Proceedings of
Telangana Academy of Sciences

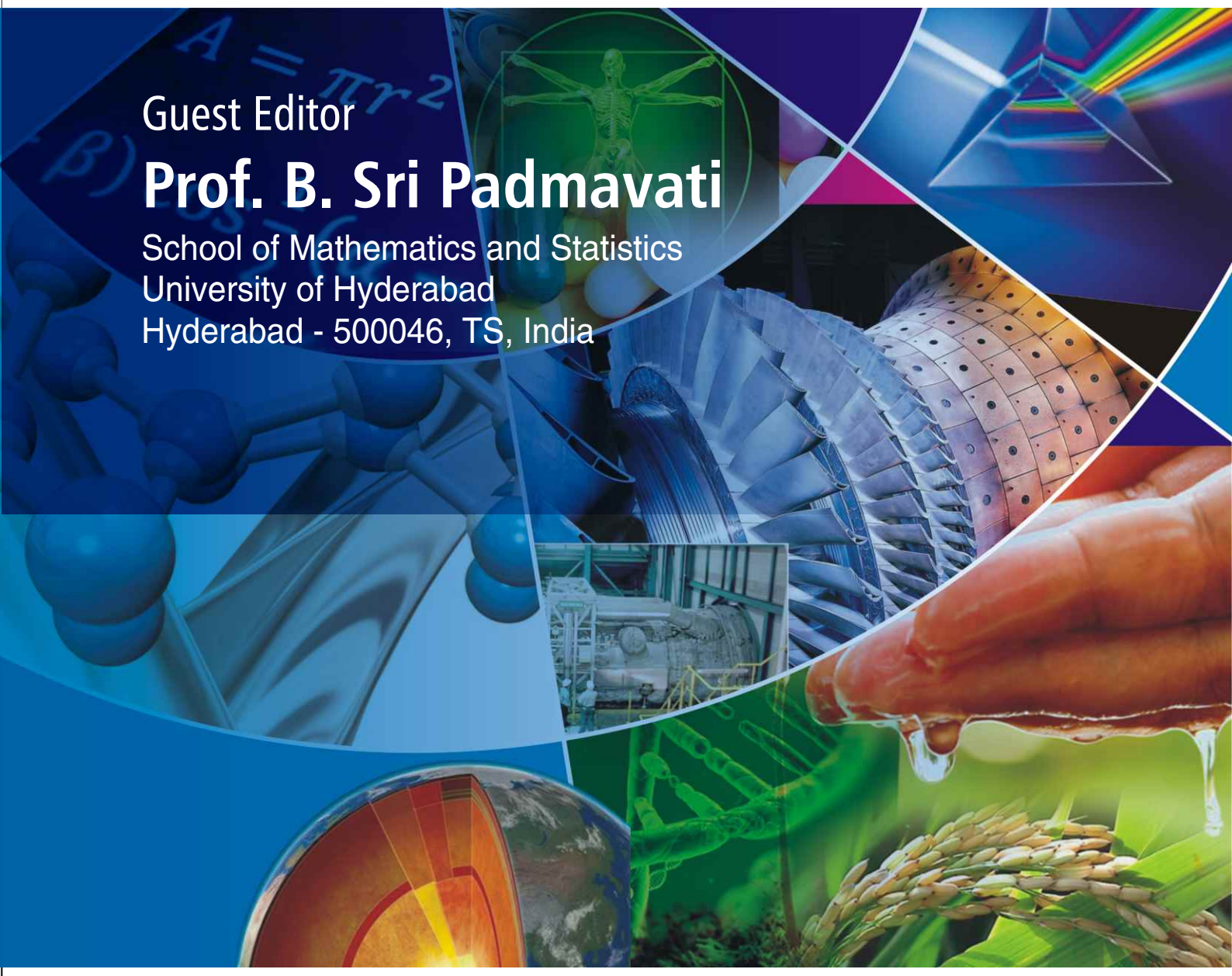
TAS

Special Issue

Vol.1, Issue 1, 2020

# Mathematical Sciences

## Frontiers in Mathematics



Guest Editor

**Prof. B. Sri Padmavati**

School of Mathematics and Statistics
University of Hyderabad
Hyderabad - 500046, TS, India

TAS

# Telangana Academy of Sciences

## Executive Council - 2018-2020

**Prof. K. Narasimha Reddy** **President**
(Former V.C., M. G Univ., Nalgonda)

**Prof. P. VenugopalReddy** **Vice-President**
(Dept. of Physics, OU, Hyderabad)

**Prof.Ch. SanjeevaReddy** **Vice-President**
(Dept. of Chemistry, K.U. Warangal)

**Dr. S. Chandra Sekhar** **Hon. Secretary**
(Director, CSIR-IICT, Hyderabad)

**Dr. Y. Purushotham** **Hon. Treasurer**
(C-MET, Hyderabad)

**Prof. S.M. Reddy** **Editor of Publications**
(Dept. of Botany, K.U. Warangal)

### Executive Council

**Prof. G. Bagyanarayana**
(Former VC, Palamuru University)

**Dr. G. Bhanuprakash Reddy**
National Institute of Nutrition (NIN), Hyderabad

**Dr. V. Chakravarthi**
(University of Hyderabad, Hyderabad)

**Dr. G. Madhusudhana Reddy**
(Outstanding Scientist & Director, DMRL, Hyderabad)

**Dr. G. Sundararajan**
(IIT-Madras, Hyderabad)

**Dr. G. Parthasarathy**
(NGRI, Hyderabad)

**Dr. K. Jeevan Rao**
(PJTSAU, Hyderabad)

**Prof. K. Janardhan Reddy**
(Dept. of Botany, OU, Hyderabad)

### Ex-Officio Members

**Dr. Ch. Mohan Rao**
(Former Director, CSIR-CCMB, Hyderabad)

**Principal Secretary**
(Higher Education Dept. Govt. of Telangana)

## Editorial Board

Editor of Publications

**Dr. S. M. Reddy**
H.No: 12-13-1126/B-305
Kalakriti Tarnaka - 500 017,
Profsmreddy@yahoo.com.,

**Dr. G.V. Madhava Sharma**
Chief Scientist, CSIR-IICT, Hyderabad-500 007
sharmagvm@gmail.com

**Dr.G. Bhanuprakash Reddy**
Scientist-F, NIN, Hyderabad - 500007
geereddy@yahoo.com

**Dr. M. Sujatha**
Pr. Scientist, Oil seeds Research,
Hyderabad - 5000030
mulpurisulata@yahoo.com

**Prof D. Prasada Rao**
Chairman, Indo-US Super Specialty Hospital,
Hyderabad
prasadaro.dasari@reddiffmail.com

**Prof. V. Malla Reddy**
H.No.-1-9-30(C-77),
Ravindranagar Colony, Road No - 1,
Hyderabad - 500 007
vangamailareddy@yahoo.coin

**Dr. G. Madhusudhana Reddy**
Scientist G, DMRL, Hyderabad-500 058
gmreddy dmrl@yahoo.coin

**Dr. G Parthasarathy**
Chief Scientist, NGRI, Hyderabad - 500 007
drg.parthasarathy@gmail.com

**Prof. M. Ghanashyam Krishna**
School of Physics, UOH, Hyderabad - 500 046
mgksp@uohyd.acin

**Prof Bir Bahadur**
The Odyssey 4a, 10-2-289/46,
Shanthinagar, Masab Tank, Hyderabad - 500 028
brirbahadur5april@gmail.com

**Dr.R.B.N. Prasad**
Chief Scientist, Iict Hyderabad - 500007
rbnprasad@gamil.com

**Dr. B. Sri Padmavati**
Professor,
School of Mathematic & Statistics, University of
Hydebad, Gachinowli, Hyderabad - 500046.
Email: bs.padmavathi@gmail.com

Telangana Academy of Sciences

## Regional Coordinators

**Hyderabad & Ranga Reddy**
**Prof. K. Janardhan Reddy**
H.No.11-14- 234/B3, Road No.11,
Haripriya Colony, Saroornagar, Hyderabad-500 035.
kjreddy50@yahoo.co.in

**Karimnagar & Adilabad**
**Prof. M.Vithal**
Dept. of Chemistry,
Osmania University, Hyderabad-500 007
mugavithal@gmail.com

**Nizamabad & Medak**
**Dr. Y. Purushotham**
Scientist 'D' C- MET, Hyderabad-051
ypurushotham@cmet.gov.in

**Warangal & Khammam**
**Prof. S. M. Reddy**
H.No: 12-13-1126/B-305 Kalakruti, Tarnaka - 500017,
profsmreddy@yahoo.com

**Nalgonda & Mahboobnagar**
**Dr. G. Bhanuprakash Reddy**
Scientist-F, NIN, Hyderabad - 500007
geereddy @yahoo.com

# FOREWORD

The Telangana Academy of Sciences (TAS) has been engaged in the Science popularization activities including Publications, Journals and Special Issues on different occasions. The recent publication activities include "*Diabesity: The Unacknowledged Indian Knowledge"*, "*Jivanayanam lo Rasayanalu* " and "*21st Century Noble Awards in Chemistry"*. As a part of its continued efforts in this direction, a Special Issue of the Proceedings of the Telangana Academy of Sciences entitled ''Frontiers In Mathematics'' has been brought out now with Prof. B. Sri Padmavati, School of Mathematics and Statistics, University of Hyderabad as a Guest Editor.

This issue contains different research topics of Mathematics with leading authors drawn from different national and international Centers of Excellence including California Institute of Technology, Emory University of USA, besides premier institutions like TIFR, Chennai Mathematical Institute, University of Hyderabad, IIT Delhi- Kanpur- Palakkad-Kharagpur-Hyderabad , University of Mumbai, Indian Statistical Institute, Kolkata. I hope this Proceedings will serve as a pace setter for detailed investigations in Mathematics and Applied Mathematics and be beneficial for the young researchers in the field and also for aspiring mathematicians.

I express my gratitude to Prof. B. Sri Padmavati, Guest Editor of the special issue and to Prof. S.M. Reddy, Editor of Publications for bringing out such an excellent special issue entitled "Frontiers in Mathematics".

**S. Chandrasekhar**
Hon. Secretary, TAS

# From the Guest Editor

It gives me immense pleasure to bring out this special issue of the Proceedings of Telangana Academy of Sciences entitled “Frontiers in Mathematics” as a Guest Editor. This issue includes articles by some very eminent and senior mathematicians and also by some promising young mathematicians.

This issue contains and assortment of topics in mathematics and applied mathematics that will appeal to serious researchers in quest of original results, to aspiring mathematicians trying to get a flavour of some research topics in mathematics and also to those who just enjoy and love reading about mathematics and mathematicians.

I thank all the authors who contributed their original results, all those who contributed expository articles and also those who wrote some other very interesting, insightful and illuminating articles for this special issue. I also thank all the reviewers who spared their precious time to send their valuable comments and suggestions to the authors.

**Prof. B. Sri Padmavati**
Guest Editor

## From the Editor of Publications

Telangana Academy of Sciences, the erstwhile AP Akademi of Sciences, is one of the oldest of the state Academies of the country, and has been serving the cause of science in particular and society at large in a befitting manner. It has been publishing scientific documents of international, national and regional relevance since its inception in 1963. It has been publishing the official journal of the academy covering the areas of Life Sciences, Chemical, Physical, Engineering and Earth Sciences. In view of its long standing contributions, we are making earnest efforts to make it vibrant which can bring name and fame to the Telangana Sciences in particular and to India at large.

India has made great strides in different fields of sciences and made remarkable efforts to realize different revolutions such as green, white, blue, yellow and rainbow which made India one among different countries in the world to become self-sufficient in agricultural products. Even though our R&D contributions in the world stand at 3.2%, no institute stands in the list of top 200 research organizations in the world, even though the contributions of some individuals are of high standard. There is a need for concerted efforts in basic research which can be translated into world class technologies.

The Academy has initiated printing of special issues and books on popular sciences in which established researchers were requested to contribute for this purpose. We are now ready with some special issues including the present one on 'Mathematics'. I thank Prof. B. Sri Padmavati, Guest Editor, and all the contributors for their valuable articles.

We also look forward to the whole hearted co-operation from the Fellows of TAS through their contributions of high quality scientific and research articles for publication in these Proceedings.

**S.M. Reddy**
Editor of Publications, TAS

## Contents

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 1–11

# *Some Highlights of Indian Contributions to Mathematics in the* $20^{th}$ *Century* ‡

**M S Raghunathan***

*University of Mumbai - DAE Centre for Excellence for Basic Sciences, Mumbai 400098.*

**Abstract:** In this paper we describe some of the major research works of Indian mathematicians in the twentieth century. These researches attracted considerable international attention and in some cases are in fact landmarks in their mathematical areas.

## 1. Prognosis of Andre Weil

Andre Weil (1906-98), one of $20^{th}$ century's greatest mathematicians, addressing the Moscow Mathematical Society in 1935, said:

"The intellectual potentialities of the Indian nation are unlimited and not many years would perhaps be needed before India can take a worthy place in world mathematics".

It is now 85 years since that pronouncement. Have our mathematicians lived up to Weil's expectations of them? Have we taken a "worthy place" in world mathematics? Before answering that question one needs to ask what is a "worthy place". Our mathematicians have made highly significant contributions over those 85 years and some of those contributions had a big role in the very evolution of the field. That said, the influence our mathematics has had is not comparable to those of the US, Russia, Britain, France, Germany and Japan. But we are ahead of the rest of the Western world and the developing countries including China (which however is currently challenging our primacy there). A nuanced response to the question then is "No, not yet, but we are getting there.". The progress achieved in the $20^{th}$ century was considerable and the momentum of that progress has been maintained in the last two decades.

Weil was aware of what was going on in mathematics in India: he had spent two years (1930-32) at the Aligarh Muslim University (AMU). He had contacts with several Indian mathematicians and was basing himself not just on the work of the legend Ramanujan (1887-1920) (who was instrumental in making the world sit up and take notice of us). There is a Telangana connection to Weil's Indian sojourn. Syed Ross Masood (1889-1937) who was the education minister of the princely state of Hyderabad in 1930 quit his job to take up the vice-chancellorship of AMU and one of his first initiatives as vice-chancellor was the appointment of the (then

‡This is a slightly expanded version of a talk with the same title given in the School of Mathematics and Statistics of the University of Hyderabad on 17th January, 2020

*Email: madabusisr@gmail.com

26-year old) Weil as Professor and Head of the Department of Mathematics at that university.

This article's aim is to highlight some Indian achievements of high significance in world mathematics in the $20^{th}$ century. It is constrained by three factors, all personal. I have written only about areas in pure mathematics with which I have some acquaintance; secondly even in those areas, it is on the cards that I have missed out some significant work by oversight; also what I rate as significant is evidently a matter of personal judgement which of course is far from infallible; third: I may have missed out on work done in the last decade of the last century - I was much less alert to what was going on in mathematics in those years than in the previous years. I have not touched upon most areas of applied mathematics (that includes Statistics where India has a big presence) at all and as the title indicates any mathematics of the present century.

Finally, in most cases I have not been able to state the theorems proved and even where I have attempted to do it, the statements would be accessible only to mathematicians with some knowledge of the relevant field - the technical concepts involved are not explained. This is inevitable as cutting edge research in most mathematical areas requires considerable back-ground in the area.

I have not given any references. This is partly because listing the various works would make a rather unwieldy bibliography. In any case since I have given only a broad idea of the work of the various people who were important players on the mathematical scene, pin-pointing references may give a misleading picture. In any case the interested reader can always go to the Mathscinet and look through the reviews to gt a better idea of the contributions of these mathematicians.

## 2. Before Ramanujan

There was one piece of work that drew international attention long before Ramanujan came on the scene and has had a sustained impact in mathematics: the "Four vertex theorem" due to Shyamadas Mukhopadhyaya (1866-1937). In 2009 there was a conference to commemorate the centenary of Mukhopadhyaya's paper in which the theorem was proved, an indication of the importance attached to the work. I will briefly describe what Mukhopadhyaya proved. Recall that for a smooth plane curve $C : [0, 1] \to \mathbb{R}^2$ and a point $C(t)$ on it the osculating circle at $C(t)$ is the circle passing through $C(t)$ and having a second order contact with it; the curvature $\kappa(t)$ of $C$ at $C(t)$ is the reciprocal of the radius of the osculating circle at $C(t)$. Mukhopadhyaya's theorem says that *if $C : [0, 1] \to \mathbb{R}^2$ is a simple closed convex curve (convex means that $\kappa(t) > 0$ for all $t \in [0, 1]$), the curvature function $\kappa(t)$ has at least four stationary points.* This was the first result in global differential geometry proved by an Indian mathematician. The result has since been generalized to *any* simple closed curve (without any restriction on the curvature).

It needs to be noted that Mukhopadhyaya's work was done before Ramanujan appeared on the scene. One can be sure that Weil's optimistic prognosis took into account Mukhopadhyaya's work.

## 3. Ramanujan

Ramanujan is a once in a century phenomenon. His work was essentially in Number theory; it covered three distinct themes in each of which his work broke new ground. There is a large body of work proving various identities between essen-

tially formal infinite algebraic expressions (products or series), most of them with implications in Number theory. A second theme is about the partition functions where he introduced what is known as the circle method which became a powerful tool for solving many problems, some of which are quite unrelated to the partition function. The partition function $p : \mathbb{N} \to \mathbb{N}$ is the function that associates to $n \in \mathbb{N}$ the number $p(n)$ of ways in which $n$ can be written as a sum of positive integers. And Ramanujan, in collaboration with G H Hardy (1877-1947) proved beautiful theorems about the behaviour of the function $p$.

Arguably the most important work of Ramanujan relates to the theme of modular forms. A modular form of weight $k$ is a holomorphic function $f : \mathbb{H} \to \mathbb{C}$ on the upper half plane $\mathbb{H} = \{z \in \mathbb{C} |\ Im(z) > 0\}$ 'satisfying the following condition: for $z \in \mathbb{H}$ and $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ in $SL(2, \mathbb{R})$, $f((a \cdot z + b)/(c \cdot z + d) = (c \cdot z + d)^k \cdot f(z)$. Taking $a = b = d = 1$ (and $c = 0$), we see that $f$ is periodic with period 1: $f(z + 1) = f(z)$ for all $z \in \mathbb{H}$. It follows that $f$ has a fourier expansion $\sum_{n=0}^{\infty} a_n(f) \cdot e^{2\pi inz}$. Ramanujan defined a fuction $\tau : \mathbb{Z} \to \mathbb{Q}$ by the formal identity:

$$\sum_{n=1}^{\infty} \tau(n) q^n = q \cdot \prod_{n=1}^{\infty} (1 - q^n)^{24}.$$

Then the function $\Delta : \mathbb{H} \to \mathbb{C}$ defined by setting for $z \in \mathbb{H}$, $\Delta(z) = \sum_{-\infty}^{\infty} \tau(n) \cdot e^{2\pi inz}$ is a modular form of weight 12. Moreover every modular form $f$ weight 12 and with $a_0(f) = 0$ is a scalar multiple of $\Delta$. Ramanjan conjectured that certain congruence relations hold for the $\tau(n)$ and that $|\tau(n)| \leq n^{11/2}$. The conjecture evoked a lot of interest in many leading number theorists and one of them, Mordell (1888-1972), proved the first part of the conjecture. It took half a century for the second part to be proved: it was done by the Fields Medallist Deligne (1944-). Ramanujan had other insights related to modular forms, but I will not go into these.

## 4. Number Theory: Diverse topics

During the three decades 1925-1955 a number of Indian mathematicians were making interesting contributions to diverse topics in Number Theory. Notable among these were S S Pillai (1901-50), T Vijayaraghavan (1902-55), Hansraj Gupta (1902-88) and R P Bambah (1925-). I will later describe some of Pillai's work. The works of the other three attracted considerable attention, even if not to the same extent as Pillai's work on the Waring Problem. Bambah's work is largely in Geometry of Numbers and he (along with Hansraj Gupta) built a school in the area at Panjab University, Chandigarh.

## 5. Number Theory: Automorphic Forms

In the early sixties K Chandrasekharan (1920-2017) and Raghavan Narasimhan (1937-2015) wrote a series of joint papers on Dirichlet Series in the prestigious Annals of Mathematics. Dirichlet Series have close connections with modular forms.

Another landmark in Indian mathematics to have a big impact in number theory is the publication of an important paper by K Ramachandra (1933-2011) in the sixties. In a paper entitled "Some applications of Kronecker's Limit Formu-

las"published in the Annals of Mathematics), Ramachandra showed that a "ray class field" over an imaginary quadratic field $K/\mathbb{Q}$ is generated by the value of a certain modular form at a suitable point. He also obtained beautiful results about units in ray class fields.

The theory of modular forms has undergone a big transformation since Ramanujan. In the first instance the notion was generalized to functions on what is known as the Siegel upper half space. In the sixties Harish-Chandra introduced the very general concept of an automorphic form as a function on a semi-simple Lie group invariant under the left action of an arithmetic discrete subgroup and transforming according to certain rules under the action the maximal compact subgroup on the right. Investigations of these automorphic forms led people to study representations of real and $p$-adic algebraic groups. In the sixties there was some work of significance on Siegel modular forms by Indian mathematicians, notably S Raghavan (1934 -2014). Work of much greater significance on representations of $p$-adic groups connected with automorphic forms came from Dipendra Prasad (1960-), (currently working in IIT Bombay), in the last decade of the $20^{th}$ century. I will not attempt to describe the work - it is far too technical for this account. However the importance of the work can be gauged by the fact that in recent years there have been several conferences on the "Gross-Prasad conjectures" which were formulated already in Dipendra Prasad's thesis (done under the supervision of Benedict Gross of Harvard).

## 6. Number Theory: Waring's Problem

The Waring's problem in Number Theory is the following question raised by Edward Waring a British mathematician in 1770:

*given a positive integer $k$, does there exist a (minimal) integer $g(k)$ such that every positive integer is the sum of $g(k)$ $k^{th}$ powers of non-negative integers.*

This question had already been raised in the case $k = 2$ by the Greek mathematician Diophantus (circa 300 CE) who had in fact conjectured that $g(2) = 4$ and J L Lagrange (1736-1818) proved that conjecture in 1770 (which is probably what led Waring to ask his question). In 1909 David Hilbert (1862-1943) answered Waring's question in the affirmative. After that mathematicians were after giving an exact formula for the number $g(k)$ for all $k$. By the thirties, when the Indian mathematician S S Pillai got interested in the problem, $g(3)$ and $g(5)$had been determined. Pillai, basing himself on some work of the Russian mathematician Vinogradov (1891-1983) came up with a formula for $g(k)$ for all $k \geq 6$. The same result was also obtained by the American mathematician L E Dickson (1874-1954). That has ensured that Pillai's name will figure in any history of mathematics. Pillai who was born in 1901 died tragically in an air accident over Cairo in 1950 when he was on his way to the International Congress of Mathematician where he was to give a talk. So by 1950 $g(k)$ had been determined for all $k \neq 4$. It was an Indian mathematician - R Balasubramanian (1951-), former Director, Institute of Mathematical Sciences, who came up with the proof that $g(4) = 19$ (in collaboration with two Frenchmen).

That is not the end of the engagement of Indians with the Waring problem. C L Siegel (1896-1981) generalizing the work of Hilbert proved the following

THEOREM 1 *Let $F$ be a number field $\mathfrak{O}_F$ its ring of integers. Then for any integer $k \geq 1$, there is an integer $g(k, F)$ such that any element in $\mathfrak{O}_F$ which can be expressed as a sum (of an unspecified number) of $k^{th}$ powers of elements of $\mathfrak{O}_F$ can*

*be expressed as a sum of $g(k, F)$ of $k^{th}$ powers of elements of $\mathfrak{O}$.*

Siegel went on to conjecture the following:

*There is a positive integer $G(k)$ (independent of the number field $F$) and a positive integer $\nu(F)$ such that all elements $\alpha \in \mathfrak{O}$ with $N_{F/\mathbb{Q}}(\alpha) \geq \nu(F)$ which can be expressed as a sum of $k^{th}$ powers of elements of $\mathfrak{O}_F$ can be expressed as a sum of $G(k)$ $k^{th}$ powers of elements of $\mathfrak{O}_F$.*

Around 1964, C P Ramanujam (1937-1974) gave a brilliant short proof of a stronger result for $p$-adic fields and deduced Siegel's conjecture. Altogether, Indian mathematicians had a big role in advancing this area substantially.

## 7. Differential Geometry / Differential Equations

The heat equation is the differential equation that governs the propagation of heat: it is the partial differential equation:

$$\partial u(x,t)/\partial t - \Delta(u(x,t)) = F(x,t)$$

Ganesh Prasad (1876-1935) studied this equation and obtained some interesting results and a paper he wrote in 1903 incorporating this was solicited by the famous mathematician- F Klein for publication in a German journal of which he was an editor. Ganesh Prasad's work however is perhaps more of a contribution to physics than to mathematics.

Minakshisundaram (1913-1968) was more interested in purely mathematical questions relating to the heat equation and that in a more general context. There is a natural elliptic differential operator $\Delta$ of order 2 on any closed compact Riemannian manifold $M$ called the Laplacian of $M$. This enables one to talk of the heat equation $\partial u(t,x)/\partial t - \Delta(u,x) = f(t,x)$ on $\mathbb{R}^+ \times M$. The heat kernel is the function $K(t,x,y)$ on $\mathbb{R}^+ \times (M \times M \backslash D)$ where $D = \{(x,x)|\ x \in M\}$ such that $\partial K(t,x,y)/\partial t - \Delta_x(K(t,x,y) = 0$ and $Lim_{t\to 0} \int_M K(t,x,y)f(x)dx = f(y)$ for all $y \in M$. Here $dx$ is the volume form on $M$. In 1949, Minakshisundaram, in collaboration with a Swedish mathematician Pleijel, gave an asymptotic expansion in powers of $t$ of the form $t^{-n/2+r}$, $r \in \{0\} \cup \mathbb{N}$ for the function $K(t,x,y)$ as $t$ tends to 0. As a corollary they could provide interesting information on the growth of the eigenvalues of the Laplacian on compact Riemannian manifolds. The work is at the interface of Differential Equations and Differential geometry and has many interesting off-shoots in both the areas and so had a big impact on both areas.

I M Singer and H P Mckean (1930-) proposed in 1969 a new approach to proving the Atiyah-Singer index theorem - one of the great theorems of the 20th century - making use of the theorem of Minakshisundaram and Pleijel. An Indian mathematician V K Patodi (1945-1976) showed that the proposed approach works in a series of brilliant papers proving many special cases of the index theorem culminating in a joint work with M F Atiyah (1929-2019) and R Bott (1923-2005) where the theorem itself was proved. Patodi went on to produce more interesting work; sadly he was cut off in his prime by health problems. Unfortunately I cannot go into the statement of the index theorem as it needs a lot of back-ground material.

In a totally different direction M S Narasimhan (1932-)and S Ramanan (1937-) proved the existence of a universal connection on the classifying bundle of a Lie group: If $G$ is a Lie group and $B_G(n)$, the classifying space for $G$-bundles on manifolds of dimension $\leq n$ there is a connection $\omega_n$ on the classifying bundle $U_G(n)$ such that any connection on a $G$-bundle $E_G$ on a manifold $M$ of dimension

$n$ is the pull back of $\omega_n$ by a bundle map $\tilde{F} : E_G \to U_G(n)$.

## 8. Algebra

I have so far talked areas of mathematics where significant advances were made in the pre-independence days and continued by the generation that came of age post-independence. Commutative Algebra saw its first major Indian contributions of importance in the late fifties when C S Seshadri (1932-) made the first dent in the following conjecture of J-P Serre (1926-):

Conjecture. *Every projective module over a polynomial ring $k[X_1, X_2 \cdots, X_n]$ in $n$ variables over a field $k$ is free.*

Many leading algebraists were after this problem when Seshadri came up with a proof of the conjecture when $n = 2$ (the case $n = 1$ had been known for a long time, well before Serre made the conjecture). This was followed up with many interesting related results by M Pavaman Murthy (1935-) (a student of Seshadri at the Tata Institute of Fundamental Research). Incidentally Pavaman Murthy is an alumnus of Osmania University and is to-date the best mathematician to come out of Telangana.

The Serre conjecture for polynomial rings in 3 variables was proved by Pavaman Murthy in the early seventies. Soon after (in 1976) the full conjecture was settled by D Quillen (1940-2011) and A Suslin (1940-) independently; and this triggered a new interest in the problem, or rather in related problems at TIFR. Specifically the question attacked is the following one:

*Is any non-degenerate quadratic form on a free module over a polynomial ring over a field $k$ equivalent to one with coefficients in $k$?*

Raman Parimala (1948-) came up with a negative answer to the question soon after the Quillen-Suslin theorem in a paper closely related to an earlier work of hers done in collaboration with R Sridhran (1935-). A little later M S Raghunathan (1941-) came up with an affirmative answer to the question when $k$ is algebraically closed. The question can be reformulated in algebraic geometric language as a question about principal orthogonal bundles on affine space; and in that formulation, it can be posed for any principal $G$-bundle, $G$ an algebraic group. Raghunathan handled this problem giving comprehensive answers.

Significant contributions have come from S M Bhatwadekar (1946-) and Ravi A Rao (1954-).

## 9. Algebraic Geometry: Affine varieties

The problem discussed in the last section as I said,really belongs with *affine* algebraic geometry, but affine algebraic geometry is essentially an avatar of commutative algebra; even so many questions it deals with are difficult to formulate elegantly in the language of commutative algebra. Here is a beautiful result proved by Ramanujam.

THEOREM 2 *Let $X$ be an affine variety over $\mathbb{C}$ of complex dimension 2 which is contractible and simply connected at infinity. Then $X$ is isomorphic to $\mathbb{C}^2$ as an affine variety.*

This set in motion extensive studies of the topology of complex algebraic surfaces.There are interesting results due to R V Gurjar (1950) and A R Shastri

(1948-) to do with the topology of affine surfaces (that is complex affine algebraic varieties of complex dimension 2). An important paper of Gurjar, written in collaboration with the Japanese mathematician M Miyanishi (1940-) eventually led to the classification of homology planes i.e, affine surfaces which have the same homology as the affine plane.

Gurjar also proved interesting results about finite group actions on affine varieties as also did Shrawan Kumar (1953-).

## 10. Algebraic Geometry: Cohomology Vanishing Results

A result of Ramanujam shed light on what is known as the Kodaira Vanishing Theorem, a theorem which asserts the vanishing of cohomology groups of certain kinds of line bundles on a smooth projective variety over complex numbers. The original proof of Kodaira of this theorem made use of Hodge theory. Ramanujam came up with a proof that uses only topology, and no analysis (Hodge theory involves some deep analysis). This result has been very influential.

A path-breaking paper of Vikram Mehta (1946-2014) and A Ramanathan (1946-93) in which they showed that a large class of varieties (over fields of positive characterestic) are "Frobenius-split", opened up a completely new approach to cohomology vanishing theorems for varieties in positive characteristic. Analogous results are proved using the Kodaira vanishing theorem in characteristic 0.

## 11. Algebraic Geometry: Moduli

Mathematicians are often interested in classifying mathematical structures up to isomorphisms. This circle of problems are called moduli problems. One such problem is the study holomorphic vector bundles on a compact Riemann surface up to isomorphism. The American mathematician David Mumford introduced the notion of a "stable" vector bundle and showed that the set of isomorphism classes of stable bundles of fixed rank and degree has a natural structure of an algebraic variety. M S Narasimhan and C S Seshadri gave a remarkable "transcendental" characterization of stable bundles. Andre Weil had shown that an in-decomposable holomorphic vector bundle of rank $r$ and degree $d$ with $0 \leq d < r$ on a surface $X$ of genus $g$ arises in a natural fashion from an $r$-dimensional representation of a certain Fuchsian group $\Gamma_r$ acting on the upper half plane $\mathbb{H}$ so that $\Gamma_r \backslash \mathbb{H} \simeq X$. The Narasimhan-Seshadri theorem asserts that a bundle of rank $r$ and degree $d$ is stable if and only if it arises from a *irreducible unitary* representation of $\Gamma_r$. This beautiful result has triggered a lot of interesting work on vector bundles by other Indian mathematicians, notably S Ramanan, A Ramanathan, Vikram Mehta, Nitin Nitsure (1957-), Indranil Biswas (1964-), V Balaji (1962-) ..... and is being continued by a number of younger mathematicians. Some of this extends the results of Narasimhan and Seshadri to principal bundles while others are contributions to the study of vector bundles on higher dimensional varieties.

Both Seshadri and Narasimhan have also studied non-stable bundle *in extenso*. Narasimhan in a joint paper with G Harder (1938-) proved interesting results about the topology of moduli spaces of stable vector bundles and their compactifications (introduced by Seshadri). In that paper they introduced what is now called the Harder- Narasimhan filtration which is a key concept in the study of non-stable bundles. Altogether Indian work on this circle of problems constitutes a very substantial part of all developments in the second half of the $20^{th}$ century.

## 12. Algebraic Geometry: More on Projective Varieties

C S Seshadri initiated a study of complete homogeneous spaces of reductive algebraic groups taking off from what the British school had done earlier. The "standard monomials" theory developed by him in collaboration with C Musili (1941-2005) and V Lakshmibai (1950) has shed a great deal of light on these varieties and their sub-varieties known as Schubert varieties. Incidentally Musili was on the Mathematics Faculty of the University of Hyderabad.

Madhav Nori (1949-) defined the Fundamental Group Scheme of a scheme over a field (thereby settling a conjecture of Grothendieck). His elegant construction of this object drew instant attention and is a pioneering work. M V Nori has numerous other important and elegant results in algebraic geometry, among them new and elegant proofs of the Hirzebruch Riemann Roch Theorem, an example of a projective variety whose fundamental group is not residually finite; but a more detailed description would be somewhat lengthy and so I will not attempt it here. I must mention in particular his outstanding work on "motives" which is getting its long over due attention now.

## 13. Representation Theory

Representation theory of semi-simple Lie groups has held a central place in mathematics since the fifties. In the early days the focus was on finite dimensional representations; infinite dimensional representations of the Lorentz group was of interest to the physicists, but did not attract much attention from mathematicians. In the mid fifties Harish-Chandra (1923-83) with his profound insights into infinite dimensional representation theory brought it to centre-stage in mathematics where it has continued to stay. However Harish-Chandra had moved to the U S before he turned to mathematics and all his work was done in that country. The first piece of important work in (infinite dimensional) representation theory that came from India came from the Indian Statistical Institute. In a paper that appeared in the Annals of Mathematics (1966), K R Parthasarathy (1936-), Ranga Rao (1936 - ) and V S Varadarajan (1937-2019) obtained deep and beautiful results on infinite dimensional representations of a complex semi-simple Lie algebra (with implications for infinite dimensional unitary representations for Lie groups). In that paper they made a conjecture which came to be called the PRV conjecture which drew considerable attention. It was eventually proved in the eighties by the Indian mathematician Shrawan Kumar.

Harish-Chandra's work in representation theory culminated in his construction of the characters of the so called discrete series representations of a semi-simple Lie group $G$. These are irreducible unitary representations that occur as subrepresentations of $L^2(G)$. One of the immediate problems that presented itself after this work was to construct concrete realizations of these representations; and the first such construction beyond the case of $G = SL(2, \mathbb{R})$ was given in a beautiful paper by M S Narasimhan (written jointly with the Japanese mathematician H Okamoto (1956-)) in 1970 for groups $G$ whose associated symmetric spaces are hermitian symmetric. The representations were realized on the space square integrable holomorphic sections of suitable holomorphic vector bundles on the symmetric space.

R Parthasarathy (1945-) followed this up with an equally beautiful paper that showed that whenever $G$ admits a discrete series, discrete series representations can be realized in the space on square integrable "spinors" in suitable vector bundles

with a spin reduction. This settled the problem of realization of discrete series completely.

Parthasarathy went on to establish several deep results which placed him among the handful of people who are responsible for the very direction in which representation theory has moved since the advent of the discrete series. I should mention S Kumaresan (1950-) whose thesis was written under the direction of R Parthasarathy: the main result of the thesis is a cohomology vanishing theorem known by Kumaresan's name. Kumaresan retired as a Professor from University of Hyderabad a few years back.

## 14. Lie Groups and their Discrete Subgroups

This is another area where Indian contribution has been both deep and prodigious. In 1960 A Selberg (1917-2007) made the remarkable discovery that unlike in the case of $SL(2,\mathbb{R})$, discrete co-compact subgroups of $SL(n,\mathbb{R})$ for $n \geq 3$ cannot be moved continuously inside $SL(n,\mathbb{R})$ except by inner conjugation. Soon after A Weil extended the result to all semi-simple Lie groups without compact factors; he also pointed out a connection between this "rigidity" and the cohomology of the discrete subgroup. M S Raghunathan proved a number of interesting vanishing results for the cohomology of discrete subgroups $\Gamma$ of semi-simple Lie groups $G$ with $G/\Gamma$ of finite volume introducing new techniques to handle the case when $\Gamma$ is not co-compact.

He also made considerable progress towards a conjecture asserting that all non-co-compact discrete subgroups of finite co-volume in higher rank semi-simple groups are "arithmetic". The Russian mathematician G A Margulis (1946-) in 1974 proved the arithmeticity conjecture for both co-compact and non-co-compact lattices and was awarded the Fields Medal essentially for that work. T N Venkataramana (1958-) later extended Margulis's arithmeticity result to positive characteristics.

R Parthasarathy (1945-) proved some deep results about the cohomology of compact locally Hermitian symmetric in the eighties. Venkataramana embarked on this topic somewhat later and has proved very interesting and profound results some of it in collaborations with L Clozel (1953) and B Speh (1949-). A description of the work would need elaboration of many concepts and definitions which I do not want to embark on.

## 15. Algebraic groups and arithmetic

Some interesting work on Galois cohomology of algebraic groups related to the so called Kneser-Tits conjecture were obtained by Gopal Prasad (1945-) and Raghunathan in the early eighties. Raman Parimala in collaboration with Eva Bayer-Fluckiger (1951-) settled a conjecture of J.-P Serre on Galois cohomology for classical groups over fields of cohomological dimension 2.

Another topic in which Indian work has had considerable impact is the "Congruence subgroup problem". A subgroup $\Gamma$ in $SL(n,\mathbb{Z})$ is a congruence subgroup if there is a non-zero ideal $\mathfrak{a}$ in $\mathbb{Z}$ such that

$$\Gamma \supset SL(n,\mathfrak{a}) \stackrel{def}{=} \{T \in SL(n,\mathbb{Z}) |\ T \equiv 1_n (mod\ \mathfrak{a})\}$$

where $1_n$ is the identity matrix. The congruence groups are evidently subgroups of

finite and the congruence subgroup problem for $SL(n,\mathbb{Z})$is the question whether these exhaust the family of all subgroups of finite index. This question can be posed in a more general context: $Z$ can be replaced by $S$-integers $\mathcal{O}_{k,S}$ in a global field (in particular number field) $k$, $S$ being a set of places of $k$ including all the Archimedian ones, and $SL(n,\mathbb{Z})$ replaced by $G_{\mathcal{O}_{k,S}} = G \cap SL(n,\mathcal{O}_{k,S})$. Karl E R Fricke (1862) and Felix Klein (1841-1925) had answered the question negatively at the fag end the $19^{th}$ century in the case $k = \mathbb{Q}$, $G = SL(2)$ and $S$ is the unique real place.

Interest in the problem revived with an answer in the affirmative by H Bass (1932-), M Lazard(1924-85) and and J-P Serre and independently by J Mennicke in 1962. The problem for "isotropic groups" over a global fields was completely solved in a series of papers of M S Raghunathan, some of them jointly with Gopal Prasad; the answer is not always in the affirmative but when not in the affirmative there is a measure of the failure (suggested by Serre) and in these papers that measure is determined.

Using the ideas used in the solution of the congruence subgroup problem Raghunathan and T N Venkataramana were able to show that the virtual first Betti number of an arithmetically defined compact Riemannian manifold of constant curvature of dimension $\geq 4$ is non-zero, a much sought after result (also proved independently by Li and Milson by very different methods).

## 16. Homogeneous Dynamics

This area is in the interface of Lie Theory and Ergodic Theory. Let $G$ be a Lie group with a bi-invariant Haar measure and $\Gamma$ a discrete subgroup such that the measure on $G/\Gamma$ induced by the Haar measure is invariant under the left action of $G$. Homogeneous dynamics is the study of actions of sub-groups, in particular 1-parameter subgroups of $G$ on $G/\Gamma$. S G Dani (1947-) wrote a series of papers in the area starting in the seventies which are foundational in nature. A conjecture of Raghunathan on unipotent subgroups of $G$ rendered accessible to ergodic theory techniques by a twist given by Dani, was pursued by many mathematicians over more than a decade till Marina Ratner (1938-2017) proved the conjecture in 1989. Dani himself had written a number of papers connected with the conjecture which had an influence on the final solution. Nimish Shah (1967-) proved the conjecture in an important special case. After the conjecture was proved, Nimish Shah went on to prove using it some very nice results about distribution of integral points on homogeneous spaces of algebraic groups defined over number fields. The conjecture opened up the possibility of proving many number theoretic results of this nature.

## 17. Differential Equations

I have already mentioned some work on Differential Equations in the section Differential Geometry/ Differential Equations; Ganesh Prasad's work mentioned there properly belongs in this section, but was mentioned there as a preliminary to the description Minakshisundaram's work which belongs to the interface of Differential Geometry and Differential Equations.

A paper of M S Narasimhan (jointly with the Japanese mathematician) published in 1959 is a truly important work in the area to come out of India. The paper among other things proves that a (distribution) solution of a linear elliptic partial differential equation with analytic coefficients is necessarily analytic.

Adimurthy (1952-), Kesavan (1952-), Vanninathan (1951-) and Srikanth (1950-) at the TIFR Centre for Applicable Mathematics have all done excellent work dealing with non-linear partial differential equations. My acquaintance with this area is inadequate for me to say more about their work. I will content myself saying that they are all internationally recognized experts in their area.

## 18. Complex Function Theory

Raghavan Narasimhan's early work done in India during 1959-65 already established him as a leading expert in Complex Function Theory in the world. In his very first paper in this area proved the following theorem (which was a 50 year old conjecture).

THEOREM 3 *A non-compact (= open) Riemann surface admits a proper imbedding in* $\mathbb{C}^3$.

He followed it up proving that a Stein manifold of (complex) dimension $n$ can be imbedded in $C^{(2n+1)}$. Being Stein is a somewhat technical condition which I will not go into. It is a necessary condition for imbeddability; an open Riemann surface is Stein. He also wrote a series of papers dealing with Levi Convexity in the short space of time at TIFR. He continued with excellent work in the area but all that was done outside the country.

R R Simha (1936-2019) is another mathematician with significant contributions in the area.

## 19. Other Areas

I have written about Indian contribution to mathematics in areas with which I have more than a fleeting acquaintance which naturally limits the coverage of this article. Functional Analysis and Combinatorics are two areas conspicuous by their absence in this account. I will content myself with mentioning some names of people who have contributed significantly to these areas: Rajendra Bhatia (1952-) and V S Sunder (1952-) to Functional Analysis and S S Shrikhande (1917-2020) and N M Singhi(1949-) to Combinatorics. There have been highly significant contributions in these areas but I have not attempted to say anything about them as I am somewhat diffident about handling them. Probability is an area where Indian mathematicians have done us proud, but I am not in a position to sift out the more outstanding contributions from a very large output. S R S Varadhan is a name to reckon with in this area (He was awarded the Abel Prize in 2007). He has been at the Courant Institute New York since before the mid sixties; however his early work, done when he was at ISI Kolkota, had already established him as an outstanding probabilist.

One other piece work that I should mention is the disproof of a conjecture of Euler on Latin Squares by S S Shrikhande jointly with R C Bose (1901-87) and E T Parker (1926-91). News of this work had the distinction (not shared by any other mathematical publication by an Indian) of making it to the front page New York Times.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 12–21

# *Diophantine Equations An Introduction*

**Dinakar Ramakrishnan***

*California Institute of Technology*

**Abstract:** This is a redaction of the Inaugural Lecture the author gave at the University of Hyderabad in January 2019 in honor of the late great Geometer (and Fields medalist) Maryam Mirzakhani.

What is presented here is a limited perspective on a huge field, a meandering path through a lush garden, ending with a circle of problems of current interest to the author. No pretension (at all) is made of being exhaustive or current.

## 1. Something light to begin with

When Nasruddin Hodja claimed that he could see in the dark, his friend pointed out the incongruity when Hodja was seen carrying a lit candle at night. "Not so," said Nasruddin, "the role of the light is for others to be able to see me."

The moral is of course that one needs to analyze all possibilities before asserting a conclusion.

Maryam Mirzakhani, whom this Lecture is named after, would have liked the stories of Hodja.

Mirzakhani's mathematical work gave deep insights into the structure of geodesic curves on hyperbolic surfaces. Such surfaces also play a major role in the field of Number theory, often through an analysis of Diophantine equations.

*Etymology*: *Hod* (or *Khod*) is of Persian origin meaning *God*, and 'Hodja' serves God, signifying a Mullah, Priest, Rabbi, Minister or Pundit (depending on one's favorite religion).

The expression *Khoda Hafez* (or 'Khuda Hafiz' in Urdu) of course means 'May God protect you' or just 'Goodbye' in the modern usage.

*Hafiz* is of Arabic origin meaning 'protector'.

## 2. A basic Definition

By a **Diophantine equation**, one means an equation of the form

$$f(X_1, X_2, \ldots, X_n) = 0$$

*Corresponding author. Email: dinakar@caltech.edu

where $f$ is a polynomial (in $n$ variables) with *coefficients in the ring of integers* $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots, \pm n, \dots\}$. Denote as usual by $\mathbb{Q}$ the field of rational numbers.

One wants to find integral (or rational) vectors $x = (x_1, \dots, x_n)$ such that $f(x) = 0$.

A study of these equations was initiated by *Diophantus of Alexandria*, who lived in the third century AD. He wrote a series of books titled *Arithmetica*, whose translation into Latin by Bachet influenced many including Pierre de Fermat. See [D], which gives a link to an English translation, and [Sch] which links to an interesting essay on Diophantus.

Diophantus may have lived earlier, and a key commentary on him by *Hypatia* is missing. Also one of Diophantus's works is missing, as he quotes some Lemmas from there in *Arithmetica*.

The consensus seems to be that he was Greek. He was likely well versed in Ancient Greek, as many learned people probably were in Alexandria, but could he have been Egyptian (or Jewish or Caldean)?

Of particular interest ar *homogeneous* Diophantine equations, i.e., with $f(x_1, \dots, x_n) = 0$ with $f$ a homogeneous polynomial. In this case, any integral solution $a = (a_1, \dots, a_n)$ leads to infinitely many integral solutions $(ba_1, \dots, ba_n)$ as $b$ varies in $\mathbb{Z}$. One calls the solutions *a primitive* if the gcd of $\{a_1, \dots, a_n\}$ is 1.

More generally, one may consider *Diophantine systems*, which are finite collections of Diophantine equations, and look for *simultaneous* integral (or rational) solutions.

## 3. Pythagorean triples

These are (positive) Integral Solutions of $X^2 + Y^2 = Z^2$.

The first sixteen primitive Pythagorean triples are
$(3, 4, 5), (5, 12, 13), (8, 15, 17), (7, 24, 25), (20, 21, 29), (12, 35, 37),$
$(9, 40, 41), (28, 45, 53), (11, 60, 61), (33, 56, 65), (16, 63, 65),$
$(48, 55, 73), (36, 77, 85), (13, 84, 85), (39, 80, 89),$ and $(65, 72, 97)$.

A larger triple is $(403, 396, 565)$.

Many old civilizations (in Babylon, China, India, for example) studied this equation long before Pythagoras. The Babylonians even found the non-trivial triple $(3367, 3456, 4825)$.

All primitive solutions can in fact be *parametrized* by:
$(2mn, m^2 - n^2, m^2 + n^2)$, $m > n$, with
$m, n$ of opposite parity, $(m, n) = 1$.

To get at this, one looks for rational solutions of $u^2 + v^2 = 1$, which are geometrically realized as **rational points** on the unit circle $S$.

They are obtained by intersecting $S$ with secant lines with rational slope emanating from $(-1, 0)$.
This illustrates the basic idea of embedding rational solutions inside real, or complex, points of the *variety* $V$ defined by the diophantine equation $f(x) = 0$.

Once we have the rational solutions $(u, v)$ of $u^2 + v^2 = 1$, one can clear the denominators and get integral solutions of $x^2 + y^2 = z^2$. A bit more work yields all the primitive Pythagorean triples.

A quick subjective comment. The approach of the Greeks in such problems stressed the importance of a proof (of completeness), which forms the basis of modern mathematics, while that of the earlier ones was more algorithmic.

## 4. $X^2 - dY^2 = 1$

Fermat's challenge of 1657 to find an integral solution for $d = 61$ brought this equation, attributed to Pell, to prominence.

However, three centuries earlier, Bhaskara in India had derived the solution $(1766319049, 226153980)$
using the **Chakravaala Vidhi** (*Cyclic method* or 'rule') due to him and (earlier) Jeyadeva.

This method provided an *algorithm* to construct from one solution many other solutions, infinitely many, and one gets all solutions this way, though there was no proof at that time.

In fact, already in the seventh century, Brahmagupta had solved this equation for $d = 83$. He derived a composition law and also 'shortcuts' like going from a solution of $u^2 = dv^2 = -4$) to $u^2 - dv^2 = 1$; for $N = 61$, $39^2 - 61(5^2) = -4$.

For an instructive and beautiful discussion of this method of the Indian mathmaticians of olden times, see [We].

## 5. Sums of three squares

Diophantus investigated the representation of a positive integer $n$ as a sum of three squares, i.e., looked at the equation

$$X^2 + Y^2 + Z^2 = n.$$

For $n = 10$, he found the elegant solution in positive integers:

$$x = \frac{1321}{711}, \; y = \frac{1285}{711}, \; z = \frac{1288}{711}.$$

His method is still interesting to peruse. He also wanted the minimum of $\{x, y, z\}$ to be $\sqrt{3}$, which he achieved.

In 1797/8, Legendre proved that the only positive integers $n$ which are *not sums of three squares* are those of the form

$$n = 4^a(8b + 7), \text{ with } a, b > 0.$$

By contrast, one knows by Lagrange that *every positive integer is a sum of four squares.*

## 6. $X^4 + Y^4 = Z^2$ and Fermat

Fermat proved that this equation, and hence $X^4 + Y^4 = Z^4$, has no positive integral solutions, and in the process introduced the *Method of infinite descent.*

*His argument*: By the previous section, any solution $(x, y, z)$ will need to satisfy $x^2 = 2mn$, $y^2 = m^2 - n^2$, implying that $m$ or $n$ is even, say $m$; then $y^2 + n^2$ is 0 modulo 4, forcing $n$ to be even as well, leading to a smaller solution. One can continue this ad infinitum, resulting in a contradiction.

This case led Fermat to claim (in the 1630's) that $X^N + Y^N = Z^N$ has no positive integral solutions for $N \geq 3$. He claimed that the margin was too small to contain

his reasoning, but there seems to be a general scepticism that he had a proof. For $n = 3$, substantial progress was made a century late by Euler.

It is now elementary to observe that to establish FLT, it suffices to settle it for odd prime exponents.

## 7. Sophie Germain

Given that this lecture is in honor of Maryam Mirzakhani, it is imperative to point out a terrific female mathematician who made significant progress on the Fermat problem. Sophie Germain, born in 1776 in Paris, was extremely talented in Math, and since at that time the Ecole Polytechnique would not admit women, she could not attend the lectures of Lagrange there, but still followed them by getting the notes under a male psuedonym!

In the early eighteen hundreds she made a real breakthrough and proved the following:
*Let $p$ be any prime such that $2p+1$ is also a prime. Then there is no solution $(x, y.z)$ in whole numbers with $p \nmid xyz$ satisfying the Fermat equation $X^p + Y^p = Z^p$.*

These primes are now called *Sophie Germain primes*, with obvious examples being $p = 5$ and $p = 11$. It is expected that there are infinitely many such primes, but this is still open.

## 8. Faltings

In 1983 the German mathematician Gerd Faltings supplied a dramatic proof (in [F]) of a *Conjecture of Mordell*, implying:

*There are only a finite number of rational solutions (up to scaling) of the Fermat Equation $F_N : X^N + Y^N = Z^N$ for all $N \geq 4$.*

In fact he proved this for solutions in any *number field*, i.e., a finite extension field $K$ of $\mathbb{Q}$, and moreover, one could replace $F_N$ by any plane curve defined by an irreducible polynomial equation of degree $> \sqrt{3}$ (so that each square is greater than 3, but they all add up to 10, making each square roughly of the same size).

This also showed the stark contrast between the number of (projective, meaning up to scaling) rational solutions of $F_N$ for $N \leq 2$ and $N \geq 4$. One sees a *dichotomy* here. But in fact, there is a *trichotomy.*

## 9. View from Riemann Surfaces

Given a homogeneous polynomial $f(X, Y, Z)$, one always has the trivial (zero) solution, and any multiple of a given solution is another.

So one scales the solutions, to get an algebraic curve $C$ defined by $f$ in the *Projective plane* $\mathbb{P}^2$, which can be thought either as the space of lines through the origin in the (affine) three space, or as the compactification of the (affine) plane by adding a line at infinity.

When $C$ is smooth, its complex solutions form a compact Riemann surface $M$, which has a genus $g$. (Simply speaking, a Riemann surface is a real surface on which one can measure angles.) Note that $M$ is a *real surface* and a *complex curve*! (It of course makes sense as $\mathbb{C}$ is a 2-dimensional vector space over $\mathbb{R}$.)

One can think of $g$ as the number of handles one can attach to the Riemann sphere to obtain $M$ (up to homeomorphism) or as the number of independent holomorphic differential 1-forms $\omega$ on $M$.

When $M$ is defined by a homogeneous equation $f(X, Y, Z)$ of degree $n$, then $g$ is given by $(n-1)(n-2)/2$. In particular the genus is $> 1$ when $N \geq 4$ and is 0 when $N \geq 2$.

What Mordell conjectured was that when $g \geq 2$, the number of rational points of $C$, embedded in $M$, is finite. This is what Faltings proved in its full generality!

## 10. The Trichotomy

In genus zero, as soon as one has a rational point, then there are infinitely many, in fact in bijection with the points on a projective line.

For example, the projective curve $X^2 + Y^2 + Z^2 = 0$ has no rational point at all, while $F_2$ defined by $X^2 + Y^2 - Z^2 = 0$ has infinitely many points.

And by Mordell (proved by Faltings), the number of rational points is finite for $g \geq 2$.

The case $g = 1$ is special; it has either no rational point, or else it is an *elliptic curve*, whose $\mathbb{Q}$-points form an abelian group $E(\mathbb{Q})$, known by Mordell to be isomorphic to $\mathbb{Z}^r \times G$, for a *finite group* $G$ and $r$ a non-negative integer, called the rank.

So in this intermediate (boundary) case, the number of points could in general be finite or infinite! For $F_3$, it happens to be finite.

## 11. Wiles and FLT

One would be remiss to not mention the deep and successful program of *Andrew Wiles*, completed in 1995, resulting in the establishment (in [W]) of **FLT for all** $N > 2$, partly relying on an important joint work with Richard Taylor ([TW]).

*The proof is ingenious, involving a series of difficult arguments, but quite complicated* for us to attempt to describe it here! It also involves deep results on elliptic curves and modular forms, and proceeds by establishing a modularity conjecture for elliptic curves over $\mathbb{Q}$, the sufficiency of which had earlier been established by K. Ribet using some ideas of G. Frey. The starting point of the strategy is to make use of the theorem of Langlands and Tunnell (cf. [La], [Tu]) that Artin's conjecture holds for Galois representations with image in $\mathrm{GL}_2(\mathbb{F}_3)$ (which is solvable), resulting in the modularity modulo 3 of any $E$.

For a thousand-word exposition, see https://simonsingh.net/books/fermats-last-theorem/the-whole-story/

In a related vein, a deep conjecture of J.-P. Serre asserting the modularity conjecture for odd 2-dimensional Galois representations was settled in 2005/8 by the elegant works of C. Khare and J.-P. Wintenberger [KW].

## 12. *L*-functions

To be concrete, let us look at elliptic curves $E$ over $\mathbb{Q}$, defined by $Y^2 = f(X)$, with $f$ a cubic polynomial with $\mathbb{Q}$-coefficients and distinct roots (in $\mathbb{C}$); For FLT, one

is interested in $E$ such that $f$ has three distinct roots in $\mathbb{Z}$. One can look at the number of points $\nu_p$ of the reduction of $E$ modulo $p$, which will be non-singular at all $p$ not dividing its conductor $N$. One sets $a_p = p + 1 - \nu_p$ for each $p$, and defines the L-function by the infinite (Euler) product

$$L(s, E) = \prod_p \frac{1}{1 - a_p p^{-s} + \omega(p) p^{1-2s}}$$

with $\omega(p) = 1$ iff $p \nmid N$ and $= 0$ otherwise. By a basic result of Hasse, one knows that $|a_p| \leq 2\sqrt{p}$, and this implies the normal convergence of $L(s, E)$ in $\Re(s) > 2$. The modularity of $E$ signifies the existence of a (normalized new) cusp form $f$ of weight 2, level $N$ , with $\mathbb{Q}$-coefficients and trivial character, such that for each $p \nmid N$ , $a_p$ is the $p$-th Hecke eigenvalue of $f$. In other terms, the L-functions of $E$ and $f$ coincide (where an argument is needed at the bad $P$).

The utility of modularity for arithmetic is that by Hecke theory one knows that $L(s, f)$ admits a holomorphic continuation to the whole $s$-plane, and satisfies a functional equation relating $s$ to $2 - s$, making $s = 11$ the critical center.

The modularity of arbitrary, not just semistable, elliptic curves over Q was accomplished (extending [TW], [W]), by the works of Brueil, Conrad, Diamond and Taylor (cf. [BCDT]).

For general $V$ over $\mathbb{Q}$, the Langlands philosophy predicts a modularity, for each degree $j \geq 0$, of the degree $j$ $L$-function $L^{(j)}(s, V)$ of the $j$-th cohomology $H^j(V)$ in terms of automorphic forms on $\mathrm{GL}(b_j)$ which are Hecke eigenforms (generating automorphic representations), where $b_j$ is the $j$-th Betti number ($= \dim(H^j(V))$. (When $V$ is an elliptic curve $E$, $L^{(1)}(s, E)$ is the $L$-function defined above, while $L^{(0)}(s, E) = \zeta(s)$ and $L^{(2)}(s, E) = \zeta(s-1)$, where $\zeta(s)$ is the Riemann Zeta function defined by the Dirichlet series $\sum_{n \geq 1} n^{-s}$ (in $\Re(s) > 1$). Some positive, striking results are known beyond elliptic curves, mostly tied up with modular (or Shimura) varieties ([Pic], [SSA]). Moreover, a fundmental new viewpoint has been brought to the subject by P. Scholze; see his recent works with A. Caraiani and others on the arxiv.

## 13. BSD

Let $E$ be an elliptic curve over $\mathbb{Q}$. Then as noted earlier, one knows by Mordell that the (commutative) group $E(\mathbb{Q})$ of $\mathbb{Q}$-rational points on $E$ is finitely generated, i.e., of the form $\mathbb{Z}^r \times H$, with $H$ a finite group. The exponent $r$ is the *rank* of $E(\mathbb{Q})$. It turns out, by a major theorem of Mazur that there are only a finite number of possibilities for $H$ as one varies $E$ over all elliptic curves over $\mathbb{Q}$; see [B1].

So the remaining (very) difficult problem is to understand the rank $r$. The famous *Conjecture of Birch and Swinnerton-Dyer*, colloquially referred to as BSD, predicts that $r$ equals the order of zero at $s = 1$ of $L(s, E)$.

This is one of the Clay Millennial problems; see

https://www.claymath.org/millennium-problems/birch-and-swinnerton-dyer-conjecture

In particular, BSD predicts that when $r$ is positive $L(s, E)$ must vanish. Here is the heuristic argument in that case: Suppose we ignore that fact that the Euler product expansion of $L(s, E)$ does not make sense at $s = 1$, and formally plug in

$s = 1$, we see that

$$L(1, E) \text{ “} ='' \prod_p \frac{p}{p - a_p + \omega(p)} = C \prod_{p \nmid N} \frac{p}{\nu_p}$$

with $C \neq 0$, where we have used $a_p = p + 1 - \nu_p$. When $r > 0$, one expects a lot of points mod $p$ for lots of primes $p$, which by the Hasse bound implies that $\nu_p$ is close to $p + 1 + 2\sqrt{p}$ for many $p$, which results in the infinite product being zero, suggesting the same for $L(1, E)$.

There is an enormous body of literature on this fundamental conjecture with several partial, but striking, theorems. We will content ourselves to describing one recent result of Bhargava, Skinner and Wei Zhang [BSZ].

In this paper the authors show that over 60 percent of elliptic curves $E$ over $\mathbb{Q}$, when ordered by the height, have $r = \text{ord}_{s=1} L(s, E) \leq 1$. Their method is to analyze the Selmer group at $p = 5$. If that can be done at an arbitrary $p$, then they can reach 100 percent (statistically).

## 14. Sato-Tate

Given any elliptic curve $E$ over $\mathbb{Q}$, we may, thanks to the Hasse bound, write $a_p = 2\sqrt{p} \cos \theta_p$, for a phase $\theta_p \in [0, \pi] \subset \mathbb{R}$. When $E$ admits complex multiplication (by an imaginary quadratic number), the distribution of the angles $\theta_p$ has been understood for some time.

In the non-CM case, an elegant conjecture of Sato and Tate, independently made, asserts that the angles $\theta_p$ are equidistributed on $[0, \pi]$ according to the measure $\frac{2}{\pi} \sin^2 \theta d\theta$. This conjecture was brilliantly solved by L. Clozel, M. Harris, N. Shepherd-Barron and R. Taylor (under a multiplicative reduction condition at a prime p) in an elaborate joint program, with the proof stretched over a series of three papers [CHT], [HST], [T].

Roughly speaking, these authors vastly generalize [TW] in the higher dimensional case, utilizing unitary Shimura varieties, and deduce the requisite analytic properties of $L(s, E, \text{sym}^n)$ ,the symmetric power $L$-functions of $E$.

A generalization valid for non-CM holomorphic newforms $f$ of weight $k \geq 2$, removing also the multiplicative reduction condition for $E$ attached to $f$ of weight 2, was established in [BGHT].

A beautiful recent preprint of J. Newton and J.A. Thorne has made a breakthrough and established the modularity of all the symmetric powers of all semistable elliptic curves $E/\mathbb{Q}$, and of all newforms $f$ of level 1 (cf. [NT]).

## 15. Hyperbolicity and Lang's Conjecture

The *Uniformization theorem* implies that every compact Riemann surface $M$ of genus $g \geq 2$ is covered by the *upper half plane* $\mathcal{H} := \{x + iy \in \mathbb{C} \,|\, y > 0\}$, or equivalently the open unit disk in $\mathbb{C}$.

For $g = 0$ (resp. $g = 1$, the universal cover is the sphere $S^2$ (resp. the complex plane $\mathbb{C}$)

The natural Poincaré metric $dxdy/y^2$ on $\mathcal{H}$ furnishes a hyperbolic structure to $M$ of genus $\geq 2$, i.e., gives it *negative sectional curvature*. Note that for $g = 0$, (resp. $g = 1$), the curvature is positive (resp. 0).

Suppose $V$ is a smooth projective variety of dimension $n$ defined by a system of diophantine equations.

Let us call $V$ *hyperbolic* if there is a non-constant holomorphic map $\varphi : \mathbb{C} \to M$, where $M$ is the complex manifold (of complex dimension $N$) defined by the complex points of $V$.

Note that in dimension one, being hyperbolic is the same as the genus being $\geq 2$.

**Conjecture** (Lang) *When $V$ is hyperbolic, it is* **Mordellic**, *i.e., has only a finite number rational points, and in fact over any number field.*

This was partly inspired by groundbreaking work of Paul Vojta ([V]), who, through his analogy between Nevanlinna theory and Diophantine approximation, made his own strong conjectures.

For an insightful discussion of general conjectures on rational points, see [M2].

## 16. The Bombieri-Lang Conjecture

An algebro-geometric generalization of algebraic curves of genus $g \geq 2$ is given by the algebraic varieties of general type.

The Bombieri-Lang Conjecture asserts that for $n$-dimensional $V$ of general type, the Zariski closure $Z$ of the rational points has irreducible components of dimension $< n$.

When $n = 2$, i.e., when $V$ is a surface, this conjecture is closely related to Lang's conjecture above. Indeed, for $V$ a surface of general type, the *Bombieri-Lang Conjecture* asserts that the irreducible components of $Z$ are all of dimension $\leq 1$. If $C$ is (the normalization of) a dimension one component, then $C$ must have genus $\geq 2$ if $V$ is hyperbolic, as any $C$ of genus $\leq 1$ will have universal cover $S^2$ or $\mathbb{C}$, inducing a non-zero holomorphic map from $\mathbb{C}$ to $\mathcal{B}$, which is impossible. Then by Faltings, $Z$ could have only a finite number of rational points, thereby yielding Mordellicity.

A very interesting situation is when $V = Y \cup D$ with $Y$ open and hyperbolic, with $D$ a divisor with normal crossings. Such a situation arises for the celebrated surfaces of Picard.

## 17. Picard Modular surfaces

Now we will focus on dimension 2, i.e., when $V$ is a smooth projective surface which is itself hyperbolic or contains an open surface $Y$ which is hyperbolic.

Here, hyperbolicity does not guarantee the universal cover being the unit disk $\mathcal{B}$ in $\mathbb{C}^2$.

However, many beautiful examples are furnished by the *Picard modular surfaces* $Y(\mathbb{C}) = \Gamma\backslash\mathcal{B}$, which have smooth compactifications $V(\mathbb{C})$ with complement a divisor $D$ with normal crossings.

$\Gamma$ is a discrete subgroup of finite covolume in $SU(2,1)$ defined by a hermitian form on $K^3$ with $K$ an imaginary quadratic field. It is known that such quotients admit models over number fields.

The divisor $D$ at infinity turns out to be a finite union of elliptic curves with complex multiplication by $K$.

Much is known about these surfaces - due to J. Rogawski, R. Kottwitz, J.S. Milne and others [Pic], [Ro].

Here is something this lecturer proved jointly with Mladen Dimitrov [DR].

**Theorem** *Let $V = Y \cup D$ be a Picard modular surface as above relative to an arithmetic subgroup $\Gamma$ of $SU(2,1)$. Then Lang's conjecture holds for a finite cover $Y'$ of $Y$.*

As a consequence, one gets Mordellicity of surfaces $Y$ which arise this way.

There is also a version establishing an analogue for compact arithmetic quotients $X$ of $\mathcal{B}$. In that case, the result had earlier been known (by a different method) by Emmanuel Ullmo.

One gets examples this way of general type surfaces arising as intersections of hypersurfaces in $\mathbb{P}^n$. A particularly simple one is the surface in $\mathbb{P}^5$ given by the solution set of the Diophantine system of equations:

$$x_1^5 + y^5 = z^5,\ x_2^5 + z^5 = w^5,\ x_3^5 + w^5 = y^5,$$

which involves the familiar Fermat equations.

If a beginner wants more information on the rudiments of Number theory, zie could look at the author's Notes: *Introduction to Number Theory* at http://www.its.caltech.edu/ dinakar/

## References


[BGHT] T. Barnet-Lamb, D. Geraghty, M. Harris, Michael and R. Taylor, A family of Calabi-Yau varieties and potential automorphy II, Publ. Res. Inst. Math. Sci. 47 (2011), no. 1, 29-98.

[BSZ] M. Bhargava, C. Skinner and Wei Zhang, *A majority of elliptic curves over Q satisfy the Birch and Swinnerton-Dyer conjecture*, (2014). https://arxiv.org/abs/1407.1826v2

[BCDT] C. Breuil, B. Conrad, Brian, F. Diamond and R. Taylor, On the modularity of elliptic curves over Q: wild 3-adic exercises, J. Amer. Math. Soc. 14 (2001), no. 4, 843-939.

[CHT] L. Clozel, M. Harris and R. Taylor, Automorphy for some l-adic lifts of automorphic mod l Galois representations, with Appendix A, summarizing unpublished work of Russ Mann, and Appendix B by Marie-France Vignras, Publ. Math. Inst. Hautes tudes Sci. No. 108 (2008), 1-181.

[DR] M. Dimitrov and D. Ramakrishnan, Arithmetic quotients of the complex ball and a conjecture of Lang. (English summary) Doc. Math. 20 (2015), 11851205.

[D] Diophantus, *Arithmetica* (Greek); Latin translation by Claude Gaspard Bachet de Mziriac; English translation by Sir Thomas L. Heath (1910): archive.org/details/diophantusofalex00heatiala

[F] G. Faltings, *Endlichkeitsstze fr abelsche Varietten ber Zahlkrpern* (German) [Finiteness theorems for abelian varieties over number fields] Invent. Math. 73 (1983), no. 3, 349-366.

[HST] M. Harris, N. Shepherd-Barron and R. Taylor, A family of Calabi-Yau varieties and potential automorphy, Ann. of Math. (2) 171 (2010), no. 2, 779-813.

[KW] C. Khare and J.-L. Wintenberger, Serre's modularity conjecture. Proceedings of the International Congress of Mathematicians. Volume II, 280-293, Hindustan Book Agency, New Delhi, 2010.

[La] R.P. Langlands, *Base change for GL(2)*, Annals of Mathematics Studies **96**. Princeton University Press (1980).

[B1] B. Mazur, Rational isogenies of prime degree (with an appendix by D. Goldfeld), Invent. Math. 44 (1978), no. 2, 129-162.

[M2] B. Mazur, *Open problems regarding rational points on curves and varieties*, Galois representations in arithmetic algebraic geometry (Durham, 1996), 239265, London Math. Soc. Lecture Note Ser., 254, Cambridge Univ. Press, Cambridge, 1998.

[NT] J. Newton and J.A. Thorne, Symmetric Power Functoriality For Holomorphic Modular Forms, arXiv:1912.11261v1 [math.NT] 24 Dec 2019.

[Pic] *T*he zeta functions of Picard modular surfaces, edited by R.P. Langlands and D. Ramakrishnan, CRM Publications, Univ. Montral, Montreal, QC (1992).

[Ro] J. Rogawski, *A*utomorphic representations of unitary groups in three variables, Annals of Mathematics Studies, **123**, Princeton University Press, Princeton, NJ, 1990.

[Sch] N. Schappacher, "Wer war Diophant?" (German) ["Who was Diophantus?"] Math. Semesterber. 45 (1998), no. 2, 141-156 (English translation: http://irma.math.unistra.fr/ schappa/NSch/Publications files/1998cBis Dioph.pdf)

[SSA] *S*tabilization of the Trace Formula, Shimura Varieties, and Arithmetic Applications, edited by L. Clozel, M. Harris, J.-P. Labesse and B.-C. Ngo, Books 1,2, International Press, Sommerville, MA (2011).

[T] R. Taylor, Automorphy for some l-adic lifts of automorphic mod l Galois representations II, Publ. Math. Inst. Hautes tudes Sci. No. 108 (2008), 183-239.

[TW] R. Taylor and A. Wiles, *Ring-theoretic properties of certain Hecke algebras*, Ann. of Math. (2) 141 (1995), no. 3, 553-572.

[Tu] J. Tunnell, Artin's conjecture for representations of octahedral type. Bull. Amer. Math. Soc. (N.S.) 5 (1981), no. 2, 173-175.

[V] P. Vojta, Nevanlinna theory and Diophantine approximation. Several complex variables (Berkeley, CA, 19951996), 535564, Math. Sci. Res. Inst. Publ., 37, Cambridge Univ. Press, Cambridge, 1999.

[W] A. Wiles, *Modular elliptic curves and Fermat's last theorem*, Ann. of Math. (2) 141 (1995), no. 3, 443-551.

[We] A. Weil, *Number theory. An approach through history from Hammurapi to Legendre*, Reprint of the 1984 edition. Modern Birkhuser Classics. Birkhuser Boston, Inc., Boston, MA (2007).

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 22–25

# *Fermat's Last Theorem in a Polynomial Ring*

**Rajat Tandon***

*Rajat Tandon, Bitra, 35 B N Reddy Colony, Rd No 14 Banjara Hills, Hyderabad 500034*

Fermat's last theorem states that if there exist integers $x, y, z$ such that $xyz \neq 0$ and $x^n + y^n = z^n$ for some natural number $n$ then $n < 3$. There are, of course, infinitely many Pythagorian triples, i.e., natural numbers $x, y, z$ such that $x^2+y^2 = z^2$, like (3,4,5), (5,12,13) etc. More than 350 years after the conjecture was originally made Wiles and Taylor proved FLT in 1996 published in their famous paper in the Annals of Mathematics.

A natural question to ask is whether there are FLT-type results in other integral domains besides $\mathbb{Z}$. The first integral domain that comes to mind are the polynomial rings $F[X]$ where $F$ is a field. This domain, like $\mathbb{Z}$, is also a Euclidean domain and therefore a unique factorization domain. Analagous to the set of natural numbers sitting inside the set of integers we have the set of MONIC polynomials (leading coefficient is 1) sitting inside $F[X]$. The units (multipicatively invertible elements) in $\mathbb{Z}$ are $\{\pm 1\}$ whilst the units in $F[X]$ are the non-zero constants. Just as every non-zero integer is the product of a unit and a natural number so also every non-zero polynomial is the product of a constant and a monic polynomial. Just as every natural number other than 1 can be written uniquely (upto order) as a product of primes so also every monic non-constant polynomial can be written uniquely (upto order) as a product of monic irreducible polynomials, i.e, those monic polynomials $p(X)$ such that if $p(X) = u(X)v(X)$, where $u(X), v(X)$ are monic polynomials then either $u(X)$ or $v(X)$ must be 1.

To simplify let us assume that $F$ is of characteristic 0. In general the monic irreducible polynomials in $F[X]$ can be quite difficult to determine and may require more than a single parameter to define but there is one case in which a single parameter defines a monic irreducible in $F[X]$, viz., when $F$ is algebraically closed. Then the monic irreducibles are of the form $X - a$ where $a \in F$ (and so uniquely determined by the single element $a \in F$). And so just as we know that every natural number $n \neq 1$ can be written uniqely (upto order) in the form $p_1^{e_1} p_2^{e_2} ....p_r^{e_r}$ where the $p_i$'s are distinct prime numbers we also know that every non-constant monic polynoial $p(x)$ can be be written uniquely (upto order) in the form $(X - a_1)^{e_1}(X - a_2)^{e_2} ....(X - a_r)^{e_r}$ where the $a_i$'s are distinct elements of $F$. Define then $ord_{p_i} n = e_i$ and analogously $ord_{a_i} p(X) = e_i$. Then we have $\log n = \sum_1^r \log p_i . ord_{p_i} n$ and analogously deg $p(X) = \sum_i^r 1 . ord_{a_i} p(X)$.

What follows is part of some lectures given by Richard Mason at Cambridge. It needs to get wider publicity. So suppose that $A(X), B(X), C(X)$ are non-zero polynomials in $F[X]$ and $n$ is a natural number such that $A(X)^n + B(X)^n = C(X)^n$.

*Corresponding author. Email: rajattan@gmail.com

Will there be any restrictions on $n$ ? If $A(X) = a, B(X) = b, C(X) = c$ are non-zero constants then since we are assuming that $F$ is algebraically closed and of characteristic zero (so every element of $F$ has an $n$th root) there will be plenty of solutions of $a^n + b^n = c^n$ whatever $n$ may be. So let us assume that not all of $A(X), B(X), C(X)$ are constants. Again if $a, b, c \in F - \{0\}$ are such that $a^n + b^n = c^n$ then for any polynomial $p(X)$ we have $(ap(X)^n) + (bp(X)^n = (cp(X))^n$. With these kind of trivial exceptions in mind let us reformulate our question as follows: If we have non-zero polynomials $A(X), B(X), C(X) \in F[X]$ such that they are MUTUALLY COPRIME and not all constants and if $A(X)^n + B(X)^n = C(X)^n$ are there any restrictions on $n$ ? Remarkably the answer is that $n$ must be less than 3. Even more remarkably the proof is almost trivial (especially compared to the Wiles-Taylor proof !). So let us formally state our theorem:

**Theorem:** Let F be an algebraically closed field of characteristic zero. Suppose we have mutually coprime non-zero polynomials $A(X), B(X), C(X) \in F[X]$, not all constants, and for some natural number $n$ we have $A(X)^n + B(X)^n = C(X)^n$. Then $n < 3$.

Before we go to the proof let me remark that there are infinitely many triples $(A(X), B(X), C(X))$ satisfying the conditions of our theorem and such that $A(X)^2 + B(X)^2 = C(X)^2$. For instance if $t$ is a natural number then $((1 - X^{2t}), 2X^t, (1 + X^{2t}))$ is such a triple.

For the proof we first introduce some notation. If $p(X)$ is a non constant polynomial in $F[X]$ we denote by $Z_p = \{a \in F | p(a) = 0\}$, in other words $Z_p$ is the set of roots of $p(X)$ in $F$.

**Lemma:** Let $A(X), B(X), C(X) \in F[X]$ be non-zero polynomials, not all constant and mutually coprime such that $A(X) + B(X) = C(X)$. Let $Z = Z_A \cup Z_B \cup Z_C$. Assume degree of $A(X)$ is greater than or equal to the degree of $B(X)$. Then $|Z| >$ degree $A(X)$:

**Proof:** We first remark that the fact that $A(X), B(X), C(X)$ are mutually coprime simply means that they have no common roots, i.e., that the union that defines $Z$ is a disjoint union. For a polynomial $p(X)$ by $p'(X)$ we simply mean the (formal) derivative of $F$, i.e., if $p(X) = \sum_0^n a_i X^i$ then $p'(X) = \sum_1^n i a_i X^{i-1}$. We define the Wronskian of $A, B$ to be the polynomial

$$W(A, B)(X) = \det \begin{bmatrix} A(X) & B(X) \\ A'(X) & B'(X) \end{bmatrix}$$

Hence $W(A, B)(X) = A(X)B'(X) - A'(X)B(X)$. By the simple properties of determinants and the fact that $A(X) + B(X) = C(X)$ we know that

$$W(A, B)(X) = W(C, B)(X) = W(A, C)(X)$$

We call this polynomial just W(X). We claim that if $a \in F$ then

$$ord_a W \geq ord_a A + ord_a B + ord_A C - 1$$

If $a \notin Z$ then the left hand side of the inequality is greater than or equal 0 whilst the right hand side is -1. Suppose $a \in Z_A$. Then $a \notin Z_B$ and $a \notin Z_C$ and so $ord_a B = ord_a C = 0$. Suppose $A(X) = (X - a)^e u(X)$ Then $W(X) = (X - a)^e u(X)B'(X) - e(X-a)^{e-1}u(X)B(X) - (X-a)^e u'(X)B(X)$. This is clearly divisible by $(X-a)^{e-1}$ and so $ord_a W \geq e - 1 = ord_a A + ord_a B + ord_a C - 1$. A similar proof can be given

if $a \in Z_B$ or $a \in Z_C$. Summing over all $a \in Z$ we have

$$\sum_{a\in Z} ord_a W \geq \sum_{a\in Z} ord_a A + \sum_{a\in Z} ord_a B + \sum_{a\in Z} ord_a C - |Z|$$

But $\sum_{a\in Z} ord_a A = \sum_{a\in Z_A} ord_a A =$ degree $A$ because for $a \notin Z_A$ we have $ord_a A = 0$. Similarly for $B$ and $C$. Hence we have

$$\sum_{a\in Z} ord_a W \geq \text{degree}A + \text{degree}B + \text{degree}C - |Z|$$

If $Z'$ is the set of roots of $W(X)$ then

$$\text{degreeW} = \sum_{\text{a}\in\text{Z}'} \text{ord}_\text{a}\text{W} = \sum_{\text{a}\in\text{Z}\cup\text{Z}'} \text{ord}_\text{a}\text{W} \geq \sum_{\text{a}\in\text{Z}} \text{ord}_\text{a}\text{W}$$

Hence

$$\text{degreeW} \geq \text{degreeA} + \text{degreeB} + \text{degreeC} - |\text{Z}|$$

From the way $W(X)$ is defined it is clear that

$$\text{degree}W(X) = \text{degree}W(C,B)(X) \leq \text{degree}B(X) + \text{degree}C(X) - 1$$

Putting these last two inequalities together we get $|Z| - 1 \geq$ degree $A(X)$ or $|Z| >$ degree $A$. q.e.d.

We are now in a position to prove our main theorem. Observe that for any non-constant polynomial $A(X)$ we have $Z_{A^n} = Z_A$. Let now $A(X), B(X)$ and $C(X)$ be non-zero polynomials in $F[X]$ which are not all constants and are mutually coprime such that $A(X)^n + B(X)^n = C(X)^n$. Let, as before, $Z = Z_A \cup Z_B \cup Z_C$ (disjoint union). Assume without loss of generality that degree $A(X) \geq$ degree $B(X)$. Then by the previous lemma we have $|Z| > n\text{degree}A$. But

$$|Z| = |Z_A| + |Z_B| + |Z_C| \leq \text{degree}A + \text{degree}B + \text{degree}C \leq 3\text{degree}A$$

since, by our assumption the degree of $A(X)$ is the maximum of the degrees of $A(X), B(X), C(X)$. Hence $n\text{degree}A < |Z| \leq 3\text{degree}A$ and so $n < 3$. q.e.d

A natural qustion to ask is whether there is. in the context of integers, any lemma like the lemma above. Recall that if $n$ is a natural number other than 1 and if $n = p_1^{e_1} p_2^{e_2} ... p_r^{e_r}$ is the unique factorization of $n$ as a product of distinct prime powers then $\log n = \sum_{i=1}^{r} logp_i \ ord_{p_i} n$. Similarly if $n(X)$ is a non-constant polynomial in $F[X]$ where $F$ is a field and $n(X) = cp_1(X)^{e_1} p_2(X)^{e_2} ... p_r(X)^{e_r}$ is the unique factorization of $n(X)$ as a product of powers of distinct monic irreducible polynomials then degree $n(X) = \sum_{i=1}^{r}$ degree $p_i(X).ord_{p_i(X)} n(X)$. So $\log n$ is analagous to degree $n(X)$. $|Z|$ may thus be interpreted as the sum the degrees of all the distinct, irreducible factors of $A(X)B(X)C(X)$. (Remember these are all of degree 1 when $F$ is algebraically closed.) Define now for a natural number $n \neq 1$ the RADICAL of $n$ (denoted by rad($n$)) to be the set of distinct prime factors of $n$ and define rad(1) to be $\{1\}$. Then an exact replica of lemma 1 in the context of natural numbers would state that of $a, b, c$ are mutually coprime natural numbers

such that $a+b=c$ then

$$\sum_{p\in \mathrm{rad}(abc)} \log p > \max(\log a, \log b, \log c) = \log c$$

or $\prod_{p\in \mathrm{rad}(abc)} p > c$. This, of course, is not true as the example of $a=1, b=8, c=9$ shows. However, in the mid eighties of the last century Masser and Oesterle made the following conjecture: define for a natural number $a$ $Rad(a) = \prod_{p\in \mathrm{rad}(a)} p$. Let $\epsilon > 0$. Then there are only finitely many natural numbers $c$ such that for any natural number $a < c$, coprime to $c$, we have $c > Rad(a(c-a)c)^{1+\epsilon}$. This can be reformulated as follows:
**The abc conjecture:** Let $\epsilon > 0$. Then there exists an absolute constant $k(\epsilon)$ such that for any triple of natural numbers $(a, b, c)$, mutually coprime, such that $a+b=c$ we have $c < k(\epsilon)Rad(abc)^{1+\epsilon}$.
The fact that the $\epsilon$ is necessary can be seen from the following example. Let $p$ be an odd prime and let $a=1,\ b=2^{p(p-1)}-1$ and $c=2^{p(p-1)}$. Then by Euler's theorem we know that $p^2$ divides $b$. Then $Rad(abc) = 2Rad(b) = 2Rad(\frac{b}{p}) \leq 2\frac{b}{p} < 2\frac{c}{p}$. Clearly no matter what absolute constant $k$ we take we cannot have $c < 2\frac{k}{p}c$ irrespective of $p$.
Suppose the conjecture is true and that $k(\frac{1}{2}) = 1$. Then if $a, b, c$ are mutually coprime natural numbers such that $a^n + b^n = c^n$ the conjecture would imply that $c^n < Rad(a^n b^n c^n)^{\frac{3}{2}} = Rad(abc)^{\frac{3}{2}} < (abc)^{\frac{3}{2}} < c^{\frac{9}{2}}$. Hence $n < 5$ and we would only need to prove FLT for $n = 3, 4$.

In 2012 Shinichi Mochizuki of Kyoto University announced that he had a proof of the conjecture but very few people have been able to understand it. Even the 2018 Fields medallist Peter Scholze thinks that there are gaps in the proof. So we are now in the strange position that whilst many in Japan think that the conjecture has been proved the rest of the world thinks not !

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 26–37

# *The Folding Mathematics*

**Archana S. Morye**[a] *
[a] ***University of Hyderabad***

**Abstract:** Origami is the art of paper folding, and it borrows its name from two Japanese words *ori* and *kami*. In Japanese, ori means folding, and the paper is called kami. While origami is just a hobby to most, there is a lot more to it. If you fold a square sheet of paper into any of the traditional origami model (for example the flapping bird) and unfold it, you can see crease patterns. These crease patterns tell us that there is a lot of geometry hidden behind the folds.

In this article, we investigate the symbiotic relationship between mathematics and origami. The first part of this article explores the utility of origami in education. We will see how origami could become an effective way of teaching methods of geometry, mainly because of its experiential nature. Complex origami patterns cannot be created out of thin air. They usually involve understanding deep mathematical theories and the ability to apply them to paper folding. In the second part of the article, we attempt to provide a glimpse of this beautiful connection between origami and mathematics.

**Keywords:** Origami, geometry, paper folding, fold-able numbers, tree-maker, cubic polynomials.

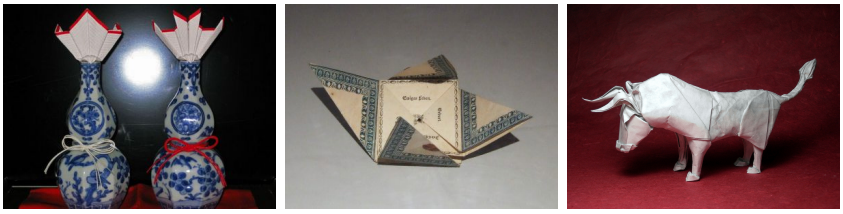
## 1. Introduction

Origami is a technique of folding paper into a variety of decorative or representative forms, such as animals, flowers etc. The origin of origami can be traced back to Japan. Originally, the art of paper folding was called as *orikata*, the craft acquired its current name in 1880 [5].

The early evidence of origami in Japan suggests that origami was primarily used as a ceremonial wrapper called the Noshi. *Noshi* is a wrapper which is attached to a gift, expressing good wishes (similar to greeting cards of today). A popular such Noshi is a pair of paper butterflies known as *Ocho* and *Mecho* that were used to decorate sake bottles (see Figure 1(a)[1]). Origami initially was an art of the elite, mainly because the paper was a luxury item. As the paper became more accessible, origami also became a well-practiced art.

The practice of origami can also be traced to Europe, the baptismal certificates

Email:sarchana.morye@gmail.com
[1]Pic source: `https://www.origami-resource-center.com/regular-mecho.html`



(a) Japanese Origami (b) European Origami (c) Wet Folding

Figure 1.

issued during the sixteenth century were folded in a specific way. (see Figure 1(b)[2]). Here, the four corners of the paper was folded repeatedly to the center. Interestingly this techniques is very different from the ones used in Japan. It is said that such a crease pattern closely resembles an old astrological horoscopes. For this reason, historians believe that folding in Europe developed more-or-less independently [5]. Some of the popular origami models from Europe are the Pajarita, the Cocotte and the boat .

### 1.1. *The modern origami*

The modern era of origami can mainly be attributed to the grand master of origami, Akira Yoshizawa. It was because of his relentless efforts that origami has transformed into a living art, from a mere craft. Apart from contributing towards developing more than 50,000 origami models, he also pioneered the wet-folding technique (see Figure 1(c)[3]). This technique involves slightly dampening the paper before using it for folding. This technique allowed the paper to be manipulated more easily, resulting in models with rounded and sculpted looks. The famous Yoshizawa-Randlett diagramming system is also his invention. Since the introduction of this system, origami has seen several advancements. Origami is now one of the well-established topics of research in major universities across the world.

Origami as it exists today has several variations, we list some of them below.

(1) *Pure Origami* : This form of origami is arguably one of the oldest and well studied form, the models here are made from a single square sheet of paper, without the use of scissors and glue. Coloring the final model is also strictly discouraged. Several models along with instructions can be found in `www.happyfolding.com`. My personal favorites and recommendations include the traditional crane, the swallowtail butterfly and the fawn models.

A more stringent variation is the *Pureland Origami* developed by John Smith. This version disallows certain types of folds allowed by the pure origami. Interestingly this type of origami is considered to be disable friendly.

(2) *Action Origami*: This form of origami involves developing models that can be animated. The most famous among the models of this type is the flapping bird. The bird flaps its wings when its tail is moved. Other interesting models of this type include origami airplanes and the instrumentalist created by Prof. Robert Lang.

(3) *Modular Origami:* In this form of origami, several identical units are folded and then assembled into a more complex origami model. An of this type is the Kusudama flower. In this model, sixty identical units are folded and arranged

[2]Pic source: `https://www.origami-resource-center.com/history-of-origami.html`
[3]Pic source: `https://en.wikipedia.org/wiki/Wet-folding`

(a) Fujimoto Hydrangeas (b) Kusudama (c) knotology torus

Figure 2.

into twelve flowers. These flowers are then arranged such that they form a regular dodecahedron, the pieces are held in place using glue or a thread. Many models have pocket and flap in each unit, so that units bind together without a glue or a thread.

(4) *Origami Tessellation:* An origami tessellation is created by repeating a pattern multiple times in all the directions, and this creates a mosaic. The kind of folds in this process predominantly includes pleats and twists. The invention of this technique can be attributed to Shuzo Fujimoto. This type of origami has an additional feature, and they produce a beautiful effect when they are backlit (when they are held against the light). Fujimoto Hydrangeas (see Figure 2(a)) is an interesting model of this type.

(5) *Strip Origami*: Strip folding is a technique that involves both paper folding and paper weaving. A fascinating model of this type is that of knotology torus (see Figure 2(c)[4]) by Dáša Ševerová.

While origami certainly has evolved as an amazing art, its applicability has also been phenomenal. Origami-inspired techniques are being sought after by almost every engineering field, ranging from space science to medical equipment and even automobile manufacturers. In medicine, origami techniques are often applied to stent designs. Stents are collapsible tubes that can be inserted into a patient's veins or arteries. When deployed, the stent expands to open the veins or arteries to improve blood flow. Origami design techniques are instrumental in developing thin and small stents. NASA's James Webb Telescope (JWST), the planned successor of Hubble space telescope, is a rather sizable infrared space telescope with a primary mirror of 6.5-meters. Origami techniques are being deployed to fold such a large telescope compactly so that it can be airlifted to space, where it can be unfolded again. Automobile manufacturers are pursuing efficient flat-folding techniques for airbags so that it occupies less space and yet unfurls quickly enough when needed.

The use of origami techniques in the science and engineering field is endless. Hence it is only prudent to study and understand origami as a science. One of the aims of this article is to illustrate the deep connections of origami with mathematics. The article is divided into two parts; the first part deals with using origami as a means to understand mathematics, geometry in particular. For example, a proof for the Pythagoras theorem merely is folding a square sheet of paper. Trisecting a line using just folds is possible. Interestingly, even trisecting an angle can be achieved. The latter is of significant interest due to its impossibility within the realms of Euclidean geometry. The second part of the article briefly dwells upon the need for mathematics to design complex origami models.

[4]Pic source:`https://www.flickr.com/photos/dasssa/3426754850`

Before we delve into the technical details, I would like to share my experience with the folding and origami community. While this subsection is unconventional, the reason I write this is to get an opportunity to acknowledge individuals who helped and inspired me during my journey. I also hope that this will nudge others to start their journey into the world of folding.

### 1.2. *My Origami Journey*

My first exposure to origami was like any other kid when I learned to fold simple models like box, flower, purse, and so on in school. The first complicated model I folded was that of a fish, from a magazine I found in my relatives' place. Learning this model gave me a huge sense of accomplishment and satisfaction. My first formal tryst with origami was during my undergraduate days when I chanced to find a Marathi and English origami series written by Indu Tilak [16]. This series is part of the textbooks prescribed by Maharashtra board for primary education. Even though these are school books, they are quite interesting and extensive. The book includes the necessary foundations to start creating basic and complex structures in origami.

While I was helped and inspired by many in the origami community, one who was extremely helpful and kind to me was Dáša Ševerová. Papers are the lifeline for origami artists. Some particular models require individual papers. Dáša Ševerová was kind enough to help me obtain certain papers that were difficult to get in India. She has also helped me fold some intricate origami tessellations.

These are some other origami books that one may wish to read.

- Origami Tessellations: Awe-Inspiring Geometric Designs, by Eric Gjerde [3].
- Origami Boxes by Tomoko Fuse [2].
- Origami Butterflies by Micheal LaFusse [11].
- Origami Journey: Into the Fascinating World of Geometric Origami, by Dáša Ševerová [15].
- Origami Inspiration by Meenakshi Mukerji [14].

The website `www.happyfolding.com`, owned by Sara Adam, is a one-point source of abundant information, beginners or otherwise.

## 2. Origami for Mathematics

Mathematics has unfortunately attained the notoriety of not just dreading the youngsters but also the adults. How many times have we not heard the phrase, "Thank god, I do not have to study mathematics anymore." This feeling has been captured aptly by the following Marathi couplet.

भोलानाथ उध्या आहे गणिताचा पपेर
पोटात माझ्या कळ येऊन दुखेल का रे ढोपर

Bholanath is a mystical, mythological, and benevolent Ox. The children pray to him for ill-health, for it is their mathematics paper the next day. It is not very surprising that mathematics is viewed so unfavorably, as it involves many abstract concepts challenging to comprehend. Even a simple definition such as an area or volume is very abstract, for it is a measurement irrespective of the shape. One reason why mathematics lacks the popularity of the other sciences is that it lacks visual appeal. Prof. John J Hopfield (the recipient of the ICTP

Dirac medal, 2001) in one of his articles [6] wrote that the reason for him being a scientist was because of the encouragement he received to do experiments. Labs and experiments are never associated with mathematics. While one could still argue that mathematics does provide the same experience through puzzles and problems, my personal experiences and memories negate such claims. Solving mathematical problems in the current day scenario has degenerated to learning to apply formulas to score high. Here is where origami can fill in to provide the missing fun.

*Example* 2.1 Consider a simple problem of proving that the sum of interior angles of a triangle adds up to 180°. While there are many ways to prove this, the folding in Figure 3 demonstrates this clearly and crisply. Here the desired rectangle is

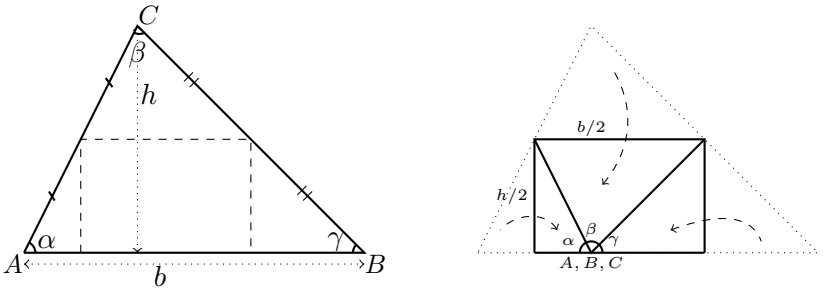



Figure 3.

identified, and the cones of the triangle are folded so that their tips meet. Since the angles $\alpha, \beta, \gamma$ covers a straight line, this proves that their sum is indeed 180°. Now the same folding also provides us with the argument of why the area of a triangle is $1/2 \times$ base $\times$ height. Notice that the fold covers the rectangle with height $h/2$ and length $b/2$ two times. So the area of the triangle is two times the area of this rectangle. This immediately provides us with the relation

$$\text{Area}(\triangle ABC) = 2 \times (b/2) \times (h/2) = 1/2 \times b \times h.$$

This demonstrates that every folding has a deep mathematical connection, possibly many. Discovering them depends on the creativity of the folder. After all, creativity is in the eye of the beholder, beautifully summarized by our beloved Dr. A. P. J. Abdul Kalam as

**"Creativity is seeing the same thing but thinking differently."**

We will demonstrate the deep connection of origami with mathematics through another example, this time, the famous Pythagorean Theorem. The Pythagorean theorem is one of the oldest known theorems and was studied by Babylonian, Egyptian, Indian, and Greek mathematicians centuries earlier. It states that the square of the hypotenuse of any right-angled triangle is equal to the sum of the squares of the other two sides. This geometric theorem probably has the most number of proofs. The standard proof which is given in the most high-school books is using similar triangles. Another exciting proof is as follows. Let $\triangle ABC$ be any right-angle triangle with $c$ as its hypotenuse and its other two sides being $a, b$. Let $C$ be a square-shaped bucket with length and height being $c$-units and its width 1-unit. Similarly, let $A$ ( respectively $B$) be a square-shaped bucket with length and height $a$ (respectively $b$) and with width exactly 1-unit. It can be demonstrated that the $C$ bucket can be filled using the water in buckets $A$ and $B$, respectively. While this is undoubtedly a fun proof, the knowledge of volumes is needed to understand the proof. Further conducting such an experiment in a class is cumbersome.

*Example* 2.2 Now consider the folding in Figure 4, it clearly demonstrates the proof of Pythagoras Theorem. In the figure, we take a square sheet of paper and

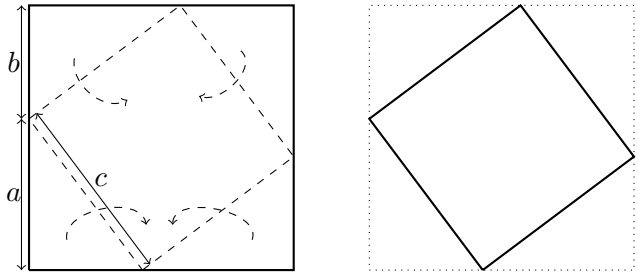



Figure 4.

mark out four right-angled triangles of equal sizes along the four corners. Each of these triangles has $c$ as its hypotenuse and $a, b$ as the size of its other sides. Each of these triangles has its hypotenuse in the inner part of the square, touching each other. Dotted lines in the figure denote these. Folding along the hypotenuse gives us a square of length $c$. Notice that the area of such a square is $c^2$, the area of the original square we started with was $(a+b)^2$. What was folded in were 4 triangles each of area $1/2 \times a \times b$. With this, we get the following equation, which also proves the Pythagoras Theorem.

$$c^2 = (a+b)^2 - (4 \times (1/2 \times a \times b)) = a^2 + b^2.$$

While we could go on demonstrating the utility of origami in proving theorems involving simple properties (for example see [4]), one may ask whether origami can also be used to solve more involved problems and theorems. For this, we need to formalize folding, i.e., make precise what kind of folds are allowed and what are not. This would allow us to investigate the constructible geometric objects through origami. This is in the same lines as the classical *ruler-and-compass* construction.

## 2.1. *Huzita-Hatori Axioms*

The formal axioms for origami is given by the *Huzita–Hatori axioms*. We briefly recall them here and direct the interested readers to [7, 10] for a comprehensive coverage on the subject.

(1) Given two distinct points $p_1$ and $p_2$, there is a unique fold that passes through both of them.
(2) Given two distinct points $p_1$ and $p_2$, there is a unique fold that places $p_1$ onto $p_2$.
(3) Given two lines $l_1$ and $l_2$, there are folds that places $l_1$ onto $l_2$.
(4) Given a point $p_1$ and a line $l_1$, there is a unique fold perpendicular to $l_1$ that passes through point $p_1$.
(5) Given two points $p_1$ and $p_2$ and a line $l_1$, there are folds (possibly empty) that places $p_1$ onto $l_1$ and passes through $p_2$.
(6) Given two points $p_1$ and $p_2$ and two lines $l_1$ and $l_2$, there are folds (possibly empty) that places $p_1$ onto $l_1$ and $p_2$ onto $l_2$.
(7) Given a point $p$, and two lines $l_1$ and $l_2$, there are folds (possibly empty) perpendicular to $l_2$ that places $p$ onto line $l_1$.

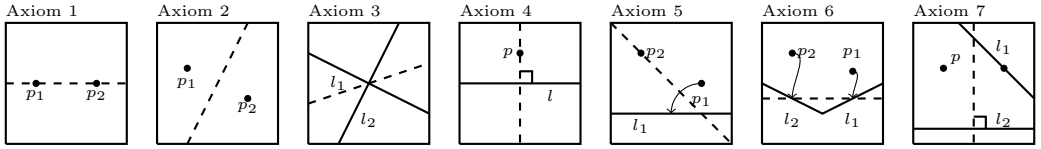



Figure 5.

*Remark* 2.3 The relevant question here is, what can these postulates achieve in comparison to the classical ruler-and-compass. While the above set of operations are termed as *axioms*, they do not necessarily indicate that such a fold is achievable or unique (see [8, 9] for details). For example, in Axiom 5, it is impossible to obtain a fold when $p_1 = p_2$. Similarly in Axiom 6, it is impossible to obtain a fold if $p_1 = p_2$ and $\ell_1 \neq \ell_2$ two parallel lines. Rest of the article only deals with the scenarios where the relevant fold is achievable.

Towards understanding the axioms, we note that the first four postulates are self-explanatory. We will examine the fifth and sixth postulates more closely. Interestingly the fifth postulate can be used to solve a quadratic equation and the sixth a cubic equation. We will demonstrate the latter in the sequel. We first prove the following lemma which states that the dotted line in the Axiom 5 of Figure 5 is actually a tangent to a parabola with its focus on $p_1$ and the directrix on $l_1$. This proof is an adaptation from [10]. Recall that a parabola is those set of points (called the *locus*) that are equidistant from a fixed point (called the *focus*) and a fixed-line (called the *directrix*).

*Notation* 2.4 If $d(x, y)$ denotes the *Euclidean distance* between two points $x$ and $y$, then the distance between the point $x$ and the line $l$ is equal to $\min\{d(x, y)|y \in l\}$, and is denoted by $d(x, l)$. Notice that, $d(x, l) = d(x, y)$ where $y$ is the foot of a perpendicular drawn from the point $x$ to the line $l$. This distance $d(x, l)$ is called the *perpendicular distance* between $x$ and $l$.

Lemma 2.5 *Given two points $p_1$ and $p_2$ and a line $l_1$, assume that there exists a fold $\ell$ that places $p_1$ onto $l_1$ and passes through $p_2$. Then $\ell$ is tangent to the parabola $\mathcal{P}$ defined by the focus $p_1$ and the directrix $l_1$.*

*Proof.* To prove the lemma, we prove that there is a unique point $x$ on the line $\ell$ which is equidistant from $p_1$ and $l_1$. By definition, this point will lie on the parabola. Since this point is unique, no other points of the line will lie on the parabola. This will prove that, the line will be tangential to $\mathcal{P}$.

For this, we prove two things. Firstly we prove that there is a point $x$ in $\ell$, which is equidistant from $p_1$ and $l_1$. Then we will prove that for any point $y \neq x$ in $\ell$, $d(y, p_1)$ is not equal to $d(y, l_1)$. Without loss of generality, we will assume that the line $l_1$ is the bottom edge of a square and that the point $p_2$ lies on the left edge of the square. Observe that given any $l_1, p_1, p_2$, one could arrange them in this manner in an appropriate large enough square. By assumption, the line $\ell$ folds $l_1$ in such a way that a point in it coincides with $p_1$. Let this point on $l_1$ be $\bar{p_1}$ (see Figure 6 for an illustration). Now draw a line perpendicular to the line $l_1$, starting at $\bar{p_1}$, let this line intersect $\ell$ at $x$. We claim that this intersection point $x$ is the required point. This is easy to observe since $\bar{p_1}$ coincides with $p_1$ when the paper is folded along $\ell$. This immediately implies that $d(x, p_1) = d(x, \bar{p_1})$.

For proving the second part (uniqueness), we again use the fact that $\bar{p_1}$ coincides with $p_1$ when the paper is folded along $\ell$. This immediately also tells us that for any point $y$ on the line $\ell$, $d(y, p_1) = d(y, \bar{p_1})$. Let $p_y$ denote the foot of the

perpendicular drawn from $y$ to line $l_1$. Hence $d(y, l_1) = d(y, p_y)$. Now for any point $y \neq x$, consider the right angle triangle $\triangle y p_y \bar{p_1}$. In this triangle since $y\bar{p_1}$ is the hypotenuse, it follows that $d(y, \bar{p_1}) > d(y, p_y) = d(y, l_1)$. This proves that $y \notin \mathcal{P}$. Hence $\ell$ is tangent to $\mathcal{P}$. ■

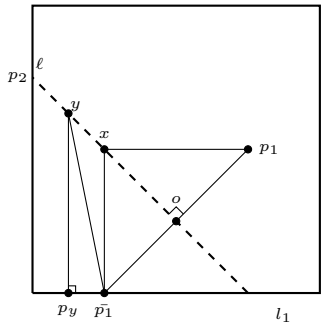



Figure 6.: Illustration for Lemma 2.5

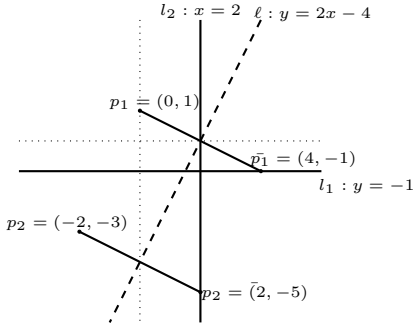



Figure 7.: Illustration of Theorem 2.7 on Eqn:$x^3 + 0x^2 - 3x - 2$

*Remark* 2.6 Notice that in the proof above, the role of $p_2$ is non-existent. The fact that $\ell$ is a tangent to the parabola $\mathcal{P}$ is invariant to $p_2$. Indeed, any fold obtained by placing the point $p_1$ onto the line $\ell_1$ creates a tangent to the parabola $\mathcal{P}$. Also, notice that there are infinite tangent lines to the parabola, one for every point on it. This collection of tangent lines is called the *tangent bundle*. The point $p_2$ picks out a set (of size $\leq 2$) of tangent lines from this tangent bundle.

Using the same technique of Lemma 2.5, one could also prove that the fold $\ell$ obtained in Axiom 6 is a line tangent to two parabolas $\mathcal{P}_1, \mathcal{P}_2$, determined by the focus and directrix $p_1$, $l_1$ and $p_2$, $l_2$ respectively. We will instead illustrate how to use Axiom 6 to solve a cubic equation. More specifically, we will show how to obtain a solution for equations of the form $x^3 + ax^2 + bx + c = 0$ where $a$, $b$, $c \in \mathbb{R}$. The idea here is to use the points $p_1 = (a, 1)$, $p_2 = (c, b)$, and the lines $l_1$ given by $y = -1$ and $l_2$ given by $x = -c$ in Axiom 6 and show that the slope of the fold obtained is a solution to the equation. The idea behind the Theorem is illustrated in Figure 7.

THEOREM 2.7 *Let $x^3 + ax^2 + bx + c = 0$ be any cubic equation with $a, b, c \in \mathbb{R}$. Consider a large enough square paper with origin at its center, let $p_1$ and $p_2$ be two point in it with the coordinates given by $(a, 1)$ and $(c, b)$ respectively. Let $l_1$ and $l_2$ be lines defined by the equation $y = -1$ and $x = -c$ respectively. Then the slope $t$ of the line $\ell$ that folds the point $p_1$ onto $l_1$ and $p_2$ onto $l_2$ is a solution to the given cubic equation.*

*Proof.* We wish to prove that any solution to the slope of the line $\ell$ obtained as a result of applying Axiom 6 to the points and lines specified below, is a solution to the given cubic equation $x^3 + ax^2 + bx + c = 0$. The points are specified as $p_1 = (a, 1)$ and $p_2 = (c, b)$ and the lines given by the equation $l_1 : y = -1$ and $l_2 : x = -c$. Let the equation defining the line $\ell$ be $\ell : y = tx + u$. Recall that, $t$ is the slope and $u$ is the $y$ intercept here.

Firstly notice that in both Axiom 5 and 6, the point $p_1$ is placed onto the line $l_1$. By Lemma 2.5 and Remark 2.6, we obtain that the line $\ell$ is tangent to the parabola defined by its focus on $p_1$ and the directrix on the line $l_1$. Let this parabola be $\mathcal{P}$. Using the focus given by $p_1$ and the directix given by line $l_1$, we obtain the equation

of $\mathcal{P}$ as

$$y = \frac{1}{4}(x-a)^2. \tag{1}$$

Since the line $\ell$ is tangent to the parabola, we obtain the equation of its slope as $t = \frac{\partial y}{\partial x} = \frac{1}{2}(x-a)$. Let $(x_1, y_1)$ be the point on $\ell$ where it makes contact with the parabola. Evaluating the equation of slope at this point provides us with:

$$t = \frac{1}{2}(x_1 - a). \tag{2}$$

Since we know the slope of $\ell$ and also that $(x_1, y_1)$ lies on $\ell$, we obtain the equation of $\ell$ as follows.

$$y = tx - tx_1 + y_1. \tag{3}$$

From this we obtain that $u = -tx_1 + y_1$. Further from Equation 1, we know that $y_1 = t^2$. We also know the equation of $x_1$ in terms of $t$ ( $x_1 = 2t + a$ from the Equation 2), as a result we get the following equation.

$$u = -t^2 - ta. \tag{4}$$

Notice that Axiom 6 mandates that the line $\ell$ is tangential to another parabola $\mathcal{P}'$, defined by the focus $p_2 = (c, b)$ and the line and $l_2 : x = -c$. The equation of this parabola is given by

$$x = \frac{1}{4c}(y-b)^2.$$

Applying the technique seen earlier, we obtain that the equation of the slope of $\ell$ in this case to be $t = \frac{2c}{(y-b)}$. Using this, we get $u = b + \frac{c}{t}$. Equating the value of $u$ obtained in Equation 4 with this, we obtain the following equation.

$$b + \frac{c}{t} = -t^2 - ta. \tag{5}$$

Notice that this translates to the equation $t^3 + at^2 + bt + c = 0$, matching with the original equation that we started with, this completes the proof. ■

## 3. Mathematics for Origami

While in the previous section, we saw that origami could be very handy to learn mathematical concepts, in this section effectively, we will see that the relationship is symbiotic. Many of the origami models require deep mathematical insights and techniques, and we will survey some of them.

The complexity of an origami model depends on the number of attachments it has. For example, an origami model resembling a bird with a head, a tail, and two wings is less complex than a model resembling a beetle with a head, two horns, and six legs. One approach to building an origami model is to come up with what is called a base. A *base* is a geometric object which roughly resembles the end product. Historically most of the origami models were constructed by trial and error basis.

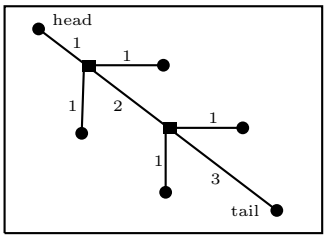



Figure 8.

That is, the paper is folded roughly in the direction of the base until the desired objective is achieved. One of the most difficult and important tasks in coming up with an origami construction is to identify how and when to create a fold. This translates to identifying the crease pattern required for the construction. While for simpler models, the trial and error approach is possible, it is highly inefficient for complex models. One important and pertinent question in this regard is whether one can come up with the crease patterns that can be folded into the desired base. The computer programs TreeMaker and Origamizer achieve this objective. Here, we survey the techniques used in the Tree Maker program. We closely adopt the definitions and language used in the TreeMaker manual [13]. To keep the article simple, we are making our exposition is very brief and informal. We direct the interested readers to [1, 12] for an extensive exposition on the subject. TreeMaker is a computer program developed by a famous origami artist and scientist Robert Lang. The software has been used to develop several complex origami models, see Figure 9[5].

Figure 9.

Given any base, let its tree diagram be the graph obtained by shrinking its skeleton to straight lines. In reality, the edges of such a tree diagram can have weights; these weights correspond to the size of foldings required in the base. For example, Figure 8[6] is the tree diagram of the lizard model next to it. Notice that it has a head, a tail, two forelegs, and two hind legs, not all have the same size. A tree diagram is said to be uni-axial if firstly it is tree-like (i.e., it is connected and has no cycles) and further, it has a main stem, and every branch only originates from this stem. The TreeMaker algorithm works when the required base is of type uni-axial.

The tree diagram is crucial to obtaining the desired algorithm. Every vertex (in a graph sense) of such a tree diagram, which has no outgoing edges, is called the terminal node. In the figure, these are represented by circular nodes. Every other vertex is called the *internal node*; in the figure, these are the rectangular nodes. An important mathematical property that allows a given uni-axial tree diagram to be folded to its base is as follows:

[5]Pic source: `https://langorigami.com/artworks/`
[6]Pic source: `https://origami.me/lizards/`

*If one can find in a square, a set of points, each corresponding to a vertex in the tree diagram such that the distance between any two points is greater or equal to the distance between the corresponding vertices in the tree diagram, then it is possible to fold such a set of points into the required base.*

Notice that the property is existential and does not immediately provide a way to recover the folding pattern. To obtain the crease pattern, the first observation is as follows:

*If the distance between any two terminal points is equal to the distance between their corresponding vertices in the tree diagram, then the line between them is definitely a crease in the final base.*

The algorithm crucially identifies points in the square so that maximal such creases occur, partitioning the square into polygons. This already is a computationally hard problem and has deep connections to a famous graph theoretical problem called the *cycle packing problem*.

The algorithm then depends on yet another mathematical property that: creases for each polygon partition of the square can be identified separately. The handling of a triangle and quadrilateral is well known, i.e., there are well-known methods to find creases for such simple polygons. However, other complex polygons can create complexities. One way to get around this is to simplify these polygons by adding extra terminal vertices to the tree diagram. This roughly translates to refining the creases and hence, partitioning of the complex polygons into simpler polygons. The algorithm can now be summarized as follows:

- Obtain the tree diagram of the desired base.
- Obtain points on the square such that the points have appropriate distances.
- Mark all the creases with minimum lengths.
- Identify the polygons partitioning the square, based on the crease marking obtained earlier.
- Process the crease pattern for the polygons individually. Simplify the polygons if needed.

While this algorithm provides the necessary crease pattern to obtain the base, identifying whether a crease is an inside fold (valley fold) or an outside fold (mountain fold) is another problem entirely. While there are effective heuristics for this, the problem, in general, is still open.

Another interesting problem that one encounters in origami is that of identifying whether a given crease pattern can be *flat-folded*. An origami model is said to be *flat-folded* if it can be compressed without making any additional creases. The question is whether it can be determined automatically, just by looking at the crease pattern, if it can be flat-folded. While there are results for sub-classes of the crease pattern due to Maekawa Jun and Kawasaki Toshikazu, the general problem still remains open. Thus, there are very many mathematical properties of origami that are being actively studied, some of which are still waiting to be solved. This demonstrates the profound impact that mathematics has on the folding art.

## 4. Conclusion

In this article, we investigated the connections between origami and mathematics. We first surveyed different types of origami models that are in vogue. We then showed how origami can effectively be used to prove mathematical theorems. We

also explored briefly the use of mathematics in constructions of complex origami models.


### *Acknowledgment*

The author would like to thank UGC SAP (DSA I) for providing financial support. The author would like to thank the anonymous referee for helpful comments and suggestions.


## References


[1] Erik D. Demaine and Joseph O'Rourke. *Geometric Folding Algorithms: Linkages, Origami, Polyhedra*. Cambridge University Press, 2008.

[2] T. Fuse. *Amazing Origami Boxes*. Dover Publications, 2018.

[3] E. Gjerde. *Origami Tessellations: Awe-inspiring Geometric Designs*. A K Peters, 2009.

[4] K. Haga, J. Fonacier, and M. Isoda. *Origamics: Mathematical Explorations Through Paper Folding*. World Scientific, 2008.

[5] Koshiro Hatori. History of origami in the east and the west before interfusion. In *Origami 5*. CRC Press, 2016.

[6] John J. Hopfield. Growing up in science. In *One hundred reasons to be a scientist*, 2004.

[7] Thomas Hull. Solving cubics with creases: The work of beloch and lill. *American Mathematical Monthly*, 2011.

[8] Asem Kasem, Fadoua Ghourabi, and Tetsuo Ida. Origami axioms and circle extension. In *Symposium on Applied Computing*, 2011.

[9] H. R. Khademzadeh and H. Mazaheri. Some results to the huzita axioms.

[10] Jaema L. Krier. Mathematics and origami : The ancient arts unite. 2007.

[11] M.G. LaFosse and R.L. Alexander. *Michael LaFosse's Origami Butterflies: Elegant Designs from a Master Folder*. Tuttle Publishing, 2013.

[12] Robert J. Lang. A computational algorithm for origami design. In *Proceedings of the Twelfth Annual Symposium on Computational Geometry*, SCG '96. ACM, 1996.

[13] Robert J. Lang. A program for origami design, treemaker 4.0. 1998.

[14] Meenakshi Mukerji. *Origami Inspirations*. CRC Press, 2010.

[15] Dáša Ševerová. *Origami Journey: Into the Fascinating World of Geometric Origami*. CreateSpace, 2018.

[16] Indu Tilak. *Origami Vol 1,2 and 3*. Tilak Brothers, 1989.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 38–45

# *Enumeration of Groups in Varieties of $A$-groups: A Survey*

**Geetha Venkataraman**[a*]

[a] *Ambedkar University Delhi, Lothian Road, Kashmere Gate, Delhi 110006, India.*

**Abstract:** Let $\mathcal{S}$ be a class of groups and let $f_{\mathcal{S}}(n)$ be the number of isomorphism classes of groups in $\mathcal{S}$ of order $n$. Let $f(n)$ count the number of groups of order $n$ up to isomorphism. The asymptotic bounds for $f(n)$ behave differently when restricted to abelian groups, $A$-groups and groups in general. We survey some results and some open questions in enumeration of finite groups with a focus on enumerating within varieties of $A$-groups.

**Keywords: finite groups, varieties of groups, enumeration, $A$-groups**.

**AMS Subject Classifications**: **20D10, 20C20, 20E10**

## 1. Introduction

Let $f(n)$ denote the number of isomorphism classes of groups of order $n$. Let $\mathcal{S}$ be a class of groups and let $f_{\mathcal{S}}(n)$ be the number of isomorphism classes of groups in $\mathcal{S}$ of order $n$. Some interesting classes that have been studied are the class of abelian groups, the class of solvable groups, varieties of groups, $p$-groups and $A$-groups, etc (see [1]). ($A$-groups are groups whose nilpotent subgroups are abelian.)

The asymptotic bounds for $f(n)$ behave differently when restricted to abelian groups, $A$-groups, and groups in general, primarily due to whether the group itself or its Sylow subgroups are abelian or not. In 1993, L Pyber proved a result which settled a conjecture about $f(n)$. The result [2] was published in The Annals of Mathematics and uses results related to Classification of Simple Finite Groups. Pyber showed that $f(n) \leq n^{(\frac{2}{27}+o(1))\mu(n)^2}$ as $\mu(n) \to \infty$, where $\mu(n)$ represents the highest power to which a prime divides $n$.

While Pyber's upper bound has the correct leading term, it is certainly not the case for the error term. The key to this puzzle may lie in deeper investigation of $A$-groups. A correct leading term for an upper bound of $f_A(n)$ could lead to the correct error term for an upper bound for $f(n)$. The best we do know about $A$-groups is that $f_{A,\,\mathrm{sol}}(n) \leq n^{7\mu(n)+6}$, where $f_{A,\,\mathrm{sol}}(n)$ is the number of isomorphism classes of solvable $A$-groups of order $n$ (see [3]). The method to finding the correct upper bound for the enumeration function for $A$-groups may be via enumerating within varieties of $A$-groups. (For varieties of groups see [4].)

*Corresponding author. Email: geetha@aud.ac.in

In this article we survey some of the results in enumeration of finite groups to provide a context for enumeration within a small variety of $A$-groups. The aim of the article is to provide a flavour of the representation theory and counting techniques used in such enumeration problems. Much of the material here is a selection of existing published work. Some unpublished material is also used to illustrate the structure of groups in a particular variety of $A$-groups.

The article unfolds as follows. The next section surveys main results in enumeration of finite groups with brief commentaries. The third section provides some basic background related to varieties of $A$-groups and specifically the varieties $\mathfrak{U} = \mathfrak{A}_p\mathfrak{A}_q$, and $\mathfrak{V} = \mathfrak{A}_p\mathfrak{A}_q \vee \mathfrak{A}_q\mathfrak{A}_p$, where $p$ and $q$ are distinct primes. A discussion on bounds for $f_{\mathfrak{V}}(n)$ is also presented with a sketch of the main steps of the proof. In the fourth and final section, we present unpublished material related to the structure of a finite group in the variety $\mathfrak{A}_p\mathfrak{A}_q\mathfrak{A}_r$, where $p, q$ and $r$ are distinct primes. The last section will also discuss some open questions related to enumeration in varieties of $A$-groups. Logarithms are taken to the base 2, unless stated otherwise.

## 2. A brief history of group enumerations

Enumeration questions and questions about classification of groups have been topics of study from even before the time when the notion of an abstract group was defined. From the latter part of the 19th century to the first quarter of the 20th century, several mathematicians worked on classifying and hence enumerating groups of specific types or order. For a good survey of these results see [5].

In 1895, Otto Hölder [6] gave a precise answer to the question about number of groups of order n when n was a product of distinct primes. If $n$ is square-free, that is, a product of distinct primes, then any group $G$ of order $n$ is meta-cyclic and so $G$ will be a semidirect product of a cyclic group by a cyclic group. He was able to use this to prove that if $n$ is square-free then

$$f(n) = \sum_{m|n} \prod_{p} \frac{p^{c(p)} - 1}{p - 1}$$

where, in the product, $p$ ranges over prime divisors of $n/m$ and $c(p)$ denotes the number of primes $q$ dividing $m$ such that $q \equiv 1 \bmod p$.

A lot of the modern work on group enumerations dates back to a paper [7], published in 1960, by Graham Higman. In the paper he showed that

$$f(p^m) \geq p^{\frac{2}{27}m^3 - O(m^2)}$$

where $p$ is prime. In 1965, Charles C Sims [8] showed that

$$f(p^m) \leq p^{\frac{2}{27}m^3 + O(m^{\frac{8}{3}})}.$$

An unpublished result of Mike Newman and Craig Seely, referred to and shown in [1], brings down the error term to $O(m^{\frac{5}{2}})$. This led to much speculation on what would be the corresponding estimates for $f(n)$.

Let $n = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$ be the prime decomposition of $n$. Define $\lambda(n) = \alpha_1 + \cdots + \alpha_k$. Note that when $n = p^m$, $p$ prime, then $\mu(n) = \lambda(n) = m$.

The speculation or conjecture that arose and which was quoted by McIver and Neumann [9] was that

$$f(n) \leq n^{(\frac{2}{27}+\epsilon)\lambda(n)^2}$$

where $\epsilon \to 0$ as $\lambda(n) \to \infty$.

The conjecture stemmed from the feeling that the reason for the number of groups of order $n$ must be because of the large number of choices we have for $p$-groups to chose as Sylow $p$-subgroups rather than the number of ways of putting the groups together.

This sentiment was shown to be right in 1991 by Laszlo Pyber [2]. He showed that the number of groups of order $n$ with specified Sylow subgroups is at most $n^{75\mu+16}$. This together with the choices available for $p$-groups to be Sylow subgroups, gives the result that $f(n) \leq n^{\frac{2}{27}\mu(n)^2+O(\mu(n)^{\frac{5}{3}})}$. Again, due to the results for $p$-groups we see that the above upper bound has the right leading term.

Broadly speaking, the leading term comes from the choices of $p$-groups that are available to be Sylow subgroups and the error term arises from the number of ways in which the Sylow subgroups can be put together to create the required group of order $n$. The error term above is certainly not the best.

Recall that a finite $A$-group is a group whose Sylow subgroups are abelian. A bound for solvable $A$-groups was first given in 1969 by Gabrielle Dickenson in [10]. She showed that $f_{A,\,\mathrm{sol}}(n) \leq n^{c \log n}$ for some constant $c > 0$. It was improved by McIver and Neumann in 1987 in [9], where they showed that $f_A(n) \leq n^{\lambda(n)+1}$. Since the number of choices for an abelian group of order $m$ up to isomorphism is at most $m$, Pyber's result on number of groups with specified Sylow subgroups, gives us $f_A(n) \leq n^{75\mu+17}$. His proof shows that for solvable $A$-groups, we can get $f_{A,\,\mathrm{sol}}(n) \leq n^{40\mu+17}$. The best bound know till date is $f_{A,\,\mathrm{sol}}(n) \leq n^{7\mu(n)+6}$, shown in [3]. However the bounds in the case of $A$-groups or even solvable $A$-groups is certainly not best possible. So a question still open is that what are the best possible constants $c > 0$ and $d > 0$ such that $f_A(n) \leq n^{c\mu+d}$?

We can see now that $A$-groups become an important class of groups from the enumeration point of view. Further if a 'best possible' value of $c$ is found in the upper bound of $f_A(n)$ then we will be closer to the correct error term in the bound for $f(n)$. Note that, whenever $n$ is a prime power, $f_A(n) \leq n$. It is therefore not possible to hope that there is a constant $c > 0$ such that $n^{c\mu} \leq f_A(n)$ for all $n$. Consequently we need to elaborate further on what is meant by a best possible $c$. We refer to [3] for this. Let $\mathcal{S}$ be a class of $A$-groups such that there are infinitely many $n$ for which $f_{\mathcal{S}}(n) \leq n$. Then $c > 0$ will be called best possible if there exists $d > 0$ such that $f_{\mathcal{S}}(n) \leq n^{c\mu+d}$ and given any $\epsilon > 0$ there exist infinitely many $n$ with $\mu(n)$ unbounded such that $n^{(c-\epsilon)\mu} < f(n)$.

In [9] and [2] it was shown that $f_{A,\,\mathrm{sol}}(n) \geq n^{c\mu}$ for certain specific types of $n$. In both the cases $0 < c < 0.08$. So the gap between these lower bounds and the best upper bounds we have is huge. The task therefore is to consider certain classes of solvable $A$-groups where the structure of the groups will allow for enumeration techniques leading to 'good' upper bounds and by that process also get 'good' lower bounds.

Let $\mathfrak{A}_n$ denote the variety of abelian groups with exponent $n$ and let $\mathfrak{U} = \mathfrak{A}_p\mathfrak{A}_q$, and $\mathfrak{V} = \mathfrak{A}_p\mathfrak{A}_q \vee \mathfrak{A}_q\mathfrak{A}_p$, where $p$ and $q$ are distinct primes. In [11], it was shown that $\mathfrak{U}$ and $\mathfrak{V}$ are both classes of solvable $A$-groups and that there exist positive constants $c = c_{p,q}, d = d_{p,q}, c' = c'_{p,q}, d' = d'_{p,q}$ such that $f_{\mathfrak{U}}(n) \leq n^{c\mu+d}$ and $f_{\mathfrak{V}}(n) \leq n^{c'\mu+d'}$, where $c$ and $c'$ are best possible in the sense discussed above.

Further it was shown in [1] that $c' = \max\{c_{p,q}, c_{q,p}\}$, where $c_{q,p}$ is the leading term for the product variety $\mathfrak{A}_q\mathfrak{A}_p$. The value of $c_{p,q}$ is given by

$$c_{p,q} = \frac{1}{d} - 2\sqrt{\left(\frac{\log p}{\log q}\right)^2 + \frac{\log p}{d \log q}} + 2\frac{\log p}{\log q},$$

where $d$ is the order of $p$ modulo $q$. In [12], Sophie Germain primes were considered. These are primes $q$ such that $p = 2q+1$ is also prime. For such $p, q$ we see that $d = 1$ and so the chance of getting larger value for $c_{p,q}$ occurs. It is not known if there are infinitely many Sophie Germain primes, but if we assume this and let $q$ tend to infinity, then $\frac{\log p}{\log q} \to 1$ and so $c_{p,q} \to 3 - 2\sqrt{2} = 0.17157\ldots$. If $q$ is the largest currently known Sophie Germain prime, namely $q = (2618163402417 \times 2^{1290000}) - 1$ and $p = 2q + 1$, then $c_{p,q}$ already agrees with $3 - 2\sqrt{2}$ to several decimal places. So $c = 0.171$ is the 'best' possible lower bound that we have currently for $A$-groups.

## 3. Varieties of $A$-groups

In this section we consider the varieties $\mathfrak{U}$ and $\mathfrak{V}$ and give an idea of some of the techniques used to enumerate within these varieties. The original work for this was done in [11] and then presented with modifications in [1].

A variety $\mathfrak{D}$ is said to be a variety of $A$-groups if all its nilpotent groups are abelian. It is sufficient to check if the finite nilpotent groups in the variety $\mathfrak{D}$ are abelian. Consider the product variety $\mathfrak{U} = \mathfrak{A}_p\mathfrak{A}_q$, where $p$ and $q$ are distinct primes. Any finite group in this variety is an extension of an elementary abelian $p$-group by an elementary abelian $q$-group and by the Schur-Zassenhaus Theorem it is a semi-direct product.

Let us now consider $\mathfrak{V} = \mathfrak{A}_p\mathfrak{A}_q \vee \mathfrak{A}_q\mathfrak{A}_p$. The finite nilpotent groups in $\mathfrak{V}$ are abelian and any group in $\mathfrak{V}$ is solvable. Indeed $\mathfrak{V}$ is locally finite. Thus $\mathfrak{V}$ is a locally finite solvable variety of $A$-groups. Further it can be shown that $\mathfrak{V}$ has exponent $pq$, any finite group of order $p^\alpha q^\beta$ and elementary abelian Sylow subgroups lies in $\mathfrak{V}$. The converse is also true. Any finite group $G \in \mathfrak{V}$ has order $p^\alpha q^\beta$ and its Sylow-subgroups will be elementary abelian.

Next we present a sketch proof of the main steps involved in finding the 'best' bounds for $\mathfrak{U}$ and $\mathfrak{V}$. As mentioned above any finite group in $\mathfrak{U} = \mathfrak{A}_p\mathfrak{A}_q$ is a semi-direct product of its elementary abelian Sylow $p$-subgroup by an elementary abelian Sylow $q$-subgroup. That is, if $G \in \mathfrak{U}$, then $G = P \rtimes Q$ where $P$ is a Sylow $p$-subgroup of $G$ and $Q$ is a Sylow $q$-subgroup of $G$.

To count the number of isomorphism classes of groups of order $p^\alpha q^\beta$ in $\mathfrak{U}$, it suffices to count the number of isomorphism classes of groups $P \rtimes_\theta Q$ where $P$ is a fixed elementary abelian group of order $p^\alpha$, $Q$ is a fixed elementary abelian group of order $q^\beta$ and $\theta$ runs over all homomorphisms from $Q$ to $\mathrm{Aut}P$. Indeed, two possible approaches could be taken at this stage. One would be to regard $\theta(Q)$ as a subgroup of $\mathrm{Aut}P$, and count the possibilities. The other would be to regard $P$ as an $\alpha$-dimensional $\mathbf{F}_pQ$-module and to then count the possibilities for $P$ up to isomorphism. The approach described below, essentially follows the second route.

We explore the structure of finite groups in $\mathfrak{U}$ further. Define $\mathcal{X} := \{G \in \mathfrak{A}_p\mathfrak{A}_q \mid G \text{ is finite and } Z(G) = 1\}$ and let $\mathcal{Y}$ be the same as $\mathcal{X}$ with $p$ and $q$ interchanged. Then for any finite group $G$ in $\mathfrak{U}$, there exists a group $G_1 \in \mathcal{X}$ such that $G = G_1 \times Z(G)$. Further, the isomorphism class of $G_1$ in $\mathcal{X}$ determines the isomorphism class of $G$ in $\mathfrak{U}$.

Let $G$ be a group in $\mathcal{X}$. Then $G$ has a normal Sylow $p$-subgroup. Let us denote it by $P$ and let $Q$ be any Sylow $q$-subgroup of $G$. As $Z(G) = 1$, we have $Q$ acting faithfully by conjugation on $P$. Thus $P$ is a $\mathbf{F}_pQ$-module and it has no non-zero trivial submodules.

Now let $\alpha$ and $\beta$ be natural numbers. Let $Q$ be an elementary abelian $q$-group of order $q^\beta$ and let $P$ be a $\mathbf{F}_pQ$-module of dimension $\alpha$. We shall say that the $\mathbf{F}_pQ$-module $P$ is of type (1) if $Q$ acts faithfully on $P$ and $P$ has no non-zero trivial $\mathbf{F}_pQ$-submodule.

Let $f_{\mathcal{X}}(p^\alpha q^\beta)$ denote the number of groups of order $p^\alpha q^\beta$ in $\mathcal{X}$ up to isomorphism. Then $f_{\mathcal{X}}(p^\alpha q^\beta)$ is the number of orbits under the action of $\mathrm{Aut}Q$ on the isomorphism classes of $\mathbf{F}_pQ$-modules of dimension $\alpha$ and type (1). Note that, by Maschke's Theorem the type (1) modules are completely reducible.

It is shown in [11] that each orbit under the action of $\mathrm{Aut}Q$ contains another special type of $\alpha$-dimensional $\mathbf{F}_pQ$-module that are specifically constructed using certain irreducible $\mathbf{F}_pQ$-modules which have certain chosen subgroups of $Q$ as kernels of the action of $Q$ on these modules. If we call these special representatives as type (2) modules, then $f_{\mathcal{X}}(p^\alpha q^\beta)$ will be bounded above by the number of type (2) modules up to isomorphism.

Using the above it can be shown that $f_{\mathfrak{U}}(n) \leq n^{c_{p,q}\mu(n)+1}$. Here

$$c_{p,q} = \frac{1}{d} - 2\sqrt{\left(\frac{\log p}{\log q}\right)^2 + \frac{\log p}{d \log q}} + 2\frac{\log p}{\log q},$$

and $d$ is the order of $p$ modulo $q$. Further for every $\epsilon > 0$ there exist infinitely many $n$ such that $f_{\mathfrak{U}}(n) > n^{(c_{p,q}-\epsilon)\mu(n)}$. So $c_{p,q}$ is best possible.

Let $G$ be a finite group in $\mathfrak{V} = \mathfrak{A}_p\mathfrak{A}_q \vee \mathfrak{A}_q\mathfrak{A}_p$ then $G \cong X \times Y$ where $X \in \mathfrak{A}_p\mathfrak{A}_q$ and $Y \in \mathfrak{A}_q\mathfrak{A}_p$. Using this, we can show that $f_{\mathfrak{V}}(n) \leq n^{d\mu(n)+2}$ where $d = \max\{c_{p,q}, c_{q,p}\}$. Further for every $\epsilon > 0$ it can be shown that there exist infinitely many $n$ such that $f_{\mathfrak{V}}(n) > n^{(d-\epsilon)\mu(n)}$. So $d$ is best possible.

## 4. Structure of groups in $\mathfrak{A}_p\mathfrak{A}_q\mathfrak{A}_r$

As seen in the earlier discussions we still do not have 'best' bounds for $A$-groups or even solvable $A$-groups. While enumerating in small varieties of $A$-groups did gives a class of $A$-groups for which 'best' bounds exist, these still do not help us bridge the gap between the upper and lower bounds for $f_{A,\,\mathrm{sol}}(n)$. The problem with the small variety of $A$-groups considered above is that they did not provide a large enough collection of $A$-groups, up to isomorphism of a given order, to be able to build a good lower bound.

To decrease the gap beween 7 and $3-2\sqrt{2}$ we could enumerate in another larger variety of $A$-groups, namely, $\mathfrak{T} = \vee_{\sigma\in S_3}\mathfrak{A}_{\sigma(p)}\mathfrak{A}_{\sigma(q)}\mathfrak{A}_{\sigma(r)}$ where $p, q, r$ are distinct primes and $S_3$ represents the permutation group on 3 letters. A first step towards analysing the join variety $\mathfrak{T}$ would be to understand the structure of finite groups in $\mathfrak{W} = \mathfrak{A}_p\mathfrak{A}_q\mathfrak{A}_r$. The next would be to analyse the structure of groups in $\mathfrak{T}$. In this section we present some of the unpublished work done in [12] on the structure of groups in $\mathfrak{W}$. The first lemma shows us that finite groups in $\mathfrak{W}$ are solvable $A$-groups.

LEMMA 4.1 *Let $p$, $q$ and $r$ be distinct primes and let $G$ be a finite group in $\mathfrak{W} = \mathfrak{A}_p\mathfrak{A}_q\mathfrak{A}_r$. Then*

(*i*) $|G| = p^\alpha q^\beta r^\gamma$ *for some* $\alpha, \beta, \gamma$ *in* $\mathbf{N}$*;*
(*ii*) $G = PQR$ *where* $P$ *is the* (*unique*) *normal Sylow* $p$*-subgroup of* $G$*;* $Q$ *and* $R$ *are some Sylow* $q$ *and Sylow* $r$*-subgroups respectively. Further* $P$*,* $Q$ *and* $R$ *are elementary abelian groups and* $QR \in \mathfrak{A}_q\mathfrak{A}_r$*.*
(*iii*) $PQ$ *is a normal subgroup of* $G$*.*
(*iv*) $G$ *is solvable.*

**Proof** Product of varieties is associative and so $G \in \mathfrak{A}_p(\mathfrak{A}_q\mathfrak{A}_r)$. Therefore by the Schur-Zassenhaus Theorem, there exists a subgroup $H$ in $G$ such that $G = P \rtimes H$ where $P$ is the normal Sylow $p$-subgroup of $G$. A similar argument shows that $H = Q \rtimes R$ where $Q$ is the normal Sylow $q$-subgroup of $H$ and $R$ a Sylow $r$-subgroup of $H$. Thus $G = PQR$ and $P$, $Q$ and $R$ are elementary abelian $p$, $q$ and $r$ Sylow subgroups of $G$ respectively. Therefore $G$ has the required order. In order to show that $PQ$ is normal in $G$, we note that $R$ normalises both $P$ and $Q$ and hence it normalises $PQ$. For the last part we note that $P$ is abelian and hence solvable. Further $G/P$ is in the metabelian variety $\mathfrak{A}_q\mathfrak{A}_r$ and so is solvable. Therefore $G$ is solvable.

THEOREM 4.2 *Let* $G$ *be a finite group in* $\mathfrak{A}_p\mathfrak{A}_q\mathfrak{A}_r$ *where* $p$*,* $q$ *and* $r$ *are distinct primes. Then there exists a* $G_0 \in \mathfrak{A}_p\mathfrak{A}_q\mathfrak{A}_r$ *such that* $Z(G_0)$*, the centre of* $G_0$*, is identity and* $G = G_0 \times Z(G)$*. Further the choice for* $G_0$ *is unique up to isomorphism.*

**Proof** By Lemma 4.1 we know that $G = PQR$. We show that there is a subgroup $G_0$, as required, by essentially showing that each Sylow subgroup of $G$ decomposes as a direct product with the corresponding Sylow subgroup of $Z(G)$ being a part.

Now $P$ can be regarded as a $\mathbf{F}_pQR$-module, where $p$ does not divide the order of $QR$. So by Maschke's Theorem, $P$ is completely reducible. Let $P_1$ be the Sylow $p$-subgroup of $Z(G)$. Then $P_1$ is a $\mathbf{F}_pQR$-submodule of $P$. Since $P$ is completely reducible there exists a normal subgroup $P_0$ of $G$ such that $P = P_0 \times P_1$. Further it is obvious that $P_0QR \cap P_1 = \{1\}$. Since $P_1$ is a subgroup of $Z(G)$ every element of $P_1$ commutes with every element of $P_0QR$. Thus $G = P_0QR \times P_1 = \hat{G} \times P_1$ where $\hat{G} = P_0QR$.

Now $Q$ is a completely reducible $\mathbf{F}_qR$- module with a submodule $Q_1$ where $Q_1$ is the Sylow $q$-subgroup of $Z(G)$. Thus we can write $Q = Q_0 \times Q_1$ such that $R$ normalises $Q_0$. By a similar argument as above $\hat{G} = P_0Q_0R \times Q_1 = \bar{G} \times Q_1$ where $\bar{G} = P_0Q_0R$.

For the final step we note that since $R$ is elementary abelian we can write $R = R_0 \times R_1$ where $R_1$ is the Sylow $r$-subgroup of $Z(G)$. Since $P_0Q_0$ is a normal subgroup of $\bar{G}$, we get that $\bar{G} = P_0Q_0R_0 \times R_1 = G_0 \times R_1$ where $G_0 = P_0Q_0R_0$. Consequently $G = G_0 \times P_1 \times Q_1 \times R_1 = G_0 \times Z(G)$. Further since $G_0$ is isomorphic to the quotient group, $G/Z(G)$, the choice of $G_0$ is unique up to isomorphism.

We have the following corollaries to the above theorem. The proof follows obviously from Theorem 4.2.

COROLLARY 4.3 *Let* $p$*,* $q$ *and* $r$ *be distinct primes. Let*

$$\mathfrak{X} = \{X \in \mathfrak{A}_p\mathfrak{A}_q\mathfrak{A}_r \mid X \text{ is finite and } Z(X) = 1\}$$

*and let* $G$ *be a finite group in* $\mathfrak{A}_p\mathfrak{A}_q\mathfrak{A}_r$*. Then there exists* $X \in \mathfrak{X}$ *and an abelian group* $Z$ *such that* $G = X \times Z$*. Further if* $G = X_1 \times Z_1$ *for some* $X_1 \in \mathfrak{X}$ *and some abelian group* $Z_1$*, then* $X \cong X_1$ *and* $Z \cong Z_1$*.*

COROLLARY 4.4 *Let $p$, $q$ and $r$ be distinct primes. Let $\mathfrak{V} = \mathfrak{A}_p\mathfrak{A}_q\mathfrak{A}_r$. Then*

$$f_{\mathfrak{V}}(p^\alpha q^\beta r^\gamma) = \sum f_{\mathfrak{X}}(p^{\alpha_1} q^{\beta_1} r^{\gamma_1})$$

*where the sum is over ordered triples of natural numbers $(\alpha_1, \beta_1, \gamma_1)$ such that $\alpha_1 \leq \alpha$, $\beta_1 \leq \beta$ and $\gamma_1 \leq \gamma$.*

We can see from Corollary 4.4 that we now need to investigate the enumeration of groups in $\mathfrak{X}$ of a given order, up to isomorphism. Let $G$ be a group in $\mathfrak{X}$. Then by Lemma 4.1 we know that $G = PQR$ where $P$ is the normal Sylow $p$-subgroup of $G$. Further $P$ can be regarded as a $\mathbf{F}_pQR$-module. Since $G$ has a trivial centre, as a $\mathbf{F}_pQR$-module $P$, has no non-zero trivial submodule. We end this section with a theorem that explains the module structure of $P$ further.

THEOREM 4.5 *Let $G$ be a group in $\mathfrak{X}$ with $G = PQR$ where $P$, $Q$ and $R$ satisfy the conditions of Lemma 4.1. Let $H = QR$. Then*

(i) *$P = P_1 \oplus P_2$, where $P_1$ and $P_2$ are $\mathbf{F}_pH$-submodules of $P$ satisfying*

$$P_1 = \{x \in P \mid h(x) = x \text{ for all } h \in Q\} .$$

*Further $P_1$ has no non-zero trivial $R$-submodule and $P_2$ does not have any non-zero trivial $Q$-submodule;*

(ii) *If $P'$ is a $\mathbf{F}_pH$-module such that $P' = {P_1}' \oplus {P_2}'$ where*

$${P_1}' = \left\{x \in P' \mid h(x) = x \text{ for all } h \in Q\right\}$$

*then $P \cong P'$ as $\mathbf{F}_pH$-modules if and only if $P_1 \cong {P_1}'$ and $P_2 \cong {P_2}'$ as $\mathbf{F}_pH$-modules.*

**Proof** Let $P_1 = C_G(Q) \cap P$. Then $P_1 = O_p(Z(PQ))$. Since $PQ$ is a normal subgroup of $G$, we get that $P_1$ is a normal subgroup of $G$. Thus $P_1$ is a $\mathbf{F}_pH$-submodule of $P$ and it is obvious that as a $\mathbf{F}_pH$-module, $P_1 = \{x \in P \mid h(x) = x \text{ for all } h \in Q\}$.

Further, since $p$ does not divide $|H|$, by Maschke's Theorem we have that $P$ is completely reducible. Thus there exists a $\mathbf{F}_pH$-submodule $P_2$ of $P$ such that $P = P_1 \oplus P_2$. Since we know that $P$ has no non-zero trivial $\mathbf{F}_pH$-submodule. Therefore $P_1$ cannot have a non-zero trivial $R$-submodule and $P_2$ cannot have any non-zero trivial $Q$-submodule.

Let $P \cong P'$ as $\mathbf{F}_pH$-modules via the mapping $\phi$. Then $\phi(P_1) = {P_1}'$. From this we get $P_2 \cong \phi(P_2) \cong P'/{P_1}' \cong {P_2}'$ as required. The converse is obvious.

From Corollary 4.4 and the above Theorem, to find a 'good' bound for groups in $\mathfrak{A}_p\mathfrak{A}_q\mathfrak{A}_r$ we need to count in $\mathfrak{X}$. This in turn depends on counting $\mathbf{F}_pH$-modules $P_1$ and $P_2$ up to isomorphism. Since these are completely reducible, and $H \in \mathfrak{A}_q\mathfrak{A}_r$, we need to investigate the irreducible $\mathbf{F}_pH$-modules. A start towards this process was made in [13]. However, much more needs to be done before we are able to get the bounds of the 'best' kind.

We end with a few open questions which were posed in Chapter 22 of [1], which are relevant to the discussions in this article. We reproduce them here.

*Question 22.21* Is it the case that $f_A(n)/f_{A,\,\mathrm{sol}}(n) \to 1$ as $\lambda(n) \to \infty$? How big is $f_A(n) - f_{A,\,\mathrm{sol}}(n)$ compared with $f_A(n)$?

*Question 22.22* Define $\alpha = \limsup_{n\to\infty} \frac{\log f_A(n)}{\mu(n)\log n}$. What is $\alpha$? Could it perhaps be $3 - 2\sqrt{2}$?

*Question 22.23* For which varieties $\mathfrak{V}$ of $A$-groups is it true that the leading term of the enumeration function $f_{\mathfrak{V}}(n)$ is equal to the leading term of $f_{\mathfrak{U}}(n)$ for some minimal non-abelian subvariety $\mathfrak{U}$ of $\mathfrak{V}$?

## References


[1] Simon R Blackburn, Peter M. Neumann & Geetha Venkataraman, *Enumeration of finite Groups*, Cambridge Tracts in Mathematics, 173, Cambridge University Press, UK, 2007.

[2] L. Pyber, *Enumerating finite groups of a given order*, Annals Of Mathematics, 137: 203-220, 1993.

[3] Geetha Venkataraman, *Enumeration of finite solvable groups with abelian Sylow subgroups*, Quart J. Math. Oxford (2), 48 (1997), 107-125.

[4] Hanna Neumann, *Varieties of groups*, Springer-Verlag, 1967.

[5] Hans Ulrich Besche, Bettina Eick & E. A. O' Brien, *A millennium project: constructing small groups*, International Journal of Algebra and Computation, Vol. 12, No. 5 (2002) 623-644.

[6] Otto Hölder, *Die Gruppen mit quadratfreier Ordnungszahl*, Nachr. Gesellsch.Wiss. zu Göttingen. Math.-phys. Klasse, 1895, pp. 211-229.

[7] Graham Higman, Enumerating $p$-groups. I: Inequalities,Proc. London Math. Soc. (3) 10 (1960) 24-30.

[8] Charles C.Sims, Enumerating $p$-groups, Proc. London Math. Soc.(3)15(1965) 151-166.

[9] Annabelle McIver and Peter M. Neumann,Enumerating finite groups, Quart. J. Math. Oxford (2) 38 (1987) 473-488.

[10] Gabrielle A. Dickenson, On the enumeration of certain classes of solvable groups, Quart. J. Math. Oxford (2) 20 (1969) 383-394.

[11] Geetha Venkataraman, Enumeration of Types of Finite Groups, DPhil Thesis, Oxford, 1993.

[12] GeethaVenkataraman, Enumeration of finite solvable groups in small varieties of A groups and associated topics, Tech. Report, Centre for Mathematical Sciences, St. Stephen's College, University of Delhi, 1999; https://aud-in.academia.edu/GeethaVenkataraman/Research-Report.

[13] Geetha Venkataraman, *On irreducibility of induced modules and an adaptation of the Mackey-Wigner method of little groups*, J. Korean Math. Soc. 50 (2013), No. 6, pp. 1213-1222.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 46–51

# *Local Global Principle for Quadratic Forms*

**V. Suresh***

**Abstract:** In this article we explain a result on local global principle for quadratic forms over function fields of curves over complete discrete valued fields and an application to the computation of a field invariant associated to quadratic forms.

**Keywords:** quadratic forms, local global principal, $u$-invariant

## 1. Introduction

Let $K$ be a field of characteristic not 2. By a *quadratic form* over $K$ we mean a homogeneous polynomial of degree 2 with coefficients in $F$, i.e $\sum_{1\leq i\leq j\leq n} a_{ij}X_iX_j$ with $a_{ij} \in F$. We say that $q$ is *isotropic* over $K$ if there exist $x_i \in K$, $1 \leq i \leq n$, not all zero, such that $q(x_1, \cdots, x_n) = \sum_{1\leq i\leq j\leq n} a_{ij}x_ix_j = 0$. If $q$ is not isotropic, then we say that $q$ is *anisotropic*, i.e $q$ is anisotropic if $q(x_1, \cdots, x_n) = 0$ implies $x_i = 0$ for all $i$. One of the central problem is to decide when a given $q$ quadratic form over a field $F$ is isotropic (or anisotropic). Given a quadratic form $q(X_1, \cdots, X_n) = \sum_{1\leq i\leq j\leq n} a_{ij}X_iX_j$ with $a_{ij} \in F$, we associate a $n \times n$ symmetric matrix $B_q = (a_{ij}) \in M_n(K)$, where for $i > j$, $a_{ij} = a_{ji}$. We say that $q$ is *non singular* if $\det(B_q)) \neq 0$ and *singular* if $\det(B_q) = 0$. The number of variables in $q$ is called the *dimension* of $q$ and $\det(B_q)$ is called the *determinant* of $q$.

If $q$ is singular, then it is easy to see that $q$ is isotropic. Hence, from now on, we assume that all quadratic forms are non singular. We say that two quadratic forms $q$ and $q'$ in $n$ variable are *isometric* over $K$ if there exists a non singular matrix $P = (p_{ij}) \in M_n(K)$ such that $q(X_1, \cdots, X_n) = q'(Y_1, \cdots, Y_n)$ with $Y_j = \sum_i p_{ij}X_i$. If $q$ and $q'$ are isometric, then we denote by $q \cong q'$.

Let $q_1$ and $q_2$ be two quadratic forms over $K$ of dimension $n$ and $m$ respectively. The *orthogonal sum* of $q_1$ and $q_2$ is defined as $(q_1 \perp q_2)(X_1, \cdots, X_{n+m}) = q_1(X_1, \cdots X_n) + q_2(X_{n+1}, \cdots, X_{n+m})$.

Let $q$ be a quadratic form. Since $\mathrm{char}(K) \neq 2$, by a change of variables, we can assume that $q(X_1, \cdots, X_n) = a_1X_1^2 + \cdots + a_nX_1^2$ for some $a_i \in K^*$ ([7]). We denote this diagonal form by $< a_1, \cdots, a_n >$. We have $\det(B_q) = a_1 \cdots a_n$ ([7]).

Let $q(X_1, \cdots, X_n) = a_1X_1^2 + \cdots + a_nX_n^2$ be a quadratic form over $K$ ( with $a_1 \cdots a_n \neq 0$). Suppose $n = 1$. Since $a_1 \neq 0$, $q(x_1) = 0$ if and only if $x_1 = 0$. Hence if $n = 1$, then $q$ is anisotropic. Suppose $n \geq 2$. Suppose $K$ is algebraically closed, e.g. the field of complex numbers. Then every element in $K^*$ is a square and hence there exists $x_1 \in K^*$ such that $x_1^2 = \sqrt{\frac{-a_2}{a_1}} \in K^*$. Since $q(x_1, 1, 0, \cdots, 0) =$

* Email: suresh.venapally@emory.edu

$a_1x_1^2 + a_2 = a_1(-\frac{a_2}{a_1}) + a_2 = 0$, $q$ is isotropic over $K$.

Let $\mathbb{R}$ be the field of real numbers and $q(X_1, \cdots, X_n) = a_1X_1^2 + \cdots + a_nX_n^2$ be a quadratic form over $\mathbb{R}$. Since the square of a non zero real number is positive and every positive real number is a square of a real number, it follows that $q$ is isotropic over $\mathbb{R}$ if and only if at least one $a_i$ is positive and at least one $a_j$ is negative ([7]). Let $\mathbb{Q}$ be the field of rationals and $q(X_1, \cdots, X_n) = a_1X_1^2 + \cdots + a_nX_n^2$ be a quadratic form over $\mathbb{Q}$. Since $\mathbb{Q} \subset \mathbb{R}$, it follows that if $q$ is isotropic, then at least one $a_i$ is positive and at least one $a_j$ is negative. Since not every positive rational number is a square of a rational number, the converse is not true. For example, $q(X_1, X_2) = X_1^2 - 2X_2^2$ is anisotropic over $\mathbb{Q}$. Let $q(X_1, X_2, X_3) = X_1^2 - 2X_2^2 - 3X_3^2$. Is $q$ isotropic over $\mathbb{Q}$? Suppose $q$ is isotropic. Then by the definition of isotropic quadratic forms, there exist $x_1, x_2, x_3 \in \mathbb{Q}$ such that at least one $x_i \neq 0$ and $x_1^2 - 2x_2^2 - 3x_3^2 = 0$. Since every rational number is ratio of two integers, write $x_i = \frac{y_i}{z_i}$ for some $y_i, z_i \in \mathbb{Z}$ and $z_i \neq 0$. Since $(\frac{y_1}{z_i})^2 - 2(\frac{y_2}{z_2})^2 - 3(\frac{y_3}{z_3})^2 = 0$, multiplying by $z_1^2z_2^2z_3^2$, we get that $c_1^2 - 2c_2^2 - 3c_3^2 = 0$ for some integers $c_i$ and at least one $c_i$ is non zero. By going modulo 3 and using the fact that 2 is not a square modulo 3, one can show that $q$ is anisotropic over $\mathbb{Q}$. Similarly, by going modulo 5, one can show that the quadratic form $q(X_1, X_2, X_3, X_4) = X_1^2 - 3X_2^2 - 5X_3^2 + 15X_4^2$ is anisotropic.

Let $q(X_1, \cdots, X_n) = a_1X_1^2 + \cdots + a_nX_n^2$ be a quadratic form with $a_i$ integers. It is easy to see that if there is no non trivial solution to $q(X_1, \cdots, X_n) = a_1X_1^2 + \cdots + a_nX_n^2 = 0$ modulo some $m \in \mathbb{N}$, then $q$ is anisotropic over $\mathbb{Q}$. This leads to a very natural question. If there is a non trivial solution to $q(X_1, \cdots, X_n) = a_1X_1^2 + \cdots + a_nX_n^2 = 0$ modulo every $m \geq 2$, is $q$ isotropic over $\mathbb{Q}$? It is known that $q(X_1, X_2, X_3, X_4, X_5) = X_1^2 + X_2^2 + X_3^2 + X_4^2 + X_5^2 = 0$ has non trivial solution modulo every $m \geq 2$ ([7, Theorem 4.2, p.217]). Hence for $q$ to be isotropic over $\mathbb{Q}$, it is not enough that $q(X_1, \cdots, X_n) = 0$ has a non trivial solution modulo every $m \geq 2$. Suppose there is a non trivial solution to $q(X_1, \cdots, X_n) = a_1X_1^2 + \cdots + a_nX_n^2 = 0$ modulo every $m \geq 2$ and $q$ is isotropic over $\mathbb{R}$. Is $q$ isotropic over $\mathbb{Q}$? This question is answered affirmatively by Hasse and Minkowskii ([7, Theorem 7.2, p.192]). Let $K$ be a number field (i.e. a finite extension of $\mathbb{Q}$). Let $\Omega_K$ be the set of discrete valuations of $K$ and all Archimedean valuations of $K$. For each $\nu \in K$, let $K_\nu$ be the completion of $K$ at $\nu$ (cf. §2). A theorem of Hasse and Minkowskii ([7, Theorem 6.5, p.223]) asserts that a quadratic form over $K$ is isotropic if and only if $q$ is isotropic over $K_\nu$ for all $\nu \in \Omega_F$. A natural question is whether there are any other fields which have this property. In this direction we have the following theorem

THEOREM 1.1 *([1]). Let $K$ be a complete discretely valued field with residue field $\kappa$ and $F$ the function field of a curve over $K$. Suppose char($\kappa$) $\neq 2$. Let $q$ be a quadratic form over $F$ of dimension at leas 3. If $q$ is isotropic over $F_\nu$ for all discrete valuations of $F$, then $q$ is isotropic.*

## 2. Quadratic forms over complete discretely valued fields

In this section we recall the definition of discretely valued fields and structure of quadratic forms over such fields.

Let $K$ be a field and $K^* = K \setminus \{0\}$. A *discrete valuation* on $F$ is a surjective homomorphism $\nu : F^* \to \mathbb{Z}$ such that $\nu(a + b) \geq \min\{\nu(a), \nu(b)\}$ for all $a, b \in F^*$ with $a + b \neq 0$. A field with a discrete valuation is called a discretely valued field.

Let $K$ be a discretely valued field and $\nu$ a discrete valuation on $K$. Let $R =$

$\{x \in K^* \mid \nu(x) \geq 0\} \cup \{0\}$. Then using the properties of $\nu$, it can be checked that $R$ is a sub ring of $K$. The sub ring $R$ is called the *valuation ring* of $K$. Let $\wp = \{x \in K^* \mid \nu(x) > 0\} \cup \{0\}$. Then $\wp$ is a maximal ideal of $R$. The field $R/\wp$ is called the *residue field* at $\nu$. Since $\nu$ is surjective, there exists $\pi \in R$ such that $\nu(\pi) = 1$. Any such $\pi \in R$ is called a *parameter*. In fact $\wp$ is the unique maximal ideal of $R$ with $\wp = R\pi$ for any parameter $\pi$. An element $u \in R$ is a unit if and only if $\nu(u) = 0$. Any element of $K^*$ can be written as $u\pi^m$ for some $u \in R$ a unit and $m \in \mathbb{Z}$.

Let $K$ be a discretely valued field and $\nu$ a discrete valuation on $K$. Let $d : K \times K \to \mathbb{R}$ be the function given by $d(x, y) = (\frac{1}{2})^{\nu(x-y)}$ if $x \neq y$ and $d(x, x) = 0$. Then, it can be shown that $d$ is a metric on $K$. Since $K$ is a metric space, we have the completion $\hat{K}$ of $K$ with respect to the metric $d$. The addition and multiplication on $K$ extends to $\hat{K}$, making $\hat{K}$ into a field. The valuation $\nu$ also extends to a valuation on $\hat{K}$. We say that $K$ is a *complete discretely valued field* if $K \simeq \hat{K}$. Note that the number 2 in the definition of $d$ is irrelevant and one can take any real number bigger than 1. We denote the completion of $K$ at $\nu$ by $K_\nu$ and the residue field at $\nu$ by $\kappa(\nu)$.

Let $p \in Z$ be a prime and $a \in \mathbb{Z}$, $a \neq 0$. Then $a = p^m b$ for some $m \geq 0$ and $b \in \mathbb{Z}$ which is not divisible by $p$. Let $\nu_p(a) = m$. For any $x \in \mathbb{Q}$, $x \neq 0$, write $x = \frac{a}{b}$ and define $\nu_p(x) = \nu_p(a) - \nu_p(b)$. Then it is easy that $\nu_p$ is a discrete valuation on $\mathbb{Q}$. The completion of $\mathbb{Q}$ at $p$ is denoted by $\mathbb{Q}_p$ and called the field of $p$-adic numbers. Any discrete valuation on $\mathbb{Q}$ is equal to $\nu_p$ for some prime $p$. Let $K$ be any finite extension of $\mathbb{Q}$. Let $\nu$ be a discrete valuation on $K$. Then the restriction of $K$ to $\mathbb{Q}$ is a discrete valuation on $\mathbb{Q}$ and hence is equal to $\nu_p$ for some prime $p$. Further the completion of $K$ at $\nu$ is a finite extension of $\mathbb{Q}_p$. A *place* of $K$ is either a discrete valuation of $K$ or an Archimedean valuation of $K$. If $\nu$ is an Archimedean place of $K$, then the completion $K_\nu$ of $K$ at $\nu$ is either $\mathbb{R}$ or $\mathbb{C}$.

Let $k$ be any field and $K = k(x)$ be the field of fractions of the polynomial ring $k[X]$. Let $p(x) \in k[x]$ be an irreducible polynomial. Then, as in the case of $\mathbb{Q}$, we have a discrete valuation $\nu_{p(x)}$ on $K$. Let $K = k[X]/(p(x))$. Then $K$ is field and the completion of $k(x)$ at $\nu_{p(x)}$ is isomorphic to $K((t))$. Since $\frac{1}{x}$ is irreducible in the ring $k[\frac{1}{x}]$ and the field of fractions of $k[\frac{1}{x}]$ is $k(x)$, we have a discrete valuation $\nu_\infty$ on $k(x)$. Any discrete valuation on $k(x)$ which is trivial on $k$ is equal to $\nu_{p(x)}$ for some irreducible polynomial $p(x)$ or $\nu_\infty$.

Let $K$ be a complete discretely valued field with discrete valuation $\nu$. Let $\kappa$ be the residue field at $\nu$. Suppose that $\mathrm{char}(\kappa) \neq 2$. Then $\mathrm{char}(K) \neq 2$. Let $R$ be the valuation ring of $K$. Let $q =< a_1, \cdots, a_n >$ be a (non singular ) quadratic form over $K$. Since each $a_i \in K^*$, we have $a_i = u_i \pi^{r_i}$ for some $u_i \in R$ a unit and $r_i \in \mathbb{Z}$. Since, $< ab^2 > \simeq < a >$ for all $a, b \in K^*$, with out loss of generality, we assume that $r_i = 0$ or 1. Hence, by reindexing $a_i$, we have $< a_1, \cdots, a_n > \simeq < u_1, \cdots, u_r > \perp \pi < u_{r+1}, \cdots, u_n >$ . For a unit $u \in R$, let $\overline{u} \in \kappa = R/m$ be the image of $u$. Since $K$ is complete, by a theorem of Springer ([7, Corollary 2.6, p. 209]), $q$ is isotropic over $K$ if and only if either $< \overline{u_1}, \cdots, \overline{u}_r >$ or $< \overline{u}_{r+1}, \cdots, \overline{u}_n >$ is isotropic over $\kappa$. Hence, if we know when a quadratic form is isotropic over $\kappa$, then we know when a quadratic form over $K$ is isotropic. Thus complete discretely valued fields play a very important role in the study of quadratic forms.

## 3. Function fields of curves

Let $K$ be a field and $\Omega_K$ the set of discrete valuations of $K$. Let $\nu \in \Omega_K$ and $K_\nu$ the completion of $K$ at $\nu$. As we have seen in the previous section, the study of quadratic forms over $K_\nu$ is well understood in terms quadratic forms over $\kappa(\nu)$. Thus it is natural to ask whether one can derive some properties of quadratic forms over $K$ using the structure of quadratic forms over $K_\nu$ for all $\nu$. We say that the *local global principle* for quadratic forms over $K$ holds if a quadratic form $q$ over $K$ is isotropic is and only if $q$ is isotropic over $K_\nu$ for all $\nu \in \Omega_K$. Let $K$ be a totally imaginary number field (i.e. a finite extension of $\mathbb{Q}$ which is not isomorphic to a subfield of $\mathbb{R}$). Then the classical Hasse-Minkowskii theorem asserts that local global principle for quadratic forms holds for $K$. In this section we give a class of fields for which the local global principle for quadratic forms holds.

Let $k$ be a field and $F$ a finite extension of $k(t)$ (such a field is called the function field of a curve over $k$). Let $\Omega_F$ be the set of discrete valuations of $F$. Let $\nu \in \Omega$ and $F_\nu$ the completion of $F$ at $\nu$. Our main theorem is the following

THEOREM 3.1 *([1]). Let $K$ be a complete discretely valued field with residue field $\kappa$ and $F$ the function field of a curve over $K$. Suppose char$(\kappa) \neq 2$. Let $q$ be a quadratic form over $F$ of dimension at least 3. If $q$ is isotropic over $F_\nu$ for all discrete valuations of $F$, then $q$ is isotropic.*

To give an idea of the proof of our main theorem, we need to recall the patching techniques developed by Harbater, Hartmann and Krashen ([2]).

Let $K$ be a complete discretely valued field with valuation ring $R$ and residue field $\kappa$. Let $F$ be a function field of a curve over $K$. Then there exists a regular integral two dimensional scheme $\mathscr{X}$ over $R$ with the function field $F$ ([5], [6] ). Let $X$ be the special fibre of $\mathscr{X}$. For each point $x$ of $X$, let $F_x$ be the field of fractions of the completion of the local ring at $x$ on $\mathscr{X}$. We have the following

THEOREM 3.2 *([3]). Let $K$, $\kappa$ and $F$ be as above. Let $q$ be a quadratic form over $F$. Then $q$ is isotropic over $F$ if and only if $q$ is isotropic over $F_x$ for all $x \in X$.*

Note that if $x$ is not a closed point of $X$, then $F_x$ is the completion of $F$ at a discrete valuation of $F$. On the other hand if $x$ is a closed point of $X$, then $F_x$ is not a complete discretely valued field. In particular we do not know much about the structure of quadratic forms over $F_x$. Thus we are still interested in the local global principle for quadratic forms. We now give an outline of the proof of our main theorem.

*Proof.* Let $q =< a_1, \cdots, a_n >$ be a quadratic form. Then, first we choose a regular proper model $\mathscr{X}$ of $F$ such that the special fire $X$ of $\mathscr{X}$ and the union of the support of $\mathrm{div}_{\mathscr{X}}(a_i)$ for all $i$ is a union of regular curves with normal crossings. Suppose that $q$ is isotropic over $F_\nu$ for all $\nu \in \Omega_F$. Let $x \in X$. Suppose that $x$ is not a closed point of $X$. Then $F_x \simeq F_\nu$ for some $\nu \in \Omega_F$. Since $q$ is isotropic over $F_\nu$, $q$ is isotropic over $F_x$.

Suppose $x$ is a closed point of $X$. Let $A_x$ be the local ring at $x$. Then $A_x$ is a two dimensional regular local ring. By the choice of on $\mathscr{X}$, the maximal ideal of $A_x$ is $(\pi, \delta)$ for some primes $\pi, \delta \in A_x$ and $a_i = u_i \pi^{r_i} \delta^{s_i}$ for some $u_i \in A_x$ a units and $r_i, s_i \in \mathbb{Z}$. Then, as in the case of complete discretely valued field (cf. §2), after reindexing $a_i$, we have $q =< u_1, \cdots, u_{n_1} > \perp \pi < u_{n_1+1}, \cdots, u_{n_2} > \perp \delta < u_{n_2+1}, \cdots, u_{n_3} > \perp \pi\delta < u_{n_3+1}, \cdots, u_{n_4} >$. Since $u_i \in A_x$ are units, $\hat{A}_x$ is a complete ring and char$(\kappa) \neq 2$, using Hensel's Lemma, one can show that $q$ is isotropic

over $F_x$ if and only if one of the forms $< u_1, \cdots, u_{n_1} >$, $< u_{n_1+1}, \cdots, u_{n_2} >$, $< u_{n_2+1}, \cdots, u_{n_3} >$, $< u_{n_3+1}, \cdots, u_{n_4} >$ is isotropic over $F_x$.

Let $\hat{A}_x$ be the completion of $A_x$ at its maximal ideal. Then $F_x$ is the field of fractions of $\hat{A}_x$. Since $\pi \in A$ is a prime, $\pi$ gives a discrete valuation $\nu_\pi$ on $F_x$ and its restriction to $F$ is also a discrete valuation on $F$. Since $F \subset F_x$, $F_\nu \subset F_{x,\nu_\pi}$. Since $q$ is a isotropic over $F_\nu$, $q$ is isotropic over $F_{x,\nu}$. Since $F_{x,\nu_\pi}$ is a complete discretely valued field with $\pi$ as a parameter and $q$ is isotropic over $F_{x,\nu}$, it follows that either $< u_{n_1}, \cdots, u_{n_2} > \perp \delta < u_{n_2+1}, \cdots, u_{n_3} >$ or $< u_{n_1+1}, \cdots, u_{n_2} > \perp \delta < u_{n_3+1}, \cdots, u_{n_4} >$ is isotropic. The residue field $\kappa(\pi)$ of the discrete valuation $\nu_\pi$ on $F_{x,\nu}$ is the field of fractions of $\hat{A}_x/(\pi)$. Since $(\pi, \delta)$ is the maximal ideal of $\hat{A}_x$, $(\overline{\delta})$ is the maximal ideal of $\hat{A}_x/(\pi)$ and hence $\kappa(\pi)$ is a complete discretely valued field with $\overline{\delta}$ as a parameter. Hence, using the fact that either $< u_{n_1}, \cdots, u_{n_2} > \perp \delta < u_{n_2+1}, \cdots, u_{n_3} >$ or $< u_{n_1+1}, \cdots, u_{n_2} > \perp \delta < u_{n_3+1}, \cdots, u_{n_4} >$ is isotropic, one concludes that one of the forms $< u_1, \cdots, u_{n_1} >$, $< u_{n_1+1}, \cdots, u_{n_2} >$, $< u_{n_2+1}, \cdots, u_{n_3} >$, $< u_{n_3+1}, \cdots, u_{n_4} >$ is isotropic over $F_x$. Thus, by (3.2), $q$ is isotropic over $F$. ■

Let $K$ be a number field (i,e a finite extension of ) and $F/K(t)$ a finite extension. It is known that the local global principle for quadratic forms over $F$ does not hold. We end this section with the following

CONJECTURE 3.3 *Let $K$ be a totally imaginary number field (e.g. $\mathbb{Q}(\sqrt{-1})$ ) and $F/K(t)$ a finite extension. Let $q$ be a quadratic form over $F$ of dimension at least 6. If $q$ is isotropic over $F_\nu$ for all $\nu \in \Omega_F$, then $q$ is isotropic over $F$.*

## 4. *u*-invariant of fields

Let $K$ be a field with $\mathrm{char}(K) \neq 2$. The *u-invariant* of $K$, denoted by $u(K)$, is the supremum of dimensions of anisotropic quadratic forms over $K$. For example if $K$ is algebraically closed (e.g. the field of complex numbers), then $u(K) = 1$. If $K$ is a finite field, then $u(K) = 2$. If $K$ is a finite extension of $\mathbb{Q}_p$, then $u(K) = 4$. More generally, let $K$ be a complete discretely valued field with residue field $\kappa$. Suppose $\mathrm{char}(\kappa) \neq 2$. Then $u(K) = 2u(\kappa)$. By the Hasse Minkowskii result it follows that for any totally imaginary field $K$, $u(K) = 4$. As a consequence of our main theorem we have the following

THEOREM 4.1 *([2], [1]) Let $K$ be a complete discretely valued field with residue field $\kappa$. Suppose char$(\kappa) \neq 2$. Suppose there exists an integer $n$ such that for every finite extension $k$ of $\kappa(t)$, $u(k) \leq n$. Let $F$ be the function field of a curve over $K$. Then $u(F) \leq 2n$.*

Let $K$ be totally imaginary number field. Then, as we mentioned above, $u(K) = 4$. However it is not known whether $u(K(t))$ is finite. We have the following

THEOREM 4.2 *([4]) Let $K$ be a totally imaginary number field and $F/K(t)$ a finite extension. If a conjecture of Colliot-Thélne on the existence of zero-cycles of degree 1 holds, then $u(F) < \infty$.*

Since we do not know the validity of conjecture of Colliot-Thélne on the existence of zero-cycles of degree 1, we do not know whether $u(F)$ is finite or not. On the other hand if the conjecture (3.3) in §3 holds, then it follows that $u(F) = 8$.

We end with the following

CONJECTURE 4.3 *Let $K$ be a totally imaginary number field and $F/K(t)$ a finite extension. Then $u(F) = 8$.*

The author thanks S. Ilangovan for carefully reading the manuscript and suggesting improvements.

## References

[1] J.-L. Colliot-Thélène, R. Parimala, V. Suresh, *Patching and local global principles for homogeneous spaces over function fields of p-adic curves*, Commentari Math. Helv. **87** (2012), 1011–1033.

[2] D. Harbater, J. Hartmann and D. Krashen, *Applications of patching to quadratic forms and central simple algebras*, Invent. Math. **178** (2009), 231–263.

[3] D. Harbater, J. Hartmann and D. Krashen, *Local-global principles for torsors over arithmetic curves*, American Journal of Mathematics, **137** (2015), 1559–1612

[4] Max Lieblich, R. Parimala and V. Suresh, *Colliot-Thélène's conjecture and finiteness of u-invariants*, Math. Ann. **360** (2014), 1-22.

[5] Lipman, J., *Introduction to resolution of singularities, Proc. Symp. Pure Math.* **29** (1975), 187-230.

[6] Lipman, J., *Desingularization of two-dimensional schemes, Ann. Math.* **107** (1978), 151-207.

[7] Scharlau, W., *Quadratic and Hermitian Forms*, Grundlehren der Math. Wiss., Vol. 270, Berlin, Heidelberg, New York 1985.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 52–64

# *On the Non-Vanishing of the Fourier Coefficients of Primitive Forms*

**Tarun Dalal**[a] and **Narasimha Kumar**[b] *

[a] *Department of Mathematics, Indian Institute of Technology Hyderabad, Kandi, Sangareddy 502285, INDIA; ma17resch11005@iith.ac.in*;
[b] *Department of Mathematics, Indian Institute of Technology Hyderabad, Kandi, Sangareddy 502285, INDIA; narasimha@math.iith.ac.in*

**Abstract:** In this semi-expository article, we discuss about the non-vanishing of the Fourier coefficients of primitive forms. We shall make a note of a discrepancy in the statement of [5, Lemma 2.2].

## 1. Introduction

In 1947, Lehmer conjectured that Ramanujan's tau function $\tau(n)$ is non-vanishing for all $n$. In [6], he proved that the smallest $n$ for which $\tau(n) = 0$ must be a prime and showed that $\tau(n) \neq 0$ for all $n < 33,16,799$. It is well-known that the Fourier coefficients of Ramanujan's Delta function $\Delta(z)$ are in fact $\tau(n)(n \in \mathbb{N})$. Note that $\Delta(z)$ is a cuspidal Hecke eigenform of weight 12 and level 1. It is a natural question to ask if a similar phenomenon continue to hold for cusp forms of higher weight and higher level.

In this semi-expository article, we study the non-vanishing of the Fourier coefficients of primitive forms of any weight and any level. We take this opportunity to make a correction in the statement of [5, Lemma 2.2].

## 2. Preliminary

In this section, we shall define modular forms and recall some basic facts about them. For more details, we refer the reader to consult [3], [7].

### 2.1. *Congruence subgroups*

The modular group $\mathrm{SL}_2(\mathbb{Z})$ is defined by

$$\mathrm{SL}_2(\mathbb{Z}) := \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) : a, b, c, d \in \mathbb{Z}, ad - bc = 1 \right\}.$$

*Corresponding author. Email: narasimha@math.iith.ac.in

For any $N \in \mathbb{N}$, we shall define a subgroup of $\mathrm{SL}_2(\mathbb{Z})$ by

$$\Gamma(N) = \{\gamma \in \mathrm{SL}_2(\mathbb{Z}) \mid \gamma \equiv \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \pmod{N}\}.$$

*Definition* 2.1 We say that a subgroup $\Gamma$ of $\mathrm{SL}_2(\mathbb{Z})$ is a congruence subgroup, if $\Gamma$ contains $\Gamma(N)$ for some $N \in \mathbb{N}$.

In this theory, the following congruence subgroups play an important role

$$\Gamma_1(N) = \{\gamma \in \mathrm{SL}_2(\mathbb{Z}) \mid \gamma \equiv \left(\begin{smallmatrix} 1 & * \\ 0 & 1 \end{smallmatrix}\right) \pmod{N}\},$$
$$\Gamma_0(N) = \{\gamma \in \mathrm{SL}_2(\mathbb{Z}) \mid \gamma \equiv \left(\begin{smallmatrix} * & * \\ 0 & * \end{smallmatrix}\right) \pmod{N}\} \quad \textit{for any } N \in \mathbb{N}.$$

The subgroup $\Gamma(N)$ is called the principal congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$. Note that $\Gamma(N) \leq \Gamma_1(N) \leq \Gamma_0(N) \leq \mathrm{SL}_2(\mathbb{Z})$, and $\Gamma(1) = \Gamma_1(1) = \Gamma_0(1) = \mathrm{SL}_2(\mathbb{Z})$.

The modular group $\mathrm{SL}_2(\mathbb{Z})$ acts on the complex upper half plane $\mathfrak{H} = \{\tau \in \mathbb{C} \mid \mathrm{Im}(\tau) > 0\}$ via

$$\gamma\tau = \frac{a\tau + b}{c\tau + d},$$

where $\tau \in \mathfrak{H}$, $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z})$. For more details, please refer to [3, §1.2].

## 2.2. *Modular forms*

In this section, we shall define modular forms and recall some results related to them.

Let $X$ be the space of all complex valued holomorphic functions on $\mathfrak{H}$. We can define an action of $\mathrm{SL}_2(\mathbb{Z})$ on $X$ by using the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathfrak{H}$ as follows. For any $k \in \mathbb{N}$, $f \in X$ and $\gamma \in \mathrm{SL}_2(\mathbb{Z})$, we define the slash operator

$$(f|_k\gamma)(\tau) := (c\tau + d)^{-k} f(\gamma\tau),\ \tau \in \mathfrak{H},$$

where $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Now, we define the notion of modular forms for any congruence subgroup $\Gamma$ of $\mathrm{SL}_2(\mathbb{Z})$.

*Definition* 2.2 Let $\Gamma$ be a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$. A function $f \in X$ is said to be a **modular form** of weight $k$ with respect to $\Gamma$ if

(1) $f|_k\gamma = f, \forall\gamma \in \Gamma$,
(2) $f|_k\alpha$ is holomorphic at $\infty$, $\forall\alpha \in \mathrm{SL}_2(\mathbb{Z})$.

*Remark* 2.3 Note that one needs to verify condition (2) only for the representatives of distinct cosets of $\Gamma$ in $\mathrm{SL}_2(\mathbb{Z})$.

Now, we explain the meaning of $f$ being holomorphic at $\infty$. From condition (1), it is clear that then $f$ will be $h\mathbb{Z}$-periodic, where $h$ is the smallest integer such that $\left(\begin{smallmatrix} 1 & h \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma$ (such $h$ exists since $\Gamma(N) \leq \Gamma$). This implies that there exists a function $g : D' \longrightarrow \mathbb{C}$, where $D'$ is unit puncture disk, such that $f(\tau) = g(q_h)$ for all $\tau \in \mathfrak{H}$, where $q_h = e^{\frac{2\pi i\tau}{h}}$. It is clear that, the function $g$ is holomorphic on $D'$, since $f$ is so on $\mathfrak{H}$. The function $f$ is said to be **holomorphic at** $\infty$ if $g$ extends holomorphically to $q = 0$. Similarly, one can define the meaning of $f|_k\alpha$ being holomorphic at $\infty$. For more details, please refer to [3, §1.1, §1.2].

We denote the space of all modular forms of weight $k$ and level $\Gamma$ by $M_k(\Gamma)$.

### 2.3. *Fourier expansion*

Let $f \in M_k(\Gamma)$. Let $h$ be the smallest integer such that $\left(\begin{smallmatrix} 1 & h \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma$. Since $f$ is holomorphic at $\infty$, then $f$ has a Fourier expansion

$$f(\tau) = \sum_{n=0}^{\infty} a_f(n) q_h^n, \quad \text{where} \quad q_h = e^{\frac{2\pi i \tau}{h}}$$

for $\tau \in \mathfrak{H}$.

*Definition* 2.4 Let $f \in M_k(\Gamma)$. We say that $f$ is a **cusp form** if $a_{f|_k\alpha}(0) = 0$ for all $\alpha \in \mathrm{SL}_2(\mathbb{Z})$. We denote the space of all cusp forms of weight $k$ and level $\Gamma$ by $S_k(\Gamma)$.

Note that $M_k(\Gamma), S_k(\Gamma)$ are vector spaces over $\mathbb{C}$. By [3, Theorem 3.5.1 and Theorem 3.6.1], these are in fact finite dimensional vector spaces over $\mathbb{C}$. Now, we shall give some examples of modular forms and cusp forms.

*Example* 2.5 For any $k \geq 2$, we define the Eisenstein series of weight $2k$

$$G_{2k}(\tau) = \sum_{(c,d) \in \mathbb{Z}^2 - \{(0,0)\}} \frac{1}{(c\tau + d)^{2k}} \in M_{2k}(\mathrm{SL}_2(\mathbb{Z})).$$

It is easy to check that $G_{2k}$ is a modular form of weight $2k$ and level 1 (cf. [3, Page 4]). The Fourier expansion of $G_{2k}$ at $\infty$ is given by

$$G_{2k}(\tau) = 2\zeta(2k) + 2\frac{(2\pi i)^{2k}}{(2k-1)!} \sum_{n=1}^{\infty} \sigma_{2k-1}(n) q^n, \quad k \geq 1, \tag{2.1}$$

where $\sigma_{2k-1}(n) = \sum_{m|n, m>0} m^{2k-1}$. The normalized Eisenstein series is defined by $E_{2k}(\tau) := \frac{G_{2k}(\tau)}{2\zeta(2k)}$. Therefore, the Fourier expansion of $E_{2k}$ at $\infty$ is given by

$$E_{2k}(\tau) = 1 - \frac{4k}{B_{2k}} \sum_{n=1}^{\infty} \sigma_{2k-1}(n) q^n,$$

where $B_k$'s are the Bernoulli numbers (cf. [3, Page 10]).

*Example* 2.6 From the dimensions of $S_k(\mathrm{SL}_2(\mathbb{Z}))$, one can see that 12 is the least integer for which there is a non-zero cusp form for $\mathrm{SL}_2(\mathbb{Z})$. Moreover, dimension of $S_{12}(\mathrm{SL}_2(\mathbb{Z}))$ is 1 and it is spanned by

$$\Delta(z) = (60G_4(z))^3 - 27(140G_6(z))^2 \in S_{12}(\mathrm{SL}_2(\mathbb{Z})), \quad z \in \mathfrak{H}.$$

The product formula for $\Delta(z)$ is given by $\Delta(z) = q\prod_{n\geq 1}(1-q^n)^{24} = \sum_{n\geq 1} \tau(n) q^n$, where $q = e^{2\pi i z}$.

*Example* 2.7 ([8], Example 2.28) For $N \in \{2, 3, 5, 11\}$, $(\Delta(z)/\Delta(Nz))^{1/(N+1)} \in S_{24/(N+1)}(\Gamma_0(N))$. Moreover, the space $S_{24/(N+1)}(\Gamma_0(N))$ is one dimensional and it is spanned by $(\Delta(z)/\Delta(Nz))^{1/(N+1)}$.

## 2.4. *Modular forms with character*

A Dirichlet character modulo $N$ is a group homomorphism $\chi : (\mathbb{Z}/N\mathbb{Z})^* \longrightarrow \mathbb{C}^*$.

*Definition* 2.8 The space of all modular forms of weight $k$ level $N$ with character $\chi$ is defined by

$$M_k(N,\chi) = \{f \in M_k(\Gamma_1(N)) | f|_k \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \chi(d)f, \forall \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(N)\}.$$

The space $M_k(\Gamma_1(N))$ decomposes as

$$M_k(\Gamma_1(N)) = \bigoplus_{\chi} M_k(N,\chi),$$

where $\chi$ varies over all Dirichlet characters of $(\mathbb{Z}/N\mathbb{Z})^*$ such that $\chi(-1) = (-1)^k$ (cf. [7, Lemma 4.3.1]). Similarly one can define the space of cusp forms of weight $k$ level $N$ with character $\chi$ and they are denoted by $S_k(N,\chi)$. One can easily check that $S_k(N,\chi) = S_k(\Gamma_1(N)) \cap M_k(N,\chi)$. Moreover, a similar decomposition holds as well, i.e.,

$$S_k(\Gamma_1(N)) = \bigoplus_{\chi} S_k(N,\chi),$$

where $\chi$ varies over all Dirichlet characters of $(\mathbb{Z}/N\mathbb{Z})^*$ with $\chi(-1) = (-1)^k$ (cf. [7, Lemma 4.3.1]).

*Example* 2.9 (Poincaré series) Let $\Gamma_\infty = \{\left(\begin{smallmatrix} 1 & b \\ 0 & 1 \end{smallmatrix}\right) \mid b \in \mathbb{Z}\}$, and $\chi$ be any Dirichlet character modulo $N$. For $m \geq 1$, we define

$$P_m(z) := \sum_{\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_\infty \backslash \Gamma_0(N)} \overline{\chi}(\gamma) \frac{1}{(cz+d)^k} \exp(2\pi i m \gamma z).$$

for any integer $k \geq 2$. By [4, Proposition 14.1], $P_m(z) \in S_k(N,\chi)$.

Now we will define two types of operators on the space of modular forms (resp., cusp forms). They are known as Hecke operators.

## 2.5. *Hecke operators*

Let $M_k(\Gamma_1(N))$ be a space of modular forms of weight $k$, level $N$. For any $(n,N) = 1$, we define the **diamond operator**

$$\langle n \rangle : M_k(\Gamma_1(N)) \longrightarrow M_k(\Gamma_1(N))$$

as

$$\langle n \rangle f := f|_k \alpha, \ \textit{for any } \alpha = \left(\begin{smallmatrix} a & b \\ c & \delta \end{smallmatrix}\right) \in \Gamma_0(N) \text{ with } \delta \equiv n \pmod{N}.$$

We can also extend the definition of diamond operator to $\mathbb{N}$ via $\langle n \rangle = 0$ if $(n,N) > 1$. Observe that for any character $\chi : (\mathbb{Z}/N\mathbb{Z})^* \longrightarrow \mathbb{C}^*$,

$$M_k(N,\chi) = \{f \in M_k(\Gamma_1(N)) | \langle n \rangle f = \chi(n) f, \forall n \in (\mathbb{Z}/N\mathbb{Z})^*\}.$$

Note that, the diamond operator acts trivially on $M_k(\Gamma_0(N))$, since $M_k(\Gamma_0(N)) = M_k(N, \chi_N^\circ)$, where $\chi_N^\circ$ is the trivial character modulo $N$.

Now, we will define the second type of **Hecke operator** for any prime $p$, and they are denoted by $T_p$. If $f(\tau) = \sum_{n=0}^{\infty} a_f(n)q^n \in M_k(\Gamma_1(N))$, then

$$(T_p f)(\tau) = \sum_{n=0}^{\infty} a_f(np)q^n + \chi_N^\circ(p)p^{k-1}\sum_{n=0}^{\infty} a_{\langle p\rangle f}(n)q^{np} \in M_k(\Gamma_1(N)).$$

Similarly, one can also defined the action of $T_p$ on $M_k(N, \chi)$ as follows: If $f(\tau) = \sum_{n=0}^{\infty} a_f(n)q^n \in M_k(N, \chi)$, then

$$(T_p f)(\tau) = \sum_{n=0}^{\infty} a_f(np)q^n + \chi(p)p^{k-1}\sum_{n=0}^{\infty} a_f(n)q^{np} \in M_k(N, \chi),$$

In fact, for $n \in \mathbb{N}$, one can define the Hecke operators $T_n$ as follows:

(1) For any prime $p$ and $r \geq 2$ we define $T_{p^r} = T_p T_{p^{r-1}} - p^{k-1}\chi(p)T_{p^{r-2}}$.
(2) For $n = p_1^{e_1} \dots p_k^{e_k}$ we define $T_n = T_{p_1^{e_1}} \dots T_{p_k^{e_k}}$.

One can check that, any two primes $p \neq q$, $T_pT_q = T_qT_p$. In fact, the Hecke operators respects the spaces $S_k(N, \chi)$ and $S_k(\Gamma_0(N))$. For more details, we refer the reader to [3, §5.3].

## 2.6. *Petersson inner product*

To study the space of cusp forms $S_k(\Gamma_1(N))$ further, we make it into an inner product space. In order to do so, we need to define an inner product on the space of cusp forms.

The **hyperbolic measure** on the upper half plane is defined by

$$d\mu(\tau) := \frac{dxdy}{y^2},\ \tau = x + iy \in \mathfrak{H}.$$

For any congruence subgroup $\Gamma \leq \mathrm{SL}_2(\mathbb{Z})$, the **Petersson inner product**

$$\langle , \rangle_\Gamma : S_k(\Gamma) \times S_k(\Gamma) \longrightarrow \mathbb{C}$$

is given by

$$\langle f, g\rangle_\Gamma = \frac{1}{V_\Gamma}\int_{\Gamma\backslash\mathfrak{H}} f(\tau)\overline{g(\tau)}(\mathrm{Im}(\tau))^k d\mu(\tau),\ \text{where } V_\Gamma = \int_{\Gamma\backslash\mathfrak{H}} d\mu(\tau).$$

This inner product is linear in $f$, conjugate linear in $g$, Hermitian symmetric and positive definite. By [3, Theorem 5.5.3], the Hecke operators $\langle n\rangle$ and $T_n$ are normal operators for $(n, N) = 1$. By [3, Theorem 5.5.4], we have that

THEOREM 2.10 *The space $S_k(\Gamma_1(N))$ has an orthogonal basis of simultaneous eigenforms for the Hecke operators $\{\langle n\rangle, T_n : (n, N) = 1\}$.*

Now, we shall introduce the theory of old forms and new forms. This in fact leads to define the notion of primitive forms. (cf. [3, §5.4] for more discussion on this).

### 2.7. *Old forms and New forms*

For $d|N$, we define the mapping

$$i_d : (S_k(\Gamma_1(Nd^{-1})))^2 \longrightarrow S_k(\Gamma_1(N)) \text{ by}$$

$$(f, g) \longrightarrow f + g|_k \left(\begin{smallmatrix} d & 0 \\ 0 & 1 \end{smallmatrix}\right).$$

The space of **old forms** is defined by

$$S_k(\Gamma_1(N))^{\text{old}} = \sum_{p|N} i_p((S_k(\Gamma_1(Np^{-1})))^2).$$

The space of **new forms** (denote by $S_k(\Gamma_1(N))^{\text{new}}$) is defined to be the orthogonal complement of $S_k(\Gamma_1(N))^{\text{old}}$ with respect to the Petersson inner product. By [3, Proposition 5.6.2], we see that the spaces $S_k(\Gamma_1(N))^{\text{old}}$ and $S_k(\Gamma_1(N))^{\text{new}}$ are stable under the action of $T_n$ and $\langle n \rangle$ for all $n \in \mathbb{N}$.

*Definition* 2.11 A primitive form is a normalized eigenform in $f \in S_k(\Gamma_1(N))^{\text{new}}$, i.e., $f$ is an eigenform for the Hecke operators $T_n, \langle n \rangle$ for all $n \in \mathbb{N}$, and $a_f(1) = 1$.

By [3, Theorem 5.8.2], the set of primitive forms in the space $S_k(\Gamma_1(N))^{\text{new}}$ forms an orthogonal basis. Each such primitive form lies in an eigen space $S_k(N, \chi)$ for an unique character $\chi$. In fact, its Fourier coefficients are its $T_n$-eigenvalues.

*Note* 2.12 When we say that $f \in S_k(N, \chi)$ is a primitive form of weight $k$, level $N$, with character $\chi$, actually we mean $f \in S_k(\Gamma_1(N))^{\text{new}}$ is a primitive form and it belongs the eigenspace $S_k(N, \chi)$.

Proposition 2.13 ([3], Proposition 5.8.5) *Let $f = \sum_{n=1}^{\infty} a_f(n) q^n \in S_k(N, \chi)$. Then $f$ is a normalized eigenform if and only if its Fourier coefficients satisfy the following relations*

*(1)* $a_f(1) = 1$,
*(2)* $a_f(m)a_f(n) = a_f(m)a_f(n)$ *if* $(m, n) = 1$,
*(3)* $a_f(p^r) = a_f(p)a_f(p^{r-1}) - p^{k-1}\chi(p)a_f(p^{r-2})$, *for all prime $p$ and $r \geq 2$.*

For more details on this content, please refer to [3, §5.7, §5.8].

## 3. Classical modular forms

Recall that, Lehmer proved that the smallest $n$ for which $\tau(n) = 0$ must be a prime. We are interested in studying a similar question for the Fourier coefficients of primitive forms of higher weight and higher level. Let $f = \sum_{n=1}^{\infty} a_f(n) q^n \in S_k(N, \chi)$ be a primitive form of even weight $k$, level $N$, with character $\chi$.

Suppose that $a_f(n) = 0$ for some $n = \prod_i p_i^{r_i} \geq 1$. By Proposition 2.13, we see that $a_f(p_i^r) = 0$ for some prime $p_i$. In this section, we shall explore the relation between the vanishing (resp., non-vanishing) of $a_f(p)$ and $a_f(p^r)$ for $r \geq 2$. We begin this discussion with a lemma of Kowalski, Robert, and Wu (see [5, Lemma 2.2]).

PROPOSITION 3.1 *Let $f = \sum_{n=1}^{\infty} a_f(n)q^n \in S_k(N, \chi)$ be a primitive form of even weight $k$, level $N$, with character $\chi$. There exists an integer $M_f \geq 1$, such that for any prime $p \nmid M_f$, either $a_f(p) = 0$ or $a_f(p^r) \neq 0$ for all $r \geq 1$.*

*Proof.* If $p \mid N$ then $a_f(p^r) = a_f(p)^r$ for any $r \geq 1$, so in this case the conclusion holds trivially. Let $p$ be a prime number such that $p \nmid N$. If $a_f(p) = 0$, then there is nothing prove. Suppose that $a_f(p) \neq 0$ but $a_f(p^r) = 0$ for some $r \geq 2$. Since $f$ is a primitive form, then by Hecke relations, we have

$$a_f(p^{m+1}) = a_f(p)a_f(p^m) - \chi(p)p^{k-1}a_f(p^{m-1})$$

for any $m \in \mathbb{N}$. These relations can be re-interpreted as

$$\sum_{r=0}^{\infty} a_f(p^r)X^r = \frac{1}{1 - a_f(p)X + \chi(p)p^{k-1}X^2}. \tag{3.1}$$

Suppose that

$$1 - a_f(p)X + \chi(p)p^{k-1}X^2 = (1 - \alpha(p)X)(1 - \beta(p)X). \tag{3.2}$$

By comparing the coefficients, we get that

$$\alpha(p) + \beta(p) = a_f(p) \qquad \text{and} \qquad \alpha(p)\beta(p) = \chi(p)p^{k-1} \neq 0,$$

since $p \nmid N$ and hence $\chi(p) \neq 0$. If $\alpha(p) = \beta(p)$, then

$$a_f(p^t) = (t + 1)\alpha(p)^t \neq 0, \tag{3.3}$$

for any $t \geq 2$ and this cannot happen. Therefore, $\alpha(p) \neq \beta(p)$. Then, by induction, we have the following

$$a_f(p^t) = \frac{\alpha(p)^{t+1} - \beta(p)^{t+1}}{\alpha(p) - \beta(p)}.$$

for any $t \geq 2$. Recall that $a_f(p^r) = 0$ for some $r \geq 2$. Therefore,

$$a_f(p^r) = 0 \text{ if and only if } \left(\frac{\alpha(p)}{\beta(p)}\right)^{r+1} = 1, \tag{3.4}$$

which implies that the ratio $\frac{\alpha(p)}{\beta(p)}$ is a $(r+1)$-th root of unity. Since $a_f(p) \neq 0$, we get that $\alpha(p) = \zeta\beta(p)$ where $\zeta$ is a root of unity and $\zeta \neq -1$ . By the product relation, we get that $\alpha(p)^2 = \zeta\chi(p)p^{k-1}$, hence $\alpha(p) = \pm\gamma p^{(k-1)/2}$, where $\gamma^2 = \zeta\chi(p)$. Therefore,

$$a_f(p) = (1 + \zeta^{-1})\alpha(p) = \pm\gamma(1 + \zeta^{-1})p^{(k-1)/2} \neq 0.$$

In particular, $\gamma(1+\zeta^{-1})p^{(k-1)/2} \in \mathbb{Q}(f)$, where $\mathbb{Q}(f)$ is the number field generated by the Fourier coefficients of $f$ and by the values of $\chi$. Since $k$ is even, we have

$$\gamma(1 + \zeta^{-1})\sqrt{p} \in \mathbb{Q}(f). \tag{3.5}$$

We have that the number of such primes $p$ are finite, since $\mathbb{Q}(f)$ is a number field. Take $M_f$ to be the product of all such primes $p$. Thus, for any prime $p \nmid M_f$, we have either $a_f(p) = 0$ or $a_f(p^r) \neq 0$ for all $r \geq 1$. ■

COROLLARY 3.2 *Let $f, M_f$ be as in the above Proposition. Then the smallest $m \in \mathbb{N}$ with $(m, M_f) = 1$ with $a_f(m) = 0$ is a prime.*

If $M_f = 1$, then the corollary is exactly the generalization of Lehmer's result that that the smallest $n$ for which $\tau(n) = 0$ must be a prime. Now, this leads to the question of calculating $M_f$ for $f$. In the second part of [5, Lemma 2.2], it was stated as follows:

PROPOSITION 3.3 *Let $f, M_f$ be as in Proposition 3.1. If the character $f$ is trivial and the Fourier coefficients of $f$ are integers, then one can take $M_f = N$.*

However, we are able to produce examples which contradicts this statement.

*Example* 3.4 Let $E$ be an elliptic curve defined by the minimal Weierstrass equation $y^2 + y = x^3 - x$. The Cremona label for $E$ is $37a1$. Let $f_E$ denote the primitive form (of weight 2 and level 37) associated to $E$ by the modularity theorem. The Fourier expansion of $f_E$ is given by

$$f_E(q) = \sum_{n=1}^{\infty} a_{f_E}(n)q^n = q - 2q^2 - 3q^3 + 2q^4 - 2q^5 + 6q^6 - q^7 + 6q^9 + O(q^{10}).$$

Note that $(2, 37) = 1$ and $a_{f_E}(2)$ is non-zero but $a_{f_E}(8) = 0$.

*Example* 3.5 Let $E$ be an elliptic curve defined by the minimal Weierstrass equation $y^2 + xy + y = x^3 - x^2$. The Cremona label for $E$ is $53a1$. Let $f_E$ denote the primitive form (of weight 2 and level 53) associated to $E$ by the modularity theorem. The Fourier expansion of $f_E$ is given by

$$f_E(q) = \sum_{n=1}^{\infty} a_{f_E}(n)q^n = q - q^2 - 3q^3 - q^4 + 3q^6 - 4q^7 + 3q^8 + 6q^9 + O(q^{10}).$$

Note that $(3, 53) = 1$ and $a_{f_E}(3)$ is non-zero but a simple calculation using the relations among the Fourier coefficients shows that $a_{f_E}(3^5) = 0$.

For the convenience of the reader, we shall recall their proof of Proposition 3.3.

*Proof.* Let $p$ be a prime number such that $p \nmid N$. If $a_f(p) = 0$, then there is nothing prove. Suppose $a_f(p) \neq 0$ but $a_f(p^r) = 0$ for some $r \geq 2$. Arguing as in Proposition 3.1, the argument is valid till (3.5). After that, they wished to show that (3.5) does not hold for any prime $p \nmid N$.

By (3.2), (3.4), we get that $\frac{\alpha(p)}{\beta(p)} = \zeta$ is a root of unity in a quadratic extension of $\mathbb{Q}$, hence $\zeta \in \{-1, \pm i, \pm\omega_3, \pm\omega_3^2\}$. All those except $\zeta = -1$ contradict the fact that $f$ has integer coefficients by simple considerations such as the following, for $\zeta = \omega_3$ say: we have $\alpha(p)^2 = \omega_3 p^{k-1}$, $\gamma = \pm\omega_3^2 p^{\frac{k-1}{2}}$ and $\lambda_f(p) = (1 + \omega_3^{-1})\gamma = \pm(1 + \omega_3^{-1})\omega_3^2 p^{\frac{k-1}{2}} = \pm(\omega_3^2 + \omega_3)p^{\frac{k-1}{2}} \notin \mathbb{Z}$. Therefore, (3.5) does not hold for any prime $p \nmid N$. ■

In the last part of the above proof, when we calculated the expression in (3.5) for $\zeta \neq \pm 1$, it seem to hold for $p = 2$ (resp., $p = 3$) with some special values of

$a_f(2)$ (resp., $a_f(3)$). In the next proposition, we have calculated the optimal value of $M_f$ and the correct version of Proposition 3.3 is

PROPOSITION 3.6 *Let $f, M_f$ be as in Proposition 3.1. If the character $\chi$ is trivial and the Fourier coefficients of $f$ are integers, then $M_f$ can be so chosen that $(M_f, N) = 1$ and $M_f \mid 6$.*

*Proof.* If $p \mid N$ then $a_f(p^r) = a_f(p)^r$ for any $r \geq 1$, so in this case the conclusion of Proposition 3.1 holds trivially. Hence, the number $M_f$ is relatively prime to $N$.

If $p \nmid N$, we argue as in the proof of Proposition 3.3 till the last step. Now, we compute (3.5) for all values of $\zeta$ to prove our proposition. Let $\omega_n$ denote $e^{\frac{2\pi i}{n}}$ for any $n \in \mathbb{N}$.

(1) The root of unity $\zeta$ cannot be 1 because of (3.3).
(2) The root of unity $\zeta$ cannot be $-1$ because $0 \neq a_f(p) = \alpha(p) + \beta(p)$.
(3) If $\zeta = \omega_3$, then $\alpha(p)^2 = \omega_3 p^{k-1} \Rightarrow \alpha(p) = \pm\omega_3^2 p^{\frac{k-1}{2}}$. This implies that $a_f(p) = \pm(1+\omega_3^2)\omega_3^2 p^{\frac{k-1}{2}} = \mp p^{\frac{k-1}{2}} \notin \mathbb{Z}$. For $\zeta = \omega_3^2$, we will get the same conclusion.
(4) If $\zeta = i$, then $\alpha(p)^2 = ip^{k-1} \Rightarrow \alpha(p) = \pm\omega_8 p^{\frac{k-1}{2}} \Rightarrow a_f(p) = \pm(1-i)\omega_8 p^{\frac{k-1}{2}} = \pm\sqrt{2}p^{\frac{k-1}{2}}$. This implies that

$$\sqrt{2}p^{\frac{k-1}{2}} \in \mathbb{Z} \iff p = 2,$$

in which case $a_f(2) = \pm 2^{k/2}$. For $\zeta = -i$, we will get the same conclusion.
(5) If $\zeta = -\omega_3$, then $\alpha(p)^2 = -\omega_3 p^{k-1} \Rightarrow \alpha(p) = \pm\frac{\sqrt{3}-i}{2} p^{\frac{k-1}{2}} \Rightarrow a_f(p) = \pm(1 + \frac{1+i\sqrt{3}}{2})\frac{\sqrt{3}-i}{2} p^{\frac{k-1}{2}} = \pm\sqrt{3}p^{\frac{k-1}{2}}$. This implies that

$$\sqrt{3}p^{\frac{k-1}{2}} \in \mathbb{Z} \iff p = 3,$$

in which case $a_f(3) = \pm 3^{k/2}$. If $\zeta = -\omega_3^2$, then we will get same conclusion.

This case by case analysis would imply that $M_f$ is a divisor of 6. This means that the possible values of $M_f$ are $1, 2, 3, 6$. ■

For any prime $p$, $\chi_p^\circ$ denote the trivial character on $(\mathbb{Z}/p\mathbb{Z})^*$, i.e., for any $N \in \mathbb{N}$, we have

$$\chi_p^\circ(N) := \begin{cases} 0 & \text{if} \quad p \mid N, \\ 1 & \text{if} \quad p \nmid N. \end{cases}$$

Based on the proof of the above proposition, we can re-interpret the above result as follows:

LEMMA 3.7 *Let $f, M_f$ be as in Proposition 3.6. Then $M_f$ can be taken to be $2^{\chi_2^\circ(N)} 3^{\chi_3^\circ(N)}$. Further if*

- *$2 \mid M_f$, $a_f(2) \neq \pm 2^{k/2}$, then 2 can be dropped from $M_f$, i.e., $M_f$ can be taken to be $3^{\chi_3^\circ(N)}$,*
- *$3 \mid M_f$, $a_f(3) \neq \pm 3^{k/2}$, then 3 can be dropped from $M_f$, i.e., $M_f$ can be taken to be $2^{\chi_2^\circ(N)}$,*
- *$6 \mid M_f$, $a_f(p) \neq \pm p^{k/2}$ (for $p = 2, 3$), then 6 can be dropped from $M_f$, i.e., $M_f$ can be taken to be 1.*

Note that the above lemma gives an optimal $M_f$ for which Proposition 3.6 continues to hold. The following corollaries describes the nature of the first vanishing of Fourier coefficients of primitive forms of higher weight $k$ and higher level $N$.

COROLLARY 3.8 *Let $f = \sum_{n=1}^{\infty} a_f(n)q^n \in S_k(\Gamma_0(N))$ be a primitive form of even weight $k$ and level $N$ with $a_f(n) \in \mathbb{Z}$. Let $M_f$ be as in Lemma 3.7. Then the smallest $n \in \mathbb{N}$ with $(n, M_f) = 1$ with $a_f(n) = 0$ is prime.*

*Proof.* Let $n$ be the smallest integer with $(n, M_f) = 1$ such that $a_f(n) = 0$. Since $f$ is a primitive form, we know that the Fourier coefficients of $f$ satisfy

$$a_f(n_1 n_2) = a_f(n_1)a_f(n_2) \text{ if } (n_1, n_2) = 1. \tag{3.6}$$

This forces that $n = p^r$, where $p$ is a prime with $(p, M_f) = 1$. By Proposition 3.6, we get that $r = 1$. Therefore $n$ has to be a prime. ∎

The following two corollaries can be thought of as a generalization of the result of Lehmer which states that the smallest $n$ for which $\tau(n) = 0$ must be a prime.

COROLLARY 3.9 *Let $f = \sum_{n=0}^{\infty} a_f(n)q^n \in S_k(\Gamma_0(N))$ be a primitive form of even weight $k$, level $N$ with $a_f(n) \in \mathbb{Z}$. If 6 divides $N$, then the smallest $n$ for which $a_f(n) = 0$ is a prime.*

*Proof.* Since $M_f \mid 6$, and $6 \mid N$, we have that $M_f \mid N$. Since $(M_f, N) = 1$, we have that $M_f = 1$. By Corollary 3.8, the result follows. ∎

In order to get a similar conclusion as above for cusp forms when $6 \nmid N$, e.g., for $\Delta$-function, we need to impose some conditions on $a_f(2), a_f(3)$, which is the content of the following Corollary. It follows from Lemma 3.7 and coincides with [10, Proposition 4.2],

COROLLARY 3.10 *Let $f = \sum_{n=0}^{\infty} a_f(n)q^n \in S_k(\Gamma_0(N))$ be a primitive form of even weight $k$, level $N$ with $a_f(n) \in \mathbb{Z}$. Suppose $a_f(2) \neq \pm 2^{\frac{k}{2}}$ and $a_f(3) \neq \pm 3^{\frac{k}{2}}$. Then the smallest $n$ for which $a_f(n) = 0$ is a prime.*

*Proof.* We know that $M_f \mid 6$ and $(M_f, N) = 1$. By Lemma 3.7, it follows that $M_f$ can be improved to 1. Therefore, the result follows by Corollary 3.8. ∎

## 4. Hilbert modular forms

There is a generalization of Proposition 3.1 available in the context of Hilbert modular forms. In fact, we used this generalization to study the simultaneous non-vanishing of Fourier coefficients of distinct primitive forms at powers of prime ideals (cf. [2]). We shall state that generalization in this section.

Let $K$ be a totally real number field of odd degree $n$ and $\mathbb{P}$ denote the set of all prime ideals of $\mathcal{O}_K$ with odd inertia degree. Let $\mathbf{P}$ denote the set of all prime ideals of $\mathcal{O}_K$.

Let $\mathbf{f}$ be a primitive form over $K$ of level $\mathfrak{c}$, with character $\chi$ and weight $\boldsymbol{2k} = (2k_1, \ldots, 2k_n)$. Let $2k_0$ denote the maximum of $\{2k_1, \ldots, 2k_n\}$. For each integral ideal $\mathfrak{m} \subseteq \mathcal{O}_K$, let $C(\mathfrak{m}, \mathbf{f})$ denote the Fourier coefficients of $\mathbf{f}$ at $\mathfrak{m}$.

Now, we state the result which is analogous to Proposition 3.1 for $\mathbf{f}$.

PROPOSITION 4.1 *Let $\mathbf{f}$ be a primitive form over $K$ of level $\mathfrak{c}$, with character $\chi$ and weight $\mathbf{2k}$. Then there exists an integer $M_{\mathbf{f}} \geq 1$ with $N(\mathfrak{c}) \mid M_{\mathbf{f}}$ such that for any prime $p \nmid M_{\mathbf{f}}$ and for any prime ideal $\mathfrak{p} \in \mathbb{P}$ over $p$, we have either $C(\mathfrak{p}, \mathbf{f}) = 0$ or $C(\mathfrak{p}^r, \mathbf{f}) \neq 0$ for all $r \geq 1$.*

*Proof.* Let $p$ be a prime number such that $p \nmid N(\mathfrak{c})$. Let $\mathfrak{p} \in \mathbb{P}$ be a prime ideal of $\mathcal{O}_K$ over $p$ and $\mathfrak{p} \nmid \mathfrak{c}$. If $C(\mathfrak{p}, \mathbf{f}) = 0$, then there is nothing prove. If $C(\mathfrak{p}, \mathbf{f}) \neq 0$, then we need to show that $C(\mathfrak{p}^r, \mathbf{f}) \neq 0$ for all $r \geq 2$, except for finitely many prime ideals $\mathfrak{p} \in \mathbb{P}$.

Suppose that $C(\mathfrak{p}, \mathbf{f}) \neq 0$ but $C(\mathfrak{p}^r, \mathbf{f}) = 0$ for some $r \geq 2$. Since $\mathbf{f}$ is a primitive form, then by Hecke relations, we have

$$C(\mathfrak{p}^{m+1}, \mathbf{f}) = C(\mathfrak{p}, \mathbf{f})C(\mathfrak{p}^m, \mathbf{f}) - \chi(\mathfrak{p})N(\mathfrak{p})^{2k_0-1}C(\mathfrak{p}^{m-1}, \mathbf{f}).$$

These relations can be re-interpreted as

$$\sum_{r=0}^{\infty} C(\mathfrak{p}^r, \mathbf{f})X^r = \frac{1}{1 - C(\mathfrak{p}, \mathbf{f})X + \chi(\mathfrak{p})N(\mathfrak{p})^{2k_0-1}X^2}. \tag{4.1}$$

Suppose that

$$1 - C(\mathfrak{p}, \mathbf{f})X + \chi(\mathfrak{p})N(\mathfrak{p})^{2k_0-1}X^2 = (1 - \alpha(\mathfrak{p})X)(1 - \beta(\mathfrak{p})X).$$

By comparing the coefficients, we get that

$$\alpha(\mathfrak{p}) + \beta(\mathfrak{p}) = C(\mathfrak{p}, \mathbf{f}) \qquad \text{and} \qquad \alpha(\mathfrak{p})\beta(\mathfrak{p}) = \chi(\mathfrak{p})N(\mathfrak{p})^{2k_0-1} \neq 0,$$

since $\mathfrak{p} \nmid \mathfrak{c}$ and hence $\chi(\mathfrak{p}) \neq 0$. If $\alpha(\mathfrak{p}) = \beta(\mathfrak{p})$, then

$$C(\mathfrak{p}^r, \mathbf{f}) = (r+1)\alpha(\mathfrak{p})^r \neq 0,$$

which cannot happen for any $r \geq 2$. So, $\alpha(\mathfrak{p})$ cannot be equal to $\beta(\mathfrak{p})$. Then by induction, for any $r \geq 2$, we have the following

$$C(\mathfrak{p}^r, \mathbf{f}) = \frac{\alpha(\mathfrak{p})^{r+1} - \beta(\mathfrak{p})^{r+1}}{\alpha(\mathfrak{p}) - \beta(\mathfrak{p})}.$$

In this case, we have

$$C(\mathfrak{p}^r, \mathbf{f}) = 0 \text{ if and only if } \left(\frac{\alpha(\mathfrak{p})}{\beta(\mathfrak{p})}\right)^{r+1} = 1,$$

which implies that the ratio $\frac{\alpha(\mathfrak{p})}{\beta(\mathfrak{p})}$ is a root of unity. Since $C(\mathfrak{p}, \mathbf{f}) \neq 0$, we get that $\alpha(\mathfrak{p}) = \zeta\beta(\mathfrak{p})$ where $\zeta$ is a root of unity and $\zeta \neq -1$ . By the product relation, we get that $\alpha(\mathfrak{p})^2 = \zeta\chi(\mathfrak{p})N(\mathfrak{p})^{2k_0-1}$, hence $\alpha(\mathfrak{p}) = \pm\gamma N(\mathfrak{p})^{(2k_0-1)/2}$, where $\gamma^2 = \zeta\chi(\mathfrak{p})$. Therefore,

$$C(\mathfrak{p}, \mathbf{f}) = (1 + \zeta^{-1})\alpha(\mathfrak{p}) = \pm\gamma(1 + \zeta^{-1})N(\mathfrak{p})^{(2k_0-1)/2} \neq 0.$$

In particular, $\mathbb{Q}(\gamma(1 + \zeta^{-1})N(\mathfrak{p})^{\frac{2k_0-1}{2}}) \subseteq \mathbb{Q}(\mathbf{f})$, where $\mathbb{Q}(\mathbf{f})$ is the field generated by $\{C(\mathfrak{m}, \mathbf{f})\}_{\mathfrak{m} \subseteq \mathcal{O}_K}$ and by the values of the character $\chi$. Since $\mathfrak{p} \in \mathbb{P}$, $N(\mathfrak{p}) = p^f$,

where $f \in \mathbb{N}$ odd. Hence, we have

$$\mathbb{Q}(\gamma(1+\zeta^{-1})p^{\frac{f(2k_0-1)}{2}}) \subseteq \mathbb{Q}(\mathbf{f}). \tag{4.2}$$

Since $2k_0 - 1$, $f$ are odd, we have that

$$\mathbb{Q}(\gamma(1+\zeta^{-1})\sqrt{p}) \subseteq \mathbb{Q}(\mathbf{f}). \tag{4.3}$$

By [9, Proposition 2.8], the field $\mathbb{Q}(\mathbf{f})$ is a number field. Hence, the number of such primes $p$ are finite. Take $M_{\mathbf{f}}$ to be the product of all such primes $p$ and $N(\mathfrak{c})$. Thus, for any prime $p \nmid M_{\mathbf{f}}$ and for any prime ideal $\mathfrak{p} \in \mathbb{P}$ over $p$, we have either $C(\mathfrak{p}, \mathbf{f}) = 0$ or $C(\mathfrak{p}^r, \mathbf{f}) \neq 0$ for all $r \geq 1$. ■

We end this article with the following statement:

LEMMA 4.2 *Let* $\mathbf{f}$ *and* $K$ *be as in Proposition 4.1. Further, if* $K$ *is Galois over* $\mathbb{Q}$*, then there exists an integer* $M_{\mathbf{f}} \geq 1$ *with* $N(\mathfrak{c}) \mid M_{\mathbf{f}}$ *such that for any prime* $p \nmid M_{\mathbf{f}}$ *and for any prime ideal* $\mathfrak{p} \in \mathbf{P}$ *over* $p$*, we have either* $C(\mathfrak{p}, \mathbf{f}) = 0$ *or* $C(\mathfrak{p}^r, \mathbf{f}) \neq 0$ *for all* $r \geq 1$*.*

We note that in a recent work of Bhand, Gun and Rath (cf. [1, Theorem 2]), they have computed the lower bounds of the Weil heights of $C(\mathfrak{p}^r, f)$, when non-zero, for prime ideals $\mathfrak{p}$ away from an ideal $\mathbf{M}$. In particular, the above lemma is a consequence of their Theorem.

## Acknowledgements

The authors are thankful to the anonymous referee for the valuable suggestions towards the improvement of this paper. The first author thanks University Grants Commission (UGC), India for the financial support provided in the form of Research Fellowship to carry out this research work at IIT Hyderabad. The second author's research was partially supported by the SERB grant MTR/2019/000137.

## References

[1] Bhand, Ajit; Gun, Sanoli; Rath, Purusottam; A note on lower bounds of heights of non-zero Fourier coefficients of Hilbert cusp forms. Arch. Math. (Basel) 114 (2020), no. 3, 285–298.

[2] Dalal, Tarun; Kumar, Narasimha. On the non-vanishing and sign changes of the Fourier coefficients of two Hilbert cusp forms. To appear in the Proceedings of the Conference on "Number theory: Arithmetic, Diophantine and Transcendence" at IIT Ropar.

[3] Diamond, Fred; Shurman, Jerry. A first course in modular forms. Graduate Texts in Mathematics, 228. Springer-Verlag, New York, 2005.

[4] Iwaniec, Henryk; Kowalski, Emmanuel. Analytic number theory. American Mathematical Society Colloquium Publications, 53. American Mathematical Society, Providence, RI, 2004.

[5] Kowalski, Emmanuel; Robert, Olivier; Wu, Jie. Small gaps in coefficients of $L$-functions and $B$-free numbers in short intervals. Rev. Mat. Iberoam. 23 (2007), no. 1, 281–326.

[6] Lehmer, D. H. The vanishing of Ramanujan's function $\tau(n)$. Duke Math. J. 14 (1947), 429–433.

[7] Miyake, Toshitsune. Modular forms. Springer-Verlag, Berlin, 2006.

[8] Shimura, Goro. Introduction to the arithmetic theory of automorphic functions. Kanô Memorial Lectures, No. 1. Publications of the Mathematical Society of Japan, No. 11. Iwanami Shoten, Publishers, Tokyo; Princeton University Press, Princeton, N.J., 1971.
[9] Shimura, Goro. The special values of the zeta functions associated with Hilbert modular forms. Duke Math. J. 45 (1978), no. 3, 637-679.
[10] Tian, Peng; Qin, Hourong. Non-vanishing Fourier coefficients of $\Delta_k$. Appl. Math. Comput. 339 (2018), 507–515.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 65–74

# A Partial Survey on Some Characteristic $p$ Invariants

**V. Trivedi***

*School of Mathematics, Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai 400005, India*

**Abstract:** Here we survey some results on characteristic $p$ invariants like Hilbert-Kunz multiplicity, Hilbert-Kunz density function and its relation to $F$-thresholds.

## 1. Introduction

In this expository article we discuss some characteristic $p$ invariants related to *Hilbert-Kunz multiplicity* for commutative Noetherian rings. For an extensive survey on Hilbert-Kunz multiplicity and related invariants one can refer to the survey article by Huneke ([15]). First we begin with some generalities. To be less abstract we can restrict to 'geometric rings'. By a *geometric ring* $R$ we mean $R$ is a ring without nilpotent elements, which is a quotient of a polynomial ring over an algebraically closed field, *i.e.*,

$$R = \frac{k[X_1, \ldots, X_n]}{(f_1, \ldots, f_m)}, \quad \text{where} \quad f_1, \ldots, f_m \in k[X_1, \ldots, X_n] \ \text{ and } \ k = \bar{k}.$$

Such a ring $R$ comes with a variety $X_R$. As a set, the variety $X_R$ is given as

$$\begin{aligned} X_R &= \text{the zero set of } \ \{f_1, \ldots, f_m\} \ \text{ in } \ k^n \\ &= \{(a_1, \ldots, a_n) \in k^n \mid f_i(a_1, \ldots, a_n) = 0, \ \ \forall \ \ 1 \leq i \leq m\}. \end{aligned}$$

The topology on $X_R$ is the Zariski topology on $X_R$. This means the closed sets of $X_R$ are precisely the sets $\{V(I)\}_I$, where $I \subseteq R$ are ideals and

$$V(I) = \{(a_1, \ldots, a_n) \in k^n \mid g(a_1, \ldots, a_n) = 0, \ \ \forall \ \ g \in I\}.$$

In particular a point $x_{\mathbf{m}} = (a_1, \ldots, a_n) \in X_R$ corresponds to the maximal ideal $\mathbf{m}_x = (X_1 - a_1, \ldots, X_n - a_n)$.

Therefore in the Zariski topology, the elements of $X_R$ are closed sets, in fact it is the weakest topology on $X_R$, where the elements are closed sets. Moreover the Hilbert Nullstellensatz allows us to recover $R$ from $X_R$.

To study a property of the ring we attach numerical invariants to the ring, which relate to the property of the ring (or to the variety).

*Corresponding author. Email: vija@math.tifr.res.in

We say a property $P$ of a ring $R$ (or of a variety $X_R$) satisfies the open condition if whenever $P$ holds at $R$ localized at $\mathbf{m}_x$ (at $x_{\mathbf{m}} \in X_R$) then it holds for all points in a Zariski open neighoubourhood of $\mathbf{m}_x$ (in a Zariski open neighbourhood of $x_{\mathbf{m}}$ in $X_R$).

Consider a pair $(R, I)$, where $R$ is a Noetherian ring of dimension $d$ and $I \subset R$ is an ideal of finite colength.

Then *the Hilbert-Samuel function* of $R$, with respect to $I$, is a function given by

$$HS(R, I) : \mathbb{N} \to \mathbb{N}, \quad \text{given by} \quad n \mapsto \ell(R/I^n).$$

It is a polynomial function of degree $d$, *i.e.*, for $n >> 0$,

$$HS(R, I)(n) = e_0(R, I)\binom{n+d-1}{d} - e_1(R, I)\binom{n+d-2}{d-1} + \cdots + (-1)^d e_d(R, I),$$

where

$$e_0(R, I) = \lim_{n \mapsto \infty} \frac{d!}{n^d} HS(R, I)(n)$$

is the Hilbert-Samuel multiplicity of $R$ with respect to $I$ and is a positive integer. For a maximal ideal $\mathbf{m} \subset R$, the integer $e_0(R, \mathbf{m})$ is a numerical invariant of $(R, \mathbf{m})$ which characterizes the singularity of $X_R$ at the point $x_{\mathbf{m}}$ (corresponding to $\mathbf{m}$).

For example

(1) If $(R, \mathbf{m})$ is an integral domain, then

$$e_0(R, \mathbf{m}) = 1 \iff X_R \quad \text{is smooth at the point} \quad x_{\mathbf{m}}.$$

(2) In general, larger the multiplicity $e_0(R, \mathbf{m})$, more singular is the variety $X_R$ at $x_{\mathbf{m}}$. For the illustrations of the following curves and surfaces one can refer to Chapter I, Excercises 5.1 and 5.2 from [13].

(i) Following two are the examples of curves. In both the examples if $x_{\mathbf{m}} \neq (0, 0)$ then $e_0(R, \mathbf{m}_x) = 1$ and the curve $X_R$ is smooth at $x_{\mathbf{m}}$.
   (a) $R = k[x, y]/(xy - x^6 - y^6)$: if $x_{\mathbf{m}} = (0, 0)$ then $e_0(R, \mathbf{m}_x) = 2$ and $X_R$ has a node at $x_{\mathbf{m}}$.
   (b) $R = k[x, y]/(x^2y + xy^2 - x^4 - y^4)$: if $x_{\mathbf{m}} = (0, 0)$ then $e_0(R, \mathbf{m}_x) = 3$ and $X_R$ has a triple point at $x_{\mathbf{m}}$.

(ii) Following two are the examples of surfaces.
   (a) $R = k[x, y, z]/(x^2 + y^2 - z^2)$: if $x_{\mathbf{m}} \neq (0, 0, 0)$ then $e_0(R, \mathbf{m}_x) = 1$ and the surface $X_R$ is smooth at $x_{\mathbf{m}}$. If $x_{\mathbf{m}} = (0, 0, 0)$ then $e_0(R, \mathbf{m}_x) = 2$ and $X_R$ has a conical double point at $x_{\mathbf{m}}$.
   (b) $R = k[x, y, z]/(xy - x^3 - y^3)$: if $x_{\mathbf{m}} \neq (0, 0, \lambda)$, for some $\lambda \in k$ then $e_0(R, \mathbf{m}_x) = 1$ and the surface $X_R$ is smooth at $x_{\mathbf{m}}$. If $x_{\mathbf{m}} = (0, 0, \lambda)$ then $e_0(R, \mathbf{m}_x) = 2$ and $X_R$ is has a double line at $x_{\mathbf{m}} = (0, 0, \lambda)$.

Moreover $e_0(R, \mathbf{m})$ is a well behaved invariant:

(1) It does not change after taking a general hyperplane section e.g., for *a general* choice of an element $h \in \mathbf{m}$, $e_0(R, \mathbf{m}) = e_0(R/(h), \mathbf{m}/(h))$. This allows us to use induction method on the dimension of the ring.

(2) It remains constant in a flat family, and hence we can consider any other member of a flat family where $R$ belongs.
(3) It has a cohomological interpretation, which gives us a powerful machinery to handle $e_0(R, \mathbf{m})$.

## 2. Hilbert-Kunz multiplicity

Henceforth we will discuss rings with positive characteristic (in fact geoemtric rings with positive characteristic). We recall that the characteristic of a ring (denoted as char $R$) is the least positive integer $m$ such that $m \cdot 1_R = 0$. For a geoemtric ring $R$, either the char $R = 0$ or $p > 0$, where $p$ is prime number.

In many situations it is easier to solve a problem by going to *reduction mod p*. One of the first example we encounter of such technique is in the proof of Eisenstein's criteria for checking irreducibility of a polynomial in $\mathbb{Q}[X]$, which involves going to $\mathbb{Z}/p\mathbb{Z}[X]$.

However, in characteristic 0, every ring $R$ (variety $X_R$) has a resolution of singularity (*i.e.*, there is a proper map $Y \longrightarrow X_R$ such that $Y$ is smooth and the map is an isomorphism on a nonempty open set) and hence a variety can be approximated by a smooth variety. But in characteristic $p > 0$, the existence of the resolution of singularity is not known in general (it is a long standing open problem).

On the other hand, in characteristic $p$, we have the Frobenius map

$$F : R \to R \text{ given by } x \mapsto x^p,$$

which is a ring homomorphism as $(x + y)^p = x^p + y^p$.

Now, analogous to the Hilbert-Samuel function and the Hilbert-Samuel multiplicity, Monsky had defined (in [22]) a characteristic $p$ numerical invariant of a ring $R$ (with respect to an ideal of finite colength)

*Definition* 2.1 For a Noetherian ring $R$ of dimension $d$, with char $R = p > 0$ and an ideal $I \subset R$ such that $\ell(R/I) < \infty$, the *Hilbert-Kunz function* $HK(R, I) : \mathbb{N} \to \mathbb{N}$ is given by

$$HK(R, I)(p^n) = \ell(R/I^{[p^n]}) = \ell(R/(f_1^{p^n}, \ldots, f_s^{p^n})),$$

where $\{f_1, \ldots, f_s\}$ is any set of generators of $I$. Let $q = p^n$, then

$$e_{HK}(R, I) = \lim_{q \mapsto \infty} \ell\left(R/I^{[q]}\right)/q^d$$

is called the *Hilbert-Kunz multiplicity* of $R$ with respect to $I$.

Interestingly it was E. Kunz [18] who introduced the colengths $\ell(R/\mathbf{m}^{[q]})$ (to characterize the regularity property of the ring) and at the same time he gave a counterexample to show $\lim_{q \mapsto \infty} \ell(R/I^{[q]})/q^d$ *does not* exist. Later P.Monsky [22] unaware of the counterexample, gave a proof that $\lim_{q \mapsto \infty} \ell(R/I^{[q]})/q^d$ exists in general. So it was discovered that the counterexample was not correct and then a series of work began on this invariant.

P. Monsky named this limit as the *Hilbert-Kunz multiplicity* $e_{HK}(R, I)$. He

proved (in the same paper)

$$HK(R,\mathbf{m})(q) = e_{HK}(R,\mathbf{m})q^d + O(q^{d-1})^1, \quad \text{where} \quad q = p^n$$

where $e_{HK}(R,\mathbf{m}) \in \mathbb{R}^+$.

One can easily see that

$$e_0(R,\mathbf{m})/d! \leq e_{HK}(R,\mathbf{m}) \leq e_0(R,\mathbf{m}).$$

In particular, for one dimensional rings, $e_0(R,\mathbf{m}) = e_{HK}(R,\mathbf{m})$.

**Open question** (Monsky 1983 [22]): *Is $e_{HK}(R)$ a rational number?*

Though it is easy to see that for a hypersurface ring $R = k[X_1,\ldots,X_n]/(f)$, where $f$ is a homogeneous polynomial of degree $m$, we have $e_0(R,\mathbf{m}) = m$, for $\mathbf{m} = (X_1,\ldots,X_n)$, it took many people and quite a few years to compute $e_{HK}$ for (specific) hypersurfaces, and in general it is still not known.

We recall some examples for which $e_{HK}(R,I)$ or $HK(R,I)$ has been computed.

(1) $R =$ a polynomial ring over a field (Kunz [18], this was easy of course)
(2) $R = k[X,Y,Z]/(f)$ a plane curve. Then
   (a) if $R$ a nodal plane curve (Monsky [25]).
   (b) If $R$ an elliptic plane curve and char $k \neq 2$ (Buchweitz-Chen [5], Pardue [27]), if $R$ an elliptic plane curve and char $k = 2$ (Monsky [23]).
(3) Diagonal hypersurfaces (Hans-Monsky [17]).
(4) Monomial ideals and binomial hypersurfaces (Conca [7]).
(5) Monoid rings, toric ring ( Eto [10], Watanabe [34], Bruns [4]) .
(6) Trinomial plane curves, *i.e.*, $k[X,Y,Z]/(f)$, where $f$ is a trinomial (Monsky [24]).
(7) $R$ a homogeneous cordinate ring of
   (1) $X =$ an elliptic curve with respect to line bundles $\mathcal{L}$ of degree $\geq 3$ ([11]), or
   (2) $X =$ a full flag variety with respect to anticanonical line bundle $\mathcal{L}$ ([11]), or
   (3) $X = F_a$ a Hirzebruch surface, for $a \geq 1$, with respect to any ample line bundle ([31]).

   In these cases we have

$$HK(X,\mathcal{L})(q) = e_{HK}(R)q^d + C_1(n)q^{d-1} + \cdots + C_d(n),$$

   where $q = p^n$ and $C_i(n)$ are periodic functions of $n$.

We note that the above examples (except (5), (6) and (7)) are hypersurfaces of special types, or monomial rings (for which one is able to use combinatorial techniques etc.).

In particular we still do not know a general formula for $e_{HK}(R,\mathbf{m})$ when

(1) $(R,\mathbf{m})$ is a 2-dimensional local ring or
(2) $R \simeq k[X_0,X_1,X_2,X_3]/(f)$, where $f$ is a general homogeneous polynomial.

---

[1]For two functions $f(x)$ and g(x), the equality $f(x) = O(g(x))$ means there is a real number $C$ such that $|f(\lambda)| < Cg(\lambda)$ for all $\lambda >> 0$.

One reason, for difficulty in the computations, is that the standard reduction techniques (as used for Hilbert-Samuel multiplicity) are known to fail for HK multiplicity.

A natural question one can ask: *why is* $e_{HK}$ *interesting*?

One important reason: $e_{HK}(R)$ *is a subtler invariant than* $e_0(R)$ *and it reveals more information about the char* $p$ *features of the ring* $R$. We will illustrate this below. It was hoped at one point that $e_{HK}$ could be used in devising a proof of resolution of singularities in char $p > 0$, based on this characterization of nonsingularity. So far this has not worked.

(1) If $R$ is an integral domain then $e_{HK}(R, \mathbf{m}) = 1 \iff R$ is smooth at $\mathbf{m}$.
(2) If $R$ is Gorenstein then ([2])

$$e_{HK}(R, \mathbf{m}) < 1 + (1/d!) \implies R \text{ is } F\text{- rational},$$

where $d$ is the dimension of $R$.

We recall that the $F$-rationality property is a substitute for the rational singularity property in char $p$ (which is useful as we do not yet have a resolution of singularity for $R$ in char $p$).

(3) Conjecture (Watanabe-Yoshida [35]): If $R$ is a $d$-dimensional ring of char $p > 2$ such that it is not smooth at a maximal ideal $\mathbf{m}$ and $R/\mathbf{m} = \bar{\mathbb{F}}_p$. Then
   (a) $e_{HK}(R, \mathbf{m}) \geq e_{HK}(A_{p,d}, \mathbf{n}) \geq 1 + a_d$,
   where $A_{p,d}$ is a quadratic $d$-dimesional hypersurface in char $p$,

$$A_{p,d} = \bar{\mathbb{F}}_p[[X_0, \ldots, X_d]]/(X_0^2 + \cdots + X_d^2)$$

   with maximal ideal $\mathbf{n} = (X_0, \ldots, X_d)$, and $a_d$ is the cofficient of $z^d$ in the power series expansion of $\sec z + \tan z$ around 0.
   (b) if the first equality holds then $R \cong A_{p,d}$ analytically (*i.e.*, the completion of $R$ at $\mathbf{m}$ is isomorphic to $A_{p,d}$).

In a later paper [36], Watanabe-Yoshida themselves proved the conjecture upto dimension 4. The first part of the conjecture has been verified upto dimension 6 by [1] and also for complete intersection rings (*e.g.*, for $R = k[X_1, \ldots, X_n]/(h_1, \ldots, h_s)$, where dimension of $R = n - s$) by [9].

## 3. Techniques from projective geoemetry

Now onwards we consider standard graded rings, *i.e.*,

$$R = k[X_1, \ldots, X_n]/(f_1, \ldots, f_m), \text{ where } f_1, \ldots, f_m \text{ are homogeneous polynomials}$$

and $k$ is an algebraically closed field. We denote the homogeneous maximal ideal by $\mathbf{m}$. (For details of the following terminology see Chapter II of [13]) Let

$$X = \text{Proj } R = \{x \in X_R \setminus \{0\}\}/\{x \simeq \lambda x \mid \lambda \in k \setminus \{0\}\}$$

be the associated projective algebraic variety, thought of as a quotient space. We note that $X$ is a complete but not an affine variety, nevertheless $X$ has a covering by affine varieties $\{X_{R_{(f)}} \mid f \text{ homogeneous element in } R\}$, where

$$R_{(f)} = \{x/f^i \mid x \in R \text{ is homogeneous of degree} = \text{degree}(f^i)\}.$$

The sheaves of rings (modules) on $X$ is a compatible set of rings (modules) assigend to each open set of $X$. The structure sheaf of rings $\mathcal{O}_X$ on the open set $X_{R_{(f)}}$ is the ring $R_{(f)}$. A vector bundle of rank $r$ on $X$ is a locally free sheaf of $\mathcal{O}_X$-modules such that $V$ on $X_{R_{(f)}}$ is $\oplus^r R_{(f)}$. A line bundle is a vector bundle of rank 1 on $X$.

Let $\mathcal{O}_X(1)$ be the ample line bundle corresponding to the embedding $X \longrightarrow \mathbb{P}^n$ induced by the canonical surjective map

$$k[X_1, \ldots, X_n] \longrightarrow k[X_1, \ldots, X_n]/(f_1, \ldots, f_m).$$

Now to compute $e_{HK}(R, \mathbf{m})$, we apply cohomological techniques as follows.

Let $h_1, \ldots, h_s$ be a set of degree one homogeneous generators of $\mathbf{m}$. Consider the canonical short exact sequence

$$0 \to V \to \oplus^s \mathcal{O}_X \to \mathcal{O}_X(1) \to 0,$$

where the map $\oplus^s \mathcal{O}_X \longrightarrow \mathcal{O}_X(1)$ is given by $(a_1, \ldots, a_s) \to \sum_i a_i h_i$.

Note that $V$ is a vector bundle on $X$ and for $m >> 0$, we have the exact sequence of $k$-vector spaces

$$0 \to H^0(X, F^{s*}(V)(m)) \to R_1^{[q]} \otimes R_m \xrightarrow{\phi_{m,q}} R_{m+q} \to H^1(X, F^{s*}(V)(m)) \to 0,$$

where

$$\ell(R/I^{[q]}) = \ell(R_0) + \ell(R_1) + \cdots + \ell(R_{q-1}) + \sum_{m \geq 0} \ell(\text{coker } \phi_{m,q}).$$

Thus the computation of $e_{HK}$ is reduced to the computation of the cohomologies of a vector bundle whose rank and degree we know.

Now a vector bundle on a normal variety $X$ (replacing $R$ by its integral closure we can assume that $X$ is a normal variety) has a unique filtration by subbundles known as the Harder-Narasimhan (HN) filtration of $V$

$$0 = V_0 \subset V_1 \subset \cdots \subset V_l = V.$$

This filtration has the property that each subquotient $V_i/V_{i-1}$ is *semistable*. We recall that, on a projective curve, a vector bundle $V$ is semistable if for every subbundle

$$W \subset V \implies \mu(W) := \deg\ W/\text{rank } W \leq \mu(V) = \deg\ V/\text{rank } V,$$

where $\mu(W)$ is called the slope of $W$. In particular, a bundle $V$ is semistable if and only if the HN filtration of $V$ is trivial. A semistable bundle has several nice properties: for example, $V$ semistable implies that, for any line bundle $\mathcal{L}$, the dual of $V$ and $V \otimes \mathcal{L}$ both are semistable. A semistable vector bundle $W$ of negative degree has $H^0(X, W) = 0$.

However it is not easy to construct the HN filtration for a vector bundle. A bundle $V$ is *strongly semistable* if $F^{s*}V$ is semistable for every $s \geq 1$, where we recall that the $s$-the iterated map $F^s : R \longrightarrow R$ given by $x \to x^{p^s}$ induces the $s$-th iterated Frobenius map $F^s : X \to X$. Though a semistable bundle may not be strongly semistable, there exists (see [19]) $s_0 >> 0$ such that, for $s \geq s_0$, the

bundle $F^{s*}V$ has HN filtration where each subquotient of the filtration is strongly semistable (we call it the strong HN filtration).

In particular if $R$ is a standard graded two dimensional ring (so $X$ is a projective curve) over a field of char $p > 0$, then (see [3], [28])

$$e_{HK}(R) = \frac{\deg\ X}{2}(\mu_{HK}(V) - \text{embdim}(R)),$$

where $\mu_{HK}(V)$ is a number given in terms of the normalized slopes of the strong HN filtration of $V$.

In the case of plane curves $e_{HK}$ gives a numerical characterization of the Frobenius semistablity behaviour of the syzygy bundle, *i.e.*, the minimum $s_0$ such that $F^{s*}V$ has strong HN filtration for $s \geq s_0$. In particular the computations of $e_{HK}$ for plane trinomials curves (in [24]) give examples of semistable bundles $V$ such that the semistablity of $F^{m*}V$ is not an open property *reduction mod p*. On the other hand the semistability property is known to be the open property *reduction mod p* ([20]).

## 4. Hilbert-Kunz density function

The previous relation between the semistability of the syzygy vector bundle $V$ and the $e_{HK}$, does not hold for dimension $d \geq 3$.

In this section we fix a graded pair $(R, I)$, *i.e.*, $R = \oplus_{m\geq 0} R_m$ a standard graded ring and $I \subset R$ a graded ideal such that $\ell(R/I) < \infty$. Moreover dim $R = d \geq 2$ and char $R = p > 0$. For such a graded pair $(R, I)$, we define ([30]) the *Hilbert-Kunz density function* (HK density function) $f_{R,I} : [0, \infty) \longrightarrow [0, \infty)$ as follows.

For $q = p^n$, consider the step function

$$f_n(R, I) : [0, \infty) \longrightarrow [0, \infty) \quad \text{given by} \quad x \to \frac{1}{q^{d-1}} \ell \left( \frac{R}{I^{[q]}} \right)_{\lfloor xq \rfloor}$$

and define $f_{R,I}(x) = \lim_{n\to\infty} f_n(R, I)(x)$.

Note that this makes sense as the sequence $\{f_n(R, I)\}_n$ is convergent (in fact uniformly convergent). This assertion strongly uses the fact that, for a prime $p$, the series $\sum_{i\geq 0} 1/p^i$ is convergent. Moreover $f_{R,I}$ is a compactly supported continuous function such that

$$\int_0^\infty f_{R,I}(x)dx = e_{HK}(R, I).$$

Since $f_{R,I} \in L^1(\mathbb{R})$ we have the association

$$\{f_{R,I} \mid (R, I) \ \text{ is a graded pair}\} \longleftrightarrow \{\hat{f}_{R,I} : \mathbb{C} \longrightarrow \mathbb{C}\},$$

where $\hat{f}_{R,I}$ is the holomorphic Fourier transform of $f_{R,I}$ given by

$$\hat{f}_{R,I}(z) = \int_0^\infty f_{R,I}(x)e^{-ixz}dx, \quad \text{for} \ \ z \in \mathbb{C}, \ \ \text{and} \ \ \hat{f}_{R,I}(0) = e_{HK}(R, I).$$

In other words, for a given graded pair $(R, I)$, to study the number $e_{HK}(R, I)$

we consider the function $f_{R,I} : [0,\infty) \longrightarrow [0,\infty)$, where

$$\int_0^\infty f_{R,I}(x)dx = e_{HK}(R,I).$$

Though this seems a rather convoluted approach, it has several advantages.

(1) The HK density function is additive (like $e_{HK}$), *i.e.*,

$$f_{M,I} = \sum_{P\in\Lambda} f_{R/P,I+P/P}\ell(M_P),$$

where $\Lambda = \{\text{prime ideals } P \text{ of } R \mid \dim R/P = \dim R\}$.

(2) The HK density function also has a multiplicative property (unlike $e_{HK}$): let $(R,I)$ and $(S,J)$ be two graded pairs defined over the same field $k$. Then $(R\#S, I\#J)$ is a graded pair, where $R\#S = \oplus_{n\geq 0}(R_n \otimes_k S_n)$ and $I\#J = \oplus_n(I_n \otimes J_n)$. The ring $R\#S$ is called the Segre product of $R$ and $S$ as Proj $R\#S =$ Proj $R \times$ Proj $S$, *i.e.*, the Segre product of rings corresponds to the product of the projective varieties. We have ([30])

$$F_{R\#S} - f_{R\#S,I\#J} = [F_R - f_{R,I}] \cdot [F_S - f_{S,J}],$$

where, if dim $R = d$ then $F_R(x) := e_0(R)x^{d-1}/(d-1)!$.

### 4.1. *Some applications of the HK density function*

It is easy to see that $e_0(R,I^k) = k^d e_0(R,I)$. However any relation of this type between $e_{HK}(R,I^k)$ and $e_{HK}(R,I^k)$ was not known earlier. The asymptotic behaviour of $e_{HK}(R,I^k)$ as $k \to \infty$ (studied in [35] and [12]) can be expressed as follows:

$$0 \leq e_{HK}(R,I^k) - k^d e_0(R,I)/d! = O(k^{d-1}).$$

Using the HK density function (with a modified uniformly converging sequence to the HK density function) one can prove ([29]) that

$$A_{R,I} := \lim_{k\to\infty} \frac{e_{HK}(R,I^k) - k^d e_0(R,I)/d!}{k^{d-1}}$$

exists. Morever

$$[e_0(R,\mathbf{m})/(d-1)!]^{\frac{2-d}{d-1}}(A_{R,\mathbf{m}}) \geq (d-1)/d. \tag{4.1}$$

This invariant gives an algebraic characterization of the tiling of a convex polytope with respect to the given lattice as follows.

Recall that the set of toric pairs $(X,D)$, (*i.e.*, $X$ is a $d-1$ dimensional projective toric variety with a very ample Cartier divisor $D$) is in one to one correspondence with the set of canonical $d-1$ dimensional integral very ample convex polytopes $P_{X,D}$.

Now, for any (rational) convex polytope in $\mathbb{R}^{d-1}$, one can choose $m >> 0$ such that $mP$ is a very ample integral convex polytope and therefore $mP = P_{X,D}$, for

some toric pair $(X, D)$. Then from the theory of the HK density function for the toric pair $(X, D)$ it follows that the polytope $\lambda P$ (for some (unique) $\lambda \in \mathbb{R} \setminus \{0\}$) tiles the space $\mathbb{R}^{d-1}$ (with respect to the given lattice) if and only if the equality holds in (4.1), where $(R, \mathbf{m})$ denotes the homogeneous coordinate ring of the toric pair $(X, D)$. In other words, in the toric case, the normalized asympototic growth of $e_{HK}(I^k)$ is the slowest if and only if the convex polytope $P_{X,D}$ tiles the ambient space $\mathbb{R}^{d-1}$.

So far we have seen that the integral of the HK density function $f_{R,I}$ is $e_{HK}(R, I)$. Now we consider another invariant attached to $f_{R,I}$ (see [33]), namely the maximum support $\alpha(R, I)$ of the function $f_{R,I}$, *i.e.*, $\alpha(R, I) = \text{Sup } \{x \mid f_{R,I}(x) > 0\}$.

We recall the following notion of $F$-threshold, as defined in [14] and proved in full generality in [8].

*Definition* 4.1 Let $I$ and $J$ be two ideals such that $J \subseteq \sqrt{I}$. Then the $F$-threshold of $J$ with respect to $I$ is

$$c^I(J) = \lim_{q \to \infty} \frac{\min \{r \mid J^{r+1} \subseteq I^{[q]}\}}{q}.$$

Now if Proj $R$ is a smooth variety then $\alpha(R, I) = c^I(\mathbf{m})$, where $c^I(\mathbf{m})$ is the $F$-threshold of $\mathbf{m}$ with respect to $I$. On the other hand, when dim $R = 2$ then the HK density function is a piecewise linear polynomial with coefficients as the slopes of the strong HN filtration of associated syzygy vector bundle.

The formula of $c^I(\mathbf{m})$ in terms of the slopes of the strong HN filtration of the syzygy bundle (applied to a modified old example of D. Gieseker [16]) gives (see [32]) a projective curve where the set of $F$-thresholds of the maximal ideal is not discrete, which answers a question in [26].

## References


[1] Aberbach, I., Enescu, F., *New estimates of Hilbert-Kunz multiplicities for local rings of fixed dimension*, Nagoya Math. J. 212 (2013), 59-85.

[2] Blickle, M., Enescu, F., *On rings with small Hilbert-Kunz multiplicity*, Proc. Amer. Math. Soc. **132**, (2004), 2505-2509.

[3] Brenner, H. *The rationality of the Hilbert-Kunz multiplicity in graded dimension two*, Math. Ann. 334 (2006), 91-110.

[4] Bruns, W. *Conic divisor classes over a normal monoid algebra Commutative Algebra and Algebraic Geometry*, Contemp. Math., vol. 390, Amer. Math. Soc., Providence, RI (2005), pp. 63-71.

[5] R.O. Buchweitz, Q. Chen, *Hilbert-Kunz functions of cubic curves and surfaces*, J. Algebra **197** no. 1, (1997), 246-267.

[6] R.O. Buchweitz, Q. Chen, K. Pardue, *Hilbert-Kunz functions*, preprint (1996)

[7] A. Conca, *Hilbert-Kunz functions of monomial ideals and binomial hypersurfaces*, Manuscripta Math. **90**, (1996), 287-300.

[8] Stefani, A., Núñez-Betancourt, L., Pérez, F., *On the existence of F-thresholds and related limits*, Trans. Amer. Math. Soc. 370 (2018), no. 9, 6629-6650.

[9] Enescu, F., Shimomoto, K., *On the upper semi-continuity of the Hilbert-Kunz multiplicities*, Journal of Algebra **285** (2005), 222-237.

[10] Eto, K., *Multiplicity and Hilbert-Kunz multiplicity of monoid rings*, Tokyo J. Math. 25 (2002), no. 2, 241-245.

[11] Fakhruddin, N.; Trivedi, V., *Hilbert-Kunz functions and multiplicities for full flag varieties and elliptic curves*, J. Pure Appl. Algebra **181** (2003), no. 1, 23-52.

[12] Hanes, D., *Notes on the Hilbert-Kunz function*, J. Algebra. 265 (2003) 619-630.

[13] Hartshorne, R., *Algebraic geoemetry*, Springer-Verlag NY (1977).
[14] Huneke, C., Mustaţă, M., Takagi, S., Watanabe, K.I., *F-thresholds, tight closure, integral closure and multiplicity bounds*, Michigan Math. J. 57, in Special Volume in Honor of Melvin Hochster, Univ. Michigan Press, Ann Arbor, (2008), 463-483.
[15] Huneke, C., *Hilbert-Kunz multiplicity and the F-Signature*, Commutative algebra, 485-525, Springer, New York, 2013.
[16] Gieseker, D., *Stable vector bundles and the Frobenius morphism*, Ann. Sci. cole Norm. Sup. (4) 6 (1973), 95-101.
[17] Han, C., Monsky, P., *Some surprising Hilbert-Kunz functions*, Math. Z., 214 (1993), no. 1, 119-135.
[18] E. Kunz, *Characterizations of regular rings of characteristic p*, Amer. J. Math. **41** (1969), 772-784.
[19] Langer, A., *Semistable sheaves in positive characteristic*, Ann. Math., 159 (2004).
[20] Maruyama, M., *Openness of a family of torsion free sheaves*, J. Math. Kyoto Univ. 16 (1976), no. 3, 627-637.
[21] Mondal, M., Trivedi, V., *Hilbert-Kunz density function and asymptotic Hilbert-Kunz multiplicity for projective toric varieties*, J. Algebra 520 (2019), 479516.
[22] Monsky, P., *The Hilbert-Kunz function*, Math. Ann. **263** (1983), 43-49.
[23] Monsky, P., *The Hilbert-Kunz function of a characteristic* 2 *cubic*, Jouranl of Algebra **197** (1997) 268-277.
[24] Monsky, P., *The Hilbert-Kunz multiplicity of an irreducible trinomial*, Journal of Algebra **304** (2006) 1101-1107.
[25] Monsky, P., *The Hilbert-Kunz theory for nodal cubics, via sheaves*, Journal of Algebra **346** (2011), 180-188.
[26] Mustaţă, M., Takagi, S., Watanabe, K.I., *F-thresholds and Bernstein-Sato polynomials*, European congress of mathematics, 341-364, Eur. Math. Soc., Zurich, 2005.
[27] Pardue, K., *Nonstandard Borel-Fixed Ideals*, Doctoral Thesis, Brandeis University, 1994.
[28] Trivedi, V., *Semistability and Hilbert-Kunz multiplicities for curves*, Journal of Algebra 284 (2005), no.2, 627-644.
[29] Trivedi, V., *Asymptotic Hilbert-Kunz multiplicity*, Journal of Algebra, 492 (2017), 498-523.
[30] Trivedi, V., *Hilbert-Kunz density Function and Hilbert-Kunz multiplicity*, Trans. Amer. Math. Soc. 370 (2018), no. 12, 8403-8428.
[31] Trivedi, V., *Hilbert-Kunz functions of a Hirzebruch surface*, Journal of Algebra **457** (2016), 405-430.
[32] Trivedi, V., *Nondiscreteness of F-thresholds*, arXiv:1808.07321, to appear in Math. Research letters.
[33] Trivedi, V., Watanabe, K., *Hilbert-Kunz density functions and F-threshold*, arXiv:1808.04093v1.
[34] Watanabe, K., *Hilbert-Kunz multiplicity of of toric rings*, Tokyo J. Math, 35 (2000), 173-177.
[35] Watanabe, K., Yoshida, K., *Hilbert-Kunz multiplicity of two-dimensional local rings*, Nagoya Math. J. 162 (2001) 87-110.
[36] Watanabe, K., Yoshida, K., *Hilbert-Kunz multiplicity of three-dimensional local rings*, Nagoya Math. J. 177 (2005), 47-75.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 75–84

# *Symbolic Powers, Set-Theoretic Complete Intersection and Certain Invariants*

**Clare D'Cruz***

*Chennai Mathematical Institute, Plot H1 SIPCOT IT Park, Siruseri, Kelambakkam 603103, Tamil Nadu, India*

**Abstract:** In this survey article we give a brief history of symbolic powers and its connection with the interesting problem of set-theoretic complete intersection. We also state a few problems and conjectures. Recently, in connection to symbolic powers is the containment problem. We list a few interesting results and related problems on the resurgence, Waldschmidt constant and Castelnuovo-Mumford regularity.

## 1. Introduction

Let $A$ be a Noetherian ring and $I$ an ideal in $A$ with no embedded components. Then the ideal $I^{(n)} := A \cap (\bigcap_{\mathfrak{p} \in Ass(A/I)} I^n A_{\mathfrak{p}})$ is the $n$-th symbolic power of $I$. The study of $n$-th symbolic power has been important for the last few decades mainly because of its connection with algebraic geometry. In recent years, it has become even more active area of research mainly because several interesting associated invariants. We list some of the interesting questions and open problems.

To understand the connection with algebraic geometry, let $\Bbbk$ be a field and let $\mathbb{A}^d$ (or $\mathbb{A}^d_{\Bbbk}$) denote the set or all $d$-tuples $\underline{a} = (a_1, \ldots, a_d)$ where $a_i \in \Bbbk$ for all $i = 1, \ldots, d$. The set $\mathbb{A}^d$ is called the affine $d$-space of dimension $d$ over $\Bbbk$. We say that a subset $Y$ in $\mathbb{A}^d$ is a zero set if it is the set of common zeros of a collection of polynomials $f_1, \ldots, f_m \in R := \Bbbk[X_1, \ldots, X_d]$ and we denote it by $Y = Z(f_1, \ldots, f_m)$. We can define a topology on $\mathbb{A}^n$ by defining the closed sets to be the zero sets. If $I = (f_1, \ldots, f_m)$, then $Y = Z(I)$. To every subset of $Y \subset \mathbb{A}^n$ we can define the ideal of $I(Y) := \{f \in \Bbbk[X_1, \ldots, X_n] | f(P) = 0 \text{ for all } P \in Y\}$. An irreducible closed subset $Y$ of $\mathbb{A}^n$ called an affine algebraic set.

Let $Y$ be an algebraic set. We say that $Y$ is defined set-theoretically by $n$ elements if there exists $n$ elements $f_1, \ldots, f_n \in R$ such that $I(Y) = \sqrt{(f_1, \ldots, f_n)}$. In 1882, Kronecker showed that $I$ can be set-theoretically defined by $d + 1$ equations [43]. Later, this result was improved by Storch [60] and by Eisenbud and Evans [24]. It follows from their work that if $\Bbbk$ is algebraically closed and $I$ is an homogenous ideal, then $I$ can be defined set-theoretically by $d$ elements. Hence it was of interest to know which ideals $I$ could be defined set-theoretically by $d - 1$ elements. If $I$ is locally complete intersection of pure dimension one, then $I$ can be defined set-theoretically

* Partially supported by a grant from Infosys Foundation. Email: clare@cmi.ac.in

by $d-1$ elements ([26], [6], [49], [64]). In 1992, Lyubeznick showed that if $V$ is an algebraic set in $\mathbb{A}^d_{\Bbbk}$ and $\text{char}(\Bbbk) = \mathfrak{p} > 0$, then $V$ can be defined set-theoretically by $d-1$ elements [45].

In 1978, Cowsik and Nori proved a remarkable result. They showed that if $\text{char}(\Bbbk) = \mathfrak{p} > 0$, then any affine curve is a set-theoretic complete intersection ([11, Theorem 1]). If $\text{char}(\Bbbk) = 0$, then one of the best known results in $\text{char}(\Bbbk) = 0$ is the result of Herzog which was later also proved by Bresinsky [7]. He showed that all monomial curves in $\mathbb{A}^3$ are set-theoretic complete intersection.

In 1981 Cowsik proved an interesting result which connects commutative algebra and algebraic geometry:

THEOREM 1.1 *[10] Let $(R, \mathfrak{m})$ a Noetherian local ring and $\mathfrak{p} \neq \mathfrak{m}$ a prime ideal. If the symbolic Rees algebra $R_s(\mathfrak{p}) := \oplus_{n\geq 0}\mathfrak{p}^{(n)}$ is Noetherian, then $\mathfrak{p}$ can be defined set-theoretically by $d-1$ elements.*

However, the converse need not be true. Cowsik's result motivated several researchers to investigate the Noetherian property of the symbolic Rees algebra. In 1987, Huneke gave necessary and sufficient conditions for $\mathcal{R}_s(\mathfrak{p})$ to be Noetherian when $\dim\ R = 3$ [40]. Huneke's result was generalised in 1991 for $\dim\ R \geq 3$ by Morales [50]. All these results paved a new way to study the famous problem on set-theoretic complete intersection. In section 2 we discuss some of these problems.

We also have the famous result due to Zariski [67] and Nagata [51] which states that if $\Bbbk$ is an algebraically closed field, then the $n$-th symbolic power of a given prime ideal consists of the elements that vanish up to order $n$ on the corresponding variety [23]. This result has been generalised to perfect fields $\Bbbk$ and radical ideals [16, Proposition 2.14, Exercise 2.15].

In general, symbolic powers of ideals are hard to compute. Hence recently, associated to symbolic powers of ideals, Bocci and Harbourne introduced a quantity called the resurgence [4]. In section 3, we will state the recent developments on this quantity and related invariants like the Waldschmidt constant and Castelnuovo-Mumford regularity.

## 2. Set-theoretic complete intersection and symbolic Rees algebra

### 2.1. *Set-theoretic complete intersection in $\mathbb{A}^n$ and $\mathbb{P}^n$*

Throughout this section $R = \Bbbk[X_1, \ldots, X_d]$ where $X_1, \ldots X_d$ are variables. We say that a radical ideal $I \subset R$ is a set-theoretic complete intersection of there exists $h = \text{ht}(I)$ elements such that $I = \sqrt{(f_1, \ldots . f_h)}$. Let $\underline{a} := (a_1, \ldots, a_d)$ be positive integers such that $\gcd(a_1, \ldots, a_d) = 1$. Let $\mathcal{C}(\underline{a}) := \{\underline{t^a} = (t^{a_1}, t^{a_2}, \ldots, t^{a_d}) | t \in \Bbbk\}$ be a curve in $\mathbb{A}^n$. If $\phi : R \longrightarrow \Bbbk[T]$ is the homomorphism given by $\phi(X_i) = T^{a_i}$ for all $i = 1, \ldots, d$. Then $\mathfrak{p}(\mathcal{C}(\underline{a})) := \ker(\phi)$ is the prime ideal defining the curve $\mathcal{C}(\underline{a})$. In other words, $I(\mathcal{C}(\underline{a})) = \mathfrak{p}(\mathcal{C}(\underline{a}))$. We say that $\mathcal{C}(\underline{a})$ can be defined set-theoretically by $d-1$ elements if there exists $d-1$ elements $f_1, \ldots, f_{d-1} \in \mathfrak{p}$ such that $\mathfrak{p} = \sqrt{(f_1, \ldots, f_{d-1})}$.

Let $d = 3$, then we have the interesting result:

THEOREM 2.1 *Let* $\gcd(a_1, a_2, a_3) = 1$.

*(1) [37] Then one of the following is true:*
  *(a) $\mathfrak{p}(\mathcal{C}(\underline{a}))$ is a complete intersection.*
  *(b) There exists integers $\alpha_i, \beta_i, \gamma_i$; $i = 1, 2$ such that $\mathfrak{p}(\mathcal{C}(\underline{a}))$ is generated by*

$2 \times 2$ *minors of the matrix* $\begin{pmatrix} X_1^{\alpha_1} & X_2^{\beta_1} & X_3^{\gamma_1} \\ X_2^{\beta_2} & X_3^{\gamma_2} & X_1^{\alpha_2} \end{pmatrix}$.

*(2)* [Herzog (Unpublised work)] *and [7]* $\mathfrak{p}(\mathcal{C}(\underline{a}))$ *is a set-theoretic complete intersection.*

Such a result is not known for $\mathbb{A}^d$, $d \geq 4$. The first result in higher dimension was given by Bresinsky where he showed that certain Gorenstein curves in $\mathbb{A}^4$ are set-theoretic complete intersection [8]. In 1990, Patil proved the following result [53]:

THEOREM 2.2 *Let* $a_1, a_2, \ldots, a_{d-2}$ *be an arithmetic sequence. Then* $\mathfrak{p}(\mathcal{C}(\underline{a}))$ *is a set-theoretic complete intersection.*

In 1980, Valla showed that certain determinantal ideals were set-theoretic complete intersection [65]. As a consequence they prove the following:

THEOREM 2.3 *[65, Example 3.3] Let* $q, m$ *be positive integers with* $\gcd(q, m) = 1$. *Put* $a_i = 2q + 1 + (i-1)m$, *for* $i = 1, 2, 3$. *Then* $\mathfrak{p}(\mathcal{C}(\underline{a}))$ *is a set-theoretic complete intersection.*

In the past forty years several researchers have given interesting examples of affine varieties which are set-theoretic complete intersection. However, the following question is still open.

*Question* 2.4 Let $\Bbbk$ be a field of characteristic zero and $d \geq 4$. Is every curve $\mathcal{C}(\underline{a}) \subseteq \mathbb{A}^d$ a set-theoretic complete intersection?

We would like to bring to the attention a paper of Moh where he considered the set-theoretic complete intersection problem of analytic space curves over an algebraically closed field. Let $\Bbbk[[X, Y, Z]]$ and $\Bbbk[[T]]$ be power series rings and $\phi : \Bbbk[[X, Y, Z]] \longrightarrow \Bbbk[[T]]$ be given by $\phi(X) = T^a + \cdots$, $\phi(Y) = T^b + \cdots$ and $\phi(Z) = T^c + \cdots$. Let $\mathfrak{p} = \ker(\phi)$. Such curves are called Moh curves. Moh showed that if $(a-2)b < c$, then $\mathfrak{p}$ is a set- theoretic complete intersection [47]. In [40] Huneke showed that the symbolic Rees algebra of the Moh curve parameterized by $(t^6, t^7 + t^{10}, t^8)$ is Noetherian. However, it is not easy to describe the defining ideal of a Moh curve. The following question is still open:

*Question* 2.5 Let $\phi : \Bbbk[[X, Y, Z]] \longrightarrow \Bbbk[[T]]$ be given by $\phi(X) = T^6 + T^{31}$, $\phi(Y) = T^8$ and $\phi(Z) = T^{10}$. Let $\mathfrak{p} = \ker(\phi)$. Is $\mathfrak{p}$ a set-theoretic complete intersection?

We now focus our attention on curves in the projective space $\mathbb{P}_k^n$. It is a long standing question whether every connected subvariety in $\mathbb{P}_k^n$ is a set-theoretic complete intersection [35]. The answer is not known even for curves in $\mathbb{P}^3$. We list a few results in this direction.

Let $\underline{a} := (a_1, \ldots, a_d)$ be integers such that $\gcd(a_1, \ldots, a_d) = 1$ and $0 = a_0 < a_1 < a_2 \cdots < a_d$. Put $S = \Bbbk[X_0, X_1, \ldots, X_d]$ and let $\psi : S \longrightarrow \Bbbk[T, U]$ be the homomorphism given by $\phi(X_i) = T^{a_d - a_i} U^{a_i}$ for all $i = 0, \ldots, d$. Then $\mathfrak{p}(\overline{\mathcal{C}}(\underline{a})) := \ker(\phi)$ is the prime ideal defining ideal of the curve $\overline{\mathcal{C}}(\underline{a})$. In other words, $I(\overline{\mathcal{C}}(\underline{a}) = \mathfrak{p}(\overline{\mathcal{C}}(\underline{a}))$.

One of the most simplest and interesting example was given by Hartshorne.

THEOREM 2.6 *[36] Let* $\Bbbk$ *be a field of positive characteristic* $\mathfrak{p}$ *and* $\underline{a} = (1, a-1, a)$, *where* $a \geq 4$. *Then* $\mathfrak{p}(\overline{\mathcal{C}}(\underline{a}))$ *is a set-theoretic complete intersection.*

Hartshorne's result was generalised by Ferrand [26]. Robbiano and Valla studied the curve $\overline{\mathcal{C}}(\underline{a})$ for $\underline{a} = (1, 3, 4)$ [55]. In 1991, Moh generalised the work of Hartshorne

and Ferrand [48].

An interesting result in $\mathbb{P}^3$ is:

THEOREM 2.7 *[55], [61] Let $\mathcal{C}$ be a curve in $\mathbb{P}^3$. Let $I(\mathcal{C}) \subset S = \Bbbk[X_0, X_1, X_2, X_3]$ be the ideal of the $\mathcal{C}$. If $S/I(\mathcal{C})$ is Cohen-Macaulay, then $I(\mathcal{C})$ is a set-theoretic complete intersection.*

A considerable amount of research has been done in this area. It is impossible to list all of them there. The list of references is not exhaustive. For a good collection of articles on set-theoretic complete intersection one can read [32]. An interesting survey on this topic was also given by Lyubneznik [44].

## 2.2. *Symbolic Rees Algebra*

The work of Cowsik, Huneke and Morales motivated several researchers to work on the symbolic powers of a prime ideal and the symbolic Rees algebra $\mathcal{R}_s(\mathfrak{p}) = \oplus_{n\geq 0}\mathfrak{p}^{(n)}$.

Let $(R, \mathfrak{m})$ be a Noetherian local ring of dimension $d$ and $\mathfrak{p}$ a prime ideal of height $d-1$. Some of the interesting questions on the symbolic Rees algebra are: (1) Is it Noetherian? (2) Is it Cohen-Macaulay (3) Is it Gorenstein? An answer to question (1) would imply that $\mathfrak{p}$ is a set-theoretic complete intersection, by Cowsik's result.

If $\phi : \Bbbk[[X_1, \ldots, X_d]] \longrightarrow \Bbbk[[T]]$ is the homomorphism given by $\phi(X_i) = T^{a_i}$ for all $i = 1, \ldots, d$, then $\mathfrak{p}(\underline{a}) := \ker(\phi)$.

In 1982, Huneke gave examples prime ideals $\mathfrak{p}$ in $\Bbbk[[X_1, X_2, X_3]]$ whose symbolic Rees algebra $R_s(\mathfrak{p})$ is Noetherian [39]. In fact he showed the following result.

THEOREM 2.8 *$R_s(\mathfrak{p}(\underline{a}))$ is Noetherian in the following cases:*

*(1) $\underline{a} = (2t+1, 2r+s, s+r+rs)$, where either $s \leq r$ or $s > r$ and $t = 1$.*
*(2) $\underline{a} = (s+2, 2r+1, s+1+rs)$, $2 \leq r \leq s$.*
*(3) $\underline{a} = (t+s+1, tr+t+1, rs+r+s)$, $r \leq t$ and $s \geq 1$.*

The first result which gave necessary and sufficient conditions for $R_s(\mathfrak{p})$ to be Noetherian was by Huneke [40, Theorem 2.1]. This result was later generalised by Morales [50, Theorem 2.1]. A consequence of their result is:

THEOREM 2.9 *Let $(R, \mathfrak{m})$ be a regular local ring of dimension $d$ and $\mathfrak{p}$ a prime ideal of height $d-1$. Then $R_s(\mathfrak{p})$ is Noetherian if and only if there exists $x \in \mathfrak{m} \setminus \mathfrak{p}$ and elements $f_i \in \mathfrak{p}^{(k_i)}$, $i = 1, \ldots, f_{d-1}$, such that*

$$\ell\left(\frac{R}{(\mathfrak{p}, f_1, \ldots, f_{d-1}, x)}\right) = \ell\left(\frac{R}{(p, x)}\right) k_1 \cdots k_{d-1}.$$

Using this criteria several researchers have produced examples of monomial curves $\mathfrak{p} \in \Bbbk[[X_1, X_2, X_3]]$ such that the symbolic Rees algebra $R_s(\mathfrak{p})$ is Noetherian. We cite a few examples here [58], [12], [29] [59], [30], [41], [28], [57].

One interesting question is: If $R$ is a Noetherian ring, and $\mathfrak{p}$ is a prime ideal, is the symbolic Rees algebra $R_s(\mathfrak{p})$ Noetherian? Rees provided an example which implies that this question does not have a positive answer [54]. Later Cowsik conjectured that if $R$ is a regular local ring and $\mathfrak{p}$ is a prime ideal, then $R_s(\mathfrak{p})$ is Noetherian. In 1990, Roberts gave a counter example to Cowsik's conjecture [56]. In 1994, Goto, Nishida and Watanabe gave an infinite class of monomial curves whose symbolic Rees

algebra $R_s(\mathfrak{p})$ is Noetherian if the characteristic of $\Bbbk$ is $\mathfrak{p}$, but if the characteristic of $\Bbbk$ is zero, then $R_s(\mathfrak{p})$ is not Noetherian [31].

We end this section stating a problem which is still open:

*Problem* 2.10 Let $\Bbbk$ be an algebraically closed field. Can one classify all monomial curves in $\mathbb{A}^3_{\Bbbk}$ for which $R_s(\mathfrak{p})$ is Noetherian?

An interesting paper in this direction is [12].

## 3. Resurgence, Waldschmidt constant and regularity

One of the reasons the symbolic Rees algebra is hard to analyse is because it is not easy to describe the symbolic powers even for curves in $\mathbb{A}^3_{\Bbbk}$. Hence, one would like to compare the symbolic powers and ordinary powers of an ideal. If $I$ is an ideal in a Noetherian ring $R$, then from the definition of symbolic powers it follows that $I^n \subseteq I^{(n)}$ and in fact for any proper ideal nonzero ideal $I$, $I^r \subseteq I^{(n)}$ holds if and only if $r \geq n$. A challenging problem to determine for which positive integers $n$ and $r$ the containment $I^{(n)} \subseteq I^r$ holds true. In [63] Swanson compared the symbolic powers and ordinary powers of several ideals.

For the rest of this section we will assume that $\Bbbk$ is algebraically closed, $S = \Bbbk[X_0, \ldots, X_d]$ and $I$ is an homogenous ideal in $S$. Hence in 2001, Ein, Lazarsfeld and Smith proved a very interesting result. It follows from their result

THEOREM 3.1 *[22]. Let $I \subset S$ be a proper ideal. If $h$ is the largest height of an associated prime of $I$, then $I^{(hn)} \subseteq I^n$ for all $n \geq 0$.*

In 2002, Hochster and Huneke proved a stronger result [38]. It follows from the above results that if $d = \dim\ R$, then $I^{(n)} \subseteq I^r$ for $n \geq (d-1)r$. In this direction, Harbourne raised the following: conjecture in:

CONJECTURE 3.2 *[1, Conjecture 8.4.2] For any homogeneous ideal $0 \neq I \subset S$, $I^{(n)} \subseteq I^r$ if $n \geq rd - (d-1)$.*

In the same paper they remark that from the methods in [38] there is enough evidence for this conjecture to be true at least when the characteristic of $\Bbbk$ is $\mathfrak{p}$ and $r = \mathfrak{p}^t$ for $t > 0$ (see [1, Example 8.4.4]. This has led to the study of the least integer $n$ for which $I^{(n)} \subseteq I^r$ holds for a given ideal $I$ and for an integer $r$. To answer this question C. Bocci and B. Harbourne defined an asymptotic quantity [5] called resurgence which is defined as

$$\rho(I) := \sup\{m/r \mid I^{(m)} \not\subseteq I^r\}.$$

Hence if $m > \rho(I)r$, then $I^{(m)} \subseteq I^r$.

In general resurgence is not easy to compute. Hence it is useful to give bounds. From the results in [22] it follows that $\rho(I) \leq d - 1$.

Another interesting invariant is the Waldschmidt constant. This constant was introduced by Waldschmidt in [66]. Let $I$ be an homogenous ideal and let $\alpha(I)$ denote the least degree of a homogeneous generator of $I$. Then we have the famous conjecture due to Nagata:

CONJECTURE 3.3 *[52] Let $V$ be a finite set of $n$ points in $\mathbb{P}^2_{\mathbb{C}}$ and $I(V)$ be the corresponding homogenous ideal in $\mathbb{C}[X_0, X_1, X_2]$. Then $\alpha(I^{(m)}) \geq m\sqrt{n}$.*

This conjecture is still open in general. It is know only in a few cases.
Define

$$\gamma(I) := \lim_{n\to\infty} \frac{\alpha(I^{(n)})}{n}.$$

$\gamma(I)$ is called the Waldschmidt constant ([66], [4]). Since $\alpha$ is subadditive, i.e., $\alpha(I)^{(n+m)} \leq \alpha(I^{(n)}) + \alpha(I^{(m)})$, it follows that $\gamma(I)$ exists ([66], [5]). Moreover, there is a lower bound for $\rho(I)$. It follows from [5, Lemma 2.3.2]:

LEMMA 3.4 *Let $0 \neq I \subset S$ be a homogenous ideal. Then $\gamma(I) \geq 1$ and*

$$1 \leq \frac{\alpha(I)}{\gamma(I)} \leq \rho(I).$$

Related to the Waldschmidt constant is the following conjecture:

CONJECTURE 3.5 (Chudnovsky) *Let $V$ be a set of points in $\mathbb{P}^d$ and $I(V)$ be the corresponding homogenous ideal in $S$. Then*

$$\gamma(I) \geq \frac{\alpha(I) + d - 1}{d}.$$

Recently, Chudnovsky's conjecture has attracted the attention of researchers ([21], [27], [46]).

For any homogenous ideal $I$ we can define the Castelnuovo-Mumford regularity as follows. Let $M$ be a finitely generated graded $S$-module. Let

$$F_\bullet : 0 \longrightarrow F_r \longrightarrow F_{r-1} \cdots \longrightarrow F_1 \longrightarrow F_0 \longrightarrow 0$$

be a minimal free resolution of $M$ where $F_i = \oplus_j S[-j]^{b_{ij}}$. Put $b_i(M) = \max\{j | b_{ij} \neq 0\}$. Then $\operatorname{reg}(M) = \max_i\{b_i(M) - i\}$.

Bounds on the Castelnuovo-Mumford regularity has been of interest. As the list is long we state only a few results. In 1997, Swanson proved that if $I$ is a homogenous ideal, then there exists an integer $r$ such that $reg(I^m) \leq mr$ for any m [62]. Later, it was proved that asymptotically $\operatorname{reg}(I^m)$ is a linear function of $m$ by [42], [14].

The behaviour of Castelnuovo-Mumford regularity of symbolic powers is not easy to predict. From a result of Cutkosky, Herzog and Trung, it follows that if $I$ is an ideal of points in a projective space and the symbolic Rees algebra $R_s(I) : \bigoplus_{n\geq 0} I^{(n)}$ is Noetherian, then $\operatorname{reg}(I^{(n)})$ is a quasi-polynomial ([15, Theorem 4.3]). Moreover, $\lim_{n\to\infty} \left(\frac{\operatorname{reg}(I^{(n)})}{n}\right)$ exists and can even be irrational [13]. For a nice survey article on Castelnuovo-Mumford regularity see [9].

Bocci and Harbourne showed showed that if $I$ is a zero dimensional subscheme in a projective space, then $\alpha(I)/\gamma(I) \leq \rho(I) \leq \operatorname{reg}(I)/\gamma(I)$ [5, Theorem 1.2.1]. Hence, if $\alpha(I) = \operatorname{reg}(I)$, then $\rho(I) = \alpha(I)/\gamma(I)$. Later, Harbourne and Huneke raised the following Conjecture:

CONJECTURE 3.6 *[34, Conjecture 2.1] Let $I$ be an ideal of fat points in $S$ and $\mathfrak{m} = (X_0, \ldots, X_d)$. Then $I^{(nd)} \subseteq \mathfrak{m}^{n(d-1)} I^n$ holds true for all $I$ and $n$.*

In the same paper they showed that the conjecture is true for fat point ideals aris-

ing as symbolic powers of radical ideals generated in a single degree in $\mathbb{P}^2$. Recently, there has been a renewed interest on the Waldschmidt constant mainly due to the containment problem. In [2] the Waldschmidt constant for square free monomial ideals was computed. In fact they showed that if $\gamma(I)$ can be expressed as the value to a certain linear program arising from the structure of the associated primes of $I$. The Waldschmidt constant has also been computed for Stanley-Risner ideals [3].

The resurgence and the Waldschmidt constant has been studied in a few cases: for certain general points in $\mathbb{P}^2$ [4], smooth subschemes [33], fat linear subspaces [25], special point configurations [20] and monomial ideals [2].

We now briefly state our results on resurgence, Waldschmidt constant and Castelnuovo-Mumford regularity. Putting weights on monomial curves $\mathcal{C}(\underline{a})$ in $\mathbb{A}^d$, we can consider them as weighted points in a weighted projective space $\mathbb{P}_{\Bbbk}(\underline{a})$. Hence the bounds for resurgence in [5] hold true. For $\underline{a} = (3, 3+m, 3+2m)$ these invariants have been computed in [18]. For $q \geq 1$, $\gcd(2q+1, m) = 1$ and $\underline{a} = (2q+1, 2q+1+m, 2q+1+2m)$, these invariants have been computed in [17]. In these cases the generators of the symbolic powers of $\mathfrak{p}(\mathcal{C}(\underline{a}))$ has been computed. In [19], for monomial curves $\mathfrak{p}(\overline{\mathcal{C}}(\underline{a}))$ in $\mathbb{P}^3$, where $\underline{a} = (m, 2m, 2q+1+2m)$, $q, m$ are positive integers and $\gcd(2q+1, m) = 1$, these invariants have been computed.

## References


[1] T. Bauer, S. Di Rocco, B. Harbourne, M. Kapustka, A. Knutsen, W. Syzdek and T. Szemberg, *A primer on Seshadri constants.* Interactions of classical and numerical algebraic geometry, 33-70, Contemp. Math., **496**, Amer. Math. Soc., Providence, RI, 2009.

[2] C. Bocci, S. Cooper, E. Guardo, Elena, B. Harbourne, M. Janssen, U. Nagel, Uwe, A. Seceleanu, A. Van Tuyl, Adam and T. Vu, Thanh, *The Waldschmidt constant for squarefree monomial ideals.* J. Algebraic Combin. **44** (2016), no. 4, 875-904.

[3] C. Bocci, Cristiano and B. Franci, *Waldschmidt constants for Stanley-Reisner ideals of a class of simplicial complexes.* J. Algebra Appl. **15** (2016), no. 7, 1650137, 13 pp.

[4] C. Bocci and B. Harbourne, *The resurgence of ideals of points and the containment problem.* Proc. Amer. Math. Soc. **138** (2010), no. 4, 1175-1190.

[5] C. Bocci and B. Harbourne, *Comparing powers and symbolic powers of ideals*, J. Algebraic Geom. **19** (2010), no. 3, 399-417.

[6] M. Boratyński, *A note on set-theoretic complete intersection ideals.* J. Algebra **54** (1978), no. 1, 1-5.

[7] H. Bresinsky, H. Monomial space curves in $\mathbb{A}^3$ as set-theoretic complete intersections. Proc. Amer. Math. Soc. **75** (1979), no. 1, 23-24.

[8] H. Bresinsky, *Monomial Gorenstein curves in $\mathbb{A}^4$ as set-theoretic complete intersections.* Manuscripta Math. **27** (1979), no. 4, 353-358.

[9] M. Chardin, *Some results and questions on Castelnuovo-Mumford regularity. Syzygies and Hilbert functions,* 1-40, Lect. Notes Pure Appl. Math., 254, Chapman & Hall/CRC, Boca Raton, FL, 2007.

[10] R. C. Cowsik, *Symbolic powers and number of defining equations.* Algebra and its applications (New Delhi, 1981), 13-14, Lecture Notes in Pure and Appl. Math., 91, Dekker, New York, 1984.

[11] R. C. Cowsik and M. V. Nori, *Affine curves in characteristic* p *are set theoretic complete intersections.* Invent. Math. **45** (1978), no. 2, 111-114.

[12] S. D. Cutkosky, *Symbolic algebras of monomial primes.* J. Reine Angew. Math. **416** (1991), 71-89.

[13] S. D. Cutkosky, *Irrational asymptotic behaviour of Castelnuovo-Mumford regularity.* J. Reine Angew. Math. **522** (2000), 93-103.

[14] S D. Cutkosky, J. Herzog, and N. V. Trung, *Asymptotic behaviour of the Castelnuovo-*

*Mumford regularity.* Compositio Math. **118** (1999), no. 3, 243-261.
[15] S. D. Cutkosky and K. Kurano, *Asymptotic regularity of powers of ideals of points in a weighted projective plane.* Kyoto J. Math. **51** (2011), no. 1, 25-45.
[16] H. Dao, A. De Stefani, E. Grifo, C. Huneke and L. Núñez-Betancourt, *Symbolic powers of ideals. Singularities and foliations. geometry*, Topology and applications, 387-432, Springer Proc. Math. Stat., 222, Springer, Cham, 2018.
[17] C. D'Cruz, *Resurgence and Castelnuovo-Mumford regularity of certain monomial curves in* $\mathbb{A}^3$. To appear in Acta Mathematica Vietnamica
[18] C. D'Cruz and S. Masuti, *Symbolic Blowup algebras and invariants of certain monomial curves in an affine space.* To appear in Comm. in Algebra.
[19] C. D'Cruz and M. Mandal, *Symbolic blowup algebras and invariants associated to certain monomial curves in* $\mathbb{P}^3$. To appear in Comm. in Algebra.
[20] M. Dumnicki, B. Harbourne, U. Nagel, A. Seceleanu, T. Szemberg and H. Tutaj-Gasińska, *Resurgences for ideals of special point configurations in* $\mathbb{P}^N$ *coming from hyperplane arrangements.* J. Algebra **443** (2015), 383-394.
[21] M. Dumnicki, and H. Tutaj-Gasińska, *A containment result in* $\mathbb{P}^n$ *and the Chudnovsky conjecture.* Proc. Amer. Math. Soc. **145** (2017), no. 9, 3689-3694.
[22] L. Ein, R. Lazarsfeld, and K. E. Smith, *Uniform bounds and symbolic powers on smooth varieties.* Invent. Math. **144** (2001), no. 2, 241-252.
[23] D. Eisenbud, *Commutative algebra. With a view toward algebraic geometry.* Graduate Texts in Mathematics, 150. Springer-Verlag, New York, 1995.
[24] D. Eisenbud and E. G. Evans *Every algebraic set in n-space is the intersection of n hypersurfaces.* Invent. Math. **19** (1973), 107-112.
[25] G. Fatabbi, Giuliana, B. Harbourne and A. Lorenzini, *Inductively computable unions of fat linear subspaces.* J. Pure Appl. Algebra **219** (2015), no. 12, 5413-5425.
[26] D. Ferrand, *Courbes gauches et fibrés de rang 2.* C. R. Acad. Sci. Paris Sér. A-B 281 (1975), no. 10, Aii, A345-A347.
[27] L. Fouli, Louiza, P. Mantero, Paolo and Y. Xie, *Chudnovsky's conjecture for very general points in* $\mathbb{P}^N_{\Bbbk}$ J. Algebra **498** (2018), 211-227.
[28] S. Goto, *The Cohen-Macaulay symbolic Rees algebras for curve singularities.* The Cohen-Macaulay and Gorenstein Rees algebras associated to filtrations. Mem. Amer. Math. Soc. **110** (1994), no. 526, 1-68.
[29] S. Goto, K. Nishida, and Y. Shimoda, *Topics on symbolic Rees algebras for space monomial curves.* Nagoya Math. J. **124** (1991), 99-132.
[30] S. Goto, K. Nishida, Y. Shimoda, *The Gorensteinness of symbolic Rees algebras for space curves.* J. Math. Soc. Japan **43** (1991), no. 3, 465-481.
[31] S. Goto, K. Nishida, K. Watanabe, *Non-Cohen-Macaulay symbolic blow-ups for space monomial curves and counterexamples to Cowsik's question.* Proc. Amer. Math. Soc. **120** (1994), no. 2, 383-392.
[32] S. Greco and R. Strano *Complete Intersections* Lectures given at the 1st 1983 Session of the Centro Internationale Matematico Estivo (C.I.M.E.) held at Acireale (Catania), Italy, June 13-21, 1983.
[33] E. Guardo, B. Harbourne and A. Van Tuyl, *Asymptotic resurgences for ideals of positive dimensional subschemes of projective space.* Adv. Math. **246** (2013), 114-127.
[34] B. Harbourne and C. Huneke, *Are symbolic powers highly evolved?* J. Ramanujan Math. Soc. **28A** (2013), 247-266.
[35] R. Hartshorne, *Ample subvarieties of algebraic varieties.* Notes written in collaboration with C. Musili. Lecture Notes in Mathematics, Vol. 156 Springer-Verlag, Berlin-New York 1970 xiv+256 pp.
[36] R. Hartshorne, *Complete intersections in characteristic* $\mathtt{p} > 0$. Amer. J. Math. **101** (1979), no. 2, 380-383.
[37] J. Herzog: *Generators and relations of abelian semigroups and semigroup rings.* Manuscripta Math. 3 (1970) 175-193.
[38] M. Hochster and C. Huneke, *Comparison of symbolic and ordinary powers of ideals.* Invent. Math. **147** (2002), no. 2, 349-369.

[39] C.Huneke On the finite generation of symbolic blow-ups. Mathematische Zeitschrift **179** (1982) 465-472
[40] C. Huneke, *Hilbert functions and symbolic powers.* Michigan Math. J. **34** (1987), 293-318.
[41] G. Knödel, P. Schenzel and R. Zonsarow, *Explicit computations on symbolic powers of monomial curves in affine space.* Comm. Algebra **20** (1992), no. 7, 2113-2126.
[42] V. Kodiyalam, *Asymptotic behaviour of Castelnuovo-Mumford regularity.* Proc. Amer. Math. Soc. **128** (2000), no. 2, 407-411.
[43] L. Kronecker, *Zur Theorie der Abelschen Gleichungen.* J. Reine Angew. Math. **92**, 1-123 (1882).
[44] G. Lyubeznik, *A survey of problems and results on the number of defining equations.* Commutative algebra (Berkeley, CA, 1987), 375-390, Math. Sci. Res. Inst. Publ., 15, Springer, New York, 1989.
[45] G. Lyubeznik, *The number of defining equations of affine algebraic sets.* Amer. J. Math. **114** (1992), no. 2, 413-463.
[46] G. Malara, T. Szemberg, and J. Szpond, *On a conjecture of Demailly and new bounds on Waldschmidt constants in* $\mathbb{P}^N$. J. Number Theory **189** (2018), 211-219.
[47] T. T. Moh, *A result on the set-theoretic complete intersection problem.* Proc. Amer. Math. Soc. **86** (1982), no. 1, 19-20.
[48] T. T. Moh *Set-theoretic complete intersections.* Proc. Amer. Math. Soc. **94** (1985), no. 2, 217-220.
[49] N. Mohan Kumar, *On two conjectures about polynomial rings.* Invent. Math. **46** (1978), no. 3, 225-236.
[50] M. Morales, *Noetherian symbolic blow-ups.* J. Algebra **140** (1991), no. 1, 12-25.
[51] M. Nagata. Local rings Interscience Tracts in Pure and Applied Mathematics, No. **13** Interscience Publishers a division of John Wiley & Sons. New York-London 1962 xiii+234 pp.
[52] M. Nagata *Lectures on the fourteenth problem of Hilbert.* Tata Institute of Fundamental Research, Bombay 1965 ii+78+iii pp.
[53] D. Patil, *Certain monomial curves are set-theoretic complete intersection.* Manuscripta Math. **68**, (1990) 399-404.
[54] D. Rees, D. *On a problem of Zariski.* Illinois J. Math. **2** 1958 145-149.
[55] Robbiano L., Valla G. *Some curves in* $\mathbb{P}^3$ *are set-theoretic complete intersections.* Algebraic geometry-open problems (Ravello, 1982), 391-399, Lecture Notes in Math., 997, Springer, Berlin-New York, 1983.
[56] P. Roberts, *A prime ideal in a polynomial ring whose symbolic blow-up is not Noetherian.* Proc. Amer. Math. Soc. **94** (1985), no. 4, 589-592.
[57] Şahin, Mesut, *Producing set-theoretic complete intersection monomial curves in* $\mathbb{P}^n$. Proc. Amer. Math. Soc. **137** (2009), no. 4, 1223-1233.
[58] P. Schenzel, Examples of Noetherian symbolic blow-up rings, Rev. Rom. Math Pures et appl. **33** (1988) 375-383.
[59] H. Srinivasan, *On finite generation of symbolic algebras of monomial primes.* Comm. Algebra **19** (1991), no. 9, 2557-2564.
[60] U. Storch, *Bemerkung zu einem Satz* von M. Kneser. Arch. Math. (Basel) **23** (1972), 403-404.
[61] J. Stückrad and W. Vogel, *On the number of equations defining an algebraic set of zeros in n-space.* Seminar D. Eisenbud/B. Singh/W. Vogel, Vol. 2, pp. 88-107, Teubner-Texte zur Math., 48, Teubner, Leipzig, 1982.
[62] I. Swanson, *Powers of ideals. Primary decompositions, Artin-Rees lemma and regularity.* Math. Ann. **307** (1997), no. 2, 299-313.
[63] I. Swanson, *Linear equivalence of ideal topologies.* Math. Z. **234** (2000), no. 4, 755-775.
[64] L. Szpiro, "Lectures on equations defining space curves," Notes by N. Mohan Kumar. Tata Institute of Fundamental Research, Bombay; by Springer-Verlag, Berlin-New York, (1979).
[65] G. Valla, *On determinantal ideals which are set-theoretic complete intersections.*

Compositio Math. **42** (1980/81), no. 1, 3-11.

[66] M. Waldschmidt, *Propriétés arithmétiques de fonctions de plusieurs variables. II.* (French) Séminaire Pierre Lelong (Analyse) année 1975/76, pp. 108-135. Lecture Notes in Math., Vol. 578, Springer, Berlin, 1977.

[67] O Zariski. *A fundamental lemma from the theory of holomorphic functions on an algebraic variety.* Ann. Mat. Pura Appl. (4) **29** (1949), 187-198.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 85–102

# *Differential Equations and Monodromy*

**T.N.Venkataramana**

*School of Mathematics, TIFR, Homi Bhabha Road, Colaba, Mumbai 400005, India*

**Abstract:** In these expository notes, we describe results of Cauchy, Fuchs and Pochhammer on differential equations. We then apply these results to hypergeometric differential equation of type ${}_nF_{n-1}$ and describe Levelt's theorem determining the monodromy representation explicitly in terms of the hypergeometric equation. We also give a brief overview, without proofs, of results of Beukers and Heckman, on the Zariski closure of the monodromy group of the hypergeometric equation. In the last section, we recall some recent results on thin-ness and arithmeticity of hypergeometric monodromy groups

## 1. Introduction

In these notes, we recall (in sections 1 to 3) the basic theory of differential equations on the unit disc and on the punctured unit disc. For references see [6] and [8].

In sections 4 and 5 we apply the theory developed in sections 1 through 3 to (state and) prove a result of Levelt [8] on the monodromy of hypergeometric differential equations of type ${}_nF_{n-1}$.

In the next few sections we use this description to prove some results (some are not proved completely because the proofs are lengthy) of Beukers and Heckman [3] on the Zariski closure of the monodromy of the foregoing hypergeometric equation. In particular, they completely determine when the monodromy is finite.

**Acknowledgments** I thank Professors Amarnath and Padmavati for inviting me to contribute an article to the proceedings of the Telangana Academy of Sciences. I also thank Professors F.Beukers, Madhav Nori and P.Sarnak for very helpful conversations related to the material presented in the paper. I am grateful to the JC Bose fellowship for the period 2019-2023 and the Max Planck Institute for Mathematics in Bonn,Germany for its hospitality and financial support while this work was prepared for publication.

*Corresponding author. Email: venky@math.tifr.res.in

## 2. Monodromy Groups

The concept of monodromy arises in many seemingly different situations. We will deal with some of the simplest ones, namely the monodromy associated to linear differential equations on open subsets in the complex plane.

### 2.1. *Differential Equations on Open Sets in the Plane*

Let $U$ be a connected open subset of the complex plane. Fix holomorphic functions $f_i : U \to \mathbb{C}$ with $0 \leq i \leq n-1$. Consider the differential equation

$$\frac{d^n y}{dz^n} + \sum_{i=0}^{n-1} f_i(z) \frac{d^i y}{dz^i} = 0.$$

If $y_1, y_2$ are solutions, then so is $c_1 y_1 + c_2 y_2$ with $c_1, c_2 \in \mathbb{C}$. That is, the space of solutions is a vector space. A fundamental result of Cauchy says that when $U$ is the unit disc, there are holomorphic functions $y_1, \cdots, y_n$ on the disc which are solutions of this differential equation, which are linearly independent and such that all solutions are linear combinations of these solutions. These solutions are called **fundamental solutions**.

THEOREM 2.1 *(Cauchy) Let $f_0, \cdots, f_{n-1} : \Delta \to \mathbb{C}$ be holomorphic functions on the open unit disc $\Delta$ and consider the differential equation*

$$\frac{d^n y}{dz^n} + f_{n-1}(z) \frac{d^{n-1} y}{dz^{n-1}} + \cdots + f_1(z) \frac{dy}{dz} + f_0(z) y = 0.$$

*Suppose that $z_0, \cdots, z_{n-1}$ are arbitrary complex numbers. Then there exists a solution $y$ to the differential equation, which is holomorphic in the whole of the disc, such that $\frac{d^j y}{dz^j}(0) = z_j$ for all $j$ with $0 \leq j \leq n-1$.*

*In particular, the differential equation has $n$ fundamental solutions.*

*Proof.* We first prove this when $n = 1$. Suppose then that we have the equation

$$\frac{dy}{dz} + f_0(z) y = 0,$$

where $f_0(z) = \sum_{k=0}^{\infty} a_k z^k$ a power series which converges in $\mid z \mid < 1$. Suppose $y(z) = \sum_{k=0}^{\infty} x_k z^k$ is a formal power series with $x_k$ a sequence of elements of $\mathbb{C}$. By looking at the coefficient of $z^{k-1}$ on both sides (which are formal power series) of the differential equation, it follows that, if the formal power series $y$ is to be a solution of the differential equation, then the $x_k$ (for $k \geq 1$) must satisfy the recursive relation

$$-k x_k = x_{k-1} a_0 + x_{k-2} a_1 + \cdots + x_0 a_{k-1}. \tag{1}$$

Let $R < 1$; then the convergence of $f_0(z)$ in $\mid z \mid < 1$ implies that there is a constant $M \geq 1$ such that $\mid a_k \mid R^k < M$ for all $k \geq 0$. Suppose $r < R$ is fixed. Let, for each $j$, $M_j$ denote the supremum

$$M_j = sup\{\mid x_j \mid R^j, \mid x_{j-1} \mid R^{j-1}, \cdots, \mid x_1 \mid R, \mid x_0 \mid\}.$$

The equation (1) shows that for each $k \geq 1$ we have

$$k \mid x_k \mid \leq \frac{M_{k-1}}{R^{k-1}}M + \frac{M_{k-1}}{R^{k-2}}\frac{M}{R} + \cdots + \frac{M_{k-1}}{R}\frac{M}{R^{k-2}} + M\frac{M}{R^{k-1}} = k\frac{M_{k-1}M}{R^{k-1}}.$$

Therefore, $\mid x_k \mid R^k \leq M_{k-1}M$ for all $k \geq 0$. In particular, since by assumption $M \geq 1$, we have $M_k \leq M_{k-1}M$ and hence $\frac{M_k}{M^k}$ is a decreasing sequence, and hence a bounded sequence. We may assume (increasing $M$ if necessary), that $M_k \leq MM^k$ for all $k$. Therefore, $\mid x_k \mid r^k \leq \frac{MM^k}{R^k}r^k$. Therefore, if $\frac{Mr}{R} < 1$ then $\sum \mid x_k \mid r^k$ is dominated by the convergent geometric series $M\sum(\frac{Mr}{R})^k$. Hence the formal power series $\sum x_k z^k$ converges in the smaller disc $\mid z \mid < \frac{R}{M}$.

We may similarly solve the differential equation in every small disc inside the unit disc $\Delta$ ( as a convergent power series around $z_0 \in \Delta$) and by the uniqueness of the power series - thanks to the recursion (1) - the two power series coincide as functions on the intersections of the smaller discs. Therefore, by the principle of analytic continuation, there is a holomorphic function $y$ on all of the disc which is a solution of the differential equation $\frac{dy}{dz} + f_0(z)y = 0$.

Exactly the same proof shows that if now $y : \Delta \to \mathbb{C}^n$ has values in a vector space, and $f_0$ is replaced by a matrix valued *holomorphic* function $A(z) : \Delta \to M_n(\mathbb{C})$ (i.e. an $M_n(\mathbb{C})$-valued convergent holomorphic function on $\Delta$) then there is a power series $y$ with coefficients $x_k$ in $\mathbb{C}^n$ which is a solution of the differential equation $\frac{dy}{dz}+A(z)(y(z)) = 0$ and which converges on all of $\Delta$.

Suppose $y_1, y_2 : \Delta \to M_n(\mathbb{C})$ are two solutions, with $y_1(0) = Id_n$ and $y_2(0) = x_0 \in GL_n(\mathbb{C})$. The solution $y_2$ is uniquely determined by its constant term $x_0 \in M_n(\mathbb{C})$. It follows that $y_2$ is the product of the matrix valued function $y_1$ and the constant matrix $x_0$:

$$y_2 = y_1 x_0.$$

This establishes that every vector valued solution $y : \Delta \to \mathbb{C}^n$ to the equation $\frac{dy}{dz} = A(z)y$ is a linear combination of the rows of the matrix $y_1$.

We now choose

$$A(z) = \begin{pmatrix} 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & -1 \\ f_0(z) & f_1(z) & f_2(z) & \cdots & f_{n-1}(z) \end{pmatrix} \quad and \quad y = \begin{pmatrix} y_1(z) \\ y_2(z) \\ \cdots \\ y_n(z) \end{pmatrix},$$

where $y$ is viewed as a column vector. Then the vector valued equation $\frac{dy}{dz} + A(z)y(z) = 0$ yields the $n$ scalar valued equations $y_1'(z) = y_2(z), \cdots y_{n-1}'(z) = y_n(z)$, and $y_n'(z) + f_0(z)y_1(z) + \cdots + f_{n-1}(z)y_n(z) = 0$. In other words, $w = y_1(z)$ is the solution to the scalar valued differential equation

$$\frac{d^n w}{dz^n} + f_{n-1}(z)\frac{d^{n-1}w}{dz^{n-1}} + \cdots f_1(z)\frac{dw}{dz} + f_0(z)w = 0.$$

This proves Cauchy's theorem in all cases. ■

If $U$ is now taken to be any connected open set in $\mathbb{C}$, then the foregoing result of Cauchy says that at each point $p$ of the open set, there are holomorphic functions $y_1, \cdots, y_n$ defined in an open neighborhood of $p$ which are fundamental solutions to the above differential equation. If $\Gamma$ denotes a closed loop in the open set $U$ starting and ending at $p$, then analytic continuation of the solutions along the path $\Gamma$ is possible and when we return to the original point, we get new fundamental solutions $w_1, \cdots, w_n$. This means that there is a matrix $M = M(\gamma)$ depending on the path, such that $w = My$ in a neighborhood of $p$. One can check that the matrix $M$ depends only on the homotopy class of the path $\gamma$ based at $p$ and not on the path itself.

Moreover, if $\gamma_1, \gamma_2$ are *two* paths based at $p$, and $\gamma$ is the composition of these paths, then one can check that $M(\gamma) = M(\gamma_1)M(\gamma_2)$. Thus, the association $\gamma \to M(\gamma)$ yields a group homomorphism from the fundamental group of the open set $U$ based at $p$, into $GL_n(\mathbb{C})$. This homomorphism is called the "monodromy representation" and the image is called the "monodromy group".

## 2.2. *Finiteness*

Let $U \subset \mathbb{C}$ be a connected open set. Then the space of holomorphic functions on $U$ is an integral domain and the corresponding field of of fractions, i.e. ratios of holomorphic functions, is a field, called the field of meromorphic functions on $U$. Let $U^* \to U$ (given by $\tau \mapsto z$) be the universal cover $U^*$ of $U$ and let $\Gamma$ be the deck transformation group.

We say that a function $f : U^* \to \mathbb{C}$ is **algebraic** if it satisfies a polynomial relation $f^n(\tau)+\sum_{i=0}^{n-1} \phi_j(z)f^i(\tau) = 0$ with coefficients $\phi_j$ in the field $K$ of meromorphic functions on $U$.

LEMMA 2.2 *A function $f$ on $U^*$ is algebraic if and only if its orbit under the deck transformation group $\Gamma$ is finite.*

*Proof.* The polynomial relation holds if $f$ is replaced by any translate under an element $\gamma \in \Gamma$. But since there are only finitely many roots to any polynomial, it follows that the orbit of $f$ under $\Gamma$ is finite.

On the other hand, if a function $f$ on $U^*$ is invariant under $\Gamma$, then it defines a holomorphic function on the base $U$. Therefore, if the orbit under $\Gamma$ is finite, then the polynomial $P(t) = \prod_{\gamma \in \Gamma/\Gamma_f}(t - \gamma(f))$ has coefficients in $K$. Hence $f$ is algebraic. ∎

COROLLARY 2.3 *Suppose*

$$\frac{d^n y}{dz^n} + f_{n-1}(z)\frac{d^{n-1}y}{dz^{n-1}} + \cdots + f_1\frac{dy}{dz} + f_0 y = 0$$

*is a differential equation with coefficients $f_i$ holomorphic on $U$. Suppose that the monodromy representation is irreducible. Then a nonzero solution to the equation is algebraic if and only if the monodromy is finite.*

*Proof.* If $f$ is a solution, then so is $\gamma(f)$ for $\gamma \in \Gamma$. If the monodromy is finite, it means that the orbit of $f$ under $\Gamma$ is finite, in particular, and hence it is algebraic by the lemma.

On the other hand, if some solution $f$ is algebraic, then by the lemma the orbit is finite. It means that the $\Gamma$ translates of $f$ span a subspace which is $\Gamma$ stable and these translates are algebraic. By irreducibility, this is the whole space. This means that for some basis of the space of solutions, the orbit of $\Gamma$ is finite for every element of the basis. This means that the image under the monodromy representation of $\Gamma$ is finite. ■

## 3. Punctured Disc

Now consider the open set $U = \Delta^*$, obtained by removing the point 0 from the unit disc $\Delta$.

*Example* 1 Let us look at the differential equation

$$\frac{dy}{dz} - \frac{\alpha}{z}y = 0,$$

where $\alpha \in \mathbb{C}$ is fixed. Solving, we get $y = z^\alpha$. This function is not "single valued". We view $z = e^{2\pi i\tau} = \phi(\tau)$ with $\tau \in \mathfrak{h}$ the upper half plane, with $\phi$ being a covering map. Consider the path $\omega : [0,1] \to \mathfrak{h}$ starting at $i$ and ending at $i+1$. Its composite $\gamma = \phi \circ \omega$ is a closed loop in $\Delta^*$ based at $p = e^{-2\pi}$; the effect of traversing along this path on the solution $y$ is to multiply it by $e^{2\pi i\alpha}$. Thus $M(\gamma)$ is the $1 \times 1$ matrix $e^{2\pi i\alpha}$.

*Example* 2 As another example, consider the equation

$$\frac{d^2y}{dz^2} = -\frac{1}{z}\frac{dy}{dz}.$$

Clearly the constant function $y_1 = 1$ is a solution; it is invariant under the action of the loop $\gamma$.

It is easily checked that $y_2 = \frac{1}{2\pi i}logz$ is another solution; to view this solution as a function, we write $y = \frac{1}{2\pi i}log(e^{2\pi i\tau}) = \frac{1}{2\pi i}2\pi i\tau = \tau$; hence the action of the loop $\gamma$ of the preceding example, is to take $y_2$ into the new solution $y_2 + \frac{1}{2\pi i}2\pi i = y_2 + y_1$. Hence

$$M(\gamma) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

### 3.1. *Regular Singular Points and a Theorem of Fuchs*

We now look at a more general case; suppose we have a differential equation of the form

$$\frac{d^ny}{dz^n} + f_{n-1}(z)\frac{d^{n-1}y}{dz^{n-1}} + \cdots + f_1(z)\frac{dy}{dz} + f_0(z)y, \tag{2}$$

where each $f_i(z)$ has at most a pole of order $n-i$ at $z = 0$. Then the monodromy $M$ (i.e. the action of the generator of $\pi_1(\Delta^*) \simeq \mathbb{Z}$) acts on the $\mathbb{C}^n$ space of solutions.

Write $\theta = z\frac{d}{dz}$; using the relation $\theta^2 = z^2\frac{d^2}{dz^2} + \theta$ one can show, by induction, that

$$z^k \frac{d^k}{dz^k} = \theta(\theta-1)\cdots(\theta-k+1)$$

for every $k \geq 1$. Then the differential equation (after multiplying throughout by $z^n$), takes the form

$$\theta(\theta-1)\cdots(\theta-n+1)y + zf_{n-1}(\theta(\theta-1)\cdots(\theta-n+1)y +$$

$$+\cdots + z^{n-2}f_2\theta(\theta-1)y + z^{n-1}f_1\theta y + z^n f_0 y = 0.$$

Rewriting this yields

$$\theta^n y + F_{n-1}(z)\theta^{n-1} + F_{n-2}(z)\theta^{n-2}y + \cdots + F_1(z)\theta y + F_0(z)y = 0$$

where now the functions $F_i$ are holomorphic on all of the disc (including the puncture). Write $a_i = F_i(0)$ and $f(t) = t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0 = \prod_{j=1}^n (t-\alpha_j)$. The equation $f(t) = 0$ is called the **indicial equation** and the roots $\alpha_1, \cdots, \alpha_n$ of the indicial equation (i.e. roots of the polynomial $f$) are called the indicial roots.

THEOREM 3.1 *(Fuchs) With the preceding notation, assume that* 0 *is a regular singular point of the differential equation (2). Then, the characteristic polynomial of the monodromy matrix $M$ of the differential equation (2) is the polynomial*

$$\prod_{j=1}^{n}(t - e^{2\pi i\alpha_j}).$$

*Moreover, every solution of the differential equation (2) is a linear combination of functions of the form $\phi(z)z^\alpha P(logz)$ where $\phi$ is a holomorphic function on all of the disc, $\alpha$ is a complex number and $P$ is a polynomial.*

Theorem 3.1 will be recast in terms of matrix valued solutions and the differential equation (2) will be rewritten as a *first order* equation.

## 3.2. *First order Matrix Valued Differential Equations*

Suppose now that $A : \Delta \to M_n(\mathbb{C})$ is a holomorphic map on all of the disc. Then $A(z)$ is represented by a convergent power series $A(z) = \sum A_k z^k$ where $A_k \in M_n(\mathbb{C})$. We look for local solutions $Y : \Delta^* \to M_n(\mathbb{C})$ to the first order equation $z\frac{dY}{dz} = A(z)Y(z)$.

*Notation* If $T \in M_n(\mathbb{C})$ and $z \in \Delta^*$ we write $z^T$ for the matrix represented by the (convergent) exponential power series in the matrix variable $(logz)T$:

$$z^T = exp((logz)T) = \sum_{k=0}^{\infty} \frac{(logz)^k T^k}{k!}$$

We list some properties of the matrix exponent.

[1] If $A$ and $B$ are commuting square matrices of size $n$, then $z^{A+B} = z^A z^B$.
[2] If $A \in M_n(\mathbb{C})$ and $g \in GL_n(\mathbb{C})$, then $z^{gAg^{-1}} = gz^A g^{-1}$.
[3] If $N \in M_n(\mathbb{C})$ is nilpotent, then $z^N$ is a polynomial in $logz$.
[4] If $A \in M_n(\mathbb{C})$ is a diagonal matrix whose diagonal entries are $a_1, a_2, \cdots, a_n$ then $z^A$ is also a diagonal matrix whose diagonal entries are $z^{a_1}, z^{a_2}, \cdots, z^{a_n}$.
[5] These properties imply that if $A \in M_n(\mathbb{C})$ is any matrix, then the entries of the matrix $z^A$ are linear combinations of functions of the form $z^\alpha P(logz)$ where $P$ is a polynomial and $\alpha \in \mathbb{C}$ is a fixed complex number.
[6] The derivative of $z^A$ satisfies: $z\frac{dz^A}{dz} = z^A A$.
[7] The monodromy operator on the multivalued function $z^A$ is simply $e^{2\pi i A}$, since $z^A = e^{2\pi i \tau A}$ and the generator of the Deck transformation group takes $\tau$ to $\tau + 1$.
For a reference to the following see [6], Theorem (4.1) and Theorem (4.2).

THEOREM 3.2 *(Fuchs) Suppose $A : \Delta \to M_n(\mathbb{C})$ is a holomorphic function. Let $\mathfrak{h} \to \Delta^*$ be the exponential covering map as before. Consider the differential equation in $Y(z) = Y^*(\tau) \in M_n(\mathbb{C})$:*

$$\frac{dY}{dz} = \frac{A(z)}{z} Y. \tag{3}$$

*The monodromy of the equation acts on the space of solutions $Y^*$ by the formula $Y^*(\tau + 1) = Y^*(\tau)M$ where $M \in GL_n(\mathbb{C})$. Moreover, the semi-simple part of the matrix $M$ is conjugate to the exponential $e^{2\pi i A_s}$ of $A_s$, where $A_s$ is the semi-simple part of the matrix $A_0 = A(0)$.*

*Proof.* First assume that if $\lambda, \mu$ are *distinct* eigenvalues of the matrix $A_0$, then they do not differ by an integer. This means that no eigenvalue of the adjoint transformation $adA_0$ can be a nonzero integer. We then show that there is a holomorphic function $X : \Delta \to GL_n(\mathbb{C})$ such that

$$Y(z) = X(z) z^{A_0}$$

is a solution of the differential equation (3). Write $X(z) = \sum_{k=0}^{\infty} X_k z^k$, and solve for the coefficients $X_k$. Write, as before, $\theta = z\frac{d}{dz}$. Then the differential equation for $Y$ is $\theta Y(z) = A(z)Y(z)$; moreover, by the formula for the differentiation for a product, we get

$$\theta Y(z) = \theta(X(z)z^{A_0}) = \theta(X(z))z^{A_0} + X(z)z^{A_0}A_0 =$$

$$= A(z)Y(z) = A(z)X(z)z^{A_0}.$$

We now cancel $z^{A_0}$ on both sides of the preceding equation and obtain

$$\theta(X(z)) + X(z)A_0 = A(z)X(z),$$

where now $A$ is holomorphic on all of the disc and $X$ is assumed to be holomorphic on all of the disc. Writing the power series for $X$ and $A$, we then get, for $k \geq 1$,

the recursion

$$kX_k + X_kA_0 = A_0X_k + \sum_{j=0}^{k-1} A_{k-j}X_j,$$

and for $k = 0$, the equation $X_0A_0 = A_0X_0$. We can solve for $X_0$ by taking $X_0$ to be identity. The recursion for the coefficients is

$$(k - adA_0)X_k = \sum_{j=0}^{k-1} A_{k-j}X_j.$$

This can be solved for all $k \geq 1$ since, by assumption, *non-zero integers k cannot be eigenvalues of the operator* $adA_0$; therefore, $k - adA_0$ is an invertible operator and hence $X_k$ may be written as a combination of the $X_j : j \leq k-1$.

[ We now check that the formal power series $\sum X_k z^k$ converges in a small enough neighborhood of 0. Consider the sequence $1 - \frac{adA_0}{k}$ for $k \geq 1$. For $k$ large enough, the $k$-Th term of this sequence is close to the identity matrix; by assumption, all the terms of this sequence are non-singular. Hence the sequence $(1 - \frac{adA_0}{k})^{-1}$ is bounded from above by a constant $M > 1$ say. Since the sequence $A_{k-j}R^{k-j}$ $(k \geq j)$ is bounded, we may assume that $\mid A_{k-j} \mid R^{k-j} \leq M$ for all $k, j$. Let, as in the proof of Cauchy's theorem, $M_k$ be the supremum of the matrix norms $\mid X_j \mid R^j$ for $j \leq k$. The recursive relation for the $X_k$ now implies that

$$k \mid X_k \mid \leq MMk_{k-1}.$$

Therefore, $\mid X_k \mid R^k \leq M^2M_{k-1}$. Since $M \geq 1$, we also have $\mid X_j \mid R^j \leq M^2M_{k-1}$ for all $j \leq k-1$. Hence $M_k \leq M^2M_{k-1}$ and therefore, $M_km^{-2k}$ is a decreasing sequence and is bounded. We may assume then that $\mid X_k \mid R^k \leq M_k \leq MM^{2k}$ and hence $\sum X_k z^k$ converges if $\mid z \mid < \frac{R}{M^2}$.]

Thus the monodromy action on $Y(z) = X(z)z^{A_0}$ is simply right multiplication by the exponential matrix $e^{2\pi iA_0}$ of $A_0$ since the solution $X(z)$ is holomorphic also at the puncture and is invariant under the monodromy action. This proves the Theorem in the case when distinct eigenvalues of $A_0$ remain distinct modulo 1.

The proof of the general case of the Theorem can be reduced to this case. Fix an eigenvalue $\lambda$ of the linear transformation $A_0$. Write $\mathbb{C}^n = E \oplus F$ where $E$ is the generalized $\lambda$ airspace for $A_0$, and $F$ an $A_0$ stable supplement to $E$. If $\varepsilon_1, \cdots, \varepsilon_r$ is a basis of $E$, and $\varepsilon_{r+1}, \cdots, \varepsilon_n$ a basis of $F$, then with respect to the basis $\varepsilon_1, \cdots, \varepsilon_n$ of $\mathbb{C}^n$, the matrix of the transformation which is $z$ times the identity on $E$ and identity on $F$ is given by $\begin{pmatrix} zI_r & 0 \\ 0 & I_{n-r} \end{pmatrix}$, where $I_k$ is the identity matrix of size $k$. Moreover, $A_0 = \begin{pmatrix} \lambda I_r + N_r & 0 \\ 0 & \delta_0 \end{pmatrix}$ where $\delta_0$ acts on $F$ and $N_r$ is a *nilpotent* matrix of size $r$. Write $Y(z) = \begin{pmatrix} zI_r & 0 \\ 0 & I_{n-r} \end{pmatrix} W(z) = M(z)W(z)$. Then it is easily seen that $W(z)$ satisfies the equation

$$\theta W(z) = B(z)W(z)$$

where $B(z)$ is holomorphic on $\Delta$ and $B_0 = B(0) = \begin{pmatrix} (\lambda-1)I_r + N_r & \beta_0 \\ 0 & \delta_0 \end{pmatrix}$. Thus the semi-simple part of the exponential of $B_0$ is conjugate to that of $A_0$. Moreover, the monodromy of $W$ and of $Y$ are the same. Consequently, $Y$ may be replaced by $W$ in the statement of the theorem without altering the conclusion.

We now apply the preceding repeatedly to ensure that if $\lambda$ and $\lambda'$ are two distinct eigenvalues of $A_0$ which differ by an integer, the $A_0$ is replaced by $B_0$ such that these eigenvalues become equal. That is, suppose $\lambda = \lambda' + m$ for some positive integer $m$ say. As above, we replace $\lambda$ by $\lambda - 1$ without altering the monodromy; we do this $m$ times until $\lambda$ is replaced by $\lambda'$, with monodromy unchanged.

Applying this procedure repeatedly to all eigenvalues which differ by an integer, we can thus ensure that all the distinct eigenvalues of $A_0$ remain distinct modulo 1. Then we are in the special case where $adA_0$ does not have non-zero integers as eigenvalues. In that case, the theorem has already been proved.

■

### 3.3. *Complex Reflections*

For a reference to the following, see Theorem 3.1.2 of [2].

THEOREM 3.3 (*Pochhammer*) *If as before, we have a differential equation*

$$\frac{d^n y}{dz^n} + \sum_{i=0}^{n-1} f_i(z) \frac{d^i y}{dz^i} = 0,$$

*on the punctured disc $\Delta^*$, and we assume that the functions $f_i$ have at most a simple pole at $z = 0$, then there are $n-1$ solutions which extend holomorphic ally to the puncture, and one solution which (possibly) has singularities at the puncture. Moreover, the monodromy matrix is of the form*

$$M = \begin{pmatrix} 1 & 0 & 0 & \cdots & * \\ 0 & 1 & 0 & \cdots & * \\ \cdots & \cdots & \cdots & & \\ 0 & 0 & 0 & \cdots & c \end{pmatrix}$$

*for some $c \neq 0$.*

The number $c$ is called the *exceptional eigenvalue* (it can even be 1) and the matrix $M$ is called a *complex reflection* (it is identity on a co dimension one subspace of $\mathbb{C}^n$).

*Proof.* By the same procedure as before, the differential equation of order $n$ can be converted to a differential equation of order 1 but with solutions in the vector space $\mathbb{C}^n$:

$$\frac{dy}{dz} = A(z)y(z) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \\ -f_0(z) & -f_1(z) & -f_2(z) & \cdots & -f_{n-1}(z) \end{pmatrix} (y(z)).$$

We write the vector valued formal power series expansion $y(z) = \sum x_k z^k$, with $x_k$ in $\mathbb{C}^n$. If we write $A(z) = \frac{A_{-1}}{z} + A_0 + A_1 z + \cdots + A_k z^k + \cdots$, then $A(z) - \frac{A_{-1}}{z}$ is a convergent power series (with values in $M_n(\mathbb{C})$) in $\mid z \mid < 1$. Solving term by term, for each $k \geq 0$ comparing the coefficient of $z^k$ we get (cf 1)

$$kx_k = A_{-1}x_k + A_0 x_{k-1} + A_1 x_{k-2} + \cdots + A_k x_0. \tag{4}$$

Now the first $n-1$ rows of the matrix $A_{-1}$ are all zero. the only other eigenvalue $d$ of $A_{-1}$ is the residue of $-f_{n-1}$ at 0. If $d$ is never a positive integer, then $A_{-1} - k$ is invertible for *all* positive integers $k$. Hence the above recursion shows that all the $x_k; k \geq 1$ are uniquely determined by $x_0$; the equation for $k = 0$ shows that $x_0$ satisfies the equation $A_{-1}x_0 = 0$. This is a co-dimension one subspace of $\mathbb{C}^n$ and hence the space of holomorphic solutions of the differential equation is of dimension at least $n-1$. This proves the first part.

If a solution is holomorphic, then analytic continuation along a loop around 0 does not change the function and hence the monodromy element acts trivially. This proves the second part of the Theorem.

Slightly more work is needed when the eigenvalue $d$ is a positive integer. The equation (4) may be applied to all $k \neq d$; in particular, if $A_{-1}x_0$ is zero, then $x_1, \cdots, x_{d-1}$ are uniquely determined. However, the equation (4) applied to $k = d$ shows that there exists a linear transformation $B_d$ on $\mathbb{C}^n$ such that for each $x_0 \in Keri(A_{-1})$ we have

$$(d - A_{-1})(x_d) = B_d(x_0).$$

The image $W$ of $d - A_{-1}$ has dimension $n-1$ and hence if $B_d(x_0)$ lie in this image $W$, then the recursion (4) still applies to locate an $x_d$. The other $x_j (j \geq k+1)$ are now uniquely determined by (4) and hence the space of solutions which are holomorphic at 0 has dimension at least $n-2$: this is the dimension of the space of $x_0$ satisfying $A_{-1}x_0 = 0$ and $B_k(x_0) \in W = Image(d - A_{-1})$.

Now we take $x_0 = 0$. Then by (4), $x_1 = \cdots = x_{d-1} = 0$. Moreover, $x_d$ satisfies $(A_1 - d)(x_d) = 0$. Since the kernel of $A_1 - d$ has dimension one, there does exist a non-zero $x_d$ with this property. Then by (4), all $x_j$ with $j \geq d+1$ are determined and hence there exists an extra holomorphic solution $w$ of the form $w(z) = z^D C_d + z^{d+1}x_{d+1} + \cdots + x_k z^k + \cdots$ . Hence the space of holomorphic solutions is again of dimension at least $n-1$. This proves the first part when $d$ is a positive integer. The statement about the monodromy matrix follows as before. ■

COROLLARY 3.4 *Under the assumptions of Theorem 3.3, the exceptional eigenvalue $c$ of the monodromy matrix satisfies $c = e^{-2\pi i\beta}$ where $\beta$ is the residue at 0 of the meromorphic function $f_{n-1}(z)$.*

*Proof.* It is easy to see that all the coefficients of the indicial equation $f(t) = 0$ except $n, n-1$ are 0 and that

$$f(t) = t^n + \beta t^{n-1} = t^{n-1}(t + \beta).$$

By Theorem 3.1 the corollary follows.

COROLLARY 3.5 *Under the assumptions of the preceding corollary, the monodromy matrix $M$ is unipotent if and only if the residue $\beta$ of $f_{n-1}(z)$ at 0 is an integer.*

■

### 3.4. *The Plane with Two Punctures*

Consider the twice punctured plane $\mathcal{U} = \mathbb{C} \setminus \{0, 1\}$, and a differential equation

$$\frac{d^n y}{dz^n} + \sum_{i=0}^{n-1} f_i(z) \frac{d^i y}{dz^i} = 0,$$

where $f_i : \mathbb{C} \setminus \{0, 1\} \to \mathbb{C}$ are holomorphic. Now the fundamental group of $\mathcal{U}$ is the free group $F_2$ on two generators $h_0, h_1$, given by small loops going counterclockwise once around 0 and 1 respectively. Thus the monodromy representation of the differential equation is a homomorphism from $F_2$ into $GL_n(\mathbb{C})$; it is completely described by specifying what the images of $h_0$ and $h_1$ are. Thus the monodromy is described by giving two matrices in $GL_n(\mathbb{C})$.

The open set $\mathcal{U} = \mathbb{C} \setminus \{0, 1\}$ may also be viewed as $\mathbb{P}^1 \setminus \{0, 1, \infty\}$. Thus, the fundamental group of $\mathcal{U}$ can also be thought of as the free group on the small loops $h_\infty, h_0, h_1$ going around $\infty, 0, 1$ modulo the relation $h_\infty h_1 h_0 = 1$. Denote, by $A$ the image of $h_\infty$ and by $B^{-1}$ that of $h_0$. Then $C = A^{-1}B$.

Since the universal cover of $\mathbb{C} \setminus \{0, 1\}$ is the upper half plane, it follows that the solutions to the foregoing equations are functions on the upper half plane and that the fundamental group of $\mathcal{U}$ is the deck transformation group.

## 4. The Hypergeometric Differential Equation

Suppose that $\mathcal{U} = \mathbb{P}^1 \setminus \{0, 1, \infty\}$. Put $\theta = z\frac{d}{dz}$ and let $\alpha_1, \cdots, \alpha_n$, and $\beta_1, \cdots, \beta_n$ be complex numbers. Write

$$D = (\theta + \beta_1 - 1) \cdots (\theta + \beta_n - 1) - z(\theta + \alpha_1) \cdots (\theta + \alpha_n). \tag{5}$$

This is a differential operator on $\mathcal{U}$. The equation $Dy = 0$ is called the "hypergeometric differential equation" and the solutions are called "hypergeometric functions". These are functions on the upper half plane.

THEOREM 4.1 *Under the monodromy representation considered in the preceding subsection, the monodromy of the generator $h_0$ around the puncture 0 has characteristic polynomial $\prod(t - e^{2\pi i(1-\beta_j)})$ and the monodromy action of $h_\infty$ has characteristic polynomial $\prod(t - e^{2\pi i\alpha_j})$. Moreover, the element $h_1$ acts by a complex reflection.*

*Proof.* Since the hypergeometric differential equation is already written in the "$\theta$" form, and the coefficients of powers of $\theta$ are linear polynomials in $z$, it follows that 0 is a regular singular point of the differential equation $Du = 0$ where $D$ is the operator in (5). The indicial equation at 0 is thus $\prod_{i=1}^{n}(t + \beta_j - 1)) = 0$. By the Theorem of Fuchs (Theorem 3.1), it follows that the monodromy of $h_0$ has

characteristic polynomial $\prod(t - e^{2\pi i(1-\beta_j)})$.

We now consider the point $\infty$; by changing the variable $z$ to the variable $w = \frac{1}{z}$, the operator $\theta_z = z\frac{d}{dz}$ changes to $-\theta_w = -w\frac{d}{dw}$. Multiplying throughout by $w$, the operator $D$ changes to

$$w(-\theta_w + \beta_1 - 1)\cdots(-\theta_w + \beta_n - 1) - (-\theta_w + \alpha_1)\cdots(-\theta_w + \alpha_n),$$

which is just a constant multiple of the hypergeometric operator

$$D' = (\theta_w - \alpha_1)\cdots(\theta_w - \alpha_n) - w(\theta_w + 1 - \beta_1)\cdots(\theta_w + 1 - \beta_n).$$

Therefore, $\infty$ is also a regular singular point of the equation $Du = 0$ and the monodromy statement follows as in the preceding paragraph.

Consider now the point $z = 1$. We write out the operator $D$ of (5) (which is in "$\theta$" form) in terms of powers of $\frac{d}{dz}$: this is of the form

$$D = z^n\frac{d^n}{dz^n} + P_{n-1}(z)\frac{d^{n-1}}{dz^{n-1}} + \cdots + P_0(z)$$

$$-z(z^n\frac{d^n}{dz^n} + Q_{n-1}(z)\frac{d^{n-1}}{dz^{n-1}} + \cdots + Q_0(z))$$

where $P_i, Q_i$ are polynomials in $z$. Therefore,

$$D = z^n(1-z)\frac{d^n}{dz^n} + R_{n-1}(z)\frac{d^{n-1}}{dz^{n-1}} + \cdots + R_0(z),$$

where the $R_i(z)$ are polynomials. Hence the hypergeometric equation $Dy = 0$ at $z = 1$ (after normalising the highest coefficient to be 1), has the property that all its coefficients $\frac{R_i(z)}{z^n(1-z)}$ at $z = 1$ have at most a simple pole at $z = 1$. By Theorem 3.3 it follows that $h_1$ maps to a complex reflection. ■

The following theorem says that these facts suffice to characterise the monodromy action.

### 4.1. *Statement of Levelt's Theorem*

The monodromy representation is very simply described. Suppose that $\alpha_j - \beta_k$ is not an integer for any two suffices $j, k$. Write $f(x) = \prod_{j=1}^n(x - e^{2\pi i\alpha_j})$, $g(x) = \prod_{k=1}^n(x - e^{2\pi i\beta_k})$. These are monic polynomials of degree $n$. Write $f = x^n + a_{n-1}x^{n-1} + \cdots + a_0$. The quotient ring $R = \mathbb{C}[x]/(f(x))$ is a vector space of dimension $n$ and has as basis the vectors $1, x, \cdots, x^{n-1}$. Write $A$ for the linear operator on the ring $R$ given by multiplication by $x$. With respect to the foregoing basis, the matrix of $A$ is

$$A = \begin{pmatrix} 0 & 0 & 0 & \cdots & -a_0 \\ 1 & 0 & 0 & \cdots & -a_1 \\ 0 & 1 & 0 & \cdots & -a_2 \\ \cdots & \cdots & \cdots & \cdots & \\ 0 & 0 & \cdots & 1 & -a_{n-1} \end{pmatrix}$$

and is called the "companion matrix" of $f$. Similarly, let $B$ be the companion matrix of $g$. Note that $C = A^{-1}B$ is identity on the first $n-1$ basis vectors. Hence $C$ is a complex reflection.

THEOREM 4.2 *(Levelt,1960) There exists a basis $\varepsilon_1, \cdots, \varepsilon_n$ of the space of solutions to the hypergeometric equation such that the monodromy representation sends $h_0$ to $B^{-1}$ and $h_\infty$ to $A$.*

*Moreover, if $\rho$ is any representation of the free group $h_0, h_\infty$ into $GL_n(\mathbb{C})$ such that the images of $h_\infty, h_0^{-1}$ have characteristic polynomials $f, g$, and such that $h_1$ goes to a complex reflection, then $\rho$ is the above monodromy representation.*

The representation described in Levelt's theorem is called the "hypergeometric representation" and the monodromy group is called a "hypergeometric".

We prove Levelt's theorem in the next section.

## 5. Levelt's Theorem

### 5.1. *Notation*

Denote by $R_0$ the ring $\mathbb{Z}[x_i^{\pm 1}, y_i^{\pm 1}]$ of Laurent polynomials in the variables $x_1, \cdots, x_n$ and $y_1, \cdots, y_n$ with integral coefficients, and by $K_0$ its quotient field. Let $R$ denote the sub-ring of $R_0$ generated by the elementary symmetric functions $\sigma_i$ in $x_i$ and the elementary symmetric functions $\tau_j$ in $y_j$ together with the inverses $\sigma_n^{-1}, \tau_n^{-1}$. Denote by $K$ the quotient field of $R$; then $K \subset K_0$. Put

$$f = f(t) = \prod_{i=1}^{n}(t - x_i) = t^n + \sum_{i=1}^{n-1} A_i t^{n-i},$$

$$g = g(t) = \prod_{j=1}^{n}(t - y_j) = t^n + \sum_{i=1}^{n} B_i t^{n-i}.$$

Then $f, g$ are polynomials in $t$ with coefficients in $R$. $F_2$ denotes the free group on two generators (which in the sequel, are often written $h_0, h_\infty$). Define a representation $\rho$ on $F_2 =< h_0, h_\infty >$ by $h_0 \mapsto A$ and $h_\infty^{-1} \mapsto B$ where $A, B$ are companion matrices of $f, g$ respectively. Denote by $\Gamma = \Gamma(f, g)$ the group generated by $A, B$ in $GL_n(R)$, and by $G$ the Zariski closure of $\Gamma$ in $GL_n$. As usual, $G^0$ denotes the connected component of identity of $G$; it is a normal subgroup in $G$ of finite index. Denote by $\Gamma^0$ the intersection of $\Gamma$ with $G^0$.

The group $\Gamma$ is called the *hypergeometric group* corresponding to the parameters $x_i, y_j$. Sometimes, it is simply called the **hypergeometric** corresponding to the polynomials $f, g$ above.

Let $\pi : R \to S$ a ring homomorphism with $S$ an integral domain whose quotient field is denoted $K_S$. Denote by $a_i, b_i \in S$ the images of $x_i, y_i$ under the map

$\pi : R \to S$. Denote by $f_S, g_S$ the monic polynomial in $t$ given by

$$f_S(t) = \prod_{i=1}^{n}(t - a_i), \quad g_S(t) = \prod_{i=1}^{n}(t - b_i).$$

We view the free $S$ module $S^n$ as the quotient ring $S[t]/(f_S(t))$. With respect to the basis $1, t, \cdots, t^{n-1}$ of $S^n$, $A_S$ is simply the matrix of the "multiplication by $t$" operator. Denote by $k$ the g.c.d. of $f_S$ and $g_S$. $B_S$ is the operator which acts as follows: $B_S(t^i) = A_S(t^i) = t^{i+1}$ if $i \leq n-2$ and $B_S(t^{n-1}) = t^n - g_S(t)$. Let $W$ be the ideal of $S^n = S[t]/(f_S(t))$ generated by the polynomial $k$; then $W$ is $A_S$ stable. Hence so is $W \otimes K_S$.

LEMMA 5.1 *The subspace $W_S = W \otimes K_S$ is also $B_S$ stable and under the action of $A_S, B_S$, the subspace $W_S$ is irreducible.*

*Moreover, on the quotient $V_S/W_S$ the operators $A_S, B_S$ coincide and as a module over $A$ (multiplication by $t$), the quotient $V_S/W_S$ is the ring $K_S[t]/(k(t))$.*

*In particular, if $f, g$ are co-prime (i.e. $a_i \neq b_j$ for any $i, j$), then $V_S = W_S$ is irreducible for the action of $F_2$.*

*Proof.* We temporarily write $A_S = A, B_S = B$. Put $D = A - B$. Then, the image of $D$ on $V_S$ is the line generated by $f_S - g_S$. Moreover, $D$ is zero on the monomials $1, t, \cdots, t^{n-2}$ and is $f - g$ on $t^{n-1}$. Therefore, the image under $D$ of a polynomial of degree exactly $n-1$ is a non-zero multiple of $f - g$.

Any subspace of $V_S$ which is stable under $A$ contains an eigenvector for $A$; these eigenvectors are of the form $\varepsilon_i = \frac{f(t)}{x - a_i}$ for some $i$. This a polynomial of degree exactly $n-1$. Hence $D(\varepsilon_i)$ is a non-zero multiple of $f - g$; thus any subspace of $V_S$ which is stable under $A, D$ contains $f - g$ and hence contains $W_S = K_S[t](f-g) + K_S[t]f = K_S[t]k(t) = (k(t))$. This proves the first part of the Lemma.

On the quotient $V_S/W_S$, the operator $D$ is zero, since the image of $D$ lies in $W_S$. Hence $A = B$ on the quotient $V_S/W_S = K_S[t]/(k(t))$. This proves the second part.

The third part is a corollary of the first part. ■

THEOREM 5.2 *(Levelt) Suppose that $a_i \neq b_j$ for any $i, j$. Suppose $h_0 \mapsto a$ and $h_\infty^{-1} \mapsto b$ is any other irreducible representation $\rho'$ of $F_2$ into $GL_n(\overline{K}_S)$ such that the following* two *conditions hold. (1) the characteristic polynomial of $a$ is $f_S(t) = \prod(t - a_i)$ and the characteristic polynomial of $b$ is $g_S(t) = \prod(t - b_j)$. (2) $a^{-1}b$ is identity on a co-dimension one subspace of $K^n$.*

*Then $\rho'$ is equivalent to $\rho$.*

*Proof.* Put $D' = a - b$ and Let $W$ be the kernel of $D'$. By assumption, $W$ has co-dimension one in $V$. Write

$$X = \cap_{i=0}^{n-2} a^{-i} W.$$

Since $X$ is an intersection of $n-1$ hyperplanes in $V$, $X$ is non-zero. Let $v \in X$,

with $v \neq 0$.

We claim that $v$ is cyclic for the action of $a$. Suppose not. Then, $v, av, \cdots, a^{n-1}v$ are linearly dependent. We then claim that $a^{n-1}v$ is a linear combination of $v, av, \cdots, a^{n-2}v$:

Suppose $v, av, \cdots, a^{n-2}v$ are already linearly dependent. By applying a suitable power of $a$ to a linear dependence relation, we see that $a^{n-1}v$ is a linear combination of the vectors $v, av, \cdots, a^{n-2}v$.

Suppose $v, av, \cdots, a^{n-2}v$ are linearly independent. Since the vectors $v, av, \cdots, a^{n-1}v$ are linearly dependent, it follows that $a^{n-1}v$ is a linear combination of $v, av, \cdots, a^{n-2}v$.

Since $f_S(a) = 0$, it follows that the span $E$ of $v, av, \cdots, a^{n-2}v$ is $a$ stable. Since all the vectors $v, av, \cdots, a^{n-2}v$ lie in $W$ by the definition of $X$, it follows that $a = b$ on $E$ and hence $E$ is stable under $\Gamma$. Since $E \neq 0$, it follows that the characteristic polynomial of $a = b$ on $E$ are equal and have a common eigenvalue, contradicting the assumption that $a_i \neq b_j$ for any $i, j$. Therefore, $v$ is a cyclic vector for the action of $a$.

Hence $v, av, \cdots, a^{n-1}v$ is a basis for $V$. It follows that with respect to this basis, the matrix of $A$ is the companion matrix of $f_S$.

By the construction of $X$, we have $a^i v \in W$ for $i \leq n-2$. Therefore, $ba^i v = aa^i v = a^{i+1}v$ for $i \leq n-2$. Induction on $i \leq n-2$ shows that $a^i v = b^i v$ for all $i \leq n-1$. With respect to the basis $v, av, \ldots, a^{n-1}v$ of $V$, the matrix of $a$ is the companion matrix of $f_S(t)$, and that of $b$ is the companion matrix of $g_S(t)$. This completes the proof of the Theorem. ■

## 6. Results of Beukers-Heckman

The Zariski closure of the hypergeometric also has a pleasant description. This is described by Beukers and Heckman [3]. For ease of exposition, we assume that the roots of $f(x), g(x)$ are roots of unity (i.e. $\alpha_j, \beta_k$ are rational numbers), and that $f, g$ are products of cyclotomic polynomials. Then $f(x), g(x) \in \mathbb{Z}[x]$. Moreover, $f(0) = \pm 1$ and $g(0) = \pm 1$. We recall from the previous section that the monodromy group $H(f, g)$ is generated by the companion matrices $A, B$ of the polynomials $f, g$ respectively.

### 6.1. *The Finite Case*

We may assume that the numbers $\alpha_j$ and $\beta_k$ lie in the closed open interval $[0, 1)$ and $\alpha_j \neq \beta + k$. We say that the numbers $\alpha_j$ and $\beta_k$ "interlace" if between any two $\alpha_j$ there is a $\beta_k$ and conversely. [3] give a criterion for the monodromy group to be finite in terms of the parameters $\alpha, \beta$:

THEOREM 6.1 *(Beukers-Heckman) The hypergeometric group corresponding to the parameters $\alpha_j, \beta_k$ is finite if and only if the parameters $\alpha_j$ and $\beta_k$ interlace.*

### 6.2. *Imprimitivity*

We say that $f(X), g(X)$ are "imprimitive" if there exists an integer $k \geq 2$ and polynomials $f_1, g_1$ such that $f(x) = f_1(x^k)$, $g(x) = g_1(x^k)$. Otherwise, we say that $f, g$ are a "primitive" pair.

We assume henceforth that $f, g$ are coprime, form a primitive pair and that $\alpha_j, \beta_k$ do not satisfy the interlacing condition. Let $G$ be the Zariski closure of the hypergeometric. Write $c = \frac{f(0)}{g(0)}$. Then $c = \pm 1$.

THEOREM 6.2 *(Beukers-Heckman) Suppose that the roots of $f, g$ do not interlace, $f, g \in \mathbb{Z}[x]$ are coprime and primitive.*

*If $c = -1$ then the Zariski closure of the hypergeometric is isomorphic to $O(n)$, the orthogonal group on $n$ variables.*

*If $c = 1$, then the Zariski closure is the symplectic group $Sp_n$ (under our assumptions, $n$ will necessarily be even).*

We do not prove this theorem, since that would take us too far afield. We refer to [3] for a proof of a more general result from which Theorem 6.2 follows.

## 7. Symplectic Case

In this section, we will assume that the Zariski closure of the hypergeometric is a symplectic group; i.e. assume that $f, g \in \mathbb{Z}[x]$, $f, g$ form a primitive pair and that the roots of $f, g$ do not interlace. Assume that $f(0) = g(0) = 1$ so that in Theorem 6.2 $c = 1$. By the result of Beukers and Heckman (Theorem 6.2) the hypergeometric $H(f, g)$ is a Zariski dense subgroup of $Sp_\Omega(\mathbb{Z})$ for a non-degenerate symplectic form $\Omega$ on $\mathbb{Q}^n$. It is then an interesting question to ask when $H(f, g)$ has finite index (i.e. when is $H(f, g)$ an arithmetic symplectic group). There is no complete characterisation but some cases are now known.

### 7.1. *Arithmetic Groups*

See [11] for the following result.

THEOREM 7.1 *suppose that $f, g$ are as in the beginning of this subsection and that the difference $f - g = c_0 + \cdots + c_d X^d$ with leading coefficient $c = c_d \neq 0$; assume that $\mid c \mid \leq 2$. Then $H(f, g)$ has finite index in $Sp_\Omega(\mathbb{Z})$; thus the hypergeometric group is an arithmetic group.*

As a family of examples, consider, for an even integer $n$,

$$f(X) = \frac{X^{n+1} - 1}{X - 1} = X^n + x^{n-1} + \cdots + X + 1,$$

$$g(X) = (X + 1)\frac{X^n - 1}{X - 1} = X^n + 2X^{n-1} + 2X^{n-2} + \cdots + 2X + 1.$$

The difference $f - g = -(X^{n-1} + X^{n-2} + \cdots + X)$ has leading coefficient $c = -1$

and hence the hypergeometric $H(f,g)$ has finite index in $Sp_\Omega$, by Theorem 7.1.

## 7.2. *Thin Groups*

We recall the following definition (see [9] for details)

*Definition* 1 Let $\Gamma \subset SL_n(\mathbb{Z})$ be a subgroup and $G$ its Zariski closure in $SL_n$. Then $G$ is defined over $\mathbb{Q}$ and $\Gamma \subset G(\mathbb{Z}) = G \cap SL_n(\mathbb{Z})$. We say that $\Gamma$ is **thin** if $\Gamma$ has infinite index in $G(\mathbb{Z})$. Otherwise, we say that $\Gamma$ is arithmetic.

Note that the notion of thinness and of arithmeticity depends on the embedding $\Gamma \subset SL_n(\mathbb{Z})$.

It is widely believed that most hypergeometric groups in Theorem 6.2 are thin. Theorem 7.1 says however, that not all the hypergeometrics are thin. There is no general criterion as to when the hypergeometric are thin, except when the Zariski closure is $O(n,1)$ (see [7]). In the next subsection, we will see examples of thin hypergeometrics with Zariski closure $Sp_4$ (constructed by Brav and Thomas [4]).

## 7.3. *Fourteen Families*

Of special interest are the hypergeometrics corresponding to $f = (X-1)^4$ (i.e. when the monodromy around infinity is *maximally unipotent*. The number of choices for $g$ are limited: $g \in \mathbb{Z}[X]$ must be a product of cyclotomic polynomials, and must have degree 4; moreover, $g(1) \neq 0$. With these constraints there are exactly 14 choices for $g$. It is known that the hypergeometric $H(f,g)$ is also the monodromy group associated to a family of *Calabi-Yau threefolds* fibering over the thrice punctured projective line. Moreover, these threefolds turn up in mirror symmetry.

In [1] (see also [5]), the question of thinness or arithmeticity of these groups was first raised. Theorem 6.2 and its proof enables us to prove that the monodromy group is arithmetic in three of these cases; later Singh [10] adapted the method to prove arithmeticity in four more cases. On the other hand, Brav and Thomas [4] have proved that 7 of these hypergeometric groups are thin. In particular , they prove

THEOREM 7.2 *(Brav and Thomas) Suppose* $f(X) = (X-1)^4$ *and* $g(X) = \frac{X^5-1}{X-1}$. *Then the hypergeometric group* $H(f,g) \subset Sp_\Omega(\mathbb{Z})$ *(is Zariski dense in* $Sp_\Omega$ *and) has infinite index in* $Sp_\Omega(\mathbb{Z})$*; in particular, it is a thin monodromy group.*

To sum up, out of these fourteen families, 7 are arithmetic ([11], [10]) and 7 are thin [4]. Theorem 7.2 is the first example of a "higher-rank" thin monodromy group whose Zariski closure is a simple group.

## 7.4. *Questions*

As was mentioned before, there is no general criterion to determine when a group is thin or not. Consider the hypergeometric $H(f,g)$ associated to

$$f(X) = (X-1)^n, \quad g(X) = \frac{X^{n+1}-1}{X-1},$$

say, with even $n$. It is easy to deduce from [3] that $H(f,g)$ is Zariski dense in $Sp_n$. When $n=4$, this is the group considered in Theorem 7.2 and is thin. However, for $n \geq 6$ and even, it is not known if the group $H(f,g)$ has infinite index in the integral symplectic group. In particular, let us consider the subgroup $\Gamma$ of $Sp_6(\mathbb{Z}) \simeq Sp_\Omega(\mathbb{Z})$ generated by the companion matrices of $(X-1)^6$, $\frac{X^7-1}{X-1}$. It is not known if $\Gamma$ has finite index or not.

## References


[1] Almkvist, Enckevort, Duco van Straten, W.Zudilin, Tables of Calabi-Yau Equations, October 2010, arXiv:math/0507/v2

[2] F.Beukers, Notes on Differential Equations and Hypergeometric Functions, HGF course 2009 in Utrecht.

[3] F.Beukers and G Heckman, Monodromy for the hypergeometric Function ${}_nF_{n-1}$, Invent. math **95** (1989), no. 2, 325-254.

[4] C Brav and H. Thomas, Thin monodromy in $Sp_4$, Compositio Math. **150**, Issue 3, 333-343.

[5] Y-H Chen, C Erdenberger, Y. Yang, N. Yui, Monodromy of Picard-Fuchs differential equations for Calabi-Yau threefolds, J reine.angew. Math. **616** (2008), 167-203.

[6] E.Coddington and N.Levinson, Theory of Ordinary Differential Equations, International Series in Pure and Applied Mathematics, Mcgraw Hill Book Company, NewYork - Toronto - London, 1955.

[7] E.Fuchs, C Meiri, P.Sarnak Hyperbolic Monodromy groups for the hypergeometric equation and Cartan Involutions, Journal of the European Math Society, Volume **16**, Issue 8, (2014), 1617-1671.

[8] A.H.M Levelt, Hypergeometric Functions I and II, Nederland Akad Wetensch Proc Ser A 64 Indag Math **23** (1961), 361-373, 373-385.

[9] P.Sarnak, Notes on thin matrix groups, in Thin groups and Superstrong Approximation, MSRI Publications **61**, Cambridge University Press, Cambridge, 2014.

[10] S.Singh, Srithmeticity of four hypergeometric groups associated to Calabi-Yau threefolds, IMRN (2015), **no 18**, 8874-8889.

[11] S.Singh and T.N.Venkataramana, Arithmeticity of certain symplectic hypergeometric groups, Duke Math J. **163** (2014), no 3, 591-617.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 103–112

# Zabreiko's Lemma: Unified Treatment of Four Fundamental Theorems in Functional Analysis

S Kumaresan
Visiting Professor
Indian Institute of Technology Kanpur
Kanpur 208016
kumaresa@gmail.com

**Abstract:** Zabreiko [3] proved a lemma in 1969 (50 years ago!) on the continuity of seminorms in terms of its countable subadditivity. All the four basic theorems of functional analysis, namely, the open mapping theorem, closed graph theorem, bounded inverse theorem and uniform boundedness principle can be derived in a somewhat uniform fashion from the lemma. In fact, these deductions are fun! It is strange that this lemma is not as popular as it should be with the authors of the textbooks on functional analysis. The aim of this article is to give a detailed exposition to the lemma, its proof and the derivations of the four theorems. Experts can skip a lot of routine steps and see how easy the lemma and the deductions are.

Let $X$ be a vector space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. We say that a map $p\colon X \to \mathbb{R}$ is a *seminorm* if it has the following properties:

(i) $p(x) \geq 0$ for all $x \in X$.
(ii) $p(tx) = |t|p(x)$ for $t \in \mathbb{K}$ and $x \in X$.
(iii) $p(x + y) \leq p(x) + p(y)$ for $x, y \in X$.

Note that $p(0) = 0$ since $p(0 \cdot x) = 0p(x) = 0$ where $x \in X$.

Thus $p$ is a norm if (i) is strengthened: $p(x) \geq 0$ for $x \in X$ and $p(x) = 0$ iff $x = 0$.

How to get seminorms on a vector space $X$? Let $T\colon X \to Y$ be any linear map. Assume that $Y$ is a normed linear space. Define $p(x) := \|Tx\|$. Then $p$ is a seminorm. (Do you see why it may fail to be a norm?)

Before Zabreiko's lemma, let us understand the geometry of sets associated with a seminorm $p$ on a vector space $X$.

Let $B := \{x \in X : p(x) < 1\}$. (This is something like an open unit ball if $p$ were a norm. If you keep this in mind, the next three claims are credible!) Let $A_n := \{x \in X : p(x) < n\}$. Note that $A_n = nB$.

We claim that $X = \cup_{t>0} tB$. Let $s := p(x)$. If $t > s$, then we claim that $x \in tB$. For, $p(x/t) = p(x)/t = s/t < 1$. Hence $x/t \in B$ and hence $x \in tB$. In particular, $X = \cup_n A_n$. For, $sB \subset tB$ if $0 < s < t$ and given any $t \in \mathbb{R}$, by Archimedean property of $\mathbb{R}$, there exists $n \in \mathbb{N}$ such that $t < n$. Hence if $x \in tB$, then $x \in nB$.

We next claim that $A_n$ is convex. For, let $x, y \in A_n$ and $t \in [0, 1]$. Observe

$$p((1-t)x + ty) \leq p((1-t)x) + p(ty) = (1-t)p(x) + tp(y) < (1-t)n + tn = n..$$

We claim that each $A_n$ is symmetric, that is, $A_n = -A_n$, equivalently, $x \in A_n$ iff $-x \in A_n$. This follows from the fact that $p(x) = p(-x)$.

Let $(X, \| \ \|)$ be a normed linear space and $p$ a seminorm on $X$. We say that a $p$ is continuous on $X$ if the map $p\colon (X, \| \ \|) \to \mathbb{R}$ is continuous.

An easy observation is that $p$ is continuous on $(X, \| \ \|)$ iff $p$ is continuous at $0 \in X$. Let us prove the nontrivial part. Let $p$ be continuous at 0. Let $x \in X$. We claim that $p$ is continuous at $x$. Let $\varepsilon > 0$ be given. By the continuity of $p$ at 0, there exists $\delta > 0$ be such that

$$\|z\| < \delta \implies p(z) < \varepsilon. \tag{1}$$

Let $y \in X$ be such that $\|y - x\| < \delta$. Note that $p(y) = p((y-x)+x) \le p(y-x) + p(x)$. Hence $p(y) - p(x) \le p(y-x) < \varepsilon$. Similarly, $p(x) = p(x-y+y) \le p(x-y)+p(y)$ and hence $p(x) - p(y) \le p(x-y)$. Observe that $p(x-y) = p(-(y-x)) = |-1|p(y-x) = p(y-x)$. Thus we have proved $|p(y)-p(x)| \le p(y-x) < \varepsilon$. Thus the continuity of $p$ at $x$ is established.

If $p$ is continuous, we claim that there exists $C > 0$ such that $p(x) \le C\|x\|$ for $x \in X$. Since $p$ is continuous at 0, let us assume that (1) holds. If $x$ is nonzero, let us consider $z := \frac{\delta}{2\|x\|}x$. Then $\|z\| < \delta$ and hence $p(z) < \varepsilon$. But this says

$$p\left(\frac{\delta}{2\|x\|}x\right) < \varepsilon \implies p(x) < \frac{2\varepsilon}{\delta}\|x\|.$$

So, we may take $C := \frac{2\varepsilon}{\delta}$.

Let $(x_n)$ be a sequence in the normed linear space $X$. Assume that $\sum_n x_n$ is convergent in $X$. That is, if $s_n := \sum_{k=1}^n x_k$, then there exists $x \in X$ such that $\|s_n - x\| \to 0$. We denote this by $\sum_n x_n = x$. We say that $\sum_n x_n$ is *absolutely convergent* or *norm convergent* if the series $\sum_n \|x_n\|$ (of nonnegative terms) is convergent.

It is well-known that if $X$ is a Banach space any absolutely convergent series is convergent.

Assume that $p\colon (X, \| \ \|) \to \mathbb{R}$ is a continuous seminorm. We claim that $p$ is *countably subadditive*: If $\sum_n x_n$ is convergent then $p(\sum_n x_n) \le \sum_n p(x_n)$. Note that the countable subadditivity of $p$ is relative to the norm $\| \ \|$. Also, if $\sum_n x_n$ is absolutely convergent, then $p(\sum_n x_n) \le \sum_n p(x_n)$.) Let $s_n$ be the $n$-th partial sum of the series. Assume that $s_n \to x$ and hence $x = \sum_n x_n$. Since $p$ is continuous $p(s_n) \to p(x) \equiv p(\sum_n x_n)$. Let $t_n := \sum_{k=1}^n p(x_k)$. Then $(t_n)$ is increasing sequence of real numbers and if it is bounded (above), it converges to $t := \text{LUB } \{t_n : n \in \mathbb{N}\}$. In particular, for each $n \in \mathbb{N}$, we have $t_n \le t$. If $(t_n)$ is not bounded above, then $\sum_n t_n = \infty$. Since we wish to show that $p(\sum_n x_n) \le \sum_n p(x_n)$, we may as well assume that $\sum_n p(x_n)$ is finite, say, $t$. Now, observe that

$$p(s_n) \le p(x_1) + \cdots + p(x_n) = t_n \le t.$$

Hence $p(x) = \lim p(s_n) \le t$. Thus the claim is proved.

Zabreiko's lemma says the converse of our claim is true *provided that* $X$ is a Banach space.

LEMMA 1 (Zabreiko) *Let $(X, \| \ \|)$ be a Banach space. Let $p$ be a countably subadditive seminorm on $X$: if $\sum_n x_n$ is convergent, then $p(\sum_n x_n) \le \sum_n p(x_n)$. Then $p$ is continuous.*

*Proof.* Let $A_n$ be defined as above.

Let $F_n := \overline{A_n}$, the closure of $A_n$ in $(X, \|\ \|)$. Since $A_n \subset F_n$ and $X = \cup_n A_n$, we see that $X = \cup_n F_n$. Thus the complete metric space $X$ is the countable union of closed sets $F_n$. Hence by Baire's theorem, one of them must have nonempty interior, say, $F_N$. Therefore there exists $a \in X$ and $R > 0$ such that $B_X(a, R) := \{x \in X : \|x - a\| < R\} \subset F_N$. Note that $B_X(-a, R) = -B_X(a, R)$. For,

$$x \in B_X(a, R) \iff \|x - a\| < R \iff \|-x + a\| < R \iff -x \in B_X(-a, R).$$

Since $F_N$ is symmetric (why?), as a consequence we see that $-B_X(a, R) \subset F_N$. Now if $\|x\| < R$, then $a + x \in B_X(a, R) \subset F_N$ and $x - a \in B_X(-a, R) \subset F_N$. Since $F_N$ is convex, we observe that

$$x = (1/2)(x - a) + (1/2)(x + a) \in F_N.$$

That is, using the geometry of $F_N$ we proved that if $B(a, R) \subset F_N$, then $B(0, R) \subset F_N$. (One can, in fact, show that $B(0, R) \subset A_N$. See Remark 2.) But in stead, we shall prove that $B(0, R) \subset 2A_N$.

Now we mimic the standard proof[1] of open mapping theorem to show that $B(0, R) \subset 2A_N$.

Let $x \in B(0, R) \subset F_N = \overline{A_N}$. Hence there exists $x_1 \in A_N$ such that $\|x - x_1\| < 2^{-1}R$. Note that $p(x_1) < N$.

Consider $x - x_1 \in B(0, 2^{-1}R) \subset 2^{-1}F_N = 2^{-1}\overline{A_N}$. Hence there exists $x_2 \in 2^{-1}A_N$ such that $\|(x - x_1) - x_2\| = \|x - (x_1 + x_2)\| < 2^{-2}R$.

This is not required for the proof. It is required for the stronger version Theorem 6. Note that $p(x_2) < 2^{-1}N$. Note that

$$\begin{aligned}\|x_2\| = \|x - (x_1 + x_2) - (x - x_1)\| &\leq \|x - x_1 - x_2\| + \|x - x_1\| \\ &< 2^{-2}R + 2^{-1}R = 2^{-1}R(1 + 2). \qquad (2)\end{aligned}$$

Proceeding by induction, we obtain a sequence $(x_n)$ such that $x_n \in 2^{-n+1}A_N$ and such that $\|x - (x_1 + \cdots + x_n)\| \leq 2^{-n}R$.

Hence we conclude that the series $\sum_n x_n$ is convergent and $x = \sum_n x_n$. Note that by our choice $p(x_n) < 2^{-n+1}N$.

Since $p$ is countably subadditive, we see that

$$p(x) = p\left(\sum_n x_n\right) \leq \sum_n p(x_n) \leq N \sum_n 2^{-n+1} \leq 2N.$$

So what have we proved? If $x \in B(0, R)$, then $p(x) < 2N$. From this it follows that if $\varepsilon > 0$ is given, we can choose $\delta := \frac{R\varepsilon}{2N}$. Then we have

$$\|x\| < \delta \implies p(x) < \varepsilon.$$

That is, $p$ is continuous at 0. ∎

We shall make use of the following two easy facts often:

(i) Let $X$ be a Banach space. If $(x_n)$ is a sequence in $X$ such that $\sum_n \|x_n\|$

[1]Do not worry, if you have not seen a proof of OMT.

is convergent, then $\sum_n x_n$ is convergent. (The converse is also true, but we do not need it.)

(ii) Let $T \in BL(X)$. Let $\sum_n x_n$ be convergent. Then $T(\sum_n x_n) = \sum_n Tx_n$. In particular, $\sum_n Tx_n$ is convergent.

We now use Zabreiko's lemma to deduce the Bounded Inverse Theorem (BIT), Close Graph Theorem (CGT), Uniform Boundedness Principle UBP) and Open Mapping Theorem (OMT) The fun part of the proofs is to guess the appropriate seminorm for the case on hand. We shall give the hints and work out the details below.

| | |
|---|---|
| BIT | Use $p(y) := \left\|T^{-1}(y)\right\|$ |
| CGT | Use $p(x) := \|Tx\|$ |
| OMT | Use $p(y) := \inf\{\|x\| : Tx = y\}$. |
| UBP | Use $p(x) := \sup\{\|T_i(x)\| : i \in I\}$. |

Let us deal with each of these results one by one.

THEOREM 2 (Bounded Inverse Theorem) *Let $X$ and $Y$ be Banach spaces. Let $T\colon X \to Y$ be a bijective continuous linear map. Then $T^{-1}\colon Y \to X$ is continuous.*

*Proof.* We need to estimate $\left\|T^{-1}y\right\|$ for $y \in Y$. This suggests that we consider $p(y) := \left\|T^{-1}(y)\right\|$.

Since $T^{-1}$ is a bijection, $p$ is well-defined. It is easy to verify that $p$ is a seminorm. As a sample, we shall show that $p(y_1 + y_2) \leq p(y_1) + p(y_2)$:

$$\begin{aligned} p(y_1 + y_2) = \left\|T^{-1}(y_1 + y_2)\right\| &= \left\|T^{-1}(y_1) + T^{-1}(y_2)\right\| \\ &\leq \left\|T^{-1}(y_1)\right\| + \left\|T^{-1}(y_2)\right\| \\ &= p(y_1) + p(y_2). \end{aligned}$$

We wish to show that it is countably subadditive. The nontrivial case arises only when $\sum_n p(y_n)$ is convergent. So, we assume that $\sum_n p(y_n) < \infty$. Let $Tx_n = y_n$, then this means that $\sum_n \|x_n\| < \infty$. Hence the series $\sum_n x_n$ is absolutely convergent, since $X$ is a Banach space. Let $x := \sum_n x_n$. We claim that $\sum_n y_n$ is absolutely convergent in the Banach space $Y$. For, $\sum_n \|y_n\| = \sum_n \|Tx_n\| \leq \|T\| \sum_n \|x_n\|$. Let $y := \sum_n y_n$. Note that $Tx = y$. We now have

$$p\left(\sum_n y_n\right) = p(y) = \left\|T^{-1}(y)\right\| = \|x\| \leq \sum_n \|x_n\| = \sum_n p(y_n).$$

Since $Y$ is complete, by Zabreiko's lemma $p$ is continuous. Hence there exists $C$ such that $p(y) \leq C\|y\|$ for $y \in Y$. That is, $\left\|T^{-1}y\right\| \leq C\|y\|$ for $y \in Y$. We conclude that $T^{-1}$ is continuous linear map and so $T^{-1} \in BL(Y, X)$. ∎

THEOREM 3 (Closed Graph Theorem) *Let $X$ and $Y$ be Banach space. Assume that $T\colon X \to Y$ is a linear map such that its graph $\{(x, Tx) : x \in X\}$ is a closed subset of $X \oplus Y$. Then $T$ is continuous, that is , $T \in BL(X, Y)$.*

*Proof.* We need to estimate $\|Tx\|$. Hence we let $p(x) := \|Tx\|$.

It is obvious that $p$ is a seminorm. We check whether it is countably subadditive. Let $\sum_n x_n$ be convergent. Again, the only nontrivial case is when $\sum_n p(x_n) = \sum_n \|Tx_n\|$ is finite. So we assume that $\sum_n \|Tx_n\|$ is convergent.

Since $Y$ is complete, the series $\sum_n Tx_n$ is convergent. Let $v_n := \sum_{k=1}^n x_k$. Then we have $v_n \to \sum_{n=1}^\infty x_n$. Note that $Tv_n = \sum_{k=1}^n Tx_k \to \sum_{n=1}^\infty Tx_n$. Thus, we see that $(\sum_{k=1}^n x_k, \sum_{k=1}^n Tx_k) \to (\sum_{n=1}^\infty x_n, \sum_{n=1}^\infty Tx_n)$. Since the graph of $T$ is closed, we deduce that $T(\sum_n x_n) = \sum_n Tx_n$. it follows that

$$p(\sum_n x_n) = \left\| T(\sum_n x_n) \right\| = \left\| \sum_n Tx_n \right\| \leq \sum_n \|Tx_n\| = \sum_n p(x_n).$$

Hence $p$ is continuous and hence there exists $C > 0$ such that $p(x) \leq C\|x\|$, that is, $\|Tx\| \leq C\|x\|$ for $x \in X$. Therefore, we conclude that $T$ is continuous. ■

THEOREM 4 (Uniform Boundedness Principle) *Let $X$ be Banach space and $Y$ a normed linear space. Let $T_i \in BL(X,Y)$ for each $i \in I$. Assume that for each $x \in X$, there exists $C_x$ such that $\sup\{\|T_i x\| : i \in I\} \leq C_x$. Then there exists $C > 0$ such that $\sup\{\|T_i\| : i \in I\} \leq C$.*

*Proof.* Since we wish to show that $\{\|T_i\| : i \in I\}$ is bounded above we define $p(x) := \sup\{\|T_i x\| : i \in I\}$.

It is easy to check that $p$ is a seminorm. Let $\sum_n x_n$ be convergent. Let $x := \sum_n x_n$. We verify the countable subadditivity.

$$\begin{aligned}
p\left(\sum_n x_n\right) \equiv p(x) &= \sup\left\{\left\| T_i(\sum_n x_n)\right\| : i \in I\right\} \\
&= \sup\left\{\left\|\sum_n T_i x_n\right\| : i \in I\right\}, && \text{using the continuity of } T_i \\
&\leq \sup\left\{\sum_n \|T_i x_n\| : i \in I\right\}, && \text{using the continuity of the norm} \\
&\leq \sum_n \sup\{\|T_i x_n\| : i \in I\}, && \text{since } \|T_i x\| \leq \sup_i\{\|T_i x\|\} \\
&= \sum_n p(x_n).
\end{aligned}$$

Hence $p$ is continuous by Zabreiko's lemma. Hence there exists $C > 0$ such that $p(x) \leq C\|x\|$ for all $x \in X$. If we recall the definition of $p$, we see that $\|T_i x\| \leq C$ for each $i \in I$ and each unit vector $x \in X$. The result follows. ■

THEOREM 5 (Open Mapping Theorem) *Let $X$ and $Y$ be Banach spaces. Let $T\colon X \to Y$ be a continuous linear map from $X$* onto *$Y$. Then $T$ is an open map, that is, $T$ maps any open set $X$ to an open set of $Y$.*

*Proof.* Guessing $p$ in this case is a little tricky or subtle. For motivation, see Remark 1.

We let $p(y) := \inf\{\|x\| : Tx = y\}$. Let $y_1, y_2 \in Y$. Given any $\varepsilon > 0$, we shall show that $p(y_1 + y_2) \leq p(y_1) + p(y_2) + \varepsilon$. Since $\varepsilon > 0$ is arbitrary, it yields the desired inequality, namely, $p(y_1 + y_2) \leq p(y_1) + p(y_2)$. Since $p(y_j)$ is the GLB of the set $\{\|x\| : Tx = y_j\}$, the quantity $p(y_j) + (\varepsilon/2)$ is not a lower bound for the set. Hence there exists $x_j$ such that $\|x_j\| < p(y_j) + (\varepsilon/2)$, $j = 1, 2$. Note that

$T(x_1 + x_2) = y_1 + y_2$ and hence

$$p(y_1 + y_2) \leq \|x_1 + x_2\| \leq \|x_1\| + \|x_2\| < p(y_1) + (\varepsilon/2) + p(y_2) + (\varepsilon/2).$$

We have therefore proved $p(y_1 + y_2) \leq p(y_1) + p(y_2)$.

Let $c \in \mathbb{K}$ be nonzero. Then

$$\begin{aligned} p(cy) &= \inf\{\|x\| : Tx = cy\} \\ &= \inf\{\|cx\| : Tx = y\} \\ &= |c| \inf\{\|x\| : Tx = y\} = |c|p(y). \end{aligned}$$

We now prove countable subadditivity.[2]

Let $\sum_n y_n$ be convergent. As usual, we may assume that $\sum_n p(y_n)$ is convergent. Let $\varepsilon > 0$ be given. Arguing as in the beginning of this proof, we can assert the existence of $x_n$ such that $\|x_n\| < p(y_n) + 2^{-n}\varepsilon$. It is clear that $\sum_n x_n$ is absolutely convergent:

$$\sum_n \|x_n\| \leq \sum_n (p(y_n) + 2^{-n}\varepsilon) = \sum_n p(y_n) + \varepsilon.$$

Since $X$ is complete, we see that $\sum_n x_n$ is convergent. Since $T$ is continuous, we see that $T(\sum_n x_n) = \sum_n Tx_n = \sum_n y_n$. Since $T(\sum_n x_n) = \sum_n y_n$, by the definition of $p$, we see that $p(y) \leq \|\sum_n x_n\|$. Now observe

$$p\left(\sum_n y_n\right) \leq \left\|\sum_n x_n\right\| \leq \sum_n \|x_n\| \leq \sum_n p(y_n) + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that $p$ is countably subadditive. Since $Y$ is complete, by Zabreiko's lemma, $p$ is continuous. Hence $p(y) \leq C\|y\|$ for some $C$ and for all $y \in Y$.

We shall show that $TB_X$ is open. If $y \in TB_X$, there exists $x \in B_X$ such that $Tx = y$. Hence by the definition of $p$, $p(y) \leq \|x\| < 1$. Conversely, if $p(y) < 1$, then we claim that there exists $x \in X$ such that $Tx = y$ and $\|x\| < 1$. For, since $T$ is onto, there do exist $x \in X$ such that $y = Tx$. If each such $x$ is such that $\|x\| \geq 1$, then $p(y) \geq 1$ by the very definition of $p$. Hence there exists $x \in X$ such that $Tx = y$ and $\|x\| < 1$. Thus we see see that

$$\begin{aligned} T(B_X) &= \{y \in Y : \exists x \in B_X \text{ such that } Tx = y\} \\ &= \{y \in Y : p(y) < 1\}. \end{aligned}$$

Since $p$ is continuous, the set $\{y \in Y : p(y) < 1\} = p^{-1}(-\infty, 1)$ is open. Thus, we conclude that $TB_X$ is open.

We now show that $T$ is an open map. Let $U \subset X$ be open. Let $y \in TU$. Then there exists $x \in U$ such that $Tx = y$. Since $x \in U$ and $U$ is open, there exists $r > 0$ such that $B(x, r) \subset U$. Since $B(x, r) = x + rB(0, 1)$, we see that $T(B(x, r)) = Tx + rTB_X = y + rTB_X$. Since the translation and the scalar multiplication by

[2]Do you recall how you proved the countable subadditivity of the outer measure, (or the countable union of sets of measure zero is again of measure zero) in Lebesgue theory of measure and integration? We shall adapt the same argument.

nonzero numbers are homeomorphisms, it follows that $T(B(x,r))$ is an open set and we have $y \in T(B(x,r)) \subset TU$. Thus every point of $TU$ is an interior point and hence $T$ is open. ■

The rest of the article deals with some technical remarks and improvements of Zabreiko's lemma.

*Remark* 1 The standard practice is to derive BIT from OMT and then show BIT implies OMT. Inspired by this, we can adapt the argument of BIT and offer a seemingly different proof of OMT.

Let $X$ and $Y$ be Banach spaces. Let $T\colon X \to Y$ be a continuous linear map which maps $X$ onto $Y$. Let $Z := \ker T$. Then $Z$ is closed linear subspace and $X/Z$ is a Banach space with the quotient norm. We have the induced map $A\colon X/Z \to Y$ defined by $A(x+Z) = Tx$. This is well-defined bijective continuous linear map. We claim that $A^{-1}$ is continuous. We define $p(y) := \left\| A^{-1}(y) \right\|$. Note that $A^{-1}(y)$ is a coset of the from $x + \ker T$ in $X/Z$. If you recall how the quotient norm is defined, you will now understand why we defined $p$ as we did in the proof of OMT.

Let $\sum_n y_n$ be an absolutely convergent series in $Y$. Let $y := \sum_n y_n$. To prove that $p$ is countably subadditive, we may assume as usual that $\sum_n p(y_n)$ is convergent.

$$\begin{aligned} p\left(\sum_n y_n\right) = \left\| A^{-1}\left(\sum_n y_n\right)\right\| &= \left\| A^{-1}\left(\sum_n AA^{-1}y_n\right)\right\| \\ &= \left\| A^{-1}A\left(\sum_n A^{-1}y_n\right)\right\| \\ &= \left\| \sum_n A^{-1}y_n \right\| \\ &\leq \sum_n \left\| A^{-1}y_n \right\| = \sum_n p(y_n). \end{aligned} \tag{3}$$

We are not manipulating symbols. The equality (where we pulled $A$ out) needs justification. Can you see how to justify it?

> Observe that the series $\sum_n A^{-1}y_n$ is absolutely convergent since $\left\| A^{-1}y_n \right\| = p(y_n)$ and by our assumption $\sum_n p(y_n)$ is convergent. Hence $A\left(\sum_n A^{-1}y_n\right) = \sum_n AA^{-1}y_n$.

By Zabreiko's lemma, we conclude that $p$ is continuous. Hence there exists $C > 0$ such that $p(y) \leq C\,\|y\|$ for $y \in Y$. That is, $\left\| A^{-1}y \right\| \leq C\,\|y\|$ for $y \in Y$. Hence $A^{-1}$ is continuous. Therefore $A$ is a homeomorphism and in particular, it is an open map. Let $\pi\colon X \to X/Z$ be the quotient map. Then it is well-known that $\pi$ is an open continuous linear map. Since $T = A \circ \pi$, we conclude that $T$ is open, being the composition of two open maps.

We now state the original version of Zabreiko's lemma which is stronger than our version above.

Lemma 6 (Zabreiko) *Let $(X, \|\ \|)$ be a Banach space. Let $p$ be a countably subadditive seminorm on $X$: if $\sum_n x_n$ is absolutely convergent, then $p(\sum_n x_n) \leq \sum_n p(x_n)$. Then $p$ is continuous.*

*Proof.* Why is this a stronger form? We need to check only 'less number' of condi-

tions for the seminorm, namely, the we need to show $p(\sum_n x_n) \leq \sum_n p(x_n)$ only if $\sum_n x_n$ is absolutely convergent.

The proof we have given above still works. We need only observe that the sequence $(x_n)$ is such that $\sum_n y_n$ is absolutely convergent. As in (2), we see, by induction that

$$\begin{aligned}\|x_{n+1}\| &= \|x - (x_1 + \cdots + x_{n+1}) - (x - (x_1 + \cdots + x_n))\| \\ &\leq \|x - (x_1 + \cdots + x_{n+1})\| + \|x - (x_1 + \cdots + x_n)\| \\ &\leq 2^{-n}R + 2^{-n-1}R = 3R \cdot 2^{-n-1}.\end{aligned}$$

We deduce from this inequality that $\sum_n \|x_n\| < 3R$ and hence it is absolutely convergent. ■

*Remark* 2 Keep the notation of the proof of Zabreiko's lemma. We claim that $B(0,R) \subset A_N$. We prove this following the ideas in [2].

Let $x \in B(0,R)$. Let $0 < \|x\| < r < R$. Let $y := \frac{R}{r}x$. Then $y \in B(0,R) \subset \overline{A_N}$. Choose $\delta$ such that $0 < \delta < 1 - \frac{r}{R}$. Hence there exists $y_0 \in A_N$ such that $\|y - y_0\| < \delta R$. We estimate $\|y_0\|$:

$$\|y_0\| = \|y - y_0 - y\| \leq \|y - y_0\| + \|y\| < \delta R + R = R(1+\delta).$$

Look at $\frac{1}{\delta}(y - y_0) \in B(0,R) \subset \overline{A_N}$. Hence there exists $y_1 \in A_N$ such that

$$\left\| \frac{1}{\delta}(y - y_0) - y_1 \right\| < \delta R. \tag{4}$$

From (4) we deduce the following.

$$\begin{aligned}\|y_1\| &= \left\| y_1 - \frac{1}{\delta}(y - y_0) + \frac{1}{\delta}(y - y_0) \right\| \\ &\leq \left\| y_1 - \frac{1}{\delta}(y - y_0) \right\| + \left\| \frac{1}{\delta}(y - y_0) \right\| \\ &\leq \delta R + R = R(1+\delta). \end{aligned} \tag{5}$$

Again from (4), we find that

$$\left\| \frac{1}{\delta^2}(y - y_0 - \delta y_1) \right\| < R. \tag{6}$$

Since $\frac{1}{\delta^2}(y - y_0 - \delta y_1) \in B(0,R) \subset \overline{A_N}$, there exists $y_2 \in A_N$ such that

$$\left\| \frac{1}{\delta^2}(y - y_0 - \delta y_1) - y_2 \right\| < \delta R. \tag{7}$$

As earlier, from (7) we make two observations.

$$\begin{aligned}\|y_2\| &\leq \delta R + \left\| \frac{1}{\delta^2}(y - y_0 - \delta y_1) \right\| \\ &< \delta R + R = R(1+\delta), \text{ by (6)}. \end{aligned} \tag{8}$$

The second observation is this:

$$\left\| \frac{1}{\delta^3}(y - y_0 - \delta y_1 - \delta^2 y_2) \right\| < R. \tag{9}$$

Assume that we have got a sequence $(y_k)_{k=0}^n$ such that

$$\left\| \frac{1}{\delta^n}(y - \sum_{k=0}^{n} \delta^k y_k) \right\| < R, \text{ that is, } \left\| y - \sum_{k=0}^{n} \delta^k y_k \right\| < \delta^n R \tag{10}$$

Hence there exists $y_{n+1} \in A_N$ such that

$$\left\| \frac{1}{\delta^n} \left( y - \sum_{k=0}^{n} \delta^k y_k \right) - y_{n+1} \right\| < \delta R. \tag{11}$$

From (11), we obtain

$$\left\| y - \sum_{k=0}^{n+1} \delta^k y_k \right\| < \delta^{n+1} R. \tag{12}$$

Thus we have a sequence $(y_k)_{k=0}^{n+1}$ and satisfying (10) with $n$ replaced by $n+1$. Also, from (11), we get

$$\| y_{n+1} \| < \delta R + \left\| \frac{1}{\delta^n} \left( y - \sum_{k=0}^{n} \delta^k y_k \right) \right\| < \delta R + R, \text{ by (10).} \tag{13}$$

Thus, we get a sequence $(y_n)_{n \in \mathbb{Z}_+}$ in $A_N$ such that

$$\left\| y - \sum_{k=0}^{n} \delta^k y_k \right\| < \delta^n R, \quad \text{for each } n \in \mathbb{Z}_+.$$

Hence we see that the series converges to $y$: $y = \sum_{k=0}^{\infty} \delta^k y_k$. Also, we see that

$$\sum_{k=0}^{\infty} \left\| \delta^k y_k \right\| \leq \sum_{k=0}^{\infty} \delta^k (R(1+\delta)).$$

Hence the series $\sum_{k=0}^{\infty} \delta^k y_k$ is absolutely convergent.

By hypothesis,

$$p(y) = p\left( \sum_k \delta^k y_k \right) \leq \sum_k \delta^k p(y_k) = \frac{1}{1-\delta} N.$$

Hence $p(x) = p(\frac{r}{R} y) < \frac{r}{R} \frac{1}{1-\delta} N < N$. Thus, $x \in A_N$.

Let $\varepsilon > 0$ be fixed. Let $x \in X$ be given. Let $\lambda := \frac{R}{(1+\varepsilon)\|x\|}$. Then $\lambda x \in B(0, R)$ and hence $p(\lambda x) < N$, that is,

$$p(x) < \frac{N}{\lambda} = \frac{(1+\varepsilon)N}{R} \|x\|.$$

Thus $p$ is continuous. ■

*Remark* 3 There is an interesting background to our article. When we were working on a book on functional analysis, we proved the standard observation that closed graph theorem (CGT) is equivalent to open mapping theorem (OMT). Recall that it is a standard practice to prove open mapping theorem first and then derive closed graph theorem and then show that the CGT implies OMT. Thus OMT and CGT are equivalent. We started wondering whether there is direct proof of closed graph theorem. We stumbled upon a classic [1]. Kato gives a direct proof of CGT. We found that the argument is very similar to the one for OMT (and similar to the proof of Zabreiko's lemma, more precisely where we establish $B(0,R) \subset A_{2N}$.) This made us ask the obvious question: Is there some underlying principle from which we can deduce CGT and OMT? A google search brought [2] to our attention.

If this article makes Zabreiko's lemma well-known among students of Functional Analysis, our purpose would be served.

**Acknowledgement:** This article is based on the post [2]. I thank D. Sukumar, IIT-Hyderabad and G. Santhanam, IIT-Kanpur for a careful reading of the article and valuable suggestion for its improved readability.

## References

[1] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Reprint of the 1980 edition.

[2] https://math.blogoverflow.com/2014/06/25/zabreikos-lemma-and-four-fundamental-theorems-of-functional-analysis/

[3] P. P. Zabreiko, A theorem for semiadditive functionals, Functional analysis and its applications 3 (1), 1969, 70-72)

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 113–123

# *Gromov's Theory of $h$-Principle*

**Mahuya Datta**[a*]

[a] *Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata 700108*

**Abstract:** We briefly discuss the results of Nash, Smale and Hirsch which provide a background to the theory of $h$-principle due to Gromov. Following this, we outline the general theory of $h$-principle and mention some of its applications in Symplectic and Contact geometry, focussing on the period 1969 - 1985.

**Keywords:** $h$-principle, isometric immersions, symplectic and contact immersions.

**AMS Subject Classifications**: 58B, 58D

## 1. Introduction

The theory of $h$-principle deals with partial differential equations or more general relations which arise in Topology and Geometry. Such differential relations have a very distinctive feature: Very often the system is under-determined and have many solutions. Furthermore, the solution spaces are dense in the space of all admissible functions. This is in contrast with the PDE's which arise in Physics where the solutions are few. This aspect of the analytic problems in Geometry and Topology gives rise to homotopy theoretic approach in addressing analytic questions.

General theory of $h$-principle ($h$ for homotopy) unfurled through a series of papers by M. Gromov in the late 60's and early years of 70's, the seed of which was laid by Whitney, Nash, Smale and Hirsch during the 50's and 60's. We recall some of the major results that were proved during this period:

### 1.1. *$C^1$-isometric immersion theorem*

THEOREM 1.1 (Nash-Kuiper[16],[13]) *Let $(M, g)$ be a Riemannian manifold of dimension $n$. Let $h$ denote the canonical metric on $\mathbb{R}^q$. If $q > n$ and $M$ admits a smooth immersion $f_0$ (resp. embedding) in $\mathbb{R}^q$, then there exists a $C^1$-immersion (resp., embedding) $f : M \to \mathbb{R}^q$ such that $f^*h = g$. Moreover, $f$ is homotopic to $f_0$ via smooth immersions.*

Nash proved the theorem for $q > n+1$ and conjectured that it can be improved to $q > n$. Soon after that, Kuiper showed that the conjecture is indeed true; however

*Corresponding author. Email: mahuya@isical.ac.in

his techniques were much more involved than that of Nash.

The isometric immersion relation is locally defined by a system of first order partial differential equations; hence it is a closed relation. Indeed if $\{g_{ij}\}$ are the coefficients of the metric tensor $g$, then the isometry relation can be expressed locally by a system of partial differential equations:

$$\frac{\partial f}{\partial x_i}.\frac{\partial f}{\partial x_j} = g_{ij}, \quad 1 \leq i \leq j \leq n \tag{1}$$

Theorem 1.1 is counter-intuitive at the first glance, since the above system of PDE is over-determined for $q < n(n+1)/2$. Hence, Nash's result came as a big surprise, for it implies that isometric $C^1$ immersions (resp., embeddings) are plenty and, in fact, dense in the space of all strictly short immersions (resp., embeddings). As an interesting consequence of the above result, one gets that a Torus with flat metric embeds $C^1$-isometrically in $\mathbb{R}^3$. Clearly this can not be true for $C^2$ immersions, since the curvature tensor associated to the metric $g$ must be preserved. Thus $C^1$ isometric immersions are not geometrically interesting. However, the result shows that there are plenty of such pathological solutions to the isometric immersion equation.

Nash described an explicit and very elegant way to obtain a $C^1$ solution to the isometric immersion problem. It is obtained as the $C^1$-limit of a sequence $\{f_k\}$ of approximate $C^\infty$ solutions to the given equation ($\varepsilon$-approximate solution means that the difference between the induced metric and $g$ is $\varepsilon$-small in the $C^0$ norm). The sequence is constructed recursively and the $k$-th approximation $f_k$ is obtained from the $(k-1)$-th approximate solution using some twisting maps which deforms a function only locally. Each approximate solution in the sequence is an improved one over the preceeding maps in the sequence. The process converges to a $C^1$ isometric map as the $C^1$-distance between two successive approximation is kept under control.

## 1.2. *$C^\infty$ Isometric Immersion theorem*

$C^\infty$ Isometric Immersion theorem (1956) [17] is the most celebrated result of Nash. It involves a sophisticated technique of proving the existence of $C^\infty$ isometric immersions of a Riemannian manifold $(M, g)$ into some Euclidean space $\mathbb{R}^q$, by establishing and appealing to an infinite-dimensional Implicit Function Theorem (IFT) for the isometric immersion operator

$$\mathcal{D} : C^\infty(M, \mathbb{R}^q) \to \text{ Riemannian metrics on } M$$

defined by $\mathcal{D}(f) = f^*h$, where $h$ is the canonical riemannain metric on $\mathbb{R}^q$. The operator $\mathcal{D}$ is a non-linear first order partial differential operator. Nash observed that the linearization of $\mathcal{D}$ at a free map $f$, denoted by $L_f$, has a right inverse $M_f$, which is a linear differential operator of zero-th order. A smooth map is said to be *free* if its first and the second partial derivatives at each point form a linearly independent set. Thus the set of free maps constitute an open subspace in the fine $C^\infty$-topology. It follows from the Implicit Function Theorem that $\mathcal{D}$ is locally invertible at free maps and hence the image of the space of free maps under $\mathcal{D}$ is open in the space of Riemannian metrics on $M$. A necessary condition for the existence of free maps and hence for the infinitesimal inversion is that $q > q_0 = n + n(n+1)/2$.

Starting with a free approximate solution $f$ to the equation $\mathcal{D} = g$, one can construct by Newton's method a sequence of free approximate solutions by using the infinitesimal inversions $M_f$ of $\mathcal{D}$ at free maps. But the Newton's process does not necessarily converge. Nash's major contribution is to introduce certain smoothing operators in the process to make it converge. This proves local invertibility of $\mathcal{D}$ at free maps.

By strengthening the technique employed in $C^1$-isometric immersion theorem, Nash showed that a free map $f_0$ can be homotoped in any given $C^3$ neighbourhood of it to an approximate solution $f$ so that $f^*h$ is $\varepsilon$-close to $g$ for an arbitrary $\varepsilon > 0$. This step requires that $q$ must be bigger than $q_0 + q_1$, where $q_1 \geq n(n+3)$. If $\varepsilon$ is small enough, then the implicit function theorem guarantees an $f'$ satisfying the relation $\mathcal{D}(f') = g$.

THEOREM 1.2 ([17]) *A compact $n$-manifold with a $C^k$ positive metric has a $C^k$ isometric imbedding in any small volume of euclidean $(n/2)(3n+11)$-space, provided $3 \leq k \leq \infty$.*

## 1.3. *Homotopy Classification of Immersions*

During 1958-59, S. Smale ([20], [21]) gave a complete homotopy classification of immersions of spheres in Euclidean spaces. M. Hirsch [12] followed up this work by proving a homotopy classification of immersions between two arbitrary manifolds. Unlike the previous examples, the immersions are not solutions to any equations, in fact, they satisfy an open condition.

The Smale-Hirsch theorem may be explained as follows: If $f : M \to N$ is a smooth immersion then its derivative $\mathrm{d}f : TM \to TN$ is a bundle monomorphism. Let $\mathrm{Imm}(M, N)$ denote the space of smooth immersions with $C^\infty$ compact open topology and let $\mathrm{Mono}\,(TM, TN)$ denote the space of bundle monomorphisms $(F, f) : TM \to TN$ with $C^0$-compact open topology.

THEOREM 1.3 (Smale-Hirsch Immersions Theorems [20],[21],[12]) *If* $\dim M < \dim N$ *then*

$$d : \ Imm(M, N) \to Mono\,(TM, TN)$$

*is a weak homotopy equivalence.*

In the modern language, the result can be restated as follows: That the $C^\infty$ immersions satisfy the parametric $h$-principle. This is the first instance of complete $h$-principle. The proof of the theorem utilizes a handle-body decomposition of $M$:

$$M_0 \subset M_1 \subset \cdots \subset M_{k-1} \subset M_k \subset \ldots M$$

where $M_0$ is a disc and $M_k$ is obtained from $M_{k-1}$ by attaching either a collar neighbourhood or a handle along the boundary of $M_{k-1}$. A formal immersion is first made homotopic to a $C^\infty$ immersion over $M_0$ and then extended to each $M_k$, $k = 1, 2, \ldots$ successively to get a global immersion.

Smale-Hirsch Immersion Theorem implies that the topology of the space of immersions between two manifolds is completely captured by the topology of the space of bundle monomorphisms between their tangent bundles. While the question of existence of immersions is a differential topological problem, the existence of a bundle monomorphism is purely an algebraic one which can be addressed with

algebraic topological techniques like obstruction theory.

It is a simple fact that the identity map of the sphere is not homotopic to the antipodal map. However, Smale's theorem concludes that the inclusion map of $S^2$ in $\mathbb{R}^3$ is regularly homotopic to the antipodal embedding $a : S^2 \to \mathbb{R}^3$ which takes $x$ to $-x$. This is known as the *Sphere eversion theorem.*

The above three work are foundation to the theory of $h$-principle. Among other contemporary work that preceded Gromov's 1969 article are the homotopy classification of submersions [18] and transversal maps [19] due to A. Phillips and classification of $k$-mersions by S. Feit [5].

Around 1969, Gromov [8] introduced a general theory for partial differential relations, known as the theory of $h$-principle, which brought all these work within a common framework. Theory of $h$-principle provides a homotopy theoretic approach to address the questions of existence and classification of solutions of partial differential relations which arise in topology and geometry.

## 2. Partial Differential Relations and *h*-principle

Partial differential equations are, most often, difficult to solve. In fact, there is no satisfactory theory to deal with a general PDE. The theory of $h$-principle provides a general approach to deal with a large class of partial differential relations arising in geometry and topology.

The main objective is to understand the topological types of the space of smooth maps - namely immersions, submersions, isometric maps as observed above or the space of geometric structures on manifolds - namely Riemannian metrics, symplectic forms, distributions - which are characterized by differential relations.

### 2.1.

An $r$-th order partial differential equation for functions $f : \mathbb{R}^n \to \mathbb{R}^m$ is a relation between partial derivatives of $f$ up to order $r$ which is expressed in the form of an equation. To each $C^r$ map $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ and a point $x$ we can associate the tuple

$$(f(x), \frac{\partial f}{\partial x_i}(x), \ldots, \frac{\partial^r f}{\partial x_{i_1} \ldots \partial x_{i_r}}(x))$$

which is called the *$r$-jet of $f$ at $x$.* A general $r$-th order partial differential equation for functions $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is then given as follows:

$$\Psi(x, f(x), \frac{\partial f}{\partial x_i}(x), \ldots, \frac{\partial^r f}{\partial x_{i_1} \ldots \partial x_{i_r}}(x)) = 0, \tag{2}$$

where $\Psi$ is a continuous real valued map on $\mathbb{R}^q$, $q = n + \binom{n}{1} + \cdots + \binom{n}{r}$.

Replacing the partial derivative functions in (2) by continuous functions $a_{i_1,\ldots,i_k}$'s we get an algebraic equation:

$$\Psi(x, a_0(x), a_i(x), \ldots, a_{i_1,\ldots,i_r}(x)) = 0. \tag{3}$$

A solution to equation (3) is called a formal solution of (2). If the given differential equation (2) has a solution then clearly (3) is solvable by a $q$-tuple of continuous functions, where $q = n + \binom{n}{1} + \cdots + \binom{n}{r}$. As such there is no reason to believe

that the converse - that the solvability of the algebraic equation (3) will imply the solvability of the PDE (2) - would be true. However, since equation (3) is much simpler to deal with, it is worth investigating if there is any condition on (3) which would guarantee the converse. In essence, the PDE (2) is said to satisfy the $h$-principle provided the converse has an affirmative answer.

`Main Question.` When can we reduce the solvability of equation (2) to that of (3). Equivalently, when does the existence of formal solution to equation (2) implies that (2) is solvable?

## 2.2. *Formal definitions*

A manifold $M$ is locally homeomorhic to an open subset of an Euclidean space; however, there is no canonical coordinate system on a manifold. Therefore, the definition of a partial differential equation on manifolds must follow a coordinate free language.

Let $f : M \longrightarrow \mathbb{R}$ be a smooth map. The $r$-th order infinitesimal information of $f$ at a point $x$ is encoded in the $r$-jet

$$j^r_f(x) = (f(x), Df(x), D^2f(x), \ldots, D^rf(x)),$$

where $D^kf(x)$ denotes the $k$-th derivative of $f$ at $x$ which is a symmetric $k$-multilinear map on $T_xM$ with values in $T_{f(x)}N$. The space of $r$-jets of germs of smooth maps between manifolds $M$ and $N$ is denoted by $J^r(M,N)$. There is a canonical map $p^{(r)} : J^r(M,N) \to M$ which takes $j^r_f(x)$ onto $x$. This defines a smooth bundle over $M$. Let $\Gamma(J^r(M,N))$ denote the space of sections of the $r$-jet bundle.

If we endow $C^\infty(M,N)$ with $C^\infty$ compact open topology and $\Gamma(J^r(M,N))$ with $C^0$-compact open topology then the $r$-jet map:

$$j^r : C^\infty(M,N) \to \Gamma(J^r(M,N))$$

is continuous. Moreover, $j^r$ is injective and so $C^\infty(M,N)$ embeds in $\Gamma(J^r(M,N))$. An element in the image of $j^r$ is called a holonomic section of $p^{(r)}$.

It is easy to see that the space of sections of the 1-jet bundle $J^1(M,N)$ can be identified with the space $\mathrm{Hom}\,(TM,TN)$ consisting of bundle morphisms $(F,f) : TM \to TN$, where $F_x : T_xM \to T_{f(x)}N$ is a linear map for all $x \in M$. Then the 1-jet map can be replaced by the derivative map:

$$\mathrm{d} : C^\infty(M,N) \to \mathrm{Hom}(TM,TN)$$

More generally, one can consider a smooth fibration $p : E \to M$ and the associated $r$-jet bundle $p^{(r)} : E^{(r)} \to M$, where $(p^{(r)})^{-1}(x) = E^{(r)}_x$ is the space of $r$-jets of germs of local sections of $p$ at $x$. The $r$-jet map in this case is denoted by $j^r : \Gamma^\infty(E) \to \Gamma(E^{(r)})$. If $E$ is a product bundle over $M$ with fibre $N$ then $\Gamma^\infty(E)$ can be identified with $C^\infty(M,N)$.

*Definition* 1 ([10]) An *$r$-th order partial differential relation* for sections of a fibre bundle $p : E \to M$ (in particular, functions $M \to N$) is defined as a subset $\mathcal{R}$ of $E^{(r)}$ (resp. $J^r(M,N)$).

To illustrate this definition, consider an $r$-th order partial differential equation for smooth maps $f : M \to N$ between manifolds which can be described in the

following form:

$$\Psi(f(x), Df(x), D^2f(x), \ldots, D^rf(x)) = 0, \tag{4}$$

where $\Psi : J^r(M,N) \to \mathbb{R}$ is a function on the total space of the $r$-jet bundle $J^r(M,N)$. Using the $r$-jet prolongation $j^r_f$ we can rewrite it as

$$\Psi(j^r_f(x)) = 0 \quad \text{for all } x \in M;$$

Thus, the subset $\mathcal{R} = \Psi^{-1}(0)$ in $J^r(M,N)$ encodes the partial differential equation (4), and $f$ is a solution if and only if $j^r_f$ has its image inside $\mathcal{R}$.

*Definition* 2 ([10]) A smooth section $f : M \to E$ (resp. a map $f : M \longrightarrow N$) is said to be a *solution* of $\mathcal{R}$ if its $r$-jet prolongation $j^r_f : M \to E^{(r)}$ maps into the subset $\mathcal{R}$; in other words, $j^r_f$ is a section of $\mathcal{R}$.

A section $\sigma : M \to E^{(r)}$ (resp. $\sigma : M \to J^r(M,N)$) of the jet bundle which maps $M$ into $\mathcal{R}$ is said to be a *formal solution* of $\mathcal{R}$.

A section of the jet bundle which is of the form $j^r_f$ (for some section $f$ of $E$) is said to be a *holonomic section.*

*Example* 1 The following classes of maps are solutions to first order partial differential relations: (a) Immersions, (b) Submersions, (c) transversal maps, (d) symplectic forms, (e) contact forms, (f) isometric immersions, (g) symplectic immersions, (h) contact immersions. The free maps are solutions of a second order partial differential relation.

Let $\Gamma(\mathcal{R})$ denote the subspace of $\Gamma(E^{(r)})$ consisting of formal solutions of $\mathcal{R}$. We denote the space of solutions of $\mathcal{R}$ by $\operatorname{Sol}\mathcal{R} \subset \Gamma^\infty(E)$, and endow it with the $C^\infty$ compact-open topology. Then the $r$-jet map $j^r : \operatorname{Sol}\mathcal{R} \longrightarrow \Gamma(\mathcal{R})$ is continuous. Since $j^r$ is injective, we can identify the space of solutions of $\mathcal{R}$ with the space of holonomic sections of $\mathcal{R}$.

*Definition* 3 A relation $\mathcal{R}$ is said to satisfy the *h-principle* if every formal solution $\sigma$ of $\mathcal{R}$ can be joined by a path in $\Gamma(\mathcal{R})$ to a holonomic section. In other words, there exists a solution $f$ of $\mathcal{R}$ such that $j^r_f$ is homotopic to $\sigma$ through formal solutions of $\mathcal{R}$.

$\mathcal{R}$ is said to satisfy the *parametric h-principle* if $j^r$ induces bijections between the homotopy groups of the two spaces:

$$\pi_k(j^r) : \pi_k(\operatorname{Sol}\mathcal{R}) \longrightarrow \pi_k(\Gamma(\mathcal{R})), \quad k = 0, 1, 2, \ldots.$$

Hence, $h$-principle reduces a differential topological problem to an algebraic one. It does not guarantee the existence of solutions but says that the only obstruction to the existence is topological. Existence of solution would follow from the $h$-principle provided the topological obstruction vanishes.

## 3. Sheaf theoretic and Analytic techniques in *h*-principle

Gromov introduced several global techniques [10] in the theory of $h$-principle which may be applied to a large class of differential relations.

- Sheaf theoretic technique - this has its origin in the Smale-Hirsch theory.

- Convex integration technique [9] - the theory is based on Kuiper's approach to $C^1$-isometric embedding problem
- Analytic technique - this is rooted in Nash's work on $C^\infty$ isometric immersion theorem

The theory has given a new direction to study the differential relations which appear in Geometry. It has unified many results that were already existing; at the same time it has produced a number of new results as simple corollaries. Here we shall discuss about the sheaf theoretic and the analytic techniques.

Throughout we consider $E \to M$ to be a smooth fibration and $\mathcal{R} \subset E^{(r)}$ an $r$-th order partial differential relation.

### 3.1. *Gromov's theorem for Open Diff-invariant relations and sheaf technique*

The first general result in the theory of $h$-principle concerns about relations $\mathcal{R}$ which are open subsets of jet spaces, referred as *open differential relations.*

THEOREM 3.1 ([8]) *Let $M$ be an open manifold. Then every open, Diff$(M)$-invariant differential relation satisfies the parametric h-principle.*

Here Diff$(M)$ denotes the pseudo-group of local diffeomorphisms of $M$. If the fibration $E$ is a natural bundle, e.g., a product bundle, or a tensor bundle then there is a natural pull-back action of Diff$(M)$ on the space of sections of $E$. This action, in turn, induces an action on the jet-space $E^{(r)}$ . Generally, if there is a Diff$(M)$ action on the jet space which keeps $\mathcal{R}$ invariant then we say $\mathcal{R}$ is Diff$(M)$ invariant. The hypothesis of Theorem 3.1 is purely topological. The theorem manifests how topology can control answers to analytic questions.

The structure of the proof is based on the following theorem.

THEOREM 3.2 (Sheaf Homomorphism Theorem [10]) *Let $\Phi$ and $\Psi$ be two topological sheaves on a manifold $M$ and $\alpha : \Phi \to \Psi$ is a sheaf homomorphism. If the sheaves $\Phi$ and $\Psi$ are flexible and $\alpha$ is a local weak homotopy equivalence then $\alpha$ is a weak homotopy equivalence. In other words, $\alpha$ induces isomorphisms between the homotopy groups of $\Phi$ and $\Psi$.*

A sheaf is said to be flexible if for every pair of compact subsets $(A, B), A \supset B$, the restriction map $\Phi(A) \to \Phi(B)$ is a Serre fibration. By a local weak homotopy equivalence we mean that the restrictions of $\alpha$ to the stalks, $\alpha_x : \Phi(x) \to \Psi(x)$, $x \in M$, are weak homotopy equivalences.

In order to prove Theorem 3.1, we take $\Phi$ to be the sheaf of solutions of the relation $\mathcal{R}$ and $\Psi$ the sheaf of sections of the jet bundle with images contained in $\mathcal{R}$. Then the $r$-jet map $j^r : \Phi \to \Psi$ is a sheaf homomorphism. Now, it is a topological fact that $\Psi$ is a flexible sheaf. Since $\mathcal{R}$ is an open relation, $j^r$ is a local weak homotopy equivalence, in fact, openness of $\mathcal{R}$ directly implies that if $j^1_f(x) \in \mathcal{R}$ then $f$ is a local solutions of $\mathcal{R}$ near $x$. The subtle part is the proof of flexibility of the solution sheaf which requires the full strength of the hypothesis of the theorem.

Flexibility of the solution sheaf does not hold true if the manifold $M$ is closed. An open $n$-dimensional manifold has a handle body decomposition having no handle of top index and so the manifold is homotopically equivalent to an $(n-1)$-dimensional CW complex $K$. In the proof of Theorem 3.1, one first proves flexibility of the sheaf of solutions near $K$ and here the positivity of the codimension of $K$ is crucial.

The theorem recovers Phillips trasnversality theorem [19], Smale-Hirsch Immersion theorem. At the same time it gives several new results on open manifolds in Symplectic and Contact geometry.

### 3.2. *Applications in Symplectic and Contact geometry*

A 2-form $\omega$ on $M$ is said to be *non-degenerate* if $\omega^n$ is a volume form, where $\dim M = 2n$. Such a form defines a bijection between the space of vector fields on $M$ and the space of 1-forms on $M$:

$$X \mapsto i_X \omega$$

where $X$ is a vector field on $M$ and $i_X$ denotes the contraction operator. In fact, the bijection is induced by a bundle map $TM \to T^*M$. A non-degenerate 2-form is called *symplectic* if it is closed, that is $\mathrm{d}\omega = 0$ which imposes a differential condition on $\omega$.

A 1-form $\alpha$ on an odd-dimensional manifold $M$ is called a *contact form* if $\alpha \wedge (\mathrm{d}\alpha)^n$ is non-vanishing, where $\dim M = 2n + 1$. In other words, $d\alpha$ is a symplectic form on the sub-bundle $\ker \alpha$.

Non-degeneracy is clearly an open condition. Moreover, symplectic and contact conditions are preserved under the pull-back action of $\mathrm{Diff}(M)$ on forms. Hence it follows that both symplectic forms and contact forms on open manifolds satisfy the parametric $h$-principle. These results may be interpreted as follows:

COROLLARY 3.3 *If $M$ is an open manifold then the space of non-degenerate 2-forms on $M$ has the same weak homotopy type as the space of exact symplectic forms on $M$. Furthermore, if $\alpha \in H^2_{deR}(M)$ then the space of symplectic forms representing the class $\alpha$ has the same weak homotopy type as the space of non-degenerate 2-forms on $M$.*

COROLLARY 3.4 *Let $M$ be an open manifold of dimension $2n + 1$. The space of contact forms on $M$ has the same weak homotopy type as the space of pairs $(\beta, \theta)$ consisting of a 1-form $\beta$ and a 2-form $\theta$ such that $\beta \wedge \theta^n$ is non-vanishing.*

The above results imply that the obstruction to the existence of symplectic and contact forms on open manifolds are purely topological.

## 4. Non-open, non-Diff invariant relations

The statement of Theorem 3.1 is quite illusive. It appears to have a limited scope of application as it only considers open $\mathrm{Diff}(M)$- invariant relations. However, the scope of sheaf technique goes far beyond this and is applicable to many interesting closed differential relations which arise from the partial differential equations. The open-ness condition on the relation has two direct consequences - it implies that the solution sheaf is *microflexible*, a weaker notion than flexibility (where only an initial part of the homotopy $F$ in a homotopy lifting problem is required to have a lift) and that it satisfies the local $h$-principle. In fact, only these two properties of open relations are exploited in the proof of Theorem 3.1. There are interesting examples of non-open relations which satisfy these two properties. Flexibility of solution sheaf, as observed above, is a difficult proposition to prove. However, microflexibility, a weaker notion than flexibility, follows rather easily for many

relations.

On the other hand invariance of $\mathcal{R}$ under the action of full Diff($M$) group is also not essential. What is needed is an acting subgroup $\mathcal{D} \subset \mathrm{Diff}(M)$ which contains enough 'sharp diffeotopies'. Roughly speaking, a sharp diffeotopy is a compactly supported diffeotopy $\delta_t$ such that the final map $\delta_1$ moves a given submanifold of positive codimension sharply away from itself.

THEOREM 4.1 ([10]) *Let $\mathcal{R}$ be a relation for which the solution sheaf $\Phi$ is microflexible. Let $\mathcal{R}$ be invariant under an action of a subgroup $\mathcal{D}$ of Diff($M$). If $\mathcal{D}$ contains sharp diffeotopies, then the sheaf $\Phi|_N$ is flexible for any submanifold $N$ of $M$ of positive codimension. In addition, if $\mathcal{R}$ satisfies the local h-principle then $j^r : \Phi|_N \to \Psi|_N$ is a weak homotopy equivalence.*

Note that, there is no restriction on the manifold $N$ and it may also be a closed manifold. Positive codimension condition on $N$ provides extra space for acting diffeotopies to move $N$ sharply away from itself. This is crucial for getting flexibility of the sheaf $\Phi|_N$ ($\Phi|_N$ consists of solutions which are defined on some unspecified open neighbourhood of $N$ in $M$). In the previous theorem, the open-ness condition on $M$ provides an extra dimension to move $K$ sharply by diffeotopy.

For dealing with differential relations on a closed manifold $M$, the idea is to embed the manifold in a higher dimensional manifold $\tilde{M}$ and to consider an appropriate extension relation $\tilde{\mathcal{R}}$ on it, in the sense that every $r$-jet in $\mathcal{R}$ is extendable to an $r$-jet in $\tilde{\mathcal{R}}$, and solutions to $\tilde{\mathcal{R}}$ when restricted to $M$ give solutions of $\mathcal{R}$.

## 4.1. *Applications to Symplectic and Contact immersions*

*Definition* 4 Let $(N, \sigma)$ be a symplectic manifold with a symplectic form $\sigma$ and $M$ be any manifold with a closed 2-form $\omega$ on it. A smooth immersion $f : M \to N$ will be called a *symplectic immersion* if it pulls back the form $\sigma$ onto $\omega$.

Any symplectic immersion $f$ must pull back the de Rham cohomology class of $\sigma$ onto that of $\omega$, that is $f^*[\sigma] = [\omega]$. Notice that this cohomology condition is preserved under homotopy.

Let $Symp_{\omega,\sigma}(M, N)$ denote the space of smooth symplectic immersions and $Symp_{\omega,\sigma}(TM, TN)$ denote the space of bundle maps $(F, f) : TM \to TN$ such that $F$ is injective linear, $F^*\sigma = \omega$ and $f^*[\sigma] = [\omega]$.

THEOREM 4.2 (Symplectic Immersion Theorem [10]) *If $\dim M < \dim N$ then*

$$d : Symp_{\omega,\sigma}(M, N) \to Symp_{\omega,\sigma}(TM, TN)$$

*is a weak homotopy equivalence.*

The complex projective space $\mathbb{C}P^q$ has a symplectic form $\sigma$ which is $U(q+1)$ invariant and which is normalized by $\langle [\sigma], [\mathbb{C}P^1] \rangle = 1$. As a corollary to the above theorem one can deduce that every symplectic manifold $(M, \omega)$ of dimension $n$ admits a symplectic immersion into $\mathbb{C}P^q$ provided the cohomology class of $\omega$ is integral and $q \geq \dim M$.

Next we state the Contact Immersion Theorem. Let $(N, \xi)$ be a contact manifold, where the contact structure $\xi$ is defined as the kernel of a 1-form $\beta$. Then $d\beta$ restricts to a symplectic structure on the subbundle $\xi$. Let $M$ be a manifold with a co-rank 1 distribution $\eta$ defined by a 1-form $\alpha$. A smooth immersion $f : M \to N$ is said to be a contact immersion if $(df)^{-1}\xi = \eta$. The derivative of any such $f$

maps $\eta$ into the contact distribution $\xi$ and pulls back the symplectic structure $d'\beta = d\beta|_{\ker \xi}$ onto $d'\alpha = d\alpha|_{\eta}$.

THEOREM 4.3 (Contact Immersion Theorem [10], [4]) *If* $\dim M < \dim N$ *then the space of contact immersions* $(M, \eta) \to (N, \xi)$ *is weak homotopy equivalent to the space of bundle monomorphism* $F : TM \to TN$ *which satisfy the following two conditions:* $F^{-1}\xi = \eta$ *and* $F^*(d'\beta) = d'\alpha$.

Symplectic and contact immersions are defined as solutions to certain differential equations. Hence, the associated differential relations are closed relations. However, both the sheaves - sheaf of symplectic immersions and the sheaf of contact immersions - are microflexible. This is easily seen from the Moser's stabilty theorem [14] and Gray's stability theorem [7]. In fact, the stability theorems provide local inversions of the differential operators associated with the symplectic immersions and contact immersions. On the other hand, the relations in both cases satisfy the local $h$-principle by Darboux theorem.

## 4.2. *Analytic Technique*

In view of Theorem 4.1, it is important to understand which relations give rise to microflexible solution sheaves and satisfy the local $h$-principle. As we have observed in the previous section, if the relation is open, then the sheaf of solutions is microflexible and we readily get local $h$-principle.

Among the closed relations are the ones that are associated with some smooth partial differential operator $\mathcal{D}$. The solutions in that case arise as the solution to the equation $\mathcal{D}f = g$ for a given $g$.

Nash had observed in [17] that the Implicit Function Theorem may be formulated for any smooth differential operator which admit appropriate infinitesimal inversions. Working in the general set up, Gromov proves that if the operator is infinitesimally invertible on some subspace $\mathcal{A}$ consisting of smooth solutions to some open relation, referred to as $\mathcal{A}$-regular maps, then the image of $\mathcal{A}$-regular maps under $\mathcal{D}$ is an open subset. For isometric immersion operator, $\mathcal{A}$-regular maps are the free maps.

Gromov observed that the local inversions $\mathcal{D}_f^{-1}$ of the operator $\mathcal{D}$ at $f \in \mathcal{A}$, constructed via IFT have the 'locality' property. Since the local inverses are not differential operators, they may not preserve the support of the function. However, it can be ensured that the value of $\mathcal{D}_f^{-1}(g)$ at any $v \in M$ depends only on the value of $g$ on the ball of radius 1 about $v$ with respect to a pre-fixed metric on $M$. Using this property, Gromov proves that the sheaf of $\mathcal{A}$-regular solutions of $\mathcal{D}(f) = g$ is microflexible. Further, it can be proved that an infinitesimal solution of the equation $\mathcal{D} = g$ can be homotoped to a local solution provided regularity is assumed. However, one may have to consider an infinitesimal solution in the $s$-jet space for some $s > r$. Combined with sheaf technique, this now gives global $h$-principle.

There are number of applications of the analytic and sheaf theoretic technique. Gromov proved $h$-principle for $C^\infty$ isometric immersions and thereby improved Nash's theorem by reducing the dimension of $\mathbb{R}^q$, provided the dimension of the domain is not too small. He also considered the isometric immersion problem on pseudo-riemannian manifolds. The Symplectic immersion theorem and Contact immersion theorem also follow from this general analytic technique if one wishes to avoid the specific stability theorems in the Symplectic and Contact geometry. We refer to Gromov's book [10] for plenty of applications of this theory.

### 4.3. *Concluding remark*

There is an overwhelming presence of $h$-principle in geometry and topology. Theory of $h$-principle also extends to holomorphic set up on Stein manifolds [11] [6]. Recent developments in the Symplectic and Contact geometry with several new $h$-principle type results ([3], [1], [2], [15]) have renewed the interest in theory of $h$-principle.

## References


[1] Matthew Strom Borman, Yakov Eliashberg, and Emmy Murphy. Existence and classi cation of overtwisted contact structures in all dimensions. *Acta Math.*, 215(2):281-361, 2015.

[2] Kai Cieliebak and Yakov Eliashberg. *From Stein to Weinstein and back, volume 59 of American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2012. Symplectic geometry of affine complex manifolds.

[3] Y. Eliashberg. Classification of overtwisted contact structures on 3-manifolds. *Invent. Math.*, 98(3):623-637, 1989.

[4] Y. Eliashberg and N. Mishachev. *Introduction to the h-principle*, volume 48 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

[5] Sidnie Dresher Feit. k-mersions of manifolds. *Acta Math.*, 122:173-195, 1969.

[6] Franc Forstnerič. *Stein manifolds and holomorphic mappings, volume* 56 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Cham, second edition, 2017. The homotopy principle in complex analysis.

[7] John W. Gray. Some global properties of contact structures. *Ann. of Math.* (2), 69:421-450, 1959.

[8] M. L. Gromov. Stable mappings of foliations into manifolds. *Izv. Akad. Nauk SSSR Ser. Mat.*, 33:707-734, 1969.

[9] M. L. Gromov. Convex integration of differential relations. I. *Izv. Akad. Nauk SSSR Ser. Mat.*, 37:329-343, 1973.

[10] Mikhael Gromov. *Partial differential relations*, volume 9 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1986.

[11] M. Gromov. Oka's principle for holomorphic sections of elliptic bundles. *J. Amer. Math. Soc.*, 2(4):851-897, 1989.

[12] Morris W. Hirsch. Immersions of manifolds. *Trans. Amer. Math. Soc.*, 93:242-276, 1959.

[13] Nicolaas H. Kuiper. On $C^1$-isometric imbeddings. I, II. *Nederl. Akad. Wetensch. Proc. Ser. A.* **58** = *Indag. Math.*, 17:545-556, 683-689, 1955.

[14] Jürgen Moser. On the volume elements on a manifold. *Trans. Amer. Math. Soc.*, 120:286-294, 1965.

[15] E. Murphy. Loose legendrian embeddings in high dimensional contact manifolds. *Preprint arXiv:1201.2245*, 2012.

[16] John Nash. $C^1$ isometric imbeddings. textitAnn. of Math. (2), 60:383-396, 1954.

[17] John Nash. The imbedding problem for Riemannian manifolds. *Ann. of Math. (2)*, 63:20-63, 1956.

[18] Anthony Phillips. Submersions of open manifolds. *Topology*, 6:171-206, 1967.

[19] Anthony Phillips. Smooth maps transverse to a foliation. *Bull. Amer. Math. Soc.*, 76:792-797, 1970.

[20] Stephen Smale. A classification of immersions of the two-sphere. *Trans. Amer. Math. Soc.*, 90:281-290, 1958.

[21] Stephen Smale. The classification of immersions of spheres in Euclidean spaces. *Ann. of Math. (2)*, 69:327-344, 1959.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 124–128

# Extendable Continuous Self Maps and Self Homeomorphisms on Subsets of $\omega^2$

**P. Chiranjeevi***

*School of Mathematics & Statistics; University of Hyderabad, Hyderabad*

**Abstract:** In this paper we characterize the following:

(1) The pairs $(Y, f)$ where $Y$ is a subset of $\omega^2$ such that $\overline{Y}$ is a clopen subset of $\omega^2$ and $f$ is a self homeomorphism on $Y$ which can be extended as a self homeomorphism on $\omega^2$ with no periodic points in $\omega^2 \setminus Y$
(2) The pairs $(Y, f)$ where $Y$ is a subset of $\omega^2$ and $f$ is a continuous self map on $Y$ which can be extended as a continuous self map on $\omega^2$ with no periodic points in $\omega^2 \setminus Y$.
(3) Sets of eventually periodic points of continuous self maps on $\omega^2$.

**Keywords:** Periodic point, Eventually periodic point, Ordinal Number

**AMS Subject Classifications**: 37C25

## 1. Introduction

Sets of periodic points have been characterized for self homeomorphisms and continuous self maps on $\omega^2$ in [2]. In the light of these results we consider the problem of characterizing the sets of eventually periodic points of continuous self maps on $\omega^2$. Together with these sets we also characterize the pairs $(Y, f)$ where $Y$ is a subset of $\omega^2$ and $f$ is either a continuous self map or a self homeomorphism on $Y$ which can be extended to $\omega^2$ with no periodic points in $\omega^2 \setminus Y$. In the case of homeomorphisms we assume that $\overline{Y}$ is clopen.

### *Definitions and Notations*

Dynamical system is a pair $(X, f)$ where $X$ is a topological space and $f$ is a continuous self map on $X$. A point $x \in X$ is called a periodic point if $f^n(x) = x$ for some $n \in \mathbb{N}$ and the least such $n$ is called the period of $x$. The set of all periodic points of $f$ is denoted by $P(f)$ and the sets of periods of periodic points of $f$ is denoted by $Per(f)$. A point $x \in X$ is called fixed point if $f(x) = x$. The set of all fixed points is denoted by $Fix(f)$. A point $x \in X$ is called eventually periodic if $f^n(x)$ is periodic for some $n \in \mathbb{N}$. The set of all eventually periodic points of $f$ is denoted

*Corresponding author. Email: chiru.hcu@gmail.com,

by $\overleftarrow{P(f)}$. A subset $S$ of $X$ is called forward $f$-invariant if $f(S) \subset S$. A subset of $X$ that is both closed and open is called clopen.

Hereafter $X$ denotes the space $\omega^2$ that is homeomorphic to $\overline{\{m + \frac{1}{n} : m, n \in \mathbb{N}\}}$. $X_0$ and $X_1$ denote the set of isolated points and the set of limit points of $X$ respectively. Given a subset $S$ of $X$, we define the sets $C_{i,S}$ for $i \in \{1, 2, ..., 8\}$ as follows:

(1) $C_{1,S} = S \cap X_0$
(2) $C_{2,S} = S^c \cap X_0$
(3) $C_{3,S} = (\overline{S \cap X_0} \setminus \overline{S^c \cap X_0}) \cap S \cap X_1$
(4) $C_{4,S} = \overline{S \cap X_0} \cap \overline{S^c \cap X_0} \cap S \cap X_1$
(5) $C_{5,S} = (\overline{S^c \cap X_0} \setminus \overline{S \cap X_0}) \cap S \cap X_1$
(6) $C_{6,S} = (\overline{S \cap X_0} \setminus \overline{S^c \cap X_0}) \cap S^c \cap X_1$
(7) $C_{7,S} = \overline{S \cap X_0} \cap \overline{S^c \cap X_0} \cap S^c \cap X_1$
(8) $C_{8,S} = (\overline{S^c \cap X_0} \setminus \overline{S \cap X_0}) \cap S^c \cap X_1$.

## 2. Main Results

PROPOSITION 2.1 *([1],p.35) Every continuous self map $f$ on a closed subset $S$ of $X$ can be extended as a continuous self map $g$ on $X$ such that $P(g) = P(f)$.*

THEOREM 2.2 *[2] Given a subset $S$ of $X$, there exists a continuous self map $f$ on $X$ such that $P(f) = S$ if and only if $\overline{S} \setminus S$ is either empty or infinite.*

THEOREM 2.3 *[2] A subset $S$ of $X$ occurs as the set of periodic points for some self homeomorphism on $X$ if and only if the sets $C_{2,S}$, $C_{6,S}$, $C_{7,S}$, $C_{8,S}$ are either empty or infinite.*

THEOREM 2.4 *If $S$ is a clopen subset of $X$ such that $C_{2,S}$ and $C_{8,S}$ are either empty or infinite then every self homeomorphism $f$ on $S$ can be extended as a self homeomorphism $g$ on $X$ such that $P(g) = P(f)$.*

*Proof.* Since $S$ is clopen, the sets $C_{4,S}, C_{5,S}, C_{6,S}$ and $C_{7,S}$ are empty. Therefore $S^c$ is either an infinite subset of $X_0$ or homeomorphic to $X$ and so there exists a self homeomorphism $f_1$ on $S^c$ such that $P(f) = \phi$. Now the map $g : X \longrightarrow X$ defined by

$$g(x) = \begin{cases} f(x) \text{ if } x \in S \\ f_1(x) \text{ if } x \in S^c \end{cases}$$

is a self homeomorphism on $X$ such that $P(g) = P(f)$. ■

THEOREM 2.5 *Let $S \subset Y \subset X$ be such that $\overline{Y}$ is clopen in $X$. Let $f$ be a self homeomorphism on $Y$ such that $P(f) = S$. Then $f$ can be extended as a self homeomorphism $g$ on $X$ such that $P(g) = S$ if and only if the following conditions hold true:*

*(1) $C_{2,\overline{Y}}$ and $C_{8,\overline{Y}}$ are either empty or infinite.*
*(2) For each $x \in \overline{Y} \setminus Y$ and for each sequence $(x_n)$ in $Y$ converging to $x$, the sequence $(f(x_n))$ is convergent and the sequence $(f^k(x_n))$ does not converge to $x$ for any $k \in \mathbb{N}$.*

*Proof.* The necessity part is known and we will now prove the sufficiency part. Suppose that (1) and (2) hold true. For each $x \in \overline{Y}$, let $g_1(x)$ be the limit of the sequence $(f(x_n))$ where $(x_n)$ is a sequence in $Y$ converging to $x$. Then $g_1$ defines a

self homeomorphism on $\overline{Y}$ which is an extension of $f$ such that $P(g_1) = S$. Now by the previous theorem there exists a self homeomorphism $g$ on $X$ which is an extension of $f$ such that $P(g) = S$. ■

*Remark* 2.6 In Theorem 2.5, we cannot omit the assumption that $\overline{Y}$ is clopen.

**Example.** Let $X = \overline{\{m + \frac{1}{n} : m, n \in \mathbb{N}\}}$ and $Y = X \setminus \{2m + \frac{1}{2n+1} : m, n \in \mathbb{N}\}$. Let $f$ be a self homeomorphism on $Y$ with $P(f) = Y, Per(f) = \{2\}$ and $f(m) = m + 1$ whenever $m$ is odd. Then $f$ cannot be extended as a self homeomorphism $g$ on $X$ such that $P(g) = Y$.

PROPOSITION 2.7 *For every continuous self map $f$ on $X$, there exists a continuous self map $g$ on $X$ such that $P(g) = \overleftarrow{P(f)}$.*

*Proof.* Observe that $\overleftarrow{P(f)}$ and so $\overleftarrow{\overleftarrow{P(f)}}$ is forward $f$-invariant. Also $x$ is eventually periodic with respect to $f$ if and only if $f(x)$ is so. Therefore $\overleftarrow{\overleftarrow{P(f)}} \setminus \overleftarrow{P(f)}$ is forward $f$-invariant. Since every nonempty forward- invariant finite set must have a periodic point, the set $\overleftarrow{\overleftarrow{P(f)}} \setminus \overleftarrow{P(f)}$ has to be either empty or infinite. So by Theorem 2.2, there exists a continuous self map $g$ on $X$ such that $P(g) = \overleftarrow{P(f)}$. ■

*Remark* 2.8 If $f$ is a continuous self map on $X$, then the existence of a continuous self map $g$ on $X$ such that $P(f) = \overleftarrow{P(g)}$ is not guaranteed.

**Example.** If $x$ is an isolated point of $X$, then $X \setminus \{x\}$ occurs as $P(f)$ for some continuous self map $f$ on $X$ but it cannot occur as $\overleftarrow{P(g)}$ for any continuous self map $g$ on $X$.

PROPOSITION 2.9 *([3]) A metric space $Y$ is countable and compact if and only if $Y$ has a base consisting of clopen sets and for every continuous self map $f$ on $Y$, $P(f)$ is nonempty.*

*Corollary* A subspace $Y$ of $X$ is compact if and only if for every continuous self map $f$ on $Y$, $P(f)$ is nonempty.

The characterization for the sets of eventually periodic points of continuous self maps on $X$ is the following:

PROPOSITION 2.10 *Given a subset $S$ of $X$, there exists a continuous self map* f *on $X$ such that $\overleftarrow{P(f)} = S$ if and only if the following conditions hold true:*

*(1) $S^c$ is either empty or noncompact.*
*(2) $\overline{S} \setminus S$ is either empty or infinite.*

*Proof.* The necessity part is trivial and we will now prove the sufficiency part. Suppose that (1) and (2) hold true. Let $X = \overline{\{m + \frac{1}{n} : m, n \in \mathbb{N}\}}$. If both $S$ and $S^c$ are closed, then let $f_1$ be the identity map on $S$ and let $f_2$ be any continuous self map on $S^c$ such that $P(f_2) = \phi$. If $S$ is closed but $S^c$ is not closed, let $f_1$ be the identity map on $S$ and define a self homeomorphism $f_2$ on $\overline{S^c}$ such that $P(f_2) = Fix(f_2) = \mathbb{N} \bigcap S$.

If $S$ is not closed but $S^c$ is closed, let $f_2$ be any continuous self map on $S^c$ such that $P(f_2) = \phi$ and define $f_1$ on $S$ as follows: Define a self homeomorphism $f_{11}$ on $\overline{\{m + \frac{1}{n} \in S : m, n \in \mathbb{N}, n \neq 1, m \in S^c\}}$ such that $P(f_{11}) = \{m + \frac{1}{n} \in S : m, n \in \mathbb{N}, n \neq 1, m \in S^c\}$ and $f_{11}(m) = f_2(m)$ if $m \in \mathbb{N} \bigcap S^c$ and $m + \frac{1}{n} \in S$ for some

$n \in \mathbb{N} \setminus \{1\}$. $f_1 : S \longrightarrow S$ by
$f_1(x) = \begin{cases} f_{11}(x) \text{ if } x \in \{m + \frac{1}{n} \in S : m, n \in \mathbb{N}, m \in S^c\} \\ x \quad \text{if } x \in S \setminus \{m + \frac{1}{n} \in S : m, n \in \mathbb{N}, m \in S^c\}. \end{cases}$
Suppose that neither $S$ nor $S^c$ is closed. Then $\overline{S}$ will be homeomorphic to $\omega^2$ and so there exists a self homeomorphism $f_1$ on $\overline{S}$ such that $P(f_1) = S$. Let $f_2 : \overline{S^c} \longrightarrow \overline{S} \setminus S$ be any continuous map such that $f_1(x) = f_2(x)\ \forall x \in \overline{S} \bigcap \overline{S^c}$.

Then the map $f : X \longrightarrow X$ defined by
$f(x) = \begin{cases} f_1(x) \text{ if } x \in S \\ f_2(x) \text{ if } x \in S^c \end{cases}$
is a continuous self map on $X$ such that $\overleftarrow{P(f)} = S$.

■

THEOREM 2.11 *If $S \subset Y \subset X$, then a continuous self map $f$ on $Y$ such that $P(f) = S$ can be extended as a continuous self map $g$ on $X$ such that $P(g) = S$ if and only if for each $x \in \overline{Y} \setminus Y$ and for each sequence $(x_n)$ in $Y$ converging to $x$, the sequence $(f(x_n))$ is convergent and the sequence $(f^k(x_n))$ does not converge to $x$ for any $k \in \mathbb{N}$.*

*Proof.* The necessity part is trivial and we will now prove the sufficiency part. Suppose that for each sequence $(x_n)$ in $Y$ converging to $x$, the sequence $(f(x_n))$ is convergent and the sequence $f^k(x_n)$ does not converge to $x$ for any $k \in \mathbb{N}$. For each $x \in \overline{Y}$, let $g_1(x)$ be the limit of the sequence $(f(x_n))$ where $(x_n)$ is a sequence in $Y$ converging to $x$. Then $g_1$ defines a continuous self map on $\overline{Y}$ which is an extension of $f$ such that $P(g_1) = S$. Now by Proposition 2.1, $g_1$ can be extended as a continuous self map $g$ on $X$ such that $P(g) = S$. ■

*Example* 1 The function $f : \{m + \frac{1}{n} : m, n \in \mathbb{N}, n \neq 1\} \longrightarrow \{m + \frac{1}{n} : m, n \in \mathbb{N}, n \neq 1\}$ defined by
$f(m + \frac{1}{n}) = \begin{cases} m + \frac{1}{n} \quad \text{if } m, n \in \mathbb{N} \text{ and if } n \text{ is odd} \\ m + 1 + \frac{1}{n} \text{ if } m, n \in \mathbb{N} \text{ and if } n \text{ is even} \end{cases}$
is a continuous self map on $\{m + \frac{1}{n} : m, n \in \mathbb{N}, n \neq 1\}$ which cannot be extended as a continuous self map on $\overline{\{m + \frac{1}{n} : m, n \in \mathbb{N}, n \neq 1\}}$.

*Example* 2 The function $f : \{m + \frac{1}{n} : m, n \in \mathbb{N}, n \neq 1\} \longrightarrow \{m + \frac{1}{n} : m, n \in \mathbb{N}, n \neq 1\}$ defined by
$f(m + \frac{1}{n}) = \begin{cases} m + 1 + \frac{1}{n} \text{ if } m, n \in \mathbb{N} \text{ and if } m \text{ is odd} \\ m - 1 + \frac{1}{n} \text{ if } m, n \in \mathbb{N} \text{ and if } m \text{ is even} \end{cases}$
is a continuous self map on $\{m + \frac{1}{n} : m, n \in \mathbb{N}, n \neq 1\}$ which can be extended as a continuous self map $g$ on $\overline{\{m + \frac{1}{n} : m, n \in \mathbb{N}\}}$ but for any such $g$, $P(g) = \overline{\{m + \frac{1}{n} : m, n \in \mathbb{N}\}}$.

**Acknowledgements:** The author wishes to thank Prof. V. Kannan for his valuable comments during the preparation of the manuscript. The author also wishes to acknowledge with gratitude the financial support received from UGC-SAP (DSA - I)

## References


[1] Alexander S.Kechris, "Classical Descriptive Set Theory", Graduate Texts in Mathematics, Springer Verlag, 1994.

[2] M. Archana, P. Chiranjeevi, V. Kannan, Dynamics on locally compact Hausdorff spaces, Topology Proceedings, Volume 48 (2016) 309-325.

[3] V. Kannan, P.V.S.P. Saradhi and S.P. Seshasai, Periodic point property, Proc. of A.P. Academy of Sciences vol **8**, No. 3 (2004) 305-312.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 129–133

# *Gap Formula for Symmetric Operators*

**S. H. Kulkarni**[a] and **G. Ramesh**[b*]

[a]*Department of Mathematics*
*I. I. T. Palakkad*
*Ahalia Integrated Campus*
*Kozhippara, Palakkad, Kerala, 678 557.*;
[b]*Department of Mathematics*
*I. I. T. Hyderabad*
*Kandi, Sangareddy*
*Telangana, India-502 285.*

**Abstract:** In this note we give a formula for the gap $\tilde{\delta}(T, nI)$ between a symmetric, closed densely defined operator $T$ and $nI$. We illustrate our result with an example.

**Keywords:** symmetric operator, gap of operators, closed operator.

**AMS Subject Classifications**: Primary 47A55

## 1. Preliminaries

In this note we derive a formula for the gap between a symmetric densely defined closed operator $T$ and the operator $nI$, where $n \in \mathbb{N}$ and $I$ is the identity operator on a Hilbert space.

Throughout a complex Hilbert space will be denoted by $H$. The inner product and the induced norm are denoted by $\langle\cdot\rangle$ and $||.||$, respectively.

Let $T$ be a linear operator with domain $D(T)$, a subspace of $H$ and taking values in $H$. If $D(T)$ is dense in $H$, then $T$ is called a *densely defined operator*.

The graph $G(T)$ of $T$ is defined by $G(T) := \{(Tx, x) : x \in D(T)\} \subseteq H \times H$. If $G(T)$ is closed, then $T$ is called a *closed operator*. Equivalently, $T$ is closed, if $(x_n)$ is a sequence in $D(T)$ such that $x_n \to x \in H$ and $Tx_n \to y \in H$, then $x \in D(T)$ and $Tx = y$.

If $T$ is a densely defined operator, then there exists a unique linear operator (in fact, a closed operator) $T^* : D(T^*) \to H$, with

$$D(T^*) := \{y \in H : x \to \langle Tx, y\rangle \text{ for all } x \in D(T) \text{ is continuous}\} \subseteq H$$

satisfying $\langle Tx, y\rangle = \langle x, T^*y\rangle$ for all $x \in D(T)$ and $y \in D(T^*)$.

We say $T$ is bounded if there exists $k > 0$ such that $\|Tx\| \leq k\|x\|$ for all $x \in D(T)$. Note that if $T$ is densely defined and bounded then $T$ can be extended

* Corresponding author. Email: rameshg@math.iith.ac.in

to all of $H$ in a unique way.

By the Closed Graph Theorem [9], an everywhere defined closed operator is bounded. Hence the domain of an unbounded closed operator is a proper subspace of a Hilbert space.

The space of all bounded linear operators in $H$ is denoted by $\mathcal{B}(H)$ and the class of all densely defined, closed linear operators in $H$ is denoted by $\mathcal{C}(H)$. If $T \in \mathcal{C}(H)$, then $N(T)$ and $R(T)$ denotes the null space and the range space of $T$, respectively. If $M$ is a closed subspace of $H$, then $M^{\perp}$ is the orthogonal complement of $M$ in $H$ and an orthogonal projection onto $M$ of $H$ will be denoted by $P_M$.

Let $S, T \in \mathcal{C}(H)$ be operators with domains $D(S)$ and $D(T)$, respectively. Then $S+T$ is an operator with domain $D(S+T) = D(S) \cap D(T)$ defined by $(S+T)(x) = Sx+Tx$ for all $x \in D(S+T)$. The operator $ST$ has the domain $D(ST) = \{x \in D(T) : Tx \in D(S)\}$ and is defined as $(ST)(x) = S(Tx)$ for all $x \in D(ST)$.

If $S$ and $T$ are closed operators with the property that $D(T) \subseteq D(S)$ and $Tx = Sx$ for all $x \in D(T)$, then $T$ is called the *restriction* of $S$ and $S$ is called an *extension* of $T$. We denote this by $T \subseteq S$.

An operator $T \in \mathcal{C}(H)$ is *self-adjoint* if $T = T^*$, *symmetric* if $T \subseteq T^*$, *positive* if $T = T^*$ and $\langle Tx, x\rangle \geq 0$ for all $x \in D(T)$.

Let $V \in \mathcal{B}(H)$. Then $V$ is called an *isometry* if $\|Vx\| = \|x\|$ for all $x \in H$ and a *partial isometry* if $V|_{N(V)^{\perp}}$ is an isometry. The space $N(V)^{\perp}$ is called the *initial space* or the *initial domain* and the space $R(V)$ is called the *final space* or the *final domain* of $V$.

THEOREM 1.1 *[9, theorem 13.31, page 349][2, Theorem 4, page 144] Let $T \in \mathcal{C}(H)$ be positive. Then there exists a unique positive operator $S$ such that $T = S^2$. The operator $S$ is called the square root of $T$ and is denoted by $S = T^{\frac{1}{2}}$.*

THEOREM 1.2 *[2, Theorem 2, page 184] Let $T \in \mathcal{C}(H)$. Then there exists a unique partial isometry $V : H \to H$ with the initial space $\overline{R(T^*)}$ and the final space $\overline{R(T)}$ such that $T = V|T|$.*

*Definition* 1 [9, page 346] Let $T \in \mathcal{C}(H)$. The resolvent of $T$ is defined by

$$\rho(T) := \{\lambda \in \mathbb{C} : T - \lambda I : D(T) \to H \text{ is invertible and } (T-\lambda I)^{-1} \in \mathcal{B}(H)\}$$

and

$$\begin{aligned} \sigma(T) &:= \mathbb{C} \setminus \rho(T) \\ \sigma_p(T) &:= \{\lambda \in \mathbb{C} : T - \lambda I : D(T) \to H \text{ is not one-to-one}\}, \end{aligned}$$

are called the spectrum and the point spectrum of $T$, respectively.

LEMMA 1.3 *[3, 7, 10] Let $T \in \mathcal{C}(H)$. Denote $\check{T} = (I + T^*T)^{-1}$ and $\widehat{T} = (I + TT^*)^{-1}$. Then*

*(1) $\check{T} \in \mathcal{B}(H)$, $\widehat{T} \in \mathcal{B}(H)$*

*(2) $\widehat{T}T \subseteq T\check{T}$, $||T\check{T}|| \leq \frac{1}{2}$ and $\check{T}T^* \subseteq T^*\widehat{T}$, $||T^*\widehat{T}|| \leq \frac{1}{2}$.*

One of the most useful and well studied metric on $C(H)$ is the gap metric. Here we give some details.

*Definition* 2 (Gap between subspaces) [6, page 197] Let $H$ be a Hilbert space and $M, N$ be closed subspaces of $H$. Let $P = P_M$ and $Q = P_N$. Then the gap between

$M$ and $N$ is defined by

$$\theta(M,N) = \|P - Q\|.$$

If $S, T \in \mathcal{C}(H)$, then $G(T), G(S) \subseteq H \times H$ are closed subspaces. The gap between $G(T)$ and $G(S)$ is called the gap between $T$ and $S$ and it is denoted by $\widehat{\delta}(T,S)$. It is to be noted that $\widehat{\delta}(T,S)$ is a metric on $\mathcal{C}(H)$. For a proof of this fact and other properties of $\widehat{\delta}(\cdot,\cdot)$, we refer to [6, Chapter IV] and [1, page 70].

On $\mathcal{B}(H)$, the norm topology and the topology induced by the gap metric are the same (see [8, Theorem 2.5] for details ).

## 2. Main result

In this section we prove our main result. First, we state a formula for computing the gap between two densely defined closed operators, which is useful in proving our result.

THEOREM 2.1 *[5, Theorem 3.5] Let $A, B \in \mathcal{C}(H)$ be densely defined. Then*

$$B\check{B}^{\frac{1}{2}}\check{A}^{\frac{1}{2}}, \widehat{B}^{\frac{1}{2}}A\check{A}^{\frac{1}{2}},\ A\check{A}^{\frac{1}{2}}\check{B}^{\frac{1}{2}},\ \widehat{A}^{\frac{1}{2}}B\check{B}^{\frac{1}{2}}$$

*are bounded and*

$$\widehat{\delta}(A,B) = \max\left\{\|B\check{B}^{\frac{1}{2}}\check{A}^{\frac{1}{2}} - \widehat{B}^{\frac{1}{2}}A\check{A}^{\frac{1}{2}}\|,\ \|A\check{A}^{\frac{1}{2}}\check{B}^{\frac{1}{2}} - \widehat{A}^{\frac{1}{2}}B\check{B}^{\frac{1}{2}}\|\right\}.$$

THEOREM 2.2 *Let $T \in \mathcal{C}(H)$ be densely defined and symmetric. Let $n \in \mathbb{N}$ be fixed. Then*

$$\widehat{\delta}(T, nI) = \frac{1}{\sqrt{1+n^2}} \max\left\{\|(T - nI)\check{T}^{\frac{1}{2}}\|,\ \|(T^* - nI)\widehat{T}^{\frac{1}{2}}\|\right\}.$$

*In particular, if $T = T^*$, then*

$$\widehat{\delta}(T, nI) = \frac{\|(T - nI)(I + T^2)^{\frac{-1}{2}}\|}{\sqrt{1+n^2}}.$$

*Further more, if $-\dfrac{1}{n} \in \sigma(T)$, then $\widehat{\delta}(T, nI) = 1$.*

*Proof.* We use the formula in Theorem 2.1. Let $S = nI$. Then $\check{S}^{\frac{1}{2}} = \dfrac{I}{\sqrt{1+n^2}} = \hat{S}^{\frac{1}{2}}$ and $S\check{S}^{\frac{1}{2}} = \dfrac{nI}{\sqrt{1+n^2}}$. Now

$$S\check{S}^{\frac{1}{2}}\check{T}^{\frac{1}{2}} - \widehat{S}^{\frac{1}{2}}T\check{T}^{\frac{1}{2}} = \frac{1}{\sqrt{1+n^2}}(n\check{T}^{\frac{1}{2}} - T\check{T}^{\frac{1}{2}}) = \frac{1}{\sqrt{1+n^2}}(nI - T)\check{T}^{\frac{1}{2}}.$$

And

$$T\check{T}^{\frac{1}{2}}\check{S}^{\frac{1}{2}} - \widehat{T}^{\frac{1}{2}}S\check{S}^{\frac{1}{2}} = \frac{1}{\sqrt{1+n^2}}(T\check{T}^{\frac{1}{2}} - n\widehat{T}^{\frac{1}{2}}).$$

Let $A := n\check{T}^{\frac{1}{2}} - T\check{T}^{\frac{1}{2}}$ and $B := T\check{T}^{\frac{1}{2}} - n\widehat{T}^{\frac{1}{2}}$. Then

$$\widehat{\delta}(T, nI) = \frac{1}{\sqrt{1+n^2}} \max\{\|A\|, \|B\|\}.$$

Note that $B^* = T^*\widehat{T}^{\frac{1}{2}} - n\widehat{T}^{\frac{1}{2}} = (T^* - nI)\widehat{T}^{\frac{1}{2}}$. Since $\|B^*\| = \|B\|$, we get that

$$\widehat{\delta}(T, nI) = \frac{1}{\sqrt{1+n^2}} \max\left\{\|(T-nI)\check{T}^{\frac{1}{2}}\|,\ \|(T^*-nI)\widehat{T}^{\frac{1}{2}}\|\right\}.$$

If $T = T^*$, then $A = B$ and hence the formula follows in this case. As $A^* = A$ and $A$ is bounded, we have

$$\|A\| = \sup\{|\lambda| : \lambda \in \sigma(A)\} = \sup\left\{\frac{|n-\lambda|}{\sqrt{1+\lambda^2}} : \lambda \in \sigma(T)\right\}.$$

Hence consider the function

$$f(x) = \frac{|x-n|}{\sqrt{1+x^2}},\ x \in \sigma(T) \subseteq \mathbb{R}.$$

If $x_0 = \dfrac{-1}{n} \in \sigma(T)$, then we have $f(x_0) = \sqrt{1+n^2}$ and hence $\|A\| \geq \sqrt{1+n^2}$. Hence $\widehat{\delta}(T, nI) = 1$. ■

The following example illustrates the formula.

*Example* 2.3 Let $H = \ell^2$ and $\mathcal{D} = \{(x_m) \in H : (mx_m) \in H\}$. Define $T : \mathcal{D} \to H$ by

$$T(x_1, x_2, x_3, \dots) = (x_1, 2x_2, 3x_3, \dots) \text{ for all } (x_m) \in \mathcal{D}.$$

Clearly $T$ is densely defined, $T = T^*$ and range of $T$ is closed. Let $\{e_m : m \in \mathbb{N}\}$ be the standard orthonormal basis of $H$. Then $Te_m = me_m$ for each $m \in \mathbb{N}$. Hence $\mathbb{N} \subseteq \sigma_p(T)$, the point spectrum of $T$. In fact, we can show that $\sigma(T) = \mathbb{N}$.

For each $m \in \mathbb{N}$, we have

$$\begin{aligned}
T^2 e_m &= m^2 e_m \\
(I+T^2)(e_m) &= (1+m^2)e_m \\
(I+T^2)^{\frac{1}{2}} e_m &= \sqrt{1+m^2}\, e_m \\
\check{T}^{\frac{1}{2}} e_m = (I+T^2)^{\frac{-1}{2}} e_m &= \frac{1}{\sqrt{1+m^2}} e_m \\
T\check{T}^{\frac{1}{2}} e_m &= \frac{m}{\sqrt{1+m^2}} e_m.
\end{aligned}$$

Now $(T-nI)\check{T}^{\frac{1}{2}}\, e_m = \dfrac{m-n}{\sqrt{1+m^2}}\, e_m$ for each $m \in \mathbb{N}$. Hence

$$\|(T-nI)\check{T}^{\frac{1}{2}}\| = \sup\left\{\frac{|m-n|}{\sqrt{1+m^2}} : m \in \mathbb{N}\right\} = \max\left\{1, \frac{|n-1|}{\sqrt{2}}\right\}. \tag{1}$$

*Note* 2.4 For a fixed $n \in \mathbb{N}$, the sequence $a_m := \left\{ \dfrac{|m-n|}{\sqrt{1+m^2}} \right\}$ decreases for $m = 1$ to $n$ ($a_n = 0$) and then increases with $\lim\limits_{m\to\infty} a_m = 1$.

*Remark* 1 Let $T$ be a symmetric closed densely defined operator. The formula given in Theorem 2.2, is a corrected version of the erroneous formula

$$\widehat{\delta}(T, nI) = \frac{1}{\sqrt{1+n^2}}$$

given in [4].

## References


[1] N. I. Akhiezer and I. M. Glazman, *Theory of linear operators in Hilbert space. Vol. II*, Translated from the Russian by Merlynd Nestell, Frederick Ungar Publishing Co., New York, 1963.

[2] M. Š. Birman and M. Z. Solomjak, *Spectral theory of selfadjoint operators in Hilbert space* (Russian), Leningrad. Univ., Leningrad, 1980.

[3] C. W. Groetsch, Spectral methods for linear inverse problems with unbounded operators, J. Approx. Theory **70** (1992), no. 1, 16–28.

[4] S. H. Kulkarni and G. Ramesh, The carrier graph topology, Banach J. Math. Anal. **5** (2011), no. 1, 56–69.

[5] S. H. Kulkarni and G. Ramesh, A formula for gap between two closed operators, Linear Algebra Appl. **432** (2010), no. 11, 3012–3017.

[6] T. Kato, *Perturbation theory for linear operators*, reprint of the 1980 edition, Classics in Mathematics, Springer-Verlag, Berlin, 1995.

[7] G. K. Pedersen, *Analysis now*, Graduate Texts in Mathematics, 118, Springer-Verlag, New York, 1989.

[8] R. Nakamoto, Gap formulas of operators and their applications, Math. Japon. **42** (1995), no. 2, 219–232. MR1356379

[9] W. Rudin, *Functional analysis*, second edition, International Series in Pure and Applied Mathematics, McGraw-Hill, Inc., New York, 1991.

[10] S. Gramsch and E. Schock, Ill-posed equations with transformed argument, Abstr. Appl. Anal. **2003**, no. 13, 785–791.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 134–146

# *Fixed Point Theorems and Applications to Fluid Flow Problems*

**G P Raja Sekhar**[a*] and **Meraj Alam**[b]

[a] *Professor, Department of Mathematics,*
*Indian Institute of Technology Kharagpur, Kharagpur 721 302*
[b] *Research Scholar, Department of Mathematics,*
*Indian Institute of Technology Kharagpur, Kharagpur 721 302*

**Abstract:** In this article, we introduce some basic variants of fixed point theorems namely Banach Contraction Theorem, Brouwer and Schauder fixed point theorems. We show the hierarchy structure of these theorems with respect to the hypothesis stated in each. Following this introduction, we present a couple of fluid flow models in terms of boundary value problems (BVPs) involving partial differential equations. We consider a model on fluid flow inside porous media governed by Brinkman-Forchheimer equation that involves a non-linear term and the other on convection - diffusion of mass transport. We convert these BVPs into equivalent fixed-point problems and establish the existence and uniqueness results via one of the fixed point theorems introduced.

**Keywords:** Porous media, Weak formulation, Lax-Milgram Lemma

## 1. Introduction

Fixed point theorems are of great importance not only in the field of mathematics but also in engineering, science, economics and game theory etc [1]. Many linear and nonlinear problems in engineering and science can be converted to fixed point problems. Since the aim of this article is to discuss applications of various fixed-point theorems in the context of fluid flow problems, we avoid some of the elementary forms of fixed-point theorems. Accordingly, for example, consider the initial value problem

$$x'(t) = f(t, x(t)), \quad x(t_0) = y_0. \tag{1}$$

For continuous $f$, (1) is equivalent to the integral equation

$$x(t) = y_0 + \int_{t_0}^{t} f(s, x(s))\, ds. \tag{2}$$

*Corresponding author. Email: rajas@iitkgp.ac.in

The Picard-Lindelöf Theorem [2] guarantees the existence of a unique solution to (1) when $f$ is Lipschitz continuous. If $f$ is merely continuous, then the Peano Theorem [2] guarantees the existence of a solution to (1), but nothing can be said about the uniqueness of the solution. We can write (2) as a fixed point problem, i.e., to find $x \in M \subset X$ such that

$$x = Tx, \tag{3}$$

for a mapping $T$ on a suitable function space $X$. The Banach fixed-point theorem and Schauder fixed-point theorem guarantee the existence of fixed points. Also we have *Banach fixed-point theorem* $\Rightarrow$ *Picard-Lindelof theorem*, *Schauder fixed-point theorem* $\Rightarrow$ *Peano theorem.*

In a similar manner, one may consider other problems in science and engineering in terms of a fixed point problem setting. Please refer to [1] [3] [2] for some important literature on fixed theorems and their applications. Here in this note, we focus on the application of fixed point theorems to fluid flow problems mainly Brinkman-Forchheimer and convection diffusion equation. Before going to our main problem, we would like to review the Navier-Stokes equation and related literature. It may be noted that the existence and uniqueness corresponding to Navier-Stokes equation (NSE) needed some attention. The NSE is the most studied problem in fluid mechanics. Starting from the pioneering work of Leray [4], the mathematical theory of NSE has been studied extensively. The question of existence, uniqueness and regularity of the solution to NSE has been well established when restricted to regularity and smallness of the data for bounded and unbounded domains. The literature on NSE is too vast and we recommend the reader to refer [5–10] and references therein. We note that almost all existing methods to resolve the Navier-Stokes problem are based on the Banach contraction principle. Also there is a recent literature where smallness condition on the data has been relaxed [11].

It may be noted that the dynamics involved in food processing, chemical engineering, flow inside human arteries within the context of fluid flow and nutrient transport require understanding the solution behavior of the corresponding transport equation. The fluid flow through porous media invites a lot of debate to the scientific community due to the fact that there are no unified models that can be adapted to govern the fluid flow inside a porous medium. As a result, effective medium models such as Brinkman equation are popular alternatives to Darcy equation [12]. Though these models are used extensively without much mathematical base, an attempt is made to derive these via rigorous mathematical techniques such as homogenization [13]. The Darcy and Brinkman equation (which are linear) hold for the sufficiently small velocity. In other words when the Reynolds number of flow is of order smaller than one. However, as the velocity increases the form drag due to solid obstacles is comparable with surface drag due to friction. Hence, there is a breakdown in the linearity of Darcy equation. According to Joseph et. al [14] the appropriate modification to Darcy equation is the Forchheimer equation. Here in this article we consider the Brinkman-Forchheimer equation (BFE). BFE is non-linear in nature somewhat mathematically similar to NSE. Thus, the existing literature on NSE help us to develop the corresponding existence and uniqueness results. Regarding the existing literature on the BFE one can follow Nield and Bejan [12] for physical development. Payne and Straughan [15] considered BFE for non-slow flow in a saturated porous medium and shown that the solution depends continuously on changes in the Forchheimer coefficient. They have shown convergence of the solution of the BFE to that of the Brinkman equation in the limit when the Forchheimer coefficient tends to zero. Liu et. al. [16] considered BFE and

shown the continuous dependence of the Forchheimer coefficient and the Brinkman coefficient in a bounded domain of a viscous fluid interfacing with a porous solid. One may refer the monograph by Ames and Straughan [17] to get some insights on related literature.

Regarding existence and uniqueness, Kaloni and Guo [18] have considered a steady nonlinear Brinkman-Forchheimer equation with double-diffusive convection through a porous medium. The existence, regularity, and uniqueness results are discussed using variational formulation. Skrzypacz and Wei [19] considered a nonlinear BFE with the convective term to model some porous medium flow in chemical reactors of packed bed type. The results concerning the existence and uniqueness of a weak solution are presented for nonlinear convective flows in medium with variable porosity and for small data. In this note, we consider a steady nonlinear Brinkman-Forchheimer equation without the convective term. We use a fixed point theorem to show existence and uniqueness of solution in weak sense. We present a variant proof on existence and uniqueness using fixed point theorem which is not attempted for Brinkman-Forchheimer equation as per the existing literature. Also some comments on the convection diffusion equation are presented.

## 2. Preliminaries

There are several well-known fixed point theorems and their variants in the literature. Here, we state some of the fundamental fixed point theorems having application in the field of linear and non-linear partial differential equations. We try to indicate the hierarchy with respect to the hypothesis involved.

THEOREM 2.1 *(Banach Contraction Theorem [20, 21]) Let $(M, d)$ be a complete metric space (CMS) and $F : M \to M$ be a strict contraction i.e., there exists a number $\rho$, $0 < \rho < 1$, such that*

$$d(F(x), F(y)) \leq \rho d(x, y), \quad \forall \ \ x, \ y \in M. \tag{4}$$

*Then there exists a unique fixed point $x^* \in M$ of $F$. Moreover, for any choice of $x_0 \in M$, the recursive sequence*

$$x_{m+1} = F(x_m), \quad m \geq 0$$

*converges to $x^*$.*

It may be noted that the above theorem is the most primitive version in the family of fixed-point theorems. In the above, the constant $\rho$ has to be strictly less than one in order to be a contraction map. If one relaxes this constraint, i.e. the case of Lipschitz continuous function, the constant is such that $0 < \rho < \infty$. This indicates that contraction implies Lipschitz continuity, however, the converse is not true. Further, the above theorem does not restrict the dimension of the metric space, however, demands (4) which is in general can be considered to be expensive. We now introduce a variant where the map is restricted to compact sets of finite dimensional space. Correspondingly, one has:

THEOREM 2.2 *(Brouwer [21, 22]) Let $\mathcal{S} \subset \mathbb{R}^n$ be a closed sphere and $T : \mathcal{S} \to \mathcal{S}$ be a continuous map. Then $T$ has a fixed point $x^*$.*

Relaxing the finite-dimensional restriction, Schauder extended the above

Brouwer Fixed Point Theorem to a Banach space setting, i.e. in general to an infinite dimensional space. Accordingly, one can have:

THEOREM 2.3 *(Schauder [21, 23]) Let $X$ be a Banach space. We are given:*

- $A \subset X$, *compact and convex.*
- $T : A \to A$ *is continuous.*

*Then $T$ has a fixed point $x^* \in A$.*

The application of Schauders Theorem requires the compactness of the set $A$, which is a strong requirement in an infinite dimensional space. In the applications to boundary value problems, it is much more convenient to formulate variants in which the compactness requirement is passed to the image $T(A)$ or to the operator $T$ itself, rather than to the domain $A$. Correspondingly, the first variant of Schauder Fixed Point Theorem that demands compactness of the image of the operator is the following:

THEOREM 2.4 *Let $X$ be a Banach space. Assume that:*

- $A \subset X$, *closed and convex.*
- $T : A \to A$ *is continuous.*
- $\overline{T(A)}$ *is compact in $X$.*

*Then $T$ has a fixed point $x^* \in A$.*

A second variant that uses the compactness of the operator $T$ is also possible. We recall that $T$ is compact if the image of a bounded set has a compact closure. Correspondingly, one can have:

THEOREM 2.5 *Let $X$ be a Banach space. Assume that:*

- $A \subset X$, *closed, bounded and convex.*
- $T : A \to A$ *is a continuous and compact operator.*

*Then $T$ has a fixed point $x^* \in A$.*

A further variant of Schauder's Theorem requiring the compactness of the operator $T$ and the existence of a family of operators $T_s$, $0 \leq s \leq 1$, is due to Leray-Schauder [21]). This relies mostly on the boundedness of the solution of $x = sT(x)$ which ensures fixed point of the operator $T$. We avoid presenting the same in detail. Having listed few important variants of fixed-point theorems together with the corresponding hypothesis, we now consider a boundary value problem based on fluid flow through porous medium.

## 3. Applications to Fluid flow problems

**Brinkman-Forchheimer equation:** [12] Let $\Omega \subset \mathbb{R}^d$, $(d = 2, 3)$ be a bounded domain with $\partial\Omega$ as its boundary that is Lipschitz continuous. For a homogeneous, isotropic porous medium the corresponding momentum balance equation is given by [12]

$$-\frac{\mu}{\varphi}\nabla^2\mathbf{u} + \frac{\tilde{\mu}}{K}\mathbf{u} + \frac{c_F\rho_f}{K^{1/2}}|\mathbf{u}|\mathbf{u} = \mathbf{b} - \nabla P \ \text{ in } \ \Omega, \tag{5}$$

together with the equation of continuity (mass conservation)

$$\text{div}\mathbf{u} = 0 \ \text{ in } \ \Omega. \tag{6}$$

We consider the Dirichlet data given by $\mathbf{u} = 0$ on $\partial\Omega$. The list of symbols and their physical meaning:

- $\mu$ : Fluid viscosity
- $\tilde{\mu}$ : Effective viscosity
- $K$ : permeability
- $\rho_f$ : Fluid density
- $\varphi$ : Porosity of the porous medium
- $\mathbf{u}$ : Average fluid velocity
- $P$ : Fluid pressure
- $c_F$ : Dimensionless form-drag constant
- $\mathbf{b}$ : Body force.

The following non-dimensionalization is used:

- $L$ : Characteristic length
- $U_0$ : the magnitude of the some reference velocity
- $\mathbf{x} = L\hat{\mathbf{x}}$
- $\mathbf{u} = U_0\hat{\mathbf{u}}$
- $\nabla = \frac{\hat{\nabla}}{L}$
- $\nabla^2 = \frac{\hat{\nabla}^2}{L^2}$
- $P = \hat{P}\rho_f U_0^2$.

Consequently, the governing equations (5)-(6) together with the Dirichlet boundary condition reduce to (for convenience "hat" is dropped)

$$-\frac{1}{\varphi\text{Re}}\nabla^2\mathbf{u} + \frac{1}{\text{DaRe}}\mathbf{u} + \frac{c_F}{\sqrt{\text{Da}}}|\mathbf{u}|\mathbf{u} = \mathbf{b} - \nabla P \ \text{ in } \ \Omega \tag{7}$$

$$\text{div}\mathbf{u} = 0 \ \text{ in } \ \Omega \tag{8}$$

$$\mathbf{u} = 0 \ \text{ on } \ \partial\Omega \tag{9}$$

where $\text{Re} = \frac{LU_0}{\nu}$ : Reynolds number, $\text{Da} = \frac{K}{L^2}$ : Darcy number, are the non-dimensional constants. It may be noted that the Darcy number indicates non-dimensional permeability of the porous medium.
Without debating much on the solution methods (analytical/numerical) to deal such problems, the main aim here is to show the existence of a weak solution of (7)-(9) using fixed point theorem approach. In order to achieve this, let us define the weak formulation in some suitable functions spaces. Please refer to the Appendix 4.1 for the details on function spaces. Assume that Re, Da, $\varphi$, $c_F$ are known constants and $\mathbf{b} \in L^2(\Omega)^d$ also known.

Let us choose a pair of test functions $(\mathbf{v}, q) \in H_0^1(\Omega)^d \times Q$ and multiplying with (7)-(8) and integrating by parts, we get the following weak formulation (using boundary condition (9)).
**Weak formulation:** A weak (or variational) formulation of (7)-(9) is to find a pair $(\mathbf{u}, P)$ such that $\mathbf{u} \in H_0^1(\Omega)^d$, $P \in Q$ and

$$\int_\Omega \left( \frac{1}{\varphi\text{Re}}\nabla\mathbf{u} : \nabla\mathbf{v} + \frac{1}{\text{DaRe}}\mathbf{u}\cdot\mathbf{v} + \frac{c_F}{\sqrt{\text{Da}}}|\mathbf{u}|\mathbf{u}\cdot\mathbf{v} \right) d\Omega = \int_\Omega (P\text{div}\mathbf{v} + \mathbf{b}\cdot\mathbf{v})\, d\Omega, \tag{10}$$

$$\int_\Omega q \mathrm{div}\mathbf{u}\, d\Omega = 0, \tag{11}$$

hold for all $(\mathbf{v}, q) \in H_0^1(\Omega)^d \times Q$.

**Pressure elimination:** In order to transform the equations (10)-(11) into a fixed point problem, one has to eliminate the pressure from the weak formulation. This can be done by using a subspace $\mathbf{V}_{\mathrm{div}} = \{\mathbf{u} \in H_0^1(\Omega)^d \mid \mathrm{div}\mathbf{u} = 0\}$ of $H_0^1(\Omega)^d$ which consists of divergence free functions. Thus, the weak formulation (10)-(11) reduces to find $\mathbf{u} \in \mathbf{V}_{\mathrm{div}}$ such that

$$\int_\Omega \left( \frac{1}{\varphi \mathrm{Re}} \nabla \mathbf{u} : \nabla \mathbf{v} + \frac{1}{\mathrm{DaRe}} \mathbf{u} \cdot \mathbf{v} + \frac{c_F}{\sqrt{\mathrm{Da}}} |\mathbf{u}| \mathbf{u} \cdot \mathbf{v} \right) d\Omega = \int_\Omega \mathbf{b} \cdot \mathbf{v}\, d\Omega, \quad \forall\ \mathbf{v} \in \mathbf{V}_{\mathrm{div}}. \tag{12}$$

Once the existence of a solution of (12) is guaranteed, one can recover the corresponding pressure $P$. This procedure is shown in the next section using *DeRham* Lemma 3.2 [9].

**Reduction to a fixed point problem:** It may be noted that the reduced weak formulation (12) is non-linear. In order to use a fixed-point setting, it would be intuitive to reduce it to a corresponding linear version. This can be done by fixing the non-linear term (i.e., assuming that it is known). Accordingly, we fix $\mathbf{w} \in \mathbf{V}_{\mathrm{div}}$ and replace the non-linear term $|\mathbf{u}|\mathbf{u}$ by $|\mathbf{w}|\mathbf{u}$ in (12) so that

$$\int_\Omega \left( \frac{1}{\varphi \mathrm{Re}} \nabla \mathbf{u} : \nabla \mathbf{v} + \frac{1}{\mathrm{DaRe}} \mathbf{u} \cdot \mathbf{v} + \frac{c_F}{\sqrt{\mathrm{Da}}} |\mathbf{w}| \mathbf{u} \cdot \mathbf{v} \right) d\Omega = \int_\Omega \mathbf{b} \cdot \mathbf{v}\, d\Omega \quad \forall\ \mathbf{v} \in \mathbf{V}_{\mathrm{div}}, \tag{13}$$

which is a linear problem. The existence of a solution of the above equation can be shown using Lax-Milgram Lemma 4.1. In order to do so, we introduce the following bilinear and linear mappings as follows

$$a(\mathbf{u}, \mathbf{v}) = \int_\Omega \left( \frac{1}{\varphi \mathrm{Re}} \nabla \mathbf{u} : \nabla \mathbf{v} + \frac{1}{\mathrm{DaRe}} \mathbf{u} \cdot \mathbf{v} \right) d\Omega,$$

$$\tilde{a}(|\mathbf{w}|; \mathbf{u}, \mathbf{v}) = \frac{c_F}{\sqrt{\mathrm{Da}}} \int_\Omega |\mathbf{w}| \mathbf{u} \cdot \mathbf{v}\, d\Omega,$$

$$F(\mathbf{v}) = \int_\Omega \mathbf{b} \cdot \mathbf{v}\, d\Omega.$$

Consequently, (13) reduces to find $\mathbf{u} \in \mathbf{V}_{\mathrm{div}}$ such that

$$A_{\mathbf{w}}(\mathbf{u}, \mathbf{v}) = F(\mathbf{v}), \quad \forall\ \mathbf{v} \in \mathbf{V}_{\mathrm{div}}, \tag{14}$$

where $A : \mathbf{V}_{\mathrm{div}} \times \mathbf{V}_{\mathrm{div}} \to \mathbb{R}$ given by

$$A_{\mathbf{w}}(\mathbf{u}, \mathbf{v}) = a(\mathbf{u}, \mathbf{v}) + \tilde{a}(|\mathbf{w}|; \mathbf{u}, \mathbf{v}).$$

The above setting now allows us to use the Lax-Milgram Lemma such that problem (14) has a unique solution (say $\mathbf{u} \in \mathbf{V}_{\mathrm{div}}$). Hence, we can define a mapping

(say $T : \mathbf{V}_{\text{div}} \to \mathbf{V}_{\text{div}}$) that takes each $\mathbf{w} \in \mathbf{V}_{\text{div}}$ to solution $\mathbf{u}$ of the linear problem (14) as

$$\mathbf{u} = T(\mathbf{w}) \in \mathbf{V}_{\text{div}}. \tag{15}$$

Thus, any fixed point of the map $T : \mathbf{V}_{\text{div}} \to \mathbf{V}_{\text{div}}$ is a solution of the equation (12). In other words in order to show that the problem (12) has a solution, one must show that the mapping $T$ has a fixed point. Thus our immediate task is to show the existence of a solution of the linear problem (14) and consequently that $T$ has a fixed point.

**Solution of the linear problem (14):** We claim that:

- $A_{\mathbf{w}}(\mathbf{u}, \mathbf{v})$ is bilinear in $\mathbf{u}$, $\mathbf{v}$ and continuous
- $A_{\mathbf{w}}(\mathbf{u}, \mathbf{v})$ is coercive in $\mathbf{V}_{\text{div}}$
- The functional $F$ belongs to $(\mathbf{V}_{\text{div}})^*$ (the dual of $\mathbf{V}_{\text{div}}$)

Clearly $A_{\mathbf{w}}(\mathbf{u}, \mathbf{v})$ is bilinear and continuous. Indeed, we have the following estimate

$$|A_{\mathbf{w}}(\mathbf{u}, \mathbf{v})| \leq \left( \frac{1}{\varphi \text{Re}} + \frac{1}{\text{DaRe}} + \frac{c_F c_e^2}{\sqrt{\text{Da}}} ||\mathbf{w}||_{0,\Omega} \right) ||\mathbf{u}||_{\mathbf{V}_{\text{div}}} ||\mathbf{v}||_{\mathbf{V}_{\text{div}}}.$$

Moreover, for any $\mathbf{u} \in \mathbf{V}_{\text{div}}$, we have

$$A_{\mathbf{w}}(\mathbf{u}, \mathbf{u}) \geq \frac{\alpha}{\text{Re}} ||\mathbf{u}||^2_{\mathbf{V}_{\text{div}}},$$

i.e $A_{\mathbf{w}}(\cdot, \cdot)$ is coercive in $\mathbf{V}_{\text{div}}$, where $\alpha = \min\{\frac{1}{\varphi}, \frac{1}{\text{Da}}\}$. Further, we have

$$|F(\mathbf{v})| \leq c_p ||\mathbf{b}||_{0,\Omega} ||\mathbf{v}||_{\mathbf{V}_{\text{div}}},$$

which implies that $F \in (\mathbf{V}_{\text{div}})^*$. Here $c_p$ : the Poincare's constant and $c_e$ : embedding constant. Hence Lax-Milgram Lemma proves that the problem (14) has a unique solution $\mathbf{u} = T(\mathbf{w}) \in \mathbf{V}_{\text{div}}$ which satisfies

$$||T(\mathbf{w})||_{\mathbf{V}_{\text{div}}} \leq \frac{c_p \text{Re}}{\alpha} ||\mathbf{b}||_{0,\Omega}. \tag{16}$$

**Existence of a fixed point of the mapping $T$:** Here in this section, we prove that $T$ has a fixed point. We verify the hypothesis of Banach Contraction theorem for the mapping $T$.

LEMMA 3.1 *The mapping $T : \mathbf{V}_{div} \to \mathbf{V}_{div}$ defined by (15) is a strict contraction in the closed ball $B_M = \{\mathbf{w} \in \mathbf{V}_{div} : ||\mathbf{w}||_{\mathbf{V}_{div}} \leq M\} \subset \mathbf{V}_{div}$ $\left(\text{where } M = \frac{c_p Re}{\alpha} ||\mathbf{b}||_{0,\Omega}\right)$ whenever the data satisfies the following smallness condition*

$$\frac{2 c_F c_p^2 c_e^2 Re^2 ||\mathbf{b}||_{0,\Omega}}{\alpha^2 \sqrt{Da}} < 1, \tag{17}$$

*where $c_p$ and $c_e$ are the constants those appear in the Poincare's and embedding inequalities respectively.*

**Proof of Lemma 3.1:** The closed ball $B_M = \{\mathbf{w} \in \mathbf{V}_{\text{div}} | ||\mathbf{w}||_{\mathbf{V}_{\text{div}}} \leq M\}$ endowed with the distance $d(\mathbf{w}_1, \mathbf{w}_2) = ||\mathbf{w}_1 - \mathbf{w}_2||_{\mathbf{V}_{\text{div}}}$ is a complete metric space and estimate (16) implies that $T : B_M \to B_M$. We have to prove that $T$ is a strict contraction in $B_M$, i.e., there exists $\delta < 1$ such that

$$||T(\mathbf{w}) - T(\mathbf{z})||_{\mathbf{V}_{\text{div}}} \leq \delta ||\mathbf{w} - \mathbf{z}||_{\mathbf{V}_{\text{div}}}, \quad \forall\ \mathbf{w},\ \mathbf{z} \in B_M. \tag{18}$$

Indeed, from Eqn (14) we have for any $\mathbf{w},\ \mathbf{z} \in \mathbf{V}_{\text{div}}$,

$$a(T(\mathbf{w}) - T(\mathbf{z}), \mathbf{v}) = -\tilde{a}(|\mathbf{w}|; T(\mathbf{w}), \mathbf{v}) + \tilde{a}(|\mathbf{z}|; T(\mathbf{z}), \mathbf{v}). \tag{19}$$

Adding and subtracting $\tilde{a}(|\mathbf{w}|; T(\mathbf{z}), \mathbf{v})$, we obtain

$$a(T(\mathbf{w}) - T(\mathbf{z}), \mathbf{v}) = -\tilde{a}(|\mathbf{w}|; T(\mathbf{w}) - T(\mathbf{z}), \mathbf{v}) - \tilde{a}(|\mathbf{w}| - |\mathbf{z}|; T(\mathbf{z}), \mathbf{v}), \tag{20}$$

for every $\mathbf{v} \in \mathbf{V}_{\text{div}}$. Choosing $\mathbf{v} = T(\mathbf{w}) - T(\mathbf{z})$ and using Holder's, embedding and Poincare's inequalities, we obtain

$$\left(\frac{\alpha}{\text{Re}} - \frac{c_F c_p c_e^2 M}{\sqrt{\text{Da}}}\right) ||T(\mathbf{w}) - T(\mathbf{z})||_{\mathbf{V}_{\text{div}}} \leq \frac{c_F c_p c_e^2 M}{\sqrt{\text{Da}}} ||\mathbf{w} - \mathbf{z}||_{\mathbf{V}_{\text{div}}} \quad \forall\ \mathbf{w},\ \mathbf{z} \in B_M, \tag{21}$$

or,

$$||T(\mathbf{w}) - T(\mathbf{z})||_{\mathbf{V}_{\text{div}}} \leq \frac{c_F c_p c_e^2 M}{\sqrt{\text{Da}}} \frac{1}{\left(\frac{\alpha}{\text{Re}} - \frac{c_F c_p c_e^2 M}{\sqrt{\text{Da}}}\right)} ||\mathbf{w} - \mathbf{z}||_{\mathbf{V}_{\text{div}}} \quad \forall\ \mathbf{w},\ \mathbf{z} \in B_M. \tag{22}$$

We can observe that the assumption (17) implies that

$$\frac{c_F c_p c_e^2 M}{\sqrt{\text{Da}}} \frac{1}{\left(\frac{\alpha}{\text{Re}} - \frac{c_F c_p c_e^2 M}{\sqrt{\text{Da}}}\right)} < 1.$$

Thus, the mapping $T : B_M \to B_M$ is a strict contraction in the metric space $B_M$ when $\frac{2 c_F c_p^2 c_e^2 \text{Re}^2 ||\mathbf{b}||_{0,\Omega}}{\alpha^2 \sqrt{\text{Da}}} < 1$. Hence, the Banach Contraction Theorem implies that $T$ has a unique fixed point (say $\mathbf{u}^*$) in $B_M$, which is also a solution of problem (12).
Moreover, for any choice of $\mathbf{u}_0 \in B_M$, the recursive sequence

$$\mathbf{u}_{m+1} = T(\mathbf{u}_m), \quad m \geq 0,$$

converges to $\mathbf{u}^*$.
As mentioned earlier, one of the tasks is to recover pressure $P$. For this purpose we use the following *DeRham* Lemma.

LEMMA 3.2 *[9, 21] Let $\Omega \subset \mathbf{R}^n$ be a Lipschitz domain, homeomorphic to a ball. A functional $g \in H^{-1}(\Omega)^d$ satisfies the condition*

$$\langle g, \phi \rangle_{H^{-1}, H_0^1} = 0, \quad \forall\ \phi \in \mathbf{V}_{div},$$

*if and only if there exists* $P \in L^2(\Omega)$ *such that*

$$-\nabla P = g.$$

*The functional* $P$ *is unique up to an additive constant.*

In order to use the above lemma, integrating by parts in equation (12), we obtain the following

$$\langle g, \mathbf{v} \rangle_{H^{-1}, H_0^1} = 0 \quad \forall\ \mathbf{v} \in \mathbf{V}_{\text{div}}, \tag{23}$$

where $g = -\frac{1}{\varphi \text{Re}} \nabla^2 \mathbf{u} + \frac{1}{\text{DaRe}} \mathbf{u} + \frac{c_F}{\sqrt{\text{Da}}} |\mathbf{u}| \mathbf{u} - \mathbf{b}$. The Lemma 3.2 implies that there exists $P \in L^2(\Omega)$ such that $-\nabla P = g$. That is

$$-\frac{1}{\varphi \text{Re}} \nabla^2 \mathbf{u} + \frac{1}{\text{DaRe}} \mathbf{u} + \frac{c_F}{\sqrt{\text{Da}}} |\mathbf{u}| \mathbf{u} - \mathbf{b} = -\nabla P.$$

Hence, it proves that $(\mathbf{u}^*, P) \in H_0^1(\Omega)^d \times Q$ is a unique solution of the problem (7)-(9).
The fluid mechanical model presented above is a first step to deal with existence and uniqueness results to BFE. Authors have contributed some results on application of deformable porous porous media to biological tumors [25, 26]. An attempt will be made to incorporate BFE model wherever it is appropriate to account for inertia due to form drag.

**Convection-diffusion equation:** [24] Consider a spherical porous pellet of radius $a$ inside which the following nutrient transport is valid

$$-D\nabla^2 c = -\mathbf{v} \cdot \nabla c - \alpha c^3$$

subject to the Dirichlet condition $c = c_0$ on $r = a$. The last term on the right hand side indicates a third order reaction kinetics of strength $\alpha > 0$. The negative sign indicates consumption of the nutrient. It is obvious to see that this reaction is non-linear and attempting an analytical solution would be impossible. However, the main aim here is to show a fixed-point setting of this model via a series solution. Of course, this would cost some sort of linearization. It may be noted that for a given $c^2 = f(r, \theta)$, the above non-linear problem reduces to the following linear problem

$$-D\nabla^2 c = -\mathbf{v} \cdot \nabla c - \alpha f c$$

Let the convective velocity be driven by a flow with constant magnitude $V$ along the $z$ - direction so that $\mathbf{v} = V\hat{k}$. Consequently, the transport equation reduces to

$$-D\nabla^2 c = -V\frac{\partial c}{\partial z} - \alpha f c$$

It can be shown that the above equation admits solution of the form

$$C(r, \theta) = c_0 e^{-r\cos\theta} \sum_{n=0}^{\infty} A_n r^n F_(r) P_n(\cos\theta),$$

where $A_n$ and $F_n(r)$ can be obtained, which ensures the existence of solution for the linear problem. For details one may refer [24]. We now define the general iterative problem. One may seek a Fourier series of the form $c = \sum_{n=0}^{\infty} H_n(r) P_n(\cos\theta)$ where

$$c = \lim_{n\to\infty} c_n.$$

This defines an iterative sequence via

$$c_{n+1} = T(c_n)$$

where $T$ is defined by the following

$$-D\nabla^2 c_{n+1}(x) + v(x)\cdot\nabla c_{n+1}(x) = -\alpha c_n^2 c_{n+1}(x).$$

One may need to show that $T$ is contraction mapping on a Banach space. Then Banach Contraction theorem can be used to show that the mapping $T$ has a unique fixed point (say $c$) and the sequence $c_n$ converges to $c$ for any choice of $c_0$.
The method suggested above is based on a preliminary investigations done in [24]. Of course at this stage, the ideas presented appear very theoretical and implementing fixed point setting require rigorous computations. The aim here is to introduce such an application of fixed point theorems.

## 4. Summary

Here in this article, we have presented some basic theorems on fixed points. Further, we have presented a non-linear model namely Brinkman-Forchheimer that describes the fluid through porous media. We first convert the problem into a weak formulation and then into a fixed point setting. Further, using Lax-Milgram lemma and Banach contraction theorem, we have shown the existence and uniqueness of a solution in weak sense under the smallness assumption on data. Authors are extending this work to make it more rigorous. The first attempt is to relax the assumption on smallness of data. Further, we may couple Brinkman-Forchheimer equation with the transport equation.

### 4.1. *Appendix A*

**Function spaces and useful results:** [21, 27] Let $\Omega$ be a bounded, open subset of $\mathbb{R}^d$, d=2,3. $L^2(\Omega)$ is the space of all measurable functions $u$ defined on $\Omega$ for which

$$||u||_{0,\Omega} = \left(\int_\Omega |u|^2\, d\Omega\right)^{1/2} < +\infty, \tag{24}$$

In (24) $||\ \ ||_{0,\Omega}$ defines a norm on $L^2(\Omega)$. For any $\mathbf{u} = (u_1, u_2, \ldots, u_d) \in L^2(\Omega)^d$, $||\mathbf{u}||_{0,\Omega}$ is defined as

$$||\mathbf{u}||_{0,\Omega} = \left(\int_\Omega \sum_{i=1}^{d} |u_i|^2\, d\Omega\right)^{1/2}, \tag{25}$$

and for any element $\mathbf{K} = (K_{ij})_{1\leq i,j\leq d} \in (L^2(\Omega))^{d\times d}$, we define the norm of $\mathbf{K}$ as

$$||\mathbf{K}||_{0,\Omega} = \left( \int_\Omega \sum_{i=1}^{d} \sum_{j=1}^{d} |K_{ij}|^2 \, d\Omega \right)^{1/2}. \tag{26}$$

The symbol $(\ ,\ )_\Omega$ denotes inner product in $L^2(\Omega)$, $L^2(\Omega)^d$, and $(L^2(\Omega))^{d\times d}$. For any two functions $\mathbf{u}$ and $\mathbf{v}$ the inner product $(\ ,\ )_\Omega$ is define as

$$(\mathbf{u}, \mathbf{v})_\Omega = \int_\Omega \mathbf{u} \cdot \mathbf{v} \, d\Omega.$$

The first order Sobolev space is denoted by $H^1(\Omega)^d$ and defined as $H^1(\Omega)^d = \{\mathbf{u} \in L^2(\Omega)^d | \nabla \mathbf{u} \in (L^2(\Omega))^{d\times d}\}$. The norm of a function $\mathbf{u} \in H^1(\Omega)^d$ is defined as

$$||\mathbf{u}||_{1,\Omega} = \left( ||\mathbf{u}||^2_{0,\Omega} + ||\nabla \mathbf{u}||^2_{0,\Omega} \right)^{1/2}. \tag{27}$$

$H^1_0(\Omega)^d$ denotes the space of functions in $H^1(\Omega)^d$ with zero value at boundary. The dual space of $H^1_0(\Omega)^d$ is denoted by $H^{-1}(\Omega)^d$. The norm on the dual space is defined as

$$||\mathbf{f}||_{*,\Omega} = \sup_{0\neq \mathbf{u}\in \mathbf{X}} \frac{|\langle \mathbf{f}, \mathbf{u} \rangle_*|}{||\mathbf{u}||_{1,\Omega}}, \tag{28}$$

where $\langle \cdot\ ,\ \cdot \rangle_*$ denotes the duality pairing.
We define some useful spaces

$$\mathbf{V}_{\text{div}} = \{\mathbf{v} \in H^1_0(\Omega)^d | \nabla \cdot \mathbf{v} = 0\}$$

and

$$Q = \left\{ q \in L^2(\Omega) \Big| \int_\Omega q = 0 \right\}.$$

The norm on $\mathbf{V}_{\text{div}}$ induced by the norm of $H^1_0(\Omega)^d$.
**Inequalities:**

- Cauchy-Schwarz Inequality

$$(\mathbf{u}, \mathbf{v})_\Omega \leq ||\mathbf{u}||_{0,\Omega} ||\mathbf{v}||_{0,\Omega},\ \forall\ \mathbf{u},\ \mathbf{v} \in L^2(\Omega)^d. \tag{29}$$

- Poincare's Inequality

$$||\mathbf{V}||^2_{0,\Omega} \leq c_p ||\nabla \mathbf{V}||^2_{0,\Omega},\ \forall\ \mathbf{V} \in H^1_0(\Omega)^d. \tag{30}$$

- Sobolev Inequality

$$||\mathbf{V}||^2_{L^4(\Omega)} \leq c_e ||\mathbf{V}||^2_{1,\Omega},\ \forall\ \mathbf{V} \in H^1(\Omega)^d. \tag{31}$$

**Lax-Milgram Lemma:** [28] Assume $H$ is a real Hilbert space, with norm $||\cdot||_H$ and inner product $(\cdot, \cdot)_H$. We let $\langle \cdot, \cdot \rangle_{H^*,H}$ denotes the pairing of $H$ with its dual space $H^*$.

LEMMA 4.1 *Assume that $A$: $H\times H \to \mathbb{R}$ is a bilinear mapping, which is bounded and coercive that is, there exist constants $\alpha,\ \beta > 0$ such that*
*(i)*

$$|A[u,v]| \leq \alpha ||u||_H ||v||_H \quad (\forall\ u, v \in H)$$

*and*
*(ii)*

$$\beta ||u||_H^2 \leq A[u,u] \quad (\forall\ u \in H).$$

*Finally, let $f : H \to \mathbb{R}$ be a bounded linear functional on $H$. Then there exists a unique element $u \in H$ such that*

$$A[u,v] = \langle f, v\rangle, \quad \forall\ v \in H. \tag{32}$$

*Moreover,*

$$||u||_H \leq \frac{1}{\beta}||f||_{H^*}$$

*i.e, the solution $u$ depends continuously on the given data $f$.*

## References


[1] Andrzej Granas and James Dugundji, *Fixed point theory*, Springer Science Business Media, 2003.

[2] E Zeidler, *Nonlinear functional analysis and its applications I: Fixed-point theorems*, 1990,

[3] William A Kirk and Brailey Sims. *Handbook of metric fixed point theory*. Springer Science Business Media, 2013.

[4] Jean Leray, *Etude de diverses équations intégrales non linéaires et de quelques problèmes que pose l'hydrodynamique*, Gauthier-Villars, 1933.

[5] Amrouche, Chérif and Rodríguez-Bellido, M Ángeles, *Stationary Stokes, Oseen and Navier–Stokes equations with singular data*, Archive for Rational Mechanics and Analysis, 199(2):597651, 2011.

[6] Reinhard Farwig, Giovanni P Galdi, and Hermann Sohr, *Very weak solutions and large uniqueness classes of stationary navierstokes equations in bounded domains of* $\mathbb{R}^2$, Journal of Differential Equations, 227(2):564580, 2006.

[7] GP Galdi, CG Simader, and H Sohr, *A class of solutions to stationary stokes and navier-stokes equations with boundary data in* $W^{-1/q,q}$, Mathematische Annalen, 331(1):4174, 2005.

[8] Hyunseok Kim, *Existence and regularity of very weak solutions of the stationary navier-stokes equations*, Archive for rational mechanics and analysis, 193(1):117152, 2009.

[9] Roger Temam, *Navier-Stokes equations: theory and numerical analysis*, volume 343. American Mathematical Soc., 2001.

[10] Giovanni P Galdi, *An introduction to the mathematical theory of the Navier-Stokes equations: Steady-state problems*, Springer Science Business Media, 2011.

[11] Mondher Benjemaa, Bilel Krichen, and Mohamed Meslameni, *Fixed point theory in fluid mechanics: an application to the stationary navierstokes problem*, Journal of Pseudo-Differential Operators and Applications, 8(1):141146, 2017.

[12] Donald A Nield, Adrian Bejan, *Convection in porous media*, volume 3. Springer, 2006.

[13] Ulrich Hornung, *Homogenization and porous media*, volume 6. Springer Science Business Media, 2012.

[14] DD Joseph, DA Nield, and G Papanicolaou, *Nonlinear equation governing flow in a saturated porous medium*, Water Resources Research, 18(4):10491052, 1982.

[15] Lawrence E Payne and Brian Straughan, *Convergence and continuous dependence for the brinkmanforchheimer equations*, Studies in Applied Mathematics, 102(4):419439, 1999.

[16] Yan Liu, Shengzhong Xiao, and Yiwu Lin, *Continuous dependence for the brinkman forchheimer fluid interfacing with a darcy fluid in a bounded domain*, Mathematics and Computers in Simulation, 150:6682, 2018.

[17] Brian Straughan and William F Ames, *Non-standard and improperly posed problems*, volume 194. Elsevier, 1997.

[18] PN Kaloni and Jianlin Guo, *Steady nonlinear double-diffusive convection in a porous medium based upon the brinkmanforchheimer model*, Journal of Mathematical Analysis and Applications, 204(1):138155, 1996.

[19] Piotr Skrzypacz and Dongming Wei, *Solvability of the brinkman-forchheimer-darcy equation*, Journal of Applied Mathematics, 2017, 2017

[20] Stefan Banach, *Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales*, Fund. math, 3(1):133181, 1922.

[21] Sandro Salsa, *Partial differential equations in action: from modelling to theory*, volume 99. Springer, 2016.

[22] Brouwer, Luitzen Egbertus Jan, *Über abbildung von mannigfaltigkeiten*, Mathematische Annalen, 71(1):97115, 1911.

[23] Juliusz Schauder, *Der fixpunktsatz in funktionalraümen*, Studia Mathematica, 2(1):171180, 1930.

[24] Swati Das (Supervisor G. P. Raja Sekhar), *Diffusion reaction problem with oscillating behavior generated by an advective flow- The case of a third order negative reaction rate*, M.Sc. Thesis, Department of Mathematics, Indian Institute of Technology Kharagpur, India 2013

[25] M. Alam, B. Dey, G. P. R. Sekhar, *Mathematical analysis of hydrodynamics and tissue deformation inside an isolated solid tumor*, Theoretical And Applied Mechanics 45 (2) (2018) 253-278.

[26] M. Alam, B. Dey, G. P. R. Sekhar, *Mathematical modeling and analysis of hydroelastodynamics inside a solid tumor containing deformable tissue,*, ZAMM-Journal of Applied Mathematics and Mechanics, (2019) e201800223.

[27] Lawrence Craig Evans, *Partial Differential Equations*, volume 19. American Mathematical Society, Providence, 2010.

[28] Haim Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Springer Science Business Media, 2010.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 147–155

# *Geometrical Acoustic Waves in van der Waals Gases*

**K. Ambika**[a] and **R.Radha**[b*]

[a]*Dept. of Mathematics and Humanities, Nirma University, Ahmedabad, India.*; [b]*School of Mathematics and Statistics, University of Hyderabad, Hyderabad, India.*

**Abstract:** High frequency asymptotic solution is obtained to the gas dynamic equations governing one dimensional unsteady planar and non planar flows of a van der Waals gases. The effects of van der Waals parameters on the shock formation is analyzed.

## 1. Introduction

The theory of nonlinear geometrical acoustics deals with small amplitude waves to obtain the asymptotic solutions to PDEs. The pioneering work in this context can be found in Ambika and Radha [1], Arora *et. al.* [2], Blythe [3], Chu [4], Clarke and McChesney [5], He and Moodie [6], [7], [8], [9], Sharma [10] and Sharma and Srinivasan [11].

In this paper, the gas dynamic equations governing the continuous motion of an unsteady, one dimensional, planar and radially symmetric flow of van der Waals gases are considered to study the propagation of disturbances through a uniform region. In the small amplitude, high frequency limit, a solution upto the second order is obtained using the theory of nonlinear geometrical acoustics. In particular, the effects of van der Waals gas parameters on the shock formation are analyzed.

## 2. Preliminaries

We consider the disturbances in a one dimensional flow of a more general class of van der Waals gases at low temperatures and high pressures whose governing equation of state is given by

$$(p + a\rho^2)(1 - b\rho) = RT\rho, \tag{1}$$

where $p$ is the pressure of the gas, $\rho$ the density, $T$ the temperature and $R$ the universal gas constant. Here the parameter $a$ denotes the amount of attraction between each particle that leads to added pressure due to the intermolecular forces of attraction and the parameter $b$ denotes the omitted volume and is related to the volume of the gas.

*Corresponding author. Email: repakar@yahoo.com

For the given equation of state (1) of a gas, the internal energy $e$, in view of $R = (\gamma - 1)C_V$ is given by

$$e = \frac{(p + a\rho^2)(1 - b\rho) - a(\gamma - 1)\rho^2}{(\gamma - 1)\rho},$$

where $C_V$ is the specific heat at constant volume and $\gamma$ is a constant. In general, for real gases, internal energy depends on the pressure $p$ and density $\rho$. However, for an ideal gas, i.e., when $a = 0$ and $b = 0$, the internal energy $e$ becomes a function of $(p/\rho)$; equivalently, the internal energy for an ideal gas depends only on the temperature and then $\gamma$ turns out to be the specific heat ratio of an ideal gas.

The basic equations governing a planar or a radially symmetric flow of a compressible fluid, whose equation of state is given by equation (1), can be written in the following form

$$\begin{aligned} &\rho_t + u\rho_x + \rho u_x + \frac{m\rho u}{x} = 0, \\ &u_t + uu_x + \frac{p_x}{\rho} = 0, \\ &p_t + up_x + \rho F^2 \left(u_x + \frac{mu}{x}\right) = 0. \end{aligned} \tag{2}$$

Here $x$ is the spatial coordinate (being either the axial distance in flows with planar geometry ($m = 0$) or the radial distance in cylindrically symmetric ($m = 1$) or spherically symmetric ($m = 2$) flows), $t$ is the time and $u$ is a fluid particle velocity and $F(p, \rho) = \sqrt{\dfrac{\gamma p + a\rho^2(\gamma - 2 + 2b\rho)}{\rho(1 - b\rho)}}$ is the sound speed. The flow variables with a subscript denotes partial differentiation with respect to the indicated variable.

It may be noticed that the thermodynamic stability of the flow under consideration requires that the sound speed to be real and positive which leads to the following condition

$$\gamma p + a\rho^2(\gamma - 2 + 2b\rho) > 0, \tag{3}$$

and $1 - b\rho > 0$, which indeed renders the system (2) to be hyperbolic.

## 3. Geometrical acoustic solution

We consider a small amplitude wave disturbance produced by the system (2) in high frequency or geometrical acoustic limit, i.e., when the time scale $\tau$ is large compared to the time $\tau_b$ associated with the boundary data. The geometrical acoustic limit corresponds to the high frequency condition $\epsilon = \tau_b/\tau << 1$. In this limit, the variations of $\rho$, $u$ and $p$ caused by the wave are of order $O(\epsilon)$, and they depend on the characteristic variable $\xi = \phi(x, t)/\epsilon$. Therefore, we make a change of the independent variables $(x, t) \to (x, \xi)$ by defining $x = x, t = \tilde{T}(x, \xi)$ and $u(x, t) = \tilde{u}(x, \xi)$, $p(x, t) = \tilde{p}(x, \xi)$ and $\rho(x, t) = \tilde{\rho}(x, \xi)$. In view of new coordinates, the basic

equations (2) are written as

$$\begin{aligned}
&\tilde{\rho}_\xi - \tilde{T}_x(\tilde{u}\tilde{\rho}_\xi + \tilde{\rho}\tilde{u}_\xi) + \tilde{T}_\xi(\tilde{u}\tilde{\rho}_x + \tilde{\rho}\tilde{u}_x + m\tilde{\rho}\tilde{u}/x) = 0,\\
&\tilde{\rho}\tilde{u}_\xi - \tilde{T}_x(\tilde{\rho}\tilde{u}\tilde{u}_\xi + \tilde{p}_\xi) + \tilde{T}_\xi(\tilde{\rho}\tilde{u}\tilde{u}_x + \tilde{p}_x) = 0,\\
&\tilde{p}_\xi - \tilde{T}_x(\tilde{\rho}\tilde{F}^2\tilde{u}_\xi + \tilde{u}\tilde{p}_\xi) + \tilde{T}_\xi(\tilde{\rho}\tilde{F}^2\tilde{u}_x + \tilde{u}\tilde{p}_x + m\rho\tilde{F}^2\tilde{u}/x) = 0,
\end{aligned} \tag{4}$$

where $\tilde{F} = F(\tilde{p}, \tilde{\rho})$ and $(\tilde{T}_x)^{-1} = -(\xi_t/\xi_x)$ The system (4) has a non-trivial solution if

$$(\tilde{u} + \tilde{F})(\tilde{\rho}\tilde{F}\tilde{u}_x + \tilde{p}_x) + \frac{m\tilde{\rho}\tilde{u}\tilde{F}^2}{x} = 0, \tag{5}$$

which turns out to be $(\tilde{T}_x)^{-1} = -(\xi_t/\xi_x)$ is the characteristic wave front of the system (2), whose speeds are $\tilde{u} \pm \tilde{F}$ and $\tilde{u}$. Here we consider the forward facing characteristic wave front given by

$$(\tilde{T}_x)^{-1} = \tilde{u} + \tilde{F}. \tag{6}$$

We now seek the solution of system (4) in the following form

$$\begin{aligned}
\tilde{\rho}(x,\xi) &= \rho_0 + \epsilon\rho^{(1)}(x,\xi) + \epsilon^2\rho^{(2)}(x,\xi) + O(\epsilon^3),\\
\tilde{u}(x,\xi) &= u_0 + \epsilon u^{(1)}(x,\xi) + \epsilon^2 u^{(2)}(x,\xi) + O(\epsilon^3),\\
\tilde{p}(x,\xi) &= p_0 + \epsilon p^{(1)}(x,\xi) + \epsilon^2 p^{(2)}(x,\xi) + O(\epsilon^3),\\
\tilde{T}(x,\xi) &= T^{(0)}(x) + \epsilon T^{(1)}(x,\xi) + \epsilon^2 T^{(2)}(x,\xi) + O(\epsilon^3),
\end{aligned} \tag{7}$$

subject to the following boundary conditions at $x = x_0$

$$\tilde{\rho}(x_0,\xi) = \rho_0 + \epsilon g(\xi), \qquad \tilde{T}(x_0,\xi) = \epsilon\xi, \tag{8}$$

where $\rho_0, u_0$ and $p_0$ refer to the uniform reference state and $g(\xi)$ is an arbitrary function.

Introducing the expansions (7) in (4), (5), (6) and (8), and equating the coefficients of powers of $\epsilon$ to zero, we get
$O(1)$ term:

$$T_x^{(0)} = 1/F_0, \qquad T^{(0)}(x_0) = 0, \tag{9}$$

$O(\epsilon)$ term:

$$\begin{aligned}
&\rho_\xi^{(1)} = \rho_0 T_x^{(0)} u_\xi^{(1)} = \left(T_x^{(0)}\right)^2 p_\xi^{(1)},\\
&\rho_0 F_0 u_x^{(1)} + p_x^{(1)} + \frac{m}{x}\rho_0 F_0 u^{(1)} = 0,\\
&T_x^{(1)} = -\left(u^{(1)} + F_{p_0}p^{(1)} + F_{\rho_0}\rho^{(1)}\right)/{F_0}^2,\\
&\rho^{(1)}(x_0,\xi) = g(\xi), \quad T^{(1)}(x_0,\xi) = \xi,
\end{aligned} \tag{10}$$

$O(\epsilon^2)$ term:

$$\rho_\xi^{(2)} - T_x^{(0)}\left(\rho_0 u_\xi^{(2)} + \rho^{(1)} u_\xi^{(1)} + u^{(1)}\rho_\xi^{(1)}\right) - T_x^{(1)}\rho_0 u_\xi^{(1)} + T_\xi^{(1)}\left(\rho_0 u_x^{(1)} + \frac{m\rho_0}{x}u^{(1)}\right) = 0,$$

$$\rho_0 u_\xi^{(2)} + \rho^{(1)} u_\xi^{(1)} - T_x^{(0)}\left(p_\xi^{(2)} + \rho_0 u^{(1)} u_\xi^{(1)}\right) - T_x^{(1)} p_\xi^{(1)} + T_\xi^{(1)} p_x^{(1)} = 0,$$

$$p_\xi^{(2)} - T_x^{(0)}\left(\rho_0 F_0^2 u_\xi^{(2)} + F_0^2\rho^{(1)} u_\xi^{(1)} + 2\rho_0 F_0\left(F_{p_0} p^{(1)} + F_{\rho_0}\rho^{(1)}\right) u_\xi^{(1)} + u^{(1)} p_\xi^{(1)}\right)$$
$$- T_x^{(1)}\rho_0 F_0^2 u_\xi^{(1)} + \rho_0 F_0^2 T_\xi^{(1)}\left(\frac{m}{x}u^{(1)} + u_x^{(1)}\right) = 0, \qquad (11)$$

$$F_0 p_x^{(2)} + \rho_0 F_0^2 u_x^{(2)} + u_x^{(1)}\left(F_0^2\rho^{(1)} + 2\rho_0 F_0(F_{p_0} p^{(1)} + F_{\rho_0}\rho^{(1)}) + \rho_0 F_0 u^{(1)}\right) + p_x^{(1)} u^{(1)}$$
$$+ p_x^{(1)}\left(F_{p_0} p^{(1)} + F_{\rho_0}\rho^{(1)}\right) + \frac{m}{x}\left(\rho_0 F_0^2 u^{(2)} + 2\rho_0 F_0 u^{(1)} F^{(1)} + F_0^2\rho^{(1)} u^{(1)}\right) = 0,$$

$$T_x^{(2)} = \frac{1}{F_0^3}\left(u^{(1)} + F_{p_0} p^{(1)} + F_{\rho_0}\rho^{(1)}\right)^2 - \frac{1}{F_0^2}\left(u^{(2)} + F_{p_0}\, p^{(2)} + F_{\rho_0}\, \rho^{(2)}\right)$$
$$- \frac{1}{2F_0^2}\left(F_{pp_0}\, \left(p^{(1)}\right)^2 + 2F_{p\rho_0}\, p^{(1)}\rho^{(1)} + F_{\rho\rho_0}\, \left(\rho^{(1)}\right)^2\right),$$

$$\rho^{(2)}(x_0, \xi) = 0, \quad T^{(2)}(x_0, \xi) = 0,$$

where $F_0$, $F_{p_0}$, $F_{\rho_0}$, $F_{pp_0}$, $F_{p\rho_0}$ and $F_{\rho\rho_0}$ denote the evaluation of $F(p,\rho)$, $\dfrac{\partial F}{\partial p}$, $\dfrac{\partial F}{\partial \rho}$, $\dfrac{\partial^2 F}{\partial p^2}$, $\dfrac{\partial^2 F}{\partial p \partial \rho}$ and $\dfrac{\partial^2 F}{\partial^2 \rho}$ in the uniform reference state.

### 3.1. *First order solution*

Assuming that the wavefront $\xi = 0$ is moving into a medium of uniform state at rest, the equations $(9), (10)_1$ lead to

$$\begin{aligned} T^{(0)} &= (x - x_0)/F_0, \\ u^{(1)} &= \rho^{(1)} F_0/\rho_0, \\ p^{(1)} &= F_0^2\rho^{(1)}. \end{aligned} \qquad (12)$$

In view of the equations (12), we can write the equations $(10)_2$ and $(10)_3$

$$\rho_x^{(1)} + \frac{m}{2x}\rho^{(1)} = 0,$$
$$T_x^{(1)} + \alpha_1\rho^{(1)} = 0,$$

where $\alpha_1 = \dfrac{1}{\rho_0 F_0^2}\left(F_0 + \rho_0 F_{\rho_0} + \rho_0 F_0^2 F_{p_0}\right)$. The above equations are integrated subject to the boundary conditions from $(10)_4$ to yield

$$\begin{aligned} \rho^{(1)} &= g(\xi)\left(\frac{x}{x_0}\right)^{-m/2}, \\ T^{(1)} &= \xi - \alpha_1 g(\xi) J(x), \end{aligned} \qquad (13)$$

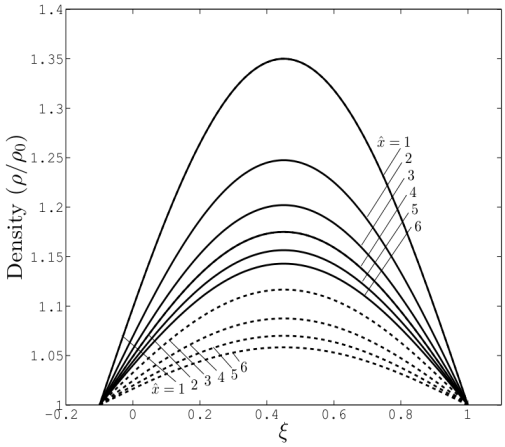



Figure 1.: The variation of the dimensionless density $\widehat{\rho}$ (defined as $(\rho_0+\rho_1)/\rho_0$) with the dimensionless variable $\xi$ (defined as $(x-F_0t)/x_0$); here $a = 0.95215$ and $b = 0.01$ which are satisfied by $\alpha_1 = 0$. The solid lines represent cylindrically symmetric flows and the dashed lines represent spherically symmetric flow configurations in a non-ideal gas. Here $\gamma = 1.4$ and $\epsilon = 0.35$.

where

$$J(x) = \begin{cases} x - x_0, & \text{if } m = 0, \\ 2x_0\left(\left(\dfrac{x}{x_0}\right)^{1/2} - 1\right), & \text{if } m = 1, \\ x_0 \log\left(\dfrac{x}{x_0}\right), & \text{if } m = 2. \end{cases}$$

Hence, the first order asymptotic solutions for the flow variables and time, correct up to $O(\epsilon)$, can be written as

$$\begin{aligned} \tilde{\rho}(x,\xi) &= \rho_0 + \epsilon g(\xi)\left(\frac{x}{x_0}\right)^{-m/2}, \\ \tilde{u}(x,\xi) &= \epsilon\frac{F_0}{\rho_0} g(\xi)\left(\frac{x}{x_0}\right)^{-m/2}, \\ \tilde{p}(x,\xi) &= p_0 + \epsilon F_0^2 g(\xi)\left(\frac{x}{x_0}\right)^{-m/2}, \\ \tilde{T}(x,\xi) &= \frac{x - x_0}{F_0} + \epsilon\left(\xi - \alpha_1 g(\xi) J(x)\right). \end{aligned} \tag{14}$$

Here the shock formation distance $x_s$ is given by,

$$\left(\alpha_1 \frac{dg(\xi)}{d\xi}\bigg|_{\xi=\xi_s}\right) J(x_s) = 1. \tag{15}$$

However, the solutions for $\rho$ and $T$, correct upto $O(\epsilon)$ are exactly same as the solutions given in [12] when $\rho = \rho_0 + \rho_1$ and $T = \tau$ which is obtained using the progressive wave approximation. Therefore the shock formation and its wave propagation follow on parallel lines when $\alpha_1 \neq 0$.

If $a$ and $b$ are chosen such that $\alpha_1 = 0$, then the wavelets are linear and do not intersect. As a result, the disturbances propagate for all time like in linear equations, without terminating into shocks (see Figure 1). Thus, it may be observed that our first order solution is not able to show the effects of nonlinear terms when $\alpha_1 = 0$. Hence, in order to study the nonlinear effects for those values of $a$ and $b$ when $\alpha_1 = 0$, we consider the second order solution.

### 3.2. *Second order solution*

In this section, to study the nonlinear effects, we consider the second order solution of the system (11) (i.e, $O(\epsilon^2)$ terms). In view of equations (12) and (13), the equations (11) and the conditions at $\xi = 0$, yield on integration

$$
\begin{aligned}
p^{(2)} =& F_0^2\rho^{(2)} + \left(F_0^3 F_{p_0} + F_0 F_{\rho_0}\right) g^2(\xi)\psi^2(x),\\
u^{(2)} =& \frac{F_0}{\rho_0}\rho^{(2)} + \left(\frac{\rho_0 F_0^2 F_{p_0} + \rho_0 F_{\rho_0} - F_0}{2\rho_0^2}\right) g^2(\xi)\psi^2(x) + \frac{mF_0^2}{2x\rho_0}\psi(x)P(\xi)\\
& - \frac{m\alpha_1 F_0^2}{4x\rho_0} J(x)\psi(x)g^2(\xi), \qquad (16)\\
T^{(2)} =& -\alpha_1 \int_{x_0}^{x} \rho^{(2)}(s,\xi)ds + \lambda_1\, g^2(\xi) \int_{x_0}^{x} (\psi(s))^2 ds - \frac{m}{2\rho_0} P(\xi) \int_{x_0}^{x} \frac{\psi(s)}{s} ds\\
& + \frac{m\alpha_1}{4\rho_0} g^2(\xi) \int_{x_0}^{x} \frac{\psi(s)J(s)}{s} ds, \qquad (17)
\end{aligned}
$$

where

$$
\begin{aligned}
\psi(x) &= \left(\frac{x}{x_0}\right)^{-m/2},\\
P(\xi) &= \int_0^{\xi} g(s)ds\ ,\\
\lambda_1 &= F_0\alpha_1^2 - \frac{\alpha_1}{2\rho_0}(1 + 2\rho_0 F_0 F_{p_0}) + \frac{1}{\rho_0^2 F_0}(1 + \rho_0 F_0 F_{p_0})\\
&\quad - \frac{1}{2F_0^2}\left(F_0^4 F_{pp_0} + 2F_0^2 F_{p\rho_0} + F_{\rho\rho_0}\right).
\end{aligned}
$$

In view of equations (12), (13) and (16), the transport equation for the second order flow variable $\rho^{(2)}$ is obtained from equation $(11)_4$ as follows

$$
\begin{aligned}
\rho_x^{(2)} + \frac{m}{2x}\rho^{(2)} - \frac{3m}{8x} F_0\alpha_1 g^2(\xi)\psi^2(x) + \frac{m(m-2)F_0}{8x^2}\psi(x)P(\xi)\\
- \frac{m(m-2)F_0}{16x^2}\alpha_1 g^2(\xi)J(x)\psi(x) = 0,
\end{aligned}
$$

which can be integrated subject to the boundary condition $(11)_6$ to get

$$
\rho^{(2)} = g^2(\xi)\psi(x)\alpha_2(x) - \frac{m(m-2)F_0}{8}\left(\frac{x - x_0}{xx_0}\right)P(\xi)\psi(x), \qquad (18)
$$

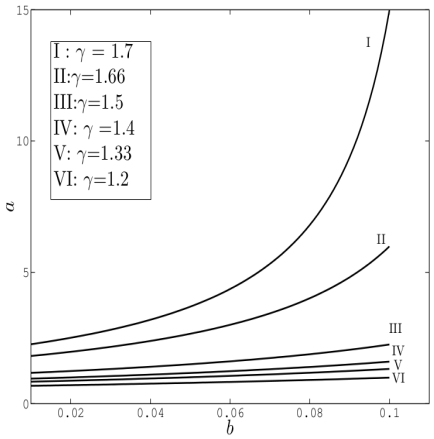



Figure 2.: The variation of the van der Waals parameter $a$ with the van der Waals parameter $b$, when $\alpha_1 = 0$, for different values of $\gamma$. Here $p_0 = 1/\gamma$ and $\rho_0 = 1$.

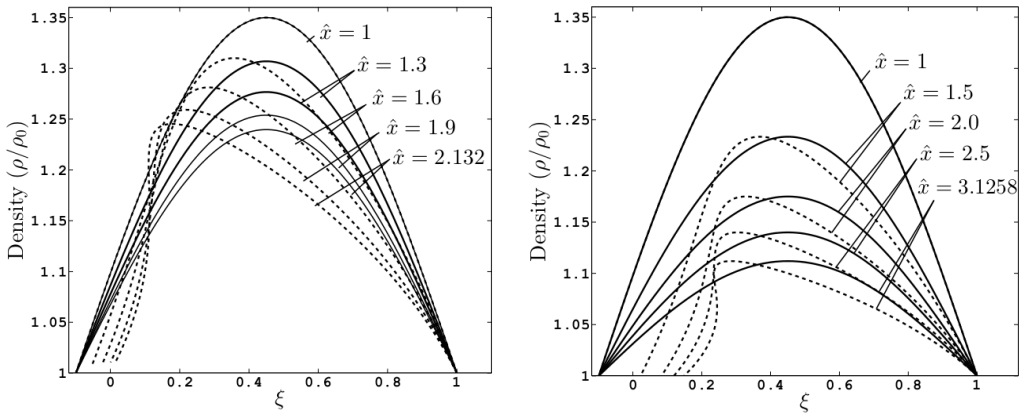
where $\alpha_2(x) = \alpha_1 \left( \dfrac{3mF_0}{8} \displaystyle\int_{x_0}^{x} \frac{\psi(s)}{s} ds + \frac{m(m-2)F_0}{16} \int_{x_0}^{x} \frac{J(s)}{s^2} ds \right)$. Equation (18) shows that the second order solution depends on the integral $P(\xi)$ which shows that the second order solution depends on the precursor wavelets also. It may be noted that the conditions on the leading wavelet remain uninfluenced by the precursor wavelets. In view of the second order solution, the shock formation distance $x_s$ on the wavelet $\xi_s$ is given by

$$\begin{aligned} \alpha_1 J(x_s) \frac{dg}{d\xi}\bigg|_{\xi=\xi_s} - \epsilon \frac{mg(\xi)}{8\rho_0} \left( (m-2)\rho_0 F_0 \alpha_1 \int_{x_0}^{x_s} \frac{s-x_0}{s x_0} \psi(s) ds - 4 \int_{x_0}^{x} \frac{\psi(s)}{s} ds \right) \\ - 2\epsilon g(\xi) \frac{dg}{d\xi}\bigg|_{\xi=\xi_s} \left( \lambda_1 \int_{x_0}^{x_s} \psi^2(s) ds - \alpha_1 \int_{x_0}^{x_s} \alpha_2(s) \psi(s) ds \right. \\ \left. \frac{m\alpha_1}{4\rho_0} \int_{x_0}^{x_s} \frac{\psi(s) J(s)}{s} ds \right) = 1. \end{aligned} \tag{19}$$

It may be observed that Figure 2 shows the variation in $a$ with respect to $b$ when $\alpha_1 = 0$, for different values of $\gamma$. As it is already mentioned in Section 3.1 that for these values of $a$ and $b$, i.e., when $\alpha_1 = 0$, the initial profile does not terminate into a shock for any time using the first order solution. However, the second order solution takes care of nonlinear terms in the formation of a shock on the wavelet $\phi = \phi_s$ at a distance $x = x_s$. The shock formation distance calculated from (19) for cylindrically and spherically symmetric flows are given in Table 1 when $\alpha_1 = 0$, for different values of $a$ and $b$. It is noticed that when $\alpha_1 = 0$, as $\gamma$ and $a$ increase, the shock formation distance obtained from equation (19) decreases in both the cylindrical and spherical waves. These effects of the second order solution, when $\alpha_1 = 0$, on the shock formation distance and distortion of the wave profile are shown in Figures 3a and 3b for a small amplitude pulse with $g(\xi) = \rho_0 \sin(\xi F_0/x_0)$, $0 \le \xi F_0/x_0 \le \pi$. For comparison, the wave propagation, using the first order solution, is also depicted in these figures. When $\alpha_1 \neq 0$, for the same initial pulse, the nonlinear distortion of the density profile, valid up to first and second order approximations is shown in Figures 4a and 4b at various distances $x$.

| | b | a | $x_{s_1}$ | $x_{s_2}$ |
|---|---|---|---|---|
| $\gamma = 1.2$ | 0.01 | 0.6789 | 2.2948 | 3.7970 |
| | 0.03 | 0.7320 | 2.2435 | 3.5790 |
| | 0.05 | 0.7928 | 2.1994 | 3.4790 |
| | 0.09 | 0.9448 | 2.1156 | 3.0384 |
| | 0.1 | 0.9910 | 2.0861 | 2.9652 |
| $\gamma = 1.33$ | 0.01 | 0.8376 | 2.1829 | 3.3181 |
| | 0.03 | 0.9150 | 2.1408 | 3.1575 |
| | 0.05 | 1.0060 | 2.1000 | 3.0124 |
| | 0.09 | 1.2462 | 2.0255 | 2.7643 |
| | 0.1 | 1.3230 | 2.0051 | 2.7090 |
| $\gamma = 1.4$ | 0.01 | 0.95215 | 2.1319 | 3.1258 |
| | 0.03 | 1.0501 | 2.0927 | 2.9869 |
| | 0.05 | 1.1679 | 2.0552 | 2.8638 |
| | 0.09 | 1.4920 | 1.9848 | 2.6505 |
| | 0.1 | 1.6000 | 1.9680 | 2.6027 |

Table 1.: Shock formation distance for different values of $\gamma$, $a$ and $b$ when $\alpha_1 = 0$; here $x_{s_1}$ and $x_{s_2}$ represent cylindrically symmetric and spherically symmetric flows respectively.

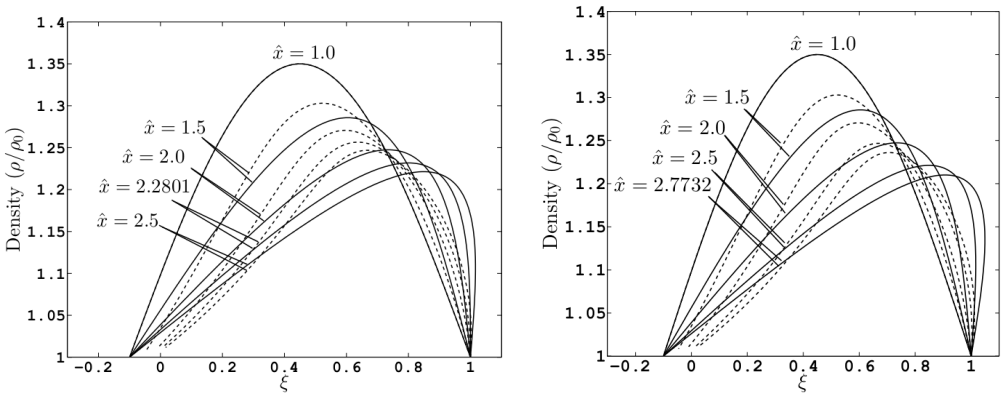



(a) Fig-3a (b) Fig-3b

Figure 3.: The variation of the dimensionless density $\widehat{\rho}$ (defined as $(\rho_0 + \rho_1)/\rho_0$) with the dimensionless variable $\xi$ (defined as $(x - F_0 t)/x_0$ ) where shock forms (a) on the wavelet $\phi_s = 2.3457$ at $\hat{x_s} = 2.1320$ in cylindrically symmetric flows, and (b) on the wavelet $\phi_s = 2.3038$ at $\hat{x_s} = 3.1258$ for spherically symmetric flows (the solid lines represent the first order solution and the dotted lines represent the second order solution). Here $a = 0.95215$, $b = 0.01$, $\gamma = 1.4$ and $\epsilon = 0.35$.

Acknowledgement

The authors K Ambika and R Radha gratefully acknowledge the financial support received from CSIR, India and UGC-SAP-DSA-I, India respectively.




(a) Fig-4a (b) Fig-4b

Figure 4.: The variation of the dimensionless density $\widehat{\rho}$ (defined as $(\rho_0 + \rho_1)/\rho_0$ ) with the dimensionless variable $\xi$ (defined as $(x - F_0 t)/x_0$ ) on the leading wavelet $\phi = 0$ in (a) cylindrically and (b) spherically symmetric flows. The density distribution, up to first and second order, is represented by the solid and the dashed lines respectively. Here $a = 0.8$, $b = 0.1$, $\gamma = 1.4$ and $\epsilon = 0.35$

## References


[1] Ambika, K. and Radha, R., Progressive waves in Non ideal gases, *International Journal of Nonlinear Mechanics* **67**, (2014) 285–290

[2] Arora, R., Tomar, A. and Singh, V. P., Shock waves in reactive hydrodynamics, *Shock Waves* **19** (2009) 145-150.

[3] Blythe, P. A., Non-linear wave propagation in a relaxing gas, *J. Fluid Mech.* **37** (1969) 31-50.

[4] Chu, B. T., Weak nonlinear waves in nonequilibrium flow in nonequilibrium flows, ed. Wegener P. P. (New York: Marcel Dekker) Vol-I, part-II, 1970.

[5] Clarke, J. F. and McChesney, A., Dynamics of relaxing gases, London: Butterworth, 1976.

[6] He, Y. and Moodie, T. B., Two-wave interactions for weakly nonlinear hyperbolic waves, *Stud. Appl. Math.* **88** (1993) 241–267.

[7] He, Y. and Moodie, T. B., Shock wave tracking for nonlinear hyperbolic systems exhibiting local linear degeneracy, *Stud. Appl. Math.* **89** (1993) 195-232.

[8] He, Y. and Moodie, T. B., Solvability and nonlinear geometrical optics for sys- tems of conservation laws having spatially dependent flux functions, *Can. Appl. Math. Q.* **2** (1994) 207-230.

[9] He, Y. and Moodie, T. B., Geometrical optics and post shock behaviour for nonlinear conservation laws, *Appl. Anal.* **57** (1995) 145-176.

[10] Sharma, V. D., Quasilinear hyperbolic systems, compressible flows, and waves, CRC Press, Taylor & Francis, USA, 2010.

[11] Sharma, V. D. and Srinivasan, G. K., Wave interaction in a non-equilibrium gas flow, *Int. J. Non-Linear Mech.* **40** (2005) 1031-1040.

[12] Ambika, K., Radha, R., and Sharma, V. D., Progressive waves in non-ideal gases, *International Journal of Non-Linear Mechanics* **67** (2014) 285-290.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 156–168

# A Survey of Age-Structured Population Models in Population Dynamics

**Joydev Halder** and **Suman Kumar Tumuluri** *

*School of Mathematics and Statistics, University of Hyderabad, Hyderabad, India.*

**Abstract:** In this article, we briefly review some age–structured models in population dynamics. In particular we focus on PDE models. In fact, in many cases the boundary conditions turn out to be nonlocal in nature which make the solutions of these models to exhibit more complex structures. Mainly we present hyperbolic and parabolic models in this article. Out of many uncanny interesting models we have chosen a few and presented here. Therefore we do not claim that the population models in this article are exhaustive.

**Keywords:** Structured population models, McKendrick–von Foerster equations, GRE inequality, semigroup theory, nonlocal boundary value problems

**AMS Subject Classifications**: 01-02; 35-02; 92-02; 92D25

## 1. Introduction

Usage of differential equations in the modeling of population dynamics can be traced back to several centuries. One of the earliest models was due to Malthus (see [45]). In that model, Malthus proposed that the rate of population growth/ decay is proportional to the size of the total population. It is widely regarded in the field of population ecology as the first principle of population dynamics. Mathematically, the model is

$$\frac{d}{dt}P(t) = \lambda P(t), \quad t \geq 0, \tag{1}$$

where $P(t)$ represents the total population size at time $t$ and $\lambda$ is the malthusian parameter of the given population. The solution of (1) is the exponential function $P(t) = e^{\lambda t}P(0),\ t \geq 0$. Thus the population blows up or decays depending on the sign of $\lambda$. The Malthus model does not refer to the effects of crowding or the limitation of resources. A population model in which the total population cannot

E-mail addresses : halderjoydev@gmail.com (J. Halder), suman.hcu@gmail.com (S. K. Tumuluri).

*Corresponding author.

grow beyond a certain limit due to resource limitations were developed by P. F. Verhulst in 1838. The Verhulst model is

$$\frac{d}{dt}P(t) = \lambda\left(1 - \frac{P(t)}{K}\right)P(t), \quad t \geq 0, \tag{2}$$

where the constants $\lambda$ and $K$ denote the intrinsic growth constant and the environmental carrying capacity respectively. Equation (2) is also known as the logistic equation. Notice that the constant $K$ is a nontrivial steady state and it is asymptotically stable. The logistic model does not consider the correlation between the population size and the mean individual fitness (often measured as per capita population growth rate) of a population. A more realistic model of population growth would allow the Allee effect (see [70]) and is given by

$$\frac{d}{dt}P(t) = \lambda P(t)\left(P(t) - A\right)\left(1 - \frac{P(t)}{K}\right), \quad t \geq 0, \tag{3}$$

where the constants $\lambda$, $K$ and $A$ are the intrinsic rate of increase, the carrying capacity and the Allee threshold respectively.

The structured population models distinguish individuals from one another according to characteristics such as age, size, location, status, and movement etc. to determine the birth, growth and death rates, interaction with each other and with the environment etc. The goal of the structured population models is to understand how these characteristics affect the dynamics of these models and thus the outcomes and consequences of the biological processes. Many authors considered age, size, spatial and maturity structured population models (see [32, 33, 65, 73, 74]).

## 2. Age-structured models : Hyperbolic PDEs

In the modeling of population dynamics, the main step is to identify some significant variables that allow the division of the population into homogeneous subgroups. Then, one can describe its dynamics through the interaction of these groups, ruled by mechanisms that depend on these variables. These variables are called structured variables. Age is one of the most natural and widely used structured variable. Let $p(x,t)$ denote the density of population that has age $x$ at time $t$. Assume that $\mu$ and $\beta$ are the age-specific mortality and age-specific fertility respectively. One of the earliest age-structured population models is due to A. G. McKendrick (see [48]) and is given by

$$\begin{cases} p_t(x,t) + p_x(x,t) = -\mu(x)p(x,t), \ x > 0, t > 0, \\ p(0,t) = \displaystyle\int_0^\infty \beta(x)p(x,t)dx, \ t > 0, \\ p(x,0) = p_0(x), \ x > 0. \end{cases} \tag{4}$$

Here $\mu$, $\beta$, $p_0$ are assumed to be non-negative functions. Model (4) is known as the renewal equation and was rediscovered by von-Foerster. It is easy to show that the solution of equation (4) is implicitly given by

$$p(x,t) = \begin{cases} p_0(t-x)e^{-\int_0^t \mu(s)ds}, & x < t, \\ p_0(x-t)e^{-\int_{x-t}^x \mu(s)ds}, & x \geq t. \end{cases} \tag{5}$$

Using the boundary condition in (4) we can write (5) as an integral equation which is also called as the renewal equation. This integral equation was studied by Lotka in detail. The steady state equation associated with equation (4) is

$$\begin{cases} \dfrac{d}{dx}\hat{p}(x) = -\mu(x)\hat{p}(x), \ x > 0, \\ \hat{p}(0) = \displaystyle\int_0^\infty \beta(x)\hat{p}(x)dx, \qquad \int_0^\infty \hat{p}(x)dx = 1. \end{cases} \tag{6}$$

From (6), it is easy to see that

$$\hat{p}(x) = \hat{p}(0)e^{\int_0^x \mu(s)ds}, \ x > 0.$$

By substituting this $\hat{p}$ in the initial condition in (6), we obtain that

$$\hat{p}(0) = \hat{p}(0)\int_0^\infty \beta(x)e^{-\int_0^x \mu(s)ds}dx.$$

Hence (6) has a nontrivial solution if and only if

$$R := \int_0^\infty \beta(x)e^{-\int_0^x \mu(s)ds}dx = 1, \tag{7}$$

where $R$ is known as the basic reproduction number which is the average off-springs produced by an individual during the reproductive period. Under appropriate assumptions on $\beta$, $\mu$ and $p_0$, it can be shown that

$$\lim_{t\to\infty} |e^{-\lambda t}p(x,t) - C_0 e^{-\lambda x - \int_0^x \mu(s)ds}| = 0,$$

where $\lambda$ is a unique solution of the characteristic equation

$$\int_0^\infty \beta(x)e^{-\lambda x - \int_0^x \mu(s)ds}dx = 1$$

and $C_0$ is a constant which depends on $p_0$. This remarkable result was conjectured by Lotka (see [73]) and proved by Feller (see [28]).

In McKendrick-von Foerster's model, the fertility and the mortality rates merely depend on the age and not on the total populations. Practically it is not the case. As there is a competition among individuals for limited resources and individuals of different ages have different advantages (disadvantages) in this competition, it is natural to assume that the fertility and mortality rates depend on the total populations. To this end, Gurtin and MaCamy introduced a nonlinear age-dependent population model where the fertility and mortality functions are density dependent (see[30]). The Gurtin-MacCamy model is given by

$$\begin{cases} p(x,t) + p_x(x,t) = -\mu(x,P(t))p(x,t), \ x > 0, t > 0, \\ p(0,t) = \displaystyle\int_0^\infty \beta(x,P(t))p(x,t)dx, \ t > 0, \\ P(t) = \displaystyle\int_0^\infty p(x,t)dx, \ t > 0, \\ p(x,0) = p_0(x), \ x > 0. \end{cases} \tag{8}$$

Using the method of characteristics, (8) can be converted into a system of to nonlinear Volterra integral equations (see [73]). The existence, uniqueness and the long time behavior of a solution of equations (8) were investigated in [30]. The steady state equation associated to (8) is

$$\begin{cases} \dfrac{d}{dx}\hat{p}(x) = -\mu(x,\hat{P})\hat{p}(x), \ x > 0, \\ \hat{p}(0) = \displaystyle\int_0^\infty \beta(x,\hat{P})\hat{p}(x)dx, \\ \hat{P} = \displaystyle\int_0^\infty \hat{p}(x)dx. \end{cases} \tag{9}$$

Under appropriate assumptions on $\mu$ and $\beta$, the solution $p(x,t)$ of equation (8) converges to the steady state $\hat{p}$ (see[30]).

## GRE Property

The notion of General Relative Entropy (GRE) is motivated by the Perron–Frobenius theory for ordinary differential equations (ODEs) and is used as a unified approach to deal with many structured population models (see [67]). Michel et al. [51, 52] have introduced the concept of GRE inequality for various structured population models including age, size, maturity. Moreover, they have used the GRE inequality (see [54, 61]) to prove *a priori* estimates and existence of solutions, long time asymptotic to the steady state, attraction to the periodic solutions etc. Here we present the GRE inequality for (4). Let $(N,\lambda,\phi)$ be a solution to the following eigenvalue problem

$$\begin{cases} N_x(x) + (\mu(x)+\lambda)N(x) = 0, \quad x > 0, \\ N(0) = \displaystyle\int_0^\infty \beta(x)N(x)dx, \end{cases}$$

and $\phi$ be a solution of the corresponding adjoint problem

$$\begin{cases} \phi_x(x) + (\mu(x)+\lambda)\phi(x) = \beta(x)\phi(0), \quad x > 0, \\ \displaystyle\int_0^\infty \phi(x)N(x)dx = 1. \end{cases}$$

If $H : \mathbb{R} \to [0,\infty)$ is a positive convex function then the following is the GRE inequality for (4):

$$\frac{d}{dt}\int_0^\infty H\left(p(x,t)e^{-\lambda t}/N(x)\right)N(x)\phi(x)dx = -D_H(pe^{-\lambda t}/N) \le 0,$$

where the dissipation of entropy $-D_H$ given by

$$-D_H(pe^{-\lambda t}/N) = N(0)\phi(0)\left(H(\int_0^\infty \frac{pe^{-\lambda t}}{N}d\mu) - \int_0^\infty H(\frac{pe^{-\lambda t}}{N}d\mu)\right),$$

and $d\mu(x) = \beta(x)N(x)/\int_0^\infty \beta(x)N(x)dx$ is a probability measure. It is one of the most powerful techniques in the analysis of long time behavior of solutions. Using the GRE inequality the asymptotic behavior of solutions of (4) is presented in the following theorem (see [60]).

THEOREM 2.1 *Assume that* $1 < \int_0^\infty \beta(x)dx < \infty$ *and* $|p_0(x)| < CN(x)$*, then the solution to* (4) *satifies*

$$\int_0^\infty |p(x,t)e^{-\lambda t} - p^0 N(x)|\phi(x)dx \downarrow 0 \quad as\ t \to \infty,$$

*with* $p^0 = \int_0^\infty p_0(x)\phi(x)dx$ *(a conserved quantity).*

In fact the GRE technique can be used to analyze nonlinear models also. For instance, consider the following nonlinear age structured population model

$$\begin{cases} p_t(x,t) + p_x(x,t) = -\mu(x)p(x,t),\ x > 0, t > 0, \\ p(0,t) = f\left(\int_0^\infty \beta(x)p(x,t)dx\right),\ t > 0, \\ p(x,0) = p_0(x),\ x > 0, \end{cases} \tag{10}$$

where $f$ is a concave functions with certain growth rate. In [50], Michel employed the GRE method to prove the asymptotic convergence of the solution of (10) to the corresponding steady state.

### *Semigroup theory methods*

The semigroup theory is another important tool to study the behaviour of the solutions of population models. The advantages of the semigroup theory over-analytical approach are that it removes the technical complexities of the proofs, allows very general nonlinearities in the model and exhibits the dynamical structure of the solution (see [49, 73]).

In [72], Webb considered equation (8) and showed that the solutions of nonlinear age-dependent population dynamics are associated with a strongly continuous semigroup of nonlinear operators in Banach space $L^1(0,\infty;\mathbb{R}^n)$. Moreover, the author proved that the nonlinear semigroup associated with the model (8) can be represented in terms of its infinitesimal generator employing an exponential formula. Using the pseudospectral differencing techniques, the infinitesimal generator associated with the semigroup of the solution operator is discretized, and the convergence results are analyzed by the authors of [14, 15]. In [15] using the numerical experiments, the authors illustrated the features of this technique. In [11], the authors formulated a more general age-structured population model as an abstract Cauchy problem in a Banach space and nonlinear semigroup theory is employed to analyze the model.

## 2.1. *Neuronal networks*

Another class of interesting age-structured models arise naturally in the study of neuronal networks (see [1, 16, 17, 35, 40]). One of the most classical models of neuronal networks is integrate-and-fire (I-F) model which was investigated by Lapicque (see [41]). In neuronal networks, another widely studied model is the Hodgkin–Huxley model (see [34]).

For a review of I-F neuron models for homogeneous synaptic inputs and for an inhomogeneous synaptic inputs see [18, 19]. For a survey in stochastic I-F models see [66]. In [20], the authors analyzed several aspects of nonlinear noisy-leaky-integrate-and-fire (NNLIF) models and studied the finite time blow-up of the weak solutions of these models. In particular, it is proved that for a suitable initial data

concentrated close to the firing potential, the weak solutions do not exist for all time. In [21], the authors investigated the well-posedness of a networked I-F model.

## 2.2. *Hematopoesies Models*

Hematopoiesis is the process by which all blood cells (red blood cells, white cells, and platelets) are produced and regulated. Mathematical modeling of hematopoietic stem cell (HSC) dynamics has been extensively studied in the past 50 years. In [42], Mackey proposed a mathematical model of HSC dynamics which is a system of two delay differential equations. This model describes the evolution of proliferating and quiescent HSCs and the delay describes the average cell cycle duration. Since then, Mackeys model has been improved by many authors. For instance, Pujo-Menjouet et al. (see [62, 63]) proved the existence of long-period oscillations, characterized situations observed in chronic myelogenous leukemia, a cancer of HSCs. Authors of [7–10] analyzed various versions of Mackeys model and investigated the effect of perturbations of the parameters in the system on the behaviour of the cell population. Adimy proposed a general model of hematopoiesis based on the Mackey model (see [9]). Further, the author used the method of characteristics to reduce the model in a system of threshold-type delay differential-difference equations. It was proved that the trivial equilibrium of the model is globally asymptotically stable if it is the only equilibrium. It was also shown that the nontrivial equilibrium, the most biologically meaningful one, can become unstable exhibiting Hopf bifurcation. These results are helpful in understanding the connection between the relatively short cell cycle durations and the relatively long periods of peripheral cell oscillations in some periodic hematological diseases.

In all these studies, the authors assumed that just after division, all cells enter the quiescent phase immediately. This restrictive assumption allows reducing the model to a delay differential system. Recently, the authors of [6]) modified the Mackey model by assuming that immediately after division only a part of daughter cells enter the quiescent phase (long-term proliferation) and the other part of cells return to the proliferating phase to divide again (short-term proliferation). Also, the authors investigated necessary and sufficient condition for the global asymptotic stability of the trivial steady state, which describes the population dying out. Moreover, the authors provided some sufficient conditions for the existence of unbounded solutions, which represent the uncontrolled proliferation of HSC population.

## 2.3. *Maturity Models*

In the age-structured population dynamics, maturity arises in various contexts. Models of the cell cycle incorporating maturity through different phases of it can be found in (see [31, 49]). Mackey et al. considered a particular time-age-maturity structured model of the biological process of hematological cell development in the bone marrow. This model is a generalization of those that had been considered previously. In [43, 44], the authors assumed that the cell cycle has two distinct phases. The cell in the first phase (rest phase) cannot divide, they mature, and they eventually enter the second phase (proliferating phase). In the second phase, the cell is committed to undegro cell division at a time $\tau$ later. Mackey et al. proved that the solution of time-age-maturity structured model exists and is globally stable (see[44]). In [22, 23], the authors analyzed this model in a particular case when maturity is independent of $\tau$. The authors also showed that the behaviour of the

solution depends on the stem cells. The numerical experiments performed earlier by Rey and Mackey suggested such a result (see [64]). A more general model than that of Mackey and Rudnicki is proposed in [7]. In that model, the rate of mortality and the rate of return from the resting phase to the proliferating phase depend on the maturity variable. Also, if the proliferating phase is long enough, then the trivial solution is the exponentially stable. Another important feature of this model is that the uniqueness and asymptotic behaviour of solutions depend only on the cells with low maturity (stem cells). For related models, one can refer [13, 24, 25].

## 3. Age-structured models : Parabolic PDEs

In [12], the authors introduced the diffusion term in McKendrik-von Foerster equation to account the variability in the DNA content which can influence the 'biological age'. In [53] the authors considered the McKendrik-von Foerster equation with diffusion (M-V-D)

$$\begin{cases} p_t(x,t) + (g(x)p(x,t))_x + \mu(x)p(x,t) = p_{xx}(x,t),\ x>0, t>0, \\ g(0)p(0,t) - p_x(0,t) = f\left(\int_0^\infty \beta(x)p(x,t)dx\right),\ t>0, \\ p(x,0) = p_0(x),\ x>0, \end{cases} \tag{11}$$

where $p$, $\mu$, $\beta$ are as in (4). The steady state equation corresponding to (11) is

$$\begin{cases} (g(x)P(x))_x + \mu(x)P(x) = P_{xx}(x),\ x>0, \\ g(0)P(0) - P_x(0) = f\left(\int_0^\infty B(x)P(x)dx\right), \\ \int_0^\infty P(x)dx < \infty, P \geq 0. \end{cases} \tag{12}$$

Apart from proving existence and uniqueness of the solution of (11), it is established that if $n_0(x) \leq CP(x)$, then the solution of (11) satisfies

$$\underline{P}(x) \leq \lim_{t\to\infty} \inf p(x,t) \leq \lim_{t\to\infty} \sup p(x,t) \leq \overline{P}(x),$$

where $\overline{P}$ (resp. $\underline{P}$) is the maximal (reps. minimal) nontrivial solution of (12).
A generalization of (11) is

$$\begin{cases} p_t(x,t) + p(x,t)_x + \mu(x,S(t))p(x,t) = p_{xx}(x,t),\ x>0, t>0, \\ g(0)p(0,t) - p_x(0,t) = \int_0^\infty \beta(x,S(t))p(x,t)dx,\ t>0, \\ S(t) = \int_0^\infty \psi(x)p(x,t)dx,\ t>0, \\ p(x,0) = p_0(x),\ x>0, \end{cases} \tag{13}$$

where the function $S(t)$ represents the weighted population which influences the mortality and fertility rates, and depends on the environmental factors.

In [36], the authors proved that, if $p_0 \in L^1(\mathbb{R}^+) \cap L^2(\mathbb{R}^+)$, then there exists a unique solution $p \in \mathbb{C}(\mathbb{R}^+; L^1(\mathbb{R}^+)) \cap L^2_{loc}(\mathbb{R}^+; W^{1,2}(\mathbb{R}^+))$. GRE inequality for linear death term and a particular type of nonlinearity in the birth term was also proved in [36].

The model in which the nonlocal term $\mu(x, S)p$ is replaced by a local reaction term in a bounded domain in $\mathbb{R}^n$ is considered in see [57, 58]. Moreover, the diffusion term is replaced by a general uniformly elliptic operator. In [57], using an upper and lower solution, the author proved that the solution to the model is non-negative and bounded whenever the initial data is so. In [37], the authors studied the well-posedness and asymptotic behavior of equation (11) posed in a finite domain $[0, a_\dagger]$ and at $x = a_\dagger$ also the Robin boundary condition is imposed similar to that of at $x = 0$.

An important feature of nonlinear parabolic equations is that thesis solutions may have finite time blow-up. In [29], the authors presented sufficient conditions to ensure the finite time blow-up/ the global existence of the solution to the following nonlocal initial boundary value problem

$$\begin{cases} p_t = \Delta p + c(x,t)p^q, \quad x \in \Omega, t > 0, \\ p(x,t) = \int_\Omega k(x,y,t)p^l(y,t)dy, \quad x \in \partial\Omega, t > 0, \\ p(x,0) = p_0(x), \quad x \in \Omega, \end{cases} \tag{14}$$

where $\Omega$ is a smooth and bounded domain in $\mathbb{R}^n$ for $n \geq 1$, and $q, l, c, k, p_0 > 0$.

## 4. Numerical methods

Finding explicit analytical solutions of population models is infeasible except in very special cases. Therefore, many authors proposed numerical schemes of these models (see [26, 27, 38, 39, 46, 59]). In this section, we present some of them which approximate the analytical solution in the time interval $[0, T]$, $T > 0$. Let, $J \in \mathbb{N}$, $h = a_\dagger/J$. We introduce the grid points $x_i = ih$, $i = 0, \ldots, J$. We denote the temporal step size by $k$, $N = [t/k]$ and time levels by $t_n = nk$, $0 \leq n \leq N$. Also, we denote by $P_i^n$ the numerical approximation of $p(x_i, t_n)$, $0 \leq i \leq J$, $0 \leq n \leq N$ and $\mathbf{P}^n = (P_0^n, \ldots, P_J^n)$, $0 \leq n \leq N$. Let $\mathbf{P}^0$ be the approximation of the initial data at the grid points when $t = 0$. Moreover, we approximate $\int_0^{a_\dagger} f(x)dx$ by the appropriate quadrature rules given by $Q_h(\mathbf{f}) = \sum_{i=0}^{J} q_i^h f(x_i)$, where $q_i^h$ are the suitable weights.

### 4.1. *Hyperbolic models*

In this section, we describe some numerical methods to find an approximate solution to the finite-age Gurtin-MacCamy model with nonlocal boundary condition in a bounded domain $[0, a_\dagger]$. For, we introduce the notation

$$x_{i-1/2} = \frac{x_{i-1} + x_i}{2}, \; DP_i^n = P_i^{n+1} - P_i^n, \; \nabla P_i^n = P_i^n - P_{i-1}^n,$$

$$P_i^{n+1/2} = \frac{P_i^{n+1} + P_i^n}{2}, \; P_{i-1/2}^n = \frac{P_i^n + P_{i-1}^n}{2}, \; 1 \leq i \leq J, \; 0 \leq n \leq N-1.$$

Lopez-Marcos has proposed several schemes on finite-age Gurtin-MacCamy model with nonlinear boundary condition. The numerical scheme given in [46] is

$$\begin{cases} \dfrac{\nabla P_i^{n-1}}{h} + \dfrac{DP_i^{n-1}}{k} + \mu(x_i, Q_h(\mathbf{P}^{n-1}))P_i^{n-1} = 0, \quad 0 \le i \le J, n \ge 1, \\ P_0^n = g(Q_h(\boldsymbol{\beta}(\mathbf{P}^n)\mathbf{P}^n), t_n), \quad n \ge 0, \\ P_i^0 = p_0(x_i), \quad 0 \le i \le J. \end{cases} \tag{15}$$

The author used the abstract framework (see [47]) to analyze scheme (15). It is proved that if $\frac{k}{h} < 1$ then (15) is a convergent scheme with first order accuracy in time and age. In [39], the authors modified this scheme to arrive at another explicit up-wind scheme which is also convergent. Another well known scheme is due to Fairweather et al. which is popularly known as box scheme (see [26]) and is given by

$$\begin{cases} \dfrac{\nabla P_i^{n+1} + \nabla P_i^n}{2h} + \dfrac{DP_i^n + DP_{i-1}^n}{2k} + \mu(x_{i-1/2}, Q_h(\mathbf{P}^{n+1/2})) \dfrac{P_{i-1/2}^{n+1/2}}{2} = 0, \\ P_0^n = g(Q_h(\boldsymbol{\beta}(\mathbf{P}^n)\mathbf{P}^n), t_n), \quad n \ge 1, \\ P_i^0 = p_0(x_i), \quad 0 \le i \le J, \end{cases} \tag{16}$$

where $Q_h$ denotes the composite trapezoidal quadrature rule. Scheme (16) is clearly implicit and at each time step the nonlinear equation must be solved by some fixed point iteration procedure. The authors proved that proposed numerical scheme is stable, convergent with second order accuracy. In [27], the authors proposed a modified box scheme which is a two step scheme in which the first level of approximation is obtained by the box-scheme. This scheme is indeed a convergent scheme and is second order accurate in time and age. For the numerical methods to the nonlinear McKendrick–von Foerster equation see [3–5, 68, 69, 71]. For more numerical schemes for age-structured population models of hyperbolic type refer to [2].

### 4.2. *Parabolic models*

Numerical methods for the nonlinear, nonlocal parabolic models have been studied by Pao (see [55–57, 59]). The author used the method of upper and lower solutions in the treatment of nonlinear and nonlocal terms. In [59], the author proposed a first order implicit numerical scheme for a reaction-diffusion equation with nonlocal boundary condition in which at each time level an iterative process is used to compute the approximate solution. In order to find the numerical solution of linear M-V-D in age the authors of [38] discretize the equation with respect to time to get a semi-implicit scheme which gives a system of ODEs. Using the energy estimates established therein and the Kačur compactness arguments the scheme proposed in [38] is proved to be convergent.

## Acknowledgements

The first author would like to thank CSIR for providing the financial support for his research. The second author would like to acknowledge UGC-SAP(DSA-1) for providing the financial assistance.

## References


[1] L. Abbott. Lapiques introduction of the integrate-and-fire model neuron (1907). *Brain Re-search Bulletin*, 50(5):303–304, 1999.

[2] L. M. Abia, O. Angulo, and J. C. Lpez-Marcos. Size-structured population dynamics models and their numerical solutions. *Discrete Contin. Dyn. Syst. Ser. B*, 4(4):1203–1222, 2004.

[3] A. S. Ackleh, K. Deng, and S. Hu. A quasilinear hierarchical size-structured model: well-posedness and approximation. *Appl. Math. Optim.*, 51(1):35–59, 2005.

[4] A. S. Ackleh, H.T.Banks, and K. Deng. A finite difference approximation for a coupled system of nonlinear size-structured populations. *Nonlinear Anal. Ser. A: Theory Methods*, 50(6):727–748, 2002.

[5] A. S. Ackleh and K. Ito. An implicit finite difference scheme for the nonlinear size-structured population model. *Numer. Funct. Anal. Optim.*, 18(9-10):865–884, 1997.

[6] M. Adimy, A. Chekroun, and T. M. Touaoula. Global asymptotic stability for an age-structured model of hematopoietic stem cell dynamics. *Appl. Anal.*, 96(3):429–440, 2017.

[7] M. Adimy and F. Crauste. Global stability of a partial differential equation with distributed delay due to cellular replication. *Nonlinear Anal.*, 54(8):1469–1491, 2003.

[8] M. Adimy, F. Crauste, M. Hbid, and R. Qesmi. Stability and hopf bifurcation for a cell population modelwith state-dependent delay. *SIAM J. Appl. Math.*, 70(5):1611–1633, 2010.

[9] M. Adimy, F. Crauste, and S. Ruan. A mathematical study of the hematopoiesis process with applications to chronic myelogenous leukemia. *SIAM J. Appl. Math.*, 65(4):1328–1352, 2005.

[10] M. Adimy, F. Crauste, and S. Ruan. Modelling hematopoiesis mediated by growth factors with applications to periodic hematological diseases. *Bull. Math. Biol.*, 68(8):2321–2351, 2006.

[11] V. Barbu and M. Iannelli. The semigroup approach to non-linear age-structured equations. *Rend. Istit. Mat. Univ. Trieste*, 28:59–71, 1997.

[12] B. Basse, B. C. Baguley, E. S. Marshall, W. R. Joseph, B. V. Brunt, G. Wake, and D. J. N. Wall. A mathematical model for analysis of the cell cycle in cell lines derived from human tumors. *Journal of Mathematical Biology*, 47(4):295–312, 2003.

[13] S. Bernard, J. Belair, and M. C. Mackey. Sufficient conditions for stability of linear differential equations with distributed delay. *Disc. Dynam. Syst. Ser. B*, 1(2):233–256, 2001.

[14] D. Breda, C. Cusulin, M. Iannelli, S. Maset, and R. Vermiglio. Stability analysis of age-structured population equations by pseudospectral differencing methods. *J. Math. Biol.*, 54(5):701–720, 2007.

[15] D. Breda, S. Maset, and R. Vermiglio. Pseudospectral differencing methods for characteristic roots of delay differential equations. *SIAM J. Sci. Comput.*, 27(2):482–495, 2005.

[16] P. C. Bressloff and J. M. Newby. Stochastic models of intracellular transport. *Review of Modern Physics*, 85(1):135–196, 2013.

[17] N. Brunel and M. C. W. van Rossum. Lapicque's 1907 paper: from frogs to integrate-and-fire. *Biol. Cybernet.*, 97(5-6):337–339, 2007.

[18] A. N. Burkitt. A review of the integrate-and-fire neuron model. i. homogeneous synaptic input. *Biol. Cybernet.*, 95(1):1–19, 2006.

[19] A. N. Burkitt. A review of the integrate-and-fire neuron model. ii. inhomogeneous synaptic input and network properties. *Biol. Cybernet.*, 95(2):97–112, 2006.

[20] M. J. Caceres, J. A. Carrillo, and B. Perthame. Analysis of nonlinear noisy integrate and fire neuron models: blow-up and steady states. *J. Math. Neurosci.*, 1(7):1–33, 2011.

[21] F. Delarue, J. Inglis, S. Rubenthaler, and E. Tanr. Global solvability of a networked integrate-and-fire model of mckean-vlasov type. *Ann. Appl. Probab.*, 25(4):2096–2133, 2015.

[22] J. Dyson, R. Villella-Bressan, and G. F. Webb. A singular transport equation modelling a proliferating maturity structured cell population. *Can. Appl. Math. Quart.*, 4(1):65–95, 1996.

[23] J. Dyson, R. Villella-Bressan, and G. F. Webb. A singular transport equation with delays. *Second International Conference on Differential Equations in Marrakesh*, 1996.

[24] J. Dyson, R. Villella-Bressan, and G. F. Webb. A nonlinear age and maturity structured model of population dynamics. i. basic theory. *J. Math. Anal. Appl.*, 242(1):93–104, 2000.

[25] J. Dyson, R. Villella-Bressan, and G. F. Webb. A nonlinear age and maturity structured model of population dynamics. ii. chaos. *J. Math. Anal. Appl.*, 242(1):255–270, 2000.

[26] G. Fairweather and J. C. L. Marcos. A box method for a nonlinear equation of population dynamics. *IMA J. Numer. Anal.*, 11(4):525–538, 1991.

[27] G. Fairweather and J. C. L. Marcos. An explicit extrapolated box scheme for the gurtin-maccamy equation. *Comput. Math. Appl.*, 27(2):41–53, 1994.

[28] W. Feller. On the integral equation of renewal theory. *Ann. Math. Statist.*, 12(3):243–267, 1941.

[29] A. Gladkov and K. I. Kim. Blow-up of solutions for semilinear heat equation with nonlinear nonlocal boundary condition. *J. Math. Anal. Appl.*, 338(1):264–273, 2008.

[30] M. E. Gurtin. The galerkin method as applied to problems in viscoelasticity. *Internat. J. Solids Structures*, 10(9):933–943, 1974.

[31] M. Gyllenberg and H. J. A. M. Heijmans. An abstract delay-differential equation modelling size dependent cell growth and division. *SIAM J. Math. Anal.*, 18(1):74–88, 1987.

[32] N. R. Hartmann. The continuity equation, a fundamental in modeling and analysis in cell kinetics. In C. Nicolini, editor, *Modeling and Analysis in Biomedicine*, volume 2, pages 1–24. World Scientific Publishing Company, Singapore, 2nd edition, 1984.

[33] H. J. A. M. Heijmans. The dynamical behaviour of the age-size distribution of a cell population. In J. A. J. Metz and . Diekmann, editors, *The Dynamics of Physiologically Structured Populations*, volume 68, pages 185–202. Springer, Singapore, 2nd edition, 1986.

[34] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, 1998.

[35] E. M. Izhikevich. Dynamical systems in neuroscience. *The MIT Press*, 2007.

[36] B. K. Kakumani and S. K. Tumuluri. On a nonlinear renewal equation with diffusion. *Math. MethodsAppl. Sci.*, 39(4):697–708, 2016.

[37] B. K. Kakumani and S. K. Tumuluri. Asymptotic behavior of the solution of a diffusion equation with nonlocal boundary conditions. *Discrete Contin. Dyn. Syst. Ser. B*, 22(2):407–419, 2017.

[38] B. K. Kakumani and S. K. Tumuluri. A numerical scheme to the Mckendrickvon Foerster equation with diffusion in age. *Numer. Methods Partial Differential Equations*, 34(6):2113–2128, 2018.

[39] Y. M. Kim and E. J. Park. An upwind scheme for a nonlinear model in age-structured population dynamics. *Comput. Math. Appl.*, 30(8):5–17, 1995.

[40] B. W. Knight. Numerical solution of structured population models. i. age structure. *The Journal of General Physiology*, 59:734–766, 1972.

[41] L. Lapicque. Recherches quantitatives sur lexcitation lectrique des nerfs traite comme une polarisation. *Journal of Physiology and Pathololgy*, 9:620–635, 1907.

[42] M. C. Mackey. Unified hypothesis for the origin of aplastic anemia and periodic hematopoiesis. *Blood*, 51(5):941–956, 1978.

[43] M. C. Mackey and R. Rudnicki. Global stability in a delayed partial differential equation describing cellular replication. *J. Math. Biol.*, 33(1):89–109, 1994.

[44] M. C. Mackey and R. Rudnicki. A new criterion for the global stability of simultaneous cell replication and maturation processes. *J. Math. Biol.*, 38(3):195–219, 1999.

[45] T. R. Malthus. *An essay on the principle of population.* J. Johnson, London, London,

1798.
[46] J. C. L. Marcos. An upwind scheme for a nonlinear hyperbolic integro-differential equation with integral boundary condition. *Comput. Math. Appl.*, 22(11):15–28, 1991.
[47] J. C. L. Marcos and J. M. Sanz-Serna. Stability and convergence in numerical analysis. iii. linear investigation of nonlinear stability. *IMA J. Numer. Anal.*, 8(1):71–84, 1988.
[48] A. G. McKendrick. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130, 1925.
[49] J. A. J. Metz and O. Diekmann. *The Dynamics of Physiologically Structured Populations*. Springer, Berlin, 1986.
[50] P. Michel. General relative entropy in a nonlinear mckendrick model. In E. H. G. Q. Chen and M. Pinsky, editors, *Stochastic Analysis and Partial Differential Equations*, volume 429, pages 205–232. Contemporary Mathematics.
[51] P. Michel, S. Mischler, and B. Perthame. General entropy equations for structured population models and scattering. *C. R. Math. Acad. Sci. Paris*, 338(9):697–702, 2004.
[52] P. Michel, S. Mischler, and B. Perthame. General relative entropy inequality: an illustration on growth models. *J. Math. Pures Appl.*, 84(9):1235–1260, 2005.
[53] P. Michel and T. M. Touaoula. Asymptotic behavior for a class of the renewal nonlinear equation with diffusion. *Mathematical Methods in the Applied Sciences*, 36(3):323–335, 2012.
[54] S. Mischler, B. Perthame, and L. Ryzhik. Stability in a nonlinear population maturation model. *Math. Models Methods Appl. Sci.*, 12(12):1751–1772, 2002.
[55] C. V. Pao. Asymptotic behavior of solutions for finite-difference equations of reaction-diffusion. *SIAM J. Numer. Anal.*, 24(1):24–35, 1987.
[56] C. V. Pao. Asymptotic behavior of solutions for finite-difference equations of reaction-diffusion. *J. Math. Anal. Appl.*, 144(1):206–225, 1989.
[57] C. V. Pao. Finite difference reaction diffusion equations with nonlinear boundary conditions. *Numer. Methods Partial Differential Equations*, 11(4):355–374, 1995.
[58] C. V. Pao. Asymptotic behavior of solutions of reaction-diffusion equations with nonlocal boundary conditions. *J. Comput. Appl. Math.*, 88(1):225–238, 1998.
[59] C. V. Pao. Numerical solutions of reaction-diffusion equations with nonlocal boundary conditions. *J. Comput. Appl. Math.*, 136(1-2):227–243, 2001.
[60] B. Perthame. *Transport equations in biology*. Birkhuser Verlag Basel, Berlin, 2007.
[61] B. Perthame and L. Ryzhik. Exponential decay for the fragmentation or cell-division equation. *Journal of Differential Equations*, 210(1):155–177, 2005.
[62] L. Pujo-Menjouet, S. Bernard, and M. C. Mackey. Long period oscillations in a $g_0$ model of hematopoietic stem cells. *SIAM J. Appl. Dyn. Syst.*, 4(2):312–332, 2005.
[63] L. Pujo-Menjouet and M. C. Mackey. Contribution to the study of periodic chronic myelogenous leukemia. *C. R. Biol.*, 327(3):235–244, 2004.
[64] A. Rey and M. Mackey. Multistability and boundary layer development in a transport equation with retarded arguments. *Can. Appl. Math. Quart.*, 1:61–81, 1993.
[65] M. Rotenberg. Transport theory for growing cell populations. *J. Theoret. Biol.*, 103(2):181–199, 1983.
[66] L. Sacerdote and M. T. Giraudo. Stochastic integrate and fire models: a review on mathematical methods and their applications. In *Stochastic biomathematical models*, pages 99–148. Springer, Heidelberg.
[67] D. Serre. *Les matrices*. Dunod, Paris, 2001.
[68] J. Shen, C. W. Shu, and M. Zhang. High resolution schemes for a hierarchical size-structured model. *SIAM J. Numer. Anal.*, 45(1):352–370, 2007.
[69] D. Sulsky. Numerical solution of structured population models. i. age structure. *J. Math. Biol.*, 31(8):817–839, 1993.
[70] G. Q. Sun. Mathematical modeling of population dynamics with allee effect. *Nonlinear Dyn*, 85(1):1–12, 2016.
[71] S. K. Tumuluri. *Age-Structured nonlinear renewal equation*. Phd thesis, UPMC (Sorbonne University), Paris, France, 2009.
[72] G. F. Webb. The semigroup associated with nonlinear age dependent population

dynamics. *Comput. Math. Appl.*, 9(3):487–497, 1983.

[73] G. F. Webb. *Theory of nonlinear age-dependent population dynamics*. Marcel Dekker, New York, 1985.

[74] R. White. A review of some mathematical models in cell kinetics. In M. Rotenberg, editor, *Biomarhematics und Cell Kinetics, Developments in Cell Biology*, volume 8, pages 243–261. Elsevier North-Holland Biomedical Press, 1981.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 169–180

# On a Question of Rawsthorne

**R. Balasubramanian**[a*] and **Priyamvad Srivastav**[b]

[a]*IMSc, Chennai and HBNI, Mumbai*; [b]*CMI, Chennai*

**Abstract:** Let $\{a_n\}_{n\geq 0}$ be a sequence of positive integers satisfying $a(0) = 1$ and $a(n) = a(\lfloor n/2\rfloor)+a(\lfloor n/3\rfloor)+a(\lfloor n/6\rfloor)$. Using Renewal theory, Erdös et al proved that $\lim_{n\to\infty} a_n/n = \frac{12}{\log 432}$. We prove the same result using tools of analytic number theory. The article is expository and highlights various useful techniques of analytic number theory.

**Keywords: Tauberian theorems, Perron formula, Linear forms in logarithms, Zero free regions**.

## 1. Introduction

Consider the sequence of integers $\{a_n\}$ given by:

$$a(0) = 1, \qquad a(n) = a\left(\lfloor n/2\rfloor\right) + a\left(\lfloor n/3\rfloor\right) + a\left(\lfloor n/6\rfloor\right), \tag{1}$$

for all natural numbers $n \geq 1$.

Now, Rawsthorne [6] raises the question whether $\lim_{n\to\infty} a_n/n$ exists and if so, what is the value? This was answered affirmatively by Erdös, Hildebrand, Odlyzko, Pudaite and Reznick [2] (and also [3]). They proved that $c = \lim_{n\to\infty} a_n/n = \frac{12}{\log 432}$. The above authors proved this as a consequence of Renewal theory.

In this article, we prove the above result using analytic number theoretic methods. In fact, we take this opportunity to make this article expository to introduce the basis of analytic number theory.

## 2. An alternate sequence

Starting from (1), we check that

$$a(0) = 1, \quad a(1) = 3, \quad a(2) = 5, \quad a(3) = 7, \quad a(4) = 9, \quad a(5) = 9, \quad a(6) = 15, \ \ldots.$$

*Corresponding author. Email: balu@imsc.res.in

Now, we define a sequence $\{b(n)\}_{n\geq 1}$ by

$$b(n) = a(n) - a(n-1).$$

Then

$$b(1) = 2, \quad b(2) = 2, \quad b(3) = 2, \quad b(4) = 2, \quad b(5) = 0, \quad b(6) = 6, \quad \ldots$$

We extend the definition of $b(n)$ to all positive real numbers by letting $b(x) = 0$ whenever $x > 0$ and $x \notin \mathbb{N}$.

We start with

LEMMA 1 *If $n \geq 2$, and $d \in \{2, 3, 6\}$, then we have*

$$a\left(\lfloor n/d \rfloor\right) - a\left(\lfloor (n-1)/d \rfloor\right) = b(n/d).$$

*Proof.* If $n$ is not a multiple of $d$, then both sides vanish. If $n$ is a multiple of $d$, say $n = kd$, then the left side is $a(k) - a(k-1) = b(k) = b(n/d)$. ■

LEMMA 2 *If $n \geq 2$, then $b(n)$ satisfies the relation*

$$b(n) = b(n/2) + b(n/3) + b(n/6). \tag{2}$$

*Proof.* Since $b(n) = a(n) - a(n-1)$, we have, using (1),

$$b(n) = \sum_{d \in \{2,3,6\}} \left(a\left(\lfloor n/d \rfloor\right) - a\left(\lfloor (n-1)/d \rfloor\right)\right),$$

and we use Lemma 1. ■

Now, we observe that $a(n)$ is the summatory function of $b(n)$. More precisely

LEMMA 3 *For any integer $N \geq 1$, we have*

$$\sum_{n \leq N} b(n) = a(n) - 1. \tag{3}$$

*Proof.* We verify it for $N = 1$, 2; The general case is by induction on $N$. ■

We also observe from (2) and by induction that

LEMMA 4 *For all $n \geq 1$, we have*

*(i)* $0 \leq a(n) \leq 3n$,
*(ii)* $0 \leq b(n) \leq 2n$,
*(iii)* $\sum_{m \leq y} b(m) \leq 3y$, *for all* $y \geq 1$.

*Proof.* For (i), it is clear that $a(n) \geq 0$. The fact that $a(n) \leq 3n$ now follows from (1) and induction. To show (ii), we see that $b(n) \geq 0$ and the rest now follows from (2) and induction. The inequality (iii) is a direct consequence of Lemma 3 and (i). ■

## 3. Generating functions

Let $g(n)$ be an arithmetic function; In order to study $g(n)$, sometimes it is convenient to study the generating function of $g(n)$; In particular if $g(n)$ is given as a linear combination of the term of the form $g(n-c)$, then we can study the growth of $g(n)$ by studying the generating power series $\sum_n g(n)x^n$.

For example, consider the Fibonacci sequence $F_n$ given by $F_0 = 1$; $F_1 = 1$; $F_n = F_{n-1}+F_{n-2}$, for all $n \geq 2$. Then, by induction, $F_n \leq 2^n$; Let $G(z) = \sum_n F_n z^n$. Then the series given is absolutely convergent in $|z| \leq 0.4$, and defines an analytic function there. In fact, the actual radius of convergence of the series is $(\sqrt{5}-1)/2$.

Now,

$$\begin{aligned} G(z)(1-z-z^2) &= \sum_{n\geq 0} F_n z^n - \sum_{n\geq 0} F_n z^{n+1} - \sum_{n\geq 0} F_n z^{n+2} \\ &= (F_0 + F_1 z + \sum_{n\geq 2} F_n z^n) - (F_0 z + \sum_{n\geq 1} F_n z^{n+1}) - \sum_{n\geq 0} F_n z^{n+2} \\ &= F_0 + (F_1 - F_0)z + \sum_{n\geq 2} F_n z^n - \sum_{n\geq 2} F_{n-1} z^n - \sum_{n\geq 2} F_{n-2} z^n, \end{aligned}$$

by a change of variable $n \to n-1$ and $n \to n-2$, so that the right hand side above equals

$$1 + \sum_{n\geq 2} (F_n - F_{n-1} - F_{n-2}) z^n = 1.$$

Thus, $G(z) = (1-z-z^2)^{-1} = (1-\alpha z)^{-1}(1-\beta z)^{-1}$, where $\alpha = (1+\sqrt{5})/2$ and $\beta = (1-\sqrt{5})/2$. By splitting into partial fractions,

$$G(z) = \frac{1}{\sqrt{5}} \left( \frac{\alpha}{1-\alpha z} - \frac{\beta}{1-\beta z} \right).$$

Now, using binomial expansion $(1-\alpha z)^{-1} = 1 + \alpha z + \alpha^2 z^2 + \dots$, we find that

$$\sum_{n\geq 0} F_n z^n = \frac{1}{\sqrt{5}} \left( \alpha(1 + \alpha z + \alpha^2 z^2 \dots) - \beta(1 + \beta z + \beta^2 z^2 \dots) \right)$$

Now, comparing the coefficients of $z^n$, we get

$$F_n = \frac{\alpha^{n+1} - \beta^{n+1}}{\sqrt{5}}, \quad \text{for all } n \geq 0.$$

## 4. Dirichlet series

Suppose that $g(n)$ is a linear combination of the terms $g(n/d)$, which is the case for the function $b(n)$ considered. In this case, we consider the generating function of the form $H(z) = \sum_{n\geq 1} \frac{g(n)}{n^z}$; These functions are called *Dirichlet series*.

When we study Dirichlet series, it is usual to represent a complex number by $s = \sigma + it$, with $\sigma$ the real part and $t$ the imaginary part instead of the customary $z = x + iy$.

Accordingly, let

$$B(s) = \sum_{n=1}^{\infty} \frac{b(n)}{n^s}.$$

Then by Lemma 4 (ii), we see the series is absolutely and uniformly convergent in $\Re(s) = \sigma \geq 3$ and defines an analytic function there.

LEMMA 5 *In $\Re(s) \geq 3$, we have*

$$B(s) = 2\left(1 - 2^{-s} - 3^{-s} - 6^{-s}\right)^{-1}.$$

*Proof.* We have

$$B(s)\left(1 - \frac{1}{2^s} - \frac{1}{3^s} - \frac{1}{6^s}\right) = \sum_{n=1}^{\infty} \frac{b(n)}{n^s} - \sum_{n=1}^{\infty} \frac{b(n)}{(2n)^s} - \sum_{n=1}^{\infty} \frac{b(n)}{(3n)^s} - \sum_{n=1}^{\infty} \frac{b(n)}{(6n)^s}$$

In the second term, replace $n$ by $n/2$ and similarly for third and fourth sum. Thus,

$$B(s)\left(1 - \frac{1}{2^s} - \frac{1}{3^s} - \frac{1}{6^s}\right) = \sum_{n=1}^{\infty} \frac{b(n) - b(n/2) - b(n/3) - b(n/6)}{n^s}.$$

Recall that $b(n) - b(n/2) - b(n/3) - b(n/6) = \begin{cases} 2, & n = 1 \\ 0, & n \geq 2. \end{cases}$ ■

LEMMA 6 *The function $(1 - 2^{-s} - 3^{-s} - 6^{-s})^{-1}$ is analytic in $\sigma > 1$.*

*Proof.* We only have to prove that $1 - 2^{-s} - 3^{-s} - 6^{-s} \neq 0$ in $\sigma > 1$. In fact,

$$\left|1 - \frac{1}{2^s} - \frac{1}{3^s} - \frac{1}{6^s}\right| \geq 1 - \frac{1}{2^\sigma} - \frac{1}{3^\sigma} - \frac{1}{6^\sigma} > 0,$$

if $\sigma > 1$. Thus, $B(s)$ is an analytic function in $\sigma > 1$. We need to find an asymptotic formula for the sum of coefficients and a standard method for doing this is to appeal to the Tauberian theorem. ■

## 5. Tauberian theorem

In 1826, Abel proved the following:

THEOREM 7 *Let $f(x) = \sum_{n=0}^{\infty} a_n x^n$, $a_n \in \mathbb{R}$; Assume that $f(x)$ converges on the real interval $(-1, 1)$. Also, assume $\sum_{n=0}^{\infty} a_n$ converges. Then $\lim_{x \to 1^-} f(x) = \sum_{n=0}^{\infty} a_n$ (here $\lim_{x \to 1^-}$ is the limit from the left).*

One would like to know whether the converse holds, i.e., if $\lim_{x\to 1} f(x)$ exists, is it true that $\sum_{n=0}^{\infty} a_n$ converges. This is clearly false. For example, take $a_n = (-1)^n$, then $f(x) = \sum_{n=0}^{\infty}(-1)^n x^n$ converges in $(-1,1)$ and equals $\frac{1}{1+x}$ there. Thus $\lim_{x\to 1} f(x)$ exists and equals $1/2$. But $\sum_{n=0}^{\infty}(-1)^n$ is not convergent.

In 1897, Tauber observed that, under some conditions, the converse holds.

THEOREM 8 *Let $f(x) = \sum_{n=0}^{\infty} a_n x^n$, with $a_n \in \mathbb{R}$. Suppose $f(x)$ converges on the real interval $(-1,1)$ and assume that $\lim_{n\to\infty} n a_n = 0$. Now, if $\lim_{x\to 1} f(x)$ exists, then $\sum_{n=0}^{\infty} a_n$ converges.*

Now, various strengthenings and generalisations are known, which all go under the name of Tauberian theory. For details, one may consult J. Korevaar [4].

## 6. Prime number theorem, A quick proof

Let $\pi(N) = \sum_{p\le N} 1$ be the number of prime numbers upto $N$. Set $\vartheta(N) = \sum_{p\le N} \log p$, be the number of primes counted with weight $\log p$. Let

$$\psi(N) = \sum_{n\le N} \Lambda(n), \qquad \text{where} \quad \Lambda(n) = \begin{cases} \log p, & n = p^m, \\ 0, & \text{otherwise.} \end{cases}$$

Then,

THEOREM 9 *The following are equivalent:*

*(a)* $\pi(N) \sim \frac{N}{\log N}$,
*(b)* $\vartheta(N) \sim N$,
*(c)* $\psi(N) \sim N$.

*Any of the statement above is called the prime number theorem.*

In 1980, D.J. Newman [5] gave a simple analytic proof of the prime number theorem, and we indicate the proof here. In fact, we follow the expository article of Korevaar [4].

We start with the following Tauberian theorem:

THEOREM 10 *Let $f(x)$ be a nonnegative, nondecreasing function on $[1,\infty]$; Assume that $f(x) = O(x)$, as $x \to \infty$. Let $g(s) = s\int_1^{\infty} f(x)x^{-s-1}\,dx$, be the Mellin transform. Assume that there exists a constant $c \in \mathbb{R}$ such that $g(s) - \frac{c}{s-1}$ can be continued analytically to a neighborhood of every point on the line $\Re(s) = 1$. Then $f(x) \sim cx$.*

We deduce PNT from Theorem 9.

Consider the Dirichlet series $\zeta(s) = \sum \frac{1}{n^s}$, called the Riemann zeta function; This is absolutely and uniformly convergent in every compact subset of $\Re(s) = \sigma > 1$ and defines an analytic function there.

Also, we have a product expansion

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1},$$

valid in $\sigma > 1$. Taking logarithmic differentiation (since $\zeta(s)$ does not vanish in this region), we get

$$\frac{-\zeta'(s)}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s}.$$

We take $f(x) = \sum_{n \le x} \Lambda(n)$ and verify the conditions of Theorem 9. The condition $f(x) = O(x)$ is easy to verify. Also,

$$\begin{aligned} g(s) &= s \int_1^{\infty} f(x) x^{-s-1}\, dx = s \int_1^{\infty} \sum_{n \le x} \Lambda(n) x^{-s-1}\, dx \\ &= s \sum_{n=1}^{\infty} \Lambda(n) \int_n^{\infty} x^{-s-1}\, dx = \sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s} = -\frac{\zeta'(s)}{\zeta(s)}. \end{aligned}$$

The product formula for $\zeta(s)$ shows that $\zeta(s) \neq 0$ in $\Re(s) > 1$ and thus $g(s)$ is analytic in $\Re(s) > 1$. Again, it is not difficult to prove that $g(s)$ has a simple pole at $s = 1$ and thus taking $c = 1$, $g(s) - \frac{1}{s-1}$ is analytic in a neighbourhood of the point $s = 1$;

In 1896, Hadamard and De la Vallee Poussin proved independently that $\zeta(1 + it) \neq 0$ if $t \neq 0$. From this, we deduce that $-\zeta'(s)/\zeta(s)$ has a pole on $\Re(s) = 1$ only at the point $s = 1$; Thus $\frac{-\zeta'(s)}{\zeta(s)} - \frac{1}{s-1}$ is analytic in a small open neighbourhood of $\Re(s) = 1$; Thus from Theorem 10, we deduce that $\psi(x) \sim x$.

## 7. The sequence $b(n)$

We want to study our problem of $\sum_{n \le x} b(n)$ by the same method; Let $f(x) = \sum_{n \le x} b(n)$; Then, as before

$$g(s) = s \int_1^{\infty} \frac{f(x)}{x^{s+1}}\, dx = \sum_{n=1}^{\infty} \frac{b(n)}{n^s} = B(s).$$

Now, let

$$D(s) = 2B(s)^{-1} = 1 - 2^{-s} - 3^{-s} - 6^{-s}; \tag{4}$$

Now, we use the fact that, in the neighbourhood of $s = 1$,

$$a^{-s} = a^{-1} \cdot a^{1-s} = a^{-1} \cdot e^{-(s-1)\log a} = a^{-1} \left(1 - (s-1)\log a + \frac{(s-1)^2 \log^2 a}{2!} + \dots\right).$$

Hence,

$$\begin{aligned} D(s) &= 1 - \sum_{d\in\{2,3,6\}} \frac{1}{d}\left(1 - (s-1)\log d + \frac{(s-1)^2}{2}\log^2 d + \dots\right) \\ &= \left(1 - \frac{1}{2} - \frac{1}{3} - \frac{1}{6}\right) + (s-1)\left(\frac{\log 2}{2} + \frac{\log 3}{3} + \frac{\log 6}{6}\right) + \dots \\ &= (s-1)\frac{\log 432}{6} + \dots . \end{aligned}$$

Thus $B(s) = 2D(s)^{-1} = \frac{12}{\log 432}\frac{1}{s-1} + \dots$, which means that $B(s)$ has a simple pole at $s = 1$ and $B(s) - \frac{c}{s-1}$ is regular at the point $s = 1$, for $c = \frac{12}{\log 432}$.

Now, we need to prove

$$1 - \frac{1}{2^s} - \frac{1}{3^s} - \frac{1}{6^s} \neq 0, \qquad s = 1 + it, \quad t \neq 0.$$

Unlike the case of the Riemann zeta function, in this case, it is easier; In fact,

$$\Re\left(1 - \frac{1}{2^{1+it}} - \frac{1}{3^{1+it}} - \frac{1}{6^{1+it}}\right) = 1 - \frac{\cos(t\log 2)}{2} - \frac{\cos(t\log 3)}{3} - \frac{\cos(t\log 6)}{6} = 0$$

if and only if $\cos(t\log 2) = \cos(t\log 3) = 1$. This implies that $t\log 2$ and $t\log 3$ are integer multiples of $2\pi$. By taking the ratios, this would mean $\frac{\log 3}{\log 2} \in \mathbb{Q}$, say $\frac{\log 3}{\log 2} = a/b$, for $a, b \in \mathbb{Z}$. This implies $3^b = 2^a$, yielding a contradiction. Thus

$$1 - \frac{1}{2^{1+it}} - \frac{1}{3^{1+it}} - \frac{1}{6^{1+it}} \neq 0, \qquad t \neq 0.$$

Thus, Theorem 10 can be applied and one can deduce

$$\sum_{n\le x} b_n \sim \frac{12}{\log 432}x.$$

## 8. The error term

Let

$$E(x) = \sum_{n\le x} b_n - \frac{12}{\log 432}x.$$

Is it possible to estimate $E(x)$? This needs:

(a) A bigger zerofree region for $B(s)^{-1}$, and
(b) Perron's formula.

In this section, we shall prove the following:

THEOREM 11 *There is a $\delta > 0$, such that*

$$E(x) = O\left(\frac{x}{(\log x)^{\delta}}\right).$$

A larger zerofree region for $D(s) = 2B(s)^{-1}$ is given by Theorem 16 in the next section. It says that for some positive constants $c_1, c_2$ and $|t|$ sufficiently large, the only pole of $B(s)$ in the region $\sigma \geq 1 - c_1|t|^{-28}$ is at $s = 1$, with residue $\frac{12}{\log 432}$. Further, it satisfies the upper bound $|B(s)| \leq c_2|t|^{28}$ for $|t|$ sufficiently large.

We start with the Perron's formula:

THEOREM 12 (Perron's formula) *Let $F(s) = \sum_{n=1}^{\infty} \frac{f(n)}{n^s}$ be absolutely convergent in $\Re(s) > 1$. Then for any $x > 0$, $x \notin \mathbb{N}$, $k \geq 0$ and $c > 1$, we have*

$$\sum_{n \leq x} f(n) \left(\log \frac{x}{n}\right)^k = \frac{k!}{2\pi i} \int_{c-i\infty}^{c+i\infty} F(s) \frac{x^s}{s^{k+1}}\, ds.$$

*In our applications, we will resort to the following truncated version:*

$$\sum_{n \leq x} f(n) \left(\log \frac{x}{n}\right)^k = \frac{k!}{2\pi i} \int_{c-iT}^{c+iT} F(s) \frac{x^s}{s^{k+1}}\, ds + O\left(\frac{k! x^c}{\pi T^k} \sum_{n=1}^{\infty} \frac{|f(n)|}{n^c} \min\left\{1, \frac{1}{T|\log \frac{x}{n}|}\right\}\right), \tag{5}$$

*where the implied constant is absolute.*

We begin with some preparatory lemmas:

For any $k \geq 0$, let

$$S_k(x) := \sum_{n \leq x} b(n) \left(\log \frac{x}{n}\right)^k. \tag{6}$$

LEMMA 13 *For all $k \geq 1$, we have*

$$S_k(x) = k \int_1^x S_{k-1}(t) \frac{dt}{t}.$$

*Proof.* This can be shown by expanding $S_{k-1}(t) = \sum_{n \leq t} b(n)(\log t/n)^{k-1}$ and evaluating the integral. ∎

LEMMA 14 *Let $k \geq 1$ and $S_k(x)$ be as given in (6). Assume further that one has an asymptotic formula $S_k(x) = cx + O(x/E(x))$, such that $E(x) \to \infty$ as $x \to \infty$. Suppose further that both $\frac{E(2x)}{E(x)}$ and $\frac{E(x/2)}{E(x)}$ remain bounded as $x \to \infty$. Then*

$$S_{k-1}(x) = \frac{cx}{k} + O\left(\frac{x}{\sqrt{E(x)}}\right),$$

*where the implied constant is absolute.*

*Proof.* Let $0 < h < x$, $h = o(x)$ and consider the quantity

$$\frac{S_k(x+h) - S_k(x)}{k} = \int_x^{x+h} \frac{S_{k-1}(t)}{t}\, dt \geq S_{k-1}(x)\frac{h}{x}\left(1 + O(h/x)\right). \tag{7}$$

since $b(n) \geq 0$ implies $S_k(t)$ is monotonically increasing. Similarly, by considering $S_k(x) - S_k(x-h)$ and following the same procedure, we get a similar lower bound.

Using the asymptotic formula for $S_k(x \pm h)$ and $S_k(x)$, we then obtain

$$kS_{k-1}(x) = \left(\frac{x}{h} + O(1)\right)\left(ch + O\left(\frac{x}{E(x)}\right)\right) = cx + O\left(h + \frac{x}{E(x)} + \frac{x^2}{hE(x)}\right)$$

Choosing $h = x/\sqrt{E(x)}$, we complete the proof. ∎

*Proof of Theorem 11.* We apply Perron's formula (5) to the sequence $\{b(n)\}$ with some $r \geq 30$ and $c = 1 + 1/\log x$. We then obtain

$$\frac{S_r(x)}{r!} = \frac{1}{2\pi i}\int_{c-iT}^{c+iT} B(s)\frac{x^s}{s^{r+1}}\, ds + O\left(\frac{x}{T^r}\sum_{n=1}^{\infty}\frac{b(n)}{n^c}\right). \tag{8}$$

Now, by partial summation and Lemma 4(iii), we find that

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{b(n)}{n^c} &= \int_1^{\infty} t^{-c} d\left(\sum_{m\leq t} b(m)\right) = t^{-c}\left(\sum_{m\leq t} b(m)\right)\Big|_1^{\infty} + O\left(\int_1^{\infty} t^{-c-1}\cdot 3t\, dt\right)\\ &= O(\log x),\end{aligned}$$

since $\sum_{m\leq t} b(m) \leq 3t$ and $c = 1 + 1/\log x$. Applying this estimate and moving the line of integration to $\sigma = 1 - c_1 T^{-28}$, we obtain

$$\begin{aligned}\frac{S_r(x)}{r!} &= \mathrm{Res}_{s=1} B(s)\frac{x^s}{s^{r+1}} + \frac{1}{2\pi i}\int_{1-c_1T^{-28}-iT}^{1-c_1T^{-28}+iT} B(s)\frac{x^s}{s^{r+1}}\, ds\\ &\quad + \int_{c-iT}^{1-c_1T^{-28}-iT} B(s)\frac{x^s}{s^{r+1}} + \int_{1-c_1T^{-28}+iT}^{c+iT} B(s)\frac{x^s}{s^{r+1}} + O\left(\frac{x\log x}{T^r}\right)\\ &= \frac{12}{\log 432}x + O\left(x^{1-c_2T^{-28}}\right) + O\left(\frac{x\log x}{T^r}\right),\end{aligned}$$

since $|B(s)| \ll |t|^{28}$ on the line $[c - iT, c + iT]$ (except for a finite range of $t$, which can be bounded separately) and that the integral of $B(s)x^s/s^{r+1}$ on the horizontal lines $[1 - c_1T^{-28} \pm iT, c \pm it]$ is negligible. Choosing

$$T = (\log x)^{1/30} \quad \text{and} \quad r = 36,$$

we find that

$$S_{36}(x) = 36!\frac{12x}{\log 432} + O\left(\frac{x}{(\log x)^{1/5}}\right).$$

Now, we apply Lemma 14 repeatedly with $k = 36, 35, 34 \ldots, 2, 1$. Assuming we have an asymptotic for $S_k(x)$, Lemma 14 gives us the asymptotic formula for $S_{k-1}(x)$ with a weaker error term. We finally obtain

$$S_0(x) = \sum_{n \le x} b(n) = \frac{12}{\log 432}x + O\left(\frac{x}{(\log x)^{\delta}}\right),$$

for some $\delta > 0$. ■

## 9. Linear forms in logarithms

We start with a special case of a theorem of Baker in 1975 [1].

THEOREM 15 *Let $\alpha_1, \ldots, \alpha_m$ be natural numbers $\geq 2$. Let $b_1, b_2, \ldots, b_m \in \mathbb{Z}$ be such that $\Lambda = b_1 \log \alpha_1 + \ldots b_n \log \alpha_n \neq 0$. Let $B = \max\{|b_1|, \ldots |b_m|\}$. Then there exists a constant $C$ depending only on $m$, $\alpha_1, \ldots \alpha_m$ such that*

$$\Lambda \geq (eB)^{-C}.$$

In fact, the theorem of Baker [1] is more general. But this version is sufficient for our purpose. From the above we easily deduce that

$$|m \log 3 - n \log 2| \geq (en)^{-C}.$$

It is known, using Padé approximation, that one can take (see [7])

$$|m \log 3 - n \log 2| \geq n^{-14}, \quad \text{for all} \quad n \geq 2. \tag{9}$$

In particular, we get $|m \log 3 - n \log 2| \geq c_1 n^{-c_2}$, if $n \geq m \geq 1$, for suitable constants $c_1$ and $c_2$. In this special case, using Padé approximation, G. Rhin has observed that one can take $c_2 = 8$.

## 10. Bigger zerofree region for $B(s)$

Now, we aim to get a bigger zero free region; Let $T$ be a large real number; Then,

THEOREM 16 *There exists $c_1 > 0$ such that the only pole of $B(s)$ in the rectangle $\sigma \geq 1 - \frac{c_1}{T^{28}}$, $|t| \leq T$, $T$ sufficiently large is at $s = 1$, and further on the line $\sigma = 1 - \frac{c_1}{T^{28}}$, one has $|B(s)| \leq c_2 T^{28}$.*

Recall that

$$D(s) = 2B(s)^{-1} = 1 - \frac{1}{2^s} - \frac{1}{3^s} - \frac{1}{6^s},$$

and let $T \geq 0$ be a large constant.

We start with the following lemmas:

LEMMA 17 *We have the following:*

*(a)* $\frac{e^x-1}{x} \leq 1.8, \quad$ *if* $\quad 0 \leq x \leq 1.$
*(b)* $\sum_{n\geq 2} \frac{x^n}{n!} \leq 0.8x, \quad$ *for* $\quad 0 \leq x \leq 1.$

*Proof.* To prove (a), note that $\frac{e^x-1}{x}$ is increasing and hence if $x \leq 1$, $\frac{e^x-1}{x} \leq e-1 \leq 1.8$.

To prove (b), we note $\sum_{n\geq 2} \frac{x^n}{n!} = e^x - 1 - x$ and the result follows from (a). ■

LEMMA 18 *If* $|s-1| \leq 1/2$, *then* $|D(s)| \geq 0.2|s-1|$.

*Proof.* Now, $D(s) = 1 - 2^{-s} - 3^{-s} - 6^{-s}$. Write

$$D(s) = D(1) + \frac{(s-1)}{1!}D'(1) + \frac{(s-1)^2}{2!}D''(1) + \dots$$

Note that we have

$$D^{(r)}(1) = \frac{(\log 2)^r}{2} + \frac{(\log 3)^r}{3} + \frac{(\log 6)^r}{6}.$$

Thus, taking $y = |s-1|$, we obtain

$$\sum_{r\geq 2} \frac{|s-1|^r}{r!}D^{(r)}(1) = \sum_{r\geq 2} \frac{(y\log 2)^r}{2\,r!} + \sum_{r\geq 2} \frac{(y\log 2)^r}{3\,r!} + \sum_{r\geq 2} \frac{(y\log 2)^r}{6\,r!},$$

Now, $0 < y < 1/2$ and hence $y \log 6 \leq 1$. Thus, by Lemma 17 (b),

$$\left|\sum_{r\geq 2} \frac{(s-1)^r}{r!}D^{(r)}(1)\right| \leq 0.8\left(\frac{y\log 2}{2} + \frac{y\log 3}{3} + \frac{y\log 6}{6}\right) \leq 0.8\,yD'(1)$$

Recalling $D(1) = 0$ and using $D'(1) = 1.0114\dots$, we have

$$|D(s)| \geq |y|D'(1) - 0.8yD'(1) \geq 0.2yD'(1) \geq 0.2y.$$

■

*Proof of Theorem 16.* We break the proof into two cases:

**Case (i)**: $|t| \leq 0.4$. In this case, we see that $|s-1| \leq 1/2$. Therefore from Lemma 18 we have $|D(s)| \geq 0.2|s-1|$, or $|B(s)| \leq 10|s-1| \leq 5$.

**Case (ii)**: $|t| > 0.4$. We want to say that $\Re(D(s))$ is nonzero and hence $D(s)$ is

nonzero. Now, assume that $D(s)$ vanishes. We have

$$\begin{aligned} 0 = \Re(D(s)) &= 1 - \frac{\cos(t\log 2)}{2^\sigma} - \frac{\cos(t\log 3)}{3^\sigma} - \frac{\cos(t\log 6)}{6^\sigma} \\ &= (1 - 2^{-\sigma} - 3^{-\sigma} - 6^{-\sigma}) + 2\left(\frac{\sin^2\left(\frac{t\log 2}{2}\right)}{2^\sigma} + \frac{\sin^2\left(\frac{t\log 3}{2}\right)}{3^\sigma} + \frac{\sin^2\left(\frac{t\log 6}{2}\right)}{6^\sigma}\right) \end{aligned}$$

Note that $1 - 2^{-\sigma} - 3^{-\sigma} - 6^{-\sigma} = O(\sigma - 1)$ and therefore

$$\sin^2\left(\frac{t\log 2}{2}\right) + \sin^2\left(\frac{t\log 3}{2}\right) = O(\sigma - 1) = O(T^{-28}). \tag{10}$$

This can happen only when $t\log 2$ and $t\log 3$ are close to rational multiples of $\pi$, which would mean $\log 3/\log 2$ is close to a rational. This would give rise to contradiction from Baker's theorem (see eq 9).

Assume that

$$t\log 2 = 2\pi a + \epsilon_1 \quad \text{and} \quad t\log 3 = 2\pi b + \epsilon_2, \tag{11}$$

where $|\epsilon_i| \leq \pi$ and $a, b = O(T)$. Substituting this in (10), we find that $\epsilon_i^2 = O(T^{-28})$, or $\epsilon_i = O(T^{-14})$. If either of $a$ or $b$ is zero, then it follows from (11) that $|t| = O(\epsilon_i) = O(T^{-14})$, which is a contradiction since $|t| > 0.4$. If $a = b = 1$, then again subtracting one equation from the other, we obtain $t\log(3/2) = \epsilon_1 - \epsilon_2 = O(T^{-14})$, again giving rise to a contradiction. We can therefore assume that $\max\{a, b\} \geq 2$. Now, again from (11), we find that

$$tb^{-14} \leq t\,|b\log 2 - a\log 3| = |b\epsilon_1 - a\epsilon_2| = O(tT^{-14}).$$

which implies $b \geq T$, a contradiction.

The lower bound $|D(s)| \geq 2c_2^{-1}T^{-28}$ for $\sigma \geq 1 - c_1/T^{28}$ (which is equivalent to the upper bound for $B(s)$) follows from the same argument since if $|D(s)|$ were smaller than $T^{-28}$, it would be absorbed in the error term $1 - 2^{-\sigma} - 3^{-\sigma} - 6^{-\sigma} = O(\sigma - 1)$ in (10) and the proof proceeds as above. ■

## References


[1] A. Baker, *Transcendental number theory*, Cambridge University Press, London-New York, 1975.

[2] P. Erdös, A. Hildebrand, A. Odlyzko, P. Pudaite, and B. Reznick, *A very slowly converging sequence*, Math. Mag., **58**, 1985, 51-52.

[3] P. Erdös, A. Hildebrand, A. Odlyzko, P. Pudaite, and B. Reznick, *The asymptotic behavior of a family of sequences*, Pacific J. Math, **126**, no. 2, 1987, 227-241.

[4] J. Korevaar, *Tauberian theory: a century of developments.*, Springer Science and Business Media **329**, 2013.

[5] D. J. Newman, *Simple analytic proof of the prime number theorem*, Amer. Math. Monthly, **87**, 1980, No. 9, 693–696.

[6] D. Rawsthorne, *Problem 1185*, Math. Mag. **57** (1984), 42.

[7] G. Rhin, *Approximants de Padé et mesures effectives dirrationalité*, Progr. Math., **71** 1987.

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 181–201

# Higher Moments of Riemann zeta-function on Certain Lines and the Abelian Group Problem

**A. Sankaranarayanan**
Email: sank@math.tifr.res.in

*School of Mathematics,*
*Tata Institute of Fundamental Research,*
*Homi Bhabha Road, Mumbai-400005, India*

**Abstract:** In this paper we improve the existing upper bound for higher moments of Riemann zeta-function on certain lines and use the techniques to improve the existing upper bound for the mean-square of the error term related to the abelian group problem.

**AMS Subject Classifications**: 11M06, 11M36.

**Keywords:** Higher moments, approximate functional equation, large values of Dirichlet sums, Γ- function.

## 1. Introduction

Let $D_k(x) = \sum_{n \leq x} d_k(n)$ where $d_k(n)$ is the number of ways of expressing $n$ as a product of $k$ factors. It can be represented as ( for $k \geq 1$ an integer )

$$D_k(x) = xP_k(\log x) + E_k(x) \tag{1}$$

where $E_k(x)$ is called the error term of the generalised divisor problem and $P_k$ is a polynomial of degree $k-1$, If we define $\beta_k$ to be the least number such that

$$\frac{1}{x}\int_0^x (E_k(y))^2\, dy = O\left(x^{2\beta_k+\epsilon}\right) \tag{2}$$

for every positive $\epsilon$, then to study $\beta_k$, naturally studying higher moments of the Riemann zeta-function on certain lines come into picture (for example see theorem 12.5 of [24]). In this connection to study $\beta_4$ and $\beta_5$, Heath-Brown proved that (see [9] and [10] respectively)

$$\int_{\frac{T}{2}}^{T} |\zeta(\frac{5}{8}+it)|^8 dt \ll T(\log T)^{38} \tag{3}$$

and

$$\int_{\frac{T}{2}}^{T} |\zeta(\frac{11}{20}+it)|^{10}dt \ll T^{\frac{3}{2}}(\log T)^{52} \tag{4}$$

for any $T \geq 10$. Because of the functional equation of $\zeta(s)$, from 3 and 4, one deduces that $\beta_4 \leq \frac{3}{8}$ and $\beta_5 \leq \frac{9}{20}$. However it is known that $\beta_k \geq \frac{k-1}{2k}$ (for $k = 2, 3, \cdots$ see theorem 12.6 (A) of [24]) and hence we obtain $\beta_4 = \frac{3}{8}$.

Let $b(n)$ denote the number of isomorphism classes of Abelian groups of order $n$. The arithmetic function $b(n)$ is multiplicative and it has a generating series

$$\sum_{n=1}^{\infty} b(n)n^{-s} = \zeta(s)\zeta(2s)\zeta(3s)\cdots \tag{5}$$

for $\Re s > 1$. If we write

$$A(x) = \sum_{n \leq x} b(n) \tag{6}$$

then, we can write

$$A(x) = \sum_{j=1}^{5} c_j x^{1/j} + \Delta(x) \tag{7}$$

where $\Delta(x)$ is known as the error term of the counting function $A(x)$. It is known on the one hand that

$$\Delta(x) \ll x^{\frac{1}{4}+\epsilon} \tag{8}$$

(see [23]) and on the other hand that

$$\int_{1}^{Y} (\Delta(x))^2 dx = \Omega\left(Y^{\frac{4}{3}}(\log Y)\right) \tag{9}$$

(see [2]). In [10], Heath-Brown proved that

$$\int_{1}^{Y} (\Delta(x))^2 dx \ll Y^{\frac{4}{3}}(\log Y)^{89} \tag{10}$$

for $Y \geq 2$. Earlier in [11], Ivić obtained an upper bound for l.h.s of (10) where he proved the exponent of $Y$ to be $\frac{39}{29}$ instead of $\frac{4}{3}$. A result similar to (10) was announced by Balasubramanian and Ramachandra in [2] before, but their claim could not be substantiated. Apart from using certain new ideas, essentially the results (3), (4) and (10) quoted above are excellent consequences of the twelfth power moments of the Riemann zeta-function on the critical line. We also refer to some important papers [12], [15] in this connection by Ivić and Motohashi, Jutila and Motohashi respectively. Developements in the mean square theory of the Riemann zeta and other zeta functions can be found in a nice article by Matsumoto in [17].

Conrey and Gonek (see [6]) presented some nice heuristic arguments to formulate several important conjetures and using these conjectures, they obtained some highly interesting results. We quote one of their conjectures (see conjecture 2 of [6]) which is very much relevant here. Let

$$\mathbb{D}_{k,y}(s) = \sum_{n=1}^{y} \frac{d_k(n)}{n^s}.$$

Then,

CONJECTURE 1 *For every positive integer $k$, we have*

$$\int_T^{2T} |\zeta(\frac{1}{2}+it)|^{2k} dt \sim \int_T^{2T} |\mathbb{D}_{k,x}(\frac{1}{2}+it)|^2 dt + \int_T^{2T} |\mathbb{D}_{k,y}(\frac{1}{2}+it)|^2 dt \tag{11}$$

*where*

$$xy = \left(\frac{T}{2\pi}\right)^k \quad \text{with} \quad xy \geq \frac{1}{2} \quad \text{or} \quad y \gg \left(\frac{T}{2\pi}\right)^k \quad \text{if} \quad x = 0. \tag{12}$$

Montgomery-Vaughan theorem asserts that

$$\int_T^{2T} \left| \sum_{n=1}^{N} b(n) n^{it} \right|^2 dt = \sum_{n=1}^{N} |b(n)|^2 \left(T + O(n)\right).$$

Thus one obtains

$$\int_T^{2T} |\mathbb{D}_{k,y}(\frac{1}{2}+it)|^2 dt \sim \frac{a_k}{\Gamma(k^2+1)} T (\log y)^{k^2}$$

for $y \ll T$ where

$$a_k := \prod_p \left( \left(1 - \frac{1}{p}\right)^{k^2} \sum_{r=0}^{\infty} \frac{(d_k(p^r))^2}{p^r} \right).$$

For higher moments , because of (12), at least one of $x$ and $y$ must be significantly larger than $T$. In our situation, one expects that (see also theorem 12.7 of [24])

CONJECTURE 2 *For $T \geq 10$, and for every integer $k \geq 2$ , the inequality*

$$\int_{\frac{T}{2}}^{T} |\zeta(1 - \frac{(k-1)}{2k} - it)|^{2k} dt \ll T (\log T)^{A_k}$$

*holds for some positive constant $A_k$.*

*Remark* 1 Conjecture 2 is true for $k = 2, 3$ and 4 due to Hardy, Cramér (see [24]) and Heath-Brown (see [9]) respectively. For $k \geq 5$, the problem is still open.

It is the aim of this paper to prove the following theorems.

THEOREM 1.1 *For $T \geq 10$, we have*

$$\int_{\frac{T}{2}}^{T} \left|\zeta(\frac{11}{20}+it)\right|^{10} dt \ll T^{\frac{3}{2}}(\log T)^{\frac{101}{4}}(\log\log T)^{9}.$$

THEOREM 1.2 *For $T \geq 10$, we have*

$$\int_{\frac{T}{2}}^{T} \left|\zeta(\frac{5}{8}+it)\right|^{8} dt \ll T(\log T)^{\frac{39}{2}}(\log\log T)^{6}.$$

THEOREM 1.3 *For $T \geq 10$, we have*

$$\int_{\frac{T}{2}}^{T} \left|\zeta(\frac{2}{3}+it)\right|^{6} dt \ll T(\log T)^{11}.$$

THEOREM 1.4 *For $Y \geq 10$, we have*

$$\int_{1}^{Y} (\Delta(x))^2 dx \ll Y^{\frac{4}{3}}(\log Y)^{\frac{129}{4}}(\log\log Y)^{5}.$$

*Remark* 2 Clearly Theorems 1.1, 1.2 and 1.4 improve the log factors of (4), (3) and (10) respectively. It is plain from the proof (see sections 3, 4 and 5) that all these improvements follow from the use of Lemmas 3.1, 3.4, 3.5 (which is 12th power moment of $\zeta(s)$ due to Heath-Brown ) and Lemma 3.6. It has to be mentioned that the method of proving Theorems 1.1, 1.2, and 1.4 is actually due to Heath-Brown except the usage of lemmas mentioned above and the facts of steps 2 of sections 4 and 5. Proof of Theorem 1.3 is comparatively simpler as can be seen at the end of section 4.

## 2. Notation and preliminaries

1. $C_1, C_2, \cdots$ denote effective absolute constants, sometimes may be positive.
2. $s = \sigma + it, w = u + iv$ unless otherwise specified.
3. $f(x) = O(g(x))$ and $f(x) \ll g(x)$ both mean that there exists a positive constant $C_1$ such that $|f(x)| < C_1 g(x)$ for $x \geq x_0$.
Also, $f(x) = \Omega(g(x))$ means that there exists a positive constant $C_1'$ such that $|f(x)| > C_1' g(x)$ for arbitrarily large value of $x$.
4. We use the following facts

$$(i). \qquad |\Gamma(\frac{w}{h}+1)| \ll \exp(-C_2\frac{|\Im w|}{h}). \tag{13}$$

(ii). If we write $\zeta(s) = \chi(s)\zeta(1-s)$, then we use

$$t^{\frac{1}{2}-\sigma} \ll \chi(s) \ll t^{\frac{1}{2}-\sigma} \tag{14}$$

for any $\sigma$ satisfying $C_3 \leq \sigma \leq C_4$ as $t \to \infty$.
(iii). For $A_i \geq 0, \alpha_i \geq 0$ and $\sum \alpha_i = 1$, we use the inequality

$$\min(A_1, A_2, \cdots, A_l) \leq A_1^{\alpha_1} \ . \ A_2^{\alpha_2} \cdots A_l^{\alpha_l}. \tag{15}$$

(iv). For $B_1, B_2$ (any two complex numbers) and $l \geq 1$ (an integer), we have

$$|B_1 + B_2|^l \leq 2^l(|B_1|^l + |B_2|^l). \tag{16}$$

## 3. Some lemmas

LEMMA 3.1 *Let $\frac{T}{10} \leq t \leq 10T$ and $X = 10000\sqrt{T}$. For $2 \geq \sigma \geq \frac{1}{2}$, we have*

$$\zeta(s) = \sum_{n \leq 100X} e^{-(\frac{n}{X})^h} n^{-s} + \chi(s) \sum_{n \leq 100X} n^{s-1} + O(T^{-1})$$

*where $h = 10 \log T$ and $\chi(s)$ is the conversion factor satisfying $\zeta(s) = \chi(s)\zeta(1-s)$.*

*Proof.* We have

$$\begin{aligned} F(s) &= \sum_{n=1}^{\infty} e^{-(\frac{n}{X})^h} n^{-s} \\ &= \frac{1}{2\pi i} \int\limits_{u=2, |v| \leq 40(\log T)^3} \zeta(s+w) X^w \Gamma(\tfrac{w}{h} + 1) \tfrac{dw}{w} + O(T^{-1}). \end{aligned} \tag{17}$$

We move the line of integration in the integral of 17 to $u = -\frac{h}{2}$. The residue arising from the pole $w = 0$ is $\zeta(s)$. Since $\zeta(s) \ll t^{\frac{1}{2}-\sigma+\frac{\sigma}{3}}$ for $\sigma \leq \frac{1}{2}$, we obtain

$$\begin{aligned} F(s) = &\frac{1}{2\pi i} \int\limits_{u=-\frac{h}{2}, |v| \leq 40(\log T)^3} \zeta(s+w) X^w \Gamma(\tfrac{w}{h} + 1) \tfrac{dw}{w} + O(T^{-1}) \\ &+ \zeta(s) + O(T^{\frac{1}{2}+\frac{10}{3}\log T + \epsilon} X^2 e^{-4(\log T)^2}). \end{aligned} \tag{18}$$

Note that we have used

$$|\Gamma(\frac{w}{h} + 1)| \ll \exp(-C_2 \frac{|\Im w|}{h}) \ll e^{-4(\log T)^2}$$

on the horizontal portions. Hence we have on using the functional equation of $\zeta(s)$,

$$\begin{aligned} F(s) = &\frac{1}{2\pi i} \int\limits_{u=-\frac{h}{2}, |v| \leq 40(\log T)^3} \chi(s+w) \left( \sum_{n \leq 100X} n^{s+w-1} + \sum_{n > 100X} n^{s+w-1} \right) \Gamma(\tfrac{w}{h} + 1) X^w \tfrac{dw}{w} \\ &+ \zeta(s) + O(T^{-1}) \\ &= J_1 + J_2 + \zeta(s) + O(T^{-1}) \ \text{ say.} \end{aligned} \tag{19}$$

Since $\chi(s) \ll t^{\frac{1}{2}-\sigma}$ and $h = 10\log T$, clearly we have

$$J_2 \ll (20T)^{\frac{1}{2}-\sigma+\frac{h}{2}}(100X)^{\sigma-\frac{h}{2}}X^{-\frac{h}{2}}(\log T)^3 \ll T^{-1}. \tag{20}$$

In $J_1$, we move the line of integration to $u = 10$. The residue arising from the pole at $w = 0$ is $\chi(s)\sum\limits_{n\le 100X} n^{s-1}$. Hence, we obtain

$$\begin{aligned} J_1 &= -\chi(s)\sum_{n\le 100X} n^{s-1} \\ &\quad + \frac{1}{2\pi i}\int\limits_{u=10,|v|\le 40(\log T)^3} \chi(s+w)\left(\sum_{n\le 100X} n^{s+w-1}\right)\Gamma(\tfrac{w}{h}+1)X^w\tfrac{dw}{w} + O(T^{-1}) \\ &= -\chi(s)\sum_{n\le 100X} n^{s-1} + J(s,\chi) + O(T^{-1}), \quad \text{say} \end{aligned} \tag{21}$$

where

$$J(s,\chi) = \frac{1}{2\pi i}\int\limits_{u=10,|v|\le 40(\log T)^3} \chi(s+w)\left(\sum_{n\le 100X} n^{s+w-1}\right)\Gamma(\frac{w}{h}+1)X^w\frac{dw}{w}.$$

We notice that

$$\frac{1}{2\pi i}\int\limits_{u=10,|v|\le 40(\log T)^3} \chi(s+w)\left(\sum_{n\le \sqrt{X}} n^{s+w-1}\right)\Gamma(\frac{w}{h}+1)X^w\frac{dw}{w}$$

$$\ll (\frac{T}{10})^{\frac{1}{2}-\sigma-u}(\sqrt{X})^{\sigma+u}X^u(\log T)^3 \ll T^{-1} \tag{22}$$

and since (see p.106 of [24])

$$\sum_{a<n\le b(\le 2a)} n^{-it} \ll a^{\frac{1}{2}}t^{\frac{1}{6}} + at^{-\frac{1}{6}},$$

we have (for $\frac{T}{10} \le t \le 10T, |v| \le 40(\log T)^3$ with $u = 10$)

$$\sum_{\sqrt{X}<n\le 100X} n^{s+w-1} \ll \sum_{n=0}^{[\frac{\log X}{2\log 2}+\frac{\log 100}{\log 2}]+1} (2^n\sqrt{X})^{\sigma+u-\frac{1}{2}}T^{\frac{1}{6}} \ll T^{\frac{1}{6}}X^{\sigma+u-\frac{1}{2}}$$

and hence we get

$$\frac{1}{2\pi i}\int\limits_{u=10,|v|\le 40(\log T)^3} \chi(s+w)\left(\sum_{\sqrt{X}<n\le 100X} n^{s+w-1}\right)\Gamma(\frac{w}{h}+1)X^w\frac{dw}{w}$$

$$\ll (\frac{T}{10})^{\frac{1}{2}-\sigma-u}T^{\frac{1}{6}}X^{\sigma+u-\frac{1}{2}}(\log T)^3 \ll T^{-1}. \tag{23}$$

Now, the lemma follows from (17) to (23) on noticing the fact that

$$\sum_{n\geq 100X} e^{-(\frac{n}{X})^h} n^{-s} \ll \sum_{n\geq 100X} \frac{X^h}{n^{\sigma+h}} \ll \frac{(100X)^{1-\sigma}}{100^h} \ll T^{-5}.$$

This proves the lemma. ■

*Remark* 3 This lemma may be compared with lemma 1 of [4].

LEMMA 3.2 *Let $f(z)$ be analytic in $|z| \leq r$ and there $\max|f(z)| \leq H$ $(H \geq 3)$. Let $C_3 \geq 1$. Then*

$$|f(0)| \leq (24C_5 \log H)\left(\frac{1}{2r}\int_{-r}^{r} |f(iy)|\, dy\right) + H^{-C_5}.$$

*Proof.* See page 2 of [3]. This is remark 2 below the corollary to the theorem stated in the introduction of [3]. ■

LEMMA 3.3 *Let $a_n, n = 1, 2, \cdots, N$ be any set of complex numbers, then*

$$\int_0^T |\sum_{n=1}^{N} a_n n^{-it}|^2 dt = \sum_{n=1}^{N} |a_n|^2 (T + O(n)).$$

*Proof.* See [18] or [21]. ■

LEMMA 3.4 *Consider any set $S^{**}$ of complex numbers with the following property*
*(i). $\Re s \geq \sigma_0$ for all $s \in S^{**}$,*
*(ii). $|\Im s| \leq T$,*
*(iii). $|\Im(s - s')| \geq 1$ for any $s, s' \in S^{**}$ with $s \neq s'$. Then for $V > 0$, we have*

$$|\{s/s \in S^{**}, |S(s)| \geq V\}| \ll_\epsilon GNV^{-2} + TG^3NV^{-6}(\log\log T)^2 + T^\epsilon + T^{1+\epsilon}G^2V^{-4}$$

*where*

$$S(s) = \sum_{n=N}^{2N} a_n n^{-s} \quad ; \quad G = \sum_{n=N}^{2N} |a_n|^2 n^{-2\sigma}.$$

*Proof.* See [22]. ■

LEMMA 3.5 *For $T \geq 10$, we have*

$$\int_0^T |\zeta(\frac{1}{2} + it)|^{12} dt \ll T^2 (\log T)^{17}.$$

*Proof.* See [8]. ■

LEMMA 3.6 *Let* $F(s)=\sum_{n=1}^{\infty}e^{-(\frac{n}{X})^h}n^{-s}$. *For* $T\geq 10$, *we have*

$$\int_{(\log T)^4}^{T}|F(\frac{1}{2}+it)|^{12}dt\ll T^2(\log T)^{17}(\log\log T)^{12}.$$

*Proof.* Gabriel's two variables convexity theorem (see [7]) which says that if

$$J(\sigma,\lambda)=\left(\int_0^T|f(\sigma+it)|^{\frac{1}{\lambda}}\right)^{\lambda},$$

then

$$J(\sigma,p\lambda+q\mu)\ll(J(\alpha,\lambda))^p(J(\beta,\mu))^q$$

where $\alpha\leq\sigma\leq\beta, p=\frac{\beta-\sigma}{\beta-\alpha}$ and $q=\frac{\sigma-\alpha}{\beta-\alpha}$. By fixing $\lambda=\frac{1}{12},\mu=\frac{1}{12},\alpha=\frac{1}{2},\beta=2,\sigma=\frac{1}{2}+\frac{1}{\log T}$, we obtain, firstly

$$\begin{aligned}\int_0^T|\zeta(\tfrac{1}{2}+\tfrac{1}{\log T}+it)|^{12}dt&\ll\left(\int_0^T|\zeta(\tfrac{1}{2}+it)|^{12}dt\right)^{1-\frac{2}{3\log T}}\left(\int_0^T|\zeta(2+it)|^{12}dt\right)^{\frac{2}{3\log T}}\\&\ll\left(T^2(\log T)^{17}\right)^{1-\frac{2}{3\log T}}T^{\frac{2}{3\log T}}\\&\ll T^2(\log T)^{17}e^{-\frac{10\log\log T}{\log T}}.\end{aligned}\tag{24}$$

Note that here we have used Lemma 3.5. Now let $s=\frac{1}{2}+it$ and $(\log T)^4\leq|t|\leq 10T$. Then

$$\begin{aligned}F(s)&=\tfrac{1}{2\pi i}\int\limits_{u=2,|v|\leq(\log T)^3}\zeta(s+w)\Gamma(\tfrac{w}{h}+1)X^w\tfrac{dw}{w}+O(T^{-1})\\&=\tfrac{1}{2\pi}\int\limits_{|v|\leq(\log T)^3}\zeta(\tfrac{1}{2}+\tfrac{1}{\log T}+i(t+v))\Gamma(1+\tfrac{1}{h\log T}+i\tfrac{v}{h})X^{\frac{1}{\log T}+iv}\tfrac{dv}{(\frac{1}{\log T}+iv)}\\&\quad+O(T^{-1})\end{aligned}\tag{25}$$

by moving the line of integration to $u=\frac{1}{\log T}$. Since

$$\int\limits_{|v|\leq(\log T)^3}\frac{dv}{|\frac{1}{\log T}+iv|}=\int\limits_{|v|\leq\frac{1}{\log T}}+\int\limits_{\frac{1}{\log T}<|v|\leq(\log T)^3}\ll\log\log T,$$

applying Hölder's inequality, we obtain,

$$|F(s)|^{12}\ll(\log\log T)^{11}\int\limits_{|v|\leq(\log T)^3}|\zeta(\frac{1}{2}+\frac{1}{\log T}+i(t+v))|^{12}\frac{dv}{|\frac{1}{\log T}+iv|}$$

and hence , we get from (24)

$$\int_{(\log T)^4}^{T}|F(s)|^{12}dt$$

$$\ll (\log\log T)^{11} \int\limits_{|v|\le(\log T)^3} \frac{dv}{|\frac{1}{\log T}+iv|}\left(\int_0^{2T}|\zeta(\tfrac{1}{2}+\tfrac{1}{\log T}+i(t+v))|^{12}dt\right) \qquad (26)$$
$$\ll T^2(\log T)^{17}(\log\log T)^{12}$$

and this proves the lemma. ■

## 4. Proofs of Theorems 1.1 , 1.2 and 1.3

First we prove Theorem 1.1 and the proofs of Theorems 1.2 and 1.3 follow in a similar manner.

### 4.1. *Step 1*

we define

$$S(L,t)=L^{\frac{1}{20}}\{\sum_{L<n\le 2L}\frac{e^{-(\frac{n}{x})^h}}{n^{\frac{11}{20}+it}}+\chi(\frac{11}{20}+it)\sum_{L<n\le 2L}n^{-\frac{9}{20}+it}\}. \qquad (27)$$

We replace the integral by a sum over well-spaced points $t_n \in [\frac{T}{2},T]$ for which $|t_m-t_n|\ge 1 \ \ (m\ne n)$. From Lemma 3.1, we have

$$\zeta(s)=\sum_{n\le 100X}e^{-(\frac{n}{X})^h}n^{-s}+\chi(s)\sum_{n\le 100X}n^{s-1}+O(T^{-1})$$

for $\frac{T}{2}\le t\le 10T, \frac{1}{2}\le \Re s\le \frac{9}{10}$. Hence, we have

$$\zeta(\frac{11}{20}+it)\ll(\log T)\max_{L\le CT^{\frac{1}{2}}}|S(L,t)|L^{-\frac{1}{20}} \qquad (28)$$

where $C$ is an effective positive constant and $L$ runs over powers of 2. For the value of $L$ giving the maximum in (28), we have $|S(L,t)|\gg 1$. Now it follows that

$$I_T=\int_{T/2}^{T}|\zeta(\frac{11}{20}+it)|^{10}dt\ll(\log T)^{10}L^{-\frac{1}{2}}\sum_n|S(L,t_n)|^{10}. \qquad (29)$$

If $N(U)$ is the number of well spaced points $t_n$ for each $U$ ( $U$ runs over powers of 2 ) for which , we have $U<|S(L,t_n)|\le 2U$, then clearly we have

$$I_T\ll(\log T)^{11}L^{-\frac{1}{2}}N(U)U^{10} \qquad (30)$$

where $L\ll\sqrt{T}, 1\ll U\ll L^{\frac{1}{2}}$.

### 4.2. *Step 2*

Since $\frac{T}{2} \leq t \leq T$ and $\chi(s) \sim t^{\frac{1}{2}-\sigma}$, we define

$$S_1(L,t) = L^{\frac{1}{20}} \sum_{L<n\leq 2L} e^{-(\frac{n}{x})^h} n^{-\frac{11}{20}-it}$$

and

$$S_2(L,t) = (\frac{L}{T})^{\frac{1}{20}} \sum_{L<n\leq 2L} n^{-\frac{9}{20}+it}.$$

Now, we observe that the inequality

$$\max(U, ||S_1| - |S_2||) \leq |S_1 + S_2| \leq \min(U, |S_1| + |S_2|)$$

holds. If $N_1(U)$ and $N_2(U)$ are the number of well-spaced points $t_n$ for each $U$ such that $\frac{U}{2} < |S_1| \leq U$ and $\frac{U}{2} < |S_2| \leq U$ respectively, then this implies that for every well-spaced point $t$ counted in $N(U)$ , there is a well-spaced point $t$ ( need not be the same ) being counted either in $N_1(U)$ or in $N_2(U)$ and hence, clearly we have

$$N(U) \leq N_1(U) + N_2(U). \tag{31}$$

### 4.3. *Step 3*

Here we consider (by Perron's formula)

$$\begin{aligned}(S_1(L,t))^3 &= \frac{L^{\frac{3}{20}}}{2\pi i}\int_{\frac{9}{20}+\frac{1}{\log T}-i\frac{T}{4}}^{\frac{9}{20}+\frac{1}{\log T}+i\frac{T}{4}} F^3(\tfrac{11}{20}+it+w)\left(\frac{(2L)^{3w}-(L)^{3w}}{w}\right)dw + O(\log T)\\ &= \frac{L^{\frac{3}{20}}}{2\pi}\int_{-\frac{T}{4}}^{\frac{T}{4}} F^3(\tfrac{1}{2}+i(t+v))\left(\frac{(2L)^{-\frac{3}{20}+3iv}-(L)^{-\frac{3}{20}+iv}}{(-\frac{1}{20}+iv)}\right)dv\\ &+ O(\frac{L^{\frac{3}{2}}T^{\frac{1}{2}}(\log T)^{\frac{17}{4}}(\log\log T)^3}{T}) + O(\log T)\end{aligned} \tag{32}$$

by moving the line of integration to $u = -\frac{1}{20}$. Of course from Lemma 3.6, since $\frac{T}{2} \leq t \leq T$ and $|v| \leq \frac{T}{4}$, we obtain

$$|F(\frac{1}{2}+i(t+v))| \ll T^{\frac{1}{6}}(\log T)^{\frac{17}{12}}(\log\log T)$$

for $\frac{T}{10} \leq |t+v| \leq 10T$ and hence the horizontal portion contributes

$$O(\frac{L^{\frac{3}{2}}T^{\frac{1}{2}}(\log T)^{\frac{17}{4}}(\log\log T)^3}{T}).$$

Now, we notice that $L \ll T^{\frac{1}{2}}$ and hence, we obtain

$$\begin{aligned}(S_1(L,t))^3 &= \frac{L^{\frac{3}{20}}}{2\pi}\int_{\frac{T}{4}}^{\frac{5T}{4}} \frac{F^3(\frac{1}{2}+i\nu)(2^{-\frac{3}{20}+3i(\nu-t)}-1)L^{-\frac{3}{20}+3i(\nu-t)}}{(-\frac{1}{20}+i(\nu-t))}d\nu\\ &+ O(T^{\frac{1}{4}}(\log T)^{\frac{17}{4}}(\log\log T)^3).\end{aligned} \tag{33}$$

Therefore, we have by Hölder's inequality,

$$\begin{aligned}|S_1(L,t)|^{12} &= |(S_1(L,t))^3|^4 \\ &\ll |\frac{1}{2\pi}\int_{\frac{T}{4}}^{\frac{5T}{4}} \frac{F^3(\frac{1}{2}+i\nu)(2^{-\frac{3}{20}+3i(\nu-t)}-1)L^{3i(\nu-t)}}{(-\frac{1}{20}+i(\nu-t))}d\nu|^4 \\ &+ O(T(\log T)^{17}(\log\log T)^{12}) \\ &\ll \left(\int_{\frac{T}{4}}^{\frac{5T}{4}} |F^3(\tfrac{1}{2}+i\nu)|^{12}\frac{d\nu}{\frac{1}{20}+|\nu-t|}\right)\left(\int_{\frac{T}{4}}^{\frac{5T}{4}}\frac{d\nu}{\frac{1}{20}+|\nu-t|}\right)^3 \\ &+ O(T(\log T)^{17}(\log\log T)^{12}).\end{aligned} \tag{34}$$

Therefore, we get from (34) and Lemma 3.6,

$$U^{12}N_1(U) \ll \sum_n |S_1(L,t_n)|^{12} \ll T^2(\log T)^{18}(\log\log T)^{12}. \tag{35}$$

### 4.4. *Step 4*

Applying Lemma 3.3 to $(S_1(L,t))^2$, from Lemma 3.2 we obtain

$$\begin{aligned}U^4N_1(U) &\ll (\log T)\int_{\frac{T}{2}}^{T}|S_1(L,t)|^4dt \\ &\ll L^{\frac{1}{5}}(\log T)\sum_{L^2<n\le(2L)^2}(d(n))^2n^{-\frac{11}{10}}(T+O(n)) \\ &\ll (T+L^2)(\log L)^3(\log T) \\ &\ll T(\log T)^4\end{aligned} \tag{36}$$

since $\sum\limits_{x\le n\le 2x}\frac{d^2(n)}{n} \ll (\log x)^3$ and $L \ll \sqrt{T}$.

### 4.5. *Step 5*

Applying Lemma 3.4 to the partial sum $(S_1(L,t))^2$, with the notation of Lemma 3.4, we have $V = U^2, N \ll L^2$ and $G = L^{\frac{1}{5}}\sum\limits_{L^2<n\le(2L)^2}(d(n))^2n^{-\frac{11}{10}} \ll (\log T)^3$ and hence we obtain

$$\begin{aligned}N_1(U) &\ll L^2U^{-4}(\log T)^3 + TL^2U^{-12}(\log T)^9(\log\log T)^2 \\ &+ T^\epsilon + T^{1+\epsilon}U^{-8}(\log T)^6.\end{aligned} \tag{37}$$

### 4.6. *Step 6*

Analogously, we obtain the same upper bounds for $N_2(U)$ on studying the partial sum $S_2(L,t)$ and hence we conclude in view of the inequality (31),

$$N(U) \ll TU^{-4}(\log T)^4, \tag{38}$$

$$\begin{aligned}N(U) &\ll L^2U^{-4}(\log T)^3 + TL^2U^{-12}(\log T)^9(\log\log T)^2 \\ &+ T^\epsilon + T^{1+\epsilon}U^{-8}(\log T)^6,\end{aligned} \tag{39}$$

and

$$N(U) \ll T^2U^{-12}(\log T)^{18}(\log\log T)^{12}. \tag{40}$$

## 5. Completion of the proof of Theorem 1.1

Let

$$\mathcal{M}^* = \max\left(L^2U^{-4}(\log T)^3, TL^2U^{-12}(\log T)^9(\log\log T)^2, T^{\epsilon}, T^{1+\epsilon}U^{-8}(\log T)^6\right).$$

### 5.1. *case (i)*

we suppose that $\mathcal{M}^* = L^2U^{-4}(\log T)^3$.

Using the inequality (15), we obtain

$$\begin{aligned} N(U) &\ll \{L^2U^{-4}(\log T)^3\}^{\frac{1}{4}}\{T^2U^{-12}(\log T)^{18}(\log\log T)^{12}\}^{\frac{3}{4}} \\ &\ll T^{\frac{3}{2}}L^{\frac{1}{2}}U^{-10}(\log T)^{\frac{57}{4}}(\log\log T)^9 \end{aligned} \tag{41}$$

and hence from (30), we get

$$I_T \ll T^{\frac{3}{2}}(\log T)^{25.25}(\log\log T)^9. \tag{42}$$

### 5.2. *case (ii)*

Suppose that $\mathcal{M}^* = TL^2U^{-12}(\log T)^9(\log\log T)^2$.
Then from (38) and (40), on using the inequality (15), we obtain

$$\begin{aligned} N(U) &\ll \{TU^{-4}(\log T)^4\}^{\frac{1}{4}}\{TL^2U^{-12}(\log T)^9(\log\log T)^2\}^{\frac{1}{4}} \\ &\quad . \ \{T^2U^{-12}(\log T)^{18}(\log\log T)^{12}\}^{\frac{1}{2}} \\ &\ll T^{\frac{3}{2}}L^{\frac{1}{2}}U^{-10}(\log T)^{13} \end{aligned} \tag{43}$$

and hence from (30), we get

$$I_T \ll T^{\frac{3}{2}}(\log T)^{24}. \tag{44}$$

### 5.3. *case (iii)*

Suppose that $\mathcal{M}^* = T^{\epsilon}$.
First of all we notice that

$$\int_{\frac{T}{2}}^{T} |\sum_{n\le T^{\frac{1}{10}}} n^{-\frac{11}{20}-it}|^{10}dt \ll \sum_{n\le T^{\frac{1}{2}}} d_5(n)n^{-\frac{11}{10}}(T+O(n)) \ll T^{\frac{3}{2}}$$

and

$$\int_{\frac{T}{2}}^{T} |\chi(\frac{11}{20}+it)\sum_{n\le T^{\frac{1}{10}}} n^{-\frac{9}{20}+it}|^{10}dt \ll T^{\frac{3}{2}}$$

and hence we can always assume that $L \gg T^{\frac{1}{10}}$. If $U \ll T^{(\frac{3}{2}-2\epsilon).\frac{1}{9}}$, then clearly we have

$$\begin{aligned} I_T &\ll (\log T)^{11} L^{-\frac{1}{2}} U^{10} N(U) \\ &\ll (\log T)^{11} L^{-\frac{1}{2}} U^{10} T^{\epsilon} \\ &\ll (\log T)^{11} U^{9} T^{\epsilon} \\ &\ll U^{9}.T^{2\epsilon} \\ &\ll T^{\frac{3}{2}} \end{aligned} \tag{45}$$

since $U \ll L^{\frac{1}{2}}$. Otherwise, suppose that $U \gg T^{(\frac{3}{2}-2\epsilon).\frac{1}{9}}$, then anyway, we have from (15)

$$\begin{aligned} N(U) &\ll \min\{T^2U^{-12}(\log T)^{18}(\log\log T)^{12}, T^{\epsilon}\} \\ &\ll \{T^2U^{-12}(\log T)^{18}(\log\log T)^{12}\}^{\frac{3}{4}}.\{L^2U^{-4}(\log T)^3\}^{\frac{1}{4}} \\ &\ll T^{\frac{3}{2}}(\log T)^{\frac{57}{4}}(\log\log T)^9 L^{\frac{1}{2}}U^{-10}, \end{aligned} \tag{46}$$

since $T^{\epsilon} \ll L^2U^{-4}(\log T)^3$. Thus, we have

$$I_T \ll (\log T)^{11} L^{-\frac{1}{2}} U^{10} N(U) \ll T^{\frac{3}{2}}(\log T)^{\frac{101}{4}}(\log\log T)^9.$$

### 5.4. *case (iv)*

Suppose that $\mathcal{M}^* = T^{1+\epsilon}U^{-8}(\log T)^6$.
Then clearly, we have from (15),

$$\begin{aligned} N(U) &\ll \min\{T^2U^{-12}(\log T)^{18}(\log\log T)^{12}, T^{1+\epsilon}U^{-8}(\log T)^6\} \\ &\ll \{T^2U^{-12}(\log T)^{18}(\log\log T)^{12}\}^{\frac{1}{2}}\{T^{1+\epsilon}U^{-8}(\log T)^6\}^{\frac{1}{2}} \\ &\ll T^{\frac{3}{2}+2\epsilon}U^{-10}. \end{aligned} \tag{47}$$

Therefore, we get

$$I_T \ll (\log T)^{11} L^{-\frac{1}{2}} U^{10} N(U) \ll T^{\frac{3}{2}+2\epsilon} L^{-\frac{1}{2}} \ll T^{\frac{3}{2}}$$

since $L \gg T^{\frac{1}{10}}$. This proves Theorem 1.1.

Theorem 1.2 follows in a similar way by defining similar $S(L_1, t)$ and on noticing the fact that

$$J_T = \int_{T/2}^{T} |\zeta(\frac{5}{8} + it)|^8 dt \ll (\log T)^9 L_1^{-1} U_1^8 N(U_1).$$

Analysis similar to the proof of Theorem 1.1 now completes the proof of Theorem 1.2.

Theorem 1.3 follows in a rather simpler way by defining similar $S(L_2, t)$ and on noticing the fact that

$$\begin{aligned} J_T' &= \int_{T/2}^{T} |\zeta(\tfrac{2}{3} + it)|^6 dt \\ &\ll (\log T)^7 L_2^{-1} U_2^6 N(U_2) \\ &\ll T(\log T)^{11} L_2^{-1} U_2^2 \\ &\ll T(\log T)^{11}, \end{aligned} \tag{48}$$

since from (38), similar estimate

$$N(U_2) \ll TU_2^{-4}(\log T)^4$$

holds and also we have

$$L_2 \ll \sqrt{T}, \quad 1 \ll U_2 \ll L_2^{\frac{1}{2}}.$$

## 6. Proof of Theorem 1.4

After suitable modifications, the analysis of Ivić (see [11], p.19-21) leads to

$$\int_{\frac{Y}{2}}^{Y} (\Delta(x))^2 dx \ll Y^{\frac{4}{3}} (\log Y)^2 \left( \max_{1 \le T \le Y} T^{-1} I_T^* \right) \tag{49}$$

where

$$I_T^* = \int_{\frac{T}{2}}^{T} |\zeta(1-\sigma+it)\zeta(1-2\sigma+2it)\zeta(3\sigma+3it)\zeta(4\sigma+4it)\zeta(5\sigma+5it)|^2 dt$$

and

$$\sigma = \frac{1}{6} + \frac{1}{\log Y}.$$

In view of the inequality, $2|ab| \le |a|^2 + |b|^2$, we have

$$I_T^* \le \max(J_T^*, J_T^{*\prime}) \tag{50}$$

where

$$J_T^* = \int_{\frac{T}{2}}^{T} |\zeta^2(3\sigma+3it)\zeta^4(4\sigma+4it)\zeta^4(5\sigma+5it)| dt \tag{51}$$

and

$$J_T^{*\prime} = \int_{\frac{T}{2}}^{T} |\zeta^2(3\sigma+3it)\zeta^4(1-\sigma+it)\zeta^4(1-2\sigma+2it)| dt.$$

Since the estimation of $J_T^{*\prime}$ is similar to $J_T^*$, we restrict our attention to $J_T^*$.

### 6.1. *Step 1*

We define

$$S_3^*(L,3t) = \sum_{L<n\le 2L} e^{-(\frac{n}{X})^h} n^{-3\sigma-3it} + \chi(3\sigma+3it) \sum_{L<n\le 2L} n^{3\sigma+3it-1}, \tag{52}$$

$$S_4^*(M,4t) = M^{\frac{1}{6}} \{ \sum_{M<n\le 2M} e^{-(\frac{n}{X})^h} n^{-4\sigma-4it} + \chi(4\sigma+4it) \sum_{M<n\le 2M} n^{4\sigma+4it-1} \} \tag{53}$$

and

$$S_5^*(N,5t) = N^{\frac{1}{3}}\{\sum_{N<n\leq 2N} e^{-(\frac{n}{X})^h} n^{-5\sigma-5it} + \chi(5\sigma+5it)\sum_{N<n\leq 2N} n^{5\sigma+5it-1}\}. \quad (54)$$

As in the proof of the previous section, we replace the integral by a sum over well spaced points $t_n \in [\frac{T}{2}, T]$ for which $|t_m - t_n| \geq 1 \ \ (m \neq n)$. From Lemma 3.1, as before, we obtain,

$$J_T^* \ll (\log T)^{10} M^{-\frac{2}{3}} N^{-\frac{4}{3}} \sum_n |(S_3^*(L,3t_n))^2 (S_4^*(M,4t_n))^4 (S_5^*(N,5t_n))^4| \quad (55)$$

for certain fixed $L, M, N$ with

$$|S_3^*(L,3t_n)| \gg 1, \quad |S_4^*(M,4t_n)| \gg 1, \quad |S_5^*(N,5t_n)| \gg 1,$$

since all other terms arising from (52), (53) and (54) contribute a small quantity to $J_T^*$.
We classify the points $t_n$ according to the ranges $U < |S_3^*| \leq 2U, V < |S_4^*| \leq 2V$ and $W < |S_5^*| \leq 2W$ in which the relevant sums lie. Here $U, V, W$ run over powers of 2 with

$$1 \ll U \ll L^{\frac{1}{2}}, 1 \ll V \ll M^{\frac{1}{2}} \ \text{ and } \ 1 \ll W \ll N^{\frac{1}{2}}. \quad (56)$$

if there are $N^*(U,V,W)$ well-spaced points $t_n$ for each triplet $(U,V,W)$, it follows that

$$J_T^* \ll (\log T)^{13} U^2 V^4 W^4 M^{-\frac{2}{3}} N^{-\frac{4}{3}} N^*(U,V,W). \quad (57)$$

Note that $L \ll T^{\frac{1}{2}}, M \ll T^{\frac{1}{2}}$ and $N \ll T^{\frac{1}{2}}$.

### 6.2. *Step 2*

we define

$$S_{31}^*(L,3t) = \sum_{L<n\leq 2L} e^{-(\frac{n}{X})^h} n^{-3\sigma-3it}, \quad (58)$$

$$S_{32}^*(L,3t) = T^{\frac{1}{2}-3\sigma} \sum_{L<n\leq 2L} n^{3\sigma+3it-1}, \quad (59)$$

$$S_{41}^*(M,4t) = M^{\frac{1}{6}} \sum_{M<n\leq 2M} e^{-(\frac{n}{X})^h} n^{-4\sigma-4it}, \quad (60)$$

$$S_{42}^*(M,4t) = M^{\frac{1}{6}} T^{\frac{1}{2}-4\sigma} \sum_{M<n\leq 2M} n^{4\sigma+4it-1}, \quad (61)$$

$$S_{51}^{*}(N,5t)=N^{\frac{1}{3}}\sum_{N<n\leq 2N}e^{-(\frac{n}{X})^{h}}n^{-5\sigma-5it} \tag{62}$$

and

$$S_{52}^{*}(N,5t)=N^{\frac{1}{3}}T^{\frac{1}{2}-5\sigma}\sum_{N<n\leq 2N}n^{5\sigma+5it-1}. \tag{63}$$

We note that $\frac{T}{2}\leq t\leq T$ and $t^{\frac{1}{2}-\sigma}\ll\chi(s)\ll t^{\frac{1}{2}-\sigma}$. If $N_{j}^{*}(U,V,W)$ $(j=1,2,\cdots,8)$ are the number of well-spaced points $t_n$ for each triplet $(U,V,W)$ such that

$$\frac{U}{2}<|S_{31}^{*}|\leq U,\quad \frac{V}{2}<|S_{41}^{*}|\leq V,\quad \frac{W}{2}<|S_{51}^{*}|\leq W \tag{64}$$

$$\frac{U}{2}<|S_{31}^{*}|\leq U,\quad \frac{V}{2}<|S_{42}^{*}|\leq V,\quad \frac{W}{2}<|S_{51}^{*}|\leq W \tag{65}$$

$$\frac{U}{2}<|S_{31}^{*}|\leq U,\quad \frac{V}{2}<|S_{42}^{*}|\leq V,\quad \frac{W}{2}<|S_{52}^{*}|\leq W \tag{66}$$

$$\frac{U}{2}<|S_{31}^{*}|\leq U,\quad \frac{V}{2}<|S_{41}^{*}|\leq V,\quad \frac{W}{2}<|S_{52}^{*}|\leq W \tag{67}$$

and

$$\frac{U}{2}<|S_{32}^{*}|\leq U,\quad \frac{V}{2}<|S_{42}^{*}|\leq V,\quad \frac{W}{2}<|S_{52}^{*}|\leq W \tag{68}$$

$$\frac{U}{2}<|S_{32}^{*}|\leq U,\quad \frac{V}{2}<|S_{41}^{*}|\leq V,\quad \frac{W}{2}<|S_{51}^{*}|\leq W \tag{69}$$

$$\frac{U}{2}<|S_{32}^{*}|\leq U,\quad \frac{V}{2}<|S_{41}^{*}|\leq V,\quad \frac{W}{2}<|S_{52}^{*}|\leq W \tag{70}$$

$$\frac{U}{2}<|S_{32}^{*}|\leq U,\quad \frac{V}{2}<|S_{42}^{*}|\leq V,\quad \frac{W}{2}<|S_{51}^{*}|\leq W \tag{71}$$

respectively, then (as before) clearly we have

$$N^{*}(U,V,W)\leq\sum_{j=1}^{8}N_{j}^{*}(U,V,W). \tag{72}$$

### 6.3. *Step 3*

Proceeding as before in steps 3, 4, 5 and 6 of the proof of Theorem 1.1, we obtain for any integer $k \geq 1$, (from Lemma 3.2 and Lemma 3.3),

$$U^{2k}N^*(U,V,W) \ll (L^k + T)(\log T)^{k^2}, \tag{73}$$

$$V^{2k}N^*(U,V,W) \ll (M^k + T)(\log T)^{k^2} \tag{74}$$

and

$$W^{2k}N^*(U,V,W) \ll (N^k + T)(\log T)^{k^2}. \tag{75}$$

(By applying Lemma 3.3), we get

$$\begin{aligned} N^*(U,V,W) \ll{}& L^2U^{-4}(\log T)^3 + TL^2U^{-12}(\log T)^9(\log\log T)^2 \\ &+ T^\epsilon + T^{1+\epsilon}U^{-8}(\log T)^6, \end{aligned} \tag{76}$$

$$\begin{aligned} N^*(U,V,W) \ll{}& M^2V^{-4}(\log T)^3 + TM^2V^{-12}(\log T)^9(\log\log T)^2 \\ &+ T^\epsilon + T^{1+\epsilon}V^{-8}(\log T)^6 \end{aligned} \tag{77}$$

and

$$\begin{aligned} N^*(U,V,W) \ll{}& N^2W^{-4}(\log T)^3 + TN^2W^{-12}(\log T)^9(\log\log T)^2 \\ &+ T^\epsilon + T^{1+\epsilon}W^{-8}(\log T)^6. \end{aligned} \tag{78}$$

Also we have

$$N^*(U,V,W) \ll T^2U^{-12}(\log Y)^{18}(\log\log Y)^{12}, \tag{79}$$

$$N^*(U,V,W) \ll T^2V^{-12}(\log Y)^{18}(\log\log Y)^{12} \tag{80}$$

and

$$N^*(U,V,W) \ll T^2W^{-12}(\log Y)^{18}(\log\log Y)^{12}. \tag{81}$$

## 7. Completion of the proof of Theorem 1.4

Because of the symmetry in the bounds of $N^*(U,V,W)$, it suffices to restrict ourselves to $N \leq M$. Otherwise

$$M^{\frac{4}{3}}N^{\frac{2}{3}} \leq M^{\frac{2}{3}}N^{\frac{4}{3}}.$$

We consider the following four cases.

**case 1.** $1 \leq N \leq T^{\frac{1}{4}}$ : $1 \leq M \leq T^{\frac{1}{4}}$ : $(N \leq M)$.

**case 2.** $T^{\frac{1}{4}} \leq N \ll T^{\frac{1}{2}}$ : $T^{\frac{1}{4}} \leq M \ll T^{\frac{1}{2}}$ : $(N \leq M)$.

**case 3.** $1 \le N \le T^{\frac{1}{16}} \quad : \quad T^{\frac{1}{4}} \le M \ll T^{\frac{1}{2}} \quad : \quad (N \le M).$

**case 4.** $T^{\frac{1}{16}} \le N \le T^{\frac{1}{4}} \quad : \quad T^{\frac{1}{4}} \le M \ll T^{\frac{1}{2}} \quad : \quad (N \le M).$

Clearly this will exhausts all possibilities when $N \le M$.

**case 1.** $1 \le N \le T^{\frac{1}{4}} \quad : \quad 1 \le M \le T^{\frac{1}{4}} \quad : \quad (N \le M).$

We take $k = 2$ in (73) and $k = 4$ in (74). Hence by using the inequality (15), we get

$$\begin{aligned} N^*(U,V,W) &\ll \{TV^{-8}(\log T)^{16}\}^{\frac{1}{2}}\{TU^{-4}(\log T)^4\}^{\frac{1}{2}} \\ &\ll TV^{-4}U^{-2}(\log T)^{10} \\ &\ll TU^{-2}V^{-4}W^{-4}(\log T)^{10}.W^4 \\ &\ll TU^{-2}V^{-4}W^{-4}(\log T)^{10}.M^{\frac{2}{3}}N^{\frac{4}{3}} \end{aligned} \tag{82}$$

since $W^4 \ll N^2 = N^{\frac{4}{3}}N^{\frac{2}{3}} \ll N^{\frac{4}{3}}M^{\frac{2}{3}}$.

**case 2.** $T^{\frac{1}{4}} \le N \le C_6T^{\frac{1}{2}} \quad : \quad T^{\frac{1}{4}} \le M \le C_7T^{\frac{1}{2}} \quad : \quad (N \le M).$

We use (79), (80) and the inequality (75) with $k = 4$. From (15), we obtain

$$\begin{aligned} N^*(U,V,W) &\ll \{T^2V^{-12}(\log Y)^{18}(\log\log Y)^{12}\}^{\frac{1}{3}}\{\max\left(T, N^4\right)W^{-8}(\log T)^{16}\}^{\frac{1}{2}} \\ &\quad . \ \{T^2U^{-12}(\log Y)^{18}(\log\log Y)^{12}\}^{\frac{1}{6}} \\ &\ll TU^{-2}V^{-4}W^{-4}(\log Y)^{17}(\log\log Y)^6\{\max\left(T^{\frac{1}{2}}, N^2\right)\} \\ &\ll TU^{-2}V^{-4}W^{-4}(\log Y)^{17}(\log\log Y)^6M^{\frac{2}{3}}N^{\frac{4}{3}}. \end{aligned} \tag{83}$$

**case 3.** $1 \le N \le T^{\frac{1}{16}} \quad : \quad T^{\frac{1}{4}} \le M \le C_8T^{\frac{1}{2}} \quad : \quad (N \le M).$

First of all we notice that $N \le M^{\frac{1}{4}}$. Let

$$\mathcal{M}_1 = \max\{M^2V^{-4}(\log T)^3, TM^2V^{-12}(\log T)^9(\log\log T)^2, T^{\epsilon}, T^{1+\epsilon}V^{-8}(\log T)^6\}.$$

(i). Suppose that $\mathcal{M}_1 = M^2V^{-4}(\log T)^3$. Then, from (74) with $k = 2$, (79) and (80), we get

$$\begin{aligned} N^*(U,V,W) &\ll \min\{M^2V^{-4}(\log T)^3, TV^{-4}(\log T)^4, \\ &\quad T^2U^{-12}(\log Y)^{18}(\log\log Y)^{12}, T^2V^{-12}(\log Y)^{18}(\log\log Y)^{12}\} \\ &\ll \{M^2V^{-4}(\log T)^3\}^{\frac{1}{4}}\{TV^{-4}(\log T)^4\}^{\frac{1}{2}} \\ &\quad . \ \{T^2U^{-12}(\log Y)^{18}(\log\log Y)^{12}\}^{\frac{1}{6}}\{T^2V^{-12}(\log Y)^{18}(\log\log Y)^{12}\}^{\frac{1}{12}} \\ &\ll TM^{\frac{1}{2}}U^{-2}V^{-4}(\log Y)^{7.25}(\log\log Y)^3. \end{aligned} \tag{84}$$

(ii). Suppose that $\mathcal{M}_1 = TM^2V^{-12}(\log T)^9(\log\log T)^2$. Then, from (73) with $k = 2$ and (74) with $k = 2$, we see that

$$\begin{aligned} N^*(U,V,W) &\ll \min\{TM^2V^{-12}(\log T)^9(\log\log T)^2, TU^{-4}(\log T)^4 \\ &\quad , \ TV^{-4}(\log T)^4\} \\ &\ll \{TM^2V^{-12}(\log T)^9(\log\log T)^2\}^{\frac{1}{4}}\{TU^{-4}(\log T)^4\}^{\frac{1}{2}} \\ &\quad . \ \{TV^{-4}(\log T)^4\}^{\frac{1}{4}} \\ &\ll TM^{\frac{1}{2}}U^{-2}V^{-4}(\log Y)^{5.25}(\log\log Y)^{\frac{1}{2}}. \end{aligned} \tag{85}$$

(iii). Suppose that $\mathcal{M}_1 = T^{\epsilon}$. Then, from (79) and (80), on using (15), we get

$$\begin{aligned} N^*(U,V,W) &\ll \{T^{\epsilon}\}^{\frac{1}{2}}\{T^2V^{-12}(\log Y)^{18}(\log\log Y)^{12}\}^{\frac{1}{3}} \\ &\quad .\ \{T^2U^{-12}(\log Y)^{18}(\log\log Y)^{12}\}^{\frac{1}{6}} \\ &\ll TM^{\frac{1}{2}}U^{-2}V^{-4} \end{aligned} \tag{86}$$

(iv). Suppose that $\mathcal{M}_1 = T^{1+\epsilon}V^{-8}(\log T)^6$. Then, from (73) with $k = 2$ and using (15), we get

$$\begin{aligned} N^*(U,V,W) &\ll \{T^{1+\epsilon}V^{-8}(\log T)^6\}^{\frac{1}{2}}\{TU^{-4}(\log T)^4\}^{\frac{1}{2}} \\ &\ll TM^{\frac{1}{2}}U^{-2}V^{-4} \end{aligned} \tag{87}$$

From (84), (85), (86) and (87), we conclude that in this case

$$\begin{aligned} N^*(U,V,W) &\ll TM^{\frac{1}{2}}U^{-2}V^{-4}(\log Y)^{7.25}(\log\log Y)^3 \\ &\ll TU^{-2}V^{-4}W^{-4}M^{\frac{1}{2}}.W^4(\log Y)^{7.25}(\log\log Y)^3 \\ &\ll TU^{-2}V^{-4}W^{-4}M^{\frac{2}{3}}N^{\frac{4}{3}}(\log Y)^{7.25}(\log\log Y)^3 \end{aligned} \tag{88}$$

since $W^4 \ll N^2$ and $N \leq M^{\frac{1}{4}}$.

**case 4.** $T^{\frac{1}{16}} \leq N \leq T^{\frac{1}{4}}$ : $T^{\frac{1}{4}} \leq M \leq C_9T^{\frac{1}{2}}$ : $(N \leq M)$.

(i). Suppose that $\mathcal{M}_1 = M^2V^{-4}(\log T)^3$. Then, clearly $M^2V^{-4}(\log T)^3 \geq TM^2V^{-12}(\log T)^9(\log\log T)^2$ holds, which leads to

$$V \geq T^{\frac{1}{8}}(\log T)^{\frac{3}{4}}(\log\log T)^{\frac{1}{4}}. \tag{89}$$

Now, from (75) with $k = 6$, (79) and (80), we obtain

$$\begin{aligned} N^*(U,V,W) &\ll \{M^2V^{-4}(\log T)^3\}^{\frac{1}{3}}\{\max\left(T, N^6\right)W^{-12}(\log T)^{36}\}^{\frac{1}{4}} \\ &\quad .\ \{T^2V^{-12}(\log Y)^{18}(\log\log Y)^{12}\}^{\frac{1}{4}}\{T^2U^{-12}(\log Y)^{18}(\log\log Y)^{12}\}^{\frac{1}{6}} \\ &\ll M^{\frac{2}{3}}U^{-2}V^{-4}W^{-4}V^{-\frac{1}{3}}W(\log T)^{10}(\log Y)^{7.5} \\ &\quad .\ (\log\log Y)^5T^{\frac{5}{6}}\left(\max\left(T^{\frac{1}{4}}, N^{\frac{3}{2}}\right)\right) \\ &\ll M^{\frac{2}{3}}U^{-2}V^{-4}W^{-4}(\log Y)^{17.25}(\log\log Y)^5 \\ &\quad .\ \left(\max\left(T^{\frac{25}{24}}N^{\frac{1}{2}}, T^{\frac{19}{24}}N^2\right)\right) \\ &\ll TM^{\frac{2}{3}}U^{-2}V^{-4}W^{-4}(\log Y)^{17.25}(\log\log Y)^5 \end{aligned} \tag{90}$$

because of (89), $W \ll N^{\frac{1}{2}}$ and since for $T^{\frac{1}{16}} \leq N \leq T^{\frac{1}{4}}$, clearly we have

$$T^{\frac{25}{24}}N^{\frac{1}{2}} \leq TN^{\frac{4}{3}} \text{ and } T^{\frac{19}{24}}N^2 \leq TN^{\frac{4}{3}}.$$

(ii). Suppose that $\mathcal{M}_1 = TM^2V^{-12}(\log T)^9(\log\log T)^2$. Then, from (75) with

$k = 8$ and (73) with $k = 2$, we obtain

$$\begin{aligned} N^*(U,V,W) &\ll \{TM^2V^{-12}(\log T)^9(\log\log T)^2\}^{\frac{1}{3}}\{\max\left(T, N^8\right)W^{-16}(\log T)^{64}\}^{\frac{1}{6}} \\ &\quad .\ \{TU^{-4}(\log T)^4\}^{\frac{1}{2}} \\ &\ll M^{\frac{2}{3}}U^{-2}V^{-4}W^{-\frac{8}{3}}T^{\frac{5}{6}}(\log T)^{15.67}(\log\log T)^{\frac{2}{3}} \\ &\quad .\ \{\max\left(T^{\frac{1}{6}}, N^{\frac{4}{3}}\right)\} \\ &\ll M^{\frac{2}{3}}U^{-2}V^{-4}W^{-4}(\log T)^{15.67}(\log\log T)^{\frac{2}{3}} \\ &\quad .\ \{\max\left(TN^{\frac{2}{3}}, T^{\frac{5}{6}}N^2\right)\} \\ &\ll TM^{\frac{2}{3}}U^{-2}V^{-4}W^{-4}(\log T)^{15.67}(\log\log T)^{\frac{2}{3}} \end{aligned} \tag{91}$$

since $W^{\frac{4}{3}} \ll N^{\frac{2}{3}}$ and $TN^{\frac{2}{3}} \ll TN^{\frac{4}{3}}, T^{\frac{5}{6}}N^2 \ll TN^{\frac{4}{3}}$ for $T^{\frac{1}{16}} \le N \le T^{\frac{1}{4}}$.

(iii). Suppose that $\mathcal{M}_1 = T^{\epsilon}$. Then, from (79) and (80), we get

$$\begin{aligned} N^*(U,V,W) &\ll \{T^2U^{-12}(\log Y)^{18}(\log\log Y)^{12}\}^{\frac{1}{6}}\{T^2V^{-12}(\log Y)^{18}(\log\log Y)^{12}\}^{\frac{1}{4}} \\ &\quad .\ \{T^{\epsilon}\}^{\frac{7}{12}} \\ &\ll T^{\frac{5}{6}+2\epsilon}U^{-2}V^{-3} \\ &\ll TU^{-2}V^{-4}W^{-4}.T^{2\epsilon-\frac{1}{6}}VW^4 \\ &\ll TU^{-2}V^{-4}W^{-4}.T^{2\epsilon-\frac{1}{6}}M^{\frac{1}{2}}N^2 \\ &\ll TM^{\frac{2}{3}}N^{\frac{4}{3}}U^{-2}V^{-4}W^{-4}.\frac{T^{2\epsilon-\frac{1}{6}}N^{\frac{2}{3}}}{M^{\frac{1}{6}}} \\ &\ll TM^{\frac{2}{3}}N^{\frac{4}{3}}U^{-2}V^{-4}W^{-4} \end{aligned} \tag{92}$$

since $V \ll M^{\frac{1}{2}}, W \ll N^{\frac{1}{2}}, N \le T^{\frac{1}{4}}, M \ge T^{\frac{1}{4}}$.

(iv). Suppose that $\mathcal{M}_1 = T^{1+\epsilon}V^{-8}(\log T)^6$. Then, from (75) with $k = 4$ and (79), we obtain

$$\begin{aligned} N^*(U,V,W) &\ll \{T^{1+\epsilon}V^{-8}(\log T)^6\}^{\frac{1}{2}}\{T^2U^{-12}(\log Y)^{18}(\log\log Y)^{12}\}^{\frac{1}{6}} \\ &\quad .\ \{TW^{-8}(\log T)^{16}\}^{\frac{1}{3}} \\ &\ll T.T^{\frac{1}{6}+\epsilon}U^{-2}V^{-4}W^{-4}.W^{\frac{4}{3}}(\log Y)^{11\frac{1}{3}}(\log\log Y)^2 \\ &\ll T.M^{\frac{2}{3}}U^{-2}V^{-4}W^{-4}(\log Y)^{11\frac{1}{3}}(\log\log Y)^2 \\ &\quad .\ T^{\epsilon}.W^{\frac{4}{3}} \\ &\ll T.M^{\frac{2}{3}}N^{\frac{4}{3}}U^{-2}V^{-4}W^{-4} \end{aligned} \tag{93}$$

since $W \ll N^{\frac{1}{2}}$ and $N \ge T^{\frac{1}{16}}$. From (90), (91), (92) and (93), in this case we conclude that

$$N^*(U,V,W) \ll TM^{\frac{2}{3}}N^{\frac{4}{3}}U^{-2}V^{-4}W^{-4}(\log Y)^{17.25}(\log\log Y)^5. \tag{94}$$

Now, from (82), (83), (88), (94), (57) and (49) the Theorem 1.4 follows.

## References


[1] R. Balasubramanian, *An improvement of a theorem of Titchmarsh on the mean square of* $|\zeta(\frac{1}{2}+it)|$, Proc. London Math. Soc., **36** (1978), 540-576.

[2] R. Balasubramanian and K. Ramachandra , *Some problems of analytic number theory*-III, Hardy-Ramanujan J., **4** (1981) , 13-40.

[3] R. Balasubramanian and K. Ramachandra , *A lemma in complex function theory*-I, Hardy-Ramanujan J., **12** (1989) , 1-5.

[4] R. Balasubramanian and K. Ramachandra , *An alternative approach to a theorem of Tom Meurman*, Acta Arith., **55** (1990), 351-364.
[5] J.B. Conrey and A. Ghosh , *A conjecture for the sixth power moment of the Riemann zeta-function*, Internat. Math. Res. Notices (1988), 775-780.
[6] J.B. Conrey and S.M. Gonek , *High moments of the Riemann zeta-function*, Duke Math. J ., **107** (2001), 577-604.
[7] R.M. Gabriel, *Some results concerning the integrals of moduli of regular functions along certain curves*, J. London Math. soc., **2** (1927), 112-117.
[8] D.R. Heath-Brown, *The twelfth power moment of the Riemann zeta-function*, Quart.J.Math Oxford Ser. (2), **29** (1978), 443-462.
[9] D.R. Heath-Brown, *Mean values of the zeta-function and divisor problems*, Recent progress in analytic number theory, (Academic Press, london, 1981), 115-119.
[10] D.R. Heath-Brown, *The number of Abelian groups of order at most $x$*, Astérisque., **198-199-200** (1991), 153-163.
[11] A. Ivić, *The number of finite non-isomorphic Abelian groups in mean-square*, Hardy-Ramanujan J., **9** (1986), 17-23.
[12] A. Ivić and Y. Motohashi, *The mean square of the error term for the fourth power moment of the zeta-function*, Proc. london Math. Soc., (3) **69** (1994), 309-329.
[13] A. Ivić and Y. Motohashi, *On the fourth power moment of the Riemann zeta-function*, J. Number Theory., **51** (1995), 16-45.
[14] A. Ivić and K. Matsumoto, *On the error term in the mean square formula for the Riemann zeta-function in the critical strip*, Monatsh. Math., **121** (1996), 213-229.
[15] M. Jutila and Y. Motohashi, *Mean value estimates for exponential sums and L-functions : a spectral - theoretic approach*, J. reine Angew. Math., **459** (1993), 61-87.
[16] G. Kolesnik, *On the number of Abelian groups of a given order*, J. Reine Angew. Math., **329** (1981), 164-175.
[17] K. Matsumoto, *Recent developements in the mean square theory of the Riemann zeta and other zeta-functions*, Number Theory Trends Math., Birkhäuser, Basel (2000), 241-286.
[18] H.L. Montgomery and R.C. Vaughan, *Hilbert's inequality*, J. London Math. Soc., (2) **8** (1974), 73-82.
[19] H.L. Montgomery, Topics in multiplicative number theory, Lecture Notes in Math. **227** (Springer, Berlin, 1971).
[20] Y. Motohashi, *The mean square of Dedekind zeta-functions of quadratic number fields*, Sieve methods, exponential sums and their applications in number theory (Cardiff, 1995), 309-324, London Math soc., lecture note Ser. **237**, Cambridge University Press, Cambridge (1997).
[21] K. Ramachandra, *Some remarks on a theorem of Montgomery and Vaughan*, J. Number Theory., **11** (1980), 465-471.
[22] K. Ramachandra, *A large value theorem for $\zeta(s)$*, Hardy-Ramanujan J., **18** (1995), 1-9.
[23] O. Robert and P. Sargos, *Three dimensional exponential sums with monomials*, J. Reine Angew. Math., **591** (2006), 1-20.
[24] E.C. Titchmarsh, The Theory of the Riemann zeta function, 2nd edition (Revised by D.R. Heath-Brown), (Oxford, 1986).

*The Special Issue of The Proceedings of Telangana Academy of Sciences*
Vol. 01, No. 01, 2020, 202–213

# *An Introduction to Ramsey's Theorem*

**Amitabha Tripathi**[a*]

[a]*Department of Mathematics, Indian Institute of Technology, Hauz Khas, New Delhi – 110016*

**Abstract:** Ramsey's theorem is an integral part of results of the type that may loosely be classified as those that satisfy the property that if a large enough system is partitioned arbitrarily into finitely many subsystems, at least one subsystem has that particular property. Although initially stated as a result in mathematical Logic, Ramsey's theorem is now considered one of the cornerstones of Combinatorics.

**Keywords:** colouring; Ramsey number; graph Ramsey number; arrows into; non-complete Ramsey Theory

**AMS Subject Classifications**: 05D10; 05C55

Frank Plumpton Ramsey
(1903 – 1930)

Frank Ramsey was a British mathematician who, in addition to Mathematics, made significant contributions in Philosophy and Economics at an early age before his death at the age of 26. Frank was born on 22 February 1903 in Cambridge where his father Arthur, also a mathematician, was President of Magdalene College. He was the eldest of two brothers and two sisters, and his brother Michael, the only one of the four siblings who was to remain Christian, later became Archbishop of

*Email: atripath@maths.iitd.ac.in

Canterbury. He entered Winchester College in 1915 and later returned to Cambridge to study Mathematics at Trinity College. With support from the economist John Maynard Keynes he became a Fellow of King's College, Cambridge in 1924, being the second person ever to be elected without having previously studied at King's College. In 1926 he became a University Lecturer in Mathematics and later a Director of Studies in Mathematics at King's College.

In 1927 Ramsey published the influential article *Facts and Propositions*, in which he proposed what is sometimes described as a *Redundancy Theory of Truth*. His other philosophical works include *Universals* (1925), *Universals of Law and of Fact* (1928), *Knowledge* (1929), *Theories* (1929), and *General Propositions and Causality* (1929). The philosopher Ludwig Wittgenstein, whose work *Tractatus Logico-Philosophicus* he helped translate into English, mentions him in the introduction to his *Philosophical Investigations* as an influence.

Ramsey's three papers in Economics were on *Subjective Probability and Utility* (1926), *Optimal Taxation* (1927) and *Optimal One–sector Economic Growth* (1928). The economist Paul Samuelson described them in 1970 as "*three great legacies – legacies that were for the most part mere by-products of his major interest in the foundations of mathematics and knowledge.*"

One of the theorems proved by Ramsey in his 1930 paper "*On a problem of formal logic*" now bears his name. While this theorem is the work Ramsey is probably best remembered for, he only proved it in passing, as a minor lemma along the way to his true goal in the paper – solving a special case of the *Decision Problem for First–order Logic*, namely the *Decidability of Bernays-Scönfinkel–Ramsey Class of First–order Logic*. A great amount of later work in Mathematics was fruitfully developed out of the ostensibly minor lemma, which turned out to be an important early result in Combinatorics, supporting the idea that within some sufficiently large systems, however disordered, there must be some order.

Easy going, simple and modest, Ramsey had many interests besides his scientific work. He was immensely widely read in English literature; he enjoyed classics though he was on the verge of plunging into being a mathematical specialist. He was very interested in politics, and well–informed; he had got a political concern and a sort of left-wing caring–for–the–underdog kind of outlook about politics.

Suffering from chronic liver problems, Ramsey contracted jaundice after an abdominal operation and died on 19 January 1930 at Guy's Hospital in London a month before turning 27.

The Decision Analysis Society annually awards the *Frank P. Ramsey Medal* to recognise substantial contributions to Decision Theory and its application to important classes of real decision problems.

## 1. Ramsey's Theorem

The newest of the three major results on Ramsey–type theorems – the theorem of Ramsey in Combinatorics that bears his name – was enunciated as a result in Logic. Ramsey's Theorem may be considered as a refinement of the *Pigeonhole Principle*, but one in which we are not only guaranteed a certain number of elements in a particular class but also guaranteed that these elements share a given property. The following problem, considered folklore, amply describes this situation.

PROBLEM 1.1 **(The Party Problem)**
*At a party consisting of six persons, there must be three mutual acquaintances or three mutual strangers.*

A simple application of the *Pigeonhole Principle* provides a proof of this problem. Consider the complete graph $\mathcal{K}_6$, with vertices $0, \ldots, 5$, each representing a party–goer, and colour the edges between acquaintances blue and those between strangers red. By a triangle we mean those three–sided figures with all vertices $0, \ldots, 5$. Thus, a triangle with all sides blue will depict the situation where the three vertices represent persons who are mutual acquaintances, and a triangle with all sides red will depict the situation where the three vertices represent persons who are mutual strangers. A proof must consist of showing that no matter how we colour each edge in one of two colurs blue, red, one of these two monochromatic triangles must arise. By the Pigeonhole Principle, at least three of the edges 01, 02, 03, 04, 05 must be of one colour, say blue. By renumbering, if necessary, suppose the edges 01, 02, and 03 are colored blue. If any one of the edges 12, 23, 13 is coloured blue, then the triangle with vertices 0 and the two endpoints of the blue edge form a blue triangle. If none of the edges 12, 23, 13 is coloured blue, then the triangle with vertices 1, 2, 3 form a red triangle. If three of the edges 01, 02, 03, 04, 05 are coloured red, the same argument with the roles of blue and red interchanged again results in a monochromatic triangle.

The mathematical statement captured by this statement of this problem is $\mathcal{R}(3,3) \leq 6$. The two 3's represent the two relationships (acquaintances, strangers) or the two colour classes (blue, red), whereas the 6 represents that fact that six people suffice to capture one or the other situation. More generally, given positive integers $m, n$, the statement

$$\mathcal{R}(m,n) = N$$

is the combination of the following two statements:

- If all the edges of $\mathcal{K}_N$ are coloured either blue or red in any manner, then there must exist $m$ vertices such that all the edges formed by the graph $\mathcal{K}_m$ on these vertices are coloured blue, or there must exist $n$ vertices such that all the edges formed by the graph $\mathcal{K}_n$ on these vertices are coloured red, and
- There is a colouring of the edges of $\mathcal{K}_{N-1}$ in blue and red such that neither of the two situations listed above arises.

The first of these situations is captured by the statement $\mathcal{R}(m,n) \leq N$, and the second by $\mathcal{R}(m,n) > N-1$. Therefore, together these imply $\mathcal{R}(m,n) = N$. Note that the roles of blue and red are interchangeable. Hence $\mathcal{R}(m,n) = \mathcal{R}(n,m)$, and it is customary to use $\mathcal{R}(m,n)$ with $m \geq n \geq 1$. It is trivial that $\mathcal{R}(m,1) = 1$. To see why $\mathcal{R}(m,2) = m$, colouring all edges of $\mathcal{K}_{m-1}$ blue simultaneously avoids a blue $\mathcal{K}_m$ and a red $\mathcal{K}_2$, and hence implies $\mathcal{R}(m,2) > m-1$. On the other hand, the only way to avoid a red $\mathcal{K}_2$ in a blue–red edge colouring of $\mathcal{K}_m$ is by colouring all edges blue, in which case there is a blue $\mathcal{K}_m$. Hence there is always either a blue $\mathcal{K}_m$ or a red $\mathcal{K}_2$ in every blue–red edge colouring of $\mathcal{K}_m$, implying that $\mathcal{R}(m,2) \leq m$, so that $\mathcal{R}(m,2) = m$. The nontrivial values of $\mathcal{R}(m,n)$ therefore start with $m \geq n \geq 3$, and $\mathcal{R}(3,3)$ is the first of these. The choice of the Party Problem as an initial example mentioned at the start of this section is therefore quite natural.

The proof of the Party Problem implies $\mathcal{R}(3,3) \leq 6$. In fact, it is true that $\mathcal{R}(3,3) = 6$. If we colour the outer five edges of $\mathcal{K}_5$ blue and the inner diagonals

red, we find no triangle of the same colour. This solitary example of 2-colouring the edges of $\mathcal{K}_5$ shows that $\mathcal{R}(3,3) > 5$, and hence also that $\mathcal{R}(3,3) = 6$. The numbers $\mathcal{R}(m,n)$ are the simplest examples of Ramsey numbers. Their existence is guaranteed by the following theorem.

THEOREM 1.2 *The Ramsey numbers $\mathcal{R}(m,n)$ satisfy the recurrence*

$$\mathcal{R}(m,n) \leq \mathcal{R}(m-1,n) + \mathcal{R}(m,n-1)$$

*for $m,n \geq 2$. Moreover, if both $\mathcal{R}(m-1,n)$ and $\mathcal{R}(m,n-1)$ are even, we have*

$$\mathcal{R}(m,n) \leq \mathcal{R}(m-1,n) + \mathcal{R}(m,n-1) - 1.$$

*The Ramsey numbers $\mathcal{R}(m,n)$ satisfy the bounds*

$$(m-1)(n-1)+1 \leq \mathcal{R}(m,n) \leq \binom{m+n-2}{m-1} = \binom{m+n-2}{n-1}$$

*for $m,n \geq 2$.*

**Proof.** Let us write $N = \mathcal{R}(m-1,n) + \mathcal{R}(m,n-1)$ for convenience. To prove the general upper bound, we must show that in any blue–red colouring of the edges of $\mathcal{K}_N$, there must exist either a blue $\mathcal{K}_m$ or a red $\mathcal{K}_n$.

Let $V$ and $E$ denote the set of vertices and edges, respectively, of $\mathcal{K}_N$, and consider any blue–red colouring of the edges of $\mathcal{K}_N$. Choose any $v \in V$, and partition the set $V \setminus \{v\}$ into sets $B = \{x \in V : xv \in E \text{ and is coloured blue}\}$ and $R = \{x \in V : xv \in E \text{ and is coloured red}\}$. Then $|B| + |R| = N - 1 = \mathcal{R}(m-1,n) + \mathcal{R}(m,n-1) - 1$, so that $|B| < \mathcal{R}(m-1,n)$ and $|R| < \mathcal{R}(m,n-1)$ is not simultaneously possible. Therefore at least one of $|B| \geq \mathcal{R}(m-1,n)$ and $|R| \geq \mathcal{R}(m,n-1)$ must hold.

Consider the case $|B| \geq \mathcal{R}(m-1,n)$; the parallel case $|R| \geq \mathcal{R}(m,n-1)$ can be argued by replacing the role of blue with red. Since the subgraph of $\mathcal{K}_B$ of $\mathcal{K}_N$ has at least $\mathcal{R}(m-1,n)$ vertices, $\mathcal{K}_B$ must contain either a blue $\mathcal{K}_{m-1}$ or a red $\mathcal{K}_n$ by definition of Ramsey number $\mathcal{R}(m-1,n)$. If the first of these cases hold, then the vertex $v$ together with those of $\mathcal{K}_{m-1}$ forms a blue $\mathcal{K}_m$ by construction of $B$. Thus, in any case, $\mathcal{K}_N$ must contain either a blue $\mathcal{K}_m$ or a red $\mathcal{K}_n$. This completes the assertion that $\mathcal{R}(m,n) \leq \mathcal{R}(m-1,n) + \mathcal{R}(m,n-1)$ for $m,n \geq 2$.

To prove the stronger upper bound in the special case where both $\mathcal{R}(m-1,n)$ and $\mathcal{R}(m,n-1)$ are even, consider any blue–red colouring of the edges of $\mathcal{K}_{N-1}$ and choose a vertex $v \in V$ of even degree; this choice is made possible because the sum of degrees of all vertices in a graph equals twice the number of edges in the graph and $N-1$ is odd. With $B$ and $R$ as defined earlier, we now have $|B|+|R| = N-2$. If $|B| \geq \mathcal{R}(m-1,n)$, the earlier argument implies $\mathcal{K}_{N-1}$ must contain a blue $\mathcal{K}_m$. Otherwise, $|R| \geq \mathcal{R}(m,n-1)$ since $\deg v$ is even, and again the earlier argument implies $\mathcal{K}_{N-1}$ must contain a red $\mathcal{K}_n$.

The proof of the upper bound the Ramsey numbers $\mathcal{R}(m,n)$ may be accomplished by induction on $k = m+n$. We may easily verify the bound for all cases where $k \leq 5$, since $\mathcal{R}(m,1) = 1$ and $\mathcal{R}(m,2) = 2$. For the same reason we may also assume $m,n \geq 3$. Assume the bound holds for all pairs of positive integers $m,n$ with $m+n < k$ and $m,n \geq 3$, and consider $\mathcal{R}(m,n)$ where $m+n = k$, $m,n \geq 3$.

By inductive hypothesis

$$\mathcal{R}(m-1,n) \le \binom{m+n-3}{m-2} \text{ and } \mathcal{R}(m,n-1) \le \binom{m+n-3}{m-1}.$$

Applying the recurrence satisfied by the Ramsey numbers $\mathcal{R}(m,n)$ yields

$$\mathcal{R}(m,n) \le \mathcal{R}(m-1,n)+\mathcal{R}(m,n-1) \le \binom{m+n-3}{m-2}+\binom{m+n-3}{m-1} = \binom{m+n-2}{m-1}.$$

This completes the proof of the upper bound for $\mathcal{R}(m,n)$ by induction.

To prove the lower bound, we need to 2-colour the edges of $\mathcal{K}_{(m-1)(n-1)}$ such that there is no blue $\mathcal{K}_{m-1}$ and there is no red $\mathcal{K}_{n-1}$. Place the $(m-1)(n-1)$ vertices of $\mathcal{K}_{(m-1)(n-1)}$ in a rectangular array in $m-1$ rows and $n-1$ columns. Colour any two vertices in the same row red, and in different rows blue. Then the red edges form $m-1$ copies of $\mathcal{K}_{n-1}$, and so there is no red $\mathcal{K}_n$. There are also no blue $\mathcal{K}_m$, for among any $m$ vertices two must be in the same row and must be joined by a red edge. Therefore the given blue-red colouring has neither a blue $\mathcal{K}_m$ nor a red $\mathcal{K}_n$. This completes the proof of the lower bound for $\mathcal{R}(m,n)$ by an example. ■

Theorem 1.2 gives $\mathcal{R}(m,3) \le \frac{1}{2}(m^2+m)$ for $m \ge 3$. This upper bound can be improved quite easily to $\mathcal{R}(m,3) \le \frac{1}{2}(m^2+3)$ for $m \ge 3$ by induction. However, actual rate of growth for the Ramsey numbers $\mathcal{R}(m,3)$ is $m^2/\log m$ for large $m$.

THEOREM 1.3 **([1, 18])** *There exist constants $c_1$ and $c_2$ such that*

$$c_1\frac{m^2}{\log m} \le \mathcal{R}(m,3) \le c_2\frac{m^2}{\log m}.$$

The lower bound is due to Kim [18]; the upper bound to Ajtai, Komlós and Szemerédi [1].

The diagonal Ramsey numbers $\mathcal{R}(n,n)$ have received considerable attention. The upper bound $\mathcal{R}(n,n)$ from Theorem 1.2 is $\binom{2n-2}{n-1}$; this is asymptotically $c4^n/\sqrt{n}$. For the lower bound, the following theorem, due to Erdős, is asymptotically sharp. This proof is significant also because probabilistic methods were introduced for the first time in Ramsey theory.

THEOREM 1.4 **([9])**

$$\mathcal{R}(n,n) > (e\sqrt{2})^{-1}n2^{n/2}\big(1+o(1)\big).$$

**Proof.** We sketch a proof of the weaker lower bound $\mathcal{R}(n,n) > 2^{(n-2)/2}$.

Let $N$ be a positive integer, which is to be specified later and which will serve as a lower bound. Let the vertices of $\mathcal{K}_N$ be labelled $1,2,3,\ldots,N$, and randomly colour all edges of $\mathcal{K}_N$ either red or blue, independently and with equal probability $1/2$. Consider any $n$-subset $X$ of $[N]$. There are $\binom{n}{2}$ edges in $X$; the probability that all are coloured either red or blue is $2^{-\binom{n}{2}}$. Therefore the probability that all edges in $X$ have the same colour is $2\cdot 2^{-\binom{n}{2}}$. Since there are $\binom{N}{n}$ ways of choosing $n$-subsets of $[N]$, the total probability that there exists a monochromatic $n$-subset of $[N]$ is $\binom{N}{n}\cdot 2^{1-\binom{n}{2}}$.

For fixed $n$, if we choose $N$ such that this probability $\binom{N}{n} \cdot 2^{1-\binom{n}{2}}$ is less than 1, then we must have a colouring which contains no monochromatic $n$-set. The weak estimates

$$\binom{N}{n} < N^n \text{ and } 1 - \binom{n}{2} < -\frac{n(n-2)}{2} \tag{1}$$

yield

$$\binom{N}{n} \cdot 2^{1-\binom{n}{2}} < N^n \cdot 2^{-n(n-2)/2} = \left(N \cdot 2^{-(n-2)/2}\right)^n.$$

Thus, the probability $\binom{N}{n} \cdot 2^{1-\binom{n}{2}}$ is less than 1 if $N = 2^{(n-2)/2}$.

The bound in the theorem is a consequence of applying stronger bounds in eqn. (1) via Stirling's formula:

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(1 + \mathrm{O}\left(\frac{1}{n}\right)\right).$$

■

Putting together the two bounds for $\mathcal{R}(n,n)$ leads to

$$\sqrt{2} \leq \liminf \sqrt[n]{\mathcal{R}(n,n)} \leq \limsup \sqrt[n]{\mathcal{R}(n,n)} \leq 4. \tag{2}$$

PROBLEM 1.5 **(Open Problem)**

- *Does* $\lim \sqrt[n]{\mathcal{R}(n,n)}$ *exist?*
- *Determine* $\lim \sqrt[n]{\mathcal{R}(n,n)}$*, if it exists.*

| $n$ \ $m$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 6 | 9 | 14 | 18 | 23 | 28 | 36 | 40<br>42 | 47<br>50 | 53<br>59 | 60<br>68 | 67<br>77 | 74<br>87 |
| 4 | | 18 | 25 | 36<br>41 | 49<br>61 | 59<br>84 | 73<br>115 | 92<br>149 | 102<br>191 | 128<br>238 | 138<br>291 | 147<br>349 | 155<br>417 |
| 5 | | | 43<br>48 | 58<br>87 | 80<br>143 | 101<br>216 | 133<br>316 | 149<br>442 | 183<br>633 | 203<br>848 | 233<br>1138 | 267<br>1461 | 269<br>1878 |
| 6 | | | | 102<br>165 | 115<br>298 | 134<br>495 | 183<br>780 | 204<br>1171 | 256<br>1804 | 294<br>2566 | 347<br>3703 | 5033 | 401<br>6911 |
| 7 | | | | | 205<br>540 | 217<br>1031 | 252<br>1713 | 292<br>2826 | 405<br>4553 | 417<br>6954 | 511<br>10578 | 15263 | 22112 |
| 8 | | | | | | 282<br>1870 | 329<br>3583 | 343<br>6090 | 10630 | 16944 | 817<br>27485 | 41525 | 865<br>63609 |
| 9 | | | | | | | 565<br>6588 | 581<br>12677 | 22325 | 38832 | 64864 | | |
| 10 | | | | | | | | 798<br>23556 | 45881 | 81123 | | | 1265 |

**Table of the Ramsey numbers $\mathcal{R}(m,n)$: Exact Values & Bounds**

With $r$ parameters, $r > 2$, the definition of the Ramsey numbers $\mathcal{R}(n_1, \ldots, n_r)$ have a natural extension. The Ramsey number $\mathcal{R}(n_1, \ldots, n_r)$ denotes the least positive integer $N$ for which the following property holds: if we colour each edge

in $\mathcal{K}_N$ with one of $r$ fixed colours randomly, there must exist a $\mathcal{K}_{n_1}$ in colour 1, or a $\mathcal{K}_{n_2}$ in colour 2, or a $\mathcal{K}_{n_3}$ in colour 3, and so on. The case $r = 2$ is obviously a special case. The following generalization of Theorem 1.2 settles the question of existence of the Ramsey numbers $\mathcal{R}(n_1, \ldots, n_r)$.

THEOREM 1.6 *The Ramsey numbers $\mathcal{R}(n_1, \ldots, n_r)$ satisfy the recurrence*

$$\mathcal{R}(n_1, \ldots, n_r) \leq \sum_{i=1}^{r} \mathcal{R}(n_1, \ldots, n_{i-1}, n_i - 1, n_{i+1}, \ldots, n_r)$$

*for $n_1, \ldots, n_r \geq 2$.*
*The Ramsey numbers $\mathcal{R}(n_1, \ldots, n_r)$ have the upper bound*

$$\mathcal{R}(n_1, \ldots, n_r) \leq \binom{n_1 + \cdots + n_r - r}{n_1 - 1, \ldots, n_r - 1}$$

*valid for $n_1, \ldots, n_r \geq 2$.*

The general version of Ramsey's theorem is considerably more complicated. Given positive integers $k$ and $r$, and sufficiently large $N$, each $k$–subset of $[N]$ is assigned one of $r$ colours. Ramsey's theorem assures the existence of such $N$. More precisely, if $k$ is any positive integer, $\ell_1, \ldots, \ell_r$ satisfy $\ell_i \geq k$ for each $i$, and we $r$–colour all $k$–subsets of $[N]$, for some sufficently large $N$, then all $k$–subsets of some $\ell_i$ numbers chosen from $[N]$ must necessarily be coloured $i$.

THEOREM 1.7 **(Ramsey's Theorem)**
*For positive integers $k, \ell_1, \ldots, \ell_r$, with each $\ell_i \geq k$, there exists a least positive integer $N = \mathcal{R}_k(\ell_1, \ldots, \ell_r)$ such that, for every $r$–colouring of all $k$–subsets of $[N]$, there exists a monochromatic set of size $\ell_i$ for some $i \in [r]$.*

When $\ell_1 = \cdots = \ell_r = \ell$, we write $\mathcal{R}_k(\ell; r)$ for $\mathcal{R}_k(\ell_1, \ldots, \ell_r)$. If $k = 2$, we usually suppress the subscript and write $\mathcal{R}(\ell_1, \ldots, \ell_r)$ for $\mathcal{R}_2(\ell_1, \ldots, \ell_r)$. Proof of existence of the generalized Ramsey numbers $\mathcal{R}_k(\ell_1, \ldots, \ell_r)$ is considerably harder to prove; see [15] for instance.

## 2. Graph Ramsey Theory

Graph Ramsey theory involves graphs in Ramsey theory, as the name suggests, and graph Ramsey numbers have graphs as inputs instead of positive integers. More specifically, given any finite collection of graphs $G_1, \ldots, G_r$, $r \geq 2$, there exists $N$ such that every edge colouring of $\mathcal{K}_N$ in $r$ colours contains a copy of $G_1$ in colour 1, or a copy of $G_2$ in colour 2, or a copy of $G_3$ in colour 3, and so on. The existence of such an $N$ follows for the Ramsey number $\mathcal{R}(n_1, \ldots, n_r)$, where $n_i$ denotes the number of vertices in the graph $G_i$, $1 \leq i \leq r$. Recall that the Ramsey number corresponding to positive integers $n_1, \ldots, n_r$ involve positive integers $N$ for which every $r$-colouring of edges in $\mathcal{K}_N$ must contain a $\mathcal{K}_{n_i}$ in colour $i$ for at least one $i$. The graph Ramsey number $\mathcal{R}(G_1, \ldots, G_r)$ is the least positive integer $N$ for which the above mentioned property holds. Since each $G_i$ is contained in $\mathcal{K}_{n_i}$, the existence of graph Ramsey numbers follow from the corresponding Ramsey

numbers. In fact,

$$\mathcal{R}(G_1, \ldots, G_r) \leq \mathcal{R}(n_1, \ldots, n_r),$$

where $n_i$ denotes the order of $G_i$, $1 \leq i \leq r$.

Graph Ramsey theory has attracted a lot of interest, specially since the late 60's. As in the case with Ramsey numbers, most of research has centered around the case $r = 2$ because of expected simplicity in the argument in this case as opposed to the cases $r > 2$. Finding exact values of graph Ramsey numbers is an extremely challenging problems, even in the case $r = 2$. For instance, the statement

$$\mathcal{R}(G_1, G_2) = N$$

is the combination of the following two statements:

- If all the edges of $\mathcal{K}_N$ are coloured either blue or red in any manner, the graph formed by considering only the blue edges must contain $G_1$ as a subgraph, or the graph formed by considering only the red edges must contain $G_2$ as a subgraph, and
- There is a colouring of the edges of $\mathcal{K}_{N-1}$ in blue and red such that neither of the two situations listed above arises.

The first of these situations is captured by the statement $\mathcal{R}(G_1, G_2) \leq N$, and the second by $\mathcal{R}(G_1, G_2) > N - 1$. Therefore, together these imply $\mathcal{R}(G_1, G_2) = N$. Note that the roles of blue and red are interchangeable.

Some of the earliest results in graph Ramsey theory include determining $\mathcal{R}(P_m, P_n)$, $\mathcal{R}(C_m, C_n)$, $\mathcal{R}(T_m, \mathcal{K}_n)$, and $\mathcal{R}(\mathcal{K}_{1,n_1}, \ldots, \mathcal{K}_{1,n_r})$. Here $P_n, C_n, T_n$ denote path, cycle, tree, respectively, each of order $n$, and $\mathcal{K}_{1,n}$ denotes a complete bipartite graph with partite sets of orders 1 and $n$, and is called a star graph.

THEOREM 2.1 **([13])**
*For integers $m, n$, with $2 \leq n \leq m$,*

$$\mathcal{R}(P_m, P_n) = m + \lfloor \tfrac{n}{2} \rfloor - 1.$$

THEOREM 2.2 **([12, 26, 27])**
*For integers $m, n$, with $3 \leq n \leq m$,*

$$\mathcal{R}(C_m, C_n) = \begin{cases} 2m - 1 & \text{if } n \text{ is odd, } (m,n) \neq (3,3); \\ m + \frac{n}{2} - 1 & \text{if } m, n \text{ are even, } (m,n) \neq (4,4); \\ \max\{m + \frac{n}{2} - 1, 2n - 1\} & \text{if } m \text{ is odd and } n \text{ is even}; \\ 6 & \text{if } m = n \in \{3, 4\}. \end{cases}$$

THEOREM 2.3 **([6])**
*If $T_m$ is any tree of order $m$ and $n$ is a positive integer, then*

$$\mathcal{R}(T_m, \mathcal{K}_n) = (m - 1)(n - 1) + 1.$$

THEOREM 2.4 **([4])**
*Let $n_1, \ldots, n_k$ be positive integers, $e$ of which are even. Then*

$$\mathcal{R}(\mathcal{K}_{1,n_1}, \ldots, \mathcal{K}_{1,n_k}) = \begin{cases} N+1 & \text{if } e \text{ is even and positive,} \\ N+2 & \text{otherwise,} \end{cases}$$

*where $N = \sum_{i=1}^{k}(n_i - 1)$.*

We close this section with a sketch of the proof of Theorem 2.3.

**Proof of Theorem 2.3.**
To establish the lower bound $\mathcal{R}(T_m, \mathcal{K}_n) > (m-1)(n-1)$, we must exhibit a colouring of each of the edges of $\mathcal{K}_{(m-1)(n-1)}$ in red or blue for which there is no red $T_m$ and no blue $\mathcal{K}_n$. Place the $(m-1)(n-1)$ vertices in a $(m-1) \times (n-1)$ rectangular grid, and join any two vertices in the same row by a blue edge and any two vertices in different rows by a red edge. The subgraph with blue edges form $m-1$ copies of $\mathcal{K}_{n-1}$, thereby avoiding a blue $\mathcal{K}_n$. On the other hand, any $m$ vertices in the subgraph with red edges must contain at least two from the same row, by Pigeonhole Principle. But these two vertices must be joined by a blue edge, which is a contradiction to our assumption that we are in the red subgraph. Therefore we have exhibited a colouring of each of the edges of $\mathcal{K}_{(m-1)(n-1)}$ in red or blue for which there is no red $T_m$ and no blue $\mathcal{K}_n$, as desired.

To establish the upper bound $\mathcal{R}(T_m, \mathcal{K}_n) \leq (m-1)(n-1)+1$, we use the following result on trees:

> If $T$ is any tree with $k-1$ vertices and $G$ is any graph with minimum vertex degree $\delta(G) \geq k$, then $T$ is a subgraph of $G$.

Consider any colouring of the edges of $\mathcal{K}_{(m-1)(n-1)+1}$ in red or blue, and let $v$ be any vertex in this graph. The proof we present runs on inducting on $n$. The base case $n = 1$ is trivial. If $v$ has more than $(m-1)(n-2)$ neighbours along the blue edges, then there must exist a red $T_m$ or a blue $\mathcal{K}_{n-1}$ among these, by induction hypothesis. Together with the vertex $v$, the graph $G$ then must contain either a red $T_m$ or a blue $\mathcal{K}_n$.

Otherwise, every vertex must have at most $(m-1)(n-2)$ incident blue edges, and hence at least $m-1$ incident red edges. The quoted result on trees now shows the existence of a red $T_m$. This completes the sketch of the proof. ■

## 3. Noncomplete Ramsey Theory

Noncomplete Ramsey theory generalize both classical Ramsey theory and graph Ramsey theory. For any collection of graphs $G_1, \ldots, G_r$, we say that a graph $G$ "*arrows into*" $(G_1, \ldots, G_r)$, and write

$$G \to (G_1, \ldots, G_r), \tag{3}$$

provided any $r$–colouring of the edges of $G$ yields a monochromatic spanning subgraph each of whose edges is coloured $i$ and that contains a $G_i$, for some $i \in [r]$.

Otherwise stated,

$$G = F_1 \oplus \cdots \oplus F_r \Longrightarrow F_i \supseteq G_i \text{ for at least one } i \in [r].$$

The graphs $F_1, \ldots, F_r$ are spanning subgraphs of $G$, and are called "factors" of $G$. By the containment $F_i \supseteq G_i$ one means simply that $G_i$ is a subgraph of $F_i$, not necessarily a spanning subgraph. Colouring the edges of $G$ by one of $r$ available colours induces a factorization of $G$, each given by the spanning subgraph with edges of one colour. Conversely, each factorizaton of $G$ leads to a colouring of the edges of $G$, with one colour assigned to all edges of each factor. Thus there is a natural correspondence between factorization of $G$ and edge–colouring of $G$ and the two terms may be used interchangeably.

The *arrows notation* may also be used to state Ramsey's theorem 1.7 concisely. Given positive integers $\ell_1, \ldots, \ell_r, k$, with each $\ell_i \geq k$, the notation

$$N \to (\ell_1, \ldots, \ell_r)^k$$

stands for the statement of Theorem 1.7, and the least such $N$ for which this statement holds is $\mathcal{R}_k(\ell_1, \ldots, \ell_r)$.

The main problem of *non–complete Ramsey Theory* is to characterize graphs $G$ that *arrow into* a given collection of graphs $G_1, \ldots, G_r$.

For any collection of graphs $G_1, \ldots, G_r$, the smallest positive integer $n$ for which $\mathcal{K}_n \to (G_1, \ldots, G_r)$ is the *graph Ramsey number* of $G_1, \ldots, G_r$, and is denoted by $\mathcal{R}(G_1, \ldots, G_r)$. Being able to characterize $G$ resolves many problems invoving the given graphs $G_1, \ldots, G_r$. For instance, the graph Ramsey number $\mathcal{R}(G_1, \ldots, G_r)$, which is the least positive integer $n$ such that

$$\mathcal{K}_n \to (G_1, \ldots, G_r),$$

may be easily determined from the characterization of $G$ in eqn. (3). In particular, the case when each $G_i$ is also a complete graph $\mathcal{K}_{\ell_i}$, the corresponding graph Ramsey number $\mathcal{R}(\mathcal{K}_{\ell_1}, \ldots, \mathcal{K}_{\ell_r})$ coincides with the Ramsey number $\mathcal{R}(\ell_1, \ldots, \ell_r)$.

One of the first instances of a solution to the main problem of characterization of $G$ in eqn. (3) is when $G_1 = G_2 = \mathcal{K}_{1,n}$, due to Murty.

THEOREM 3.1 *Let $G$ be a connected graph and $n$ a positive integer. Then*

$$G \to (\mathcal{K}_{1,n}, \mathcal{K}_{1,n})$$

*if and only if*

(i) $\Delta(G) \geq 2n - 1$, *or*
(ii) $n$ *is even and* $G$ *is a* $(2n-2)$*-regular graph of odd order.*

The result of Theorem 3.1 has been generalized by Gupta, Thulasi Rangan & Tripathi [16] to $G_1 = \mathcal{K}_{1,n_1}, \ldots, G_k = \mathcal{K}_{1,n_k}$, where $n_1, \ldots, n_k$ are any $k$ positive integers, $k \geq 2$. The characterization of $G$ satisfying

$$G \to (\mathcal{K}_{1,n_1}, \ldots, \mathcal{K}_{1,n_k}) \tag{4}$$

is described by one of four cases, and these cases involve conditions on the graph or their regularization. A $k$–factor of a graph is a factor that is $k$–regular, and a

$\Delta(G)$–regularization of $G$ is a $\Delta(G)$-regular graph $G^\star$ of which $G$ is an induced subgraph.

THEOREM 3.2 **([16])**
*Let $G$ be a connected graph, let $n_1, \ldots, n_k$ be positive integers of which $e$ are even, and let $N = \sum_{i=1}^{k}(n_i - 1)$. Let $G^\star$ be the $\Delta$-regularization of $G$. Then*

$$G \to (\mathcal{K}_{1,n_1}, \ldots, \mathcal{K}_{1,n_k})$$

*if and only if*

(i) $\Delta(G) \geq N + 1$, *or*
(ii) *$G$ is $N$–regular, of odd order and $e$ is even and non-zero, or*
(iii) *$G$ is $N$–regular, of even order, at least one $n_i$ is even, and $G$ does not have an $n_i - 1$ factor for at least one even $n_i$, or*
(iv) *$G$ is not $N$–regular, $\Delta(G) = N$, and $G^\star \to (\mathcal{K}_{1,n_1}, \ldots, \mathcal{K}_{1,n_k})$.*

The proof of Theorem 3.2 involves several basic results that deal with characterizations of graphs that have a $k$-factor, such as the ones due to Tutte [30, 31] and Petersen [22], and with edge colourings, such as the one due to Vizing [32], and independently, to Gupta [17]. Even a sketch of a proof of this result is beyond the scope of this article, but we briefly indicate how Theorem 2.4 and Theorem 3.1 may be deduced from Theorem 3.2.

**Theorem 3.2 implies Theorem 2.4.**
Observe that $G = \mathcal{K}_{N+2}$ satisfies eqn. (4) by condition (i). To complete the proof, we need to show that $\mathcal{K}_{N+1}$ satisfies eqn. (4) if and only if $e$ even and non-zero.

If $e$ is even and non-zero, condition (ii) applies to $\mathcal{K}_{N+1}$. Conversely, suppose $\mathcal{K}_{N+1}$ satisfies eqn. (4). If $N$ is even, by condition (ii), $e$ is even and non-zero. If $N$ is odd, by condition (iii), $\mathcal{K}_{N+1}$ does not have an $(n_i - 1)$–factor for at least one even $n_i$, which contradicts the well known fact that $\mathcal{K}_{2n}$ is 1–factorable for each $n \geq 1$. ■

**Theorem 3.2 implies Theorem 3.1.**
When $k = 2$ and $n_1 = n_2 = n$, $N + 1 = 2(n - 1) + 1 = 2n - 1$, so that part (i) in Theorem 3.1 is a direct translation of part (i) in Theorem 3.2. Part (ii) in Theorem 3.2 reduces to $G$ being a $(2n - 2)$–regular and of odd order, with $n$ even.

Part (iii) in Theorem 3.2 reduces to $G$ being a $(2n-2)$–regular and of even order, with $n$ even, such that $G$ does not have a $(n - 1)$-factor, and part (iv) in Theorem 3.2 reduces to $G$ being not $(2n-2)$–regular, $\Delta(G) = 2n-2$, and $G^\star \to (\mathcal{K}_{1,n}, \mathcal{K}_{1,n})$. It can be shown that neither of these cases can occur. ■

## References


[1] M. Ajtai, J. Komlós and E. Szemerédi, *A note on Ramsey numbers*, J. Combin. Theory Ser. A 29 (1980), 354–360.
[2] N. Alon and J. H. Spencer, *The Probabilistic Method*, Second Ed., Wiley–Interscience Series in Discrete Mathematics and Optimization, New York, 2000.
[3] B. Bollobás, *Extremal Graph Theory*, London Mathematical Society Monographs, Vol. 11, Academic Press Inc. (Harcourt Brace Jovanovich Publishers), London, 1978.
[4] S. A. Burr and J. A. Roberts, *On Ramsey numbers for stars*, Util. Math. 4 (1973), 217–220.

[5] S. A. Burr, *Determining generalized Ramsey numbers is NP-hard*, Ars Combin. 17 (1984), 21–25.
[6] V. Chvátal, *Tree-complete graph Ramsey numbers*, J. Graph Theory 1 (1977), 93.
[7] F. R. K. Chung and R. L. Graham, *Erdős on Graphs: His Legacy of Unsolved Problems*, A K Peters Ltd., Wellesley, MA, 1998.
[8] F. R. K. Chung and C. M. Grinstead, *A survey of bounds for classical Ramsey numbers*, J. Graph Theory 1 (1983), 25–37.
[9] P. Erdős, *Some remarks on the theory of graphs*, Bull. Amer. Math. Soc. 53 (1947), 292–294.
[10] P. Erdős and G. Szekeres, *A combinatorial problem in geometry*, Compos. Math. 2 (1935), 464–470.
[11] P. Erdős and J. H. Spencer, *Probabilistic Methods in Combinatorics*, Probability and Mathematical Statistics, Vol. 17, Academic Press, New York-London, 1974.
[12] R. J. Faudree and R. H. Schelp, *All Ramsey numbers for cycles in graphs*, Discrete Math. 8 (1974), 313–329.
[13] L. Gerencsér and A. Gyárfás, *On Ramsey-type problems*, Ann. Univ. Sci. Budapest Eötvös Sect. Math. 10 (1967), 167–170.
[14] R. Graham, K. Leeb and B. Rothschild, *Ramsey's theorem for a class of categories*, Adv. Math. 8 (1972), 417–433.
[15] R. L. Graham, B. L. Rothschild and J. H. Spencer, *Ramsey Theory*, Second Edition, Wiley–Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons Inc., New York, 1990.
[16] S. Gautam, A. K. Srivastava and A. Tripathi, *On multicolour noncomplete Ramsey graphs of Star graphs*, Discrete Appl. Math. 156 Issue 12 (2008), 2423–2428.
[17] R. P. Gupta, *The chromatic index and the degree of a graph* (Abstract 66T-429), Notices Amer. Math. Soc. 13 (1966), 719.
[18] J. H. Kim, *The Ramsey number $\mathcal{R}(3,t)$ has order of magnitude $t^2/\log t$*, Random Structures Algorithms 7 (1995), 173–207.
[19] B. M. Landman and A. Robertson, *Ramsey Theory on the Integers*, Second Edition, Student Mathematical Library 24, American Mathematical Society (AMS), Providence, RI, 2014.
[20] J. Nešetřil, Ramsey Theory, *Handbook of Combinatorics*, Vol. 1, 2 (R. L. Graham et al., Eds.), Elsevier, Amsterdam, 1995, 1331–1403.
[21] J. Nešetřil and V. Rödl, Eds., *Mathematics of Ramsey Theory*, Algorithms and Combinatorics, vol. 5, Springer–Verlag, Berlin, 1990.
[22] J. Petersen, *Die Theorie der regulären graphs*, Acta Math. 15 (1891) 193–220.
[23] S. P. Radziszowski, *Small Ramsey Numbers*, Dynamic Survey 1, 15th Revision, Electron. J. Combin., 3 March 2017.
[24] F. P. Ramsey, *On a problem of formal logic*, Proc. London Math. Soc. 30 (1930), 264–286.
[25] F. S. Roberts, *Applications of Ramsey theory*, Discrete Appl. Math. 3 (1984), 251–261.
[26] V. Rosta, *On a Ramsey-type problem of J. A. Bondy and P. Erdős, I*, J. Combin. Theory 15B (1973), 94–104.
[27] V. Rosta, *On a Ramsey-type problem of J. A. Bondy and P. Erdős, II*, J. Combin. Theory 15B (1973), 105–120.
[28] V. Rosta, *Ramsey Theory Applications*, Dynamic Survey 13, Electron. J. Combin., 2004.
[29] A. Soifer, *Ramsey theory: yesterday, today, and tomorrow*, Progress in Mathematics, Boston, Mass., Birkhäuser, New York, 2011.
[30] W. T. Tutte, *The factors of graphs*, Canad. J. Math. 4 (1952) 314–328.
[31] W. T. Tutte, *A short proof of the factor theorem for finite graphs*, Canad. J. Math. 6 (1954) 347–352.
[32] V. G. Vizing, *On an estimate of the chromatic class of a p-graph*, Diskret. Analiz. 3 (1964) 25–30.

TAS
Telangana Academy of Sciences
HYDERABAD

## TELANGANA ACADEMY OF SCIENCES GUIDELINES TO CONTRIBUTORS FOR MANUSCRIPT PREPARATION

**The Proceedings of Telangana Academy of Sciences (Proc. TS Akad. Sc. - TAS), publishes review and research articles, research and short communications, S&T News of relevance to the Academy and Letters.**

***Review Articles:*** These are articles from personal and others research, covering the topics of current relevance in science and technology including interdisciplinary areas focusing a survey on available literature, current trends and future perspectives (8-10 Pages).

***Research Articles:*** These articles usually cover complete description of current research findings from the authors own work. They provide a full account on the experimental and analytical details (5-12 Pages).

***Short Research Communications:*** These communications cover the preliminary research findings from the authors own research work. They are fast-tracked for immediate publications (2-4 Pages).

***S&T News and Letters:*** These items cover the highlights of the latest developments and information related to science, technology, scientific administration available from the APAS and various other sources. Information about historical and recent developments of well-known universities and R&D institutes in the world, and news about awards, scholarships and others are also covered (1-2 Pages).

**Manuscript Preparation**

**The manuscript should be typed on one side of the A4 size paper in two columns. The font is Times New Roman.**

***Title:*** It should be bold, 14 point, centered to the page and brief in a maximum of two lines.

***Authors:*** Names should be in bold, 12 point, in the order - family name, middle name and last name. The corresponding author(s) name should be marked with a '*' and provided with his/her email address.

***Address:*** It should be normal, 10 point, with full postal details of department, institute.

***Abstract:*** A short abstract, 12 point, 10-15 lines (~100-150 words) describing the salient features.

***Keywords:*** A list of 5-6 keywords to be presented below the abstract, in 10 point.

***Text:*** The text, 12 point, should be divided into sub-headings, such as: Introduction, Results and Discussions, Conclusions, Acknowledgements, References. The format of the references is as follows:

1. (a) M. Siedlecka, G. Goch, A. Ejchart, H. Sticht, A. Bierzynski, Proc. Natl. Acad. Sci. USA 1999, 96, 903-908. (b) A. Patgiri, A. L. Jochim, P. S. Arora, Acc. Chem. Res. 2008, 41, 1289-1300.

2. (a) S. Hecht, I. Huc, (Eds.), Foldamers: Structure, Properties and Applications, Wiley-VCH: Weinheim, Germany, 2007. (b) C. Shellman, In Protein Folding; Jaenicke, R., Ed.; Elsevier: Amsterdam, 1980, pp 53-64.

**Submission:**

**The manuscript should be submitted as a soft copy along with a cover letter to the email ID: editor.tas@gmail.com. The cover letter should describe the Title, Authors along with their email IDs and mention the type of the masnucripts submitted along with the following statements:**

- The mansucript is submitted exclusively to TAS. The contents partially or totally were not submitted to any other journal simultaneoulsy.
- I have taken the consent of all the co-authors before submitting the work, as the corresponding author.
- I have verified the manuscript for plagiarism before submiting and I have obtained necessary copyright permissions wherever necessary.

**Justification:** The covering letter should coantian a paragraph of justification for the proposed publication.

**Reviewers:** The authors should provide a list 3-5 possible reviewers along with their departmetnal address, phone/fax numbers and email address for consideration of the Editorial Board **TAS**.

**Reviewing of articles:** All articles are first assessed by an editorial board member. Unsuitable manuscripts in their current form will be sent back to the author for modifciations and resubmission for peer reviewing, while, totally unsuitable articles will be returned without reveiwing.

The mansucripts are reviewed by two referees and positive assessment from both the referees is essentail for final acceptance. The authors will be notified on the acceptance/rejection or revision after receivng the comments from reveiwers. No further correspondance on the rejected manuscripts will be entertained.

Three hard copies may submitted by post to the editor at

**the address:**

**The Editor,**
**Telangana Academy of Sciences**
**Osmania University Campus,**
**Hyderabad-500 007.**
**Phone: 040-27098029/040-27070570**
**Email: editor.tas@gmail.com**

# Mathematical Sciences

Frontiers in Mathematics

TAS

Telangana Academy of Sciences, erstwhile A.P. Akademi of Sciences, the first State Academy, established in the year 1963 has been engaged in the advancement of Sciences in the State over 50 years. The Academy has been publishing journals and books on science including seminar Proceedings as a part of its publication activities. The Proceedings of Telangana Academy of Sciences (TAS) publishes review articles, research articles, research communications and S & T News.

## TAS publishes journals in the following research areas:

- Physical & Mathematical Sciences
- Engineering Sciences
- Chemical Sciences
- Earth, Ocean Atmospheric and Environmental Sciences
- Life Sciences and Agricultural Sciences
- Medical, Health and Pharmaceutical Sciences.


## Telangana Academy of Sciences

OU Qtr. No. L-68, Besides International Hostel
Tarnaka, Hyderabad 500 017. Phone: 040-2700 8029
E-mail: taspublications2018@gmail.com; tsas2015@gmail.com


**Science to Ignite Young Minds and Enlighten the Masses**

Science for the Nation's Prosperity