoy M (2006) Ferroelectricity in spiral magnets. *Phys Rev Lett* 96, 067601 (4 pages)
- [106] Nagya A, Prokhorenkoa V, Dwayne Millera RJ (2006) Do we live in a quantum world? Advances in multidimensional coherent spectroscopies refine our understanding of quantum coherences and structural dynamics of biological systems. *Curr Opin Struct Biol* 16(5), 654–663
- [107] Nelson DR (2003) Spherical Crystallography: Virus Buckling and Grain Boundary Scars. arXiv:cond-mat/0311413v1
- [108] Newman SA, Muller GB (2000) Epigenetic mechanisms of character origination. *J Exp Zool* 288, 304–317
- [109] Nicolis G, Prigogine I (1977) *Self-organization in Non-equilibrium Systems: from dissipative structure to order through fluctuations*. Wiley, New York
- [110] Odenbach S (2004) Recent progress in magnetic fluid research. *J Phys Condens Matter* 16, R1135–R1150
- [111] Ohfuji H, Akai J (2002) Icosahedral domain structure of framboidal pyrite. *Amer Mineral* 87, 176–180
- [112] Ohfuji H, Rickard D (2005) Experimental synthesis of framboids—a review *Earth-Sci Rev* 71, 147–170
- [113] Ohldag H, Tyliczszak T, Hohne R et al. (2007) Pi-Electron ferromagnetism in metal-free carbon probed by soft X-ray dichroism. *Phys Rev Lett* 98, 187204
- [114] Orgel LE (1986) Did template-directed nucleation precede molecular replication? *Orig Life Evol Biosph* 17(1), 27–34
- [115] Orgel LE (2000) Self-organizing biochemical cycles. *Proc Natl Acad Sci USA* 97(23), 12503–12507

- [116] Orgel LE (2004) Prebiotic chemistry and the origin of the RNA world. *Crit Rev Biochem Mol Biol* 39(2), 99–123
- [117] Palm R, Korenivski V (2009) A ferrofluid based neural network: design of an analogue associative memory. *New J. Phys.* 11, 023003
- [118] Palmer HM, Greaves C (1999) Structural, magnetic and electronic properties of  $Fe_{0.5}Cu_{0.5}Cr_2S_4$ . *J Mater Chem* 9, 637–640
- [119] Patel AD (2001) Quantum algorithms and the genetic code. *Pramana* 56(2), 367
- [120] Patel AD (2006) Optimal Database Search: Waves and Catalysis, *Int. J. Quant. Inform.* 4, 815–825; Erratum, *ibid.* 5 (2007) 437; [quant-ph/0401154](#)
- [121] Patel AD (2002) Carbon—The First Frontier of Information Processing, *J. Biosc.* 27, 207–218
- [122] Pastor-Satorras R., Rubi J.M. (2000) Dipolar interactions induced order in assemblies of magnetic particles. *J. Magn. Magn. Mater.* 221(1–2), 124–131
- [123] Pattee HH (1979) Complementarity vs. reduction as explanation of biological complexity. *Am J Physiol Regul Integr Comp Physiol* 236, R241–R246
- [124] Pattee HH (1996) The Problem of Observables in Models of Biological Organizations. In: Khalil EL, Boulding KE, (Eds.) *From Evolution, Order, and Complexity*, Routledge, London.
- [125] Phillips R, Quake SR (2006) The biological frontier of physics. *Physics Today* 59(5), 38–43
- [126] Pileni MP (2003) The role of soft colloidal templates in controlling the size and shape of inorganic nanocrystals. *Nat Mater* 2(3), 145–50
- [127] Polyani M (1968) Life's irreducible structure. *Science* 160 (3834), 1308–1312
- [128] Preisinger A, Aslanian S (2004) The formation of framboidal greigites in the Black Sea. *Geophys Res Abstr* 6, 02702 (SRef-ID: 1607-7962/gra/EGU04-A-02702)
- [129] Pross A (2005) Stability in chemistry and biology: Life as a kinetic state of matter. *Pure Appl. Chem.* 77 (11), 1905–1921
- [130] Rickard D, Butler IB, Oldroyd A (2001) A novel iron sulphide mineral switch and its implications for Earth and planetary science, *Earth Planet Sci Lett* 189, 85–91
- [131] Rickard D, Luther GW III (2007) The chemistry of iron sulfides. *Chem Revs* 107, 514–562
- [132] Rickard D, Morse JW (2005) Acid volatile sulfide (AVS) *Marine Chem* 97, 141–197
- [133] Righeter K, Drake MJ, Yaxley G (1997) Prediction of siderophile element metal-silicate partition coefficients to 20GPa and 2,800° C: the effects of pressure, temperature, oxygen fugacity, and silicate and metallic melt compositions. *Phys Earth Planet Interiors* 100, 115–134
- [134] Roberts AP, Turner GM (1993) Diagenetic formation of ferrimagnetic iron sulfide minerals in rapidly deposited marine sediments, South Island, New Zealand. *Earth Planet Sci Lett* 115, 257–273
- [135] Rosensweig RE (1997) *Ferrohydrodynamics*. Dover, New York
- [136] Russell MJ (2007) The Alkaline Solution to the Emergence of Life: Energy, Entropy and Early Evolution. *Acta Biotheor* 55(2), 133–179
- [137] Russell MJ, Arndt NT (2005) Geodynamic and metabolic cycles in the Hadean. *Biogeosciences* 2, 97–111
- [138] Russell MJ, Daniel RM, Hall AJ et al. (1994) A hydrothermally precipitated catalytic iron sulphide membrane as a first step toward life. *J Mol Evol* 39, 231–243
- [139] Russell MJ, Hall AJ (1997a) The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. *J Geol Soc London* 154(pt 3) 377–402
- [140] Russell MJ, Hall AJ (1997b) *J. Geol. Soc. London* 154 (pt 3) 377
- [141] Russell MJ, Hall AJ (2006) The onset and early evolution of life. In: Kesler SE and Ohmoto H (eds) *Evolution of early earth's atmosphere, hydrosphere, and biosphere—constraints from ore deposits*. Geological Society of America, *Memoir*, 198, 1–32
- [142] Russell MJ, Hall AJ (2009) A hydrothermal source of energy and materials at the origin of life. In “*Chemical Evolution II: From Origins of Life to Modern Society*”. American Chemical Society
- [143] Russell MJ, Hall AJ, Gize AP (1990) Pyrite and the origin of life. *Nature* 344, 387

- [144] Russell MJ, Hall AJ, Turner D (1989) In vitro growth of iron sulphide chimneys: possible culture chambers for origin-of-life experiments. *Terra Nova* 1(3), 238–241
- [145] Russell MJ, Hall AJ, Boyce AJ et al. (2005) On hydrothermal convection systems and the emergence of life. *Economic Geology* 100 (3), 419–438
- [146] Russell MJ, Hall AJ, Mellersh AR (2003) On the dissipation of thermal and chemical energies on the early Earth: The onsets of hydrothermal convection, chemiosmosis, genetically regulated metabolism and oxygenic photosynthesis. In: Ikan R (ed) *Natural and laboratory-simulated thermal geochemical processes*. Dordrecht, Kluwer Academic Publishers, pp 325–388
- [147] Sawlowicz Z (1993) Pyrite framboids and their development: a new conceptual mechanism. *Int J Earth Sci* 82, 148–156
- [148] Sawlowicz Z (2000) Framboids: from their origin to application. *Prace Mineralog Pan* 88, 1–80
- [149] Schroeder T, John B, Frost R (2002) Geologic implications of seawater circulation through peridotite exposed at slow-spreading mid-ocean ridges. *Geology*; 30(4), 367
- [150] Schroedinger E (1944) *What is Life?* Cambridge University Press, Cambridge
- [151] Selvam AM (1998) Quasicrystalline pattern formation in fluid substrates and phyllotaxis. In: Barabe D, Jean RV (eds) *Symmetry in Plants*. World Scientific Series in Mathematical Biology and Medicine (4), Singapore, pp 795–809
- [152] Shapiro R (1999) Prebiotic cytosine synthesis: A critical analysis and implications for the origin of life. *Proc Natl Acad Sci* 96, 4396–4401
- [153] Skomski R (2008) *Simple models of magnetism*. Oxford University Press, Oxford
- [154] Spitzer J and Poolman B (2009) The role of biomacromolecular crowding, ionic strength, and physicochemical gradients in the complexities of life's emergence. *Microbiol Molec Biol Reviews* 73, 371–388
- [155] Stein DL (1996) Spin glasses. In: Rigden JS (ed) *Macmillan Encyclopedia of Physics*. Simon and Schuster Macmillan, New York, pp 1514–1516
- [156] Sun S, Murray CB, Weller D et al. (2000) Monodisperse FePt Nanoparticles and Ferromagnetic FePt Nanocrystal Superlattices. *Science* 287, 1989–1992
- [157] Taketomi S, Takahashi H, Inaba N et al. (1991) Experimental and theoretical investigations on agglomeration of magnetic colloidal particles in magnetic fluids. *J Phys Soc Jpn* 60(5), 1689–1707
- [158] Tenaillon C, Pring A, Estchmann B et al. (2006) Transformation of pentlandite to violarite under mild hydrothermal conditions. *Amer Mineral* 91(4), 706–709
- [159] Tishin AM and Spichkin YI (2003) *The Magnetocaloric Effect and its Applications*, IOP Publishing, Bristol and Philadelphia (2003) 475 pp
- [160] Torbet J, Ronziere M-C (1984) Magnetic alignment of collagen during self-assembly. *Biochem J* 219, 1057–1059
- [161] Traverso S (2005) Cytoskeleton as a fractal percolation cluster: Some biological remarks. In: Losa G, Merlini D, Nonnenmacher E, Weibel E (eds) *Fractals in biology and medicine*. (Part 4) Birkhauser, Basel, pp 269–275
- [162] Trevors JT, Pollack GH (2005) Hypothesis: the origin of life in a hydrogel environment. *Prog Biophys Mol Biol* 89(1), 1–8
- [163] Vaughan DJ, Craig JR (1985) the crystal chemistry of iron-nickel thiospinels. *Amer Mineral* 70, 1036–1043
- [164] Vaughan DJ, Tossell JA (1981) Electronic structure of thiospinels minerals: results from MO calculations. *Amer Mineral* 66, 1250–1253
- [165] Vestal CR (2004), Ph.D. Thesis, Georgia Institute of Technology (Atlanta, GA, USA)
- [166] Viedma C, Ortiz JE, de Torres T, Izumi T and Blackmond DG (2008) Evolution of Solid Phase Homochirality for a Proteinogenic Amino Acid. *J Am Chem Soc* 130, 15274–15275
- [167] Wächtershäuser G (1988) Before enzymes and templates: theory of surface metabolism. *Microbiol Rev* 52, 452–484
- [168] Wang Z, Holm C (2003) Structure and magnetic properties of polydisperse ferrofluids: A molecular dynamics study. *Phys Rev E* 68, 041401, pp 1–11

- [169] Wasilewski P, Kletetschka G (1999) Lodestone: Natures only permanent magnet-what it is and how it gets charged; *Geophys Res Lett*, 26(15), 2275–2278
- [170] Weaver JC, Vaughan TE, Astumian RD (2000) Biological sensing of small field differences by magnetically sensitive chemical reactions. *Nature* 405(6787), 707–709
- [171] Wilkin RT, Barnes HL (1997) Formation processes of framboidal pyrite. *Geochim Cosmochim Acta* 61, 323–339
- [172] Winkelhofer M (2005) Biogenic magnetite and magnetic sensitivity in organisms - from magnetic bacteria to pigeons. *Magneto hydrodynamics* 41(4), 295–304
- [173] Wolthers M, Van Der Gaast SJ, Rickard D (2003) The structure of disordered mackinawite. *Amer Mineral* 88, 2007–2015
- [174] Wu M, Xiong Y, Jia Y et al. (2005) Magnetic field-assisted hydrothermal growth of chain-like nanostructure of magnetite. *Chem Phys Lett* 401, 374–379
- [175] Yethiraj A (2007) Tunable colloids: control of colloidal phase transitions with tunable interactions. *Soft Matter* 3, 1099–1115
- [176] Zubarev AY, Fleischer J, Odenbach S (2005) Towards a theory of dynamical properties of polydisperse magnetic fluids: Effect of chain-like aggregates. *Physica A* 358, 475–491
- [177] Zubarev AY, Iskakova LY (2004) To the theory of rheological properties of ferrofluids: influence of drop-like aggregates. *Physica A* 343, 65–80
- [178] Zurek WH (2003) Decoherence, einselection, and the quantum origins of the classical. *Rev. Mod. Phys.* 75, 715–775

# The Ehrenfest Theorem in Quantum Field Theory

Ragavachariar Parthasarathy

*Dedicated to the memory of Prof. Alladi Ramakrishnan*

Professor Alladi Ramakrishnan founded the Institute of Mathematical Sciences (MATSCIENCE) in 1962 and attracted bright young students interested in theoretical physics. His contributions to the theory of Stochastic processes, elementary particle physics and Generalized Clifford Algebras will be remembered forever. He was instrumental in my joining MATSCIENCE in 1977 and encouraged me till his end in my research work. I consider it my duty to dedicate this article in his memory.

**Summary** The validity of the Ehrenfest theorem in Abelian and non-Abelian quantum field theories is examined. The gauge symmetries are taken to be unbroken. By suitably choosing the physical subspace, the above validity is proven in both the cases.

**Mathematics Subject Classification (2000)** 81Q05, 81T13, 81V05

**Key words and phrases** Ehrenfest's theorem · Schrödinger equation · Expectation values · Dirac equation · Abelian field theory · Gauge fixing · Quantum lagrangian · Physical subspace · Non-Abelian field theory · Quantum Chromo Dynamics · Path integral approach · Faddeev–Popov ghosts · Quantum equations · BRS transformation · Global gauge and scale transformations · Physical states

## 1 Quantum Mechanics

In quantum mechanics, it is reasonable to expect the motion of a wave packet to agree with the motion of the corresponding classical particle whenever the potential energy changes by a small amount over the dimensions of the wave packet.

---

R. Parthasarathy  
Chennai Mathematical Institute, H1, SIPCOT IT Park, Padur Post, Siruseri 603103, India  
e-mail: [sarathy@cmi.ac.in](mailto:sarathy@cmi.ac.in)

If we mean by the “position” and the “momentum” vectors of the wave packet, their expectation values, then we can show that the classical and the quantum motions agree. This important result is known as the Ehrenfest theorem [1, 2]. To illustrate this theorem, let us first consider non-relativistic quantum mechanics. We have the Schrödinger equation

$$\begin{aligned} i\hbar \frac{\partial \psi(\vec{x}, t)}{\partial t} &= -\frac{\hbar^2}{2m} \vec{\nabla}^2 \psi(\vec{x}, t) + V(\vec{x}) \psi(\vec{x}, t), \\ -i\hbar \frac{\partial \psi(\vec{x}, t)^\dagger}{\partial t} &= -\frac{\hbar^2}{2m} \vec{\nabla}^2 \psi(\vec{x}, t)^\dagger + V(\vec{x}) \psi(\vec{x}, t)^\dagger, \end{aligned} \quad (1)$$

where  $m$  is the mass of the particle and  $V(\vec{x})$  is the real potential.

We shall take the wave function  $\psi(\vec{x}, t)$  in (1) as normalized. Then the expectation value of the  $x$ -component of the position operator and its time derivative are

$$\begin{aligned} \langle x \rangle &= \int \psi^\dagger x \psi \, d\tau, \\ \frac{d}{dt} \langle x \rangle &= \int \left( \frac{d\psi^\dagger}{dt} \right) x \psi \, d\tau + \int \psi^\dagger x \left( \frac{d\psi}{dt} \right) \, d\tau. \end{aligned} \quad (2)$$

Using (1), it follows

$$\frac{d}{dt} \langle x \rangle = -\frac{i\hbar}{m} \int \psi^\dagger \frac{\partial}{\partial x} \psi \, d\tau = \frac{1}{m} \langle p_x \rangle. \quad (3)$$

Similarly, starting from  $\langle p_x \rangle = -i\hbar \int \psi^\dagger \frac{\partial}{\partial x} \psi \, d\tau$ , it is easy to find

$$\frac{d}{dt} \langle p_x \rangle = \left\langle -\frac{\partial V(\vec{x})}{\partial x} \right\rangle. \quad (4)$$

From (2) and (4), we note that the *classical* equations of motion

$$\frac{d\vec{x}}{dt} = \frac{\vec{p}}{m}; \quad \frac{d\vec{p}}{dt} = -\vec{\nabla} V(\vec{x}) \quad (5)$$

are satisfied by their expectation values in quantum mechanics. The wave packet moves like a classical particle whenever the expectation value gives a good representation of the classical variable. They provide an example of the correspondence principle [1, 2].

In the case of relativistic quantum mechanics, the manipulations are a little less direct. We consider the Dirac equation [3].

$$\begin{aligned} H &= \vec{\alpha} \cdot \vec{p} + \beta m, \\ i\hbar \frac{\partial \psi}{\partial t} &= (\vec{\alpha} \cdot \vec{p} + \beta m) \psi, \end{aligned} \quad (6)$$

where  $\vec{\alpha}$  and  $\beta$  are hermitian  $4 \times 4$  matrices and  $\psi$  is a  $4 \times 1$  column vector. We shall set the velocity of light  $c$  to unity hereafter. By using the Heisenberg equation of motion  $\frac{dx}{dt} = \frac{1}{i\hbar}[x, H]$ , it is seen that

$$\int \psi^\dagger \frac{dx}{dt} \psi d\tau = \int \psi^\dagger \alpha_x \psi d\tau. \quad (7)$$

First, we recall the plane wave solutions  $\psi^{(i)}$  [3] of the Dirac equation,

$$\begin{aligned} \psi^1(x) &= \sqrt{\frac{E+m}{2m}} e^{-ipx} \begin{pmatrix} 1 \\ 0 \\ \frac{p_z}{(E+m)} \\ \frac{p_+}{(E+m)} \end{pmatrix}; & \psi^2(x) &= \sqrt{\frac{E+m}{2m}} e^{-ipx} \begin{pmatrix} 0 \\ 1 \\ \frac{p_-}{(E+m)} \\ -\frac{p_z}{(E+m)} \end{pmatrix}, \\ \psi^3(x) &= \sqrt{\frac{E+m}{2m}} e^{ipx} \begin{pmatrix} \frac{p_z}{(E+m)} \\ \frac{p_+}{(E+m)} \\ 1 \\ 0 \end{pmatrix}; & \psi^4(x) &= \sqrt{\frac{E+m}{2m}} e^{ipx} \begin{pmatrix} \frac{p_-}{(E+m)} \\ -\frac{p_z}{(E+m)} \\ 0 \\ 1 \end{pmatrix}, \end{aligned}$$

corresponding to positive energy ( $E > 0$ ) spin-up, spin-down states and negative energy ( $E < 0$ ) spin-up, spin-down states of the electron, respectively,  $px = Et - \vec{p} \cdot \vec{x}$  and  $p_\pm = p_x \pm ip_y$ . These solutions satisfy  $\psi^{(i)\dagger}(x)\psi^{(j)}(x) = \frac{E}{m}\delta^{i,j}$  ( $i, j = 1, 2, 3, 4$ ). Using these, we construct the wave packets

$$\begin{aligned} \Psi(E > 0) &= \sum_{i=1}^2 \int A_i(\vec{p}) \psi^{(i)} d^3 p, \\ \Psi(E < 0) &= \sum_{i=3}^4 \int A_i(\vec{p}) \psi^{(i)} d^3 p, \end{aligned} \quad (8)$$

and note

$$\int \Psi^\dagger(E > 0) \Psi(E > 0) d^3 x = \int d^3 p \frac{E}{m} \{|A_1(\vec{p})|^2 + |A_2(\vec{p})|^2\}. \quad (9)$$

Similar expression can be written for  $\Psi(E < 0)$ . Using the explicit representation of the  $\alpha_x$  matrix [3], we have

$$\int \Psi^\dagger(E > 0) \alpha_x \Psi(E > 0) d^3 x = \int d^3 p \left(\frac{px}{m}\right) \{|A_1(\vec{p})|^2 + |A_2(\vec{p})|^2\}. \quad (10)$$

From (7), (9), and (10), it follows  $\frac{d}{dt}\langle x \rangle = \langle \frac{px}{m} \rangle$ , showing the validity of the Ehrenfest theorem. Further, we consider the Dirac particle in an external electromagnetic field. Setting the vector potential as zero (for simplicity), the Dirac hamiltonian is

$$H = \vec{\alpha} \cdot \vec{p} + \beta m - e\phi, \quad (11)$$

where  $\phi$  is the scalar potential. Using the Heisenberg equation of motion for a dynamical variable  $F$ ,  $\frac{dF}{dt} = \frac{1}{i\hbar}[F, H]$ , it follows that  $\frac{d\vec{p}}{dt} = -\vec{\nabla}(-e\phi)$  and so  $\langle \frac{d\vec{p}}{dt} \rangle = -\langle \vec{\nabla}(-e\phi) \rangle$ , showing the validity of the Ehrenfest theorem.

Thus in quantum mechanics, we see that the expectation values of the position and the momentum operators satisfy the classical equations of motion. We would like to extend this to quantum field theory.

## 2 Abelian Field Theory

We consider the lagrangian density for the electromagnetic field minimally coupled to a source  $j^\mu(x)$  (Dirac current)

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + eA_\mu(x)j^\mu(x), \quad (12)$$

where  $A_\mu(x)$  is the electromagnetic field,  $e$  is the coupling strength, and

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \quad (13)$$

The corresponding *classical equation* (Euler–Lagrange equation) is

$$\partial_\mu F^{\mu\nu} + ej^\nu = 0. \quad (14)$$

Equation (14) is the *classical equation of motion* and gives the Maxwell equations with source.

It is well known that the manifestly covariant theory of massless vector field is to be quantized with indefinite metric [4]. The impossibility of quantizing the electromagnetic field with positive definite metric has been shown by Mathews, Seetharaman, and Simon [5]. A physically meaningful theory is constructed by introducing a “subsidiary condition,” which is a condition defining the *physical subspace* of the indefinite metric Hilbert space of the electromagnetic field. Here, we follow the  $B$ -field formalism of Nakanishi [6]. In order to quantize the above lagrangian, one has to fix the gauge. This is carried out by considering the coefficient of the terms quadratic in  $A_\mu$  in the action  $S = \int d^4x \mathcal{L}$  (after a partial integration). This coefficient is the differential operator  $\square g^{\mu\nu} - \partial^\mu \partial^\nu$ . The two-point function  $\langle A_\mu(x)A_\nu(y) \rangle$  is governed by the above differential operator.

The Feynman propagator for the photon (quantized electromagnetic field) is the inverse of this differential operator in the momentum space. As this differential operator is not invertible, the photon propagator is not defined. This difficulty is avoided by choosing a gauge. We choose the covariant gauge  $\partial^\mu A_\mu = 0$  and implement this gauge fixing in the lagrangian by adding the “gauge fixing term”  $-1/2a(\partial^\mu A_\mu)^2$ , where  $a$  is a parameter. This modifies the coefficient of the terms quadratic in  $A_\mu$  in the action  $S$  as  $\square g^{\mu\nu} - \partial^\mu \partial^\nu + 1/a \partial^\mu \partial^\nu$ . This, in the momentum space



is,  $-p^2 g^{\mu\nu} + (1 - 1/a)p^\mu p^\nu$ , the inverse of which is  $-1/p^2\{g_{\mu\nu} + \frac{a-1}{p^2} p_\mu p_\nu\}$ , which is the Feynman propagator for the photon in the covariant gauge.

We introduce the above covariant gauge fixing via  $B(x)$ , an auxiliary hermitian scalar field, and consider the *quantum lagrangian*

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + B(x)\partial^\mu A_\mu + \frac{a}{2}B^2(x) + eA_\mu(x)j^\mu(x), \quad (15)$$

where  $a$  is a parameter. It is important to realize that the gauge field  $A_\mu(x)$  and  $B(x)$  in (15) are *operators*, while the gauge field in (12) is a classical field. The quantum equations of motion from (15) are

$$\begin{aligned} \partial_\mu F^{\mu\nu} - \partial^\nu B(x) &= -ej^\nu, \\ \partial^\mu A_\mu + aB(x) &= 0. \end{aligned} \quad (16)$$

Using the second equation to eliminate the  $B$ -field in the lagrangian, we recover the gauge fixing term  $-1/2a(\partial^\mu A_\mu)^2$ . By taking  $\partial_\nu$  of the first equation and using the conservation of the current  $j^\nu(x)$ , namely,  $\partial_\nu j^\nu(x) = 0$ , we see that  $B(x)$  satisfies the equation of motion for a massless scalar field, admitting positive and negative frequency solutions. Equation (16) can be considered to be the *quantum Maxwell equations*, while (14) is the classical equation of motion. The fields in (16) are operators and act on functions (states) in the indefinite, metric Hilbert space. For this reason, this method of quantization is called “operator method of quantization.” In order to ensure that physically meaningful degrees of freedom only contribute (the longitudinal and the time-like photons are unphysical) to the observables, we impose Gupta’s subsidiary condition on the photon states by

$$B^+(x)|\phi\rangle = 0, \quad (17)$$

where the superscript  $+$  denotes the positive frequency part of  $B(x)$ . The physical subspace in the indefinite metric Hilbert space is defined in (17). *The physical subspace  $V_{phys}$  is the totality of the states  $|\phi\rangle$  satisfying (17).* Now consider the expectation value of the quantum equations of motion (16) between *physical states*  $|\phi\rangle$  defined in (17). They are

$$\begin{aligned} \langle\phi|\partial_\mu F^{\mu\nu} - \partial^\nu B(x) + ej^\nu|\phi\rangle &= 0; \quad |\phi\rangle \in V_{phys}, \\ \langle\phi|\partial_\mu A^\mu + aB(x)|\phi\rangle &= 0. \end{aligned} \quad (18)$$

Using  $B^- = (B^+)^\dagger$  and (17), (18) becomes

$$\begin{aligned} \langle\phi|\partial_\mu F^{\mu\nu} + ej^\nu|\phi\rangle &= 0; \quad \forall |\phi\rangle \in V_{phys}, \\ \langle\phi|\partial_\mu A^\mu|\phi\rangle &= 0. \end{aligned} \quad (19)$$

Comparing (19) with (14), we see that the expectation value of the quantum equation of motion taken with the states in the physical subspace reproduces the classical

equations of motion, generalizing the Ehrenfest theorem to Abelian quantum field theory. Since the classical equation of motion is linear in  $A_\mu(x)$ , one can separate the positive and negative frequency parts and then the second equation above gives  $\partial^\mu A_\mu^+(x)|\phi\rangle = 0$ , subsidiary operator condition of Gupta. This feature is not shared by the non-Abelian theory as in this the classical equation of motion for the non-Abelian gauge field is non-linear and a separation into positive and negative frequency parts is not possible.

### 3 Non-Abelian Field Theory

As an example, we consider  $SU(3)$  gauge theory relevant to Quantum Chromo Dynamics (QCD), the gauge theory of the strong interactions of quarks. The *classical* lagrangian density is given by

$$\mathcal{L}_{YM} = -\frac{1}{4}F_{\mu\nu}^a F^{\mu\nu a} + gA_\mu^a j^{\mu a}, \quad (20)$$

where  $j^{\mu a}$  is the external source (color current of the quark);  $\mu$  and  $\nu$  are the Lorentz indices;  $a, b$ , and  $c$  are the  $SU(3)$  group indices;  $g$  is the coupling strength; and

$$F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + gf^{abc}A_\mu^b A_\nu^c. \quad (21)$$

In the above,  $f^{abc}$ s are the structure constants of  $SU(3)$  and  $g$  is also the coupling strength of the self-interaction of the non-Abelian gauge fields. The above lagrangian is gauge invariant. This can be verified by using the infinitesimal gauge transformation on the gauge field  $A_\mu^a$ , namely,

$$\begin{aligned} A_\mu^a &\rightarrow A_\mu^a + D_\mu^{ab}\omega^b, \quad \omega^a \in SU(3), \\ D_\mu^{ab} &= \partial_\mu\delta^{ab} + gf^{acb}A_\mu^c. \end{aligned} \quad (22)$$

Consider the first term in the lagrangian. Then it is found, using the Jacobi identity

$$f^{bcd}f^{dae} + f^{cad}f^{dbe} + f^{abd}f^{dce} = 0, \quad (23)$$

that

$$\delta_{\text{gauge}}(F^{\mu\nu a}F_{\mu\nu}^a) = 2gf^{acb}F^{\mu\nu a}F_{\mu\nu}^c\omega^b \equiv 0. \quad (24)$$

The *classical equations of motion* from (20) are given by

$$D_\mu^{ab}F^{\mu\nu b} + gj^{\nu a} = 0. \quad (25)$$

The operator  $D_\mu^{ab}$  in (22) is called the covariant derivative in the adjoint representation and using the Jacobi identity (23), it is found that the commutator  $[D_\mu, D_\nu]^{ab} = -gf^{abq}F_{\mu\nu}^q$ . Acting on (25) by  $D_\nu^{ca}$ , using the commutator, it is seen that

$$D_\nu^{ab} j^{vb} = 0, \tag{26}$$

i.e., the current  $j^{va}$  is covariantly conserved. As the source  $j^{\mu a}$  is gauge invariant, in the action integral, the second term in the lagrangian is invariant using (22) and (26) after one partial integration. Thus the lagrangian in (20) is gauge invariant.

Using the covariant derivative, the classical equation of motion (22) can be rewritten as

$$\begin{aligned} \partial_\mu F^{\mu\nu a} + gf^{acb}A_\mu^c F^{\mu\nu b} + gj^{va} &= 0, \\ \partial_\mu F^{\mu\nu a} &= -gJ^{va}, \text{ where} \\ J^{va} &\equiv j^{va} + f^{acb}A_\mu^c F^{\mu\nu b}. \end{aligned} \tag{27}$$

The current  $J^{va}$  contains besides the matter contribution, the non-Abelian fields. The non-Abelian fields themselves act as the source (like in gravity). By inspection, we see that  $\partial_\nu J^{va} = 0$ , i.e., the current  $J^{va}$  is ordinarily conserved.

An attempt to quantize (20) along the lines of the Abelian theory, i.e., “operator method of quantization,” runs into difficulty. The auxiliary fields  $B^a(x)$  in this case do not satisfy  $\square B^a(x) = 0$  due to the self-coupling property of the non-Abelian fields. So it is not possible to write down the positive and negative frequency parts. Furthermore the classical equations of motion are nonlinear. The proper method is to use the “path integral approach.” For the reasons given in the Abelian field theory, here also we need to fix the gauge to obtain the propagator for the gauge fields  $A_\mu^a(x)$ . Furthermore, in the “path integral method,” one integrates all possible gauge field configurations. As the lagrangian (20) is gauge invariant, two gauge field configurations related by gauge transformation will give the same lagrangian. This, in the path integral, amounts to double counting in the space of gauge fields. This is avoided by fixing the gauge and integrating over the space of gauge fields modulo gauge fixing. We choose the covariant gauge  $\mathcal{F}^a = \partial^\mu A_\mu^a(x) = 0$ .

The above gauge fixing relation, however, does change by the gauge transformation and so the gauge variation of the gauge fixing relation is non-trivial in the non-Abelian gauge theory. This, in the path integral approach, brings in the Faddeev–Popov ghost (anti-commuting scalars) fields. Using the results from the path integral approach [7], the lagrangian density for *quantum* non-Abelian theory can be written as

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^a F^{\mu\nu a} - \partial^\mu B^a A_\mu^a + \frac{\alpha}{2}B^a B^a - i\partial^\mu \bar{c}^a (D_\mu^{ab} c^b) + gj_\nu^a A^{\nu a}, \tag{28}$$

where  $\alpha$  is a gauge parameter and  $c^a$  are the ghost fields. They are hermitian

$$c^a = (c^a)^\dagger; \quad \bar{c}^a = (\bar{c}^a)^\dagger, \tag{29}$$

and the ghost fields  $c^a$  and  $\bar{c}^a$  anti-commute.

A comparison of (28) with (15) reveals that now we have (for  $SU(3)$ ) eight auxiliary fields  $B^a$  and a new term involving the Faddeev–Popov ghost fields. One can also quantize the Abelian massless field by the above procedure (path integral approach) and in that case, the ghosts decouple from the gauge fields. In contrast, in (28), the fourth term contains coupling of the ghost fields with the gauge fields. This is crucial. The second and the third terms in (28) are the gauge fixing part and the fourth term is the Faddeev–Popov ghost part  $\mathcal{L}_{FP}$ . Using (29) and the anti-commuting property of the ghost fields, it is seen that  $\mathcal{L}_{FP}^\dagger = \mathcal{L}_{FP}$ . The *quantum equations of motion* following from (28) are as follows:

$$\begin{aligned} D_\mu^{ab} F^{\mu\nu b} &= \partial^\nu B^a - g j^{\nu a} - i g f^{abc} (\partial^\nu \bar{c}^b) c^c, \\ \partial_\mu A^{\mu a} + \alpha B^a &= 0, \\ D_\mu^{ab} (\partial^\mu \bar{c}^b) &= 0, \\ \partial_\mu (D^{\mu ab} c^b) &= 0. \end{aligned} \tag{30}$$

Before considering the physical states, we recall that the quantum lagrangian (28) is gauge fixed. So, we do not have the local gauge invariance in (28). However, it was found by Becchi, Rouet, and Stora (BRS) [8] that (28) is invariant under a special global transformation (First Global Transformation) involving Faddeev–Popov ghosts. This BRS transformation is given by

$$\begin{aligned} \delta A_\mu^a &= D_\mu^{ab} c^b = [iQ, A_\mu^a], \\ \delta \psi &= i g c^a t^a \psi, \\ \delta B^a &= 0 = [iQ, B^a], \\ \delta c^a &= -\frac{g}{2} f^{abc} c^b c^c = \{iQ, c^a\}, \\ \delta \bar{c}^a &= i B^a = \{iQ, \bar{c}^a\}, \end{aligned} \tag{31}$$

where  $Q$  is the BRS-charge  $Q = \int d^3x \{ B^a (D_\mu^{ab} c^b) - \partial_0 B^a c^a + i \frac{g}{2} f^{abc} \partial_0 \bar{c}^a c^b c^c \}$  (see [7] for details). From (31), it is seen that  $\delta F_{\mu\nu}^a = g f^{acb} F_{\mu\nu}^c c^b$  and the invariance of (28) under (31) can be verified.

Although the local gauge invariance is explicitly broken by the gauge fixing, (28) has global gauge symmetry. This global gauge transformation (Second Global Transformation) given by

$$\begin{aligned} \Delta A_\mu^a &= f^{abc} \theta^b A_\mu^c, \\ \Delta \psi_i &= -i (t^a)_{ij} \theta^a \psi_j, \\ \Delta \bar{\psi}_i &= i \bar{\psi}_j (t^a)_{ji} \theta^a, \\ \Delta B^a &= f^{abc} \theta^b B^c, \\ \Delta c^a &= f^{abc} \theta^b c^c, \\ \Delta \bar{c}^a &= f^{abc} \theta^b \bar{c}^c, \end{aligned} \tag{32}$$

where  $\theta^a$  is the global gauge parameter, generates the conserved Noether current

$$\begin{aligned} \mathcal{J}_\mu^a &= f^{abc} A^{\nu b} F_{\nu\mu}^c + j_\mu^a + f^{abc} A_\mu^b B^c - i f^{abc} \bar{c}^b (D_\mu^{cd} c^d) + i f^{abc} \partial_\mu \bar{c}^b c^c, \\ &= J_\mu^a + f^{abc} A_\mu^b B^c - i f^{abc} \bar{c}^b (D_\mu^{cd} c^d) + i f^{abc} (\partial_\mu \bar{c}^b) c^c, \end{aligned} \quad (33)$$

where in the last step we used the third relation in (27).

We now consider the first equation in (30) and rewrite that as

$$\partial_\mu F^{\mu\nu a} + g f^{acb} A_\mu^c F^{\mu\nu b} = \partial^\nu B^a - g j^{\nu a} - i g f^{abc} (\partial^\nu \bar{c}^b) c^c. \quad (34)$$

This in view of (33) can be written as

$$\partial_\mu F^{\mu\nu a} + g \mathcal{J}^{\nu a} = (D^{\nu ac} B^c) - i g f^{abc} \bar{c}^b (D^{\nu cd} c^d). \quad (35)$$

The right side of (35) can be expressed, using the BRS transformations (31), as  $-i\delta(D^{\nu ab} \bar{c}^b)$  and so (35) becomes

$$\partial_\mu F^{\mu\nu a} + g \mathcal{J}^{\nu a} = \{Q, D^{\nu ab} \bar{c}^b\}. \quad (36)$$

This quantum equation of motion is to be compared with the classical equation of motion (27). We note that  $J^{\nu a}$  in (27) is replaced by  $\mathcal{J}^{\nu a}$  in (36) and the right side is expressed as a BRS variation. Both  $J^{\nu a}$  and  $\mathcal{J}^{\nu a}$  are ordinarily conserved. That the quantum equation (34) can be written in the form (36) was first shown by Ojima [9].

The vector space for the non-Abelian gauge fields, on which the quantum equations act, is an indefinite metric space. A physical subspace of this is to be defined. It was shown by Kugo and Ojima [10] that the physical space is defined by the condition

$$Q|\phi\rangle = 0. \quad (37)$$

Taking the expectation value of (36) between the physical states and using (37), it follows

$$\langle\phi|\partial_\mu F^{\mu\nu a} + g \mathcal{J}^{\nu a}|\phi\rangle = 0. \quad (38)$$

This expression when compared with the classical equation of motion (27) shows that the Ehrenfest theorem is not fully satisfied. The global conserved current  $\mathcal{J}^{\nu a}$  differs from the conserved current  $J^{\nu a}$ , as seen from (33). Now we consider (33) and note that this difference is given by  $f^{abc} A_\mu^b B^c - i f^{abc} \bar{c}^b (D_\mu^{cd} c^d) + i f^{abc} (\partial_\mu \bar{c}^b) c^c$ . The first two terms can be expressed using (31) as  $\delta(i f^{abc} \bar{c}^b A_\mu^c)$ , noting that when the BRS variation crosses the ghost field, it picks up a sign. So the first two terms can be rewritten as  $\{-Q, f^{abc} \bar{c}^b A_\mu^c\}$  and this when taken between the physical states vanishes. Then, (38) becomes

$$\langle\phi|\partial_\mu F^{\mu\nu a} + g J^{\nu a} + i f^{abc} (\partial^\nu \bar{c}^b) c^c|\phi\rangle = 0. \quad (39)$$

This still differs from the classical equation of motion by a term involving ghosts only.

We now take up the quantum lagrangian (28) and note that it is invariant under the scale transformation (Third Global Transformation)

$$c^a \rightarrow e^\alpha c^a ; \quad \bar{c}^a \rightarrow e^{-\alpha} \bar{c}^a, \quad (40)$$

with  $\alpha$  being a constant. This global transformation affects only the FP-ghost fields in (28). The Noether current corresponding to this transformation is given by

$$\begin{aligned} J_{gh}^\lambda &= \delta_\alpha c^a \frac{\partial \mathcal{L}}{\partial(\partial_\lambda c^a)} + \delta_\alpha \bar{c}^a \frac{\partial \mathcal{L}}{\partial(\partial_\lambda \bar{c}^a)}, \\ &= i\bar{c}^a (D^{\lambda ab} c^b) - i(\partial^\lambda \bar{c}^a) c^a, \end{aligned} \quad (41)$$

as  $\alpha$  is arbitrary. The corresponding conserved charge  $Q_{gh} = (Q_{gh})^\dagger$  is called the FP-ghost charge generating the above scale transformation on the ghost fields, leaving other fields invariant [7]. This is given by

$$\delta_{gh} c^a = [iQ_{gh}, c^a] = c^a ; \quad \delta_{gh} \bar{c}^a = [iQ_{gh}, \bar{c}^a] = -\bar{c}^a. \quad (42)$$

Using the above, the third term in (39) can be written as

$$\begin{aligned} i f^{abc} (\partial_\mu \bar{c}^b) c^c &= -\frac{1}{2} \delta_{gh} (i f^{abc} (\partial_\mu \bar{c}^b) c^c), \\ &= \frac{1}{2} [Q_{gh}, f^{abc} (\partial_\mu \bar{c}^b) c^c], \end{aligned} \quad (43)$$

as  $\delta_{gh}$  when crosses a FP-ghost field picks up a sign.

We defined the physical subspace in (37) as the assembly of states in the indefinite metric Hilbert space annihilated by the BRS-Charge. We now restrict the physical subspace further by another subsidiary condition

$$Q_{gh} |\phi\rangle = 0. \quad (44)$$

Then, using (43) in the last term in (39) and in view of the further restriction (44) on the physical states, (39) becomes

$$\langle \phi | \partial_\mu F^{\mu\nu a} + g J^{\nu a} | \phi \rangle = 0, \quad (45)$$

showing that the expectation value of the quantum equation of motion for the non-Abelian gauge fields agrees with the classical equation of motion (27).

Now we examine the other quantum equations of motion in (30). The second equation in (30), in view of the BRS-transformation (31), can be written as  $\partial_\mu A^{\mu a} + \alpha \{Q, \bar{c}^a\} = 0$ , which when its expectation value between the physical states defined in (37) is taken gives  $\langle \phi | \partial_\mu A^{\mu a} | \phi \rangle = 0$ , giving the gauge fixing

condition. The third equation in (30), in view of the third global transformation (42), is written as  $[iQ_{gh}, (D_\mu^{ab}(\partial^\mu \bar{c}^b))] = 0$  and its expectation value taken between the physical states vanishes on account of (44). The fourth equation in (30), using the BRS-transformation (first global transformation), becomes  $[iQ, \partial_\mu A^{\mu a}]$  whose expectation value between the physical states vanishes on account of (37). This shows the validity of the Ehrenfest theorem for the quantum non-Abelian theory. We have made use of three global transformations to arrive at this conclusion.

## 4 Summary

The Ehrenfest theorem in quantum mechanics is shown to be satisfied in the quantum field theory by suitably taking the physical subspace for the gauge fields. In the Abelian quantum field theory, the one subsidiary condition on the physical states of the photon is enough to show this. In the case of non-Abelian field theory, the subsidiary condition (37) is not enough and one has to further restrict the physical space by (44). Then the expectation value of the quantum equations of motion between the physical states satisfying (37) and (44) agrees with the classical equations of motion, including the gauge fixing condition.

## References

1. P. Ehrenfest, Zeits. f. Physik, **45** (1927) 455.
2. L.I. Schiff, *Quantum Mechanics*, McGraw-Hill. Second Edition, 1955.
3. J.D. Bjorken and S.D. Drell, *Relativistic Quantum Fields*, McGraw-Hill, 1965.
4. S.N. Gupta, Proc. Phys. Soc. **A63** (1950) 681; K.Bleuler, Helv. Phys. Acta. **23** (1950) 567.
5. P.M. Mathews, M. Seetharaman and M.T. Simon, Phys. Rev. **D9** (1974) 1700.
6. N. Nakanishi, Prog. Theor. Phys. **35** (1966) 1111.
7. N. Nakanishi and I.Ojima, *Covariant Operator Formalism of Gauge Theories and Quantum Gravity*, World Scientific, 1990; L.H. Ryder, *Quantum Field Theory*, Cambridge University Press, 1996.
8. C. Becchi, A. Rouet and R. Stora, Ann. Phys. **98** (1976) 287.
9. I. Ojima, Nucl. Phys. **B143** (1978) 340.
10. T. Kugo and I. Ojima, Phys. Lett. **73B** (1978) 459. T. Kugo and I. Ojima, Prog. Theor. Phys. **60** (1978) 1869.