



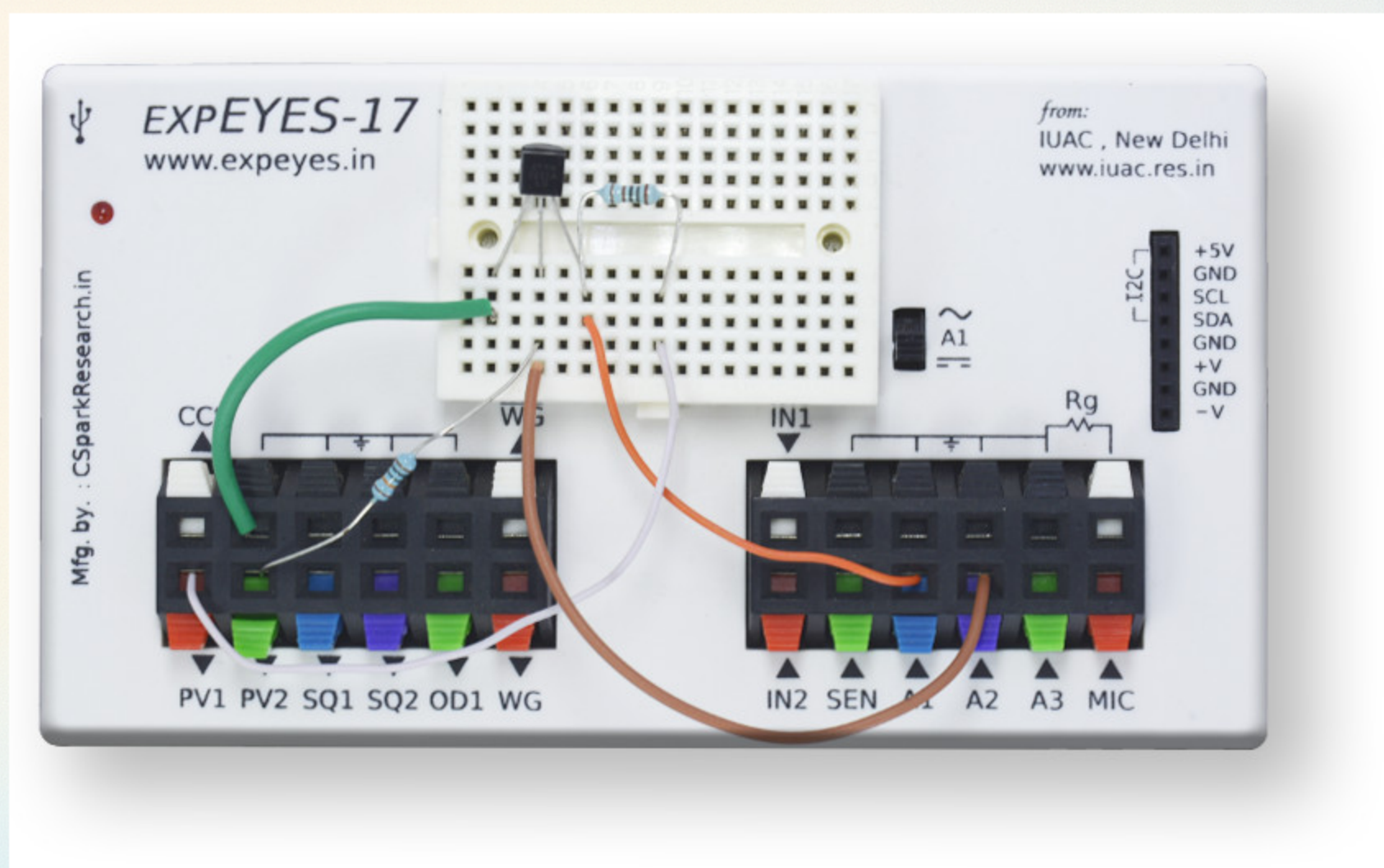
ISSN 0970-5953

Vol. 38 No.2

July-September 2024

Physics Education

Quarterly e-Journal Devoted to Physics Pedagogy



Utilizing ExpEYES-17 Kit to Determine
Boltzmann Constant

Volume 38, Number 2

In this Issue:

Editorial

1. **Demonstration of Online Practical Experiment during COVID-19 Pandemic: Determination of Boltzmann constant & Reverse Saturation Current using ExpEYES-17 kit**
Durjoy Roy and Subhrajyoti Biswas.....1-19
2. **Estimating gravimetric effects for ordinary bodies**
Pier Franco Nali.....1-12
3. **Why is a 60° prism preferred for observing dispersion**
Gautham Dathatreyan and K. M. Udayanandan.....1-5
4. **Solitons: Waves with many Attributes of Particles**
Vishwamittar.....1-29

Editorial

Revitalizing the Physics Education Journal

The Physics Education Journal of the Indian Association of Physics Teachers (IAPT) has been dormant for some time due to several significant challenges, including the COVID-19 pandemic disruptions, website management technical issues, and transition in editorial management. However, we are pleased to announce that we are undertaking a complete renewal of the journal website to meet current standards expected of a quality academic publication.

Acknowledgments and Gratitude

We express our sincere gratitude to DAE-BRNS for continuing their financial support, which will sustain the journal for the next five years. We also thank the new editorial team members who have generously volunteered their services to maintain the high quality of this journal.

Publication Schedule and Plans

To ensure that previously accepted papers receive their due recognition, we have decided to publish Volume 38 & 39 with four quarterly issues containing these papers, scheduled from April 2024 to January 2025. Almost all papers have been typeset in LaTeX by our local team, proofread by the authors, and are now ready for publication.

Enhancing Transparency and Accountability

As we continue our pursuit of excellence in academic publishing, we are implementing several new initiatives to enhance transparency, accountability, and the overall quality of our journal. These changes are designed to uphold the integrity

of the academic publishing process and provide our readers with the most accurate and reliable scientific information.

New Publication Requirements

1. Copyright Transfer Agreement: Effective immediately, all authors submitting articles must sign a copyright transfer agreement.
2. Enhanced Reference System: Authors must provide clickable links for all DOI and URL references, and include a minimum of 10-15 references per article.
3. Chicago Style Citation: Authors must follow the Chicago style of referencing to maintain consistency and clarity.
4. Anti-Plagiarism Verification: Authors must submit an anti-plagiarism report generated using reputable software with their manuscript.
5. Visual Abstracts: Each submission must include a visual abstract summarizing the article's main findings and implications.

Submission Guidelines

Authors should submit manuscripts with all required supporting documents through our online submission system once it is operational. Please ensure your manuscript adheres to our guidelines and includes all required elements, including the copyright transfer agreement, properly formatted references with clickable links, and anti-plagiarism report.

Call for Submissions

We welcome comments and scholarly rebuttals to any of the published papers from IAPT members and our broader readership. In the interim, kindly submit your papers via email to chiefeditorphyedu2025@gmail.com & secretaryphyedu2025@gmail.com.

Thank You for your continued support and cooperation.

Professor O.S.K.S Sastri

Editor-in-Chief

Physics Education Journal,

Indian Association of Physics Teachers

Demonstration of Online Practical Experiment during COVID-19 Pandemic: Determination of Boltzmann constant & Reverse Saturation Current using ExpEYES-17 kit

Durjoy Roy¹ and Subhrajyoti Biswas²

¹Department of Electronic Science, Rishi Bankim Chandra College,
Naihati 743165, India.
roy.durjoy@gmail.com

²Department of Physics, Rishi Bankim Chandra College,
Naihati 743165, India.
anjansubhra@gmail.com

Submitted on 13-01-2022

Abstract

The Boltzmann constant can be easily obtained in teaching laboratories using a traditional method. The method devised here uses the ExpEYES-17, a low-cost, microcontroller-based digital data acquisition system, where the actual experiment is performed with real diodes, and Python is used to complement the learning.

1 Introduction

The determination of fundamental physical constants plays a crucial role in understanding the underlying principles of nature. One such constant, the **Boltzmann**

constant (k_B), establishes a fundamental relationship between temperature and energy, serving as a cornerstone in statistical mechanics and thermodynamics. Traditional methods for determining k_B often involve complex experimental setups and extensive manual data collection. However, modern technological advancements have enabled the development of cost-effective and efficient methodologies that enhance accuracy while simplifying the process.

The COVID-19 pandemic and subsequent lockdowns disrupted traditional hands-on laboratory education, highlighting the need for a new paradigm of **on-line and remote practical experiments**.

The transition to virtual learning necessitated the adoption of digital tools that could facilitate real-time data acquisition, remote control of experiments, and computational analysis. In response, digital laboratory platforms such as **ExpEYES-17** have emerged as powerful alternatives, allowing students and researchers to conduct physics experiments from their homes or in hybrid learning environments. The availability of low-cost, open-source hardware and software has further expanded access to experimental learning, overcoming geographical and logistical barriers.

In this study, we present an alternative approach to determine the Boltzmann constant using a PN junction diode and the **ExpEYES-17** data acquisition system. Previously, a study [7] was published to find Boltzmann Constant. In this article, we have applied curve fitting method, and determined the reverse saturation current as well.

ExpEYES-17 is a low-cost, microcontroller-based digital data acquisition system specifically designed for educational and research applications. It offers precise control over experimental parameters, real-time data visualization, and seamless integration with Python-based computational tools. The versatility of **ExpEYES-17** has been demonstrated in various experimental setups, including measurements of electronic circuit parameters, sensor-based studies, and thermodynamic experiments [1, 10].

The experiment is based on the **current-voltage (I-V) characteristics of a PN junction diode**, which follows the well-known diode equation. By analyzing the exponential relationship between current and voltage in the forward bias region, we extract the Boltzmann constant using a curve-fitting technique. The experiment also incorporates a **Platinum Resistance Thermometer (PT100)** to measure the temperature of the diode during operation, ensuring accuracy in calculations.

Previous studies have successfully employed **ExpEYES-17** for undergraduate physics experiments, as documented in works published in *IOP Physics Education* and *IAPT Physics Education* [8, 9, 11, 14, 15]. These works highlight the effectiveness of the system in hands-on learning and experimental physics education, demonstrating its applicability in teaching fundamental physics principles. Inspired by these studies, our work further extends its utility by providing a refined approach for determining the Boltzmann constant.

This approach provides students and researchers with an accessible and efficient method for determining the Boltzmann constant while reinforcing fundamental concepts in semiconductor physics, thermodynamics, and experimental data analysis. The integration of Python-based computation further enhances the learning experience, making it a valuable addition to undergraduate and postgraduate physics laboratories.

2 The ExpEYES-17

The ExpEYES-17 is basically a data acquisition system / kit along with a four channel digital storage oscilloscope (DSO) which was developed by a group of scientists and researchers at the Inter-University Accelerator Center (IUAC), New Delhi, India. The name of the kit is short form of *expEriments for Young Engineers & Scientists*. The main architecture has been designed using the Micro-controller *PIC24EP64GP204* and runs by Python. The hardware design and necessary software are freely available [17] to share knowledge and foster interest in experiments worldwide.

The top view of the kit has been shown in Fig. 1. There are separate connector blocks - output block, input block and I^2C modules. The output block contains programmable sources for various signals and voltages. *PV1* and *PV2* are two programmable direct voltage sources with 12 bit resolution. It has two square waves generators *SQ1* and *SQ2*. The frequency of these sources can be varied from 4Hz to 100kHz as well as duty cycle. However, the frequency range of sine or triangular waves is lower (5Hz to 5kHz) compared to that of square waves. The waveforms other than square wave is obtained from *WG*. The \overline{WG} indicates the complements of the signals of *WG*. Apart from these outputs, there are One digital Output *OD1*, and one constant current source *CCS* which can supply 1mA current. Any other waveforms which are

not specified can also be generated by writing a simple python code as discussed in the user manual.

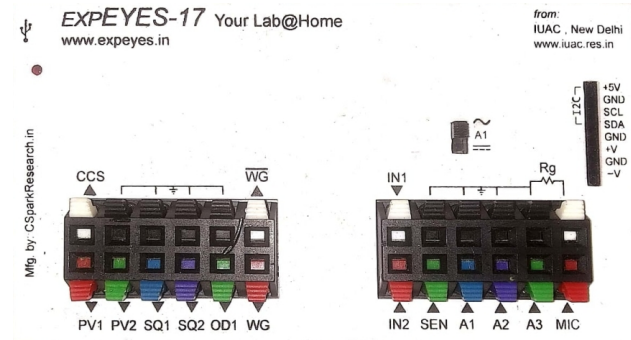


Figure 1: The inputs & outputs of the ExpEYES-17 kit.

On the other hand, inputs for signal capturing have been integrated into the input block, which contains six input terminals. The *IN1* is used to measure capacitance upto 10nF and the frequency counter *IN2* is able to measure frequencies upto several MHz. The input terminals *A1* and *A2* are capable of measuring analog voltages within $\pm 16V$ range having maximum 1Msps sampling rate. Other inputs, *A3* can measure $\pm 3.3V$ and *MIC* is capable of capturing audio signals. A detailed discussion about the device can be found in article [10].

The kit is connected to laptop / computer through usb port and no external power is needed to run the device. The Fig. 2 displays the graphical user interface GUI which is written on C and python. The GUI consists of oscilloscope display

and the control sliders for *PV1*, *PV2*, *SQ1*, *WG* and four oscilloscope channels *A1*, *A2*, *A3* and *MIC* including the time base and trigger control. There is a list of about 50 experiments from school to graduate levels that can be performed with this kit. But one may design other experiments too.

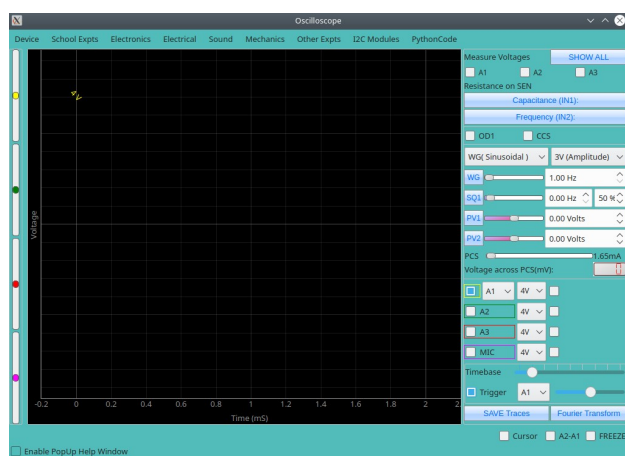


Figure 2: A view of the User Interface (Software) of the ExpEYES-17 kit.

Boltzmann constant.



Figure 3: Ludwig Eduard Boltzmann (1844-1906)

3 Theory:

3.1 Boltzmann's Constant:

Ludwig Eduard Boltzmann (1844-1906) was an Austrian theoretical physicist and philosopher. His greatest achievements were the development of statistical mechanics, and the statistical explanation of the second law of thermodynamics. In 1877 he provided the current definition of entropy, $S = K_B \ln(\omega)$, interpreted as a measure of the statistical disorder of a system.[19]. Max Planck named the constant k_B the

3.2 PN Junction diode, and finding the Boltzmann Constant from its equation:

The discovery of semiconductors ushered in a new era of electronic technology. The pure semiconductors behave like ordinary resistance and follow the Ohm's current-voltage relation and they have no practical applications in electronic devices. Some impurities are doped to produce impure semiconductors. Depending upon the doping materials, the impure semiconductors are either *p*-type or *n*-type. The former contains a large

number of positively charged holes and a small number of electrons. The n -type semiconductor contains large number of electrons and few numbers of holes. In Fig. (4) p -type and n -type semiconductors are shown.

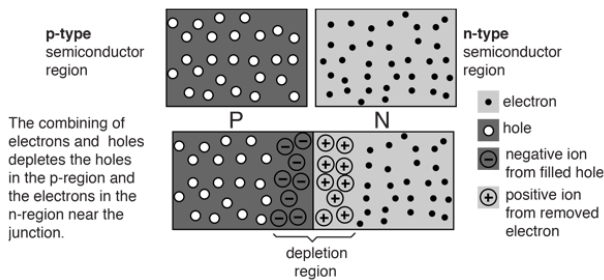


Figure 4: p type and n type semiconductors, and p-n junction diode, formed on a single crystal [18]

When different materials are doped onto a single crystal in such a way that one side forms a p -type region and the other side forms an n -type region, it creates a PN junction, as shown in Fig (4). Almost all the electronic devices contains one or more such pn junctions. Since the p -side has a high concentration of holes and the n -side has a high concentration of electrons, electrons naturally diffuse toward the p -side and holes toward the n -side. Electrons and holes recombine as they diffuse across the p - and n -sides. However, some electrons on the p -side near the junction cannot recombine. Similarly, holes on the n -side near the junction line can not recombine. The electrons near the junction line on the p -side and holes on the n -side set up an electric

field and this electric field prevents further diffusion of holes from p -side and electrons from the n -side. A charge depletion region is created across the junction, as shown in Fig (4) forming a barrier potential V_0 . This barrier potential varies from $0.1V$ to $0.3V$.

Such pn junction is called a junction diode or simply diode, which may be connected to a power supply in two different ways. When p -side is connected with positive terminal to the source and n -side to its negative terminal, is called forward biasing. In this connection, holes on p -side and electrons on n -side get repulsion by the positive and negative terminals, respectively. If the potential difference V across the junction is greater than the barrier potential V_0 , holes from p -side and electrons from n -side cross the barrier and produces the forward current I . This forward current and applied potential difference are related by

$$I = I_s \left(e^{\frac{qV}{nK_B T}} - 1 \right), \quad (1)$$

where, q is the charge of electron, K_B is Boltzmann's constant, T is absolute temperature of the junction and $n = 1$ for Germanium (Ge) crystal and 2 for Silicon (Si) crystal. The reverse saturation current is denoted by I_s . Another type of connection, called reverse bias connection when p -side is connected with the negative terminal and n -side to the positive terminal of the source. In this connection, holes from p -side and electrons from n -sides are attracted by the negative and positive terminals, respec-

tively. Only, few electrons from p -side and holes from n -side get repulsion and diffuse across the junction and produces very small amount of reverse saturation current I_s .

In this experiment we use **both** Ge and Si diode in forward bias. At the room temperature, ($T \approx 300K$), and $V = 0.5V$ to $1V$, $e^{\frac{qV}{nK_B T}}$ varies from of the order 10^4 to 10^8 REF and thus $e^{\frac{qV}{nK_B T}} \gg 1$. This enables us to neglect 1 compared with the exponential term and approximate the Eq. (1) as

$$I = I_s e^{\frac{qV}{nK_B T}}, \quad (2)$$

The above equation can be written after taking the logarithm of both side

$$\ln(I) = \frac{qV}{nK_B T} + \ln(I_s), \quad (3)$$

which is the equation of a straight line

$$y = m x + c. \quad (4)$$

In the above equation, $m = \frac{\Delta y}{\Delta x}$ is the slope of the straight line and c denotes the length where the straight line cuts the y -axis. Now comparing, the Eqs. (3) and (4), one can write

$$\begin{aligned} y &= \ln(I), \\ x &= V, \\ m &= \frac{q}{nK_B T}, \\ c &= \ln(I_s). \end{aligned}$$

In the experiment, we plot graph of $\ln(I) - V$ and determine the slope $m =$

$\frac{\Delta \ln(I)}{\Delta V}$. This value of the slope m is equal to $\frac{q}{nK_B T}$. Therefore,

$$K_B = \frac{q}{nT} \cdot m. \quad (5)$$

Eq. (5) is our working formula for Determination of the value of the Boltzmann's constant.

3.3 Temperature sensing using RTD (PT100) Sensor:

Resistor Temperature Detectors (RTD) are temperature sensors that sense change of temperature by measuring the change in the value of a resistor. Many RTD elements consist of a length of fine wire wrapped around a ceramic or glass core but other types constructions are also common. The RTD wire is a pure material, typically platinum, nickel, or copper. The material has an accurate resistance/temperature relationship which is used to provide an indication of temperature. As RTD elements are fragile, they are often housed in protective probes, mostly covered in a steel tube that is hermetically sealed.

Platinum, a noble metal and having the most stable resistance-temperature relationship over the largest temperature range, was proposed by Sir William Siemens as an element for a resistance temperature detector back in 1871. Figure (5) shows a schematic of a common wire-wound Platinum Resistor Thermal sensor.

One important property of metals that is used to construct the resistive elements of RTDs is the linear approximation of the resistance versus temperature relationship between 0° C and 100° C. This temperature coefficient of resistance is denoted by α and having units of $\Omega/(\Omega^\circ\text{C})$:

$$\alpha = \frac{R_{100} - R_0}{100 \times R_0} \quad (6)$$

where R_{100} is the resistance of the metal at 100° C, and R_0 is the resistance at 0° C.

Typically, industrial PRTs have a nominal alpha value of $\alpha = 3.85 \times 10^{-3}$ per ° C.

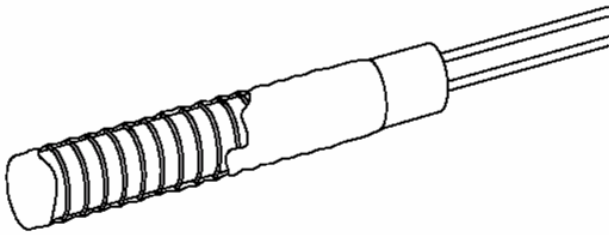


Figure 5: Schematic of a PRT.

Eq. (5) shows that the Boltzmann Constant K_B is a function of temperature in absolute scale. We have used a PT100 Platinum Resistance Thermometer to measure the temperature of the P-N Junction diode during the experiment, and then convert the measured temperature to absolute scale. PT100 has a typical resistance value of 100 Ω at the 0° C. The Resistance Temperature relation of a PT100 sensor for temperatures greater than 0° C is given by the Callendar-Van Dusen equation

$$R_t = R_0(1 + At + Bt^2) \quad (7)$$

where, t =temperature, R_t = resistance at temperature t , and R_0 = resistance at 0° C.

For industrial grade of PRT, standard EN 60751:1995 provides values for the coefficients from the value of α of Eq (6) as

$$A = 3.9083 \times 10^{-3} \text{ } ^\circ\text{C}^{-1},$$

$$B = -5.775 \times 10^{-7} \text{ } ^\circ\text{C}^{-2}$$

Since the coefficient B is very small, the resistance changes almost linearly with the temperature.

For positive changes in temperature, solution of the quadratic equation using the famous Sreedhar Acharya formula yields the following relationship between temperature and resistance:

$$t = \frac{-A + \sqrt{A^2 - 4B(1 - \frac{R_t - R_{offset}}{R_0 - R_{offset}})}}{2B} \quad (8)$$

PT100 sensors come with a length of wire, and thus, the resistance of the wire is added as an offset value with the measured temperature, which must be subtracted from the values of the measured resistances as shown in the equation (8). This equation is used in the python program to determine the temperature of the diode during the experiment.

3.4 Curve fitting to find the Boltzmann Constant:

Equation (3) shows that a straight line can be found by plotting the $\log_{10} I_d$ vs V_d data. We must fit a straight line that is best fit with the experimental data, find its slope(m) and intercept(c) to find the values of K_B and I_s .

Linear regressions of x and y , a Least Squares Regressions Method for fitting curve is used here. In this method, a straight line is fitted by the means of minimizing the vertical distances between the actual data points and the straight line. The coefficients of an equation analogous to equation (4), $y = a_1x + a_0$ are calculated from the following relations:

$$a_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} \quad (9)$$

$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \quad (10)$$

where m and c are analogous to a_1 and a_0 , respectively. x_i and y_i are the deviations between the experimental data and fitted line points. The details of the expressions are out of the scope of discussions of this article, and can be found in [4].

4 Experimental Set-up:

The whole experimental set-up consists of two parts, namely,

- Hardware, i.e., the ExpEYES-17 kit and the Circuit for the experiment, and
- Python Program, i.e., the Software for the experiment

The parts are elaborated below in the following sections.

4.1 Circuit for the experiment:

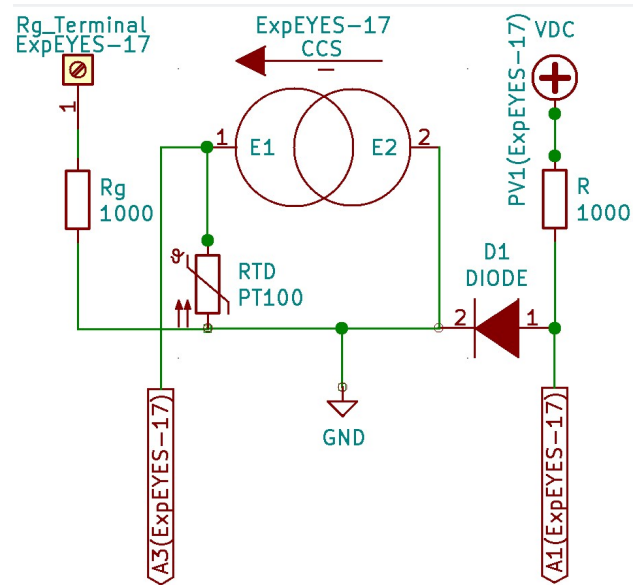


Figure 6: Circuit for finding Diode Characteristics for K_B .

The ExpEYES-17 is the main equipment used in the experiment to generate and measure the required signals. A simple circuit, as shown in the left hand side with respect to ground in the figure (6), is constructed with the help of wires and crocodile clips. We have used two diodes in this experiment. Diode 1N4148 is a Silicon diode, and 1N60 is a Germanium

diode. The diode under test is shown as D1. The diode is biased through a resistance R1. The purpose of using the resistor is to determine the diode current using an equation, as ExpEYES-17 cannot measure current directly. The voltage across the diode is sensed by the Channel A1 of the ExpEYES-17 while the diode gets dc bias from the variable voltage source PV1. PV1 is controlled by the python program listed in Listing (1).

The left hand side of the figure (6), with respect to the ground, shows the schematic of the connection of PT100 temperature sensor. In ExpEYES-17, a constant current source is available for use via the terminal CCS. It is designed to deliver 1 mA of constant current, however, due to tolerances of the components used, this current varies a little bit from device to device. The PT100 is connected between the CCS and the ground terminal, and the voltage caused by the flow of the constant current generated by CCS across the PT100 is sensed via the A3 terminal of the ExpEYES-17 kit. The A3 terminal can sense $\pm 3.3V$ and has a high input resistance of $10M\Omega$. A non-inverting amplifier built around a TL082 OP-AMP in the ExpEYES allows the gain of the A3 input to be set by connecting a resistor from terminal R_g to ground, to ground, as given by equation (11)

$$Gain = 1 + \frac{10000}{R_g} \quad (11)$$

The temperature obtained by this arrangement is used to compute the value of the

Boltzmann constant using the python programs.

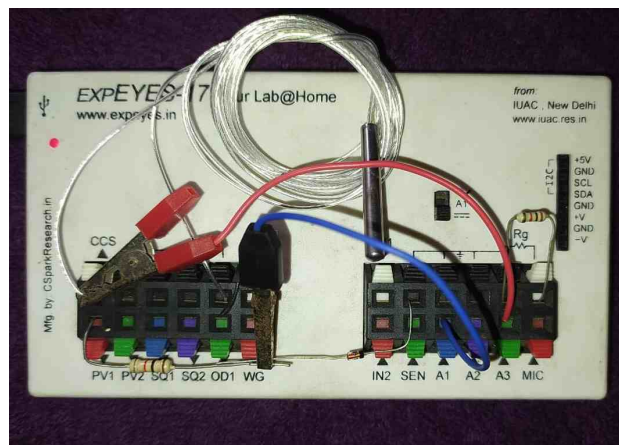


Figure 7: Actual circuit with the ExpEYES-17 for the experiment. The n side of the 1N60 Diode is connected to ground, and p side to PV1 via a $1K\Omega$ resistor. The black clip takes the V_D to A1. The PT100 is connected between CCS and ground, the red clip takes the V_t to A3. Another $1K\Omega$ resistor is connected between R_g and ground.

Figure (7) shows the actual connection of the circuit on the ExpEYES-17 kit.

4.2 Software for the experiment:

The software for this program is nothing but a couple of programs written in python. Libraries specific to ExpEYES-17 are included in the program, and the libraries required to plot the diode characteristics, transfer the data from one program to another, and to curve fit the plots.

4.2.1 Program 1:

The first Python program (1) generates a list of diode voltages and currents by varying PV1. The range of the non-linear section is selected using `vdliml` and `vdlimh` since the logarithm of the non-linear region results in a linear plot. Each voltage-current reading is accompanied by temperature data from the PT100 sensor, converted to absolute values and stored in `latemp[]` for averaging in program (2).

This program produces two sets of lists: `vda[]` and `ida[]` for plotting in figures (9) and (10), and `lvda[]` and `lida[]` for calculating K_B in program (2). The Callendar–Van Dusen equation (7) is used to compute temperature from PT100 resistance changes, accounting for device offsets. The wire resistance at 0°C , denoted as r_{offset} , is used for calibration. Data transfer between the two programs is handled via text files using the `pickle` library.

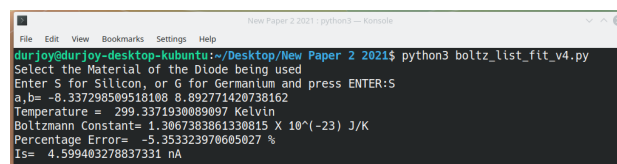
4.2.2 Program 2:

The second Python program (2) processes the voltage-current lists from program (1) to plot $\ln(I)$ vs. V and determine the slope $m = \frac{\Delta \ln(I)}{\Delta V}$ through curve fitting (3.4). The resulting straight-line fit, located in the fourth quadrant with a positive slope, is used to compute the Boltzmann constant separately for Si and Ge diodes.

Unlike assuming room temperature, this program utilizes temperature readings

recorded in program (1) to compute an accurate average. The calculated K_B values are compared with standard values, and the percentage error is determined. Additionally, the logarithm of the reverse saturation current, $\log_{10} I_s$, is derived from the y-axis intercept.

Figure (8) shows typical output for a Silicon Diode.



```

durjoy@durjoy-desktop-kubuntu:~/Desktop/New Paper 2 2021$ python3 boltz_list_fit_v4.py
Select the Material of the Diode being used
Enter S for Silicon, or G for Germanium and press ENTER:S
a,b= -8.337298509518108 8.892771420738162
Temperature = 299.3371930089097 Kelvin
Boltzmann Constant= 1.3067383861330815 X 10^(-23) J/K
Percentage Error= -5.35323970605027 %
Is= 4.599403278837331 nA

```

Figure 8: Typical Program output, as given out by program (2) for a Si diode 1N4148

5 Results:

Plots in figures (9) and (10), show the VI characteristics of a Ge and a Si diode. The data generated to plot the curves are used to compute the Boltzmann constant.

Figures (11) and (12) show the plots of the experimental data and the linear fitted curve for the Ge and Si Diodes, respectively.

Figure (12) shows the screenshot of the program output as well with the plot. The calculated values of K_B and I_s along with the average temperature in Kelvin, are displayed as output.

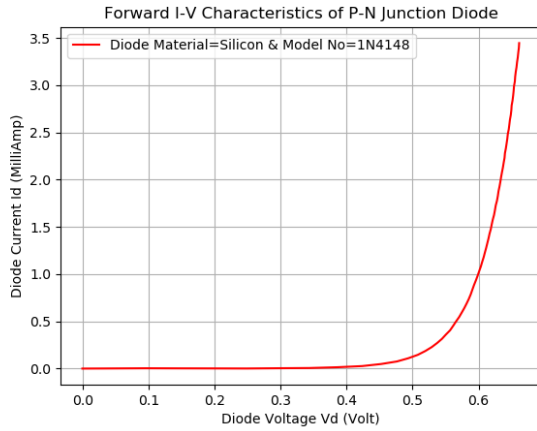


Figure 9: Ge (1N60) Diode Characteristics for K_B generated running Program (1).

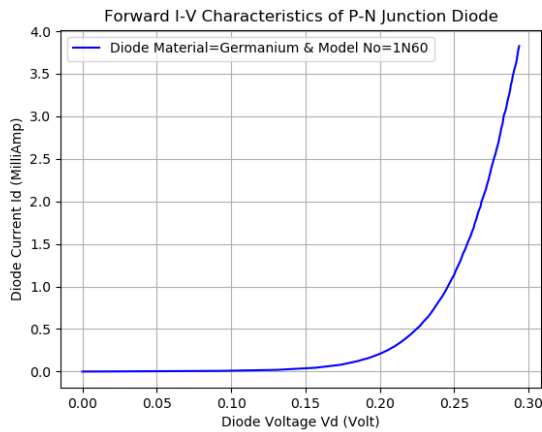


Figure 10: Si (1N4148) Diode Characteristics for K_B generated running Program (1).

The percentage error of measurement δ is also calculated by the program using the given relation:

$$\delta = \frac{v_A - v_E}{v_E} \times 100\% \quad (12)$$

Where v_A and v_E are the actual and exact value of any parameter being measured. Table (1) tabulates the value v_A obtained by the experiment, and the value of the

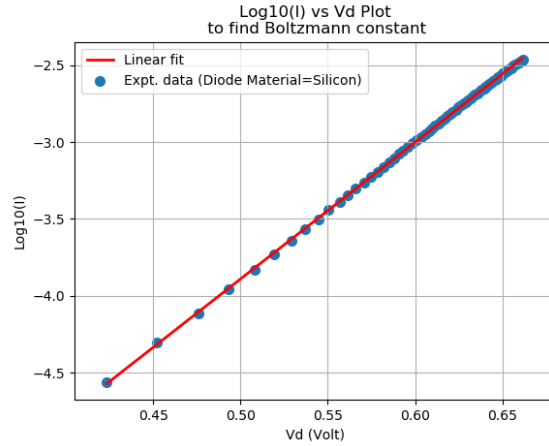


Figure 11: $\text{Log}_{10} I_s$ plot (Scatter), and fitted curve (Line) plot for Si (1N4148) Diode to find K_B , generated running Program (2).

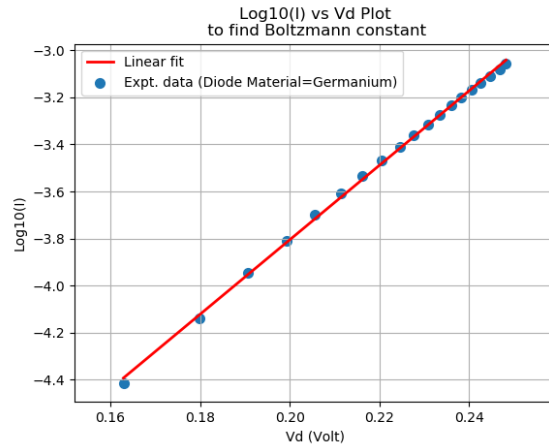


Figure 12: $\text{Log}_{10} I_s$ plot (Scatter), and fitted curve (Line) plot for Ge (1N60) Diode to find K_B , generated running Program (2).

percentage error δ in each case, given the exact value v_E of the Boltzmann Constant is $K_B = 1.38 \times 10^{-23} \text{ J K}^{-1}$.

Table (2) tabulates the values of the reverse saturation current I_s for each case. It appears from the data that in case of

Table 1: Data recorded for K_B

Model	Mat.	v_A of K_B in JK^{-1}	δ (%)
1N60	Ge	1.47×10^{-23}	6.72
1N4148	Si	1.30×10^{-23}	-5.35

Table 2: Data recorded for I_S

Model	Mat.	I_S in nano Amperes
1N60	Ge	114.11
1N4148	Si	4.59

Germanium, the amount of I_S is higher than that of a Si diode.

6 Discussions:

The experiment designed here is not new to undergraduate physics students, but the approach is. This method may provide an additional way to solve the problem, complementing the traditional hands-on experiment method, where each data point is col-

lected manually. The use of *python* is another interesting approach to design such age-old experiments in a new way. The authors are at opinion that the use of ExpEYES-17 in the undergraduate laboratories along with the traditional methods can definitely widen the horizon of learning.

Acknowledgments

The authors thank all those who have developed and contributed to the design of the kit, Especially, Dr. Ajith Kumar B. P., Satyanarayana V. V. V. of IUAC, and Jithin B. P. of Himachal Pradesh Open University. The book[3] on python has helped the authors a lot in coding in python.

Appendix

The Program listings are provided below.

```

1 # Data Collection for finding Boltzmann Constant
2
3 #Initialization of device
4 import eyes17.eyes
5 p = eyes17.eyes.open()
6 from pylab import *
7 import numpy as np
8 import matplotlib.pyplot as plt
9
10 #import libraries
11 import time
12 import math

```



```
13 import pickle
14
15 #Get the Material of the Diode being used
16 mat=input ("Type S for Silicon, or G for Germanium and press ENTER:")
17 if (mat=='S'):
18     eta=2
19     mat1='Silicon'
20     colname='red'
21     vdliml= 0.40          #the lower limit of the exponential part of
    the V-I plot.
22     vdlimh= 0.75          #the upper limit of the exponential part of
    the V-I plot.
23 elif (mat=='G'):
24     eta=1
25     mat1='Germanium'
26     colname='blue'
27     vdliml= 0.15          #the lower limit of the exponential part of
    the V-I plot.
28     vdlimh= 0.25          #the upper limit of the exponential part of
    the V-I plot.
29 else:
30     print('Diode Material not defined')
31
32 #Get the Number of the Diode being used
33 mod=input ("Type the diode model number and press ENTER:")
34
35 #define constants for Temperature recording
36 A = 3.9083e-3            #value of coefficient A
37 B = -5.7750e-7           #value of coefficient B
38
39 #measured values of offset and constant current source
40 a3offset = 0.0005908659726263022    #offset voltage at a3
41 iccs = 0.0011132813656059932        #ccs current in A
42
43 #other parameters for temperature recording
44 Rg = 977                    #value of 1k resisor
45 gain = 1 + 10000/Rg        #gain at a3 using Rg
46 r0 = 101.8                 #pt100 resistance at 0 deg c
47
```

```

48 #Finding the data for plot and calculation
49 vda = []          # Initiate Diode voltage list
50 ida = []          # Initiate Diode current list
51 lida = []         # Initiate log I list
52 lvda = []         # Initiate list for next plot
53 latemp = []       # Initialize list for Absolute Temperature recording
54
55 pv1=0.0
56 while (pv1 <= 4.00):
57     p.set_pv1(pv1)          #set bias voltage
58     p.set_state(CCS=1)     #enable ccs
59     time.sleep(0.025)      #give the device time to set the
    voltage/current
60     vs = pv1              #check bias voltage
61     vd = p.get_voltage('A1') #get Diode voltage
62     va3 = p.get_voltage('A3') #Read A3 for temperature
    measurement
63     vt = va3 - (gain * a3offset) #correct offset in measuring A3
64     vpt100 = vt/gain          # voltage across PT100
65     rt = vpt100 / iccs        #voltage for curent teperature
66     C = 1 - (rt - 1.8)/(r0 - 1.8) #find c and exlude offset
    resistor of pt100 wire
67     ctemp = (-A + math.sqrt( A*A - 4 * B * C ) ) / (2.0 * B) #find
    current temperature
68     atemp = 273.15 + ctemp    #find temperature in absolute
    scale
69     idx = (vs-vd)/969        #find Diode current for 1K Ohm in
    Amp
70     idma = idx*1000          #current in mA for I-V plot
71     if vdliml <= vd <= vdlimh: #non-linear section of the plot
72         if idx > 0 :         #discard negative value, if any
73             idlog = math.log10(idx) #convert I to log
74             lida.append(idlog)      #append in the list of id
75             lvda.append(vd)         #append in the list of vd
76             latemp.append(atemp)    #append in the list of absolute
    temperature
77     vda.append(vd)           #append in list for vd vs id plot
78     ida.append(idma)         #append in list for vd vs id plot
79     p.set_state(ccs=0)       #disable ccs

```

```

80     pv1 += 0.05                                #next value of source voltage
81
82
83 #save the lists in a file to calculate the Boltzmann Constant
84 with open("ivdata.txt", "wb") as f:
85     pickle.dump(lvda, f)                        #dump the vda values in a text
        file
86     pickle.dump(lida, f)                        #dump the log(10)ida values in a
        text file
87
88 #save the temeratures list in a file to calculate the Boltzmann
        Constant
89 with open("tempdata.txt", "wb") as f:
90     pickle.dump(latemp, f)                      #dump the temperature values in
        Kelvina in a text file
91
92 #Plot
93 plt.title('Forward I-V Characteristics of P-N Junction Diode') #Plot
        Title
94 plt.xlabel('Diode Voltage Vd (Volt)')          #x axis label
95 plt.ylabel('Diode Current Id (MilliAmp)')      #y axis label
96 plt.plot(vda, ida, color= colname, label='Diode
        Material='+str(mat1)+' & Model No='+str(mod)) #plot graph
97 plt.grid(True)                                #show gridlines
98 plt.legend(loc='upper left')                   #show legend
99 plt.show()                                    #show graph

```

Listing 1: Python Program to generate data to find Boltzmann Constant using p-n junction diode, and to plot the diode characteristics.

```

1 # Plotting of V - Log10I plot and calculation of Boltzmann Constant
2
3 #import libraries
4 import matplotlib.pyplot as plt
5 import scipy.optimize as opt

```

```
6 import numpy as np
7 import pickle
8 from math import *
9
10
11 #Get the Material of the Diode being used
12 print('Select the Material of the Diode being used')
13 mat=input ("Enter S for Silicon, or G for Germanium and press ENTER:")
14 if (mat=='S'):
15     eta=2
16     mat1='Silicon'
17     colname='red'
18
19 elif (mat=='G'):
20     eta=1
21     mat1='Germanium'
22     colname='blue'
23
24 else:
25     print('Diode Material not defined')
26
27 q = 1.602176634e-19          #standard value of unit charge
28 ks = 1.380649e-23           #standard value of k
29
30 with open("ivdata.txt", "rb") as f:      # Unpickling V-log10I data
31     Vd = pickle.load(f)
32     Id = pickle.load(f)
33
34 with open("tempdata.txt", "rb") as f:    # Unpickling Temperature data
35     atemp = pickle.load(f)
36     at=sum(atemp) / len(atemp)          # Find avg value of
37                                         Temperature in Kelvin
38
39 #intialization for curve fitting
40 x,y=[],[]
41 for i in range(len(Vd) ):
42     x.append(Vd[i])
43     y.append(Id[i])
44 n=len(x)
```



```
44
45 ## Linear equation : y=a+bx fit
46 def func(x,a,b):
47     return a+b*x
48 sx=0.0
49 sy=0.0
50 sxy=0.0
51 sx2=0.0
52 for i in range(n):
53     sx=sx+x[i]
54     sy=sy+y[i]
55     sxy=sxy+x[i]*y[i]
56     sx2=sx2+x[i]*x[i]
57 D=n*sx2-sx*sx
58 A=sy*sx2-sx*sxy
59 B=n*sxy-sx*sy
60 a=A/D
61 b=B/D
62 print ("a,b=",a,b)
63
64 fx=[]
65 for i in range(n):
66     fx.append(func(x[i],a,b))
67
68
69 k = q/(2.303 * eta * (at) * b)          #find k from fitted curve
70 pe = ((k - ks)/ks)*100
71 Is = 10 ** a
72 print("Temperature = ", at, "Kelvin")#print temperature during
    experiment
73 print("Boltzmann Constant=", k/1.0e-23, "X 10(-23) J/K") #print the
    value of k
74 print("Percentage Error= ", pe, "%")
75 print("Is= ", Is*1.0e9, "nA") # Print Is
76
77 ## Plot
78 plt.title('Log10(I) vs Vd Plot\n to find Boltzmann constant')
79 plt.xlabel('Vd (Volt)')
80 plt.ylabel('Log10(I)')
```

```

81
82 plt.scatter(x, y, label="Expt. data"+" (Diode
    Material="+str(mat1)+")", linewidth=2)
83 plt.plot(x, fx, "red", label="Linear fit", linewidth=2)
84 plt.grid(True)
85 plt.legend(loc="upper left")
86 plt.show()

```

Listing 2: Python Program to find Boltzmann Constant using curve fitting of a straight line based on the data obtained from program (1)

References

- [1] Kumar, A. B. P., and Satyanarayana, V. V. V. (2017). *The ExpEYES-17 User's Manual*. Inter-University Accelerator Center, New Delhi.
- [2] Kumar, A. B. P. *Python for Education*. The Phoenix Project, Inter-University Accelerator Center, New Delhi.
- [3] Brown, M. C. (2018). *Python: The Complete Reference* (4th ed.). McGraw Hill.
- [4] Rajaraman, V. (1999). *Computer Oriented Numerical Methods* (3rd ed.). Prentice-Hall India, pp. 119-121.
- [5] Singh, J. (2009). *Semiconductor Devices: Basic Principles*. Wiley India, pp. 476-477, ISBN 9788126511020.
- [6] Kumar, A. B. P., et al. (2009). "Innovative Science Experiments Using Phoenix." *Physics Education*, **44**, 469.
- [7] Luthra, V., Kaur, A., Saini, S., Jaglan, S., Dhasmana, S., Apoorva, and Kumar, A. B. P., 2016, "Evaluation of Boltzmann's constant: Revisit using interfaced data analysis," *Physics Education*, **32**(3), Article 2, Jul-Sep 2016.
- [8] Kumar, A. B. P. (2018). "Innovative Science Experiments Using ExpEYES-17." *IAPT Physics Education*, **35**(2), 32-40.
- [9] Satyanarayana, V. V. V. (2019). "Using ExpEYES-17 for Teaching Basic Physics Concepts." *IOP Physics Education*, **54**(6), 065015.
- [10] Roy, D. (2020). "Junction Field Effect Transistor Characteristics: A New Approach Using ExpEYES-17." *Physics Education*, **36**(3), April-June, ISSN 0970-5953.
- [11] Biswas, S., 2022, "Determination of the band gap of germanium and silicon using ExpEYES-17 kit," *Physics Education*, **57**(2), 025026. <https://doi.org/10.1088/1361-6552/ac3b95>
- [12] Biswas, S., 2022, "Study of Fourier series of user-defined waveforms us-

- ing ExpEYES-17 kit," *Physics Education*, **57**(3), 035008. <https://doi.org/10.1088/1361-6552/ac420a>
- [13] Biswas, S., and Roy, D., 2022, "Microcontroller-based study of diode thermometers for online demonstration of undergraduate laboratory classes in COVID-19 lockdown," *Physics Education*, **57**(4), 045011. <https://doi.org/10.1088/1361-6552/ac563f>
- [14] Roy, A., Roy, D., and Mallick, A., 2023, "Construction and remote demonstration of an inexpensive but efficient linear differential variable transformer (LVDT) for physics or electronics teaching during COVID-19 pandemic," *Physics Education*, **58**(1), 015007. <https://doi.org/10.1088/1361-6552/ac93de>
- [15] Roy, A., Ghosal, M., Mallick, A., Roy, D., and Biswas, B., 2024, "Construction and remote demonstration of an inexpensive but efficient experimental setup for studying self-inductance and mutual inductance between two coils," *Physics Education*, **59**(2), 025022. <https://doi.org/10.1088/1361-6552/ad22f3>
- [16] CERN OHL. Available at: <https://ohwr.org/cernohl>.
- [17] ExpEYES-17 Project Website. Available at: <https://www.expeyes.in>.
- [18] HyperPhysics, Georgia State University. Available at: <http://hyperphysics.phy-astr.gsu.edu/hbase/Solids/pnjun.html>.
- [19] Inter-University Accelerator Center Phoenix Project. Available at: <https://www.iuac.res.in/phoenix/>.
- [20] Nobel Prize in Physics 1918. Available at: <https://www.nobelprize.org/prizes/physics/1918/planck/biographical/>.
- [21] Wikipedia. Available at: <https://commons.wikimedia.org/w/index.php?curid=643178>.

Estimating gravimetric effects for ordinary bodies

Pier Franco Nali¹

¹Independent scholar, Cagliari, Italy.

pfnali@alice.it

Submitted on 27-04-2022

Abstract

The main topic of this article is a discussion about the best way to show students that the proportionality of mass and weight, strictly true for point-like particles, is an excellent approximation for objects of “normal” size. The usual way of addressing this issue, although very simple, is not entirely satisfactory. Our approach considers first and second order, coordinate dependent, gravimetric effects, connected to the internal geometry of objects; these effects, extremely small, are estimated through examples.

and the proportionality of mass and weight cannot be regarded as strictly exact. Anyway, discrepancies involved are generally tiny and can safely be overlooked in most situations we commonly experience.

It should be deemed that teachers usually show students that for common bodies the acceleration of gravity does not vary appreciably (or, otherwise said, the gravitational field is uniform) within the size of the object, which can be done very simply. At least at the college level, but also at the high-school level, after having presented relative motions and introduced the “apparent forces” in non-inertial reference frames - or “inertial forces” as we want to call them - teachers specify that the weight force on the earth’s surface is the resultant of the (true) gravitational force and of the apparent (inertia) forces, in particular the centrifugal force if the body is stationary.

Limiting the problem to the gravitational component only (let’s say it \mathbf{G}) teachers follow the usual simple path of differentiation $\Delta G/G = \Delta R^{-2}/R^{-2} = -2\Delta R/R = -2h/R$, where R is the radius of the earth and h is the height of the body (or, if stu-

1 Introduction

Direct proportionality of mass and weight is a well-established principle, proven as an experimental fact for all bodies in the same place. However, apart from the special case of uniform gravitational field, this principle is only valid locally, that is for point particles. When both the variability of the gravitational field and the bodies’ internal structure cannot be ignored, the point-particle approximation fails

dents do not know the differentials, simply calculate $G(R+h)/G(R) = R^2/(R+h)^2$. Since the difference in the first order is already very small, $2h/R$ typically being of the order of 10^{-8} , that of the second order is obviously negligible for any practical effect.

But we immediately realize that this method is flawed as the inverse proportion of the gravitational force to the square of the distance is strictly valid for point particles (and in the other special case of centrally-symmetric homogeneous bodies). In this way we implicitly assume the conclusion, as affirming uniformity of the gravitational field within point objects is tautological: our argument contains a circularity and runs into a logical fallacy (*petitio principii*). Furthermore, properly speaking, h does not correlate to the object size: it merely represents a (small) displacement of the body (or rather, its center of mass) from the surface of the earth.

Thus the argument is ineffective for extended bodies of arbitrary shape as it does not properly capture the effect on weight force (in either a rotating or a non-rotating frame) of the variation of \mathbf{g} within the object size; it would be preferable to find a different approach, allowing us to address the topic in broader generality and rigour.

2 Background

According to the “Declaration on the unit of mass and on the definition of weight; conventional value of g_n ”: *«The word “weight” denotes a quantity of the same nature as a*

“force”: the weight of a body is the product of its mass and the acceleration due to gravity; in particular, the standard weight of a body is the product of its mass and the standard acceleration due to gravity.» [1] And, contextually: *«The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram.»*¹ The value adopted for the standard acceleration due to gravity (on earth) is $g_n = 980.665 \text{ cm s}^{-2}$. Thus, for the weight force:

$$\mathbf{w} = m\mathbf{g}, \quad (1)$$

where m is the mass of the object and \mathbf{g} is the acceleration vector due to gravity (We denote vectors, like \mathbf{g} , \mathbf{w} , \mathbf{G} , as bold letters and represent their magnitudes, like g , w , G , as italic letters).

More generally, weight means the gravitational force (or this plus the centrifugal force) on a small mass compared to that of the source (e.g. “the weight of astronauts on the Moon”).

It is important to notice that “weight” and “gravitational force” are the same force but the use of either term is contextual and it is good practice to adhere to conventions on their use to avoid ambiguity. Calling the gravitational force on a celestial body “weight” creates confusion, and contradicts the convention that reserves this word for practical use (on weight vs gravitational force see e.g. [2]).

Hence the use of “weight” should be re-

¹The definition of the kilogram in terms of the international prototype is obsolete and no longer in force since 20 May 2019; it has been redefined in terms of the Planck constant.

served for the force experienced by an object with mass in a gravitational field (e.g. light is bent by gravitational fields, although it has not weight, because it is massless).

We are now ready to introduce the alternate differentiation pathway $\Delta \mathbf{g}/g = 1/g (\Delta \mathbf{g}/\Delta R) \Delta R = R/g (\Delta \mathbf{g}/\Delta R) h/R$, where h is now the linear size of the object and $\mathbf{g}(R)$ (gravitational + centrifugal) does not have a form given *a priori*. In this way we see that the less restrictive condition that the gradient $\Delta \mathbf{g}/\Delta R$ is the same order of \mathbf{g}/R would fulfill the requirement of uniformity of the gravitational field within the size of the object.

In a more formal way, if we think of the body made up of particles of masses m_i , being $m = \sum_i m_i$ the mass of the entire body and $\rho(\mathbf{r}) = \sum_i m_i \delta(\mathbf{r} - \mathbf{r}_i)$ its density, where δ is the Dirac delta function, the weight force is the generalization of eq. (1) by integration over the whole space:

$$\mathbf{w} = \iiint \rho(\mathbf{r}) \mathbf{g}(\mathbf{r}) d\tau = \sum_i m_i \mathbf{g}(\mathbf{r}_i). \quad (2)$$

For an uniform field ($\mathbf{g}(\mathbf{r}) = g\hat{\mathbf{z}}$) the equation (2) reduces to (1) and the internal size and geometry of the body are irrelevant.

But the uniformity condition for \mathbf{g} only holds approximately near the earth's surface. The earth's gravitational field is not uniform even on a small scale; modern gravimeters allow us to appreciate g with eight or nine significant digits (some μgals); this is how to say that variations in the earth's gravity between points even a few centimeters apart are detectable instrumentally.

In this article we consider very small (some ppb) coordinate dependent effects. At these scales there are several others effects, both instrumental and environmental, which are not so weak. For example, \mathbf{g} has a dependence on time: the effect of terrestrial tides alone is two orders of magnitude greater (a few hundreds μgals), to which are to be added the effects of tides in the oceans, the hydrological and barometric components, also variable over time, and so on; also the motion (if any) of the measuring instrument has to be taken into account. Moreover, although independent of the mass m of the body under consideration, \mathbf{g} generally depends on "other" masses; it will be assumed that these external masses vary very slowly.

Obviously we should not forget to mention the major non-gravitational contribution to weight, that of the centrifugal force due to diurnal rotation (which is a component of \mathbf{g}); the centrifugal force is some part per thousand of the gravitational force and the dependence of the two forces on the distance, from the center or from the axis, is different; the effect on the weight of Archimedes' thrust in the air is also significant. In the following we shall leave all these effects just mentioned aside from the present study and focus our attention on coordinate related ones, dependent on size and geometric configuration of objects.

It is easy to see that for bodies for which experiments can be established, such as for bodies near the earth's surface, these coordinate-dependent effects are far too small for the standard resolution of dynamometers

and scales. To this end we take Taylor series expansions truncated to the second-order of the functions $\mathbf{g}(\mathbf{r}_i)$ in eq. (2) centered about $\mathbf{R}_{cm} = \frac{1}{m} \sum_i m_i \mathbf{r}_i$, that is the radius vector

conducted from the origin of the coordinates to the body center of mass:²

$$\mathbf{g}(\mathbf{r}_i) = \mathbf{g}_{cm} + (\mathbf{r}_i - \mathbf{R}_{cm}) \cdot \nabla \mathbf{g} \Big|_{\mathbf{R}_{cm}} + \frac{1}{2} ((\mathbf{r}_i - \mathbf{R}_{cm}) \cdot \nabla)^2 \mathbf{g} \Big|_{\mathbf{R}_{cm}} \quad (3)$$

(where \mathbf{g}_{cm} is evaluated in the center of mass).

By fixing the origin of the coordinates in the center of the earth, the equation (3) is quite exact for all applications in which we need to evaluate the weight force on bodies located on or near the earth's surface. In fact, under these conditions $|\mathbf{r}_i - \mathbf{R}_{cm}| \ll R$ with $|\mathbf{R}_{cm}| \cong R$ (assuming $R \simeq 6.371 \times 10^6$ m for the earth's mean radius) and the third and higher-order terms can be overlooked.

3 The first order effect. Implications for precision mass measurements

For a (small) displacement of a body from P to a near point P' we can express the variation of \mathbf{g} within the body by means of the eq. (3) as:

$$\mathbf{g}(\mathbf{r}'_i) = \mathbf{g}(\mathbf{P}') + (\mathbf{r}'_i - \mathbf{P}') \cdot \nabla \mathbf{g} \Big|_{\mathbf{P}'} \quad (4)$$

(overlooking the small second order term). For a rigid body holds the distance preserving condition $\|\mathbf{r}'_i - \mathbf{P}'\| \equiv \|\mathbf{r}_i - \mathbf{P}\|$. For the

sake of simplicity we assume the special condition of a purely translational displacement (preserving distance, angle, sense, and orientation) such that $\mathbf{r}'_i - \mathbf{P}' \equiv \mathbf{r}_i - \mathbf{P}$, so that, if we take P as the center of mass, the second term of the right-hand side vanishes identically by introducing eq. (4) in eq. (2), and eq. (4) reduces to

$$\mathbf{g}(\mathbf{P}') = \mathbf{g}_{cm} + (\mathbf{P}' - \mathbf{R}_{cm}) \cdot \nabla \mathbf{g} \Big|_{\mathbf{R}_{cm}}. \quad (5)$$

In turn, eq. (2) reduces to

$$\mathbf{w}' = \mathbf{w}_{cm} + \mathbf{w}_{-1}, \quad (6)$$

where

$$\mathbf{w}_{-1} = m (\mathbf{P}' - \mathbf{R}_{cm}) \cdot \nabla \mathbf{g} \Big|_{\mathbf{R}_{cm}}. \quad (7)$$

Applying the gradient criterion $\nabla \mathbf{g} \sim \mathbf{g}/R$ we easily obtain $w_{-1}/w_{cm} \sim h/R$, so, for a body similar in size and mass to the obsolete kilogram prototype ($h \sim$ a few centimeters) $w_{-1} \sim 10^{-7} - 10^{-8}$ N.

In recent decades, in view of the redefinition of the SI units, in particular the kilogram, the goal set by the CGPM was to achieve accuracy of the order of 10^{-8} . Laboratories

²Under suitable analyticity conditions for the functions $\mathbf{g}(\mathbf{r}_i)$.

such as the National Physical Laboratory in the UK and numerous other metrology labs around the world have worked for years to achieve the required accuracy with Kibble's balances. At the meeting of 17th May 2019 of the CCM, I. A. Robinson (NPL) stated: «*Whilst, at present, it is theoretically possible to measure the principal quantities to around 2–3 parts in 10⁹ a number of other effects in the apparatus must be taken into account.*» [3] This is a number of practical reasons, which limit accuracy, such as alignments, vibrations, etc. The NPL also plans to develop simpler Kibble's balances, affordable and operable in laboratories not as highly-specialized as NPL, capable of 10^{−8} accuracy. At this accuracy level, a number of systematic effects has to be taken into account, including gravimetric contributions (see e.g. [4]). An historical account of the development of these sensitive balances in the context of the proposed reform of the SI is outlined in [5].

Apparently, gravimetric effects such as those we are talking about were first considered in the early 1970s in connection with the development at the National Bureau of Standards of the “One Kilogram Balance” NBS No. 2, whose standard deviation was approx. 4 µg. [6]

Such kind of balances, used for comparing masses, compare the attractive gravitational forces between weights (or loads) and the earth. It is assumed (often implicitly) that these forces are exactly proportional to the masses of the loads (in vacuum) and do not vary during the measurement. The force

on a standard weight used for the comparison of masses depends on the distance from the center of the earth to the center of gravity of the weight.³ A second weight, of a different configuration, may have its center of gravity at a different distance from its base and thus the distance of the weight's center of gravity from the center of the earth will be different when the weight is placed on the weighing pan (which operates with the bases of the weights to be compared virtually on the same level). In this way, the constant of proportionality between the gravitational forces and the masses of the weights on the pan will be slightly altered, leading to a systematic error in the results of the comparisons between the masses, the so-called “gravitational configuration effect” introduced by Almer and Swift. [7]

If we consider a reference weight $w_r = m_r g(R)$ and a second equal weight w_x , whose centers of gravity are spaced by a distance $\Delta h = d$ above their bases, then, from eqs. (6) and (7):

$$w_x = w_{x,cm} + w_{x,-1},$$

$$w_{x,cm} = m_x g(R),$$

$$w_{x,-1} = w_{x,cm} \cdot \frac{1}{g} \frac{\partial g}{\partial h} (\Delta h),$$

³We anticipate here the notion of center of gravity that we will resume later. To practical effects of the discussion carried out in this section we can consider the center of gravity coincident with the center of mass, although the two concepts, in general, are to be kept distinct.

or, in the approximation of the gravitational component alone,

$$w_{x,-1} = -w_{x,cm} \cdot \frac{2d}{R}. \quad (8)$$

And, having imposed $w_x = w_r$,

$$m_x = m_r + m_{-1} = m_r + \frac{2dm_r}{R}. \quad (9)$$

The term $m_{-1} = 2dm_r/R$ in eq. (9) is the (first order) corrective term that must be applied to the mass of the second weight to take into account the difference in the force of gravity on the weights placed on the weighing pan of the balance whose centers of gravity are at different distances from their bases. The corrective term can be evaluated independently of the equation (8), valid in the approximation of the gravitational component alone, by directly measuring the acceleration of the free fall g and the gradient of the gravitational field $\partial g/\partial h$ in the place in which the mass calibration takes place.

In the case $d = +1\text{ cm}$ the correction for the comparison of nominal weights of 1 kg calculated using the equation (9) is approx. $+3\text{ }\mu\text{g}$. The old Pt-Ir kilogram prototype (density 21.55 kg/dm^3) is a right circular cylinder with a volume of approx. 46.5 cm^3 and approx. the same height (39 mm) as the diameter. Stainless steel samples (density 8.00 kg/dm^3), having volume (125 cm^3) respecting the same proportions, have a height of 54.2 mm. The resulting distance of the samples' centers of mass (/gravity) from their base is higher than that of the prototype's center of mass from its base by an amount of 7.6 mm, which leads to a correction of $+2.4\text{ }\mu\text{g}$. For

comparison, as Almer and Swift stated: «*Currently, mass comparisons at the 1-kg level can be carried out with standard deviations as small as 1.5 parts in 10^9 .*» [7]

This correction is far from being the most significant; the largest volume ($\approx 80\text{ cm}^3$) of the stainless steel 1-kg samples results in a correction for the aerostatic thrust of approx. $+94\text{ mg}$ (assuming an air density of 1.2 kg/m^3), that is about 40,000 times the gravitational effect. [8] Nonetheless the gravitational correction becomes significant for high precision mass measurements. In fact, accuracy is limited not only by the achievable precision and uncertainty associated with the value of the sample, but also by systematic errors. It can be said that the accuracy of the results of the measurements is achieved only after all the relevant systematic errors have been identified and evaluated. This implies that in the design of an experiment all factors, even those that at first appear small, must be estimated to establish their potential importance as systematic factors affecting the measured results.

4 The second order effect

By introducing eq. (3) into eq. (2) and noticing that the first order term vanishes identically for the choice of \mathbf{R}_{cm} , eq. (2) reduces to

$$\mathbf{w} = \mathbf{w}_{cm} + \mathbf{w}_{-2},$$

where

$$\mathbf{w}_{cm} = m\mathbf{g}_{cm}$$

and

$$\mathbf{w}_{-2} = \frac{1}{2} \sum_i m_i ((\mathbf{r}_i - \mathbf{R}_{cm}) \cdot \nabla)^2 \mathbf{g}|_{\mathbf{R}_{cm}}$$

(or, in the continuous limit)

$$= \frac{1}{2} \iiint \rho(\mathbf{r}) ((\mathbf{r} - \mathbf{R}_{cm}) \cdot \nabla)^2 \mathbf{g}|_{\mathbf{R}_{cm}} d\tau.$$

The term \mathbf{w}_{cm} is the weight force acting on the material point to which the body is reduced, having the mass of the body and located in its center of mass.

The term \mathbf{w}_{-2} is a second order gravitational correction that takes into account the effect of the internal geometry of the body, estimated as follows:

$$w_{-2} = w_{cm} \cdot \frac{1}{2} \cdot \frac{1}{g} \frac{\partial^2 g}{\partial z^2} (\Delta z)^2,$$

or, in the purely gravitational component approximation,

$$w_{-2} = w_{cm} \cdot \frac{3d^2}{R^2}$$

(where $\Delta z = d$ is the linear size of the object). It represents the difference due to mass distribution around the center of mass compared to the situation in which all the mass is thought to be concentrated in one point. More formally, it can be shown (see e.g. [9]) that the mass distribution intervenes to second order through the inertia tensor of the body. For a right circular cylinder of mass 1 kg a few centimeters high, like a copy of the old Pt-Ir kilogram prototype, the order of magnitude of the w_{-2} term is $\sim 10^{-16} - 10^{-17}$ N, the same of the weight of the equivalent mass of 1 joule, just 1/10 of that of the mass of an *Escherichia coli* bacterium and one hundred

thousand times smaller than that of the mass of a human cell.

Although fully negligible for bodies of ordinary mass near the surface of the earth, similar but a bit more significant effects occur in various kinds of problems, often faced with methods borrowed from celestial mechanics; in these situations, all the possible contributions must be carefully evaluated both in theoretical analyses and in the design of the experiments. A typical example are tidal phenomena, whose effects depend on the gradient of the gravitational field, rather than on intensity, and the variations of the gravitational force from one part of the object to the other must be considered. Meanwhile, there is no doubt that in these situations the bodies cannot be thought of as material points; Newton had already noticed that the exact results obtained for point-like particles are only approximate in presence of gravitational force between extended bodies attracting at short distances. In celestial mechanics it is usually satisfactory to stop calculations at the second order of approximation.

Moreover, as the size of the objects under consideration are on a planetary or sub-planetary scale, i.e. a significant fraction (say, from a few thousandths to a few hundredths) of the earth's radius (think, for example, of lithosphere segments of which we want to study the isostatic conditions), or when the bodies are very close to an attracting center (a situation encountered in geophysical and astrophysical contexts), also the assumptions under which the equation (3) holds can fail

and additional contributions should be considered.

In addition, sometimes it is not even possible to set up experiments or carry out direct measurements; when this occurs, the evaluation of gravitational forces needs *ad hoc* modeling of objects, which may require, for example, the computation of quadruple or sextuple integrals and numerical integration (see, for example, [10]).

5 The elusive center of gravity. Near-uniform field

The slight variation of the gravitational field within the size of earthly objects brings us to the interesting questions of the parallel field and the center of gravity.

The earth's gravitational field can be locally modeled by a field consisting of parallel vectors of (slightly) non-uniform intensity. This picture is useful because it allows us to introduce the “scalar weight” w in a coherent way,⁴ providing a tool to face and clarify the problem of determining a unique point (if any) where you can think applied the total weight force acting on all the particles of the body, i.e. its center of gravity.⁵ A real gravitational field cannot be both parallel and non-

uniform at the same time. It is convenient to examine the case of the near-uniform field, which, in addition to being simplistic, reproduces the gravitational field near the earth's surface with an excellent degree of approximation. Furthermore, with this choice, the problem can be dealt with in one dimension. For the usual central field

$$\mathbf{g}(\mathbf{r}) = -k \frac{\mathbf{r}}{\|\mathbf{r}\|^3}$$

($k = GM$ for the earth's gravitational field) $\nabla \cdot \mathbf{g} = 0$ everywhere. In the near-uniform model we consider a small cylindrical region where there is a field of vectors parallel to $\hat{\mathbf{z}}$, having non-uniform modulus, so defined:

$$\mathbf{g}(\mathbf{r}) = g(z)\hat{\mathbf{z}} = -kz^{-2}\hat{\mathbf{z}}, \quad (10)$$

with $z \gtrsim R$.

The equation (10) does not represent a real Newtonian gravitational field as \mathbf{g} does not have zero divergence. However, for z large enough, i.e. far from the center of the field (e.g. near the earth's surface), the divergence is small and the eq. (10) is a very good approximation, locally (far from the center of the earth), of a gravitational field generated by a spherically symmetric mass distribution.⁶ In this framework, the center of gravity of a body can be defined through the “equipollent” moment condition (see [11], p. 18). The moment

⁴“scalar” here does not mean invariant under rotation; here we intend 1-dimensional 1-component scalar field.

⁵The center of gravity is susceptible to other definitions, which we will not deal with here. A definition different from that of the weighted average can be given, for example, in the case of the spherically symmetric field.

⁶We assume the simplified picture of spherical earth, uniform density, not rotating; we abstract from all possible disturbing factors (assuming absence of air, no influence of celestial bodies, etc.). The influence of the body under examination on the central gravitational field is also assumed (external field approximation).

of a single force on a particle is perpendicular to the force and the vector radius from the co-ordinate origin to the position of the particle. In general, however, this is not true for a system of forces; the total moment of a system of forces around a point O (the pole, which we will also assume as the origin of the coordinates) is generally not perpendicular to the total force vector acting on the system.

The moment \mathbf{T}_{eq} of the system of forces equipollent to a single weight force \mathbf{w} acting on the body satisfies the vector equation

$$\mathbf{T}_{eq} = \mathbf{R}_{cg} \times \mathbf{w}, \quad (11)$$

where \mathbf{w} is the total weight force acting on the body, defined by eq. (2) and \mathbf{R}_{cg} is the radius vector joining the pole with the point of application of this force, i.e. with the body center of gravity. The total moment of the forces acting on the system is by definition $\mathbf{T} = \sum_i \mathbf{r}_i \times \mathbf{w}_i$, and the total weight force $\mathbf{w} = \sum_i \mathbf{w}_i$. Imposing the perpendicularity condition to these two vectors is equivalent to making the equation (11) valid for \mathbf{T} , that we rewrite as

$$\sum_i (\mathbf{r}_i - \mathbf{R}_{cg}) \times \mathbf{w}_i = \mathbf{0}. \quad (12)$$

The equation (12) (*torque equation*) does not have solution if \mathbf{T} and \mathbf{w} are not orthogonal (and neither is zero) and in this case the center of gravity vector $\mathbf{R}_{cg} \equiv (X, Y, Z)$ cannot be determined by this method. We do not examine here the existence conditions of the solutions of the *torque equation*, whose detailed discussion can be found, for example,

in [12]. Fortunately, in the special case of parallel field the orthogonality condition is met.⁷ If we choose the z -axis in the direction of the field, then $\mathbf{w}_i = w_i \hat{\mathbf{z}}$ and eq. (12) reduces to the linear system

$$\begin{cases} \sum_i (x_i - X) w_i = 0, \\ \sum_i (y_i - Y) w_i = 0. \end{cases}$$

The moment of total weight force will have only the x and y components different from zero, from which the X and Y components of the vector \mathbf{R}_{cg} can be calculated; these define the line of action of the total weight force. There remains the z component to be determined (the *torque equation* for the z component is a null identity). We observe, however, that under the assumptions made the equation (12) can be rewritten as

$$\left(\sum_i w_i \mathbf{r}_i - w \mathbf{R}_{cg} \right) \times \hat{\mathbf{z}} = \mathbf{0}. \quad (13)$$

Then, as the pole O can be chosen arbitrarily and $\hat{\mathbf{z}}$ is a fixed vector, the equation (13) can be satisfied by choosing the vector \mathbf{R}_{cg} defined as (see [11], p. 48)

$$\mathbf{R}_{cg} = \frac{1}{w} \sum_i w_i \mathbf{r}_i = \frac{1}{w} \sum_i m_i g(\mathbf{r}_i) \mathbf{r}_i \quad (14)$$

or, in the continuous limit,

$$\mathbf{R}_{cg} = \frac{1}{w} \iiint \rho(\mathbf{r}) g(\mathbf{r}) \mathbf{r} d\tau, \quad (15)$$

which constitute the definition of the center of gravity in the case of parallel field. For a uniform field the equation (15) becomes

$$\mathbf{R}_{cm} = \frac{1}{m} \iiint \rho(\mathbf{r}) \mathbf{r} d\tau, \quad (16)$$

⁷Another case in which this condition is met is that of a planar system of forces.

that is \mathbf{R}_{cg} coincides with the center of mass radius vector. In the equations (15) and (16) it is implied that

$$w = \iiint \rho(\mathbf{r})g(\mathbf{r}) d\tau,$$

$$m = \iiint \rho(\mathbf{r}) d\tau.$$

With series expansions of w and $g(\mathbf{r})$ in eq. (15) around the center of mass,⁸ using the

equations (3) and (10), and truncating after the first order, we have (we omit the detailed steps):

$$\mathbf{R}_{cg} = \mathbf{R}_{cm} + \mathbf{R}_{-1} + \cdots,$$

$$\mathbf{R}_{-1} = -\frac{2}{mZ_{cm}} \iiint \rho(\mathbf{r}) (z - \mathbf{R}_{cm} \cdot \hat{\mathbf{z}}) (\mathbf{r} - \mathbf{R}_{cm}) d\tau. \quad (17)$$

If as an example we consider a solid in the shape of a right cylinder or a rectangle parallelepiped, very elongated with respect to its basis, resting on the earth's surface so as to approach the situation of a parallel and near-uniform field, we reduce the problem to one dimension. If h is the height of the solid, the z -coordinate of its center of mass will be

given by $Z_{cm} = R + h/2$; we also express the variable of integration as a function of the coordinate in the system of the center of mass $\zeta = z - Z_{cm}$; finally, for simplicity, suppose the solid of uniform density ρ . Then we can write the Z coordinate of the center of gravity as

$$Z = Z_{cm} + Z_{-1} + \cdots = Z_{cm} - \frac{2}{hZ_{cm}} \int_{-h/2}^{h/2} \zeta^2 d\zeta + \cdots = Z_{cm} - \frac{h^2}{6Z_{cm}} + \cdots.$$

The term $Z_{-1} = -h^2/6Z_{cm} \cong -h^2/6R$ in the first order of approximation represents the displacement of the center of gravity apart from the center of mass. This is a tiny difference: in the case of Dubai's Burj Khal-

ifa, currently the tallest building in the world ($h = 829.80$ m), the center of gravity is only about 2 cm below the center of mass! The center of gravity is a specially elusive concept. It identifies a defined point, but, unlike the center of mass, it does not have a definite position. Its position depends, in general, on the

⁸See previous note 2.

relative positions of the body under consideration and the attractive mass. As can be seen from the equation (17), when the distance $Z_{cm} \cong R$ of the body from the center of the earth increases, the center of gravity approaches the center of mass. This feature makes it difficult to work with the center of gravity and in practice this concept is seldom used. The detailed treatment of this and other interesting problems related to the center of gravity is beyond our scope; an introductory discussion on these topics can be found on the Wikipedia page “*Centers of gravity in non-uniform fields*”⁹ and related talk,¹⁰ to which the interested reader is referred.

6 Conclusions

We have established that the proportionality of mass and weight for ordinary bodies can be taken as an excellent approximation in all cases of practical interest. However, it is advisable for students to always clarify the limits of validity of this approximation, both in their theoretical meaning and for the aspects related to the sensitivity of the experiments. For this purpose, the gradient criterion $\Delta \mathbf{g} / \Delta R \sim \mathbf{g} / R$ is suitable for exploring the variation of the gravitational force within the size of the body. While it is easy to show that this gravimetric effect is negligible for ordinary bodies, special caution should be observed when, in investigating certain areas, you go beyond the

validity range of the point particle approximation. In geophysics, hydrostatics and astrophysics various situations are encountered of strongly inhomogeneous gravitational field and the gravitational effects connected to the internal geometry of the bodies cannot be neglected. Such effects must be carefully considered; for example: in celestial mechanics and astrodynamics, in the calculation of the short-distance interaction of non-spherical shaped bodies (see, e.g., [13–15]); in geophysics, in the calculation of the gravimetric field of a polyhedral plate (see, e.g., [16–18]); in hydrostatics, in the computation of the thrust, where the pressure gradient is replaced by the product of the density of the fluid and the gravitational field (see, e.g., [19]). These problems are addressed on a case-by-case basis and often require the development of specific solutions.

Acknowledgments

I thank Santo Armenià for drawing my attention to this topic. I also thank the reviewer for helpful comments/suggestions.

⁹https://en.wikipedia.org/wiki/Centers_of_gravity_in_non-uniform_fields

¹⁰https://en.wikipedia.org/wiki/Talk:Centers_of_gravity_in_non-uniform_fields

References

- [1] CGPM, Proceedings of the 3rd CGPM (1901), 1901, p. 70.
- [2] I. Galili, Int. J. Sci. Educ. 23, 1073 (2001).
- [3] I. A. Robinson, 17th meeting of the CCM (17th May 2019).
- [4] I. A. Robinson and S. Schlamming, Metrologia 53, A46 (2016).
- [5] R. P. Crease, Phys. World 24 (03), 39 (2011).
- [6] H. E. Almer, J. Res. Natl. Bur. Stand. (U. S.) 76C (01-02), 1 (1972).
- [7] H. E. Almer and H. F. Swift, Rev. Sci. Instrum. 46, 1174 (1975).
- [8] Z. J. Jabbour and S. L. Yaniv, J. Res. Natl. Inst. Stand. Technol. 106, 25 (2001).
- [9] D. Hestenes, New Foundations for Classical Mechanics (Kluwer, Dordrecht, 2002), p. 520.
- [10] J. Stirling, New J. Phys. 19, 073032 (2017).
- [11] M. F. Beatty, Principles of Engineering Mechanics, Volume 2 (Springer, New York, 2016).
- [12] K. R. Symon, Mechanics, 3rd edition (Addison-Wesley, Reading, 1971).
- [13] J. Ashenberg, Celestial Mech Dyn Astr 99, 149 (2007).
- [14] Y. Shi, Y. Wang and S. Xu, Celestial Mech Dyn Astr 129, 307 (2017).
- [15] X. Hou, D. J. Scheeres and X. Xin, Celestial Mech Dyn Astr 127, 369 (2017).
- [16] B. Banerjee and S. P. Das Gupta, Geophysics 42, 1053 (1977).
- [17] D. Nagy, Geophysics 31, 362 (1966).
- [18] R. Karcol and R. Pašteka, Pure Appl. Geophys. 176, 257 (2019).
- [19] F. M. S. Lima and F. F. Monteiro, Rev. Bras. de Ensino de Fis. 35, 3701 (2013).

Why is a 60° prism preferred for observing dispersion

Gautham Dathatreyan.¹, and K. M. Udayanandan³

¹ Neettiyath Kalam, Palakkad, 679 536

² Sree Narayana College, Vadakara, Kerala,
udayanandan@gmail.com

Submitted on 12-05-2023

Abstract

Prisms are commonly used in classrooms to study refraction and dispersion, with 60° prisms being the most prevalent in school and college laboratories. In this article, we explore the reasons behind their widespread use. Using Cauchy's law, we determine the refractive indices for violet and red light for a selected prism. Applying Snell's law and the principle of total internal reflection, we demonstrate that a 60° prism provides optimal dispersion, maximizing the separation between red and violet rays over a broad range of incidence angles.

persion led to new techniques for determining refractive indices [2, 3, 4, 5, 6]. Despite these advancements, equilateral prisms remain the standard in school and college laboratories, with textbooks consistently depicting 60° prisms in sections on refraction [7, 8]. However, the reasoning behind this preference is unclear. Do these prisms offer superior dispersion? In this article, we explore this question, beginning with a review of refraction through a prism in the next section.

1 Introduction

In 1666, Sir Isaac Newton [1] investigated the effect of passing light through triangular glass prisms with different refracting angles. His experiments led to the conclusion that white light consists of a spectrum of colors. Over time, extensive studies on dis-

2 Refraction through a prism

In Figure 1, we can see that A is the angle of the prism, i_1 is the angle of incidence, i_2 is the angle of emergence, r_1 is the angle of refraction at the first side and r_2 is the angle of refraction for the second side. For refraction of light through a prism we have $r_1 + r_2 = A$ and deviation[7]

$$d = i_1 + i_2 - A \quad (1)$$

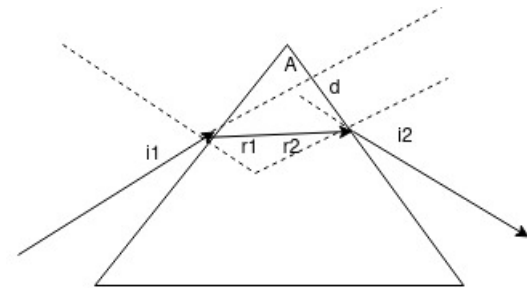


Figure 1: Refraction through a prism

From Snell's law we have $\sin i_1 = \mu \sin r_1$ and $\sin i_2 = \mu \sin r_2$, where μ is the refractive index of the prism. Then substituting for i_2 and r_2

$$d = i_1 + \sin^{-1} [\mu \sin(A - r_1)] - A$$

Finally substituting the value of r_1 , the angle of deviation

$$d = i_1 + \sin^{-1} \left[\mu \sin \left(A - \sin^{-1} \left(\frac{\sin i_1}{\mu} \right) \right) \right] - A \quad (2)$$

The Equation(2) shows that deviation depends only on refractive index, angle of incidence and angle of the prism. Next we will obtain refractive index for some colors.

3 Refractive index for red and violet rays

For each color we have different refractive index given by Cauchy's law[4]

$$\mu = A + \frac{B}{\lambda^2}$$

where A and B are Cauchy's constants. By using minimum deviation method, we can find A and B and then μ for different λ . Using a prism available in the laboratory, we

got the refractive index for red and violet light as

$$\mu_r = 1.40 \quad (3)$$

$$\mu_v = 1.58 \quad (4)$$

respectively. These values will be different for different prisms. We are interested in finding the deviation of red and violet rays which are related with μ . Hence we found μ only for red and violet. Next we will find whether for any A and angle of incidence, dispersion is possible.

4 Total internal reflection and absence of dispersion for some angle of incidences

When the light rays travel from an optically denser medium to a rarer medium, for a particular angle of incidence called critical angle C given by the relation $\sin C = \frac{1}{\mu}$ is reached, the ray will be totally internally reflected. We will check whether this will happen for red and violet rays for different angled prisms. Using the μ given by Eqn(3) and (4) the critical angle for red and violet are

$$C_{red} = 45.58^\circ$$

$$C_{violet} = 39.26^\circ$$

Using the expression $r_1 = \sin^{-1} \left(\frac{\sin i_1}{\mu} \right)$ we can get the refraction angle for the first surface and using $r_1 + r_2 = A$, we can find the refracting angle for the second surface r_2 . Let us check whether r_2 matches with the critical angle values for different angle

of incidences. We will find this for $i = 30^\circ, 35^\circ, 40^\circ$ and 60° for different A 's.

4.1 For $i = 40^\circ$

From the Table 1 we can observe that when

Table 1:

A in degree	$\mu = 1.4$ r_2 in degree	$\mu = 1.58$ r_2 in degree
10	-17.33	-18.14
20	-7.33	-8.14
30	2.67	1.86
40	12.67	11.86
50	22.67	21.86
60	32.67	31.86
70	42.67	41.86
80	52.67	51.86

A is 70° , the violet ray gets totally internally reflected and hence there will be no dispersion. The negative sign in the refracted angle for small prisms is due the refraction to the other side of the normal compared to the side of the incident ray.

4.2 For $i = 60^\circ$

The value of r_2 for different angles of the prism are given in Table 2. Here violet light gets totally internally reflected for 80° prism. We see that for larger angle of incidence the total internal reflection happens for larger angle prisms. So let us go for low angle of incidence.

Table 2:

A in degree	$\mu = 1.4$ r_2	$\mu = 1.58$ r_2
10	-28.21	-23.24
20	-18.21	-13.24
30	-8.21	-3.24
40	1.79	6.76
50	11.79	16.76
60	21.79	26.76
70	31.79	36.76
80	41.79	41.76

Table 3:

A in degree	$\mu = 1.4$ r_2 in degree	$\mu = 1.58$ r_2 in degree
10	-14.21	-11.1
20	-4.21	-1.1
30	5.71	8.9
40	15.71	18.9
50	25.71	28.9
60	35.71	38.9
70	45.71	48.9
80	55.71	58.9

4.3 For $i = 35^\circ$

We can see that the total internal reflection happens for both violet and red for 70° prism. Usually we take angle of incidence between 30° and 60° . So we find what happens at $= 30^\circ$, which is given in Table 4.

4.4 For $i = 30^\circ$

When we take $i = 30^\circ$ we see that for $A = 60^\circ$ there is no dispersion.

Table 4:

A in degree	$\mu = 1.4$ r_2 in degree	$\mu = 1.58$ r_2 in degree
10	-11.1	-8.66
20	-1.1	1.34
30	8.9	11.34
40	18.9	21.34
50	28.9	31.34
60	38.9	41.34
70	48.9	51.34
80	58.9	61.34

So we found $d_v - d_r$ for different i and A . We found that $d_v - d_r$ is maximum for a 60° prism, which ensures maximum dispersion. This is shown in Table 5

Table 5:

A in degrees	$d_v - d_r$ in degrees
10	2.1
20	3.95
30	5.76
40	7.93
50	11.17
60	19.13
70	Absent
80	Absent

5 Observations from the above studies

Above studies show that for getting dispersion without total internal reflection upto 60° prism, we have to choose the angle of incidence between $i = 30^\circ$ and $i = 60^\circ$. Next we will find for which angled prism we get maximum dispersion with the above choice of angle of incidences.

6 Separation between violet and red rays

From Eqn(2) we can find deviation for violet, d_v and deviation for red d_r and to get a maximum separation, $d_v - d_r$ must be large.

7 Conclusions

In typical laboratory experiments, the angle of incidence is chosen between 30° and 60° . Our analysis demonstrates that, within this range, a 60° prism provides the best dispersion, ensuring well-separated red and violet rays. This explains why equilateral prisms are widely used in educational settings for studying refraction and dispersion. Their optimal performance in practical conditions makes them the preferred choice for instructional experiments.

References

- [1] Newton, Opticks, Dover Publications Inc.(1952)
- [2] Dispersion and Resolving Power of Prism Spectrometers, George W. Hazard, American Journal of Physics, 19, 235-236 (1951)
- [3] Generalized prism dispersion theory, F. J. Duarte and J. A. Piper, American Journal of Physics, 51, 1132-1134 (1983)
- [4] Refraction through a prism, Albert Feldman, American Journal of Physics 51, 929-931 (1983)
- [5] Refraction through a Prism, John H. Kirby, Nature volume 44, 294 (1891)
- [6] Refractive Index Measurement and its Applications, Shyam Singh, Phys. Scr. 65, 167(2002)
- [7] Fundamentals of Optics, Francis Jenkins and Harvey Elliott White, 4th Edition, McGraw Hill Education(2017)
- [8] D. Halliday, R. Resnick, and Jearl Walker, Fundamentals of physics (John Wiley and Sons, 2014)

Solitons: Waves with many Attributes of Particles*

Vishwamittar¹

¹ Retired from Department of Physics, Panjab University,
Chandigarh 160014, India.

Contact Address: #121, Sector-16, Panchkula-134113 (Haryana).
vm121@hotmail.com

**The article is dedicated with warm regards to Prof. C. G. Mahajan, Prof. K.S. Harchand,
Prof. K.C. Mittal, Prof. Vijendra K. Agarwal, and (late) Prof. R.S. Sud, and their families.*

Submitted on 18-07-2022

Abstract

The purpose of this article is to expose the students and other readers to the wonderful world of solitons – the self-sustaining solitary waves. Beginning with a brief review of the history of their discovery as waves on water surface and their modeling by the Korteweg-de Vries nonlinear partial differential equation, and characterization as particles, we have given a concrete definition for these. A simple soliton solution for the afore-mentioned equation has been elucidated. Also included is a summary of Sine-Gordon and Nonlinear Schrödinger equations. The question ‘why care for solitons?’ has been answered by giving an overview of multifaceted theoretical and practical applications of its concepts in various branches of science, particularly physics. Effort has been made to keep the presentation as elementary as

possible omitting some mathematical subtleties of the subject.

1 Introduction

When we read or hear the word ‘wave’, the immediate thing that comes to our mind is ‘the wave moving on the surface of water’. A stone thrown into still water of a pond creates a disturbance that travels radially outwards in all directions from the point of hitting while the water particles on the surface vibrate up-and-down. Thus, as the wave propagates away from the point of its origin, the water particles remain where they were (a cork or a paper boat placed on the surface shows only up-and-down oscillations but no forward motion) and only energy is transported outwards.

A wave is a continuous disturbance from the state of equilibrium that travels from one region of space to another and transports energy / information without any translational movement of the intervening medium. The properties which characterize a wave and distinguish one from another are velocity, amplitude, angular frequency, and wavelength. Some well-known examples of waves are transverse waves on vibrating strings (which form the basis of musical instruments like veena, sarangi, violin, etc.), longitudinal or pressure waves in a gas, voltage and current waves along an electrical transmission line, electromagnetic waves (with light and radio waves as typical examples) in the free space / vacuum or a material medium, etc. Interestingly, irrespective of their diverse individual properties, these waves are treated by common mathematical formalism [1].

The partial differential equation (PDE) describing a progressive wave, travelling along x – direction with disturbance function $u(x, t)$ at point x at instant of time t , reads

$$\frac{\partial^2 u(x, t)}{\partial t^2} = v^2 \frac{\partial^2 u(x, t)}{\partial x^2}. \quad (1)$$

This celebrated wave equation was first introduced and solved in a general way by d'Alembert in 1747, while developing a mathematical model of a vibrating string. Here, v is the velocity of propagation of the wave and is also called phase velocity of the wave. $u(x, t)$ represents transverse displacement in a string or water wave, pressure in a sound wave in air, voltage or current in

an electrical transmission line (where Eq. (1) is called telegraphist's equation) and so on. Note that Eq. (1) is a linear partial differential equation so that superposition principle holds good. Accordingly, if $u_1(x, t)$ and $u_2(x, t)$ are solutions of this second-order differential equation, then any linear combination of these functions is also a solution. From physics point of view this means that if two (or more) different waves are present in a medium, the disturbance at any point at any given time is the sum of the disturbances separately produced by these individual waves.

It is common practice to denote partial derivatives of a function $u(x, t)$ with respect to time and position coordinate by using t and x as subscripts of ∂ and u , and to suppress the explicit dependence on these variables. Thus, $\frac{\partial u(x, t)}{\partial t} \equiv \partial_t u \equiv u_t$, $\frac{\partial u(x, t)}{\partial x} \equiv \partial_x u \equiv u_x$, and so on. In these short-hand notations, the classical wave equation, Eq. (1), can be written as

$$\partial_t^2 u - v^2 \partial_x^2 u = u_{tt} - v^2 u_{xx} = 0. \quad (2)$$

We shall use the second abbreviation in this article. It may be added that waves are quite commonly observed in higher spatial dimensions and the preceding equation is modified to read

$$u_{tt} - v^2 \nabla^2 u = 0, \quad (3)$$

where ∇^2 is the Laplace operator in the chosen coordinate-system.

The wave travelling in $+x$ – direction, sometimes called the forward wave, is represented by the implicit function $u(x, t) =$

$f_1(x - vt)$ while the one moving in $-x$ direction (the backward wave) is given by $u(x, t) = f_2(x + vt)$. The arguments $x \mp vt$ are usually referred to as the characteristic variables. The profile of the wave is governed by the mathematical form of f_1 or f_2 . Under ideal conditions, the waves described by f_1 and f_2 do not change their form as these propagate. Thus, the shape of the wave given by $f_1(x - vt)$ at time $t > 0$ will be the same as at $t = 0$ except that it is shifted to the right by an amount vt .

It may be added that if the function describing a progressive wave is such that the profile of the disturbance at time $t = 0$ is either a sine or a cosine function then it is known as a harmonic or a sinusoidal wave. Accordingly,

$$\begin{aligned} u(x, t) = & a \cos(k[x - vt]), \\ & a \sin(k[x - vt]), \\ \text{or} \\ & a \exp(ik[x - vt]) \end{aligned} \quad (4)$$

represent a harmonic wave of amplitude a , propagation constant or angular wave number k (which gives periodicity in the space coordinate x), wavelength $\lambda = 2\pi/k$, and angular frequency $\omega = 2\pi\nu = \frac{2\pi v}{\lambda} = vk$. Therefore, Eq. (4) can also be written as

$$u(x, t) = a \cos(kx - \omega t) \quad (5)$$

and so on. The argument $(kx - \omega t)$ is phase of the wave at point x at time t . Obviously, for a specific point x it changes linearly with time t . Of course, the phase of the wave can

be generalized to read $(kx - \omega t + \theta)$, with θ as phase at $x = 0, t = 0$.

If the medium through which a wave is passing, is such that the phase velocity of the wave is the same for all frequencies (i.e., $v = \frac{\omega}{k} = \text{constant}$, independent of both ω and k), then it is called a non-dispersive or dispersion less medium. On the other hand, a medium is said to be dispersive if the wave velocity is different for different frequencies ($\frac{\omega}{k} \neq \text{constant}$). Note that free space is non-dispersive medium for light waves while glass is a dispersive medium for these.

Now, if we consider an arbitrary pulse (a one-time disturbance or a wave of very short duration), which is a linear superposition of many harmonic waves with different angular frequencies, it will travel without deformation in its profile in a non-dispersive medium as all the constituent waves move with the same speed. However, in a dispersive medium, the phase velocities of the component harmonic waves are different so that the fast-moving constituents go ahead, and the slow ones lag behind. Consequently, the pulse changes its shape as time evolves and will spread out or disperse as it moves (leading to decrease in its amplitude and increase in width). In this case, we talk about group velocity $v_g = d\omega/dk$. It is this velocity with which energy is transported by the pulse or any wave comprising different frequencies. The expression, such as $\omega = k - k^3$ or $\omega = \omega_0 |\sin(\frac{k}{2})|$, giving the variation of ω as a function of k is known as a dispersion relation.

Before proceeding further, it is worthwhile to point out that if we redefine time such that $t' = vt$, then $u_t = \frac{\partial u(x, t')}{\partial t'} \frac{\partial t'}{\partial t} = v u_{t'}$ and, similarly, $u_{tt} = v^2 u_{t't'}$. Accordingly, Eq. (2) is transformed to read $u_{t't'} - u_{xx} = 0$. Replacing t' by t and keeping in mind that the rescaled time has dimension of length rather than time, we can rewrite the wave equation, as $u_{tt} - u_{xx} = 0$. A comparison of this equation with the original equation, viz. Eq. (2), shows that the envisaged transformation is equivalent to taking $v = 1$. Of course, dimension of u in both the equations is the same.

It is pertinent to note that like the linear differential equation describing simple harmonic oscillator, the wave equation, Eq. (1) or (2), is obtained by assuming the amplitude of wave to be small. As such, it is an idealized model for the one-dimensional wave motion. However, if the derivation is made for more realistic situations, which necessarily involve nonlinearity, we get wave equations involving dispersive as well as nonlinear terms. One such nonlinear partial differential equation (NLPDE) was derived by Korteweg and de Vries (usually abbreviated as KdV) in 1895 to describe the propagation of waves in one-dimension on the surface of a shallow canal assuming the flow to be inviscid, incompressible, steady and irrotational. In its standard dimensionless form, as commonly used in the current literature, it reads [2, 3, 5, 8]

$$u_t - 6uu_x + u_{xxx} = 0. \quad (6)$$

Here, t and x are normalized time and nor-

malized coordinate in the direction of wave propagation, respectively. If in any problem similar equation of evolution turns out to be of different form, it can be transformed into this standard form of the KdV equation by using an appropriate scale. Here, the first term gives time evolution of the disturbance proceeding in $+x$ - direction. The second term in this equation is nonlinear, which leads to steepening or narrowing of the wave. Also, because of the presence of nonlinear term, the principle of superposition of solutions does not hold good. This, in turn, makes wave structure robust in interactions / collisions with other wave structures. The third term is the dispersion part and will give rise to a nonlinear relationship between ω and k . In fact, the KdV equation is the simplest NLPDE that incorporates both nonlinearity and dispersion.

Eq. (6) admits a solution of the form $u(x, t) = A \operatorname{sech}^2(x, t, A)$ indicating presence of amplitude in the argument. This represents a bell-shaped profile, which describes a solitary wave first observed by Russell in 1834. This feature of the KdV equation and its modified form has been found to be very useful in the study of waves in elastic rods, liquid-gas bubble mixtures, plasmas, anharmonic lattices, etc., besides the water waves.

In this article, we delineate upon the fascinating and interesting topic of solitary waves from pedagogic point of view at reasonably basic level [2-10]. In Section 2, we give an overview of their discovery, devel-

opment of the subject and nomenclature as solitons. This is followed by Section 3 where a precise definition of solitons is given. Section 4 is devoted to description of a simple soliton-solution of the KdV equation which is being used as a prototypical example of exactly solvable soliton-bearing model. This approach involves easily understandable mathematics and, still, makes the concept quite transparent. Also included are some remarks regarding various solutions of this equation. Then we move on to Section 5 to give a brief information about two other NLPDEs leading to different flavours of solitons. Section 6 summarizes versatile applications of the solitons to a wide variety of systems in diverse fields. We close the article in Section 7 by making some general comments.

2 A Historical Account of the Discovery of Solitary Waves and Growth of the Subject

We shall not be able to do justice to all the spectacular developments in vast subject of solitons with an elegant history of nearly two centuries and shall concentrate mainly on those contributions that had a larger influence on the overall progress, particularly from physics point of view.

It was in 1830's that a Scottish civil engineer and naval architect named John Scott Russell, with a view to develop an efficient design for canal boats, performed experiments on moving boats in Edinburgh-

Glasgow canal to find relation between their shape, speed, and the force needed to push them. One day in August 1834, this young man (then 26 years old) was observing the motion of a boat that was being rapidly drawn along a narrow channel by a pair of horses. He found that when the boat suddenly stopped, the moving water collected around it in a state of violent agitation and then abruptly it rolled forward with great velocity of about 13 - 14 km / hr in the form of a nearly 9 m long and 30 - 50 cm high smooth and well-marked accumulation of water. This heap travelled on the surface of water without any change in its profile or speed till it was lost in the windings of the channel after a follow up of about 2 km. He called this singular wave 'the wave of translation'. Obviously, this discovery was essentially a random happenstance.

Impressed by this unexpected observation, Russell carried out extensive meticulous experiments about the nature of these waves of elevation in many canals, rivers, lakes, and in a large wave tank in his back garden. He concluded that this wave motion was unique and quite different from other types of oscillatory motions – the speed depends on its amplitude and the depth of water, and they never merge. Therefore, he started referring to them as 'solitary waves' in the sense that this wave had only a single protuberance traveling without any change in its shape, size, or speed. Treating it as a gravity wave, he found that speed of the wave on a water sur-

face with undisturbed depth h is given by $v = \sqrt{g(h+a)}$, where a is amplitude of the wave. This showed that the larger the amplitude of a wave higher its speed – a miraculous nonlinear effect.

However, he could not convince his contemporaries, particularly mathematicians, about the importance and even novelty of these waves mainly because his findings were at variance with the then accepted theories of hydrodynamics and he himself could not give an analytical formalism. What was strikingly surprising and unusual about this wave and was not appreciated by the scientists at that time is: A coherent hump of water was formed out of turbulence produced by sudden stopping of boat in shallow water and this protrusion maintained its characteristics over quite a long distance in contrast with the normal behavior of water waves that spread out and disappear after travelling over reasonably short distances.

Despite this situation, Russell's work was followed by Stokes' efforts in 1847 to get some theoretical interpretation and by Boussinesq's (1871) and Rayleigh's (1876) successful explanation of the nature of these waves. They used Euler's equations of motion for an inviscid, incompressible fluid and not only obtained Russell's formula for speed but also an expression for the wave profile, reading

$$u(x, t) = a \operatorname{sech}^2 [A(x - vt)]. \quad (7)$$

Here, A is a parameter that depends on the amplitude a and the height h of water sur-

face from the base of the canal. This expression is strictly true for $a \ll h$. However, these scientists did not derive or write the differential equation satisfied by the above expression for $u(x, t)$. This task was done by Dutch mathematician Korteweg and his student de Vries in 1895 who obtained the remarkable Eq. (6) and derived various wave properties which were similar to those observed by Russell in different experiments, though they did not refer to the work done by him. Not only this, it also seems that even they themselves did not realize the importance of their finding as they did not pursue it further. Continuing the narration of the history, it may be mentioned that in 1955, Fermi, Pasta, Ulam, and Tsingou, working at one of the world's earliest computers (the MANIAC machine), performed numerical investigation of heat transfer in a solid modeled by a one-dimensional lattice consisting of equal mass anharmonic oscillators. They observed that there was a periodic recurrence in the distribution of energy rather than expected equipartition of energy among the modes. This astounding result and the fact that the system considered by these scientists was closely related to discretization of the KdV equation, prompted Zabusky and Kruskal (1965) to undertake the initial value problem for the KdV equation [11]. Pursuing insightful numerical simulations, they (i) mimicked the Russell's solitary waves; (ii) explained the odd results of Fermi, Pasta, Ulam, and Tsingou; and (iii) found that when two or more

of the KdV solitary waves interact or collide with each other they neither break up nor disperse and rather emerge out preserving their individual shapes and velocities as if there was no interaction. These do undergo a small change in their phase on collision.

Keeping in view their last- mentioned novel finding that assigned remarkable corpuscular or particle-like characteristic to these waves, they coined the term ‘soliton’ for these solitary waves to emphasize its kinship with electron, photon, phonon, etc. that behave like both particle and wave. In fact, it was this milestone work that brought the KdV equation into limelight after being in obscurity for nearly seven decades. However, the first rigorous analytical solution to this famous equation reading $u(x, t) = -B \operatorname{sech}^2 \left[\sqrt{\frac{B}{2}} (x - 2Bt - x_0) \right]$ was given by Gardner and coworkers in 1967 [5,12]. The method developed by them involves formulating a scattering problem with desired solution as potential and solving this as a first step. The outcome of this solution is then used to reconstruct $u(x, t)$. This technique is referred to as Inverse Scattering Method. They also obtained the general multi-solitons or n -solitons solution for the KdV equation. The salient feature of this method lies in the fact that it provides exact solution for nonlinear wave equations by linear techniques and is useful in discovering solitons. Later, this ingenious approach together with its generalizations and the novel method put forward by Hirota (1971) for obtaining multi-soliton solutions,

provided powerful tools for solving many physically interesting NLPDEs and, thus, for studying solitons. However, we shall not dwell on details of these techniques or other methods developed for solving the soliton-bearing equations as these are too technical in nature. In the meantime, Toda (1967) reported existence of a soliton in a discrete, integrable system, which is now called Toda lattice.

These developments opened up fascinating vistas, and established study of solitons or solitary waves as a vibrant and flourishing topic of research among mathematicians, physicists, engineers, and others. On one hand, this boom led to discovery of numerous soliton-bearing nonlinear evolutionary PDEs in one or more space-dimensions and thus adding to mathematical richness of theory of solitons. On the other hand, solitons became objects of immense physical importance. In fact, solitons play same role in the description of nonlinear systems as harmonic waves in the linear systems. Consequently, lot of effort has been directed at exploiting fecundity of applications of this concept in different branches of science. However, before going ahead, we first define solitons in Section 3.

3 Defining a Soliton

Strictly speaking solitons are such solutions of the NLPDEs that (i) do not change profile while travelling nor do they disperse, implying complete stability; (ii) survive colli-

sions, emerging unblemished; (iii) cannot be constructed as a superposition of harmonic waves; and (iv) the speed of the wave profile depends on its amplitude. Thus, solitons are self-reinforcing, non-dissipative, and persistent solitary waves of finite amplitude, and are indubitably nonlinear entities. These propagate undistorted over long distances and maintain their speed and shape upon collision / interaction with other such waves.

However, the term 'soliton' has been used by scientists in a relatively loose manner for the objects which do not necessarily fulfil all the above requirements. It is, in a way, used to signify a spatially compact, finite field energy configuration which may or may not be time dependent. This term has also been adopted to cover a large class of solitary excitations that are localized in space-time and though long-lived, are only metastable. Thus, the condition of these being perfectly stable is relaxed. In this sense, some of the soliton solutions can be identified as elementary excitations. Such moderation of the definition has made the realm of usage of the theory of solitons quite vast.

4 Rudimentary Solution of the KdV Equation

Guided by the approach presented by Drazin and Johnson [2], to obtain a travelling or progressive wave solution to the KdV equation, Eq. (6), we introduce a new variable or parameter $\eta = x - vt$, which repre-

sents the position in a reference frame moving with the wave with speed v . Note that $\frac{\partial \eta}{\partial x} = 1$ and $\frac{\partial \eta}{\partial t} = -v$. Also, the solution can be written as $f(\eta) \equiv f$ in place of $u(x, t)$ and it represents a wave travelling with speed v in the original coordinate system. Now,

$$u_t = \frac{\partial u}{\partial t} = \frac{df}{d\eta} \frac{\partial \eta}{\partial t} = -vf',$$

$$u_x = \frac{\partial u}{\partial x} = \frac{df}{d\eta} \frac{\partial \eta}{\partial x} = f',$$

and

$$u_{xxx} = \frac{\partial^3 u}{\partial x^3} = f''''.$$

Making these substitutions into Eq. (6), we get

$$-vf' - 6ff' + f'''' = 0. \quad (8)$$

Obviously, the NLPDE has been transformed into an ordinary differential equation with the nonlinear and dispersive terms intact. Integrating the above differential equation with respect to single variable η , we have

$$-vf - 3f^2 + f'' = C_1, \quad (9)$$

where C_1 is arbitrary constant of integration. Multiplying with f' on both sides of this equation and integrating again, we obtain

$$-v\frac{f^2}{2} - 3\frac{f^3}{3} + \frac{(f')^2}{2} = C_1f + C_2. \quad (10)$$

Here, C_2 is second arbitrary constant. Eq. (10) can be rewritten as

$$(f')^2 = 2\{f^3 + \frac{v}{2}f^2 + C_1f + C_2\} \equiv 2F(f). \quad (11)$$

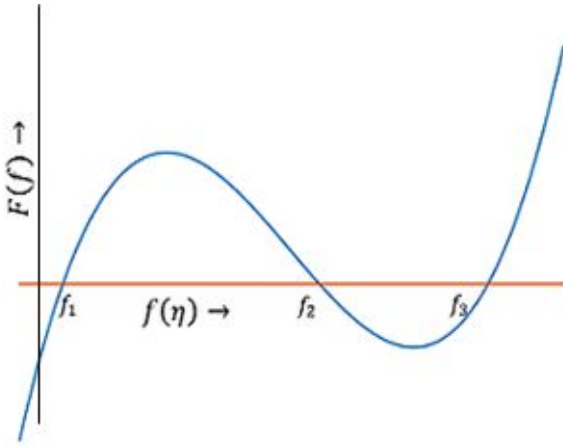


Figure 1: An arbitrary plot showing typical dependence of $F(f)$ on f as per Eq. (11).

This gives us f' in terms of a cubic polynomial in f , where C_1 and C_2 are determined by the initial conditions satisfied by the KdV equation.

Now, $f(\eta)$ being solution of wave equation, it represents a (classical) wave displacement, and, therefore, it must be real (so that it can be observed) and finite or bounded. This, in turn, implies that f' too is real so that $(f')^2 \geq 0$ and hence $F(f) \geq 0$. Thus, only those $f(\eta)$ are physically acceptable for which $F(f)$ is non-negative. Since $F(f)$ is a cubic polynomial, it will have three real-valued zeros defined by $f^3 + \frac{v}{2}f^2 + C_1f + C_2 = 0$. Let these be f_1, f_2 , and f_3 and, in general, such that $f_1 < f_2 < f_3$. Of course, sometimes two or all the three zeros may coincide with each other. Note that for a cubic polynomial with the coefficient of the cubic term as unity, the sum of its zeros equals negative of the coefficient of the

square term. Thus,

$$f_1 + f_2 + f_3 = -\frac{v}{2}. \quad (12)$$

Since v is speed of propagation of the wave, it will be positive along the $+x$ - direction and this demands that

$$f_1 + f_2 + f_3 < 0. \quad (13)$$

Obviously, f_1 will certainly be negative and the signs of f_2 and f_3 may be negative or positive depending on the values of v, C_1 and C_2 .

For extremely large $|f|$, $F(f)$ is governed by f^3 , and, therefore, $F(f)$ is negative for negative large values of f and is positive for positive large magnitudes of f . Since f_1 and f_3 are, respectively, the lowest and the largest zeros of $F(f)$, it will be negative for $f < f_1$ and it will be positive for $f > f_3$. So, at the zero f_1 , $F(f)$ goes from negative values to positive values, and at the zero f_3 , it must again go from negative to positive values. Accordingly, at the zero f_2 , sign of $F(f)$ values changes from positive to negative; Fig. 1. Thus, $F(f)$ and hence $(f')^2$ is positive for $f_1 < f < f_2$ and for $f > f_3$. But it is bounded only for $f_1 < f < f_2$. Therefore, acceptable solution $f(\eta)$ must lie between f_1 and f_2 , which must be distinct.

Since f_1, f_2 , and f_3 are zeros of $F(f)$, we can express it as product of three factors:

$$F(f) = (f - f_1)(f - f_2)(f - f_3). \quad (14)$$

This together with Eq. (11) gives us

$$\frac{df}{d\eta} = \pm [2(f - f_1)(f - f_2)(f - f_3)]^{1/2}. \quad (15)$$

Confining ourselves to the region $f_1 < f < f_2$ (and the corresponding η values $\eta_1 < \eta < \eta_2$), we get from Eq. (15)

$$\int_{\eta_1}^{\eta} d\eta = \pm \int_{f_1}^f \frac{dg}{[2(g-f_1)(g-f_2)(g-f_3)]^{1/2}}. \quad (16)$$

We have used g as variable in the integral on the right-hand side as f is being taken as upper limit. Now, we substitute $g = f_1 + (f_2 - f_1) \sin^2 \theta$ so that lower limit $g = f_1$ corresponds to $\theta = 0$, and the upper limit $g = f$ implies

$$f = f_1 + (f_2 - f_1) \sin^2 \Theta, \quad (17)$$

where Θ is upper limit value of θ . Making these substitutions together with $dg = 2(f_2 - f_1) \sin \theta \cos \theta d\theta$ on the right-hand side of Eq. (16), simplifying the resulting expression, and using the fact that left-hand side equals $\eta - \eta_1$, we finally obtain

$$\begin{aligned} \eta - \eta_1 &= \pm \sqrt{\frac{2}{f_3 - f_1}} \int_0^{\Theta} \frac{d\theta}{\sqrt{1 - m \sin^2 \theta}} \\ &= \pm \sqrt{\frac{2}{f_3 - f_1}} w \quad (\text{say}). \end{aligned} \quad (18)$$

Here,

$$m = \frac{f_2 - f_1}{f_3 - f_1}, \quad (19)$$

such that $0 \leq m \leq 1$. Also,

$$w = \int_0^{\Theta} \frac{d\theta}{\sqrt{1 - m \sin^2 \theta}}, \quad (20)$$

is incomplete elliptic integral of first kind with parameter m .

Now, we define a new pair of functions corresponding to w :

$$\text{sn } w \equiv \text{sn}(w|m) = \sin \Theta, \quad (21a)$$

$$\text{cn } w \equiv \text{cn}(w|m) = \cos \Theta, \quad (21b)$$

These are, respectively, called the Jacobi elliptic sine (snoidal) and Jacobi elliptic cosine (cnoidal) functions. Θ is usually referred to as Jacobi amplitude.

Note that for $m = 0$, $w = \int_0^{\Theta} d\theta = \Theta$, so that

$$\text{sn } w \equiv \text{sn}(w|0) = \sin \Theta = \sin w, \quad (22a)$$

$$\text{cn } w \equiv \text{cn}(w|0) = \cos \Theta = \cos w. \quad (22b)$$

Obviously, for $m = 0$, which happens when f_2 merges with f_1 from above, i.e., $f_2 \rightarrow f_1^+$, the functions $\text{sn } w$ and $\text{cn } w$ are periodic sin and cos functions, respectively.

On the other hand, for $m = 1$,

$$\begin{aligned} w &= \int_0^{\Theta} \frac{d\theta}{\sqrt{1 - \sin^2 \theta}} = \int_0^{\Theta} \sec \theta d\theta \\ &= \ln[\tan \Theta + \sec \Theta] \\ &= \ln \left[\frac{1 + \tan\left(\frac{\Theta}{2}\right)}{1 - \tan\left(\frac{\Theta}{2}\right)} \right] \end{aligned} \quad (23)$$

This, on simplification, yields

$$\tan\left(\frac{\Theta}{2}\right) = \tanh\left(\frac{w}{2}\right) \quad (24)$$

which, in turn, gives $\sin \Theta = \tanh w$ and $\cos \Theta = \text{sech } w$. These imply that

$$\text{sn } w \equiv \text{sn}(w|1) = \sin \Theta = \tanh w, \quad (25a)$$

$$\text{cn } w \equiv \text{cn}(w|1) = \cos \Theta = \text{sech } w. \quad (25b)$$

Thus, for $m = 1$, the elliptic functions $\text{sn } w$ and $\text{cn } w$ are aperiodic $\tanh w$ and $\text{sech } w$, respectively. Note that $m = 1$ if the zeros f_2 and f_3 of $F(f)$ coalesce to form a double zero ($f_2 \rightarrow f_3^-$ as $f_2 < f_3$) but are distinct from f_1 .

After this digression, we come back to Eq. (17), replace $\sin^2 \Theta$ by $1 - \cos^2 \Theta$, and get

$$f = f_2 - (f_2 - f_1) \cos^2 \Theta. \quad (26)$$

In view of Eq. (21b), this can be rewritten as

$$f = f_2 - (f_2 - f_1) \operatorname{cn}^2(w|m), \quad (27)$$

where

$$w = \pm(\eta - \eta_1) / \sqrt{\frac{2}{f_3 - f_1}}, \quad (28)$$

from Eq. (18). Since cn is an even function, we omit \pm sign and express Eq. (27) as

$$f(\eta) = f_2 - (f_2 - f_1) \operatorname{cn}^2 \left(\sqrt{\frac{f_3 - f_1}{2}} \{\eta - \eta_1\} \middle| m \right). \quad (29)$$

This is called cnoidal wave solution of the KdV equation – generalization of the sinusoidal wave.

In the limit $m \rightarrow 0$, which is achieved when $f_2 \rightarrow f_1^+$, we use Eq. (22b) for $\operatorname{cn} w$ and then the double angle trigonometric identity $\cos^2 w = \frac{1}{2}(1 + \cos 2w)$, and get $f(\eta)$ in terms of \cos function with $\frac{(f_2 - f_1)}{2}$ as coefficient. Thus, $f(\eta)$ describes an oscillatory cosine wave with amplitude $\frac{(f_2 - f_1)}{2}$, which, obviously, is quite small. It is found that the wave is dispersive in nature. This is low amplitude linear wave limit of the cnoidal solution. However, we shall not go into its further discussion.

Next, for the case $m = 1$, which is the most nonlinear limit, we use Eq. (25b) for

$\operatorname{cn} w$ and put $f_2 = f_3$ into Eq. (29). Accordingly, we have

$$f(\eta) = f_3 - (f_3 - f_1) \operatorname{sech}^2 \left(\sqrt{\frac{f_3 - f_1}{2}} \{\eta - \eta_1\} \right). \quad (30)$$

Now, from the definition $\operatorname{sech} y = \frac{2}{e^y + e^{-y}}$, we note that $\operatorname{sech} y = 1$ for $y = 0$ and equals zero for $y \rightarrow \pm\infty$. Thus, $\operatorname{sech}^2(\sqrt{\frac{f_3 - f_1}{2}} \{\eta - \eta_1\}) = 1$ for $\eta = \eta_1$ and 0 for $\eta \rightarrow \pm\infty$. The corresponding values of $f(\eta)$ are f_1 and f_3 , respectively. Since f_1 is necessarily negative and less than f_3 , $f(\eta)$ has minimum value f_1 at $\eta = \eta_1$, and maximum value f_3 for $\eta \rightarrow \pm\infty$. In other words, if we plot a graph of $f(\eta)$ as function of η , this will be a wave profile with depression (upside-down) having value f_1 at $\eta = \eta_1$, and depth $f_3 - f_1$. However, as we are looking for a model to describe a waveform above the water surface, we consider $-f(\eta)$ rather than $f(\eta)$. Accordingly, $-f(\eta)$ represents a profile with $-f_1$ at $\eta = \eta_1$ as peak and $-f_3$ as minimum value for $\eta \rightarrow \pm\infty$. Consequently, we can identify $-f_1 - (-f_3) = f_3 - f_1$ as amplitude a of the wave. Thus, Eq. (30) can be written as

$$-f(\eta) = -f_3 + a \operatorname{sech}^2 \left\{ \sqrt{\frac{a}{2}} (\eta - \eta_1) \right\}. \quad (31)$$

The velocity of this wave is given by

$$v = -2(f_1 + f_2 + f_3) = 2a - 6f_3, \quad (32)$$

where we have used $f_2 = f_3$ and $f_3 - f_1 = a$ in Eq. (12). Furthermore,

$$\eta = x - vt = x + 6f_3 t - 2at. \quad (33)$$

Note that v is directly proportional to amplitude implying that the larger the amplitude the higher the speed. Also, for v to be positive, the zero f_3 must be less than $\frac{a}{3}$.

Having obtained the solution, Eq. (31), for Eq. (6), and using Eq. (33), we can write

$$\begin{aligned} -u(x, t) &= -f_3 + a \operatorname{sech}^2\left\{\sqrt{\frac{a}{2}}(x - vt - \eta_1)\right\} \\ &= -f_3 + a \operatorname{sech}^2\left\{\sqrt{\frac{a}{2}}(x + 6f_3t - 2at - \eta_1)\right\}. \end{aligned} \quad (34)$$

This describes a wave of elevation having nonperiodic bell-shaped profile of amplitude a ($> 3f_3$), travelling with speed $v = 2a - 6f_3$, initial phase factor $-\sqrt{\frac{a}{2}}\eta_1$, and $-f_3$ as ambient or undisturbed or equilibrium level. The presence of a in the argument of sech in Eq. (34) shows that the shape of the wave depends on amplitude in a complicated manner, which, in turn, implies that $-u(x, t)$ represents a nonlinear wave. From the first equality in Eq. (34) it is found that $-u(x, t) + f_3 = a$ if $x = vt + \eta_1$. Obviously, the peak appears at $x = \eta_1$ for $t = 0$ implying that η_1 can be taken to be 0 by using the location of the peak at $t = 0$ as reference for measuring x . Furthermore, $-u(x, t) + f_3 = a/2$ when $x_{\pm} = vt + \eta_1 + \sqrt{\frac{2}{a}} \ln(\sqrt{2} \pm 1)$. Taking the distance between the points at which the height of the wave above the ambient level is half the amplitude, as width of the profile, called full width at half maximum, we have $\Delta x \equiv x_+ - x_- = \sqrt{\frac{2}{a}} \ln \frac{\sqrt{2}+1}{\sqrt{2}-1}$. Thus, width of the profile is inversely proportional to \sqrt{a} . Combining this result with the statement af-

ter Eq. (33), we note that the wave of elevation described by Eq. (34) is such that taller the wave, narrower and faster it is. It is the solitary wave discovered by Russell and its plot is depicted in Fig. 2 for three values of t . Note that profile of the wave is the same for all the three t values shown here and has $\Delta x = 3.94$.

As a follow up of the preceding discussion, suppose we launch two solitary waves having different amplitudes such that the one with smaller amplitude is leading. The wave with higher amplitude will have larger velocity so that as time passes it will come closer to the other wave, bump into it at some instant of time and ultimately overtake it. The end-result will be that the two waves pass through each other without losing their identity, i.e., they come out of the collision unscathed – a particle-like robustness. In fact, this aspect was also observed by Russell.

It may be mentioned that the actual solution, Eq. (30), representing a wave of depression rather than a wave of elevation, is a consequence of the negative sign of the nonlinear term in the standard form of the KdV equation, which has been solved here. Furthermore, the cnoidal wave solution, Eq. (29), is not the only possible solution to the KdV equation; other simple looking solutions have also been found. Besides, solutions leading to more than one soliton, have also been obtained.

It is worth emphasizing that the dispersion term u_{xxx} in Eq. (6) gives rise to ten-

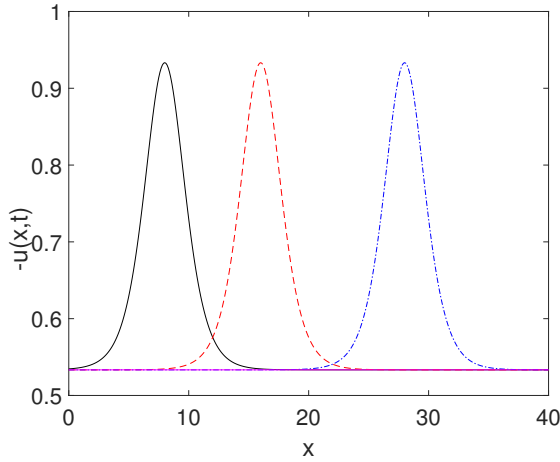


Figure 2: One-soliton solution of the KdV equation given by Eq. (34) at $t = 2.0$ (black), 4.0 (red), and 7.0 (blue) for $a = 0.4$; $v = 4.0$, $\eta_1 = 0$.

dency of flattening or spreading of the wave profile, while the nonlinear term $-6uu_x$ makes it steep and cohesive. The precise balancing of these two tendencies leads to ‘no change in the shape’ of the wave, i.e., the soliton solution. The KdV equation has been found to be very useful in modeling the dynamics of physical systems characterized by mild dispersion and weak nonlinearity.

By convention, the word ‘soliton’ is used for the wave profile with positive displacement (i.e., an elevation) and the envelope with negative displacement (i.e., a depression) is called an anti-soliton. Thus, $-u(x, t)$ given by Eq. (34) defines a soliton, while $u(x, t) = f_3 - a \operatorname{sech}^2(\sqrt{\frac{a}{2}} \{x + 6f_3t - 2at - \eta_1\})$ is an anti-soliton. When a soliton and corresponding anti-soliton collide with each other, the net displacement is zero and this is referred to

as annihilation of soliton – anti-soliton pair. However, generally these pairs collide and then separate.

5 Some Other Soliton-Bearing Nonlinear Partial Differential Equations

It has been pointed out in the preceding section that the solution of the KdV equation maintains its shape indefinitely because of exact cancellation of the spreading or broadening produced by the dispersive term and the narrowing effects of the nonlinear term. In fact, any NLPDE containing dispersive and nonlinear terms counterbalancing detrimental effects of each other will have soliton solution. Of course, these solitons can be distinctly different from the bell-shaped solitons of the KdV equation. Two such evolution equations having more than one soliton solution and finding wide range applications in physics, biology, and engineering, together with relevant brief comments, are listed below. While writing these NLPDEs, the variables involved are taken to be properly rescaled. In fact, these equations are more useful than the KdV equation, which has been discussed in detail not because it is the oldest but because it is the simplest in nature.

5.1 Sine-Gordon Equation

This NLPDE reads

$$u_{xx} - u_{tt} - \sin u = 0, \quad (35)$$

with $\sin u$ as nonlinear term. The presence of u as argument of \sin implies that this equation describes angular disturbance expressed in radians. It was originally put forward by Bour in 1862 during the investigation of surfaces of constant negative curvature in 3-dimensional space. Later, it was rediscovered by Frenkel and Kontorova in 1939 in their seminal work on study of crystal dislocations, which are defects or irregularities in the crystal structure along some direction and can even be mobile. It was in 1962 that Perring and Skyrme obtained a 2-soliton solution for Eq. (35). Subsequently, 1- and 3- soliton solutions were also obtained. This equation drew lot of attention in 1970s onwards as it was found to be useful in explaining many physical phenomena and is the simplest NLPDE in a periodic medium. It is interesting to note that the name 'sine-Gordon equation' (SGE in short) has its origin in its resemblance to the well-known Klein-Gordon equation for a free particle in relativistic quantum mechanics, which reads $\sum_{j=x,y,z} \phi_{jj} - \phi_{tt} - \phi = 0$, in natural units $m = c = \hbar = 1$, and was discovered in 1926. Of course, the Klein-Gordon equation is a linear partial differential equation, which can be considered as special case of the SGE obtained by retaining only first term in the Taylor series expansion of $\sin u$.

One of the soliton solutions of Eq. (35) is

$$u(x, t) = 4 \tan^{-1} \left[e^{\pm \frac{x - \alpha t - x_0}{\sqrt{1 - \alpha^2}}} \right], |\alpha| < 1, \quad (36)$$

where α is normalized velocity of propagation of the solitary wave. The initial position x_0 can be easily taken as 0. Note that for finite constant value of αt , $u(x, t)$ in Eq. (36) with positive exponent has values 0, π and 2π rad for $x \rightarrow -\infty$, $x = \alpha t$ and $x \rightarrow \infty$, respectively. On the other hand, the corresponding values of $u(x, t)$ with negative exponents are 2π , π and 0 rad. Thus, the solution given by Eq. (36) is monotonically varying function of x , and is such that as x increases from $-\infty$ to ∞ for fixed value of t , u changes from 0 to 2π for positive exponent and from 2π to 0 for the negative exponent. The value of u in both the cases is π when $x = \alpha t$. This feature of the solution for the SGE is interpreted as following. Eq. (36) represents a twist or kink having same sign as that of the exponent. These two situations define soliton and anti-soliton, respectively, and are known as 2π -kink and -2π -kink (or antikink); Fig. 3. In the context of nonlinear optics, these are, respectively, referred to as $+2\pi$ pulse and -2π pulse.

It is worth mentioning that a soliton solution is said to be topological if it has its origin in topological constraints and a twist with variation in the value of x is an example of this situation. As such, the SGE kink is an iconic one-dimensional topological soliton while the Russell's water wave soliton is non-topological. In fact, the structure of a system is changed after the passage of a topological-soliton wave through this.

The soliton solutions of SGE and its modified versions find numerous and in-

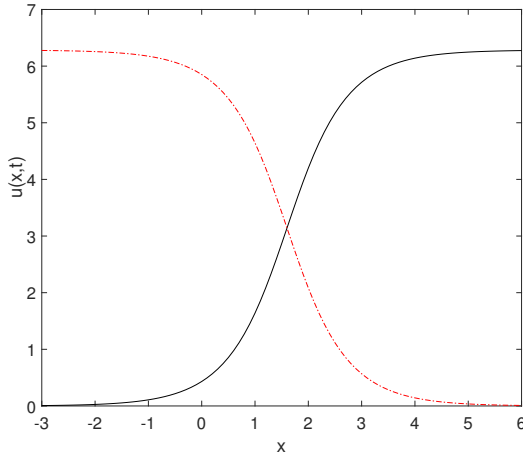


Figure 3: A sketch of the analytic solution $u(x, t)$ as function of x for the Sine-Gordon equation for $\alpha = 0.8$, $t = 2.0$, and $x_0=0$. The black line represents a 2π – kink, while the dash-dot red line depicts antikink. These have value π rad at $x = 1.6$.

valuable applications in condensed matter physics, nonlinear optics, biophysics, astrophysics, relativistic field theory, geophysics particularly seismic modeling, and in the description of mechanical transmission lines.

5.2 Nonlinear Schrödinger Equation

In 1968, Zakharov while studying nonlinear waves in the small amplitude approximation on the surface of a deep fluid introduced a NLPDE that can be written as

$$iu_t + u_{xx} \pm |u|^2 u = 0 \quad (37)$$

This equation is referred to as Nonlinear Schrödinger equation (abbreviated as NLSE) because it looks like the highly acclaimed

1-dimensional time-dependent Schrödinger equation of non-relativistic quantum mechanics (i.e., $i\hbar\psi_t + \frac{\hbar^2}{2m}\psi_{xx} - V\psi = 0$), with nonlinear term $\pm|u|^2$ corresponding to potential V . Note that, the Schrödinger equation is a linear PDE and $\psi(x, t)$ is wavefunction of the particle assumed to be spinless. Of course, generally, the derivation of NLSE has nothing to do with quantum mechanics. The exact analytic solution of NLSE, obtained by Zakharov and Shabat in 1972 by using the inverse-scattering method, showed that these describe deep-fluid wave-envelope solitons which modulate a periodic sinusoidal wave. These findings were experimentally verified by Yuen and Lake in 1975. A different solitary wave solution to this equation was reported by Ma in 1979, and a rational-cum-oscillatory solution was presented by Peregrine in 1983. Note that $\pm|u|^2$ in the nonlinear term in NLSE is a sort of self-interacting quantity, wherein upper and lower signs, respectively, represent repulsive and attractive self-interactions. In view of this feature, Eq. (37) is also known as cubic Schrodinger equation.

However, before proceeding further, it may be pointed out that NLSE is a simplified version of the equations used by Ginzburg and Landau in 1950 in their study of the macroscopic theory of superconductivity, and by Ginzburg and Pitaevskii in 1958 in the theory of superfluidity. Furthermore, in 1964, Chiao et al and Talanov employed similar equation while investigating

the phenomenon of self-focusing of optical beams and the conditions under which an electromagnetic beam can propagate without spreading in nonlinear media.

The soliton solution of Eq. (37), with + sign for the nonlinear term, determined by Zakharov and Shabat reads

$$u(x, t) = ae^{i\{\frac{v}{2}(x-vt)+bt\}} \operatorname{sech}\left\{\frac{a(x-vt)}{\sqrt{2}}\right\}. \quad (38)$$

Here, the wave amplitude a , velocity v , and real constant b are such that $a^2 = 2\left(b - \frac{v^2}{4}\right) > 0$. While writing this solution, the initial phase and the initial position appearing in the exponential and the sech terms have been assumed to be zero. The exponential term leads to an oscillatory component with amplitude dependent sech term as the envelope profile so that the resulting wave packet is a modulated one; Fig. 4. Such a solitary wave described by an envelope with an internal oscillation or pulsation, is called a breather. Sometimes, the terms envelope soliton and intrinsic localized modes are also used for this entity, particularly in nonlinear lattice dynamics. The $u(x, t)$ given by Eq. (38) represents a moving breather as it advances in space. In contrast, a breather solution has been obtained for the SGE that does not move and, hence, is referred to as a stationary breather.

In nonlinear optics, the breather solution that produces self-focusing of the carrier wave is known as bright soliton and the one giving self-defocusing is called the dark soliton.

It may be added that in addition to

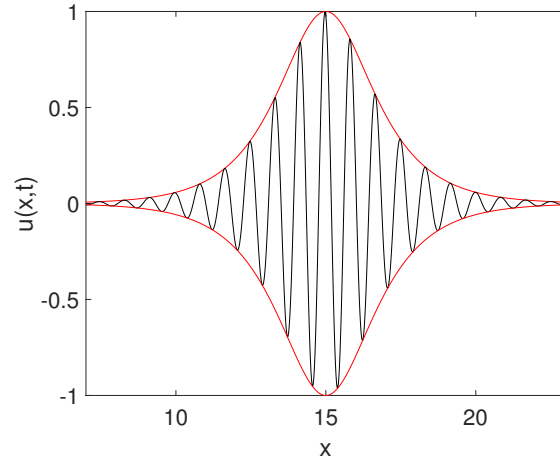


Figure 4: Profile of the breather solution given by real part of Eq. (38) for $a = 1.0$, $v = 15.0$, $t = 1.0$, $b = 56.75$. The black line depicting internal oscillations is confined within the red-line envelope soliton.

continuous NLSE, soliton solutions have also been found for the discrete nonlinear Schrödinger equation

$$iu_{n,t} + u_{n+1} + u_{n-1} \pm |u_n|^2 u_n = 0, \quad (39)$$

and some of its generalizations leading to proper elaboration of many interesting properties of nonlinear lattice chains.

The models that are compliant with NLSE and its different variants have played an important role in the developments in nonlinear optics (light waves), soft-condensed matter physics particularly Bose-Einstein condensation (matter waves), fluid dynamics, plasma physics, etc. and continue to be valuable even now.

6 Some Applications of Solitons

The concept of solitons, including their different cousins, as sophisticated mathematical constructs to explore nonlinear phenomena has not only revolutionized research in mathematics as well as mathematical physics leading to advent of many new ideas and techniques, but has also been fruitfully exploited in developing numerous practical applications in different branches of science and engineering. Besides the three NLPDEs discussed in Sections 4 and 5, many other equations such as modified KdV equation, Benjamin-Ono equation, Boussinesq equation, Davydov's equations, etc. have been found to be of immense value. Most of different soliton bearing NLPDEs have been solved analytically as well as by numerical methods. In this section, we briefly describe some typical problems in various fields where solitons with their different manifestations have been employed, and we certainly do not claim exhaustiveness of the list. Also, the topics dealt with are being listed in alphabetical order.

1. Astrophysics and Cosmology

- (a) Electrostatic solitary waves have been experimentally observed in astrophysical plasmas such as the sun, the solar wind, lunar wake, the planetary magnetospheres, etc. Also, many theoretical models have been proposed to interpret the observed characteristics of these waves.

- (b) It has been shown that low dimensional black holes can be realized as solitons of the sine Gordon equation. Furthermore, it has been ferreted out that some field theoretic models for studying black holes also have soliton solutions indicating their intimate relationship.
- (c) The Great Red Spot of Jupiter (GRS), which is slightly oval and nearly 16,000 km wide, is an anti-cyclonic vortex that has persisted for hundreds of years of continuous observation despite the highly turbulent atmosphere on the planet. Two- and three- dimensional soliton models were put forward for this in 1980s, wherein the latter were found to be in better agreement with the then available data. However, some scientists believe that these models do not capture every minute detail of the GRS as soliton.
- (d) The solutions of some cosmological models with the cosmological constant have been found to exhibit the existence of solitary waves under specific conditions besides the travelling wave periodic solutions that resemble the gravito-static waves.

2. Biological Systems

- (a) The model developed for energy

transfer and energy coupling in hydrogen-bonded spines that span the length of protein α -helices and stabilize it, encompasses the so called Davydov soliton. This soliton represents a state composed of an excitation of amide-I and its associated hydrogen-bond distortion. In fact, the relevant conjectures have been supported by different spectroscopic studies of proteins.

- (b) The concept of Davydov soliton has also been used to describe a local conformational change of the deoxyribonucleic acid (DNA) α -helix, which too has been confirmed by experiments.
- (c) Generation of force in the sliding filament model of muscular contraction has also been attributed to Davydov solitons.
- (d) The intrinsic localized modes, which arise from the anharmonicity of interatomic potentials, have been observed in proteins and identified as solitons localized in both space and time.
- (e) The studies pertaining to electron / proton transport in α - helix sections of proteins, and the signal as well as energy propagation in lipid membranes have brought out the involvement of soliton-like mechanisms.
- (f) Solitons obtained as solutions for the Peyrard–Bishop model (and its extended versions) put forward to understand the dynamics of DNA, explain important features like local opening (i.e., separation of double-stranded DNA into two single strands) and DNA transcription. A similar soliton-bearing model has also been developed to elucidate the long-distance charge transport in DNA molecule.
- (g) The behaviour of many biopolymers has been explained in terms of breathers and this aspect too has been investigated experimentally.
- (h) The concept of solitary waves has been recently used in neuroscience as an alternative to the earlier accepted ionic-hypothesis based Hodgkin–Huxley model to describe the propagation of signals along the excitable cells such as neurons and cardiac myocytes.
- (i) Soliton-related mechanisms have been reported to play an important role in the eukaryotic multicellular movements during morphogenesis and development.
- (j) It has been proposed that the blood pressure pulse is an outcome of a KdV soliton produced in the heart and its propagation in blood vessels.

3. Condensed Matter Physics

- (a) The phase-slip centres in the charge density wave condensate formed during phase transitions in which the electron density develops a small periodic distortion accompanied by a corresponding modulation of the ion equilibrium position, are solitons. These solitons are found in one-dimensional metals and organic conductors.
- (b) Solitons occur in structural phase transitions in quasi-one-dimensional ferroelectrics.
- (c) The flipping of spins in magnetic phase transitions in quasi-one-dimensional ferromagnets as well as antiferromagnets is associated with the kink solitons.
- (d) Solitons have been found to be instrumental in polymerization mechanism and creation of bond defects in polymers.
- (e) The phenomena of transport and existence of defects in two-dimensional Coulomb gases and two-dimensional spin systems (i.e., thin films) are understood in terms of solitons.
- (f) A domain wall or a Bloch wall in ferromagnets, ferrimagnets, ferroelectrics, etc. is an interface that separates magnetic or electric polarization domains of different types. These walls are exact solutions to SGE, NLSE, and their modifications and, hence, these have been identified as relevant solitons. These aspects have been experimentally verified by neutron scattering, NMR, and ESR studies in many materials.
- (g) Starting from the Frenkel-Kontorova model with on-site periodic potential, mentioned earlier in Section 5.1, the atomistic theories of crystal dislocations have been generalized to include physically more relevant non-sinusoidal and anharmonic interactions. Solitons, particularly kinks, have been found to play an important role in all these models and have been confirmed in experimental measurements.
- (h) Liquid crystals (which are used in display devices like televisions, computer monitors, laptop screens, calculators, etc.) are self-organized anisotropic fluids that are thermodynamically intermediate between the isotropic liquid and the crystalline solid, showing the fluidity of liquids and the order of crystals. Thus, these are mesophase entities. Being non-linear materials, these have been widely used for creation and description of various types of solitons since 1968. The associated as-

pects have led to new applications of the liquid crystals.

- (i) Solitons have been experimentally observed in thin superfluid ^4He films (a few atomic layers thick) adsorbed on solid substrates (two-dimensional system) as well as bulk superfluid ^4He (three-dimensional quantum material) and have been theoretically expounded using KdV equation and still better by employing phenomenological modeling based on time-dependent density functional theory.
- (j) Different types of solitons have also been observed in ultracold superfluid ^3He phases in the absence as well as presence of magnetic field (magnetic solitons). These have been explained theoretically using an NLSE like equation.
- (k) An arrangement or device obtained by sandwiching a thin layer of a non-superconducting material (up to about 3 nm thick insulator or a few μm thick non-superconducting metal) between two layers of superconducting material, is known as a Josephson junction. It has a unique and important feature that a dc (supercurrent) can pass through the junction / barrier from one superconductor to the other even in the absence of an applied volt-

age and a sinusoidal ac current is generated when a fixed voltage is applied across it. The former is a consequence of quantum tunneling of Cooper pairs (pairs of electrons with opposite momenta and spins loosely bound at very low-temperatures due to electron-lattice interactions) across the nonconducting barrier and the latter makes it a nonlinear oscillator. These junctions find applications in quantum-mechanical circuits such as superconducting quantum interference devices (SQUIDs), superconducting qubits, and rapid signal flux quantum digital devices. The dynamics of the Josephson junction is reasonably well described by a perturbed SGE, which makes it a system for the study of solitons and phenomena associated with these. In fact, discrete breathers have been observed in arrays of Josephson junctions, and the solitons in the junctions which are much longer than characteristic Josephson penetration depth (which is of the order $1 - 1000 \mu\text{m}$), are known as fluxons because they contain one quantum of magnetic flux ($h/2e = 2.07 \times 10^{-15} \text{ Wb}$; here h is Planck's constant and e is charge of an electron).

- (l) The cumulation of a macroscopic

fraction of noninteracting identical boson particles (the entities having integer spin, which is actually an integer multiple of $\hbar = h/2\pi$, and is described by symmetric wavefunction) in the lowest energy or the ground state in a system under appropriate characteristic conditions of temperature, number density, etc. is known as Bose-Einstein (BE) condensation. It represents a phase transition to a state of matter in which a good number of constituents of the system suddenly coalesce into a single coherent quantum mechanical entity that can be described by a wavefunction on nearly macroscopic scale. The condensate appears as a sharp peak in both position and momentum space. The macroscopic dynamics of BE condensates near 0 K is generally modeled by a 3-dimensional version of NLSE with a term for the trap potential and is called the Gross-Pitaevskii equation. The solutions of this and other similar equations, lead to solitons of different types which have been observed experimentally as well. Besides, investigations on manipulating the properties of solitons in the BE condensates via nonlinearity management have also been carried out.

4. Engineering

- (a) Mathematical and computational studies of a variety of problems in theoretical aerodynamics have shown that in some situations, solitons can lead to chaotic motion.
- (b) The use of electrical components with nonlinear permittivity and permeability makes a transmission line to be nonlinear. In fact, such transmission lines (constructed with easily available components) constitute reasonably simple and low-cost experimental devices for investigating various aspects of nonlinear waves. These properly designed networks have been shown to produce (electrical) soliton pulses over a wide range of frequencies and find applications in wide band focusing and shaping of signals, and in instrumentation for microwave systems, in high-speed sampling oscilloscopes, and for data transmission in high-speed digital circuits, etc.
- (c) A structure such as a rod or a pile of plates with rectangular or circular cross-section made from some metal, polymeric materials, etc. that propagates elastic waves with minimal loss of energy by restricting their transmission along its length, is called an elastic or a solid waveguide. Such waveguides provided with piezoelectric transduc-

ers are used for measurement of strain, pressure, and temperature. In addition, these waveguides find applications in energy harvesting, vibration control, health monitoring, and wave steering for actuation. If the nonlinearity produced by properties of the constituent material and by the strain is compensated by the spatial dispersion caused by the finite transverse size of the waveguide, then longitudinal density solitary waves can be generated in it. These are generally described by Boussinesq-type NLPDE of elasto-dynamics and have been observed experimentally. These so-called strain or bulk solitons represent a powerful localized wave that can transport elastic energy over reasonably large distances with negligible losses.

- (d) Granular crystals are nonlinear tailored metamaterials obtained from tight packing of macroscopic solid grains or particles like ball bearings made of a metal or an alloy or bits of polymers such as nylon, teflon, delrin, etc. (rather than atoms or molecules) that interact elastically. Like atoms in a crystal, the particles in a granular crystal can also be arranged in one-, two-, or three-dimensional lattices. The freedom to choose constituent particles with different

masses, sizes, material properties, and geometries, and possibility to arrange these in a variety of configurations in a lattice, make the granular crystals highly tunable even in respect of the extent of nonlinearity. The dynamical description of these fabricated crystals brings out existence of traveling solitons as well as discrete breathers in these, which have been observed experimentally. These aspects, in turn, have made these engineered or manipulated materials useful as the shock-absorbers in armor and sports helmets; for sound-focusing devices, acoustic switches, acoustic logic elements; for mechanical vibrational energy harvesting systems; and for converting mechanical vibrations into electrical current that could drive small sensors or transmitters.

- (e) The micro-electromechanical and nano-electromechanical systems, generally made from materials like carbon nanotubes and graphene, are artificial devices that combine electrical and mechanical processes at micro and nano scale, respectively. These find applications in automobiles, accelerometers, aerospace systems, sensors for environmental monitoring, defence systems, biomedical diagnostics, medical

devices, signal processing, wireless communications, etc. Studies pertaining to dynamics of many such systems, particularly those comprising arrays of nonlinear oscillators, have established the existence of discrete breathers in these.

5. Hydrodynamics and Geophysics

- (a) The discovery of 'great wave of translation' by Russell in shallow water and its theoretical modeling by KdV equation, have motivated many scientists to study multifarious properties of shallow-water solitary waves in the laboratory. A variety of wave-tank experiments have been performed to investigate various aspects of these waves, including different types of collisions between solitons, and these continue to be of interest even now. Generally, the experimental results exhibit good agreement with relevant theoretical predictions. In addition, it is well recognized that various properties of shallow-water waves near the beaches are successfully explained by the KdV equation.
- (b) The surface waves observed in deep water have been identified as envelope solitary waves, whose theory was developed by Zakharov in terms of an NLPDE simi-

lar to the NLSE. These waves have also been investigated experimentally using large water tanks.

- (c) The soliton solutions of the KdV as well as the Benjamin-Ono equations have been used to describe internal gravity waves in the ocean, which are large amplitude waves travelling at low speed and originate from density differences caused by variations in temperature and saline concentration. These have been observed and painstakingly studied in many seas by oceanographers.
- (d) The seemingly spontaneous and extremely large rogue or monster waves too have been modelled as solitary waves.
- (e) It has been argued that strong velocity-dependence of amplitude of a solitary wave disturbance on the surface of water in an ocean created by an underwater earthquake, volcanic eruption, etc. makes its amplitude larger as the wave advances towards a beach. If the wave energy is quite high, then amplitude becomes so large that the wave breaks down into numerous waves of very large width (few hundred kilometers) and small amplitude (1 meter or so) as it reaches the beach. The catastrophe so created at the beach results in devastating tsunamis and hurri-

canes.

- (f) Because of resemblance of some mountain ranges and layer distribution in some sedimentary rocks with envelope wave packets, NLPDEs based mathematical models have been developed to show that geo-solitons may have played an important role in their formation.

6. Nonlinear Optics

Nonlinear optics is the branch of optics that deals with the behaviour of light in the materials in which the electric polarization produced by the electric field of the light passing through it varies as higher powers of the electric field strength, i.e. nonlinearly, particularly when the light intensity is high. When a highly intense beam of laser radiation propagates through a material like silica-based glass, lithium niobate, etc., additional phase shift (called self-phase modulation) is introduced due to intensity dependent refractive index (the Kerr effect). This nonlinear phase shift in the pulse leads to its shrinkage in contrast with spreading produced by dispersion. If these two opposing effects cancel each other, we get temporal optical solitons. Besides temporal optical solitons, spatial optical solitons have also been found to exist in many nonlinear media. When an intense light beam passes through such bulk

materials along, say, x – direction, it may undergo diffraction along the two transverse directions. If the broadening produced by diffraction is counter-balanced by the narrowing caused by the nonlinearity associated with intensity dependent refractive index, spatial optical solitons are obtained. In addition to the temporal and spatial solitons, spatiotemporal optical solitons (where both the diffraction and dispersion effects are simultaneously compensated by nonlinearity) have also been created in some nonlinear optical materials. In a nutshell, an optical soliton refers to a situation where light beam or pulse (self-trapped in time or space or both) travels through a nonlinear optical material without any change in its profile and velocity. These solitons are mathematically described by NLSE (continuous as well as discrete) and are found in photonic crystal fibres, photorefractive materials, photopolymers, etc.

- (a) The idea of temporal soliton transmission in glass fibre waveguide (or an optical fibre) was put forth by Hasegawa and Tappert in 1973 on the basis of theoretical and numerical calculations, and its experimental observation in silica-glass fibre was reported by Mollenauer et al in 1980. When laser pulses are used for communication employing optical fibre, the solitons

involved are sometimes referred to as fibre-solitons. Presently, it is possible to propagate solitons without degradation over thousands of kilometers. Such a communication has zero loss and no dispersion, which explains the focus of a great research effort to understand the dynamics of soliton transmission in optical fibres. Furthermore, it brings out the importance of optical fibre communication in information technology and in the long-distance, high bandwidth communication – the well-known internet and the world-wide web.

- (b) The spatial solitons in photorefractive polymers make these highly efficient optical elements for transmission of data and for controlling coherent radiation in various electro-optical and optical communication devices.
- (c) Ultra-short pulse solitons are being used in the field of optical spectroscopy and medicine.
- (d) Optical solitons in birefringent optical fibres are used for optical switching.
- (e) The fabrication of materials with extremely strong nonlinear effect has made it possible to create optical solitons even with very low laser powers. These find applications in optical information storage

of large amount of data, all-optical switches, and significantly faster optical systems than any known electronic devices. These concepts form essential basis of the possible optical digital computer system or the photonic computer with solitons as bits.

- (f) Light or optical bullets which are three-dimensional localized pulses of electromagnetic energy and have been observed in arrangements like array of silica glass waveguides, sapphire samples, plasmas, etc., are examples of spatiotemporal solitons. However, these lose energy during interactions / collisions implying that these are not solitons in the strict sense of the term.

7. Nuclear Physics

- (a) Topological solitons have been found to describe reasonably well some properties of nuclei, including prediction of binding energies to the correct nuclear physics level. This aspect has been found to have substantial impact on the studies pertaining to nuclear matter in neutron stars and in nuclear fusion.
- (b) Nontopological soliton models based on simple phenomenological field theories have been used

to incorporate the quark structure of hadrons in nuclear physics.

- (c) It has been shown that the velocity dependent terms in the nucleon-nucleon potential lead to formation of solitons in nuclear matter, which play role in nuclear multi-fragmentation reactions.

8. Plasma Physics

Plasma physics deals with the study of matter consisting of a large number of charged particles – ions and / or electrons. The presence of inter-particle coulomb interaction makes plasmas a nonlinear system and, as such, these offer a good testing ground for the study of solitons.

- (a) The KdV and some other similar NLPDEs have been used to describe the local charge density reflecting the local departure of the charge from neutrality in the plasmas, and, thus, establishing the presence of travelling solitons in these.
- (b) Ion-acoustic solitary waves have been theoretically and experimentally studied in magnetized plasma.
- (c) The Alfvén waves, observed in plasmas on the earth and in the space, are low-frequency travelling oscillations of the ions caused by the interaction of the magnetic

fields and electric currents within the plasma. These magneto-hydrodynamic waves were among the first to be modeled using idea of solitons.

- (d) Dusty plasmas, which contain small suspended particles, have been modeled using nonlinear oscillator chains, showing the existence of discrete breathers in these.
- (e) It is well known that space debris objects, whose number in earth's orbit is estimated to be few hundred million, pose immense threat to the earth-orbiting satellites. Also, these objects get electrically charged due to their exposure to the ionospheric plasma environment. Some recent analytical, computational, and experimental investigations have shown that charged objects moving with high speed through a plasma lead to generation of plasma density solitons. Accordingly, depending on its size, charge and velocity, debris object will produce solitons, which can be detected by fixing simple instruments on the spacecraft.

9. Quantum Mechanics, Elementary Particle Physics, and Field Theory

- (a) With a view to develop a classical interpretation of quantum mechanics, Bohm and others treated

quantum processes as stochastic processes. This approach was subsequently used to derive a nonlinear relativistic Klein-Gordon equation yielding soliton solutions, which follow the average de Broglie-Bohm trajectories analogous to the linear solutions of the Schrödinger and the Klein-Gordon equations. These ideas have been extended to show that even photon can be represented as a soliton. A relationship between the electromagnetic amplitude of this soliton and photon energy or frequency has been established. Also, it has been proved that the concept of photon-soliton is in conformity with the familiar interactions in the photoelectric and Compton effects.

- (b) Recall that solitons are confinement of energy of the wave-field, propagate without change in shape, collide like particles, and a soliton-antisoliton pair may get annihilated. In view of these facts, it was conjectured that if an appropriate system of nonlinear field equations admits soliton-solutions then these may represent elementary particles. As such, 'bags' and 'lumps' in quantum fields are described in terms of solitons. However, many of these issues are still being debated [4].

- (c) The instanton solutions of Yang-Mills field equations used for unifying electromagnetic and weak forces are soliton-like because these are localized in space as well as time.
- (d) In order to explain the stability of protons, neutrons, and mesons, Skyrme (1961) developed a model in which these elementary particles could be treated as topological defect solitons in a quantum field. This stable field configuration with special topological properties came to be known as skyrmion. However, this idea did not find much ground in particle physics even though it accounted for some low-energy properties of the nuclear particles. Interestingly, skyrmion-like topologies have been found to exist in many condensed matter systems such as some liquid crystal phases, BE condensates, quantum Hall systems, and helimagnetic materials in which neighbouring magnetic moments arrange themselves in a helical or spiral pattern. In the last category of materials exemplified by FeGe, Tb, Dy, etc. these form domains as small as 1 nm and involve extremely low energy. These features make magnetic skyrmions a good option for developing very efficient memory-

storage and other spintronics devices. In fact, the activities in this direction constitute the emerging field called skyrmionics.

- (e) The Einstein field equation, which constitutes the backbone of the general theory of relativity, describes gravity to be a consequence of spacetime being curved by both mass and energy. It is, in fact, a set of ten NLPDEs in four independent space and time variables, expressed as a tensor equation. Non-linearity of the equation leads to a solution, which has soliton characteristics (confined to a finite region of spacetime and has a finite energy) and is called the gravitational soliton. It can be separated into two kinds - a soliton of the vacuum Einstein field equation and a soliton of the Einstein-Maxwell equations. Even black holes, the main sources of gravitational radiation, are two-soliton solutions of Einstein's equations in vacuum.

7 Epilogue

It is indeed very interesting to note that solitons occur over a wide range of scales. On one side, these have been found to be extremely useful in understanding various phenomena at the nuclear and atomic level, though their experimental manifestation is

not that straightforward. On the other hand, these have been extensively observed and manipulated in the macroworld. The linear dimension of nuclear solitons is few femtometer (10^{-15} m), of the optical solitons is few nm (10^{-9} m), of the solitons observed on the surfaces of water bodies is few cm to few meters or even few hundred kilometers, and of the GRS solitons is thousands of kilometers. Solitons associated with BE condensates are observed at ultra-low temperatures of 10^{-7} K or so, the hydrodynamic solitons occur at around 300 K, the temperature over the GRS is about 1600 K, and core temperature of the sun is about 10^7 K.

The distinctive behaviour of Josephson junctions including magnetic flux quantization, superfluidity of helium, and Bose-Einstein condensation (discussed under condensed matter physics in Section 6) are manifestations of quantum effects at macroscopic level. All of these have their origin in the collective coherent behaviour of constituent quantum particles with nonlinear interactions, which balance the dispersive effect of kinetic energy. As mentioned earlier also different species of solitons (which are coherent structures created by perfect balance between effects of nonlinearity and dispersion) have not only been predicted in these systems using relevant NLPDEs (like NLSE, etc.) but have also been observed experimentally. Thus, the above-mentioned systems are nonlinear quantum phenomena where the Hamiltonian is a nonlinear function of the wavefunction of the micro-

scopic entities involved. It has been argued (see, e.g. [4]) that nonlinear quantum theory based on NLPDEs with solitons as integral part be developed in proper perspective to describe such systems and to investigate related features in detail. It can be said without any exaggeration that this development will act as stimulant for a new surge of soliton-oriented activities in condensed matter physics, polymer science, and biophysics.

Lastly, to conclude the article, we quote Kasman [9] “solitons have become (*vital*) tools of scientists and engineers for understanding the universe”.

8 References

- [1] S. P. Puri, *Textbook of Vibrations and Waves*, (Macmillan India, New Delhi, 2004).
- [2] P. G. Drazin and R. S. Johnson, *Solitons: An Introduction*, (Cambridge University Press, Cambridge, 1989).
- [3] M. Remoissenet, *Waves Called Solitons: Concepts and Experiments*, (Springer-Verlag, Berlin, 2003).
- [4] B. Guo, X. Pang, Y. Wang, and N. Liu, *Solitons*, (De Gruyter, Berlin, 2018).
- [5] R. M. Miura, *SIAM Review* **18**, 412 (1976).
- [6] L. Debnath, *Int. J. Math. Edn. Sc. & Tech.* **38**, 1003 (2007).
- [7] N. J. Zabusky and M. A. Porter, *Scholarpedia* **5**(8), 2068 (2010).
- [8] J. Bundgaard, https://inside.mines.edu/fs_home/tohno/teaching/PH505.2011/A-Survey-of-The-History-and-Properties-of-Solitons.pdf (2011).
- [9] A. Kasman, *Current Science* **115**, 1486 (2018).
- [10] S. Manukure and T. Booker, *Partial Diff. Eqns. Appl. Maths.* **4**, 100140 (2021).
- [11] N. J. Zabusky and M. D. Kruskal, *Phys. Rev. Lett.* **15**, 240 (1965).
- [12] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, *Phys. Rev. Lett.* **19**, 1095 (1967).