The Fundamental Theorem of Calculus

The single most important tool used to evaluate integrals is called "The Fundamental Theorem of Calculus". It converts any table of derivatives into a table of integrals and vice versa. Here it is

Theorem 1 (Fundamental Theorem of Calculus)**.**

Let $f(x)$ be a function which is defined and continuous for $a \le x \le b$.

Part 1: Define, for $a \le x \le b$, $F(x) = \int_a^x f(t)\,dt$. Then $F(x)$ is differentiable and

$$F'(x) = f(x)$$

Part 2: Let $G(x)$ be any function which is defined and continuous on $[a, b]$ and which is also differentiable and obeys $G'(x) = f(x)$ for all $a < x < b$. Then

$$\int_a^b f(x)\,dx = G(b) - G(a) \qquad \text{or} \qquad \int_a^b G'(x)\,dx = G(b) - G(a)$$

A function $G(x)$ that obeys $G'(x) = f(x)$ is called an antiderivative of f. The form $\int_a^b G'(x)\,dx = G(b) - G(a)$ of the Fundamental Theorem is occasionally called the "net change theorem".

"Proof" of Part 1. By definition

$$F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}$$

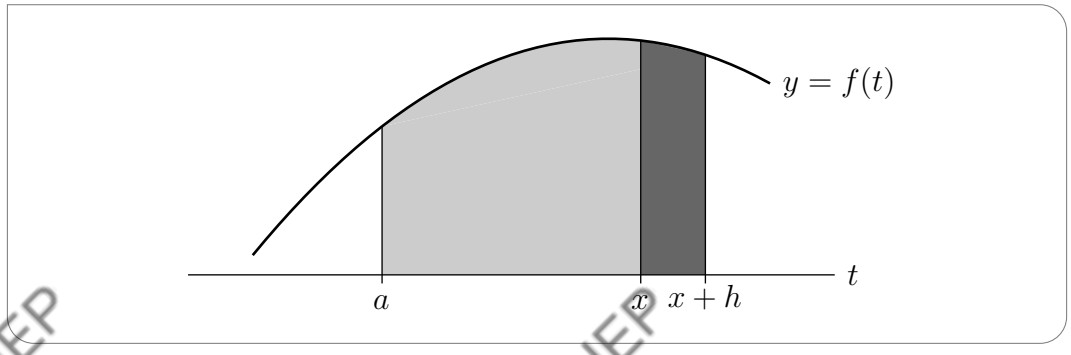
For notational simplicity, let's only consider the case that f is alway nonnegative. Then

$$\begin{aligned} F(x+h) &= \text{the area of the region } \left\{ (t,y) \;\middle|\; a \le t \le x+h,\ 0 \le y \le f(t) \right\} \\ F(x) &= \text{the area of the region } \left\{ (t,y) \;\middle|\; a \le t \le x,\ \quad 0 \le y \le f(t) \right\} \end{aligned}$$

So

$$F(x+h) - F(x) = \text{the area of the region } \left\{ (t,y) \;\middle|\; x \le t \le x+h,\ 0 \le y \le f(t) \right\}$$

That's the more darkly shaded region in the figure



As t runs from x to $x = h$, $f(t)$ runs only over a very range of values, all close to $f(x)$. So the darkly shaded region is almost a rectangle of width h and height $f(x)$ and so has an area which is very close to $f(x)h$. Thus $\frac{F(x+h)-F(x)}{h}$ is very close to $f(x)$. In the limit $h \to 0$, $\frac{F(x+h)-F(x)}{h}$ becomes exactly $f(x)$, which is exactly what we want. We won't justify this limiting argument on a mathematically rigorous level (which is why we put quotation marks around "Proof", above), but it should at least look very reasonable to you. □

"Proof" of Part 2. We want to show that $\int_a^b f(t)\ dt = G(b) - G(a)$, or equivalently that $\int_a^b f(t)\ dt - G(b) + G(a) = 0$. We'll just rename b to x and show that

$$H(x) = \int_a^x f(t)\ dt - G(x) + G(a)$$

is always zero. This will imply, in particular, that $H(b) = \int_a^b f(t)\ dt - G(b) + G(a)$ is zero.

First we'll check if $H(x)$ is at least a constant, by computing the derivative

$$\begin{aligned} H'(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(t)\ dt - G'(x) \\ &= f(x) - f(x) \qquad \text{(by Part 1 and the hypothesis } G'(x) = f(x)) \\ &= 0 \end{aligned}$$

So $H(x)$ must be a constant function and the value of the constant is

$$H(a) = \int_a^a f(t)\ dt - G(a) + G(a) = 0$$

as we want. □

We'll first do some examples illustrating the use of part 1 of the Fundamental Theorem of Calculus. Then we'll move on to part 2.

Example 2 ($\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x e^{-t^2}\ dt$)

Find $\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x e^{-t^2}\ dt$.

Solution. We don't know how to evaluate the integral $\int_0^x e^{-t^2}\ dt$. In fact $\int_0^x e^{-t^2}\ dt$ cannot be expressed in terms of standard functions like polynomials, exponentials, trig functions and so on. Even so, we can find its derivative by just applying the first part of the Fundamental Theorem of Calculus with $f(t) = e^{-t^2}$ and $a = 0$. That gives

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x e^{-t^2}\ dt = e^{-x^2}$$

Example 2

Example 3 ($\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{x^2} e^{-t^2}\,dt$)

Find $\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{x^2} e^{-t^2}\,dt$.

Solution. Once again, we will apply part 1 of the Fundamental Theorem of Calculus. But we must do so with some care. The Fundamental Theorem tells us how to compute the derivative of functions of the form $\int_a^x f(t)\,dt$. The integral $\int_0^{x^2} e^{-t^2}\,dt$ is *not* of the specified form because the upper limit of $\int_0^{x^2} e^{-t^2}\,dt$ is x^2 while the upper limit of $\int_a^x f(t)\,dt$ is x. The trick for getting around this obstacle is to define the auxiliary function

$$E(x) = \int_0^x e^{-t^2}\,dt$$

The Fundamental Theorem tells us that $E'(x) = e^{-x^2}$. (We found that in Example 2, above.) The integral of interest is

$$\int_0^{x^2} e^{-t^2}\,dt = E(x^2)$$

So by the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{x^2} e^{-t^2}\,dt = \frac{\mathrm{d}}{\mathrm{d}x}E(x^2) = 2x\,E'(x^2) = 2xe^{-x^4}$$

Example 3

Example 4 ($\frac{\mathrm{d}}{\mathrm{d}x}\int_x^{x^2} e^{-t^2}\,dt$)

Find $\frac{\mathrm{d}}{\mathrm{d}x}\int_x^{x^2} e^{-t^2}\,dt$.

Solution. Yet again, we can't just blindly apply the Fundamental Theorem. This time, not only is the upper limit of integration x^2 rather than x, but the lower limit of integration also depends on x, unlike the lower limit of the integral $\int_a^x f(t)\,dt$ of the Fundamental Theorem. Fortunately we can use the basic properties of integrals to split $\int_x^{x^2} e^{-t^2}\,dt$ into pieces whose derivatives we already know.

$$\int_x^{x^2} e^{-t^2}\,dt = \int_x^0 e^{-t^2}\,dt + \int_0^{x^2} e^{-t^2}\,dt = -\int_0^x e^{-t^2}\,dt + \int_0^{x^2} e^{-t^2}\,dt$$

So, by the previous two examples,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\int_x^{x^2} e^{-t^2}\,dt &= -\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x e^{-t^2}\,dt + \frac{\mathrm{d}}{\mathrm{d}x}\int_0^{x^2} e^{-t^2}\,dt \\ &= -e^{-x^2} + 2xe^{-x^4}\end{aligned}$$

Example 4

We're almost ready for examples using part 2 of the Fundamental Theorem. We just need a little terminology and notation.

Definition 5.

(a) A function $F(x)$ whose derivative $F'(x) = f(x)$ is called an antiderivative of $f(x)$.

(b) The symbol $\int f(x)\,dx$ is read "the indefinite integral of $f(x)$". It stands for *all* functions having derivative $f(x)$. If $F(x)$ is any antiderivative of $f(x)$, and C is any constant, then the derivative of $F(x)+C$ is again $f(x)$, so that $F(x)+C$ is also an antiderivative of $f(x)$. Conversely, the difference between any two antiderivatives of $f(x)$ must be a constant, because a function has derivative zero if and only if it is a constant. So $\int f(x)\,dx = F(x) + C$, with the consant C called an "arbitrary constant" or "constant of integration".

(c) The symbol $\int f(x)\,dx\Big|_a^b$ means

- take any function whose derivative is $f(x)$. Call the function you have chosen $F(x)$.
- Then $\int f(x)\,dx\Big|_a^b$ means $F(b) - F(a)$.

We'll later develop some strategies for computing more complicated integrals. But for now, we'll stick to integrals that are simple enough that we can just guess the answer.

Example 6

Find $\int_1^2 x\,dx$.

Solution. The main step in evaluating an integral like this is finding the indefinite integral of x. That is, finding a function whose derivative is x. So we have to think back and try and remember a function whose derivative is something like x. We recall that

$$\frac{\mathrm{d}}{\mathrm{d}x}x^n = nx^{n-1}$$

We want the derivative to be x to the power one, so we should take $n = 2$. So far, we have

$$\frac{\mathrm{d}}{\mathrm{d}x}x^2 = 2x$$

This derivative is just a factor of 2 larger than we want. So we divide the whole equation by 2. We now have

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{2}x^2\right) = x$$

which says that $\frac{1}{2}x^2$ is an antiderivative for x. Once one has an antiderivative, it is easy to compute the definite integral

$$\int_1^2 x\,dx = \overbrace{\frac{1}{2}x^2}^{\text{a function with derivative } x.}\Bigg|_1^2 = \frac{1}{2}2^2 - \frac{1}{2}1^2 = \frac{3}{2}$$

as well as the indefinite integral

$$\int x\ dx = \frac{1}{2}x^2 + C$$

Example 6

Example 7

Find $\int_0^{\pi/2} \sin x\ dx$.

Solution. Once again, the crux of the solution is guessing a function whose derivative is $\sin x$. The standard derivative that comes closest to $\sin x$ is

$$\frac{\mathrm{d}}{\mathrm{d}x}\cos x = -\sin x$$

which is the derivative we want, multiplied by a factor of -1. So we multiply the whole equation by -1.

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(-\cos x\big) = \sin x$$

This tells us that the indefinite integral $\int \sin x dx = -\cos x + C$. To answer the question, we don't need the whole indefinite integral. We just need one function whose derivative is $\sin x$, that is, one antiderivative of $\sin x$. We'll use the simplest one, namely $-\cos x$. The prescribed integral is

$$\int_0^{\pi/2} \sin x\ dx = \overbrace{-\cos x}^{\text{a function with derivative } \sin x.}\Big|_0^{\pi/2} = -\cos\frac{\pi}{2} + \cos 0 = -0 + 1 = 1$$

Example 7

Example 8

Find $\int_1^2 \frac{1}{x}\ dx$.

Solution. Once again, the crux of the solution is guessing a function whose derivative is $\frac{1}{x}$. Our standard way to get derivatives that are powers of x is

$$\frac{\mathrm{d}}{\mathrm{d}x}x^n = nx^{n-1}$$

That is not going to work this time, since to get $\frac{1}{x}$ on the right hand side we need to take $n = 0$, which gives a right hand side of 0. Fortunately, we also have

$$\frac{\mathrm{d}}{\mathrm{d}x}\ln x = \frac{1}{x}$$

which is exactly the derivative we want. We're now ready to compute the prescribed integral.

$$\int_1^2 \frac{1}{x}\ dx = \overbrace{\ln x}^{\text{a function with derivative } 1/x.}\Big|_1^2 = \ln 2 - \ln 1 = \ln 2$$

Example 8

Example 9

Find $\int_{-2}^{-1} \frac{1}{x}\, dx$.

Solution. As we saw in the last example,

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln x = \frac{1}{x}$$

But we cannot use $\ln x$ in this example because, here, x runs from -2 to -1, and in particular is negative, and $\ln x$ is not defined when x is negative. A variant of $\ln x$ which is defined when x is negative is $\ln(-x) = \ln|x|$, so let's compute

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln(-x) = \frac{1}{-x}(-1) = \frac{1}{x}$$

by the chain rule. Fortunately, this is exactly the derivative we want, so we're now ready to compute the prescribed integral.

$$\int_{-2}^{-1} \frac{1}{x}\, dx = \overbrace{\ln(-x)}^{\substack{\text{a function with} \\ \text{derivative } 1/x.}} \Big|_{-2}^{-1} = \ln 1 - \ln 2 = \ln \frac{1}{2}$$

The statements

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln x = \frac{1}{x} \qquad \text{for } x > 0$$

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln(-x) = \frac{1}{x} \qquad \text{for } x < 0$$

are often combined into

$$\frac{\mathrm{d}}{\mathrm{d}x} \ln|x| = \frac{1}{x}$$

Example 9

Example 10

Find $\int_{-1}^{1} \frac{1}{x^2}\, dx$.

Solution. Beware that this is a particularly nasty example, which illustrates a booby trap hidden in the Fundamental Theorem of Calculus. The booby trap explodes when the theorem is applied sloppily. The sloppy solution starts, as normal, with the observation that

$$\frac{\mathrm{d}}{\mathrm{d}x} \frac{1}{x} = -\frac{1}{x^2}$$

so that

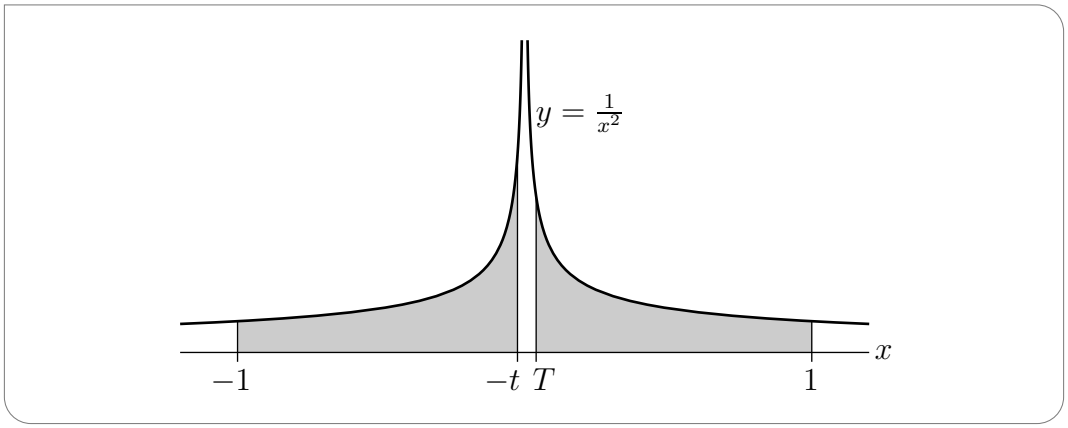
$$\frac{\mathrm{d}}{\mathrm{d}x}\Big(-\frac{1}{x}\Big) = \frac{1}{x^2}$$

and it appears that

$$\int_{-1}^{1} \frac{1}{x^2}\,dx = \Big[\overbrace{-\frac{1}{x}}^{\substack{\text{a function with}\\ \text{derivative } 1/x^2.}}\Big]_{-1}^{1} = -\frac{1}{1} - \Big(-\frac{1}{-1}\Big) = -2$$

Unfortunately, this answer cannot be correct. In fact it is ridiculous. The integrand $\frac{1}{x^2} > 0$, so the integral has to be positive. The flaw in the argument is that the Fundamental Theorem of Calculus, which says that if $F'(x) = f(x)$ then $\int_a^b f(x)\,dx = F(b) - F(a)$, is applicable only when $F'(x)$ exists and equals $f(x)$ for **all** x between a and b. In this case $F'(x) = \frac{1}{x^2}$ does not exist for $x = 0$. So we cannot apply the Fundamental Theorem of Calculus as we tried to above.

An integral, like $\int_{-1}^{1} \frac{1}{x^2}\,dx$, whose integrand is undefined somewhere in the domain of integration is called improper. We'll later give a more thorough treatment of improper integrals. For now, we'll just say that the correct way to define improper integrals is as a limit of well–defined approximating integrals. The approximating integrals have restricted domains of integration that exclude the "bad" points where the integrand is undefined. In the current example, the original domain of integration is $-1 \le x \le 1$. The domains of integration of the approximating integrals exclude from $[-1, 1]$ small intervals around $x = 0$. The shaded area in the figure below illustrates a typical approximating integral, whose domain of integration consists of the original domain of integration, $[-1, 1]$, but with the interval $[-t, T]$ excluded.



The full domain of integration is only recovered in the limit $t, T \to 0$.

For this example, the correct computation is

$$\begin{aligned}\int_{-1}^{1} \frac{1}{x^2}\,dx &= \lim_{t\to 0+} \int_{-1}^{-t} \frac{1}{x^2}\,dx \;+\; \lim_{T\to 0+} \int_{T}^{1} \frac{1}{x^2}\,dx \\ &= \lim_{t\to 0+} \Big[-\frac{1}{x}\Big]_{-1}^{-t} + \lim_{T\to 0+} \Big[-\frac{1}{x}\Big]_{T}^{1}\end{aligned}$$

$$= \lim_{t\to 0+}\left[\left(-\frac{1}{-t}\right)-\left(-\frac{1}{-1}\right)\right]+\lim_{T\to 0+}\left[\left(-\frac{1}{1}\right)-\left(-\frac{1}{T}\right)\right]$$
$$= \lim_{t\to 0+}\frac{1}{t}+\lim_{T\to 0+}\frac{1}{T}-2$$
$$= +\infty$$

Example 10

The above examples have illustrated how we can use the Fundamental Theorem of Calculus to convert knowledge of derivatives into knowledge of integrals. We are now in a position to easily built a table of integrals. Here is a short table of the most important derivatives that we know.

$F(x)$	1	x^p	$\sin x$	$\cos x$	$\tan x$	e^x	$\ln x$	$\arcsin x$	$\arctan x$
$f(x)=F'(x)$	0	px^{p-1}	$\cos x$	$-\sin x$	$\sec^2 x$	e^x	$\frac{1}{x}$	$\frac{1}{\sqrt{1-x^2}}$	$\frac{1}{1+x^2}$

And here is the corresponding short table of integrals.

$f(x)$	$F(x)=\int f(x)\,dx$
1	$x+C$
x^p	$\frac{x^{p+1}}{p+1}+C$ if $p\neq -1$
$\frac{1}{x}$	$\ln\lvert x\rvert+C$
$\sin x$	$-\cos x+C$
$\cos x$	$\sin x+C$
$\sec^2 x$	$\tan x+C$
e^x	e^x+C
$\frac{1}{\sqrt{1-x^2}}$	$\arcsin x+C$
$\frac{1}{1+x^2}$	$\arctan x+C$

2002

Fundamental theorem of algebra

Paul Shibalovich

FUNDAMENTAL THEOREM OF ALGEBRA

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

in

Mathematics

---

by

Paul Shibalovich

December 2002

FUNDAMENTAL THEOREM OF ALGEBRA

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

by

Paul Shibalovich

December 2002

Approved by:

[redacted]

Gary R. Griffing, Committee Chair — Date: 11/21/02

[redacted]

Belisario Ventura, Committee Member

[redacted]

James S. Okon, Committee Member

[redacted]

Peter Williams, Chair
Department of Mathematics

J. T. Hallett

Terri Hallett,
Graduate Coordinator
Department of
Mathematics

ABSTRACT

The Fundamental Theorem of Algebra (FTA) is an important theorem in Algebra. This theorem asserts that the complex field is algebraically closed. That is, if a polynomial of degree n has n-m real roots ($0 \leq m \leq n$), then the Fundamental Theorem asserts that the polynomial has its remaining m roots in the complex plane.

This thesis will include historical research of proofs of the Fundamental Theorem of Algebra and provide information about the first proof given by Gauss of the Theorem and the time when it was proved. Also, it will include proofs of the Fundamental Theorem using three different approaches: algebraic approach, complex analysis approach, and Galois Theory approach.

The conclusion of the thesis will explain the similarities of the three proofs as well as their differences.

ACKNOWLEDGMENTS

First of all, I want to thank Dr. Gary Griffing, Dr. Belisario Ventura, and Dr. Jim Okon for their willingness to assist me in working on my thesis. In addition, I would like to express a special appreciation to Dr. Gary Griffing for his help and time dedication in helping and working with me. Also, I want to thank all the professors at California State University, San Bernardino who helped me build strong mathematical foundation throughout the two years in the Master's program and prepared me to become successful in Mathematics.

DEDICATION

To

Nadya Pavlov

Timothy Shibalovich

Olga Shibalovich

Lidia Pelepchuk

Anna Pavlenko

Lubov Rudenko

TABLE OF CONTENTS

CHAPTER ONE

BACKGROUND

1.1 Introduction

The content of Chapter One presents an overview of the thesis. The contexts of the problem are discussed and are then followed by the purpose and significance of the thesis. Finally, the notation to be used is presented.

1.2 Purpose of the Thesis

The purpose of this thesis is to explore development of the Fundamental Theorem of Algebra. Prior to the 17th century AD there were several attempts to prove it, but they failed. It was in 1799 that Gauss, in his dissertation, proved the Theorem for the first time.

1.3 Context of the Problem

Being an instructor of Mathematics, the Fundamental Theorem of Algebra arises in College Algebra class. It is simply stated concept and the students do not have any problem understanding its importance and applicability.

1.4 Significance of the Thesis

The significance of this thesis is to inform the reader about the development of the Fundamental Theorem of Algebra and its proofs. Also, the significance of this

thesis is to educate the reader about different approaches used to prove the Theorem.

1.5 Notations

The notational conventions used in this thesis are the following:

1.5.1 Definition

FTA will stand for the Fundamental Theorem of Algebra.

1.5.2 Definition

The set of real numbers will be denoted by R.

1.5.3 Definition

The set of complex numbers will be denoted by C.

1.5.4 Definition

A set, for example P, will be denoted by P.

1.6 Organization of the Thesis

This thesis is divided into six chapters. Chapter One provides an introduction to the context of the problem, purpose of the thesis, significance of the thesis, and the notation. Chapter Two consists of historical development of the FTA and its proof. Chapter Three explores the proof of the FTA using complex analysis approach. Chapter Four presents the algebraic proof of the FTA. Chapter Five gives a proof of the FTA using Galois Theory approach.

Chapter Six examines the similarities of the three proofs as well as their differences. Finally, the references include the sources used for this thesis.

CHAPTER TWO

HISTORY OF THE FUNDAMENTAL THEOREM OF ALGEBRA

2.1 Introduction

The Fundamental Theorem of Algebra is one of the most important results in Algebra. In the past it provided a motivation to study the set of complex numbers and polynomials whose roots are in this set. In this paper we will research the development of the Fundamental Theorem of Algebra and prove it using different approaches. We will start with development of the FTA, and then lead the reader through to the first proof (1799) presented by Carl Friedrich Gauss (1777-1855). Then we will prove the FTA using three different approaches and explore their similarities and differences. We will prove most theorems used in this thesis. If a theorem is not proved, we will provide ample reference to the proof of the theorem.

2.2 Development of the Fundamental Theorem of Algebra

Early studies of the roots of equations involved only positive real roots, so the Fundamental Theorem of Algebra was not relevant at that time. "Cardano was the first to realize that one could work with quantities more general

than the real numbers" [1]. He worked with cubic equations, in particular, the equation $x^3 = 15x + 4$ which gave him an answer involving negative square root: $\sqrt{-121}$. Cardano was able to manipulate equations with 'complex numbers', but he did not understand his own mathematics. The concept of the number of roots and the degree of a polynomial was slowly developing. In 1629 Flemish mathematician Albert Girard was first to claim that for any equation of degree n there are always n roots. However, it was not considered that solutions are of the form $a + bi$ where a and b are real numbers. Girard was the first to conjecture that "a polynomial equation of degree n must have n roots" [1]. In 1637 Descartes said that "one can 'imagine' for every equation of degree n, n roots but these imagined roots do not correspond to any real quantity" [2]. However, his statement was merely a suggestion.

2.3 Generalization of the Fundamental Theorem of Algebra

The first attempt to 'prove' FTA was given by Leibniz in 1702, however, it was unsuccessful [2]. Leibniz considered the equation $x^4 + y^4 = 0$ and claimed that this equation could never be written as a product of two real quadratic factors. Unfortunately, he was mistaken not

realizing that $x^4 + y^4 = 0$ can be written as $x^4 + 2x^2y^2 + y^4 - 2x^2y^2 = 0$ or $(x^2 + y^2)^2 - 2x^2y^2 = 0$ or $\left(x^2 + y^2 - \sqrt{2}xy\right)\left(x^2 + y^2 + \sqrt{2}xy\right) = 0$. Leibniz' conclusion withheld him from further study of the general equation $x^4 + y^4 = 0$, and thus from finding the complex roots of the equation $x^4 + 1 = 0$. Then in 1742 Euler showed that Leibniz' assertion was false. Later, Euler proved that every real polynomial of degree n, $n \leq 6$, has exactly n 'complex roots'. It was only in 1749 that Euler tried to prove the FTA for the general case. At that time the FTA was generalized and stated as follows:

"Every polynomial of degree n with real coefficients has exactly n zeros in C" [1].

Nowadays, the FTA is stated in the following way:

"A polynomial with coefficients which are complex numbers has all its roots in the complex field" [3].

2.4 First Proof of the Fundamental Theorem of Algebra

There were several attempts to prove the FTA for the general case, but the first legitimate proof was given by Carl Friedrich Gauss in his doctorial thesis of 1799 [1]. This proof was topological in nature. Throughout his lifetime Gauss produced several different proofs.

2.5 Summary

The FTA was born with an intuitive idea, which was later generalized by Girard in 1629, and finally became a fundamental result in Algebra. Since the first proof of the FTA that was done by Gauss in 1799 we have, today, at least six conceptually different proofs of this important Theorem[3]. These proofs include a complex analysis approach, and analysis approach, a purely algebraic approach without Galois Theory, and algebraic approach with some analysis, a topological approach, and Galois Theory approach. In this thesis, we will explore three proofs of the FTA using a complex analysis approach, a non-Galois Theory algebraic approach, and a Galois Theory approach.

CHAPTER THREE

COMPLEX ANALYSIS PROOF OF THE FUNDAMENTAL THEOREM OF ALGEBRA

3.1 Introduction

The proof of the FTA using the complex analysis approach requires some complex analysis background. Section 3.2 provides definitions needed to prove the Theorem. In addition to these definitions we will use Liouville's Theorem and the Cauchy Inequality. The proofs for Liouville's Theorem and the Cauchy Inequality will be provided in sections 3.3.2 and 3.3.3 respectively. Section 3.4 will provide the complex analysis proof of the FTA.

3.2 Definitions and Facts

3.2.1 Definition

A function $f(z)$ of the complex variable z is analytic in an open set G if it has a derivative at each point in the set G.

3.2.2 Definition

An entire function is a function that is analytic at each point in the entire complex plane.

3.2.3 Definition

A function $f(z)$ of the complex variable z is said to be continuous in a region G if it is continuous at each point in G.

3.2.4 Definition

The function $f(x)$ is bounded on the set K if there is a number M such that $|f(x)| \leq M$ for all $x \in K$.

In the proof of Liouville's Theorem we will use Cauchy Integral Formula and Cauchy-Riemann equations stated below.

3.2.5 Fact (Cauchy Integral Formula)

If a function $f(z) = u(x,y) + iv(x,y)$ is analytic at a point $z_0 = x_0 + iy_0$ and the component functions u and v have continuous partial derivatives of all orders at that point, then

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \cdot \int_C \frac{f(z)\,dz}{(z - z_0)^{n+1}} \text{ for } n = 0, 1, 2, \ldots$$

where C is a positively oriented simple closed contour and $|f(x)| \leq M$ [4].

3.2.6 Fact (*Cauchy-Riemann Equations*) [4].

Let $f(z) = u(x,y) + iv(x,y)$. If $f'(z)$ exists where $z_0 = x_0 + iy_0$, then the following are true.

i) $u_x(x_0, y_0) = v_y(x_0, y_0)$

ii) $u_y(x_0, y_0) = -v_x(x_0, y_0)$

3.3 Theorems

Theorem 3.3.1: If $f'(z) = 0$ everywhere in a domain D, then $f(z)$ must be constant throughout D [4].

Proof: Let $f(z) = f(x,y) = u(x,y) + iv(x,y) \in C$ and $z = x + iy$. Suppose $f'(z) = f'(x,y) = 0 \ \forall \ z \in C$. In order to show that $f(z)$ must be constant throughout D, it suffices to show that $f(z) = a + ib$ with $a, b \in R$ for all z. Since $f'(z) = 0$ and $f'(z) = u_x + iv_x$, we get $u_x = 0 = v_x$. Similarly, $f'(z) = 0$ and $f'(z) = u_y + iv_y$ give us $u_y = 0 = v_y$. So, we have equality $u_x = u_y = v_x = v_y = 0$. This implies that $u = a$ and $v = b$ with $a, b \in R$ for all z. Thus, we conclude that $f(z) = a + ib$ an element of C for all z. Hence, $f(x)$ must be constant throughout D.

Theorem 3.3.2: Let $z_0 = x_0 + iy_0$, $f(z) = u(x,y) + iv(x,y)$ be an analytic function within and on a circle $|z - z_0| = r$, and component functions u and v have continuous partial derivatives of all orders at that point. Also, let C denote the positively oriented circle $|z - z_0| = r$, then

$$\left|f^{(n)}(z_0)\right| \leq \frac{n!M_r}{r^n}$$

where r is radius of the circle C and M_r is the maximum value of the function on C[4].

Proof: Assume that $z_0 = x_0 + iy_0$ and that $f(z) = u(x,y) + iv(x,y)$ is an analytic function within and on a circle $|z - z_0| = r$. Since the component functions u and v have continuous partial derivatives of all orders, and f is bounded on C, then by Cauchy Integral Formula

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \cdot \int_C \frac{f(z)dz}{(z-z_0)^{n+1}} \text{ for } n = 0, 1, 2, \ldots \qquad (1)$$

where C is a positively oriented circle $|z - z_0| = r$ and $|f(z)| \leq M$.

Now, the maximum value of $|f(z)|$ on the circle C depends on the radius of C. Let M_r denote that maximum value of $|f|$ on the circle of radius r. Using (1), we get the Cauchy Inequality $\left|f^{(n)}(z_0)\right| \leq \frac{n!}{2\pi} \cdot \frac{M_r}{r^{n+1}} \cdot 2\pi r$ for $n = 1, 2, \ldots$ or

$$\left|f^{(n)}(z_0)\right| \leq \frac{n!M_r}{r^n} \text{ for } n = 1, 2, \ldots$$

where M_R is a bound for $f(z)$ on $|z - z_0| \leq r$.

Theorem 3.3.3 (Liouville's Theorem): If $f(z)$ is entire and bounded in the complex plane C, then $f(z)$ is constant throughout the plane.

Proof: Let $f(z)$ be an entire and bounded function in the complex plane. To prove this theorem we need to show that $f(z)$ is a constant throughout C [4].

We are given that $f(z)$ is entire. By definition of 3.2.2, $f(z)$ is analytic at each point in the entire plane. Then, by definition of 3.2.1, $f(z)$ has a derivative at each point in the set C. Now, by Theorem 3.2.2 for any circle in the plane, there exists maximum value $M_r > 0$ that depends on the radius r of the circle C such that $|f'(z)| \le \frac{M_r}{r}$ for an arbitrary $z \in C$. In this theorem we are also given that $f(z)$ is bounded in the complex plane. This implies that there exists a constant $M > 0$ such that $|f(z)| \le M$ for all $z \in C$. Since M_r is maximum value of f on C and M is maximum value in entire plane, the inequality $M_r \le M$ is true independently of the radius r. Thus, $|f'(z)| \le \frac{M}{r}$ where z is any fixed point and r is arbitrary large. However, the inequality $|f'(z)| \le \frac{M}{r}$ with

an arbitrary large radius r can hold only if $f'(z) = 0$. Since the choice of z was arbitrary, the statement $f'(z) = 0$ must be true everywhere in the complex plane. Hence, function $f(z)$ is a constant.

3.4 Fundamental Theorem of Algebra

Theorem 3.4.1: Any polynomial $P(z) = a_0 + a_1 z + a_2 z^2 + \ldots + a_n z^n$ with $a_n \neq 0$ of degree $n \geq 1$ has at least one zero in C. That is, there exists at least one point z_0 such that $P(z_0) = 0$ [4].

Proof: Suppose $P(z) = a_0 + a_1 z + a_2 z^2 + \ldots + a_n z^n$, where $a_n \neq 0$. Consider $f(z) = \frac{1}{P(z)}$. Since $\forall z \in C$ there exists a derivative of $f(z)$, the function $f(z)$ is analytic. By definition 3.2.2, $f(z)$ is an entire function. Now, we want to show that $f(z)$ is bounded in C. Dividing $P(z)$ by z^n, we get $\frac{P(z)}{z^n} = \frac{a_0}{z^n} + \frac{a_1}{z^{n-1}} + \ldots + \frac{a_{n-1}}{z} + a_n$ where $a_n \neq 0$. Now, let

$$w = \frac{a_0}{z^n} + \frac{a_1}{z^{n-1}} + \ldots + \frac{a_{n-1}}{z} \qquad (2)$$

This implies that $P(z) = (w + a_n) z^n$. There exists large enough $K \in N$ such that when $|z| \geq K$, $\frac{|a_i|}{|z^{n-i}|} < \frac{|a_n|}{2n}$

$\forall\, i = 1, 2, \ldots, n$, that is for each $\frac{a_i}{z^{n-i}}$ in (2). By the triangular inequality this implies that

$|w| \le \frac{|a_0|}{|z^n|} + \frac{|a_1|}{|z^{n-1}|} + \ldots + \frac{|a_{n-1}|}{|z|} < \frac{|a_n|}{2}$ for all values of z. So when $|z| \ge K$, $|w| < \frac{|a_n|}{2}$. Thus, $|w| - |a_n| < \frac{|a_n|}{2} - |a_n|$ or $|w| - |a_n| < -\frac{|a_n|}{2}$, and hence $|a_n| - |w| > \frac{|a_n|}{2}$. Consequently, we get that

$$\big||a_n| - |w|\big| > \frac{|a_n|}{2} \tag{3}$$

However, $|a_n + w| \ge \big||a_n| - |w|\big|$ is always true. This and inequality (3) then gives us $|a_n + w| \ge \big||a_n| - |w|\big| > \frac{|a_n|}{2}$ or

$$|a_n + w| > \frac{|a_n|}{2} \tag{4}$$

Since $P(z) = (w + a_n)z^n$ and because of inequality (4), we can say that $|P(z)| = |a_n + w| \cdot |z^n| > \frac{|a_n|}{2} \cdot |z|^n \ge \frac{|a_n|}{2} \cdot K^n$ whenever $|z| \ge K$ for arbitrary positive real number. This then implies that $|P(z)| > \frac{|a_n|}{2} \cdot K^n$ or $|f(z)| = \frac{1}{|P(z)|} < \frac{2}{|a_n| \cdot K^n}$ whenever $|z| \ge K$. This shows that f is bounded in the

region exterior to the disk $|z| \geq K$. The function $f(z)$ is continuous in the closed disk $|z| \leq K$ because $f(z) = \frac{1}{P(z)}$ is differentiable at each $z \in C$. Therefore $f(z)$ is bounded in the closed disk $|z| \leq K$. This implies that $f(z)$ is bounded in the entire plane.

Since $f(z)$ is entire and bounded in C by the Liouville's Theorem, $f(z)$ is a constant. Since $f(z)$ is a constant, $f(z) = \frac{1}{P(z)}$ and $\deg P(z) \geq 1$, $P(z)$ is also a constant. But this is a contradiction. Thus, the assumption $P(z) \neq 0$ for every value of z is not true. Hence, there exists at least one point $z_0 \in C$ such that $P(z_0) = 0$.

3.5 Summary

The complex analysis proof of the FTA is concise, and is proved by contradiction. The argument in this proof goes as follows: if a non-constant polynomial has no zeros, the multiplicative inverse of this polynomial is a bounded analytic function. However, Liouville's Theorem shows that such a function is constant. Thus, the polynomial itself has to be a constant, which is a

contradiction to the assumption. Although the complex analysis proof does require additional knowledge from Complex Analysis, it is not hard to understand it.

CHAPTER FOUR

ALGEBRAIC PROOF OF THE FUNDAMENTAL THEOREM OF ALGEBRA

4.1 Introduction

The algebraic proof of the FTA requires some background from Abstract Algebra. Section 4.2 presents definitions used for theorems in this chapter. Theorems needed to prove the FTA using an algebraic approach without the use of Galois Theory are reviewed in section 4.3. The proof of the FTA is presented in section 4.4.

4.2 Definitions

4.2.1 Definition

A field F is called formally real if -1 is not expressible in it as a sum of squares.

4.2.2 Definition

A field P is called a real closed field if P is formally real, but no proper algebraic extension of P is formally real.

4.2.3 Definition

A field F is algebraically closed if every polynomial equation with coefficients in F has a solution

in F. That is, F is algebraically closed if any $P(x) \in Q[x]$ has its roots in $Q[x]$.

4.2.4 Definition

Commutative ring F in which the set of nonzero elements forms a group with respect to multiplication is called a field. Field E is said to be an extension of F, if E contains a subfield isomorphic to F.

4.2.5 Definition

Assume that E is an extension of F. An element $a \in E$ is said to be algebraic over F if a is a solution of some polynomial equation with coefficients in F.

4.2.6 Definition

A field K is called an ordered field if the property of positiveness (> 0) is defined for its elements and if it satisfies the following postulates.

i) $\forall a \in K,\ a = 0,\ a > 0,\ -a > 0$

ii) If $a > 0$ and $b > 0$, then $a + b > 0$ and $ab > 0$

4.3 Theorems

Theorem 4.3.1: In the field of complex numbers the equation $x^2 = a + bi$ with a and b being real numbers is always solvable. That is, every number of the field has a square root in the field[5].

Proof: Let $x = c + di$ where c and d are real numbers. This implies that $x^2 = (c + di)^2 = c^2 + 2cdi - d^2 =$

$= (c^2 - d^2) + (2cd)i$.

Now, let's define a and b as

$$a = c^2 - d^2 \tag{1}$$

$$b = 2cd \tag{2}$$

So that $a + bi = (c^2 - d^2) + (2cd)i$. Then $a^2 + b^2$ gives us $a^2 + b^2 = (c^2 - d^2)^2 + (2cd)^2 = c^4 - 2c^2d^2 + d^4 + 4c^2d^2 =$

$= c^4 + 2c^2d^2 + d^4$ or $a^2 + b^2 = c^4 + 2c^2d^2 + d^4 = (c^2 + d^2)^2$. From the last statement we get $c^2 + d^2 = \sqrt{a^2 + b^2}$, since $c^2 + d^2 \geq 0$. Now, from (1) we have $a = c^2 - d^2$ or $d^2 = c^2 - a$. This implies that $c^2 + d^2 = c^2 + c^2 - a$ or $2c^2 = a + c^2 + d^2$ or

$$c^2 = \frac{a + \sqrt{a^2 + b^2}}{2} \tag{3}$$

Similarly, from (1) we have $a = c^2 - d^2$ or $c^2 = d^2 + a$. This implies that $c^2 + d^2 = d^2 + a + d^2 = \sqrt{a^2 + b^2}$ or

$$d^2 = \frac{-a + \sqrt{a^2 + b^2}}{2} \tag{4}$$

From (3) we get $c = \pm\sqrt{\frac{a + \sqrt{a^2 + b^2}}{2}}$, and from (4) we get $d = \pm\sqrt{\frac{-a + \sqrt{a^2 + b^2}}{2}}$. This shows that for every choice of c and d a square root of $a + bi$ is in the field of complex numbers.

Now, we will extend previous theorem in the following way.

Theorem 4.3.2: Let K be any arbitrary ordered field with the property that if $a \in K$, $a > 0$, then $\sqrt{a} \in K$. If $a + bi \in K(i)$, where $i^2 = -1$, then there exists $c + di \in K(i)$ such that $(c + di)^2 = a + bi$ [5].

Proof: Assume that if $a \in K$ and $a > 0$, then $\sqrt{a} \in K$. Let $a + bi \in K(i)$. Need to show that there exists $c + di \in K(i)$ such that $(c + di)^2 = a + bi$. From Theorem 4.3.1 we found that $c = \pm\sqrt{\frac{a + \sqrt{a^2 + b^2}}{2}}$, $d = \pm\sqrt{\frac{-a + \sqrt{a^2 + b^2}}{2}}$.

Choose $c = \sqrt{\frac{a + \sqrt{a^2 + b^2}}{2}}$ and $d = \sqrt{\frac{-a + \sqrt{a^2 + b^2}}{2}}$, then

$$(c + di)^2 = \left(\sqrt{\frac{a + \sqrt{a^2 + b^2}}{2}} + i\sqrt{\frac{-a + \sqrt{a^2 + b^2}}{2}}\right)^2 =$$

that for every $M > 0$ there exists $0 < x_0$ such that for all $x_0 < x_1$, $0 < M < \frac{P(x_1)}{x_1^{n-1}}$ and therefore $P(x_1) > 0$.

Similarly, $\lim_{x \to -\infty} \frac{P(x)}{x^{n-1}} = \lim_{x \to -\infty} \frac{a_n x^n + \ldots + a_1 x + a_0}{x^{n-1}} =$

$= \lim_{x \to -\infty} (a_n x + a_{n-1} + \frac{1}{x}(a_{n-2} + \frac{a_{n-3}}{x^2} + \ldots + \frac{a^0}{x^n})) = -\infty$. This implies that for every $M < 0$ there exists $x_0 < 0$ such that for all $x_1 < x_0$, $\frac{P(x_1)}{x_1^{n-1}} < M < 0$ and therefore $P(x_1) < 0$. Since $P(x_1) > 0$ and $P(x_2) < 0$, from the Intermediate Value Theorem it follows that there exists an x_3 between x_1 and x_2 such that $P(x_3) = 0$. Hence, if $P(x)$ is of odd degree, then it must have a real root. Similarly, when leading coefficient $a_n < 0$, $P(x)$ also has a real root. Thus, if $P(x)$ is of odd degree, then it must have a real root.

Theorem 4.3.4: Let P be a real closed field, then P can be ordered in one and only one way[3].

Proof: In order to prove this Theorem we need to show two things. First, if $a \neq 0 \in P$, then either a or $-a$ is a square. Moreover, these cases are mutually exclusive. So we need to show that either a is a square or $-a$ is a

square. Second, we need to show that the ordering of real closed field P is unique.

Suppose $\gamma \in P$ is not the square of an element in P, then $\sqrt{\gamma}$ is a root of $x^2 - \gamma$ and it follows that $P \subset P(\sqrt{\gamma})$. This implies that $P(\sqrt{\gamma})$ is not formally real. Since $P(\sqrt{\gamma})$ is not formally real,

$$-1 = \sum_{i=1}^{n} \left(\alpha_i \sqrt{\gamma} + \beta_i\right)^2$$

This implies that $-1 = \sum_{i=1}^{n} \left(\gamma \alpha_i^2 + 2\alpha_i \beta_i \sqrt{\gamma} + \beta_i^2\right)$ with $\alpha_i, \beta_i \in P$ or $-1 = \gamma \sum_{i=1}^{n} \alpha_i^2 + \sum_{i=1}^{n} 2\alpha_i \beta_i \sqrt{\gamma} + \sum_{i=1}^{n} \beta_i^2$. Since by hypothesis $\sqrt{\gamma} \notin P$, we get $2\sum_{i=1}^{n} \alpha_i \beta_i = 0$ and therefore

$$-1 = \gamma \sum_{i=1}^{n} \alpha_i^2 + \sum_{i=1}^{n} \beta_i^2 .$$

However, we know that P is formally real. This implies that $\gamma \sum_{i=1}^{n} \alpha_i^2 \neq 0$. Consequently, $\gamma \sum_{i=1}^{n} \alpha_i^2 = -1 - \sum_{i=1}^{n} \beta_i^2$

and

$$\gamma = \frac{-1 - \sum_{i=1}^{n} \beta_i^2}{\sum_{i=1}^{n} \alpha_i^2} \qquad (5)$$

Thus, $\gamma \in P$ is not the square of an element in P, because

γ cannot be expressed as a sum of squares in P; for otherwise -1 is a sum of two squares in P. Equivalently, by contrapositive, if $-\gamma$ is a sum of two squares then $-\gamma \in P$ is the square of an element in P. So, we need to show that $-\gamma$ is a sum of two squares. Now, from (5) we obtain

$$-\gamma = \frac{1 + \sum_{i=1}^{n} \beta_i^2}{\sum_{i=1}^{n} \alpha_i^2}$$

But since $1^2 + \sum_{i=1}^{n} \beta_i^2$ is a sum of squares, both the numerator and denominator are sums of squares. Hence, both numerator and denominator are squares. This implies that $-\gamma = c^2$ for some $c \in P$.

Now, we need to prove the uniqueness of ordering on P. Let '<' be an ordering on P defined by $0 < a$ if and only if $a = b^2 \ (b \neq 0)$. Suppose there exists any other ordering '<<' on P. In order to show uniqueness of the ordering on P, we need to show two things.

i) Assume $0 << a$. By proof either a or $-a$ is a square. But squares are positive. This implies that we cannot have $-a$ as a square. Therefore, we

must have $a = b^2 \ (b \neq 0)$. Hence, by the definition of the ordering of '<', we get $0 < a$.

ii) Now suppose that $0 < a$. By definition of '<' this implies that $a = b^2 \ (b \neq 0)$.

a) If $b >> 0$, then by definition 4.2.6 $b^2 >> 0$. Therefore, we get $a = b^2 >> 0$. Hence, we conclude that $a >> 0$.

b) If $-b >> 0$, then by definition 4.2.6 $(-b)^2 >> 0$. Therefore, we get $a = (-b)^2 >> 0$. Hence, we conclude that $a >> 0$.

Therefore, if P is a real closed field, then it can be ordered in one and only one way.

Theorem 4.3.5: In a real closed field (*r.c.f.*) P every polynomial of odd degree has at least one root in P. Let P be a *r.c.f.*, then every $f(x) \in P[x]$ of odd degree has at least one root[5].

Proof: Let $f(x) = a_n x^n + \cdots + a_1 x + a_0 \in P[x]$, n is odd, $a_n \neq 0$, and P be a *r.c.f.*

Assume that all odd polynomials of odd degree less than n have at least one root.

Either $f(x)$ is reducible or irreducible in $P[x]$.

Case 1) Suppose $f(x)$ is reducible in $P[x]$, then this implies that $f(x) = f_1(x) \cdot q(x)$ where $f_1(x)$ is irreducible and $q(x)$ may or may not be reducible in $P[x]$. If $q(x)$ is reducible we can apply induction on $q(x)$ and reduce $f(x)$ to $f(x) = f_1(x) \cdots f_m(x)$ where each $f_i(x)$ is irreducible in $P[x]$. Since $f(x)$ is of odd degree, one of $f_i(x) \in P[x]$ is of odd degree and, by induction on degree of f, $f(x)$ has a root in P. Hence, $f(x)$ has a root in P.

Case 2) Suppose $f(x)$ is irreducible in $P[x]$. Then $f(x)$ has a root in an extension field $P(\alpha)$. This then implies that $f(x) = c \cdot (x-\alpha) \cdot f_1(x) \cdots f_k(x)$ with $f_i(x)$ irreducible for $1 \le i \le k \le n$ in $P(\alpha)[x]$. Since P is a *r.c.f.*, by definition 4.2.2 $P(\alpha)$, a proper extension of P, *is* not formally real. So, then we can express -1 as

$$-1 = \sum_{i=1}^{r} (h_i(\alpha))^2 \quad \text{for some } r \in N \tag{6}$$

where $h_i(\alpha) = c_{0,i} + c_{1,i}\alpha + c_{2,i}\alpha^2 + \cdots + c_{m,i}\alpha^m, c_i \in P$ with $m < n$. The degree of $h_i(x)$ is at most $n-1$ since $\deg f(x) = n$, and α is a symbolically adjoined root of $f(x)$. Now, applying the division algorithm we get

$$\sum_{i=1}^{r}(h_i(x))^2 = f(x)\cdot g(x) + r(x) \qquad (7)$$

where $\deg r(x) < \deg f(x)$ and $\deg f(x) = n$.

Evaluating (7) at α, we get

$-1 = \sum(h_i(\alpha))^2 = 0\cdot g(\alpha) + r(\alpha) \in P(\alpha)$, because $f(\alpha) = 0$. This implies that $r(\alpha) = -1$. Now, $\forall\, h_i(x) \in P(\alpha)[x]$ where α is symbolically adjoined root of $f(x)$, we have $r(x) = -1$. From equality (2) we get $\sum(h_i(x))^2 = f(x)\cdot g(x) + (-1)$. This gives us an identity

$$-1 = \sum(h_i(x))^2 + (-1)f(x)\cdot g(x) \qquad (8)$$

We know that $\sum(h_i(x))^2$ is of even degree greater than one. This means that the leading coefficients of $(h_i(x))^2$ are squares which implies that leading coefficients of $(h_i(x))^2$ can not cancel out in addition. Moreover, $\deg h_i(x)$ is less than or equal to $n-1$. This implies that $\deg \sum(h_i(x))^2 \le 2n-2$, since the leading coefficient (x^{n-1}) raised to the second power gives us $(x^{n-1})^2 = x^{2n-2}$. Since $\deg \sum(h_i(x))^2 \le 2n-2$ is even, from equality (7) we get that $\deg f(x)\cdot g(x) \le 2n-2$ also must be even. Moreover, $\deg f(x) = n$ is odd. This implies that $\deg g(x) \le n-2$ also

is odd. By induction on n there exists $a \in P$ such that $g(a) = 0$. Using identity (8) we get $-1 = \sum (h_i(a))^2$ since a is a root of $g(x)$. So this means that P is not formally real, but by assumption P is formally real. This gives us a contradiction. So then $f(x)$ is reducible in $P[x]$, and it has a root in P. Thus, every $f(x) \in P[x]$ of odd degree has at least one root in P.

Theorem 4.3.6: If F is of characteristic zero and if a, b are algebraic over F, then there exists $c \in F(a,b)$ such that $F(a,b) = F(c)$ [3].

Proof: Let $f(x), g(x) \in F[x]$ both be irreducible and let $f(a) = 0, \quad g(b) = 0$ where $a, b \notin F$. This implies that there exists an extension field of F in which $f(x)$ and $g(x)$ can be factored completely. Let $a = a_1, a_2, \ldots, a_n$ be roots of $f(x)$ and $b = b_1, b_2, \ldots, b_n$ be roots of $g(x)$. Since characteristic of F is zero, the roots of $f(x)$ and $g(x)$ are all distinct. Since, we have finitely many distinct roots for $f(x)$ and $g(x)$, so for $k \neq 1$ we must have $b_k \neq b_1$.

This implies that the equation $a_i + xb_k = a_1 + xb_1$ has at most one root x in $F(a,b)$ for every i and every $k \neq 1$. Let $\gamma \in F$ be different from the roots of each equation $a_i + xb_k = a_1 + xb_1$. This gives us an equation $a_i + \gamma b_k \neq a_1 + \gamma b_1$ for every i and every $k \neq 1$. Let $c = a_1 + \gamma b_1 = a + \gamma b$, then we have $c \in F(a,b)$. To prove $F(a,b) = F(c)$, we still need to show that $F(a,b) \subseteq F(c)$.

Let's take $c = a + \gamma b$ and solve it for a. This gives us $a = c - \gamma b$. We have $g(b) = 0$ and $f(a) = f(c - \gamma b) = 0$ with coefficients of $f(x)$ in $F(c)$. Since $c - \gamma b_k \neq a_i$ for $k \neq 1$ and $i = 1, \ldots, n$, the polynomials $g(x)$ and $f(c - \gamma x)$ have only the root b in common. This implies that $f(c - \gamma b_k) \neq 0$ for $k \neq 1$. Moreover, polynomials $g(x)$ and $f(c - \gamma x)$ have only one linear factor $x - b$ in common, because b is a simple root of $g(x)$. So we need to show that $\gcd\big(f(c - \gamma x), g(x)\big) = x - b$.

Suppose that $\gcd(f(c-\gamma x), g(x))$ has a factor other than $x-b$. Recall that $g(b_j) = 0$ for $b_j \neq b$. This implies that $f(c-\gamma b_j) \neq 0$ since $c-\gamma b_j$ for $j \neq 1$ avoids all roots a_i of $f(x)$. Thus, $x-b_j$ with $j \neq 1$ does not divide the gcd. Also $(x-b)^2$ does not divide $g(x)$ since $b = b_1, b_2, \ldots, b_n$ are all distinct roots. This implies that $(x-b)^2$ does not divide the gcd. Thus, $\gcd(f(c-\gamma x), g(x)) = x-b$ over some extension of F. Since $\deg(x-b) = 1$ and $\gcd(f(c-\gamma x), g(x)) \in F(c)[x]$, we have $(x-b) \in F(c)[x]$. This implies that $b \in F(c)$. So we have $\gamma \in F$, $c \in F(c)$, and $b \in F(c)$. Since $a = c-\gamma b$, we conclude that $a \in F(c)$. Thus, $a, b \in F(c)$ implies that $F(a, b) \subseteq F(c)$. Finally, $F(c) \subseteq F(a, b)$ and $F(a, b) \subseteq F(c)$ implies that $F(a, b) = F(c)$.

4.4 Fundamental Theorem of Algebra

Theorem 4.4.1: If, in an ordered field K, every positive element possesses a square root and every polynomial of odd degree has at least one root, then the field obtained by adjoining i, $K(i)$, is algebraically closed[5].

Proof: Let K be an ordered field and every $a > 0 \in K$ possess a square root. Assume that every $f_i(x) \in K[x]$ of odd degree has at least one root in K.

We need to show that $K(i)$ is algebraically closed. Let $a \in K$, then $\sqrt{a} \in K(i)\ \forall a \in K$. By Theorem 4.3.1 any $f(x) \in K[x]$ with $\deg f(x) = 2$ is solvable in $K(i)$. In order to show algebraic closure of $K(i)$, it suffices to show that every $f(x) \in K[x]$, $f(x)$ irreducible, has a root in $K(i)$.

Let $f(x) \in K[x]$ be a polynomial with no double roots and $\deg f(x) = n$ such that $n = 2^m \cdot q$ where q is odd. This implies that $f(x) = a_n x^n + \ldots + a_1 x + a_0$, $a_i \in K[x]$.

By induction on m we can assume that every f in $K[x]$ whose degree is divisible by 2^{m-1} but not by 2^m has a root in $K(i)$. So if $m = 1$, then $\deg f(x) = q$. Now, q is odd and, by hypothesis, there exists a root of $f(x) \in K$, which implies that there exists root of $f(x) \in K(i)$.

Now, suppose every polynomial $f(x)$ of degree $2^{m-1} \cdot q^1$ where q^1 is odd has a root in $K(i)$. Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be the roots of $f(x)$ in an extension of K. Choose $c \in K$ such that $\alpha_j \alpha_k + c(\alpha_j + \alpha_k) = d_{jk}$ are all different expressions for $1 \leq j < k \leq n$ by reasoning similar to that given in

Theorem 4.3.6. Choosing two $\alpha's$ out of n $\alpha's$ gives us $\binom{n}{2}$ different expressions for d_{jk}. But when $n = 2^m \cdot q$, $\frac{n(n-1)}{2} =$

$= \frac{2^m q(2^m q - 1)}{2} = 2^{m-1} q(2^m q - 1)$. This implies that 2^{m-1} divides $\frac{n(n-1)}{2}$ but $2^m \nmid \frac{n(n-1)}{2}$. Consider the polynomial $p(x) = \pi_{jk}(x - d_{jk})$ of degree $\frac{n(n-1)}{2}$. As shown above 2^{m-1} divides $p(x)$, but $2^m \nmid p(x)$. Therefore, by induction hypothesis there exists at least one root $d_{jk} \in K(i)$.

For ease in notation suppose $d_{12} = \alpha_1\alpha_2 + c(\alpha_1 + \alpha_2) \in K(i)$ be one of the expressions with α_1, α_2 roots of $f(x)$. By Theorem 4.3.6 $K(\alpha_1\alpha_2, \alpha_1 + \alpha_2) = K(\alpha_1\alpha_2 + c(\alpha_1 + \alpha_2)) \subseteq K(i)$. Thus, $\alpha_1 + \alpha_2 \in K(i)$ implies that $(\alpha_1 + \alpha_2)^2 \in K(i)$. But since $\alpha_1\alpha_2 \in K(i)$ and $(\alpha_1 + \alpha_2)^2 = \alpha_1^2 + 2\alpha_1\alpha_2 + \alpha_2^2$, $\alpha_1^2 + \alpha_2^2 \in K(i)$. Now, consider $(\alpha_1 - \alpha_2)^2$, $(\alpha_1 - \alpha_2)^2 = \alpha_1^2 - 2\alpha_1\alpha_2 + \alpha_2^2 =$

$= \alpha_1^2 + \alpha_2^2 - 2\alpha_1\alpha_2$. But $\alpha_1^2 + \alpha_2^2 \in K(i)$ and $\alpha_1\alpha_2 \in K(i)$ implies that $(\alpha_1 - \alpha_2)^2 \in K(i)$ and consequently by Theorem 4.3.2 $\alpha_1 - \alpha_2 \in K(i)$. So we have $\alpha_1 + \alpha_2 \in K(i)$ and $\alpha_1 - \alpha_2 \in K(i)$ which implies that $\alpha_1 + \alpha_2 + \alpha_1 - \alpha_2 = 2\alpha_1 \in K(i)$ and $\alpha_1 \in K(i)$. Since

$\alpha_1 \in K(i)$, $-\alpha_1$ is also in $K(i)$. Thus, $(\alpha_1 + \alpha_2) + (-\alpha_1) = \alpha_2$, which implies that $\alpha_2 \in K(i)$ and consequently all roots of $f(x)$ are in $K(i)$. Hence, $K(i)$ is algebraically closed.

4.5 Summary

The proof of the FTA without the use of Galois Theory requires a solid background in Abstract Algebra. The argument in the proof in this chapter goes as follows: pick an arbitrary function $f(x) \in K[x]$ where $f(x)$ is irreducible, and show that all the roots of $f(x)$ are in $K(i)$.

CHAPTER FIVE

GALOIS THEORY PROOF OF THE FUNDAMENTAL THEOREM OF ALGEBRA

5.1 Introduction

The Galois Theory proof of the FTA requires a strong background in Group Theory and Galois Theory. Section 5.2 provides the reader with definitions needed to prove the Theorem in this chapter. Then we provide statements of theorems and a lemma used to prove the FTA using Galois Theory approach. Finally, section 5.4 presents the proof of the Theorem.

5.2 Definitions and Notation

5.2.1 Definition

If α is a root of $f(x)$, then α has multiplicity $m \geq 1$ if $f(x) = (x-\alpha)^m g(x)$ where $g(\alpha) \neq 0$. If $m = 1$, then α is a simple root otherwise it is a multiple root.

5.2.2 Definition

A Galois extension of F is a finite separable splitting field over F.

5.2.3 Definition

Let G be a finite group with $|G| = p^m \alpha$, with p being a prime and with $(p, \alpha) = 1$. Then a p-Sylow subgroup is a subgroup of order p^m.

5.2.4 Definition

Let K be a finite extension of F and $\alpha \in K$. Then α is separable over F if α is a simple root of $irr(\alpha, F)$. K is a separable extension if every $\alpha \in K$ is separable over F.

5.2.5 Definition

F' is a splitting field for $f(x)$ over F if F' is the smallest extension field of F in which $f(x)$ splits completely.

5.2.6 Definition

If p is a prime, then a p-group is a group G where every element has order a power of p. If G is finite, this implies that $|G| = p^n$ for some n.

5.2.7 Definition

Let K be a Galois extension of F. Then the group of automorphisms of K that fix F is called the Galois group of K over F, denoted by $Gal(K/F)$. If H is a subgroup of

$Gal(K/F)$, we let K^H denote the elements of K fixed by H.

5.2.8 Notation

$irr(\alpha, F)$ will denote an irreducible polynomial with coefficients in F and root α.

5.3 Theorems

Theorem 5.3.1 (Fundamental Theorem of Galois Theory): Let K be Galois extension of F with Galois group $G = Gal(K/F)$. For each intermediate field E let $\tau(E)$ be the subgroup of G fixing E. Then:

i) τ is a bijection between intermediate fields containing F and subgroups of G.

ii) E is Galois over F if and only if $\tau(E) \lhd G$ where $\tau(E)$ is normal in G.

iii) $|G| = |K:F|$.

iv) $|E:F| = |G:\tau(E)|$. That is, the degree of an intermediate field over the ground field is the index of the corresponding subgroup in the Galois group[6].

Theorem 5.3.2 (Sylow Theorem): Let G be a finite group of order $p^m \alpha$ with p a prime and with $(p, \alpha) = 1$, then G has a p-Sylow subgroup[6].

5.4 Lemma

Lemma 5.4.1: If G is a finite p-group of order p^n, then G has a subgroup of order p^{n-1} and hence of index p [6].

5.5 Fundamental Theorem of Algebra

Theorem 5.5.1: The complex number field C is algebraically closed; i.e., any non-constant complex polynomial has a root in C [6].

Proof: Let $f(x) \in C[x]$ and $f(x)$ be non-constant complex polynomial. There exists splitting field K for $f(x)$ over C. Since K is a finite extension of C and C is a finite extension of R, K must be a finite extension of R. So, K is a finite, separable (characteristic zero) splitting field over C and, by the Definition 5.2.2, it is a Galois extension of R. In order to prove the FTA, we will show that any nontrivial Galois extension of C must be C itself.

For the K above, any finite extension of R with $|K:R| = 2^m q$ where $(2, q) = 1$ and $K \neq R$. Suppose $m = 0$, then $|K:R| = q$. This implies that K is an odd-degree extension of R. By Theorem 4.3.6 K is a simple extension and therefore $K = R(\alpha)$ where $irr(\alpha, R)$ is of odd degree. However, by Theorem 4.3.3 odd-degree real polynomials always have a real root. Thus, $irr(\alpha, R)$ is of degree one, since $irr(\alpha, R)$ is irreducible having a real root. But this implies that $\alpha \in R$ and $K = R$, which is a contradiction. Therefore, if K is a nontrivial extension of R with $|K:R| = 2^m q$ where $(2, q) = 1$, then $m > 0$.

Now suppose that K is a 2^{nd} degree extension of C, which means that $m = 1$ and $q = 1$. By Theorem 4.3.6 this implies that $K = C(\alpha)$ where $\deg irr(\alpha, C) = 2$. However, by Theorem 4.3.1 complex quadratic polynomials always have roots in C, which implies that we have $\deg irr(\alpha, C) = 2$. But this is a contradiction. Therefore, C has no two degree extensions.

Now let K be a Galois extension of C. Since C is finite extension over R, K is also Galois extension over R. Suppose $|K:R| = 2^m q$ with $(2, q) = 1$ and $m > 1$. Let

$G = Gal(K/R)$ be the Galois group. This implies that $|G| = |K:R| = 2^m q$ with $(2, q) = 1$ and $m > 1$. Since 2 is a prime number and $(2, q) = 1$, by Lemma 5.4.1 G has a 2-Sylow subgroup of order 2^m and index q. By Theorem 5.3.1(iv) there exists an intermediate field E with $|K:E| = 2^m$ and $|E:R| = q$, but then E is an odd-degree finite extension of R. From the argument above this means that $q = 1$ and hence $E = R$. Therefore, $|K:R| = 2^m$ and $|G| = 2^m$. Since $|K:R| = 2^m$, we must have $|K:C| = 2^{m-1}$.

Now suppose $G_1 = Gal(K/C)$, then $|G_1| = 2^{m-1}$. By the definition 5.2.6 G_1 is a 2-group. Moreover, G_1 is either trivial or a nontrivial 2-group. Suppose that G_1 is nontrivial 2-group, then by the Lemma 5.4.1 there exists a subgroup of order 2^{m-2} and index 2. Then by the Theorem 5.3.1 this implies that there exists an intermediate field E of degree two over C. However, we showed that C has no degree two extensions. So then G_1 must be a trivial 2-group and $|G_1| = 1$. Hence, $|K:C| = 1$ and $K = C$.

5.6 Summary

The Galois Theory proof of the FTA uses many facts stated in this chapter as well as in Chapter Four. The argument of this proof goes as follows. First, pick an arbitrary non-constant function $f(x) \in C[x]$. Then consider an algebraic extension of C and show that any such nontrivial extension of C must be C itself.

CHAPTER SIX

SIMILARITIES AND DIFFERENCES

6.1 Introduction

Among the three proofs of the FTA, we find some similarities and differences. It is interesting that using different mathematical tools--Complex Analysis, non-Galois Theory Algebra, and Galois Theory—we can prove the same concept. Section 6.2 will present will present similarities and differences of those proofs.

6.2 Similarities and Differences

Comparing the three proofs, we find that the Galois Theory proof and the algebraic proof have some similarities. In both of these proofs we picked a polynomial $f(x)$ in a field, and then studied the proper extension of the field associated with the roots of $f(x)$. Even though Theorems 4.4.1 and 5.4.1 have similar approaches to prove the Fundamental Theorem of Algebra, they utilize different tools to prove it.

The three proofs also have some differences. The Complex analysis proof of the Fundamental Theorem of Algebra is different from the other two. The complex

analysis proof is done by contradiction. In this proof, we picked a non-constant analytic function in the complex plane, and showed that the assumption was false.

REFERENCES

[1] http://www-groups.dcs.stand.ac.uk/~history/

HistTopics/Fund_theorem_of_algebra.html

[2] http://physics.rug.ac.be/fysica/Geschiedenis/

HistTopics/Fund_theorem_of_algebra.html

[3] Herstein, I. N. Topics in Algebra, 2nd edition, John Wiley & Sons, New York, 1974.

[4] Brown, James W. & Churchill Ruel V., Complex Variables and Applications, 6th edition, McGraw-Hill, Inc., New York, 1996.

[5] Waerden, Van Der, Algebra, vol. 1, New York, 1970.

[6] Fine, Benjamin & Rosenberger, Gerhard The Fundamental Theorem of Algebra, Springer, New York, 1997.

Little Mathematics Library

L.A.KALUZHNIN

THE FUNDAMENTAL THEOREM OF ARITHMETIC

Mir Publishers · Moscow

ПОПУЛЯРНЫЕ ЛЕКЦИИ ПО МАТЕМАТИКЕ

Л. А. Калужнин

ОСНОВНАЯ ТЕОРЕМА АРИФМЕТИКИ

Издательство «Наука» Москва

LITTLE MATHEMATICS LIBRARY

L.A.Kaluzhnin

THE FUNDAMENTAL THEOREM OF ARITHMETIC

Translated from the Russian
by
Ram S. Wadhwa

MIR PUBLISHERS
MOSCOW

First published 1979

На английском языке.

CONTENTS

INTRODUCTION

It is customary to think that arithmetic precedes algebra, and that it is a more elementary part of mathematics. At school, arithmetic is taught from the first form while algebra only from the fifth. Since a vast majority of people know about mathematics mainly from what they have learnt at school, the idea about the elementariness of arithmetic has taken deep roots. Meanwhile arithmetic, if considered as a study of properties of integers, and of operations upon them, is a difficult and far from elementary section of mathematics. True, in such a generalized sense, this section is rather known as "higher arithmetic" or "theory of numbers" so as to distinguish it from school arithmetic. But these designations do not alter facts. And the fact is that both school arithmetic and higher arithmetic belong to one and the same sphere of knowledge. In my view, it would be very useful if schoolboys from higher classes, having interest in mathematics, enriched the knowledge that they have acquired in lower classes. Actually, such an enrichment is also essential in order to get acquainted with higher arithmetic in future.

This brochure is intended to be of help in this direction.

As a starting point, we shall consider the so-called *fundamental theorem of arithmetic*. This somewhat scientific designation need not be frightening: everybody knows this theorem well and often use it for arithmetical calculations (e. g. while finding the common denominator of fractions), not realizing at the same time that this is an important theorem requiring a careful and detailed proof. We shall explain what it is all about.

Every integer can be expressed as a product of prime numbers. For example,

$$420 = 2 \times 2 \times 3 \times 5 \times 7 \qquad (1)$$

Now, if the number is sufficiently large, then for finding the corresponding factorization, it is necessary sometimes to spend a long time. Nevertheless, we can accomplish this factorization in all cases if we like. But may be, we have been just lucky so far? Are we sure that any arbitrary whole number can be represented as a product of prime numbers? It is actually so, but this fact requires a proof. The first part of the fundamental theorem in fact comprises the statement:

Every whole number can be represented as a product of prime numbers.

The proof of this statement is carried out in this brochure. In fact it is very simple and it would be useful for the reader to work it out independently. The proof of the second part of the theorem is more difficult (it is, however, considered self-evident at school).

Before going on to its formulation, let us once again consider the above example of factorization of the number 420 into prime multipliers. The procedure, well-known from school, also represented schematically thus:

420	2
210	2
105	3
35	5
7	7
1	

actually gives the factorization (1). But may be, there are also other methods of factorization? How to know whether they will also give the same result? Of course, for example, we can try to expand the given number as a product of two smaller numbers (not necessarily prime numbers) and then each of them as a product of smaller numbers and so on until we arrive at numbers which cannot be factorized further (i. e. at prime numbers). However, from the very first step it is clear that such a process is not unique. In fact, for example, for the same number 420, we have

$$420 = 20 \times 21, \qquad 420 = 15 \times 28$$

Thus it is quite natural to ask: are there whole numbers which can be expressed in different ways as products of prime numbers? It turns out that such whole numbers do not exist, and the corresponding statement about the *uniqueness* of factorization of numbers as product of prime multipliers does, actually, constitute the second part of the fundamental theorem:

If some whole number n is expanded in two ways as a product of prime multipliers

$$n = p_1 \cdot p_2 \cdot \ldots \cdot p_k = q_1 \cdot q_2 \cdot \ldots \cdot q_l$$

then these factorizations exactly coincide except for the order of multipliers: both of them contain one and the same number of

multipliers, $k = l$, and every multiplier occurring in the first factorization is repeated the same number of times in the second [1].

We shall give quite a detailed proof of this statement. It is, as we pointed out earlier, much more complicated than the proof of the first statement. This complication is not accidental but is connected with the fundamental properties of the arithmetic of whole numbers. It turns out that apart from this primary arithmetic, there are in existence, and of great use, many other 'arithmetics'. In some of the arithmetics, the statements of the fundamental theorem are valid, in others – not, more so, the statement about uniqueness of expansion is not fulfilled. We shall give examples of arithmetics of the first as well as second kind.

We shall consider in greater detail one arithmetic of the first kind – the arithmetic of complex whole numbers, or as they are often called, Gaussian whole numbers. We may mention, by the way, that we shall sometimes call the ordinary whole numbers as *rational* whole numbers (so as not to confuse them with Gaussian whole numbers). However, at places where they don't lead to confusion, we shall be speaking of just whole numbers, meaning thereby rational whole numbers. In the arithmetic of Gaussian whole numbers, the theorem is likewise applicable and this applicability carries along with it a whole lot of interesting and far from obvious properties of rational whole numbers.

At the end of this brochure, we shall give an example of the arithmetic in which the fundamental theorem is not applicable: true, the numbers being considered there may be expressed as a product of prime multipliers, but it may turn out that the prime numbers occurring in the two expansions are different. We shall not investigate this arithmetic in greater detail: this would require the introduction of a number of new concepts and a study of their properties, which is possible only in the framework of a serious university course.

For an understanding of our exposition, the reader is not required to possess more knowledge than is imparted by a school curriculum in mathematics, but for one important exception. While proving the theorems, we shall be making extensive use of the method of mathematical induction [2]. This method in mathematics is

---

[1] If we consider any artitrary whole number (positive or negative), then by the uniqueness of factorization into prime multipliers it should be understood that two factorizations $n = p_1 \cdot p_2 \ldots p_k$ and $n = q_1 \cdot q_2 \ldots q_l$ may differ not only in the order of the multipliers, but also in signs of corresponding multipliers; see § 1 – formulation of the fundamental theorem.

[2] Sometimes it is also called "complete induction".

unfortunately rarely taught at school. A detailed substantiation and elucidation of this theory would lead us too far from our topic. To the readers, who would like to get acquainted with the method of proof by induction, we may recommend the brochure by I. S. Sominsky "The Method of Mathematical Induction" (Mir Publishers, 1975) which was published in the series "Little Mathematics Library", or the book "On Mathematical Induction" in the series "Popular Lectures on Mathematics" (Nauka, 1967) by I. S. Sominsky, L. I. Golovina and I. M. Yaglom.

At the end of this brochure, we shall mention some books which explain in a comprehensible form the theoretical-numerical facts that are more or less closely linked with the question being investigated here.

§ 1. The Fundamental Theorem of Arithmetic. Proof of the First Part

We shall give a single formulation for the statements given in the introduction, i. e. formulate completely the fundamental theorem of arithmetic.

Any non-zero whole number may be represented in the form of a product of prime numbers; moreover, such a representation is unique except for the order of the multipliers and their signs.

As has already been stated, the above-mentioned theorem contains two statements: first, a statement about the existence of a representation for any number as a product of prime numbers, and second, a statement about the uniqueness of such a representation. We shall prove both these statements. In this paragraph, we shall prove only the first of these. To begin with, we shall make two simple observations:

1. One (1) is, for many reasons, not considered as a prime number in spite of the fact that it cannot be expressed as a product of smaller numbers. Then the question arises: in what way is the above-mentioned theorem valid for integer 1? Or, in other words, in what way is integer 1 represented in the form of a product of prime numbers? Mathematics, in contrast with, say, grammar, does not like exceptions. We shall consider that

$$1 = 1$$

actually is the expansion of integer 1 into a product of prime numbers. Moreover, the number of prime multipliers in the right hand side is equal to zero. This relation reminds the definition of zero order

$a^0 = 1$ (number of multipliers a is equal to zero) and is convenient in many ways. We make a similar agreement for integer -1 also.

2. As a second remark, we shall simply give an example to explain the concept of uniqueness of expansion of a whole number into prime multipliers. Two expansions for number 18

$$18 = 2 \times 3 \times 3$$

and

$$18 = (-3) \times (-2) \times 3$$

are considered indistinguishable.

Proof for the existence of expansion of a rational whole number into a product of prime multipliers. We shall first confine ourselves to the case of positive whole numbers. The possibility of their expansion into prime multipliers is proved by the method of mathematical induction:

(a) For $n = 1$ $1 = 1$ is the required representation: 1 is a product of merely a large number of prime numbers.

(b) Let us suppose that for all positive numbers m, that are less than n, expansion into a product of prime numbers is already established. We shall then go on to show that for number n also such an expansion will occur. If n is a *prime* number, then

$$n = n$$

is the required expansion (one prime multiplier).

Let n be a *complex* number. Then it is a product $n = n_1 \cdot n_2$ of two whole numbers n_1 and n_2 each of which is different from 1 or from n. Consequently, $n_1 < n$ and $n_2 < n$. But then, by the principles of induction, the expansion of numbers n_1 and n_2 as products of prime numbers is already established:

$$n_1 = p_1 \cdot p_2 \ \ldots \ p_r$$
$$n_2 = q_1 \cdot q_2 \ \ldots \ q_s$$

where p_j and q_i are prime numbers. We have $n = p_1 \cdot p_2 \ \ldots \ \ldots \ p_r \cdot q_1 \cdot q_2 \ \ldots \ q_s$, i. e. we have got the required expansion of the number n.

If n is a negative whole number, then $-n$ is a positive number. As has already been proved, $-n$ is expandable into a product of prime numbers. Let

$$-n = p_1 \cdot p_2 \ \ldots \ p_k$$

Then

$$n = (-1)\, p_1 \cdot p_2 \ \ldots \ p_k$$

or, for example, $n = (-p_1) \cdot p_2 \ \ldots \ p_k$ is the required expansion of the number n. This also proves the first part of the theorem. There are many proofs for uniqueness of expansion. The one which we shall deduce is neither the shortest nor the simplest. However, our proof has the advantage that it can be directly generalized into a number of other cases, for example, to the case of polynomials of one variable, and to the case of complex whole numbers. Apart from this, during the course of the proof, we shall obtain a number of important theorems of arithmetic as a sort of by-product.

§ 2. Division with Remainder and Greatest Common Divisor (GCD) of Two Numbers. Proof of the Second Part of the Fundamental Theorem

The statement about the possibility of "division with a remainder" in the case of whole number is the starting point for our consideration. This statement can be precisely formulated as under:

THEOREM 1. *Let a and b be whole numbers and $b \neq 0$. Then there exist whole numbers q and r* [1], *where $0 \leqslant |r| < b$,* such that

$$a = q \cdot b + r \tag{1}$$

The equality $r = 0$ in the equation (1) is equivalent to the fact that number a is divisible by b [2]. We shall denote such a fact in future by $b|a$ – this is an accepted notation in the number theory.

We shall prove the possibility of such a representation. For this, we observe that for every rational number τ a whole number t

[1] The remainder r can be any whole number – positive, negative, or zero.

[2] For two whole numbers a and b, the statements "number a is divisible by number b", "number a is a multiple of number b", "number b is a divisor of number a", or, finally, "number b divides number a" mean one and the same thing; we shall use each one of them.

can be found so that $|\tau - t| < 1$ [1]. Let $\tau = \frac{a}{b}$; a and b being whole numbers. We select a whole number q so that $\left|\frac{a}{b} - q\right| < 1$ and express

$$r = b\left(\frac{a}{b} - q\right) = a - bq$$

Thus r is a whole number $|r| = |b|\left|\frac{a}{b} - q\right| < |b| \times 1 = |b|$ and

$$a = q \cdot b + r,$$

q.e.d. [2]

Theorem (1) allows us to deduce the idea of GCD of two numbers and prove many of its properties.

DEFINITION 1. If a and b are two non-zero whole numbers and if c is a number such that $c|a$ and $c|b$, then c is called a *common divisor* of numbers a and b. We shall note that any two numbers always have common divisors. These are numbers 1 and -1. If no other divisors exist, then numbers a and b are called *mutually prime* numbers. We shall talk about the mutually prime numbers later.

DEFINITION 2. Number d is called the *greatest common divisor* of numbers a and b (GCD), if: (1) d is a common divisor of a and b and (2) d is divisible by any other common divisor of numbers a and b. (Thus, for example, 6 is GCD of numbers 18 and 30, since $6|18$ and $6|30$, and on the other hand, 6 is divisible by all common divisors of these numbers: 1, -1, 2, -2, 3, -3, 6, -6.)

The reader must be aware even from school that a GCD exists for any pair of whole numbers and must also be conversant with the method of its determination. But if we recall and carefully

---

[1] As a matter of fact the nearest whole number to τ differs from it by not more than $\frac{1}{2}$ but we shall not require the precision.

[2] We note that in representation (1), the whole numbers q and r are not determined uniquely. For example for $a = 13$ and $b = 3$ we have $13 = 4 \cdot 3 + 1$ ($q = 4$, $r = 1$) or $13 = 5 \cdot 3 + (-2)$ ($q = 5$, $r = -2$). This is also seen from our proof. In fact if a is not divisible by b, then $\frac{a}{b}$ is a fractional number but then $n < \frac{a}{b} < n + 1$ where n is a whole number. For number q we can choose $q = n$ or $q = n + 1$ which gives two representations of the form (1). Only in case where $b|a$, is the number q singularly represented, $a = q \cdot b$; in this case $r = 0$.

analyze this method we can easily deduce that it makes use of the factorization of numbers a and b into prime multipliers and of the uniqueness of such a factorization. This method is still forbidden to us since we are just going to prove the corresponding theorem.

From our definition (Definition 2) it does not directly follow that, for any two numbers a and b, a GCD always exists. We shall now prove that it is actually so; moreover, this proof shall not make use of factorization of numbers a and b into prime multipliers.

THEOREM 2. *For any pair of whole numbers* $a \neq 0$ *and* $b \neq 0$, *there exists a GCD.*

PROOF. In addition to numbers a and b, we shall consider all the possible numbers of the type $xa + yb$ where x and y are any integers. Numbers of such kind,

$$v = xa + yb \tag{2}$$

are called *linear combinations* of numbers a and b. For example, for $a = 6$, $b = 22$, the linear combination will be numbers 28 $(28 = 1 \cdot 6 + 1 \cdot 22)$, $10 (10 = (-2) \cdot 6 + 1 \cdot 22)$, $-92 (-92 = 3 \cdot 6 + + (-5) \cdot 22)$, etc. Generally, for any given numbers a and b there exists an infinitely large number of their linear combinations. We shall denote the set of such numbers through M. We observe that this set contains, in particular, also the numbers a (for $y = 0$, $x = 1$) and b (for $x = 0$, $y = 1$) as well as number 0 ($x = 0$, $y = 0$). All numbers v from the set M are obviously whole numbers. If v belongs to M then $-v$ also belongs to M (if $v = xa + yb$, then $-v = (-x)a + (-y) \cdot b$). We also notice one more property of numbers v belonging to M, that we shall need at once: all such numbers are divisible by all common divisors of numbers a and b. In fact, if $c|a$ and $c|b$ and say $a = a'c$ and $b = b'c$, then $v = xa + yb = xa'c + yb'c = (xa' + yb')c$, i. e. $c|v$. Now let $d \neq 0$ – the minimum number by taking modulus out of all non-zero numbers in M [1]. We shall prove that d is the GCD for numbers a and b. It satisfies the property of GCD as per definition (2), since all numbers from M possess this

[1] Such number in the set M actually does exist. We notice that in the set M are contained numbers which are not equal to zero (for example a or b) and their moduli are positive integers, i. e. natural numbers. But one of the fundamental properties of natural numbers, usually applied as an axiom (see I.S. Sominsky, "The Method of Mathematical Induction") is that any non-void collection of natural numbers always contains a minimum number.

property. All that is now required is to establish that it also possesses the property (1), i. e. d is a common divisor of numbers a and b. We shall show that $d|a$. Since d belongs to M, it can be expressed in the form $d = sa + tb$ where s and t are suitable integers. We shall divide a by d with remainder, i. e. we shall find such numbers q and r, $r < |d|$, so that

$$a = qd + r$$

But then the remainder r also must belong to set M. Actually,

$$r = a - qd = a - q(sa + tb) = (1 - qs)a + tb$$

We now recall that d by modulus is the minimum number among non-zero numbers of set M and $r < d$. It follows that $r = 0$ and $d|a$. In exactly similar way the divisibility $d|b$ is proved. The theorem is thus proved.

We have established the existence of GCD of two non-zero whole numbers. Apart from that we shall deduce from the proof the following fact which we shall soon require:

THEOREM 3. *GCD of numbers a and b is represented in the form of a linear combination of these numbers.*

The question arises: has the GCD of numbers a and b been singularly determined? The answer is, of course, in the negative: if number d possesses the properties (1) and (2) of the definition of GCD, then $-d$ also possesses these properties. But this exhausts the non-singularity. Actually, let d and d' be two GCDs of numbers a and b. Since d possesses the property (2) and d' – property (1), $d'|d$. But analogously $d|d'$. Thus $\alpha = \frac{d}{d'}$, and $\frac{d'}{d} = \frac{1}{d/d'} = \frac{1}{\alpha}$ are integers. But the only integers whose reciprocals are also integers are numbers 1 and -1. Thus $\alpha = 1$ or $\alpha = -1$, whence $d' = d$ or $d' = -d$. If in the definition of GCD, we require that this number were positive – it sometimes (but not always) is convenient – then it could be said that GCD of *two non-zero integers exists and is singularly determined.*

In future we shall express GCD of numbers a and b through (a, b) as is usually the practice in the literature on number theory.

Let's go over to the question of pairs of mutually prime numbers. We have already come across this concept. Now we shall repeat its definition.

DEFINITION 3. Integers $a \neq 0$ and $b \neq 0$ are called *mutually prime* if their GCD is equal to 1.

In other words, it may be said that mutually prime numbers are such numbers for which the only common divisors are numbers 1 and -1.

From the aforesaid (Theorem 3), it follows that if $(a, b) = 1$, then 1 can be expressed in the form

$$1 = sa + tb \tag{3}$$

with suitable integers s and t. Conversely, if the equality (3) holds for suitable s and t, then a and b are mutually prime. Really (see proof of Theorem 1), $d = (a, b)$ – this is the lowest number by modulus among non-zero numbers of the type $xa + yb$. Consequently, if (3) holds, then $|d| \leqslant 1$ and $d \neq 0$, so $d = \pm 1$.

From this directly follows the most important property of mutually prime numbers.

THEOREM 4. *If $a|bc$ and $(a, b) = 1$, then $a|c$* (this property reads: *if number a divides the product of two numbers and is mutually prime to one of them, then it is a divisor of the other*).

PROOF. Since $(a, b) = 1$, we can find such numbers s and t so that

$$1 = sa + tb \tag{4}$$

Multiplying both sides by c we have

$$c = (sc)a + t(bc)$$

Both items on the right-hand side are divisible by a, consequently c is divisible by a.

The following statement is also useful.

THEOREM 5. *If number a is mutually prime with numbers b and c, then it is mutually prime with the product bc.*

PROOF. Since $(a, b) = 1$, we can find whole numbers s and t satisfying the equality

$$1 = sa + tb$$

Analogously, since $(a, c) = 1$, then

$$1 = ua + vc$$

for suitable u and v. Multiplying these two equations we get

$$1 = (sa + tb)(ua + vc) = sua^2 + savc + tbua + tbvc = (sua + svc + tbu)a + (tv)\cdot(bc)$$

If $m = sua + svc + tbu$ and $n = tv$, then m and n are integers and

$$1 = ma + n(bc)$$

This shows that a and bc are mutually prime.

The statement of the last theorem can be easily extended for an indefinite number of factors.

THEOREM 6. *If a is mutually prime with numbers b_1, b_2, ..., b_k, then a is mutually prime with the product $b_1 \cdot b_2 \ldots b_k$.*

The proof of this theorem is carried out by the method of mathematical induction for k factors.

PROOF *of uniqueness* of factorization of an integer as a product of prime multipliers.

Now, at last, we can prove the second part of the fundamental theorem of arithmetic. For this, we observe that by definition of a prime number, two different prime numbers are mutually prime. The proof of uniqueness of factorization shall be carried out by induction for absolute value of number n.

(a) If $|n| = 1$, then $n = \pm 1$ and

$$1 = 1, \quad -1 = -1$$

i. e. the factorization is unique for numbers 1 and -1.

(b) Let us suppose that the property to be proved is already true for all numbers m for which $|m| < |n|$. Let

$$n = p_1 \cdot p_2 \ldots p_k = q_1 \cdot q_2 \ldots q_l$$

be two factorizations for the number n as products of prime numbers $p_1, p_2 \ldots, p_k$ and $q_1, q_2, \ldots, q_l$ respectively. We state that prime number p_k occurs among prime numbers $q_1, q_2, \ldots, q_l$ (or, may be, is opposite in sign to some one of them). Really, if it is not so, i. e. if $p_k \neq \pm q_i$, $i = 1, 2, \ldots, l$, then p_k would be mutually prime with all the numbers q_i and, consequently, according to Theorem 6, also with their product, i. e. with the number n. But this is impossible since $p_k | n$, i. e. $(p_k, n) = p_k$. Thus p_k is equal to some one of the prime numbers $\pm q_i$. We may assume that $p_k = q_l$ because if it is not so we can obtain such an equality by rearranging the multipliers q_i and then, if at all $p_k = -q_l$, by changing the sign of q_l by changing it in some other q_i also.

Thus we get

$$n = p_1 \cdot p_2 \ldots p_{k-1} \cdot p_k = q_1 \cdot q_2 \ldots q_{l-1} \cdot p_k$$

whence

$$m = \frac{n}{p_k} = p_1 \cdot p_2 \ldots p_{k-1} = q_1 \cdot q_2 \ldots q_{l-1}$$

But $|m| < |n|$ and by assumption of induction, the statement of

theorem for m has already been proved, i. e. $k-1=l-1$, the sequences in $p_1, p_2, \ldots, p_{k-1}$ and $q_1, q_2 \ldots q_{l-1}$ contain, except for the accuracy in signs, the same prime numbers and corresponding prime numbers occur in both the factorizations the same number of times, and since $p_k = q_l$, then it is also valid for sequences $p_1, p_2, \ldots, p_{k-1}, p_k$ and $q_1, q_2, \ldots, q_{l-1}, q_l$, q.e.d.

§ 3. Algorithm of Euclid and Solution of Linear Diophantine Equations with Two Unknowns

According to Theorem 2 two integers a and b have a GCD. We shall now describe a single procedure for determining GCD which was indicated even in the 'Elements of Euclid' and is called "Euclidean Algorithm".

For this we shall assume that

$$|a| \geqslant |b|$$

First step. Let us divide a by b with remainder:

$$a = q_1 \cdot b + r_1, \quad |r_1| < |b| \tag{1}$$

If $r_1 = 0$, then $b|a$ and $(a, b) = b$. If $r_1 \neq 0$, then we take the *Second step.* Let us divide b by r_1:

$$b = q_2 \cdot r_1 + r_2, \quad |r_2| < |r_1| \tag{2}$$

If $r_2 \neq 0$, then we take the *Third step.*

$$r_1 = q_3 \cdot r_2 + r_3, \quad |r_3| < |r_2| \tag{3}$$

and so on. At every step the new remainder is less than the remainder in the previous step

$$|b| > |r_1| > |r_2| > \ldots$$

and at some kth step ($k < |b|$) the remainder becomes equal to zero. *kth step.*

$$r_{k-2} = q_k \cdot r_{k-1} \tag{k}$$

We shall show that the last non-zero remainder r_{k-1} is the required (a, b). Really, we get a chain of equalities:

$$\begin{aligned}
&(1) \quad a - q_1 \cdot b + r_1 \\
&(2) \quad b = q_2 \cdot r_1 + r_2 \\
&(3) \quad r_1 = q_3 \cdot r_2 + r_3 \\
&\qquad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
&(k-1) \quad r_{k-3} = q_{k-1} \cdot r_{k-2} + r_{k-1} \\
&(k) \quad r_{k-2} = q_k \cdot r_{k-1}
\end{aligned}$$

From the last equality we get $r_{k-1} | r_{k-2}$, from the last but one – $r_{k-1} | r_{k-1}$ and $r_{k-1} | r_{k-2}$ and, consequently, $r_{k-1} | r_{k-3}$. From the previous equality we can analogously conclude that $r_{k-1} | r_{k-4}$ and thus going step by step to earlier equations, we conclude that $\ldots$, $r_{k-1} | r_2$, $r_{k-1} | r_1$, $r_{k-1} | b$, $r_{k-1} | a$. We see that r_{k-1} is the common divisor of numbers a and b.

Now let $c | a$ and $c | b$. Then from (1), (2), $\ldots$, $(k-1)$ successively, we get $c | r_1$, $c | r_2$, $\ldots$, $c | r_{k-1}$. Thus r_{k-1} is really the GCD for numbers a and b.

Let us take a numerical example: $a = 858$, $b = 253$. We have

$$
\begin{aligned}
&(1) \quad 858 = 3 \cdot 253 + 99 \\
&(2) \quad 253 = 2 \cdot 99 + 55 \\
&(3) \quad 99 = 1 \cdot 55 + 44 \\
&(4) \quad 55 = 1 \cdot 44 + 11 \\
&(5) \quad 44 = 4 \cdot 11
\end{aligned}
$$

whence $(858, 253) = 11$. Thus, with the help of Euclidean algorithm, GCD of two numbers is determined without factorizing them into prime multipliers.

In Theorem 3 we established that $(a, b) = d$ can be expressed in the form

$$d = s \cdot a + t \cdot b$$

but in the proof there was no indication as to how the corresponding s and t can be found. With the help of Euclidean algorithm, this problem is very easily solved. We won't describe the procedure for a general case, but shall explain it for the already solved numerical example.

So, we have to find whole numbers s and t such that

$$11 = s \cdot 858 + t \cdot 253$$

From (4), (3), (2), (1) successively, we get

$$
\begin{aligned}
11 &= 55 + (-1) \cdot 44 \\
44 &= 99 + (-1) \cdot 55 \\
55 &= 253 + (-2) \cdot 99 \\
99 &= 858 + (-3) \cdot 253
\end{aligned}
$$

Now substituting in the first of these equalities the expression for 44 from the second, then for 55 the expression from the next

equality and so on, we get

$$\begin{aligned} 11 &= 55 + (-1)\cdot(99 + (-1)\cdot 55) \\ &= 2\cdot 55 + (-1)\cdot 99 \\ &= 2\cdot(253 + (-2)\cdot 99) + (-1)\cdot 99 \\ &= 2\cdot 253 + (-5)\cdot 99 \\ &= 2\cdot 253 + (-5)\cdot(858 + (-3)\cdot 253) \\ &= (-5)\cdot 858 + 17\cdot 253 \end{aligned}$$

Finally: $s = -5$, $t = 17$.

The reader can easily make out how this algorithm can be used in a general case. The equalities occurring in the Euclidean algorithm while finding the GCD of numbers a and b allow us to solve equations of the type

$$d = xa + yb$$

(where $d = (a, b)$).

In general, the equation of the type

$$xa + yb = c$$

where a, b, c are the given integers for which one seeks solution x, y in integers, is called a *linear diophantine equation* with two unknowns. It is called linear since the unknowns x and y occur in it in the *first* order. The term "diophantine" [1] indicates that the coefficients of the equation are *integers* and the required solution are also *integers*.

We observe that we have really learnt how to solve the linear diophantine equations of the type

$$xa + yb = c \qquad \text{(I)}$$

But we must discuss the question about all the solutions of the equation (I) in greater detail. We shall notice first that not every equation of this type has a solution. Actually, if equation (I) does have a solution in integers, say $x = x_0$ and $y = y_0$: $c = x_0a + y_0b$, and if $d = (a, b)$, then, since $d|a$, $d|b$, d divides both terms on the right-hand side and, consequently, also divides c. From this we draw the following conclusion:

In order that a solution in terms of integers of equation (I) *may exist, it is necessary that the right-hand side of the equation is divisible by the greatest common divisor of the numbers a and b.*

[1] Named after the ancient Greek mathematician Diophantos (around 250 B.C.) who investigated equations for integers in his book "Arithmetica". At the end of our exposition we shall stop for a while on the quadratic *diophantine equations.*

For example, the equation

$$9x + 15y = 7$$

does not have a solution, since 7 is not divisible by $3 = (9, 15)$. On the contrary, if $d|c$, then the equation (I) does have a solution in terms of integers and we even know how to find such a solution. Actually let $c = c'd$, and let s and t are such integers (they can be found out with the help of Euclidean algorithm) that

$$d = as + bt$$

Then

$$c = c'd = a(sc') + b(tc')$$

i. e. $x_0 = sc'$, $y_0 = tc'$ are the solutions of the equation (I).

Let us solve, for example, the diophantine equation

$$33 = 858x + 253y \qquad \text{(II)}$$

We have already shown that

$$11 = 858 \cdot (-5) + 253 \cdot 17$$

Multiplying this equality by 3, we get

$$33 = 858 \cdot (-15) + 253 \cdot 51$$

Thus $x = -15$, $y = 51$ are the solutions of the equation (II). It should not be thought that the desired solution is unique. Generally, it turns out that *if a diophantine equation of the type* (I) *does have a solution, then it has an infinite number of solutions.* We shall now study this question in greater detail: we shall prove the formulated statement and find a general form for all possible solutions of the equation (I). Let us begin with the elucidation of the general form. Let us suppose that we already know that, in addition to the solution in terms of integers x_0, y_0, the equation (I) also has the solution x_1, y_1, we have

$$c = ax_0 + by_0$$
$$c = ax_1 + by_1$$

Subtracting the second equality from the first, we get

$$a(x_0 - x_1) + b(y_0 - y_1) = 0$$

or

$$a(x_0 - x_1) = b(y_1 - y_0) \qquad \text{(III)}$$

If $d = (a, b)$, then we put $a' = a/d$, $b' = b/d$, i. e.

$$a = a'd$$
$$b = b'd$$

where a' and b' are mutually prime numbers. Dividing the equality III by d, we arrive at the equality

$$a'(x_0 - x_1) = b'(y_1 - y_0)$$

But then since a', b' are mutually prime, $a' | (y_1 - y_0)$ and, analogously, $b' | (x_0 - x_1)$. Substituting

$$y_1 - y_0 = a'k_1$$
$$x_0 - x_1 = b'k_2$$

we get $a'b'k_1 = a'b'k_2$, whence $k_1 = k_2 = k$. Thus, finally

$$y_1 = y_0 + a'k = y_0 + \frac{a}{d}k \qquad \text{(IV)}$$

$$x_1 = x_0 - b'k = x_0 - \frac{b}{d}k \qquad \text{(V)}$$

where k is some integer. Conversely, it is easy to check that if x_0, y_0 is the solution of equation (I), then all pairs of numbers IV, V for any integer k give solution to the equation (I). Actually,

$$ax_1 + by_1 = a\left(x_0 - \frac{b}{d}k\right) + b\left(y_0 + \frac{a}{d}k\right)$$
$$= ax_0 + by_0 + \left(-\frac{ab}{d}k + \frac{ab}{d}k\right)$$
$$= c + 0 = c$$

Thus, *if* x_0, y_0, *are solutions of equation* (I), *then all numbers of the type* $x_0 - \frac{b}{d}k$, $y_0 + \frac{a}{d}k$ *are also solutions* (it means that for every case, there are infinite solutions — one for every k) *and there are no other solutions.*

§ 4. Gaussian Numbers and Gaussian Whole Numbers

The natural generalization of rational whole numbers is the complex whole numbers or, as they are usually called, "Gaussian whole numbers", after the great German mathematician K.F. Gauss, who first studied them in detail.

DEFINITION 4. A complex number is called "Gaussian whole number" if its real and imaginary parts are essentially rational whole numbers. In other words, they are complex numbers of the form α

$$\alpha = a + bi \tag{1}$$

where a and b are whole (rational) numbers. In addition to the Gaussian whole numbers, we shall also need (simple) *Gaussian numbers*, i. e. complex numbers, whose real and imaginary parts are rational numbers.

The relation between the field of Gaussian numbers and Gaussian whole numbers is analogous to the relation between rational numbers and rational whole numbers. More precisely we mean the following statement which we shall frequently use without special reservation, and which the reader can easily verify directly.

I. *Sum, difference and product of two whole Gaussian numbers are also Gaussian whole numbers* (this property is expressed in short by saying that Gaussian whole numbers form a *ring*).

II. *Sum, difference, product and quotient* (in case the divisor is not equal to zero) *of two Gaussian numbers are also Gaussian numbers.* (This property is expressed shortly as: Gaussian numbers form a *field*.)

III. *Quotient of two Gaussian whole numbers is a Gaussian number and, conversely, every Gaussian number can be represented as a quotient of two Gaussian whole numbers.*

The last statement requires a little explanation. Let $\alpha = a + bi$ and $\beta = c + di$ are Gaussian whole numbers (i. e. a, b, c, d are whole rational numbers) and let $\beta \neq 0$. We shall show that $\gamma = \alpha/\beta$ – Gaussian number. Actually

$$\gamma = \frac{a + bi}{c + di} = \frac{(a + bi)(c - di)}{(c + di)(c - di)}$$

$$= \frac{ac + dd - adi + bci}{c^2 + d^2}$$

$$= \frac{ac + bd}{c^2 + d^2} + \frac{bc - ad}{c^2 + d^2} i$$

Numbers $\frac{ac + bd}{c^2 + d^2}$ and $\frac{bc - ad}{c^2 + d^2}$ – real and imaginary parts of the number γ are rational and, consequently, γ is a Gaussian number.

We observe finally that obviously any rational number is Gaussian

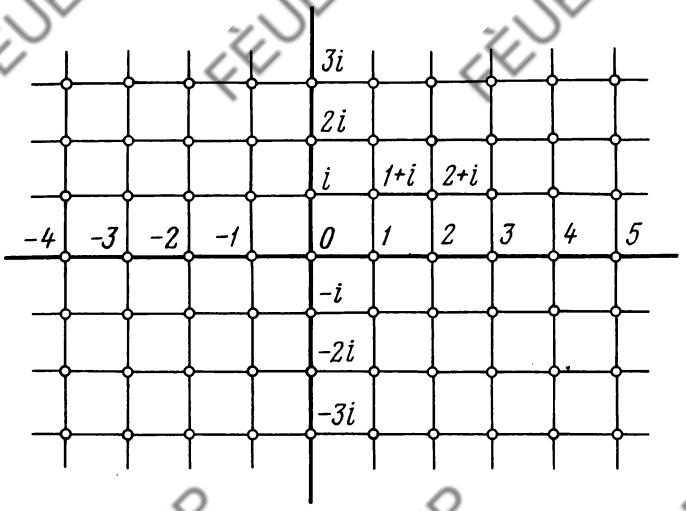


Fig. 1

(imaginary part is equal to zero) and that every rational whole number is a Gaussian whole number.

For future, it will be useful to have an idea about the arrangement of Gaussian whole numbers on a complex plane. By definition itself, the Gaussian whole numbers are represented by points with integral coordinates (Fig. 1). They are located on top of the mesh of squares with sides equal to unity, covering the complex surface.

From the theory of complex numbers we shall need the ideas of norm and modulus of a complex number. We remind that *norm* of a complex number $\alpha = x + yi$ is the non-negative real number $N(\alpha) = x^2 + y^2$, the *modulus* of the complex number α denoted by $|\alpha|$ is the real number $\sqrt{x^2 + y^2}$. Geometrically, modulus of a complex number is the distance of the corresponding point on the complex surface from the origin of coordinates. Norm $N(\alpha)$ of a number α is represented as the product $N(\alpha) = \alpha \cdot \bar{\alpha}$, where $\bar{\alpha}$ is the complex conjugate $x - iy$ of number α. *The property of multiplicability* of norm is also supposed to be a familiar property, i. e.

$$N(\alpha \cdot \beta) = N(\alpha) \cdot N(\beta) \qquad (2)$$

We shall at once note that if α is a Gaussian number then $N(\alpha)$

is a non-negative rational number and even if α is a Gaussian whole number, then $N(\alpha)$ is a non-negative whole number [1].

However, not every positive rational whole number is a norm of a Gaussian whole number. In fact we shall now prove the following theorem.

THEOREM 7. *A positive rational whole number c is norm of some Gaussian whole number if and only if the number c, can be represented in the form of sum of squares of two integers.*

PROOF. If $\alpha = a + bi$ is a Gaussian whole number, then $N(\alpha) = a^2 + b^2$ is the sum of squares of whole numbers a and b. Conversely, if $c = x^2 + y^2$ where x and y are rational whole numbers, then $c = N(x + yi)$ where $x + yi$ is a Gaussian whole number. The theorem is thus proved.

It is not difficult to show that not every positive whole number can be represented as a sum of two squares. We shall show, for example, that a positive odd integer t, which can be represented as a sum of two squares of integers, gives a remainder equal to 1 upon division by 4, i. e. is a number of the type $t = 4k + 1$. Actually let $t = x^2 + y^2$, then one of the numbers, say, x, must be even, the other y – odd. Let $x = 2m$ and $y = 2n + 1$. Then $x^2 = 4m^2$ and $y^2 = 4(n^2 + n) + 1$ and, finally, $t = 4(m^2 + n^2 + n) + 1$, which proves our statement. In this way, numbers 7, 11, 15 and others which cannot be represented in the form of a sum of two squares are, consequently, not norms of Gaussian numbers.

We shall explain the question, precisely which whole numbers can be represented in the form of sum of two squares or, in other words, which numbers are the norms of Gaussian whole numbers, after studying the arithmetic of Gaussian whole numbers. We shall now go on to a study of this arithmetic.

As in the domain (ring) of rational whole numbers, so also in the domain of Gaussian whole numbers, the question of divisibility is of main interest.

We shall say that a Gaussian whole number α divides a Gaussian whole number β and denote this fact as $\alpha \mid \beta$ – if for some Gaussian whole number γ, the equation

$$\beta = \alpha \cdot \gamma \tag{3}$$

holds. Since from (3) follows $N(\beta) = N(\alpha) \cdot N(\gamma)$, the necessary condition for $\alpha \mid \beta$ is the divisibility $N(\alpha) \mid N(\beta)$ where $N(\alpha)$ and $N(\beta)$ are rational whole numbers.

---

[1] Modulus $|\alpha|$ of a Gaussian number is not necessarily a rational number; therefore, in future we shall mainly use norm instead of modulus.

In case of rational whole numbers, there are only two numbers which divide all integers: $+1$ and -1. In case of Gaussian whole numbers, there are four such numbers: $+1$, -1, $+i$, $-i$. It is easily seen that these four numbers satisfy this property. Actually,

$$\alpha = \alpha \cdot 1$$
$$\alpha = (-\alpha) \cdot (-1)$$
$$\alpha = (-\alpha i) \cdot i$$
$$\alpha = (\alpha i) \cdot (-i)$$

There are no other numbers among Gaussian whole numbers with the given properties. In fact, if some Gaussian whole number ξ divides all Gaussian whole numbers, then this must, in particular, divide number 1 (therefore such numbers are called *unitary divisors*). From $N(\xi) | 1$ it follows that $N(\xi) = 1$. If $\xi = x + yi$, then $x^2 + y^2 = 1$. It is obvious that this equation has precisely four solutions among rational whole umbers: $x = 1$, $y = 0$; $x = -1$, $y = 0$; $x = 0$, $y = 1$; $x = 0$, $y = -1$. These four solutions exactly correspond to Gaussian whole numbers $+1$, -1, i, $-i$.

For Gaussian whole numbers, in a way analogous to rational whole numbers, we develop the concept of *common divisor*, *greatest common divisor*, *mutually prime numbers* and *prime numbers*. The first three concepts are determined exactly in the same way as in the case of rational whole numbers. However, we must deal with the definition of simple Gaussian integers at a little greater length.

DEFINITION 5. A Gaussian whole number π is called *prime* if in all its factorizations $\pi = \tau \cdot \sigma$ as product of two Gaussian whole numbers, one of the factors (τ or σ) is a unitary divisor. (Here the unitary divisors are not considered simple numbers.)

This property may be expressed in other words as follows: a simple Gaussian number π is a nonzero whole Gaussian number whose norm is greater than unity and which cannot be expanded as a product of two Gaussian whole numbers whose norms are less than the norm of number π.

According to this definition, simple Gaussian numbers are, for example, numbers $\pi_1 = 2 + i (N(\pi_1) = 5)$, $\pi_2 = 3 + 2i (N(\pi_2) = 13)$. In general all numbers, whose norms are simple rational numbers, are simple numbers. Below we shall see that simple Gaussian whole numbers are not exhausted by these examples. We shall describe all simple Gaussian numbers. As for now, we go on to the formulation and proof of the fundamental theorem of arithmetic for Gaussian whole numbers.

FUNDAMENTAL THEOREM. *Any Gaussian whole number* $\alpha \neq 0$ *can be expressed as a product of simple Gaussian numbers*

$$\alpha = \pi_1 \cdot \pi_2 \ \ldots \ \pi_k \tag{4}$$

(π_i are simple Gaussian numbers not necessarily different from one another). *Such an expansion is unique in the following sense: if*

$$\alpha = \sigma_1 \cdot \sigma_2 \ \ldots \ \sigma_l \tag{5}$$

is another expansion of number α *into a product of simple Gaussian numbers* σ_j, *then both these expansions have one and the same number of multipliers,* $k = l$, *and factors* (4) *and* (5) *may differ from each other only by the order of factors and by multipliers which are unitary divisors.*

As regards the part of formulation concerning uniqueness of expansion, we shall make yet another observation. If, say,

$$\alpha = \pi_1 \cdot \pi_2 \cdot \pi_3$$

is a product of simple numbers π_1, π_2, π_3, then, for example,

$$\alpha = (-\pi_3) \cdot (i\pi_2) \cdot (i\pi_1) = (\pi_1 \cdot \pi_2 \cdot \pi_3)$$

is the other representation of number α as a product of simple numbers $-\pi_3$, $i\pi_2$, $i\pi_1$ as differing from simple numbers π_1, π_2, π_3. However, it is easy to notice that any of the numbers $-\pi_3$, $i\pi_2$, $i\pi_1$ is obtained by multiplying one of the numbers π_1, π_2, π_3 by some unitary divisor; moreover, the initial order of numbers is also changed. Such differences in the expansion of one and the same number are allowed. The second part of the formulation of the theorem actually states that non-uniqueness of such kind in different expansions vanishes. This case is not different from the case of rational whole numbers in arithmetic. It is simply complicated by the fact that in case of arithmetic of Gaussian whole numbers, we are provided with a large number of unitary divisors [1]. The statement about the uniqueness of the expansion may be formulated more briefly by introducing the idea of associability of Gaussian whole numbers.

DEFINITION 6. Two Gaussian whole numbers are called *associative* if they differ from each other by a factor equal to a unitary divisor, i. e. β, $-\beta$, $i\beta$, $-i\beta$ are associative Gaussian whole numbers if β is an arbitrary Gaussian whole number.

---

[1] We note that uniqueness of expansion, except for the signs of the multipliers about which we talked in the case of whole *rational* numbers, also means uniqueness except for multipliers which are unitary divisors, since $+1$ and -1 are the only unitary divisors in this case.

By using this definition, the statement about uniqueness in the fundamental theorem is formulated as under:

If $\alpha = \pi_1 \cdot \pi_2 \ldots \pi_k$ *and* $\alpha = \sigma_1 \cdot \sigma_2 \ldots \sigma_l$, *where* $\pi_i (i = 1, 2, \ldots, k)$ *and* $\sigma_j (j = 1, 2, \ldots, l)$ *are prime numbers, then* $l = k$ *and the multipliers* σ_j *may be expressed so that every* σ_j *will be associative with the corresponding prime number* π_j.

We shall outline the *proof of the fundamental theorem.* It is done in the same way as the proof of the corresponding statements for rational whole numbers. Therefore we shall not do it exhaustively but strongly recommend the reader to do it himself.

The first statement of the theorem – about the existence of an expansion – may be done by induction for norm of the number:

(a) If $N(\alpha) = 1$, then $\alpha = 1, -1, i, -i$, i. e. α can be expanded into a product of an empty set of prime numbers [1].

(b) Let $N(\alpha) = n$, and for all Gaussian whole numbers with minimum norm, the statement has already been proved. Then either α is a prime number and everything is proved, or $\alpha = \rho \cdot \tau$ where $N(\rho) < n$ and $N(\tau) < n$. According to assumptions of induction, factorization for ρ and τ do exist: $\rho = \pi_1 \cdot \pi_2 \ldots \ldots \pi_k$ and $\tau = \sigma_1 \cdot \sigma_2 \ldots \sigma_l$. Then $\alpha = \pi_1 \cdot \pi_2 \ldots \pi_k \cdot \sigma_1 \cdot \sigma_2 \ldots \ldots \sigma_l$ is the factorization for α.

The proof of the statement about the uniqueness may be carried out by means of establishing the properties of GCD and properties of mutually prime numbers in the case of Gaussian whole numbers. The statement about the possibility of division with a remainder for the case of Gaussian whole numbers provides the clue to the whole proof. Here it is formulated as under:

Let $\alpha, \beta, (\beta \neq 0)$ *be two Gaussian whole numbers, then there exist Gaussian whole numbers* γ *and* ρ, *where* $N(\rho) < N(\beta)$, *so that*

$$\alpha = \gamma \cdot \beta + \rho$$

The proof is based on a very simple geometrical fact: if P is a point lying in a square with side a, or on one of its sides, then the distance of this point P from the nearest corner is less than a. Really, the centre of the square is the point farthest from all corners. But its distance from any corner is equal to $\frac{1}{\sqrt{2}} a < a$. Any other point in the square is situated even closer to the nearest corner.

[1] About 'factorizability' of unitary divisors into products of prime multipliers, we accept the same terms as for ± 1 in case of rational whole numbers, see p. 11.

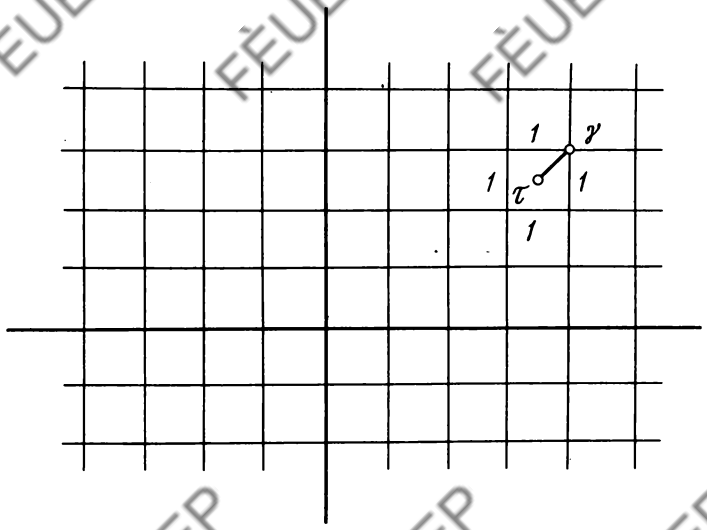


Fig. 2

It is now clearly visible from this simple proof that for any point τ on a complex surface, we can find a point γ with integral coordinates – representing a Gaussian whole number – and removed from τ by a distance less than 1 (Fig. 2). In other words, for any complex number τ there exists a Gaussian whole number γ so that $N(\tau - \gamma) < 1$. Let us find such a γ for number $\tau = \alpha/\beta$ and put $\rho = \alpha - \gamma\beta$. Then ρ is Gaussian whole number

$$N(\rho) = N(\beta) \cdot N\left(\frac{\alpha}{\beta} - \gamma\right) < N(\beta)$$

and

$$\alpha = \gamma\beta + \rho$$

The statement is proved.

Having already got the theorem on division with remainder, we can prove all other properties in the same way as we did above in the case of rational numbers: (1) we prove the existence of a GCD for two Gaussian whole numbers α and β in the form of numbers $\delta \neq 0$ with minimum norm from a set of numbers that can be represented in the form $\alpha\xi + \beta\eta$ (ξ and η are Gaussian whole numbers), (2) the concept of Gaussian whole numbers that

are prime to each other is introduced and the fundamental lemma is proved: if α is mutually prime to β_1 and if α is mutually prime to β_2, then α is mutually prime to $\beta_1 \cdot \beta_2$. After this it is very easy to prove by induction of norm, the uniqueness of factorization into prime multipliers.

§ 5. Gaussian Prime Numbers and Representation of Rational Whole Numbers as Sum of Two Squares

We now go on to a description of all Gaussian prime numbers. We shall first prove some auxiliary statements – lemmae.

LEMMA 1. *Every Gaussian prime number is a divisor of a prime rational number* [1].

Actually, since $N(\alpha) = \alpha \cdot \bar{\alpha}$, *any Gaussian whole number divides its norm*: $\alpha | N(\alpha)$. Now let π be a prime Gaussian number, then $\pi \mid N(\pi)$, and let $N(\pi) = p_1 \cdot p_2 \ldots p_r$ is the factorization of number $N(\pi)$ as a product of prime rational numbers. We have $\pi \mid p_1 \cdot p_2 \ldots p_r$, hence π divides one of the prime numbers p_i. In fact, if the prime Gaussian whole number π did not divide any of the numbers p_i then it would be mutually prime to each of them and consequently to their product $N(\pi)$. But this is impossible since $\pi | N(\pi)$. So π is a divisor of one of the prime rational whole numbers p_i and the lemma is proved.

LEMMA 2. *Norm $N(\pi)$ of a prime Gaussian number π is either a prime rational number or the square of a prime rational number.*

Really, as we already know, π divides some prime rational number p. Let $p = \pi \cdot \gamma$. Taking the norms $N(\pi) \cdot N(\gamma) = p^2$, only two possibilities exist: (1) $N(\pi) = N(\gamma) = p$ and (2) $N(\pi) = p^2 = N(p)$ and $N(\gamma) = 1$. The lemma is thus proved.

The second case means that γ is a unitary divisor and one of the equalities $\pi = p$, $\pi = -p$, $\pi = ip$, $\pi = -ip$ is valid. Consequently, p is such a prime rational number that it is also a prime

---

[1] We observe that a prime rational number is always a whole Gaussian number also; however, as a Gaussian number it is not necessarily prime, but may be divided into Gaussian whole numbers with a lower norm. Thus, for example, 2 is a prime number if it is considered as a rational whole number. But it is not prime if we consider it as a Gaussian whole number. Actually, in the domain of Gaussian whole numbers, 2 can be factorized as $2 = (1 + i)(1 - i)$ and neither of the factors $1 + i$ and $1 - i$ is a unitary divisor. It is obvious that 5 is also not prime in the domain of Gaussian numbers, since $5 = (2 + i)(2 - i)$.

Gaussian number. In case (1) γ is a prime Gaussian number since $N(\gamma) = p$. It may be stated that $\gamma = \bar{\pi}$. Actually, $N(\pi) = p = \pi \cdot \bar{\pi}$ and $\bar{\pi}$ is a prime number. But we also have $p = \pi \cdot \gamma$ so that $\bar{\pi} = \gamma$.

On the other hand, let p be some prime rational number. Then if it is not prime Gaussian number, it is divisible by some prime Gaussian number other than p and, in addition, as we have seen, $p = \pi \cdot \bar{\pi}$. So p is a product of two prime Gaussian complex conjugate numbers. In this case p is the norm of a Gaussian whole number and can therefore be represented as a sum of two squares. Such a prime number if it is odd (i. e. $p \neq 2$) is a number of the form $4n + 1$, representable as a sum of two squares. It can be shown that *all prime numbers of the form* $4n + 1$ *can be represented as a sum of two squares*, i. e. they are the norms of some Gaussian whole numbers and, consequently, belong to the class of such prime rational numbers which can be factorized into products of two complex conjugate prime Gaussian numbers. We shall not carry out proof of this statement [1]. It is all prime rational numbers other than numbers of the type $4n + 1$ or 2, i. e. numbers of the type $4n + 3$ that form the set of prime rational numbers which are both prime and are in the domain of Gaussian numbers.

Two (2) holds a special position. It is easy to see that

$$2 = i \cdot (1 - i)^2, \quad N(1 - i) = 2$$

Thus 2 is divisible by the square of a prime Gaussian number $(1 - i)$.

Assuming that all prime numbers of the type $4n + 1$ can be represented as a sum of two squares, we can now determine all the rational whole numbers which can be represented as a sum of two squares. As we already know, the only necessary and sufficient condition for any such number t is that it should be norm of some Gaussian whole number α: $t = N(\alpha)$. Number α is expanded as a product of prime Gaussian numbers

$$\alpha = \pi_1 \cdot \pi_2 \ \ldots \ \pi_r \tag{6}$$

We divide all prime numbers $\pi_i (i = 1, 2, \ldots, r)$ into two classes. The first class contains such numbers π_i whose norms are prime, and correspondingly the second class contains numbers whose

[1] The proof of this fact based on theory of comparisons and given by L. Euler may be found in any textbook on number theory. It is dealt with in great detail in ref. [3] in the list of literature at the end of the brochure.

norms are squares of prime integers [1]. We denote the various numbers of first class as $\sigma_j (j = 1, 2, \ldots, l)$ and those of second class as $\rho_k (k = 1, 2, \ldots, s)$. We have: $N(\sigma_j) = p_j$, $N(\rho_k) = q_k^2$, where p_j is a prime number of the type $4n + 1$ or 2, and q_k is a prime number of the type $4n + 3$. Combining the equal prime numbers in the right-hand side of (6) we can write the product in the form of powers of prime numbers σ_j and ρ_k

$$\alpha = \sigma_1^{a_1} \ldots \sigma_l^{a_l} \cdot \rho_1^{b_1} \ldots \rho_s^{b_s} \tag{7}$$

Changing to norms, we get

$$N(\alpha) = t = N(\sigma_1^{a_1}) \ldots N(\sigma_l^{a_l}) \cdot N(\rho_1^{b_1}) \ldots N(\rho_s^{b_s})$$

$$t = p_1^{a_1} \ldots p_l^{a_l} q_1^{2b_1} \ldots q_s^{2b_s} \tag{8}$$

We see that prime numbers q_k enter the factorization of number t in *even powers*. Conversely, suppose number t can be represented in the form (8) where each p_j is a prime number of the type $4n + 1$ or number 2, q_k are prime numbers of the type $4n + 3$ and $a_1, \ldots, a_l, b_1, \ldots, b_s$ are non-negative whole numbers. Then, since every p_j is a sum of two squares, we may select σ_j so that $N(\sigma_j) = p_j$. Further, putting $\rho_k = q_k$ and $\alpha = \sigma_1^{a_1} \ldots \sigma_l^{a_l} \cdot \rho_1^{b_1} \ldots \rho_s^{b_s}$, we get $t = N(\alpha)$, i. e. t can be represented as a sum of two squares. Finally, we have the following theorem:

THEOREM 8. *The necessary and sufficient condition that a rational whole number could be represented as a sum of two squares is that prime numbers of the type $4n + 3$ in the factorization of this number should occur in even orders* [2].

We observe that this theorem gives a criterion for a diophantic equation of 2nd order, $x^2 + y^2 = t$, to have a solution (whole number). We shall not stop here to explain how such a solution is actually found out.

---

[1] Of course, it may turn out that one of the classes is empty. This, however, does not substantially allect the course of our discussions. We shall only have to consider that all numbers a_j or all numbers b_k (in factorization (7) or (8)) may be zeros.

[2] Such a formulation also covers the case when the factorization of the number under consideration does not include any prime numbers of the type $4n + 3$, because 0 is also even!

In general, a study of diophantic equations of the type

$$ax^2 + 2bxy + cy^2 = t$$

is closely linked with arithmetics whose number domains are analogous to the domains of Gaussian whole numbers.

In such investigations, the following astonishing fact, which mathematicians encountered in the middle of the last century, is important. *The theorem about uniqueness of factorization of numbers into prime numbers does not hold in all like arithmetics.* Without going into the details of the situation arising here, we shall cite an example of one "arithmetic" in which the fundamental theorem is not valid.

§ 6. Yet Another "Arithmetic"

We shall consider complex numbers of the type

$$\alpha = x + y\sqrt{-5} \tag{1}$$

where x and y are rational whole numbers. It is easily seen that sum, difference and product of numbers of type (1) are also numbers of the same type. We denote the set of all such numbers by Γ. Obviously, Γ contains all rational whole numbers (for $y = 0$). Just as in the case of rational and Gaussian whole numbers, we can talk about divisibility of Γ: α divides β ($\alpha \mid \beta$), if $\beta \mid \alpha$ is again a number from Γ, i. e. representable in form (1). As also in the case of Gaussian whole numbers, the norms of numbers from Γ play an important role in the question of divisibility

$$N(\alpha) = N(x + y\sqrt{-5}) = (x + y\sqrt{-5})(x - y\sqrt{-5})$$
$$= x^2 + 5y^2$$

In this way, the norm of any number from Γ is a rational whole number and since $N(\xi \cdot \eta) = N(\xi) \cdot N(\eta)$, the condition $N(\alpha) \mid N(\beta)$ is necessary (though generally not sufficient) so that $\alpha \mid \delta$.

Just like the case of Gaussian whole numbers, the idea of unitary divisors and prime numbers is introduced. As regards unitary divisors, things are even simpler here than for Gaussian whole numbers. That is, only numbers ± 1 are unitary divisors. Actually, for unitary divisors $\xi = u + v\sqrt{-5}$, the condition $N(\xi) = u^2 + 5v^2 = 1$ must hold. But this diophantic equation obviously cannot have any other solution except $u = \pm 1$ and $v = 0$.

The fact that each number from Γ can be expressed as a product of prime numbers from Γ is proved by induction for norms in exactly the same way as for Gaussian whole numbers. But the *statement about uniqueness* of such a factorization *is not valid* here, and we shall prove it by a simple example.

We shall first show that numbers $2 = 2 + 0 \cdot \sqrt{-5}$, $3 = 3 + 0 \times$ $\times \sqrt{-5}$, $1 + \sqrt{-5}$, $1 - \sqrt{-5}$ are prime numbers in Γ. Actually, $N(2) = 4$, $N(3) = 9$, $N(1 + \sqrt{-5}) = N(1 - \sqrt{-5}) = 6$. If any of these numbers were not prime in Γ, then it could be divisible only by some number $\alpha = x + y\sqrt{-5}$ for which $N(\alpha) = x^2 + 5y^2 = 2$ or $N(\alpha) = x^2 + 5y^2 = 3$. But there are no such numbers in Γ since, obviously, the equations

$$x^2 + 5y^2 = 2 \tag{2}$$

and

$$x^2 + 5y^2 = 3 \tag{3}$$

do not have whole number solutions.

Thus, the given 4 numbers are prime numbers in Γ. We now consider an easily verifiable equality

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}) \tag{4}$$

It shows that number 6 from Γ has two different factorizations into prime numbers.

The German mathematician E. Kummer (1810-1893) encountered this effect while trying to solve the so-called great Fermi theorem. The difficulties that arose later in connection with the non-validity of the fundamental theorem of arithmetic in some important domains of numbers were successfully overcome by Kummer himself as well as by other mathematicians – R. Dedekind, E. Zolotarev, L. Kronecker, etc. Thus arose a vast new branch of mathematics, called the theory of algebraic numbers, which is being successfully developed right till the present time.

LITERATURE

We shall indicate here a number of works through which the reader may get additional information about the number theory, and in particular about the subject of our booklet. The proposed list does not claim to be exhaustive. Some of these works, in turn, carry references to literature on number theory. We shall list these books in the order of increasing difficulty. Thus, while the first few works do not require any special background beyond the framework of school mathematics, the latter are textbooks for students in universities and teachers-training institutes. Experience has shown that these can also be used for self-study.

1. Khinchin, A.Ya., Elements of Number Theory: Encyclopaedia of Elementary Mathematics, Book 1, *Arithmetic*, Gostekhizdat, 1951.

This article by the eminent Soviet mathematician A.Ya. Khinchin is written with great skill and may be recommended both for teachers as well as senior schoolboys. Chapter I "Divisibility and Prime Numbers" contains, among other things, a detailed proof of the fundamental theorem of arithmetic. Chapter II "Method of Comparisons" is, in our view, one of the best introductions to the theory of comparisons – a section of number theory having a very large number of applications in arithmetic as well as modern algebra.

2. Markushevich, A.I., Division with Remainder in Arithmetic and Algebra (Series *Pedagogical Teachers' Library*), pub. Academy of Pedagogical Sciences of RSFSR, 1949.

The book contains almost all the material presented in this booklet, as also many other sections of algebra and arithmetic directly connected with the theory of divisibility.

3. Davenport, H., Higher Arithmetic, Nauka, 1965.

The subtitle of the book "An Introduction to Number Theory" itself indicates that it contains a systematic treatment of the elements of this field. Written by an eminent English specialist on number theory, Davenport's book does not require a background beyond the framework of school mathematics. It may be specially recommended for junior mathematics students, but may also be used by non-mathematicians. The latter, of course, must possess the skill to properly understand descriptions. The readers of our booklet will certainly find Chapter II on "Comparisons" and Chapter V on "Sums of Squares" (giving a complete proof of representability of prime numbers of the type $4n + 3$ as a sum of two squares) of special interest.

The following books may be recommended for mathematics students for self-study in the number theory.

4. Vinogradov, I. N., Principles of Number Theory, Nauka, 1965.
5. Arnold, I. V., Theoretical Arithmetic, Uchpedgiz, 1939.
6. Bukhshtab, A. A., Number Theory, Prosveshchenie, 1966.
7. Hasse, G., Lectures on Number Theory, For. Lang. Publ. House, 1953.

TO THE READER

Mir Publishers would be grateful for your comments on the content, translation and design of this book. We would also be pleased to receive any other suggestions you may wish to make.

Our address is:
USSR, 129820, Moscow, I-110, GSP
Pervy Rizhsky Pereulok, 2
Mir Publishers

Printed in the Union of Soviet Socialist Republics

11179

Other books for your library

L. GOLOVINA, D.SC. AND I. YAGLOM, D.SC.

Induction in Geometry

This booklet deals with various applications of the method of mathematical induction to solving geometric problems and was planned by the authors as a natural continuation of I. S. Sominsky's booklet "The Method of Mathematical Induction" published by Mir Publishers in 1975. It contains 37 worked examples and 40 problems accompanied by brief hints.

Contents. The Method of Mathematical Induction. Calculation by Induction. Proof by Induction. Construction by Induction. Finding Loci by Induction. Definition by Induction. Induction on the Number of Dimensions.

N. VILENKIN, D.SC.

Method of Successive Approximations

This book explains in a popular form the methods of approximation, solutions of algebraic, trigonometric, model and other equations.

Intended for senior schoolchildren, polytechnic students, mathematics teachers and for those who encounter solutions of equations in their practical work. In the course of exposition some elementary ideas about higher mathematics are introduced in the book. About 20 solved exercises are included in the appendix.

The book has already been translated into Spanish and French.

Contents. Successive Approximations. Achilles and the Tortoise. Division by Computers. Determination of Square Roots by the Method of Successive Approximations. Determination of Roots with a Natural Index by Means of Successive Approximations. Iterational Method. Geometrical Meaning of Iterational Method. Compressible Representations. Compressible Representations and Iterational Method. Method of Chords. Perfected Method of Chords. Derivative of a Polynomial. Newton's Method of Approximate Solution of Algebraic Equations. Geometrical Meaning of Derivative. Geometrical Meaning of Newton's Method. Derivatives of Any Function. Calculation of Derivatives. Determination of First Approximations. Combined Method for Solution of Equations. Criterion for Convergence of Iterational Process. Speed of Convergence of Iterational Process. Solution of a System of Linear Equations by the Method of Successive Approximations.

E 0 0 8 1

Educated at the Paris University, Lev Arkadievich Kaluzhnin D.Sc., is a Professor at the Kiev State University. He is the author of some 100 published works, which include a number of textbooks for university and secondary school students.

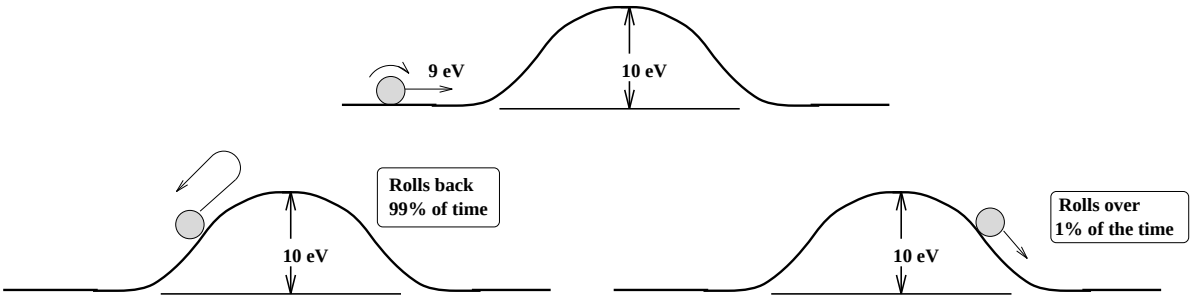
In this booklet, Prof. Kaluzhnin deals with one of the fundamental propositions of arithmetic of rational whole numbers - the uniqueness of their expansion into prime multipliers. Having establisehd a conncetion between arithmetic and Gaussian numbers and the question of representing integers as sum of squares, Prof. Kaluzhnin has shown the uniqueness of expansion also holds in the arithmetic of complex (Gaussian) whole numbers.

The author hopes that the booklet will not only be of interest to senior schoolboys but will also be useful for teachers.

Mir Publishers
Moscow

Quantum Tunneling

In this chapter, we discuss the phenomena which allows an electron to quantum tunnel over a classically forbidden barrier.



This is a strikingly non-intuitive process where small changes in either the height or width of a barrier create large changes the tunneling current of particles crossing the barrier. Quantum tunneling controls natural phenomena such as radioactive α decay where a factor of three increase in the energy released during a decay is responsible for a 10^{20} fold increase in the α decay rate. The inherent sensitivity of the tunneling process can be exploited to produce photographs of individual atoms using scanning tunneling microscopes (STM) or produce extremely rapid amplifiers using tunneling diodes. It is an area of physics which is as philosophically fascinating as it is technologically important.

Most of this chapter deals with *continuum* rather than bound state systems. In bound state problems, one is usually concerned with solving for possible stationary state energies. In tunneling problems, one has a continuous spectrum of possible incident energies. In these problems we are generally concerned with solving for the probability that an electron is transmitted or reflected from a given barrier in terms of its known incident energy.

Quantum Current

Tunneling is described by a transmission coefficient which gives the ratio of the current density emerging from a barrier divided by the current density incident on a barrier. Classically the current density $\vec{J}$ is related to the charge

density ρ and the velocity charge velocity $\vec{v}$ according to $\vec{J} = \rho\ \vec{v}$. Its natural to relate the current density ρ with the electron charge e and the quantum PDF(x) according to $\rho = e\Psi^*(x)\Psi(x)$. It is equally natural to describe the velocity by $\hat{p}/m$ where (in 3 dimensions) $\check{p} = -i\hbar(\partial/\partial x) \rightarrow -i\hbar\vec{\nabla}$. Of course $\vec{\nabla}$ is an operator which needs to operate on part of ρ. Recalling this same issue from our discussion on **Quantum Measurement** we expect:

$$\vec{J} \sim \frac{e}{m}\Psi^*\vec{\nabla}\Psi \tag{1}$$

This form isn't totally correct but fairly close as we shall see.

The continuity equation which relates the time change of the charge density to the divergence of the current density, provides the departure point for the proper derivation of the quantum current.

$$\frac{\partial\rho}{\partial t} + \vec{\nabla}\cdot\vec{J} = 0 \tag{2}$$

By integrating both sides of the continuity current over volume (d^3x) and using Gauss's theorem, one can show that the continuity equation is really just an elegant statement of charge conservation or the relationship between the rate of change of the charge within a surface and the sum of the currents flowing out of the surface.

$$\frac{\partial}{\partial t}\int_{\mathcal{V}} d^3x\ \rho + \int_{\mathcal{V}} d^3x\ \vec{\nabla}\cdot\vec{J} = 0$$

$$0 = \frac{\partial Q}{\partial t} + \int_{\mathcal{S}} d\vec{a}\cdot\vec{J} = \frac{\partial Q}{\partial t} + I_{\text{out}} \tag{3}$$

Rather than talking about the charge and current current density; one often removes a factor of e and talks about the probability density (PDF = ρ) and probability current. We know that the probability density is given by just $\rho = \Psi^*\Psi$ and can get a formula like the continuity equation by some simple, but

clever manipulations of the time dependent Schrödinger Equation. We begin pre-multiplying the SE by Ψ^*:

$$i\hbar\Psi^*\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\Psi^*\frac{\partial^2}{\partial x^2}\Psi + V(x)\Psi^*\Psi \tag{4}$$

We next pre-multiply the complex conjugate of the SE by Ψ and assume a real potential.

$$-i\hbar\Psi\frac{\partial\Psi^*}{\partial t} = -\frac{\hbar^2}{2m}\Psi\frac{\partial^2}{\partial x^2}\Psi^* + V(x)\Psi^*\Psi \tag{5}$$

Subtracting Eq. (5) from (4) we have:

$$i\hbar\left(\Psi^*\frac{\partial\Psi}{\partial t} + \Psi\frac{\partial\Psi^*}{\partial t}\right) = -\frac{\hbar^2}{2m}\left(\Psi^*\frac{\partial^2}{\partial x^2}\Psi - \Psi\frac{\partial^2}{\partial x^2}\Psi^*\right) \tag{6}$$

By applying the rules for differentiating a product it is easy to show:

$$\frac{\partial\Psi^*\Psi}{\partial t} = \left(\Psi^*\frac{\partial\Psi}{\partial t} + \Psi\frac{\partial\Psi^*}{\partial t}\right)$$

$$\frac{\partial}{\partial x}\left(\Psi^*\frac{\partial}{\partial x}\Psi - \Psi\frac{\partial}{\partial x}\Psi^*\right) = \left(\Psi^*\frac{\partial^2}{\partial x^2}\Psi - \Psi\frac{\partial^2}{\partial x^2}\Psi^*\right) \tag{7}$$

Inserting the Eq. (7) expressions into Eq. (6) and dividing by $i\hbar$ we have:

$$\frac{\partial\Psi^*\Psi}{\partial t} + \frac{\partial}{\partial x}\frac{\hbar}{2mi}\left(\Psi^*\frac{\partial}{\partial x}\Psi - \Psi\frac{\partial}{\partial x}\Psi^*\right) = 0 \tag{8}$$

As you can see Eq. (8) is of the form of the 1 dimensional continuity Eq. (1) once one makes the identification:

$$\rho = \psi^*\psi \quad , \quad \vec{J} = \frac{\hbar}{2mi}\left(\Psi^*\frac{\partial}{\partial x}\Psi - \Psi\frac{\partial}{\partial x}\Psi^*\right) \rightarrow \frac{\hbar}{2mi}\left(\Psi^*\vec{\nabla}\Psi - \Psi\vec{\nabla}\Psi^*\right)$$

$$\vec{J} = \frac{\hbar}{m}\mathcal{I}m\left(\Psi^*\vec{\nabla}\Psi\right) \tag{9}$$

where the latter form follows from the observation that $a - a^* = 2i\,\mathcal{I}m(a)$ where $\mathcal{I}m()$ denotes an imaginary part.

In computing the current using Eq. (9) one must consider both the time dependence as well as the space dependence. In order to produce a non-vanishing current density, the wave function must have **a position dependent phase**. Otherwise, the phase of $\Psi(x,t)$ will be the same as the phase of $\vec{\nabla}\Psi(x,t)$ and therefore $\Psi^*\vec{\nabla}\Psi$ will be real. The current density for an electron in a stationary state of the form $\Psi(x,t) = \psi(x)\ \exp(-i\omega t)$ is zero since the phase dependence has no spatial dependence. This makes a great deal of sense since the PDF of a stationary state is time independent which indicates no charge movement or currents. A combination of two stationary states with different energies such as: $\Psi(x,t) = a\ \psi_1(x)\ \exp(-i\omega_1 t) + b\ \psi_2(x)\ \exp(-i\omega_2 t)$ will have a position dependent phase , an oscillating PDF, and a non-zero current density which you will explore in the exercises.

To reinforce the idea that a position dependent phase is required to support a quantum current, consider writing the wave function in polar form $\psi(x) = |\psi(x)|\ \exp(i\phi(x))$ where we have a real modulus function $|\psi(x)|$ and a real phase function $\phi(x)$. Using the chain rule it is easy to show that:

$$\vec{J} = \frac{\hbar}{m}|\psi(x)|^2\ \vec{\nabla}\phi(x) \tag{10}$$

Hence the quantum current is proportional to the gradiant of the phase – a constant phase implies no current.

A particularly simple example of a state with a current flow is a quantum traveling wave of the form: $\Psi(x,t) = a\ \exp(ikx - i\omega t)$. Direct substitution of this form into Eq. (9) or (10) gives us:

$$\vec{J} = (a^*a)\ \frac{\hbar k}{m}\ \hat{x} \tag{11}$$

A Strategy For Solving Tunneling Problems

We will limit ourselves to one-dimensional tunneling through a various potential barriers. An important consequence of working in one dimension is that

the current must be the same at every point along the x-axis since there is no where for the charges to go. We can insure this automatically by using a single, stationary state wave function corresponding to a particle with a definite energy to describe the current flow everywhere. Let us see why this works. In one dimension, the (probability) continuity equation becomes:

$$\frac{\partial}{\partial t}\{\Psi^*\Psi\} + \frac{\partial J}{\partial x} = 0 \quad , \quad \frac{\partial}{\partial t}\{\Psi^*\Psi\} = 0 \ \text{ for a stationary state } \rightarrow \frac{\partial J}{\partial x} = 0 \quad (12)$$

Eq. (12) implies that J is independent of position, and since it is constructed from a stationary state wave function, Eq. (9) tells us that J is independent of time. We thus automatically get a constant current with the same value *everywhere* along the x axis.

How do we find the stationary state wave function? Usually we choose a constant (often zero) potential region on the left of any barriers to "start" a wave function of the form $\psi = \exp(ikx) + r\ \exp(-ikx)$ where $k = \sqrt{2mE}/\hbar$. We think of the $\exp(ikx)$ piece as the incident wave which travels along the positive x axis and the $r\exp(-ikx)$ piece as the reflected wave. One can show[†] that the total current in this zero potential region (or any other region) is $J = \hbar k/m - |r|^2\hbar k/m$. We can think of the total current as the algebraic sum of the incident and reflected currents where each contribution is computed by Eq. (11). One then uses continuity of the wave function and derivative continuity to find the unknown r coefficient. The result of the calculation is generally expressed by a *reflection coefficient* $R \equiv |r|^2$ which is equivalent to $R = |J_r|/|J_i|$ where J_r is the reflected current due to $r\exp(-ikx)$ piece and J_i is the incident current due to the $\exp(ikx)$ piece. Our goal is to calculate R as a function of E which determines the k value we start with.

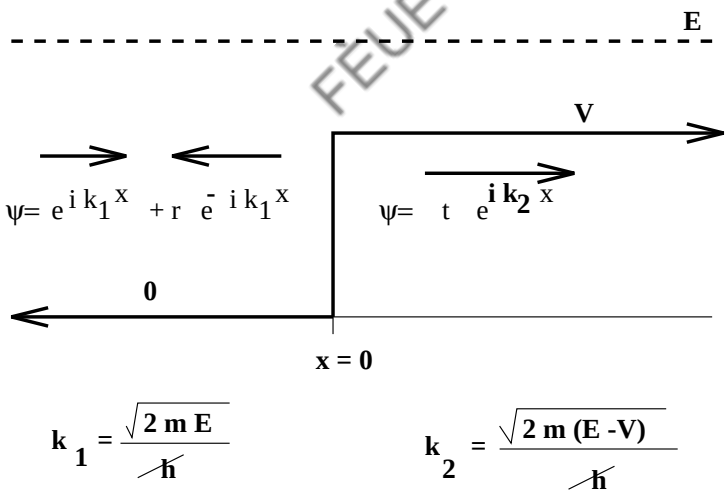
If it turns out that $R < 1$, there will be a net current at $x = +\infty$ (and everywhere else) which we will call the transmitted current or J_t. This will be a

† In homework you show that the interference between the incident and reflected parts of the wave function carries no current which is far from obvious without an explicit calculation

single current since we assume there is nothing at infinity to reflect this current. Current conservation reads $J_i + J_r = J_t$ or $|J_i| - |J_r| = |J_t|$ or dividing by $|J_i|$, $1 - R = T$ where T is the transmission coefficient defined as $T = |J_t|/|J_i|$. We will illustrate this approach in the next section.

The Quantum Curb

The quantum curb as illustrated below involves a traveling wave of the form $\exp(ikx)$ carrying an incident (probability) current $\vec{J} = \frac{k}{m}\,\hat{x}$ which travels to the right in the $x < 0$ region of zero potential. It strikes a potential step of height V at $x = 0$ producing both a reflected wave of amplitude r which travels to the left along with a transmitted wave of amplitude t which travels to the right. In order to insure traveling waves in both the $x > 0$ and $x < 0$ regions, the electron must be classically allowed and have sufficient energy such that $E > V$.



Following our strategy we write $\psi = \exp(ikx) + r\exp(-ikx)$ in our constant potential region at $x < 0$. We can then solve for r and t which are the amplitudes of the reflected and transmitted wave relative to the incident wave of unit amplitude by envoking continuity of ψ, Eq. (13), and its derivative, Eq. (14), at the point $x = 0$.

$$e^{ik_1 0} + r\ e^{-ik_1 0} = t\ e^{ik_2 0} \rightarrow 1 + r = t \tag{13}$$

$$\frac{\partial}{\partial x}\, e^{ik_1 x} + \frac{\partial}{\partial x}\, r\, e^{-ik_1 x} = \frac{\partial}{\partial x}\, t\, e^{ik_2 x}|_{x=0} \rightarrow \quad ik_1 - ik_1\, r = ik_2\, t \rightarrow 1 - r = \frac{k_2}{k_1}\, t \quad (14)$$

Adding Eq. (13) and Eq. (14) we get

$$2 = 1 + \frac{k_2}{k_1}\; t \rightarrow t = \frac{2k_1}{k_1 + k_2} \tag{15}$$

$$\text{From} \quad 1 + r = t \text{ we can find } r = 1 - \frac{2k_1}{k_1 + k_2} \rightarrow r = \frac{k_1 - k_2}{k_1 + k_2} \tag{16}$$

We note that $k_1 = \sqrt{2mE}/\hbar$ and $k_2 = \sqrt{2m(E-V)}/\hbar$ which means for the step up curb: $k_1 > k_2$ and $r > 0$. If the curb were a step down curve $r < 0$.

We turn next to the R and T coefficients. These are **not** the relative amplitudes t and r, but rather are the ratio of the currents carried by the transmitted or reflected waves over the incident wave. Following Eq. (11) we have:

$$T = \frac{(t^* t)\; k_2/m}{k_1/m} = \frac{4k_1 k_2}{(k_1 + k_2)^2} \tag{17}$$

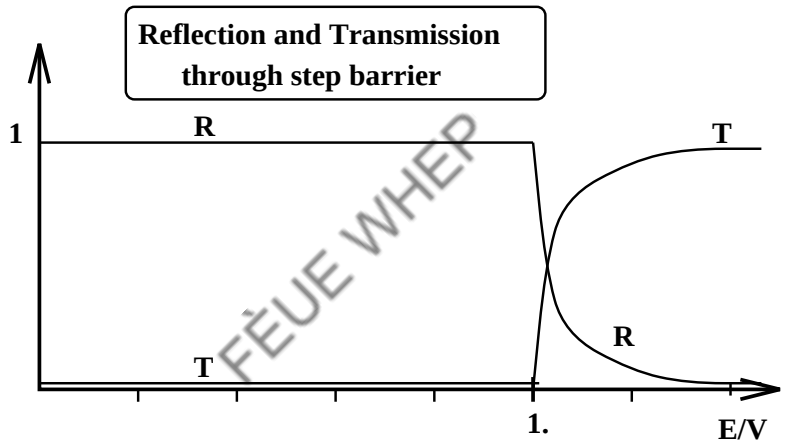
$$R = \frac{(r^* r)\; k_1/m}{k_1/m} = \frac{(k_1 - k_2)^2}{(k_1 + k_2)^2} \tag{18}$$

Using algebra you can show from Eq. (17) and Eq. (18) that $T + R = 1$ as we expect. The current would not be conserved if we had (incorrectly) written the transmission coefficient as the just ratio of the transmitted over incident squared moduli ($T = t^* t$) rather than the correct expression $T = t^* t\; k_2/k_1$. The formula for the reflection and transmission coefficients for a light wave passing from air to glass is exactly the same as Eq.(17) and Eq. (18) which one gets from classical electrodynamics.

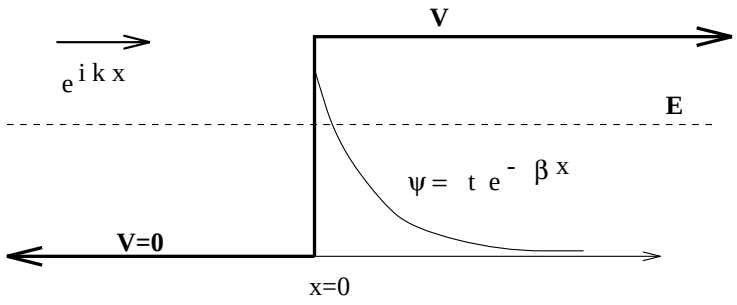
We can use $k_1 = \sqrt{2mE}/\hbar$ and $k_2 = \sqrt{2m(E-V)}/\hbar$ to write the transmission and reflection coefficients in terms of the dimensionless ratio E/V.

$$T = \frac{4\sqrt{E/V}\,\sqrt{E/V-1}}{\left(\sqrt{E/V}+\sqrt{E/V-1}\right)^2} \quad , \quad R = \frac{\left(\sqrt{E/V}-\sqrt{E/V-1}\right)^2}{\left(\sqrt{E/V}+\sqrt{E/V-1}\right)^2} \tag{19}$$

The below figure shows a sketch of the R and T coefficient as a function of E/V.



The above plot suggests that once $E < V$ there is 100 % reflection and 0 % transmission. This case is formally discussed in one of the exercises, but is reasonably easy to understand. If $E < V$, $x > 0$ becomes a classically forbidden region with an exponential wave function of the form $\psi = te^{-\beta x} = |t|\ e^{i\delta}\ e^{-\beta x}$.



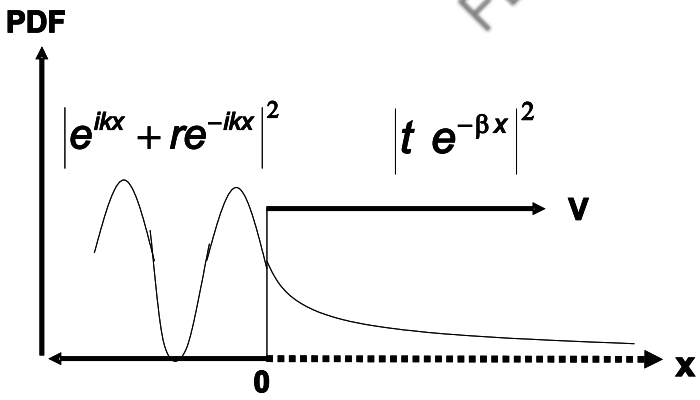
The current $\vec{J} = (\hbar/m)\mathcal{I}m\left(\Psi^*\vec{\nabla}\Psi\right)$ must vanish in this region since the complex phase is a constant (δ) independent of x. Informally there is no transmission

current since the wave function of the electron exponentially dies away in the region $x > 0$ leaving no possibility of finding the electron at large values of x. Since there is no current at $x > 0$, there can be no current at $x < 0$ either which means the reflected current must cancel the incident current and thus $R = 1$.

It is interesting to compute the (unnormalized) PDF in the two regions. In the region $x > 0$ the PDF is of the form $\text{PDF}(x > 0) = |\psi(x)|^2 = |t|^2 \exp(-2\beta x)$. In the region $x < 0$ we have $\text{PDF(x} < 0) = \left|e^{ikx} + |r|\ e^{i\delta}\ e^{-ikx}\right|^2$ We have explicitly written the reflection amplitude as a modulus $|r|$ and a phase δ. In this case $|r| = 1$ and, as you show in homework, δ is a function of E/V. Using $|A + B|^2 = |A|^2 + |B|^2 + 2\text{Re}\{A^* B\} = |A|^2 + |B|^2 + 2\text{Re}\{B^* A\}$ we have:

$$\text{PDF(x} < 0) = 1 + |r|^2 + 2|r| \cos(2kx - \delta) = 2 + 2\cos(2kx - \delta) \qquad (20)$$

Below is a crude sketch of the PDF where the PDF has continuity and derivative continuity at $x = 0$. In the $x < 0$ region one has a standing wave pattern.

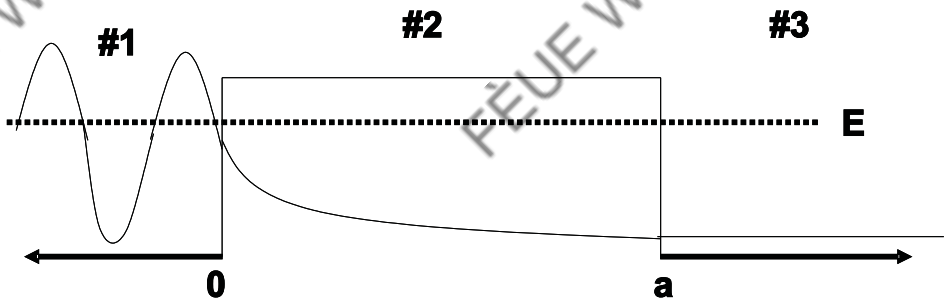


Quantum Tunneling

As we just saw, there is no transmission through a classically forbidden barrier step since beyond $x > 0$ we have a single, exponentially decaying wave function which cannot create the position dependent phase required to have a non-zero quantum current. If , on the other hand, we restore the potential back to ground as shown below, the classically forbidden region can have both $\exp(+\beta x)$ as well as $\exp(-\beta x)$ contributions. As long as the boundary condition equations give a

different complex phase between these two contributions, the complex phase will develop a position dependence and the classically forbidden region will carry a current, implying there will be a current in the $x < 0$ region as well and hence $R < 1$, $T > 0$ and there will be a current at infinity.

Another way of looking at this is based on the fact that the current according to Eq. (9) is proportional to the wave function as well as its derivative. If there is just a classically forbidden region past $x > 0$, the wave function will die out to zero and there will be no possibility of a non-zero ψ at infinity. Since, $J \propto \psi^* \frac{\partial \psi}{\partial x}$ according to Eq. (9), this means there can be no quantum current at $x \rightarrow \infty$ and thus no net current anywhere. If , however, we restore the potential back to ground, we can "catch" the dying wave function before it totally decays away, and have thus have $\psi \neq 0$ at infinity and thus a current everywhere.



We have crudely sketched the PDF $= \psi^* \psi$ of the electron in the limit of low transmission. We can estimate the transmission coefficient in this limit by making use of the classically forbidden, quantum curb PDF's discussed in the last section. The PDF's in region #1 , #2, and #3 will be

$$\text{PDF}_1 \approx 2 + 2\cos(2kx - \delta) \quad ; \quad \text{PDF}_2 \approx |c|^2 e^{-2\beta x} \quad ; \quad \text{PDF}_3 = |te^{ikx}|^2 = |t|^2 \quad (20b)$$

The region #1 PDF is approximate since we set $|r| = 1$ whereas $|r|$ is slightly less than 1 in the low transmission limit. The region #2 PDF is approximate since we threw away the $\exp(\beta x)$ piece that must be present in region #2 to convey the current but it should be small in this limit. We can now estimate $T = |t|^2$

by matching the approximate PDF's at $x = 0$ and $x = a$.

$$\text{PDF}_1(0) = 2 + 2\cos(\delta) = f(E/V) = \text{PDF}_2(0) = |c|^2 \rightarrow |c|^2 = f(E/V)$$

We write $\text{PDF}_1(0) = f(E/V)$ since the phase δ is a function of E/V for the classically forbidden quantum current.

$$\text{PDF}_2(a) = f(E/V)e^{-2\beta a} = \text{PDF}_3(a) = |t|^2 \rightarrow T = f(E/V)e^{-2\beta a} \qquad (20c)$$

This is indeed the correct form of the exact solution when $\beta a \gg 1$ given by Eq. (26) that we discuss in the next section. We will show that $f(E/V)$ is a relatively smooth function that is approximately 1 meaning that the transmission coefficient is roughly $T \approx \exp -2\beta x$ where $\beta = \sqrt{2m(V-E)}/\hbar$. As we will argue later this means that the tranmitted current is very sensitive to very small (atomic scale) changes in a and the forbidden gap $V - E$. Here is an illustration. Consider an energy gap of $V - E = 2\ eV$. This is typical of the work function of metals which forms the barrier preventing metal electrons from escaping into space. The β corresponding to this work function is:

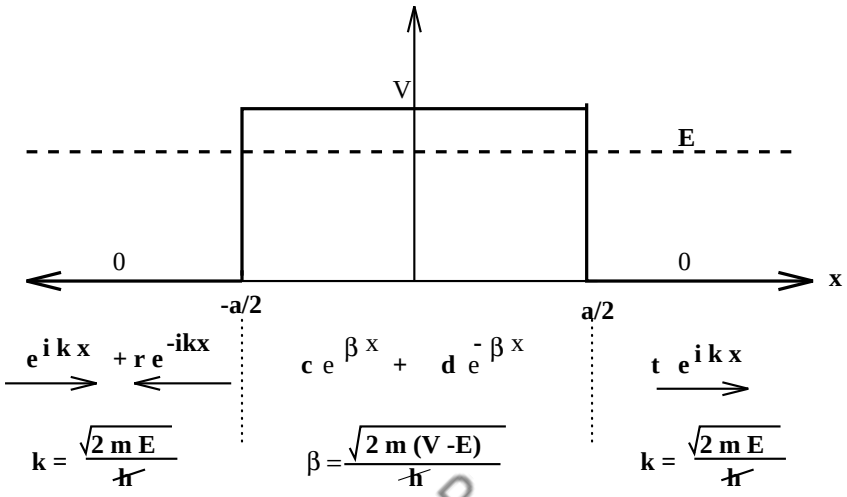
$$\beta = \frac{\sqrt{2m(V-E)}}{\hbar} = \frac{\sqrt{2mc^2(V-E)}}{\hbar c} = \frac{\sqrt{2(0.511 \times 10^6\ eV)(2\ eV)}}{197\ eV nm} = 7.26\, nm^{-1}$$

Now consider varying the tunneling length a from $a = 0.25\ nm$ to $a = 0.20\ nm$, the ratio of the tunneling current is:

$$\frac{T(0.20)}{T(0.25)} = \frac{f(E/V)\ \exp(-2(7.26)(0.20))}{f(E/V)\ \exp(-2(7.26)(0.25))} = e^{14.5\ (0.25-0.20)} = 2.1$$

To put this into perspective, we found that changing the tunneling length by the radius of a hydrogen atom (0.05 nm) changes the tranmission coefficient or tunneling current by 210 %. This extreme sensitivity of tunneling to distance changes on the scale of atomic dimensions forms the foundation of the STM or scanning tunneling microscope that we will describe later.

Formal Solution of the Classically Forbidden Barrier



In close analogy with our treatment of the step, in the region $x < -a/2$ we have an incident and reflected wave $\psi = e^{ikx} + r\ e^{-ikx}$. In the forbidden region, the wave function is constructed out of a decaying and growing exponential with unknown coefficients. We can envoke continuity and derivative continuity at the boundary $x = -a/2$ to obtain:

$$e^{-ika/2} + r\ e^{ika/2} = c\ e^{-\beta a/2} + d\ e^{+\beta a/2} \tag{21}$$

$$ik\ e^{-ika/2} - ik\ r\ e^{ika/2} = \beta\ c\ e^{-\beta a/2} - \beta d\ e^{+\beta a/2} \tag{22}$$

In the region $x > a/2$ we have a single transmitted wave and can envoke continuity and derivative continuity at the boundary $x = +a/2$ to obtain:

$$c\ e^{\beta a/2} + d\ e^{-\beta a/2} = t\ e^{ika/2} \tag{23}$$

$$\beta\ c\ e^{+\beta a/2} - \beta\ d\ e^{-\beta a/2} = ik\ t\ e^{ika/2} \tag{24}$$

Eq. (21) - Eq. (24) represent four equations in four unknowns: (r c d t). The solution to this series is not terribly instructive so I will just quote the results

for the transmission coefficient:

$$T = \left(1 + \frac{\sinh^2(\beta a)}{4(E/V)(1-E/V)}\right)^{-1} \quad \text{where} \quad \beta = \frac{\sqrt{2m(V-E)}}{\hbar} \tag{25}$$

The reflection coefficient follows from $R = 1 - T$. To get some insight into this we will go to various limits.

The $\beta a \gg 1$ limit

In this limit $\sinh\beta \to \frac{1}{2}e^{\beta a} \gg 1$. This means that $\sinh^2(\ \ )/[\ \ ]$ dominates the expression $1 + \sinh^2(\ \ )/[\ \ ]$ and Eq. (25) becomes:

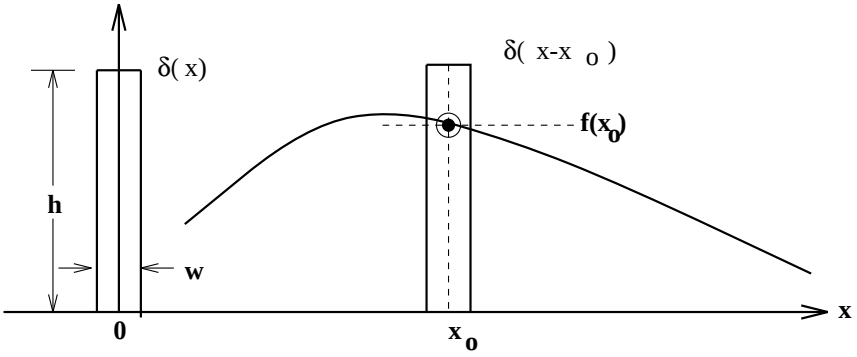
$$T \approx 16(E/V)(1-E/V)\ \ e^{-2\beta a} \tag{26}$$

This expression agrees with the form we deduced in Eq. (20c) in the $T \ll 1$ limit.

The Delta function limit

First a few words about δ-functions in case you haven't encountered them before. A δ function is a function with an infinitesimal width and an infinite height but a unit area. A force described as δ-function in time such as $F(t) = \delta(t)$ is known as a unit impulse which occurs at time $t = 0$. As you probably know from both mechanics and circuit theory; it is often relatively easy to describe a the behavior of a circuit or mechanical system to a voltage or force impulse. The same is true of quantum mechanical systems.

In quantum mechanics we often think of the a δ-function potential. We can think of this potential as a rectangular function of width w and height h in the limit that $w \to 0$, $h \to \infty$, and $wh = 1$ although there are many other limiting forms which approach the δ-function as well. The δ-function centered at $x = 0$ is written as $\delta(x)$. To shift the δ-function to the right so that it centers on x_o we translate the function by subtracting x_o from its argument or $\delta(x - x_o)$.



The operational definition of the δ-function is as follows:

$$\int_{x \in x_o} dx\ f(x)\ \delta(x - x_o) = f(x_o) \tag{28}$$

In words, the integral of the product of $f(x) \times \delta(x - x_o)$ over any interval containing the point x_o is just the function evaluated at x_o. Its easy to see how our rectangular representation of $\delta(x - x_o)$ has this property.

$$\int_{x \in x_o} dx\ f(x)\ \delta(x - x_o) = \lim_{w \to 0} \int_{x_o - w/2}^{x_o + w/2} dx\ h \times f(x) = wh \times f(x_o) = f(x_o) \tag{29}$$

<u>δ-function potentials in the Schrödinger Equation</u>

We write the time independent Schrödinger Equation for the case of a δ-function potential of strength g or $V(x) = g\ \delta(x - x_o)$. In writing this, we note that dimensions of strength g are energy $\times$ distance ($eg\ eV \cdot nm$) since the dimensions of $\delta(x - x_o)$ are distance^{-1} in order that $\int\ dx\ \delta(x - x_o) = 1$ (dimensionless).

$$\frac{-\hbar^2}{2m}\ \frac{\partial^2 \psi}{\partial x^2} + g\ \delta(x - x_o)\psi = E\psi \tag{30}$$

If we restrict ourselves to the region in an infinitesimal neighborhood of x_o, it is

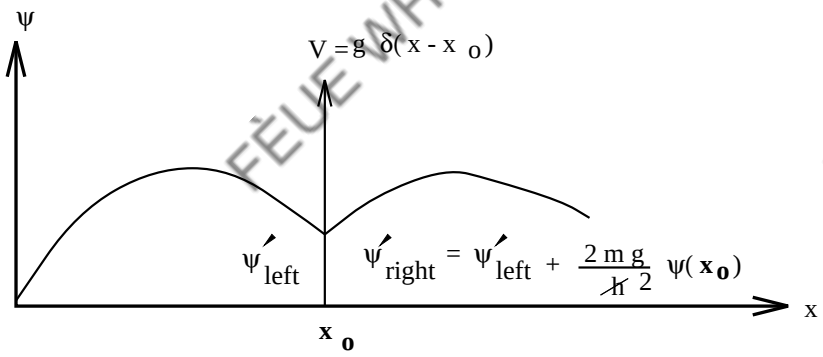
clear that $g\delta(x-x_o) \to \infty \gg E$ so we can ignore the righthand side.

$$\frac{\hbar^2}{2m}\,\frac{\partial^2\psi}{\partial x^2} = g\;\delta(x-x_o)\psi \tag{31}$$

Integrating both sides of the equation from $x_o - \Delta \to X_o + \Delta$ where $\Delta \to 0$ we have:

$$\frac{\hbar^2}{2m}\int_{x_o-\Delta}^{x_o+\Delta} dx\;\frac{\partial^2\psi}{\partial x^2} = \int_{x_o-\Delta}^{x_o+\Delta} dx\;g\;\delta(x-x_o)\psi = g\psi(x_o)$$

$$\frac{\partial\psi}{\partial x}_{x_o+\Delta} - \frac{\partial\psi}{\partial x}_{x_o-\Delta} = \frac{2mg}{\hbar^2}\psi(x_o) \tag{32}$$



Hence the δ-function potential creates a discontinuity in the slope of the wave function which is proportional to the δ-function strength and the value of ψ at the δ-function location. I will ask you in the exercises to apply Eq. (32) to compute the transmission coefficient through a δ-function barrier. Here is a check for you.

The $a \to 0$ but $Va = g$ limit of Eq. (25)

Lets consider this limit for:

$$T = \left(1 + \frac{\sinh^2(\beta a)}{4(E/V)(1-E/V)}\right)^{-1} \quad \text{where} \quad \beta = \frac{\sqrt{2m(V-E)}}{\hbar}$$

Since Va is finite, $a\sqrt{V}$ or $\beta a \to 0$ and therefore $\sinh^2(\beta a) \to \beta^2 a^2$. We also

have $4(E/V)(1-E/V) \to 4E/V$ since $V \to \infty$. Hence

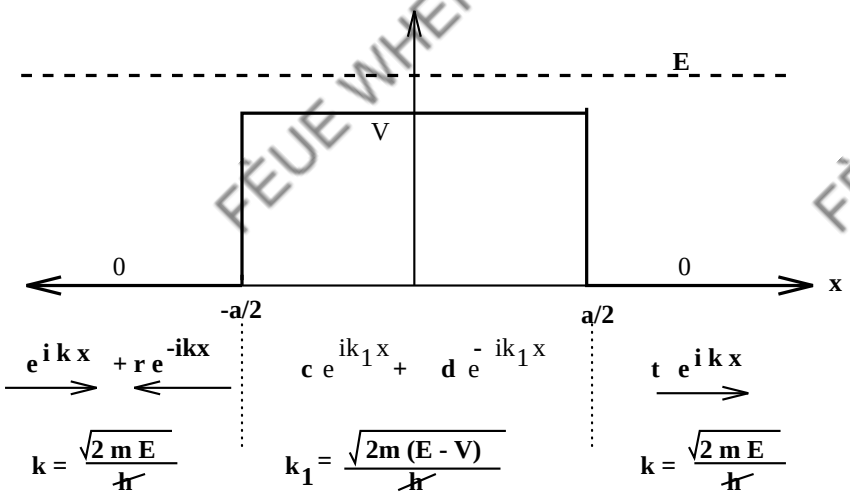
$$T = \left(1 + \frac{2ma^2(V-E)V}{4E\hbar^2}\right)^{-1} \to \left(1 + \frac{ma^2V^2}{2E\hbar^2}\right)^{-1} \tag{33}$$

Inserting $aV = g$ and $E = \hbar^2k^2/2m$ we have:

$$T = \left(1 + \frac{m^2g^2}{\hbar^4k^2}\right)^{-1} \tag{34}$$

Classically allowed barrier

We next consider the case of a traveling wave incident on a classically allowed barrier with $E > V$ as illustrated below.

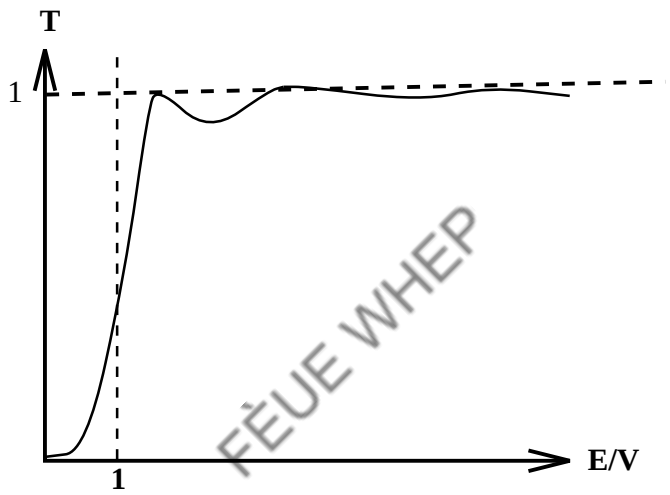


The transmission coefficient for this barrier is:

$$T = \left(1 + \frac{\sin^2(k_1a)}{4(E/V)(E/V-1)}\right)^{-1} \quad \text{where} \quad k_1 = \frac{\sqrt{2m(E-V)}}{\hbar} \tag{35}$$

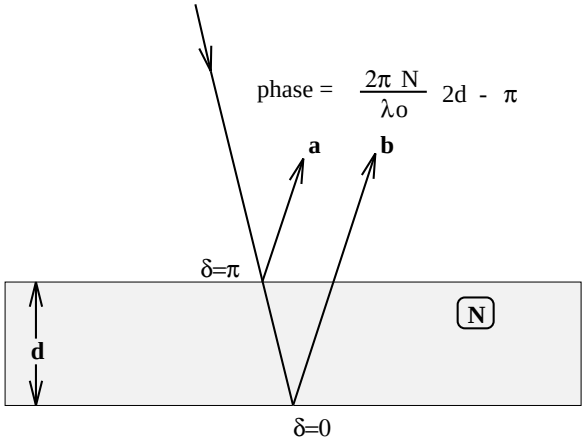
The real difference between this case and the classically forbidden case is the use of $c\ \exp(ik_1x) + d\ \exp(-ik_1x)$ rather than $c\ \exp(\beta x) + d\ \exp(-\beta x)$ for the wave function in the $0 < x < a$ region. Essentially the exponential argument $\beta \to ik_1$. We note that we can get to Eq. (35) from the forbidden T in Eq. (31) by the substitution $\sinh i\beta \to i \sin k_1$ which describes how a hyperbolic sine of an imaginary number is related to the usual sine.

We note from the form of Eq. (35), that we have perfect tranmission whenever $k_1 a = n\pi \ , n = 1, 2, ...$ This condition is equivalent to $n(\lambda_1/2) = a$ where λ_1 is the electron wavelength in the barrier region. Here is a crude sketch of the transmission coefficient as a function of E/V:



The phenemona of 100 % tranmission through a barrier at specific magic energies or wavelengths is often called **transmission resonance**. Examples occur in both atomic and nuclear physics. In atomic physics , one has the Ramsauer effect (discovered 1908) where noble gas atoms become nearly transparent to several volt electrons of of specific energies. A very similar phenomena, known as "size resonance". occurs for several MeV neutrons which can pass transparently through the nucleus at resonant energies.

Transmission resonance at magic wavelengths also occurs in reflections of electromagnetic waves from thin films as shown below. We have angled the incident ray a bit for clarity but we will discuss the case of normal incidence.



The dielectric reflection from the top surface (**Ray a**) acquires a boundary phase shift of π, while the reflection from the bottom surface (**Ray b**) has no boundary phase shift but acquires an "distance" phase shift of $k_1(2d) = 2\pi d/\lambda$, where $\lambda = \lambda_o/N$ If $n(\lambda/2) = d$, the two reflected rays will cancel and destructively interfere leading to 100% transmission. This is essentially what happens in the quantum mechanical case as well: the waves reflected as the barrier is entered and exited interfere destructively.

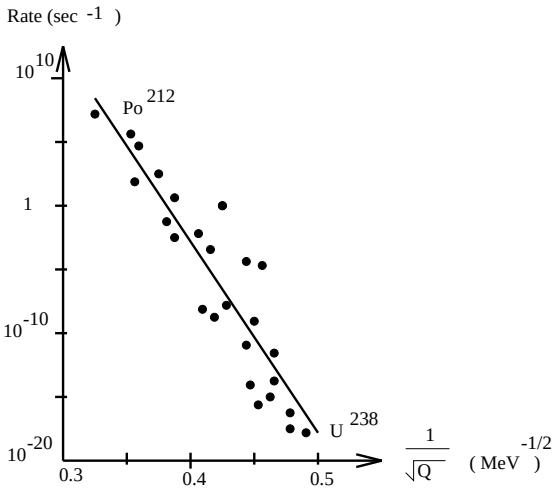
Quantum Tunneling in the Real World

There are a wide variety of real world phenemena which can be pictured in terms of tunneling processes.

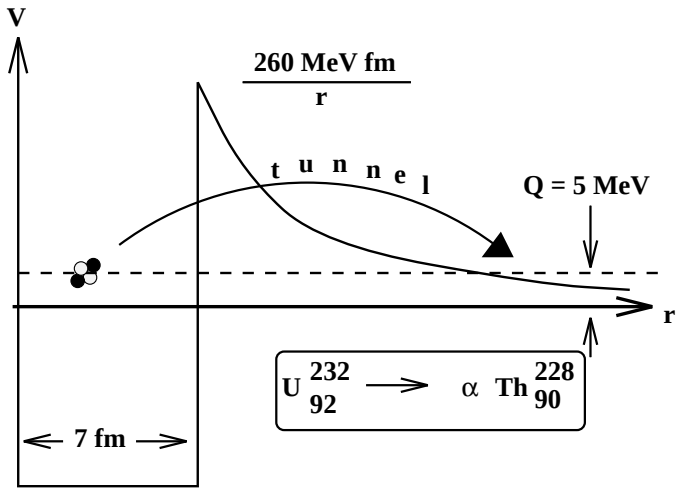
A example on the nuclear level involves alpha decay whereby a heavy parent nucleus becomes more stable by losing some electrical charge by ejecting an α particle. An example is provided by the decay of uranium isotopes:

$$U_{92}^{A} \rightarrow \alpha_2^4 + Th_{90}^{A-4}$$

In this notation, A is the atomic weight (the number of neutrons and protons), the subscript is the atomic number (the number of protons). An α particle is a helium nucleus which is an unusually stable nucleus consisting of 2 neutrons and 2 protons. The half-life of the various radioactive isotopes depends exponentially on the energy release as crudely sketched below:

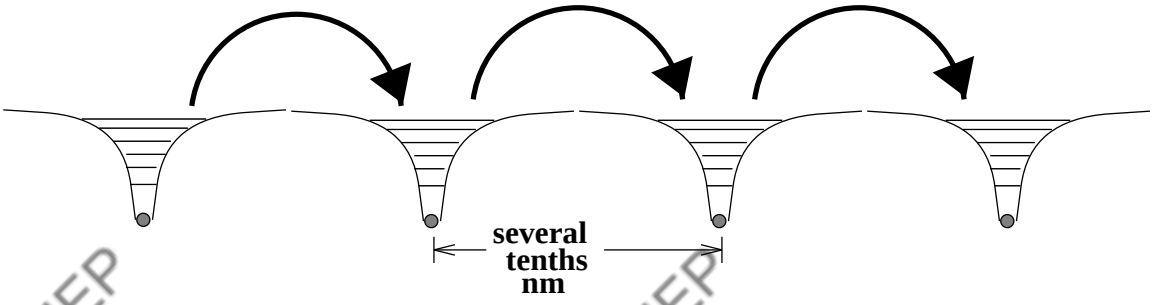


George Gamow in the 1930's proposed a quantum tunneling explanation for nuclear α decay. In this model, the α particle is initially bound by the strong interaction within a nuclear well created by the Thorium nucleus to form Uranium. The Uranium decays by having the α particle tunnel through the Coulomb barrier of the Thorium nucleus. As depicted below, the gap by which the tunneling is classically forbidden decreases as the energy release (Q) increases. The decay rate is proportional to the transmission coefficient through the barrier which depends exponentially on the energy gap.



Nuclear α decay is controlled by very, small tranmission coefficients. At the other extreme, we consider the nearly free motion of conduction electrons

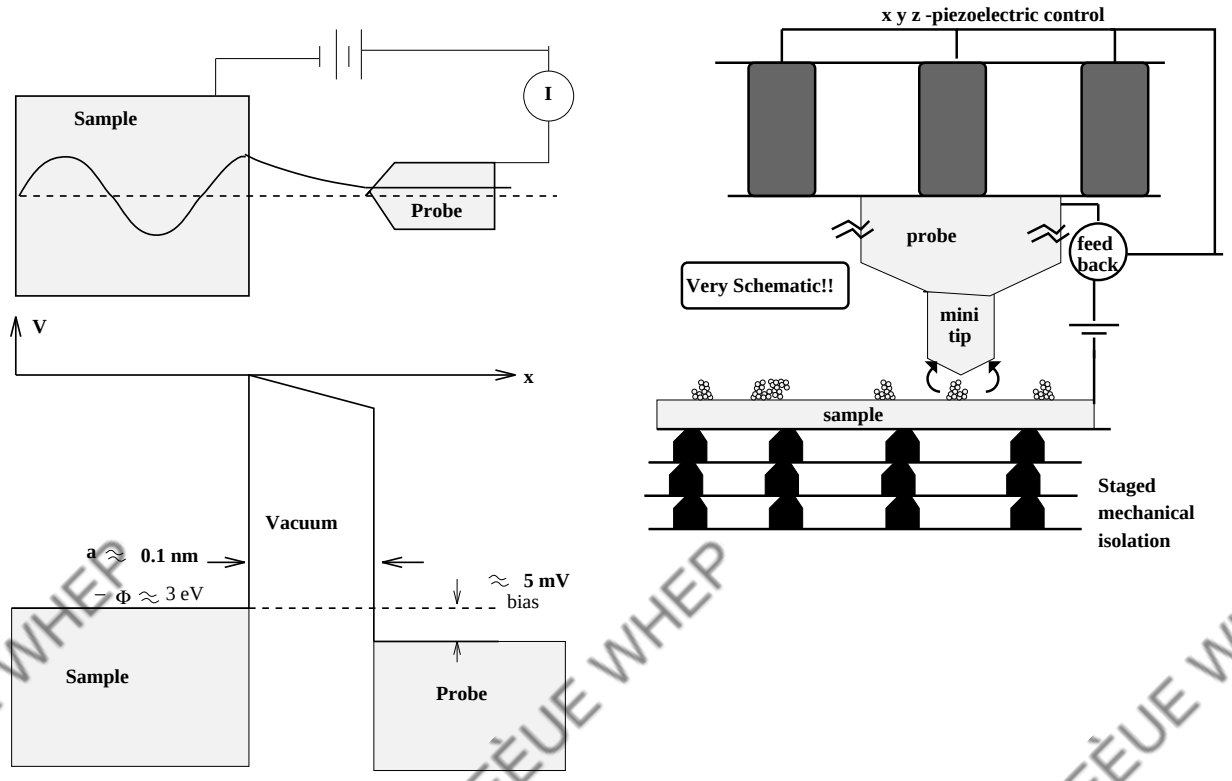
through a metal. Although electrons are bound within individual metal atoms on quantum levels of the atomic discrete spectrum, they exhibit nearly free motion in a metal crystal where there is regularly spaced lattice of ions on a spacing of a few tenths of a nanometer.



We can think of the loosely bound valence electrons in outer orbitals as jumping from ion to ion by quantum tunneling through fairly weak barriers owing to the narrow width and small depth of the effective interatomic barrier. We will discuss this in depth on our chapter on Crystals.

The Scanning Tunneling Microscope

A very impressive device which exploits the extreme sensitivity of quantum tunneling is called the Scanning Tunneling Microscope (STM) developed by Gerd Binnig and Heinrich Rohrer of the IBM Zurich Research Laboratory in 1981. They won the 1986 Nobel Prize for Physics for this achievement. The basic idea of the STM is sketched below:



The STM determines the distance between a surface and probe to atomic distance scales by measuring the quantum tunneling current of electrons leaving the sample and entering the probe. The quantum barrier to electron flow is essentially the work function of the metal. Recall from the photoelectric effect that it requires a certain minimum energy to photoeject an electron from a metal surface. In terms of a potential, we can say that the metal surface has a potential which lies below the potential of free space (vacuum). This minimum potential or work function (Φ) is typically several electron volts for most surfaces. Because of the work function, electrons will be bound within the surface. A naked surface will contain its electrons in analogy with a step barrier. The wavefunction outside the metal will be a classically forbidden wave function $\psi(x) = \exp(-\beta x)$ where $\beta = \sqrt{2m\Phi}/\hbar$, and there will be no electron current out of the surface. If one brings up a metal probe within a few atomic dimensions of the surface, one will form a quantum barrier rather than a step and it will be possible for electrons to quantum tunnel from the sample to the probe. Essentially the vacuum gap

region (V = 0) acts as the quantum barrier to the electrons which lie at a few volts below vacuum potential.

The STM measures the gap between the surface and probe by measuring the size of the quantum tunneling current. In order to achieve a net current, it is necessary to induce an asymmmetry such that the tunneling current from the sample surface to the probe is not cancelled by an opposite tunneling current from the probe to the surface. To insure this asymmetry, the probe is typically biased a few millivolts below the surface being studied. The tunneling current is proportional to the surface electron density, the small bias potential, and the barrier transmission coefficient. This transmission coefficient depends exponentially on the gap between the surface and the probe. The probe is "scanned" across the sample surface in much the same way as a television raster pattern, and the tunnel current forms either a gray scale or pseudo-color picture of the surface height. It is possible to resolve variations in the tunneling current due to individual atoms changing the surface to probe barrier gap!

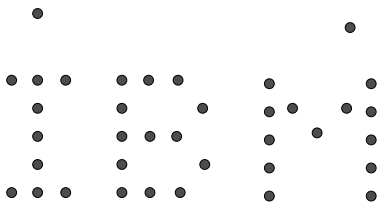
This is the concept but like any technological innovation, the devil is in the details. For example, in order to resolve individual atoms, one needs a probe tip which is only a few atoms wide. It is relatively easy to electochemically etch metals to achieve micron diameter wires, but this is about 10000 times too coarse. My understanding is that Binnig and Rohrer began with micron diameter tungsten needles which were then micro-pitted by placing them in a strong electric field which dislodge atoms on the surface leaving sharp points behind. The object was not to get one precisely polished micro-tip as depicted above, but one particularly sharp point which dominates the tunneling.

Another problem is mechanical oscillation due to room vibration. Typical vibration amplitudes are about a micron which is again 10000 times larger than the ≈0.1 nm required gap for good tunneling. These vibrations are damped by using a series of mechanical stages with carefully controlled natural vibration frequencies selected to block transmission of vibrations over a wide band width.

In addition they used copper plates and magnets designed to damp vibrations by dissipating eddy currents.

Finally one has the apparently formidable problem of controlling the probe position on the scale of atomic dimensions. The elegant solution employed by Binnig and Rohrer was to position the probe using a three point support consisting of three piezoelectric ceramics which expands or contracts by a few tenths of nanometer per applied volt. The tripod support allows the probe to be advanced to the surface (equal expansion of the ceramics) angled (unequal expansion) to affect a transverse scan. In one incarnation, one feeds the tunneling current through a control loop designed to maintain a constant tunnel current by having the probe follow the surface topography and thus maintain a constant gap. The display is then tied to the piezoelectric control current.

Not only can surfaces be measured, but the electrostatics of the probe can create forces which enable manipulation on the atomic level. A very impressive demonstration of this involved using an STM probe to nudge individual Xenon atoms to spell a word on a substate. Here is an crude artist's conception.



Because the distance between the probe and the sample is so small, there is little chance of air molecules slipping in between the tunneling gap. STM studies can be performed in air, oils, and even electrolytic solutions. This extends the reach of the instrument to physics, chemistry, engineering and microbiology.

Important Points

1. The quantum mechanical (probability) current density is given by

$$\vec{J} = \frac{\hbar}{2mi}\left(\Psi^*\vec{\nabla}\Psi - \Psi\vec{\nabla}\Psi^*\right) = \frac{\hbar}{m}\mathcal{I}m\left(\Psi^*\vec{\nabla}\Psi\right)$$

   The usual electric current density is equal to the charge $\times$ probability current density. In one dimension $\vec{\nabla} = \hat{x}\partial/\partial x$.

2. The traveling wave $\Psi(x,t) = a\ \exp(ikx - i\omega t)$ carries the probability current $\vec{J} = (a^*a)\ \frac{k}{m}\ \hat{x}$.

3. In barrier transmission problems we send in an incident traveling wave of the form $\psi(x) = e^{+ik_1x}$ which strikes the barrier producing a reflected wave (re^{-ik_1x}) and a transmitted wave (te^{+ik_2x}). The reflection and transmission coefficients are the ratios of the modulus of the reflected or transmitted currents to the incident current.

$$R = r^*r \quad , \quad T = \frac{k_2}{k_1}t^*t \quad , \quad R + T = 1$$

By matching boundary conditions for the case of a step boundary we obtained for a classically allowed step:

$$T = \frac{4k_1k_2}{(k_1+k_2)^2} \quad R = \frac{(k_1-k_2)^2}{(k_1+k_2)^2}$$

where k_1 and k_2 are the wave vectors in the two regions. For a classically forbidden step T=0 and R=1 but the reflected wave is phase shifted.

4. For a classically forbidden square bump barrier with a low transmission coefficient:

$$T \approx 16(E/V)(1-E/V)\ \ e^{-2\beta a}$$

   where a is the width of the barrier and $\beta = \sqrt{2m(V-E)}/\hbar$.

5. One can have 100 % transmission through a classically allowed square bump barrier when the width is an integral number of half wavelengths (wavelengths in the barrier). This is the same condition for 100 % transmission through a thin film.

6. We discussed several real work applications of quantum tunneling. The conduction of electrons in a metal can be thought of as quantum tunneling between the closely spaced atoms of the lattice. In the low tranmission limit one has the Gamow model for α emission, and the scanning tunneling microscope. All these phenomena are hypersensitive to small differences in the energy or width of the barrier.