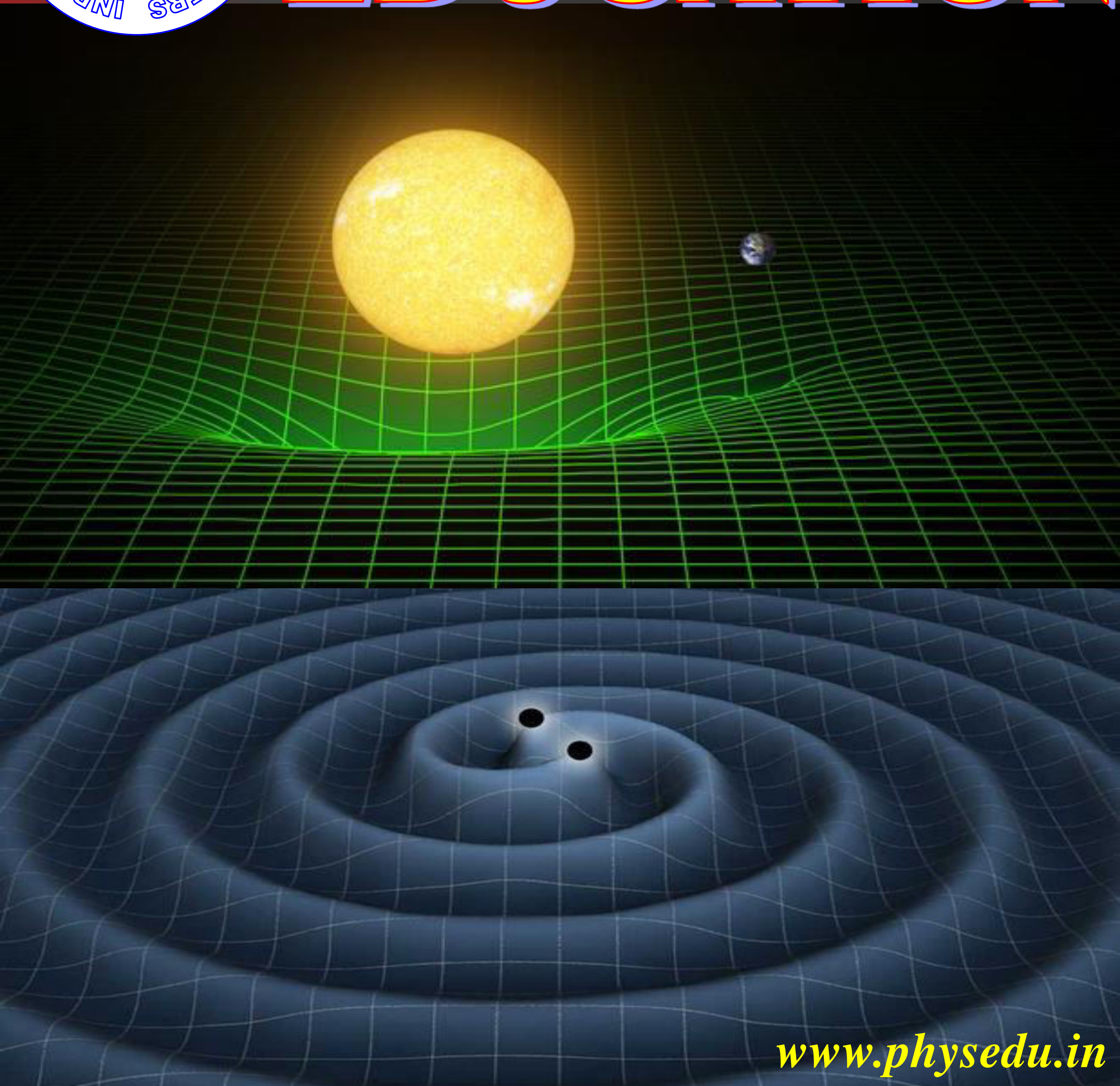
Vol 32 No. 4 Oct - Dec 2016 ISSN 0970-5953

# PHYSICS EDUCATION



www.physedu.in

Volume 32, Number 4

## In this Issue

# EDITORIAL

As this issue of *Physics Education* reaches you, the new year 2017 would have already begun. Nevertheless, I wish all our contributors and readers a happy and prosperous new year.

When *Physics Education* went online in 2012, it was presumed that the entire editorial processing would be done online. It is indeed true that the entire processing, right from submission, refereeing until publication, is completely online. This also requires that our authors help us by submitting manuscripts in the prescribed format as well as take responsibility to provide clear figures and other related materials before publication. As we enter the new year, I hope that authors will extend their fullest cooperation in this process. This would not only speed up processing but also helps us concentrate on other matters related to this journal.

Talking of time, the year 2017 was heralded in India with a one second delay to synchronise the solar time and the atomic time. India's National Physical Laboratory at New Delhi did this exercise in their capacity as the standard time keepers for the nation. Once again, it underlines the crucial, though often silent, role played by physics in many areas of national interest.

M. S. Santhanam
Chief Editor
Physics Education

# How a Diode Tames the Sun

Yashika Gupta[1] [2]

[1]Material Science Research Lab,
S.G.T.B. Khalsa College, University of Delhi, Delhi, India.

[2]Department Of Electronic Science, University of Delhi, South Campus, Delhi, India.
g.yashi01@gmail.com

(Submitted 28-09-2016)

---

**Abstract**

With growing interest in renewable energy, a large work force would be required to engineer, install and maintain devices for energy harvesting. The current trend points to the non-polluting use of solar energy. Recognising this, it becomes necessary to have a large audience understand the basic working of solar cells. This article focuses on the conceptual understanding of a p-n semiconductor junction and how by material engineering the junction can be used to harvest solar energy. The manuscript provides an insight into the simple semiconductor physics and its implementation at the device level.

---

## 1 Introduction

To meet the growing demands of energy and with limited reserve of fossil fuels, there is a great need for low cost, enviornmental friendly renewable energy resources. Among other natural energy resources like wind, hydrothermal, biomass etc solar energy stands out to be most promising candidate with the Sun being the never-ending source of light energy with wavelength range spanning the electromagnetic spectrum from UV to Infrared [1, 2]. When a material is incident with light having energy greater than or equal to the band-gap of the material, the electrons absorb this energy and is excited from the valance band to the conduction band, becoming free for conduction process. This phenomenon of producing voltage or electric current is known as photovoltaic effect and the device working on this principle is known as the solar cell. It was first discovered by a french scientist, Edmund Becquerel in 1839 who also give this phenomenon its name [3]. However, the world got its first practical solar cell much later in 1954 when Chapin et al at Bell laboratories, UK developed a simple p-n junction Silicon-solar cell with 6 % conversion efficiency of solar energy into elecrticity [4]. Till date Si-solar cells with efficiency ≈ 25 % dominate the photovoltaic industry. However, due to high processing cost and expensive Si-technology, focus has been shift towards a large number of other organic and inorganic materials eg. P3HT, CdS, SnS, CIGS etc. as solar cell absorbing layer [5, 6, 7, 8, 9, 10], leading to a rise of thin film solar cell technology. Thin film solar cells although have cut down the manufacturing cost by great margins but they still have to go a long way as far as conversion efficiency is concerned.

In this manuscript, we will discuss how a simple p-n junction solar cell works and what important parameters one has to consider while designing a thin film solar cell. But before going into the details of photovoltaics, lets begin with an over-view os a simple p-n junction.

## 2 p-n junction

A p-n junction is formed when a p-type semiconductor (having holes as majority carriers) is grown over a n-type semiconductor (electrons as majority carriers) or vice-versa, using one of the fabrication methods such as thermal evaporation, lithography, wet chemical methods etc.. Because of the difference in electron's and hole's concentration in the two layers, a concentration gradient is set leading to diffusion of charge carriers. Electrons move from n-type region (where they are the majority charge carriers) to p-type region (where they are in minority) and similarly holes move from p-type to n-type region. Since this type of carrier movement results in an increase in number of minority carriers on both p and n-side of the junction, this process is called minority-carrier injection and the resulting current is known as diffusion current.

These diffused minority charge-carriers recombine with the majority carriers present in the layers and eventually deplete all the free charge carriers in the vicinity of junction (on both sides) leaving behind immobile charge ions, donors ions ($\mathrm{N_d}$, positive) on n-side and acceptor ions ($\mathrm{N_a}$, negative) on p-side of junction. This region, depleted of mobile charge-carriers is known as depletion region/space charge region and extends upto $\mathrm{x_p}$ and $\mathrm{x_n}$ in the two layers as shown in fig 1. The loss of charge neutrality in depletion region give rise to an electric field with direction from n-side (positive donor ions) to p-side (negative acceptor ions).

This electric field prevents the further diffusion of charge carriers and itself exerts a force on the carriers causing a drift of electrons from p-side to n-side and that of holes from n-side to p-side i.e. in the direction opposite to the diffusion current. Therefore, the electric field builds upto a point where equilibrium is reached and both diffusion and drift currents balance each other i.e.

$$I_{diffusion} = I_{drift} = I_o \tag{1}$$

At equilibrium, the electrochemical potential[1] represented by fermi-levels, on either side of the junction are equal. This allignment of fermi levels results in bending of bands in the depletion region and hence to an electric potential known as built-in voltage '$V_o$' or '$V_{bi}$'.[2] Since the electric field is maximum at the junction while zero in the neitral region, it implies that there exists a potential gradient in the depletion region. The distribution of space-charge and the potential in this region is given by Poisson's equation [11] For depletion region $0 < x < x_n$

$$\frac{d^2V}{dx^2} = \frac{-q}{\epsilon_s}(N_d) \tag{2}$$

And for depletion region $-x_p < x < 0$

$$\frac{d^2V}{dx^2} = \frac{q}{\epsilon_s}(N_a) \tag{3}$$

Where, '$\epsilon_s$' is the dielectric constant of the semiconductor material and $x = 0$ at the junction and increases as we move away from the junction into the depletion region. The potential is '$V_p$' and '$V_n$' at '$-x_p$' and '$x_n$' respectively. $V_{bi}$ is given as

$$V_o = V_n - V_p \tag{4}$$

i.e. the potential difference between the constant potentials in the neutral regions.

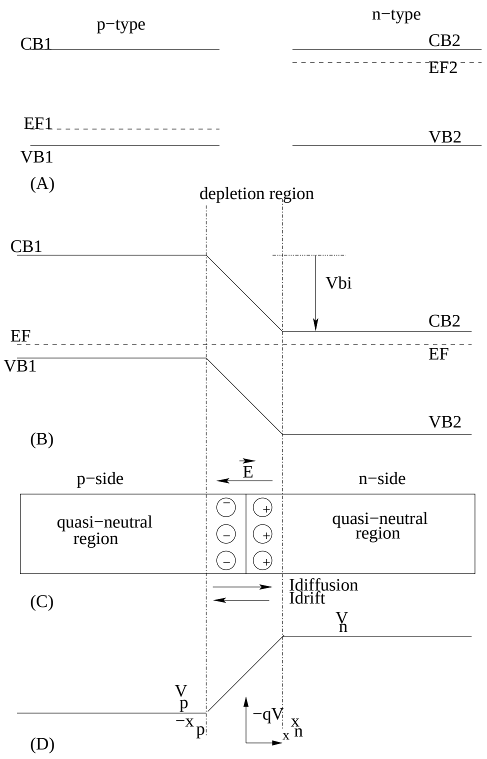


Figure 1: Figure showing (A) energy-band diagrams of isolated p-type and n-type semiconductor (B) energy-band diagrams of a p-n junction (C) different regions of a diode with direction of currents flowing through it (D) distribution of electrostatic potential across the diode.

Therefore, any p-n junction device (as shown in the fig 1) consists of three regions:

(i) Depletion or the space-charge region

---

[1] It describes the average energy of carriers, generally given in terms of electron energy.

[2] In order for the charge carriers to diffuse further, they have to overcome the potential barriers '$V_{bi}$' which is also known as knee voltage i.e. $V_{bi}$ is the minimum voltage that must be applied across the junction in order for the current to start flowing through the diode. It is necessary for maintaining equilibrium at the junction and is not an external potential and therefore can't be measured using a voltmeter.

(ii) Quasi-neutral region on both sides of depletion region, where charge carrier concentration equals the carrier concentration of neutral p or n-region.

(iii) Contact electrodes to collect the charge-carriers generated in the p-n junction device and connect it to external circuit.

# 3 p-n junction under applied bias

## Forward bias and Reversed bias

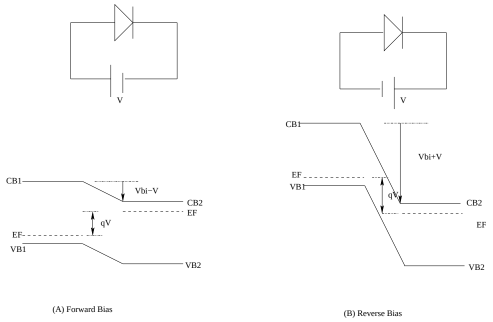
When an external bias (V) is connected across the junction such that the p-side is connected to positive terminal of the battery and n-side to negative terminal, the junction is said to be forward bias. Forward bias lowers the junction potential to ($\mathrm{V_{bi}} - \mathrm{V}$) and thereby increases the probability of carrier diffusion across the junction, by a factor of $\mathrm{e}^{\frac{\mathrm{qV}}{\mathrm{KT}}}$ (known as Boltzman factor). Therefore, diffusion current under FB is given as its equilibrium value multiplied by the Boltzman factor. Under reverse bias condition, p-side is connected to lower potential than n-side, the barrier height at the junction is increased to ($\mathrm{V_{bi}} + \mathrm{V}$) thereby decreasing the diffusion current.

In both the cases, drift current ($\mathrm{I_o}$) being independent of the voltage remains the same. It mainly depends on the number of charge carriers crossing the junction, which depends on the elctron-hole pairs being produced in the depletion region. Therefore, it is also known as the generation current or reverse saturation current as it is the only current flowing through the diode when it is reversed biased.

Therefore under an applied bias, total current flowing through the junction is given as

$$I = I_o(e^{\frac{qV}{KT}} - 1) \tag{5}$$

Where, $\mathrm{V} = +\mathrm{V}$ for forward bias and $\mathrm{V} = -\mathrm{V}$ for reverse bias. The above equation is known as Schokley equation. It is the fundamental equation for micro-electronic devices.



(A) Forward Bias

(B) Reverse Bias

Figure 2: Figure depicting energy-band diagram of a diode under (A) Forward bias (B) Reverse bias

In both the biasing, the drop (V) in the built-in voltage causes an equivalent shift (qV) in fermi level positions from its equilibrium position in absence of biasing on both sides of junction with respect to each other as shown in fig 2.

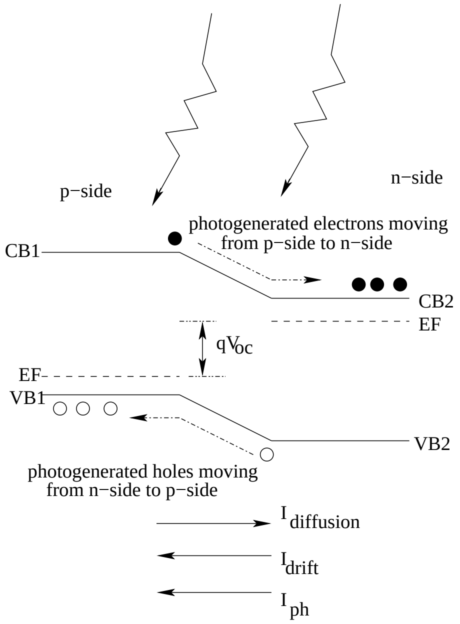


Figure 3: Figure showing energy-band diagram of a p-n junction under illumination.

# 4 p-n junction under illumination

When a p-n junction is exposed to light having energy greater than the band-gap of the semiconductor in use, electron-hole pairs are produced both at the junction (i.e. in depletion region) as well as in the quasi-neutral region. However, due to presence of large concentration of majority (mobile) carriers in neutral region, rapid recombinations take place. Thus, the total number of photogenerated charge carriers are effectively those generated in depletion region only. In the space-charge region on p-side of the junction, holes (from generated electron-hole pairs) get recaptured by the acceptor ions present there and the elctrons are swept towards the n-side. Similarly the electrons generated on the n-side depletion region, get recaptured by the donor ions there and holes cross the junction to reach p-side. This decreases the number of donor and acceptor ions in the space-charge region and hence decreases the total potential across the junction to ($V_{bi} - V_l$), just like in FB case.

$$I = I_o(e^{\frac{qV_l}{KT}} - 1) - I_{ph} \qquad (6)$$

'$V_l$' in above equation is the drop in the built-in volatage across the junction on exposure to light, the maximum value it can take is called open-circuit voltage or $V_{oc}$ beyond which the diffusion current will start dominating. The maximum drop in the built-in potential under illumination or the value of $V_{oc}$ is decided by the difference in the position of Fermi-levels of the two materials before junction formation [12] i.e. $E_{fn} - E_{fp}$ (see fig 3), This is the limiting value of open-circuit voltage as fermi-levels can not be shifted beyond these positions in absence of ant external voltage bias. $I_{ph}$ being a negative current i.e. in opposite direction to the convential forward-bias current ($I_{diff}$), it will shift down the I-V curve to IV$^{th}$ quadrant, as shown in fig 4. $V_{oc}$ can also be seen as the maximum voltage that can be generated across the solar cell terminals i.e. when it is not connected to external load (open-circuit condition). In this case, the charge carriers crossing the junction will reach the respective electrodes and get accumulated there resulting in a potential drop (positive on p-side and negative on n-side). Hence, it is called the open-circuit

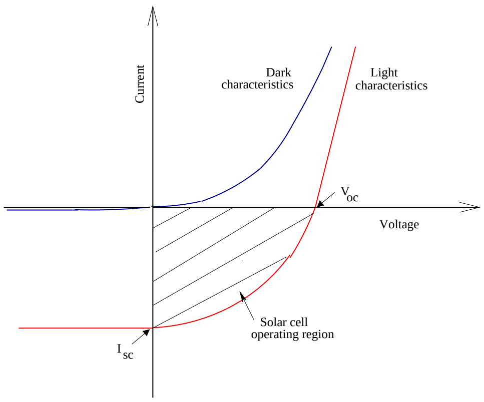


Figure 4: I-V characteristics of a solar cell.

voltage ($V_{oc}$). It can be derived from eqn (5) as

$$V_{oc} = \frac{KT}{q} ln\left(\frac{I_{ph}}{I_o} + 1\right) \qquad (7)$$

Similarly, the maximum light-current that can be produced on illumination is called short-circuit current or $I_{sc}$. As the name suggests, it is the maximum current which will flow through the external circuit when p- and n-sides of the junction are short-circuited[3]. Since this current depends on the area of solar cell, J-V curve is generally reported instead of I-V curve, where J is known as the current density and is given as (J = I/area). Such a device which converts solar energy into electrical energy is known as phtovoltaic device or more simply a solar cell.

A p-n junction works as a solar cell in the forth quadrant, where power is negative ($P = V \times (-I)$) i.e. instead of dissipating power, the device is generating the power. But, does that mean all the p-n junctions will work as solar cells under illumination? The answer is 'NO'. For better understanding, we have to look into the details of solar cell physics from a material enggineering point of view. The ability of a p-n diode to work as a solar cell depends on a large number of factors such as

(i) Material's ability to generate carriers upon excitation.

(ii) Length of p- and n-layers and other parameters influencing the charge-carrier transport.

(iii) Selection of electrodes for collection of generated charge carriers.

Before discussing the above points into details, please note that uptil now, we have considered both p and n-layers to be absorbers as in case of crystalline-Silicon solar cells (c-Si). We will now shift our discussion from c-Si type solar cells to thin film solar cells for simplicity. These solar cells have only one absorber layer (p-type or n-type) and the other is just a complimentory layer (transparent to sunlight) forming the junction with the absorber layer.[4] Lets discuss the itemized aspects of device physics one by one.

---

[3] In this case, the potential barrier remains unchanged as the total work done is zero i.e. $qV = 0$

[4] In this article, we are basically dealing with inorganic solar cells which rely on the junction property for charge carrier generation and separation. There are a large number of other solar cells also like oraganic solar cells, dye-sensitized solar cells etc. having different working principle than the inorganic ones. However, the basic material engineering and selection criteria remains the same.

# 5 Material selection for juction formation

For a p-n junction to operate as solar cell, the two layers forming the junction must fulfil the following criteria.

### 5.0.1 Optimum band-gap

When a material is exposed to sun, it will absorb light having energy (E) greater than or equal to its band-gap ($E_g$) while being transparent to other wavelengths. Energy equal to the band-gap is used by the charge carriers to move from valance band to conduction band and rest of the energy ($E$-$E_g$) is lost/dissipated as heat. Therefore, for applications in solar cells, the material should have an optimum band-gap to absorb maximum portion of solar spectrum with minimum heat dissipation. As we all know that the amount of sunlight reaching the earth's surface depends on the latitude and longitude of the place, for testing purposes the solar spectrum is standardized as AM 1.5,[5] major portion of the spectrum comprises of VIS-IR region. Theoretically, the band-gap for maximum solar cell output is estimated to be $\approx 1.5$ eV for AM 1.5 spectrum [11].

## 5.1 Minimum lattice-mismatch between layers

The layers forming the junctons should have minimum lattice mismatch in their crystal structure so as to avoid the formation of any defect at the interface. These defects hinder the movement of charge-carriers and might result in trapping or recombination and hence degrade the solar cell performance.

## 5.2 Band discontinuity between lattice-matched p and n-layers

Under the section 'p-n junctions', we have already discussed the band-bending considering both the p- and n-layers are of the same material eg.(p-Si/n-Si) and therefore have their conducton band and valance band at same level in energy-diagram and the only difference is in the position of fermi-levels due to different doping type in the two layers. Such p-n junctions are called homo-junctions (as shown in fig 1). However, when a juncton is formed using p- and n-layers of different materials eg.(p-SnS/n-ZnO), having different band-gap and electron effinities,[6] a discontinuity is generated in their band-structure as Fermi levels line up at equilibrium [13]. The discontinuity in conduction band of the two layers is known as the conduction band offset ($E_c$) and is given as the difference in the electron effinities of the two layers. Sim-

---

[5] 'AM' stands for air mass and is given as the secant of angle that the sun makes with the zenith, for AM 1.5 spectrum, this angle is 45°.

[6] electron effinity of a semiconductor is defined as the energy difference between it's conduction band edge and the vacuum level

ilarly, the discontinuity in valance band is known as valance band offset ($\mathrm{E_v}$) (see fig 5). These junctions are called heterojunctions. If the absorber layer is of p-type, then the $\mathrm{E_c}$ (with respect to n-type junction layer) plays a major role [14] while for a solar cell having n-type absorber layer, $\mathrm{E_v}$ (with respect to p-type junction layer) decides the junction quality [15]. If ($\mathrm{E_c}$) or ($\mathrm{E_v}$) is large (greater than 0.3 V), it will effect the band-bending and would lead to the formation of spike or cliff structures at the junction as shown in fig 5. The shape of the energy bands can be easily found out by solving Poisson's equation with boundary conditions of continuous electric flux density at the junction (i.e. $\epsilon_{\mathrm{p}}\mathrm{E_p} = \epsilon_{\mathrm{n}}\mathrm{E_n}$), where '$\epsilon_{\mathrm{p}}$' and '$\epsilon_{\mathrm{n}}$' are the dielectric constants of the p- and n- layers. Formation of spike or cliff at the junction degrade the junction quality by trapping the carriers instead of promoting the charge-separation.

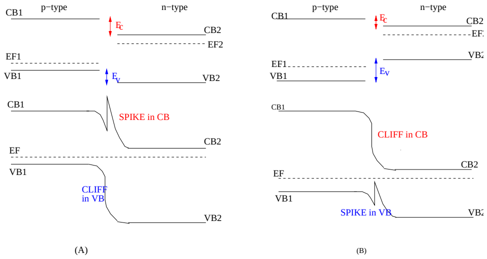


Figure 5: Figure showing the spike and cliff formation in a heterojunction.

# 6 Length of different regions of p-n junction

After separation at the junction, the generated charge carriers have to travel different regions (depletion region and quasi-neutral regions) to reach the external circuit. The length of each region effects the carrier transport and in turn effects the overall solar cell efficiency. Some of the factors effecting the thickness of these leyers are

## 6.1 Absorption coefficient

The absorption of light by a material does not only depend on the band-gap and energy of incident photon but also depends on the ability of a material to absorb and is given by the absorption coefficient ($\alpha$).

Absorption coefficient (unit $\mathrm{cm}^{-1}$) determines the length traversed by the light (of a perticular wavelength, $\lambda$) in a material before getting absorbed. It can be easily derived from Beer Lambert's law [16]

$$I_z = I_o e^{-\alpha z} \tag{8}$$

where, '$\mathrm{I}'_{\mathrm{z}}$ is the light intensity after travelling distance 'z' in the material. '$\mathrm{I}'_{\mathrm{o}}$ is the incident light intensity. Penetration depth is given as ($\frac{1}{\alpha}$ cm).

Absorption coefficient of the material should be large enough to absorb maximum sunlight incident on it so that the minimum amount of material is required (it will minimize the material's cost). Therefore, the length of absorber layer should be optimized according to its absorption coefficient such that considerable amount of light reach the junction before getting completely absorbed by the material since effective number of charge carriers are generated in depletion region (near the junction) only as discussed in sections above.

### 6.2 Diffusion length of carriers

Another major factor deciding the lengths of quasi-neutral and space-charge regions in solar cell is the diffusion-length of the charge carriers. Diffusion length by principle, is the distance travelled by a charge carrier between subsequent collisions or before getting recombine. It is propotional to the product of mobility $\mu$ and life-time $\tau$ of charge-carriers i.e. $\sqrt{\mu\tau}$ [11].

The electron-hole pair produced inside space-charge region should cross the junction before getting recombine i.e. electrons have to reach n-side and holes p-side without recombination. Since these carriers are minority carriers in the region where they are produced, the depletion layer width should be less than the diffusion length of the minority carriers produced on that side of the junction. This implies that the optimum depletion layer width on p-side of the junction would be decided by the diffusion length of the electrons in p-type material and similarly the optimum width of depletion region on n-side would depend on the diffusion length of holes in n-type material.

Upon crossing the junction, holes and electrons have to travel p-region and n-region respectively (where they are majority carriers) to reach the electrodes. However, during their journey across the quasi-neutral region, these charge carriers can encounter a large number of collisions and recombinations[7] [11] with mobile carriers already present there implying that even in these regions, they have a finite lifetime associated with them. Thus, conductivity of the material becomes another decisive factor as it will effect the mobility of the respective charge-carrier and hence the majority carrier's diffusion length. To sum up, the optimum length of the depletion region is decided by the minority carrier's diffusion length while that of quasi-neutral regions by majority carrier's diffusion length, making conductivity (or doping)[8] a major decisive factor.

## 7 Contact electrodes

To collect the photo-generated carriers and to feed them to external circuit, metal contacts are made on both sides of semiconductor layers. Contact on the side from where light enters into the cell is known as front contact and the contact on the opposite side is known as the back contact.

Depending on the work function of the metal and the semiconductor, two types of junctions/contacts are formed between them

(i) Ohmic contact

---

[7] There are mailny four different types of recombinations in a semiconductor namely band to band recombination, R-G cenetrs recombination, excitonic recombination and Augur recombination.

[8] conductivity ($\sigma$) of a material is related to carrier concentration (N) by relation $\sigma = \mathrm{qN}\mu$

(ii) Rectifying contact

Ohmic contacts follow the linear V = IR relation and are best for collection of charge carriers. While in case of rectifying contacts, a junction is formed at the metal-semiconductor interface (i.e. ohm's law does not apply anymore) leading to loss of carriers in overcoming the junction barrier. For a p-type absorber layer, metal work function should be more as compared to the p-type semiconductor material used to obtain an ohmic contact while it should be less for making an ohmic contact with n-type semiconductor as shown in fig 6. Conventionlly, metals like Al, In, Cu, Ag and Au are used for making contacts. Front contact is generally made in form of very thin metal strips to allow maximum light to enter the cell while thin metal sheets are grown as back contacts to prevent light from escaping the cell. Now-a-days special type of wide-band semiconductors having low sheet resistance ($< 10\ \Omega/\square$) and high transmittance are gaining populaity. These semiconductors are known as Transparent conducting oxides or TCO's. Some of the widely used TCO's are ZnO, Al:ZnO, FTO, ITO etc. While for back contacts, metals remain the best choice for their good reflectivity.

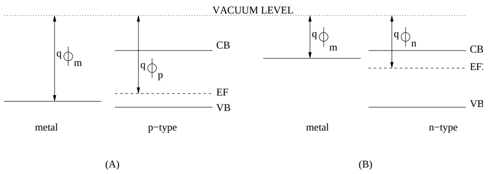


Figure 6: A representative of energy-band allignment suitable for ohmic-junction formation between (A) metal and p-type semiconductor (B) metal and n-type semiconductor.

## 8 Ideal solar cell

Concluding from above discussion, an ideal solar cell should only have depletion region with no (or zero length) neutral region and electrodes placed at the junction boundary at $\mathrm{x_n}$ and $\mathrm{x_p}$ as shown in fig 7. If diffusion length of holes and electrons in n- and p-region be $\mathrm{L_p}$ and $\mathrm{L_n}$ respectively. Then

$$x_n \leq L_p$$
$$x_p \leq L_n$$

or

$$(x_n + x_p) \leq (L_p + L_n) \tag{9}$$

Now, if penetration length is $\left(\frac{1}{\alpha}\right)$, then following condition has to be satisfied for all the depletion region to be illuminated

$$\left(\frac{1}{\alpha}\right) \geq (L_p + L_n) \geq (x_n + x_p) \tag{10}$$

However, to have a zero neutral region is not possible but ts length can be minimized by careful material engineering.

## 9 Characteristics of a solar cell

Once a p-n junction is made, it is tested to check whether it will function as a solar cell

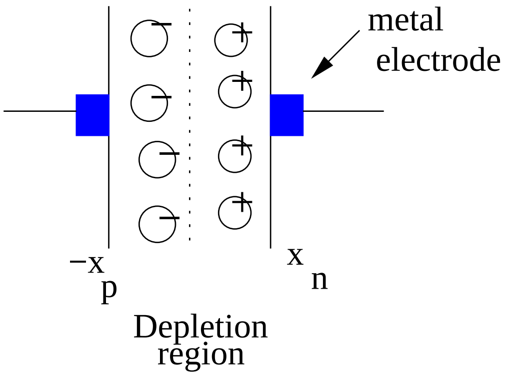


Figure 7: An ideal solar cell.

on exposure of light or not? for this purpose, I-V or (J-V) characteristics of the junction are recorded for a moving bias of (-V to +V) under AM 1.5 spectra.

Negative bias (-V to 0) is applied to ensure the quality of the junction i.e. it should not break down under the reverse bias conditions. Also under reverse bias conditions, charge carriers are “sucked out” of the device resulting in a current, flowing opposite to the direction of conventional diode current ($I_{diff}$ during forward bias). Therefore, reverse bias ensures that the current flowing through the cell is purely due to generation of charge carriers in the device.

A p-n junction will continue to function as a solar cell (i.e. I-V in IV$^{th}$ quadrant) until the applied bias reaches $V_{oc}$, beyond which the diode gets forward bias or in other words the barrier at junction get so reduced that the forward current starts dominating.

To check the quality of the fabricated solar cells, following quantities are defined

## Fill Factor (FF)

To draw maximum power from a device, the area of the rectangle enclosed[9] by the J-V curve should be maximised or in other words the squareness of the I-V curve should be increased. The point at which the rectangle touches the I-V curve (fig 8)is known as the maximum power-point and the corresponding output voltage and current are labelled as $V_m$ and $I_m$. The parameters used to define/quantize the squareness of a solar-cell I-V curve is called Fill factor (FF) and is given as

$$FF = \frac{areaoflargestrectangle}{totalareaunderI - Vcurve} = \frac{V_m I_m}{V_{oc} I_{sc}} (11)$$

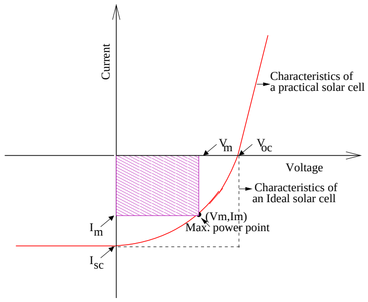


Figure 8: I-V characteristics of a practical and an ideal solar cell.

## Power Efficiency ($\eta$)

The power efficiency of a solar cell is simply the ratio of maximum power that can be

---

[9] The power generated/dissipated by any device is the area of rectangle completely enclosed by the I-V curve.

extracted from it to the total optical power incident.

$$\eta(\%) = \frac{V_m I_m}{P_{in}} = \frac{FF \times V_{oc} I_{sc}}{P_{in}} \times 100(\%) \quad (12)$$

A p-n junction with good Fill factor will ensure a higher efficiency and hence would be a good solar cell.

Power efficiency ($\eta$) as discussed above is basically concerned with the total output power from a solar cell which along with material's ability to convert photons to electron-hole pairs include a lot of other factors such as design of solar cells such as shape and size of different solar cell layers, biasing of solar cell etc.

To characterize the intrinsic nature of solar cells (i.e. number of electrons produced per incident photon) and to separate it from rest of the factors, another term has been defined by the researchers known as the Quantum efficiency. Two types of efficiencies have been defined

### Internal Quantum efficiency (IQE)

It tells about the junction quality, separating it from the material's ability of absorbing the photon. It is defined as the ratio of number of electrons collected to the totatl number of photons absorbed.

### External Quantum efficiency (EQE)

It takes into account the solar cell design along with the property of the material. It is given as the ratio of the number of electrons collected to the total number of photons incident.

## 10 Modeling of a solar cell

To model an ideal solar cell, a current source is used connected in parallel with a diode [11] as shown in fig 9. The current source is used as the photogenerated current remains constant for a given illumination level irrespective of the load connected across the solar cell while the diode element indicates the directional property of the current (i.e. from p to n-side). This is the main difference between a solar cell and a battery (apart from battery's limited lifetime) that the battery provides a constant emf across its terminals while varying the current depending on the circuital elements. In solar cells, the current depends on the number of photogenerated charge carriers while the voltage is determined mainly by the

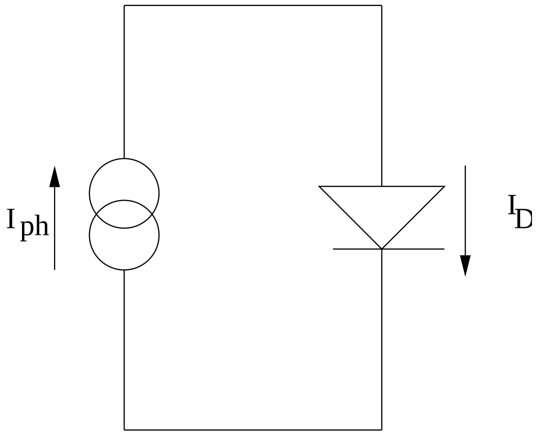


Figure 9: Equivalent model of an ideal solar cell.

load connected[10]. However, during their journey from junction to electrodes, these charge carriers have to travel various regions having different resistances and can face recombination at grain boundaries, bulk and at interfaces between layers etc. All thses losses are modelled as series and shunt resistances connected across the solar cell and indicate a decrease in its efficiency by dissipating power [11]. Series resistance represents the difficulty of hole and electron motion due to bulk resistances of semiconductors and metals used and the contact resistances between the two. Shunt resistance arises from current leakage through the device (around the edges and between the contacts), it also includes the effect of foreign impurities and crystal defects induced while fabrication. The resultant equivalent circuital model of solar cell is shown in fig 10. This is a standard one-diode model. Two or more diodes are also used in some cases to account for increased recombinations due to grain boundaries etc. However, in this article we will limit our discussion to one-diode model only. In the figure ‘$I'_{ph}$ is the photo-current, ‘$V'_D$ and ‘$I'_D$ are the diode voltage and current given, ‘$V'$ and ‘$I'$ are total voltage and current through the circuit.

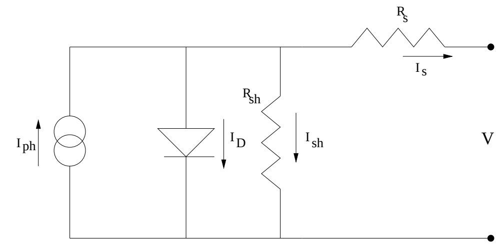


Figure 10: One-diode equivalent model of a solar cell.

Applying KCL gives,

$$I_{ph} - I_D - I_{sh} - I = 0 \tag{13}$$

Using eqn (5)

$$I = I_{ph} - I_o(e^{\frac{qV_D}{KT}} - 1) - I_{sh} \tag{14}$$

Also $\mathrm{V_D = V + IR_s}$ and

$$I_{sh} = \frac{V + IR_s}{R_s h} \tag{15}$$

substituting in eqn (14) gives

$$I = I_{ph} - I_o(e^{\frac{q(V+IR_s)}{KT}} - 1) - \frac{V + IR_s}{R_s h} \tag{16}$$

As seen in previous sections, the photogenerated charge carriers undergo recombination in the depletion region as well as the neutral regions. To include this effect, the eqn 5 is modified as

$$I = I_o(e^{\frac{qV}{nKT}} - 1) \tag{17}$$

Where, ‘n’ is the ideality factor, it determines the departure of a p-n junction characteristcs from the ideal diode characteristics. Generally, n varies from 1 to 2 for a practical solar cell with n = 1 for an ihere deal junction solar cell. However, there have been reports of p-n junction, with n = 2/3 (due to Auger recombination taking place) and n > 2 (due to recombinations at the grain boundaries and tunneling effects). Therefore, eqn 18 becomes

$$I = I_{ph} - I_o(e^{\frac{q(V+IR_s)}{nKT}} - 1) - \frac{V + IR_s}{R_s h} \tag{18}$$

This is the characteristic equation of a practical solar cell.

---

[10] Assuming that all the generated carriers are collected at the electrodes.

### 10.1 Effect of $\mathrm{R_s}$ and $\mathrm{R_{sh}}$ on I-V characteristics of a solar cell

$\mathrm{R_s}$ being a series connection between the exetrnal circuit and the cell, controls the value of maximum current flowing through the device (under the assumption $\mathrm{R_{sh}} \gg \mathrm{R}_s$). For a very small value of $\mathrm{R_s}$, $\mathrm{I} \approx \mathrm{I_{sc}}$. While controlling $\mathrm{I_{sc}}$, it does not affect the open-circuit voltage of the cell as at this point, the photo-generated current and hence the current through $\mathrm{R_s}$ is zero. As seen from the equivalent circuit in fig 10, $\mathrm{R_{sh}}$ provides an alternate path for current and thus affects the voltage across the solar cell terminals. This voltage will be equal to $\mathrm{V_{oc}}$ for $\mathrm{R_{sh}} \approx \infty$. Generally, $\mathrm{R_{sh}} \gg \mathrm{R}_s$ therefore, it does not affect the total current (I) as already discussed.The values of $\mathrm{R_{sh}}$ and $\mathrm{R_s}$ can be estimated using the I-V curve as

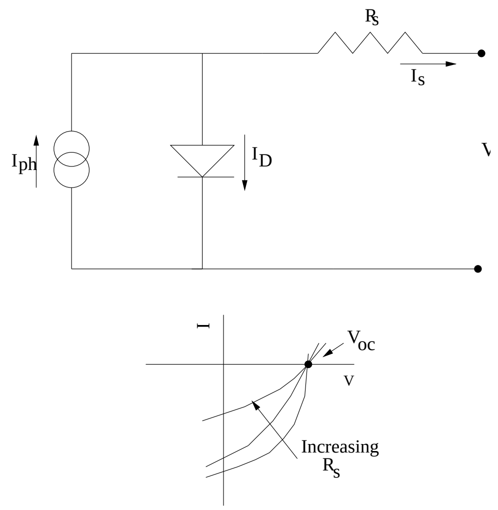


Figure 11: Effect of $\mathrm{R_s}$ on solar cell characteristics.

$$R_s = \left(\frac{dV}{dI}\right)_{V=V_{oc}} \tag{19}$$

$$R_{sh} = \left(\frac{dV}{dI}\right)_{V=I_{sc}} \tag{20}$$

Values of both the resistances greatly affects the sharpness of I-V curve. $\mathrm{R_s}$ alters the short-circuit current, which will decrease as the value of $\mathrm{R_s}$ increases, keeping open-circuit voltage constant fig 11. On the other hand, as the value of shunt resistance decreases, the voltage across the diode will decrease and hence the $\mathrm{V_{oc}}$ will decrease as shown in fig 12. The I-V curve will take the shape of a straight line intersecting the two axis at $\mathrm{V_{oc}}$ and $\mathrm{I_{sc}}$ if the values of the two resistances diverge greatly from their ideal values.

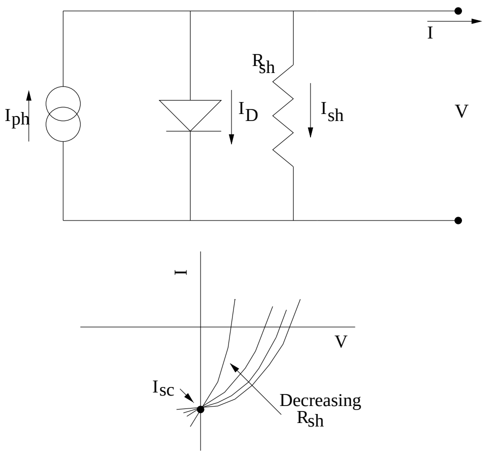


Figure 12: Effect of $\mathrm{R_{sh}}$ on solar cell characteristics.

## Acknowledgments

I would like to acknowledge Dr. P.Arun, discussions with him greatly helped in shaping up the article to its present form.

## References

[1] Fu.Qiang Radiation (solar, 1859).

[2] J.H.Scofield The Solar Spectrum, (PHYS-068, 2009).

[3] A.E.Becquerel (1839). Compt. Rend. Acad. Sci. 9, 561

[4] D.M.Chapin, C.S.Fuller, G.L.Pearson (1954). J. Appl. Phys. 25, 676

[5] B. J.Stanbery (2003). Critical Rev. Solid State Mater. Sci. 27, 73

[6] D.A.Jenny, R.H.Bube (1954). Phy. Rev. 96, 1190

[7] P.K.Bhat, S.R.Das, D.K.Pandya, K.L.Chopra (1979). Solar Energ. Mater. 1, 215

[8] C.Piliego, L.Protesescu, S.Z.Bisri, M.V.Kovalenkobc, M.A.Loi (2013). Energ. Env. Sci. 6, 3054

[9] M.Ristov, G.Sinadinovski, M.Mitreski, M.Ristova (2001). Sol. Energ. Mater. Sol. Cells. 69, 17

[10] G.Wagner, R.Kaden, V.Lazenka, K.Bente (2011). Phy. Status Solid 9, 2150

[11] S.M.Sze., K.K.Ng Physics of Semiconductor Devices. (3rd Ed. John Wiley and Sons, Inc., U.S.A, 2007)

[12] J.C.Blakesley, D.Neher (2011). Phys. Rev. B 84, 075210

[13] B.G.Streetman Solid State Electronic Devices. (4th Ed. Prentice Hall, India, 1999)

[14] P.Sinsermsuksakul, K.Hartman, S.B.Kim, J.Heo, L.Sun, H.H.Park, R.Chakraborty, T.Buonassisi, R.G.Gordon (2013). Appl. Phys. Lett. 102, 053901

[15] K.A.Nagamatsu, S.Avasthi, J.Jhaveri, J.C.Sturm (2014). IEEE J. Photovolt. 4, 260

[16] H.Fujiwara Spectroscopic Ellipsometery- Principles and Applications. (John Wiley, New York, 2007)

# Visualizing Electromagnetic Fields Using Gnuplot
## Part 2
## Field from Accelerating Charged Particle

Somnath Datta
Professor of Physics (Retired), National Council of Educational Research and Training,
New Delhi-110016
*Res:* 656, “Snehalata”, 13th Main, 4th Stage, T K Layout, Mysore 570009, India
datta.som@gmail.com; http://sites.google.com/site/physicsforpleasure

(Submitted 09-08-2016)

**Abstract**

This article is a continuation of our efforts to demonstrate how to plot the **E** field from a time varying source, using Gnuplot. In the present case the source has been taken to be a charged particle in accelerated motion, moving with relativistic speeds. Three specific examples of acceleration have been taken, namely, (a) rectilinear acceleration, (b) rectilinear deceleration, (b) centripetal acceleration. The formulas used for the plots have been taken from a 1972 paper by Roger Y. Tsien. However, these formulas have been re-derived in details to make the article reader friendly and pedagogical. We have highlighted the lessons that the reader can derive from these plots: (1) Clarification of concepts related to retarded time, and propagation of the **E** field from the source at the retarded time $t_r$ to the observer at the present time $t$. (2) Confirmation of the “Purcell condition” $\tan\phi_\beta = \gamma \tan\phi_0$ in which $\phi_\beta$ and $\phi_0$ are the angles that a given field line makes with the direction of motion, after and before a period of acceleration. (3) Insight into the phenomenon of synchrotron radiation, which is also an important component of the pulsed radiations from pulsars and the Crab nebula. The commands used for writing the relevant equations, and plotting them, have been copied from the Console into an Appendix, so that an interested reader can replicate all the plots on his personal computer.

# 1 Introduction

This article is a follow-up of our earlier article published in this journal[1], in which we had shown application of Gnuplot for plotting Electromagnetic Fields originating from an oscillating Electric Dipole. In this article we shall concentrate on another important source of Electromagnetic fields, namely, a charged particle in accelerated motion. We shall take only three examples of this acceleration: (a) Rectilinear acceleration, (b) Rectilinear deceleration (c) Centripetal acceleration. In our examples, the particle under our consideration is moving with relativistic velocity $c\boldsymbol{\beta}$, where $c$ is the speed of light.

Formulas of the Electric Field $\mathbf{E}$ and the magnetic field $\mathbf{B}$, originating from moving charges are familiar[2, 3]. However, plots of these fields are not so common. Purcell[4], in presenting a beginner's course in Electricity and Magnetism, has drawn $\mathbf{E}$ field lines from a charged particle, originally at rest, and then picking up a relativistic velocity under an extremely large acceleration existing for a very short time. Roger Y. Tsien[5], in his 1972 paper has shown actual plots of field lines for a several interesting cases. These plots were drawn using "an IBM 360/65 computer programmed in Fortran IV and a ... drum plotter". Such heavy tools are things of the past. Now any interested student can make all these plots on his desk using his PC and free software, like Gnuplot, and get wonderful, almost unbelievable results, in all colours.

Our objective in this article is two-fold. First, we want to demonstrate the power of Gnuplot in plotting such difficult fields with so much ease. The plot commands, copied from the Console to Appendix B, can encourage the reader to get more practice which he can use profitably in his study of Classical Electrodynamics. Secondly, we would like to illuminate some difficult concepts surrounding solution Maxwell's equation, particularly *retarded time* $t_r$, propagation of the field from the source at the retarded time $t_r$ to the field point at the present time $t$, covering the *charge-to-field* distance $R$, abbreviated as *CtF* distance, with the speed of light $c$, and illustrate how, with the plotting of the examples cited by us, he clarifies, and sharpens his understanding of these concepts.

Gnuplot can plot mathematical functions, even the most difficult ones, with ease, if we write the function clearly in the command line. There are two excellent guide books[6, 7] which the interested reader should keep as his constant companion for quick reference.

In our case the function is a parametric function of the form $x = f(t),\ y = g(t)$ representing the $\mathbf{E}$ field line on the $XY$ plane. The parameter $t$ in our case is the CtF distance $R$ having range $[0 : R_{\max}]$. In the final example we have switched to retarded time $t'$ as the parameter which is related to $R$ as $t' = -R/c$, and having the inverse range $[-R_{\max}/c : 0]$.

We have written a sample of commands from the actual command line, in Appendix B, so that the reader can replicate *all* the plots presented in this article.

The mathematics of plotting is somewhat difficult. It requires not only a crisp understanding of the $(\mathbf{E}, \mathbf{B})$ fields, their relation to the retarded time, but also a clear understanding of transformation of coordinates leading to the formulation of the differential equation for the field line, and its solution. To facilitate the reading of this article without too much abstraction, we have shifted most of this mathematics to Appendix A.

Sec.2 begins with determination of the $(\mathbf{E}, \mathbf{B})$ field from a point charge $q$ in arbitrary acceleration. Our task in this article is to plot only the $\mathbf{E}$ field, disregarding the $\mathbf{B}$ field completely. The reason: (1) Whereas the motion of the charged particle $q$ and the $\mathbf{E}$ field are confined to the same plane, taken as the $XY$ plane, the $\mathbf{B}$ field is perpendicular to this plane. Hence we cannot show the geometry of the motion and the field arising from it on the same diagram. (2) The $\mathbf{E}$ field and the $\mathbf{B}$ field are intimately connected by the equation $c\mathbf{B} = \mathbf{n} \times \mathbf{E}$, where $\mathbf{n}$ is a unit vector in the direction of the CtF vector, lying on the $XY$ plane. Hence $\mathbf{E}$ carries all the information about $\mathbf{B}$.

The next two sections, i.e., Secs. 3, 4 have been devoted to obtaining the differential equation of the field line and the boundary condition. These equations have been given by Tsien. However, we have spent extra efforts to re-derive these results in detailed steps which may be easier to follow.

Sec. 5 presents the case of the charge $q$ moving under a constant force $\mathbf{F}$, acting in the positive x-direction. We have considered four examples:

(1) *Accelerated* motion on the *positive* x-axis, $q$ moving *away* from the origin towards $+\infty$:

Ex.1(a) $\mathbf{F}$ continues for ever;

Ex.1(b) $\mathbf{F}$ starts and stops

(2) *Decelerated* motion on the *positive* x-axis, $q$ moving *towards* the origin from $+\infty$:

Ex.2(a) $\mathbf{F}$ continues for ever;

Ex.2(b) $\mathbf{F}$ starts and stops

By measuring the angles made by the field lines in Ex.1(b) the reader can confirm the Purcell condition.

Each set of plottings teaches some lessons about how the field travels from the charge to the observer covering the CtF distance. These have been summarized by us.

Finally we arrive at the conclusions: (a) the field lines for the deceleration case are very similar to those for the acceleration case; (b) they curve down from straight lines, as if acted on by induced gravity, opposite to the direction of acceleration.

Sec. 6 presents the field from a charged particle moving in a circle with relativistic and ultra-relativistic speeds, and offers some useful insight into *Synchrotron Radiation*, which is an important component of the pulsed radiations coming from from pulsars and the Crab nebula.

# 2 Motion with Arbitrary Acceleration

## 2.1 General Formula

We shall write the $(\mathbf{E}, \mathbf{B})$ field of a charged particle in arbitrary motion as a first step towards making their plots. In Fig. 1(a) we have explained the configuration, with reference to the coordinate system $XYZ$ of which O is the origin.

$\Gamma$ is the trajectory of the particle, given by the parametric representation $\tilde{\mathbf{r}} = \tilde{\mathbf{r}}(t')$, implying that each point B on $\Gamma$ has a radius vector $\tilde{\mathbf{r}} = \xi(t')\mathbf{i} + \zeta(t')\mathbf{j} + \chi(t')\mathbf{k}$ in which $\xi, \zeta, \chi$ are known functions of time $t'$. We shall soon identify $t'$ with *retarded time.* A is the location of the particle at the *present* time $t$ (observer's time), and P is the location of the observer (where the $(\mathbf{E}, \mathbf{B})$ field is measured).

The $\mathbf{E}(\mathbf{r}, t)$ field originates from the retarded location B of the particle at the retarded time $t'$, reaches P at the present time $t$, travelling with the speed $c$, covering a distance $R(t')$ which is the length of the radius vector $\mathbf{R}(t')$ from B to P. We shall call the vector $\mathbf{R}$ "*charge to field vector*", in brief "CtF" vector. Therefore,

$$
\begin{aligned}
&R(t') = c(t - t');\\
&\Rightarrow \quad t' = t - \frac{R(t')}{c} = t - \frac{|\mathbf{r} - \tilde{\mathbf{r}}(t')|}{c}.
\end{aligned} \tag{1}
$$

When the particle is at B, it has velocity $\mathbf{v}(t') = c\boldsymbol{\beta}(t')$ and acceleration $\boldsymbol{a}(t') = c\dot{\boldsymbol{\beta}}(t') = c\frac{d\boldsymbol{\beta}}{dt'}$. Even though $\boldsymbol{\beta}$ is a dimensionless vector, $\dot{\boldsymbol{\beta}}$ is not. It has the dimension $\mathrm{T}^{-1}$.

Let $\mathbf{n}(t')$ the unit vector in the direction of the CtF vector $\mathbf{R}(t')$. Hence the relations.

$$
\begin{aligned}
\mathbf{R}(t') &= \mathbf{r} - \tilde{\mathbf{r}}(t') = R(t')\,\mathbf{n}(t'); && \text{(a)}\\
\mathbf{v}(t') &= \tfrac{d\tilde{\mathbf{r}}(t')}{dt'} = c\boldsymbol{\beta}(t'); && \text{(b)}\\
\boldsymbol{a}(t') &= \tfrac{d\mathbf{v}(t')}{dt'} = c\dot{\boldsymbol{\beta}}(t'). && \text{(c)}
\end{aligned} \tag{2}
$$

The $(\mathbf{E}, \mathbf{B})$ field of the particle at the observer P at the time $t$ is now given by the following formulas[3, 2].

$$
\begin{aligned}
\mathbf{E}(\mathbf{r}, t) &= \frac{q}{4\pi\varepsilon_0}\left[\left(\frac{(1-\beta^2)(\mathbf{n}-\boldsymbol{\beta})}{\kappa^3 R^2}\right)\right.\\
&\qquad \left. + \left(\frac{\mathbf{n} \times \{(\mathbf{n}-\boldsymbol{\beta}) \times \dot{\boldsymbol{\beta}}\}}{c\kappa^3 R}\right)\right]_{t'=t_r}.\\
c\mathbf{B}(\mathbf{r}, t) &= \mathbf{n}(t_r) \times \mathbf{E}(\mathbf{r}, t).\\
\kappa(t') &= 1 - \boldsymbol{\beta}(t') \cdot \mathbf{n}(t').
\end{aligned} \tag{3}
$$

## 2.2 Special Case: Motion confined to a plane

We shall specialize the $(\mathbf{E}, \mathbf{B})$ field for the special case in which the path $\Gamma$, as well as the observer P, exist on a plane, which we shall take as the $XY$ plane. The equation of the path is

$$
\tilde{\mathbf{r}} = \tilde{\mathbf{r}}(t') \quad \Rightarrow \quad \tilde{x} = \xi(t'),\ \tilde{y} = \zeta(t'). \tag{4}
$$

The field point P has Cartesian coordinates $(x, y)$ with respect to the *fixed* origin O. However, for the purpose of drawing the field lines, we shall find it convenient to propose a set of alternative, curvilinear coordinates $(R, \Phi)$ with respect to the *moving* origin B$(t')$. Here $R$ is the length of the CtF vector $\mathbf{R}$ and $\Phi$ is the angle this radius vector makes with the $X$ axis.

Let $(\mathbf{e}_R, \mathbf{e}_\Phi)$ be the unit vectors associated with the new coordinates $(R, \Phi)$. As per convention, $\mathbf{e}_R$ and $\mathbf{e}_\Phi$ are the directions in which the respective coordinates are increasing. In particular, $\mathbf{e}_R$ is *identical* with the unit vector $\mathbf{n}$ introduced earlier and $\mathbf{e}_\Phi$ is perpendicular to $\mathbf{e}_R$.

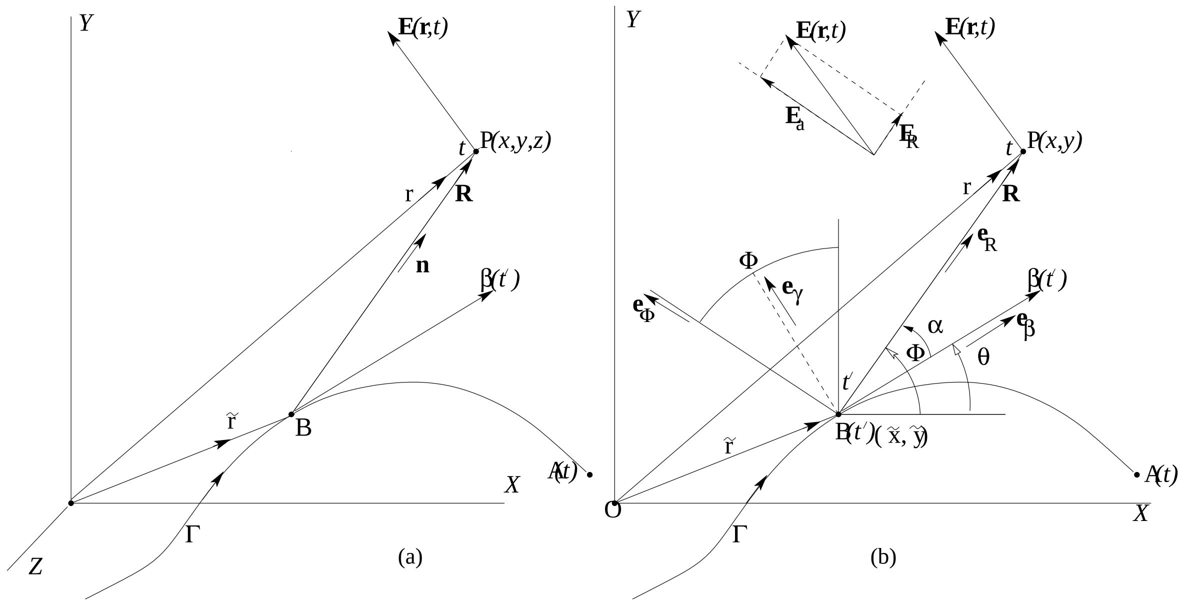


Figure 1: *path of charge q in arbitrary motion*

We have thus employed *three* sets of radius vectors, coordinates, unit vectors:
(1) $\mathbf{r} = x\mathbf{i}+y\mathbf{j}$, Cartesian coordinates, for the observer P with respect to the *fixed* origin O;
(2) $\tilde{\mathbf{r}}(t') = \tilde{x}(t')\mathbf{i} + \tilde{y}(t')\mathbf{j}$, Cartesian coordinates, for the moving charge $q$ with with respect to the *fixed* origin O;
(3) $\mathbf{R}(t') = (R(t'), \Theta(t'))$, polar coordinates, and the corresponding unit vectors $(\mathbf{e}_R(t'), \mathbf{e}_\Phi(t'))$ of the observer P with respect to the *moving* point charge $q$.

We shall rewrite the **E** vector, after breaking it up into two components, the "velocity field" $\mathbf{E}_v$ (depending entirely on the velocity) and the "acceleration field" $\mathbf{E}_a$ (involving both velocity and acceleration of the particle).

$$\begin{aligned}
\mathbf{E} &= \mathbf{E}_v + \mathbf{E}_a, \quad \text{where,}\\
\mathbf{E}_v &= \frac{q}{4\pi\varepsilon_0}\left[\frac{(1-\beta^2)(\mathbf{e}_R - \boldsymbol{\beta})}{\kappa^3 R^2}\right]_{t'=t_r}.\\
\mathbf{E}_a &= \frac{q}{4\pi\varepsilon_0}\left[\frac{\mathbf{e}_R \times \{(\mathbf{e}_R - \boldsymbol{\beta}) \times \dot{\boldsymbol{\beta}}\}}{c\kappa^3 R}\right]_{t'=t_r}.
\end{aligned} \tag{5}$$

It is seen that the acceleration field $\mathbf{E}_a$is purely *transverse*, i.e., it has only $\mathbf{e}_\Phi$ component, perpendicular to the line BP.

## 2.3 Velocity and Acceleration of $q$ with reference to the new unit vectors

We shall have to write the **E** field with reference to the new unit vectors $(\mathbf{e}_R, \mathbf{e}_\Phi)$. Let the velocity vector $c\boldsymbol{\beta}$ make angle $\theta$ from the $X$-axis, and the CtF vector $\mathbf{R}(t')$ make angle $\alpha$ from $\boldsymbol{\beta}$, as illustrated in Fig 1(b), both angles taken positive if measured anticlockwise from the respective reference lines. Let us, temporarily, introduce two orthogonal unit vectors $(\mathbf{e}_\beta, \mathbf{e}_\gamma)$, along and perpendicular to the direction of the velocity $\boldsymbol{\beta}$, and write all the new unit vectors with reference to the Cartesian unit vectors $(\mathbf{i}, \mathbf{j}, \mathbf{k})$. Note that $\Phi = \alpha + \theta$.

$$\begin{aligned} \mathbf{e}_R &= \cos\Phi\,\mathbf{i} + \sin\Phi\,\mathbf{j}. \\ \mathbf{e}_\Phi &= -\sin\Phi\,\mathbf{i} + \cos\Phi\,\mathbf{j}. \\ \mathbf{e}_\beta &= \cos\theta\,\mathbf{i} + \sin\theta\,\mathbf{j}. \\ \mathbf{e}_\gamma &= \mathbf{k} \times \mathbf{e}_2 = -\sin\theta\,\mathbf{i} + \cos\theta\,\mathbf{j}. \end{aligned} \tag{6}$$

Either using (6), or straight from the drawings in Fig.1(b), one obtains $(\mathbf{e}_\beta, \mathbf{e}_\gamma, \dot{\mathbf{e}_\beta})$ in terms of $(\mathbf{e}_R, \mathbf{e}_\Phi)$.

$$\begin{aligned} \mathbf{e}_\beta &= \cos\alpha\,\mathbf{e}_R - \sin\alpha\,\mathbf{e}_\Phi. \\ \mathbf{e}_\gamma &= \sin\alpha\,\mathbf{e}_R + \cos\alpha\,\mathbf{e}_\Phi. \\ \dot{\mathbf{e}_\beta} &= (-\sin\theta\,\mathbf{i} + \cos\theta\,\mathbf{j})\,\dot\theta = \dot\theta\,\mathbf{e}_\gamma. \end{aligned} \tag{7}$$

Hence,

$$\begin{aligned} \boldsymbol{\beta} &= \beta\mathbf{e}_\beta = \beta\,(\cos\alpha\,\mathbf{e}_R - \sin\alpha\,\mathbf{e}_\Phi) \\ \dot{\boldsymbol{\beta}} &= \dot\beta\,\mathbf{e}_\beta + \beta\,\dot{\mathbf{e}_\beta} = \dot\beta\,\mathbf{e}_\beta + \beta\,\dot\theta\,\mathbf{e}_\gamma \\ &= (\dot\beta\,\cos\alpha + \beta\,\dot\theta\,\sin\alpha)\,\mathbf{e}_R \\ &\quad +(-\dot\beta\,\sin\alpha + \beta\,\dot\theta\,\cos\alpha)\,\mathbf{e}_\Phi. \end{aligned} \tag{8}$$

## 2.4 The E field with reference to $(\mathbf{e}_R, \mathbf{e}_\Phi)$

Let us now go back to Eq. (5), and write $\boldsymbol{\Omega} = \mathbf{e}_R \times \{(\mathbf{e}_R - \boldsymbol{\beta}) \times \dot{\boldsymbol{\beta}}\} = \boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_2$, where

$$\begin{aligned} &\boldsymbol{\Omega}_1 = \mathbf{e}_R \times (\mathbf{e}_R \times \dot{\boldsymbol{\beta}}) = \mathbf{e}_R(\mathbf{e}_R \cdot \dot{\boldsymbol{\beta}}) - \dot{\boldsymbol{\beta}} \\ &= -\mathbf{e}_\Phi(\mathbf{e}_\Phi \cdot \dot{\boldsymbol{\beta}}) = (\dot\beta\,\sin\alpha - \beta\,\dot\theta\,\cos\alpha)\,\mathbf{e}_\Phi. \\ &\boldsymbol{\Omega}_2 = -\mathbf{e}_R \times (\boldsymbol{\beta} \times \dot{\boldsymbol{\beta}}) = -\mathbf{e}_R \times \beta^2\dot\theta\,\mathbf{k} = \beta^2\dot\theta\,\mathbf{e}_\Phi \\ &\boldsymbol{\Omega} = [\dot\beta\,\sin\alpha + \beta\,\dot\theta(\beta - \cos\alpha)]\,\mathbf{e}_\Phi. \end{aligned}$$

We thus get the following expression for the "acceleration field":

$$\mathbf{E}_a = \frac{q}{4\pi\varepsilon_0}\left[\frac{\dot\beta\,\sin\alpha + \beta\,\dot\theta(\beta - \cos\alpha)}{c\kappa^3 R}\mathbf{e}_\Phi\right]_{t'=t_r}. \tag{9}$$

Back to Eq. (5), note that the expression in the numerator is $(1-\beta^2)[(1-\beta\cos\alpha)\,\mathbf{e}_R + \beta\sin\alpha\,\mathbf{e}_\Phi]$, so that the "velocity field" is:

$$\mathbf{E}_v = \frac{q}{4\pi\varepsilon_0}\left[\frac{(1-\beta\,\cos\alpha)\,\mathbf{e}_R + \beta\sin\alpha\,\mathbf{e}_\Phi}{\kappa^3\gamma^2 R^2}\right]_{t'=t_r}, \tag{10}$$

where $1/\gamma^2 = 1 - \beta^2$.

Adding the two fields given in Eqs.(9) and (10), we get the complete field:

$$
\begin{aligned}
\mathbf{E}\,(\mathbf{r},t) &= \frac{q}{4\pi\varepsilon_0}\left[\frac{\dot\beta\,\sin\alpha+\beta\,\dot\theta(\beta-\cos\alpha)}{c\kappa^3 R}\mathbf{e}_{\Phi}+\frac{(1-\beta\,\cos\alpha)\,\mathbf{e}_R+\beta\sin\alpha\,\mathbf{e}_{\Phi}}{\kappa^3\gamma^2R^2}\right]_{t'=t_r}\\
&= E_R\,\mathbf{e}_R+E_{\Phi}\,\mathbf{e}_{\Phi},\quad \text{where}\\
E_R &= \frac{q}{4\pi\varepsilon_0}\left[\frac{(1-\beta^2)(1-\beta\,\cos\alpha)}{\kappa^3R^2}\right]_{t'=t_r} && (a)\\
E_{\Phi} &= \frac{q}{4\pi\varepsilon_0}\left[\frac{\frac{R}{c}\{\dot\beta\,\sin\alpha+\beta\,\dot\theta(\beta-\cos\alpha)\}+(1-\beta^2)\beta\sin\alpha}{\kappa^3R^2}\right]_{t'=t_r} && (b)\\
\kappa(t') &= 1-\boldsymbol{\beta}(t')\cdot\mathbf{e}_R(t')|_{t'=t_r}=1-\beta\cos\alpha(t'). && (c)
\end{aligned}
\tag{11}
$$

It can be observed from the above equations that "close" to the charge, the field is predominantly "velocity", and "far" from it, predominantly "acceleration". How close or how far? We shall set up a criterion in Sec. 5.2

## 3 The differential equation for the Field Lines

Let us go back to Eq. (1), relating the time of observation $t$ to the retarded time $t'$. Given any pair of values of $(\mathbf{r},t)$, determining the field $\mathbf{E}\,(\mathbf{r},t)$ at the location $\mathbf{r}$ at time $t$, this equation can be solved to yield a unique value of $t'$, and hence a unique retarded location $\tilde{\mathbf{r}}(t')$ of $q$.

Let us consider a typical field line $\Omega$ passing through the observation point P$(\mathbf{r},t)$ at any arbitrary time $t$, as shown in Fig. 2(a). If we take the observation time $t$ as fixed, then there will exist a $1 \leftrightarrow 1$ correspondence between $\mathbf{r}$ on $\Omega$ and the retarded point $\tilde{\mathbf{r}}(t')$.

Different points on the field line $\Omega$: $t, 1, 2, 3, 4$, will correspond to different points on the path $\Gamma$ at the corresponding retarded times: $t' < t'_1 < t'_2 < t'_3 < t'_4$, and different CtF vectors: $\mathbf{R}(t'), \mathbf{R}(t'_1), \mathbf{R}(t'_2), \mathbf{R}(t'_3), \mathbf{R}(t'_4)$, their magnitudes steadily increasing along the path: $R(t') > R(t'_1) > R(t'_2) > R(t'_3) > R(t'_4)$, as shown in the figure. At A, $t' = t'_A = t$ and $\mathbf{R}(t'_A) = \mathbf{0}$.

Here it should be noted that the $\mathbf{E}$ field caused by the moving charge $q$, and now existing over all space, can be pictured as field lines, all of which converge at A (present location of $q$). Infinite number of field lines, pointing in all directions, emanate from A. Only one of them, marked $\Omega$, passes through the observer at P, and is tangential to the electric field $\mathbf{E}\,(\mathbf{r},t)$. Our immediate task is to obtain the differential equation, and the

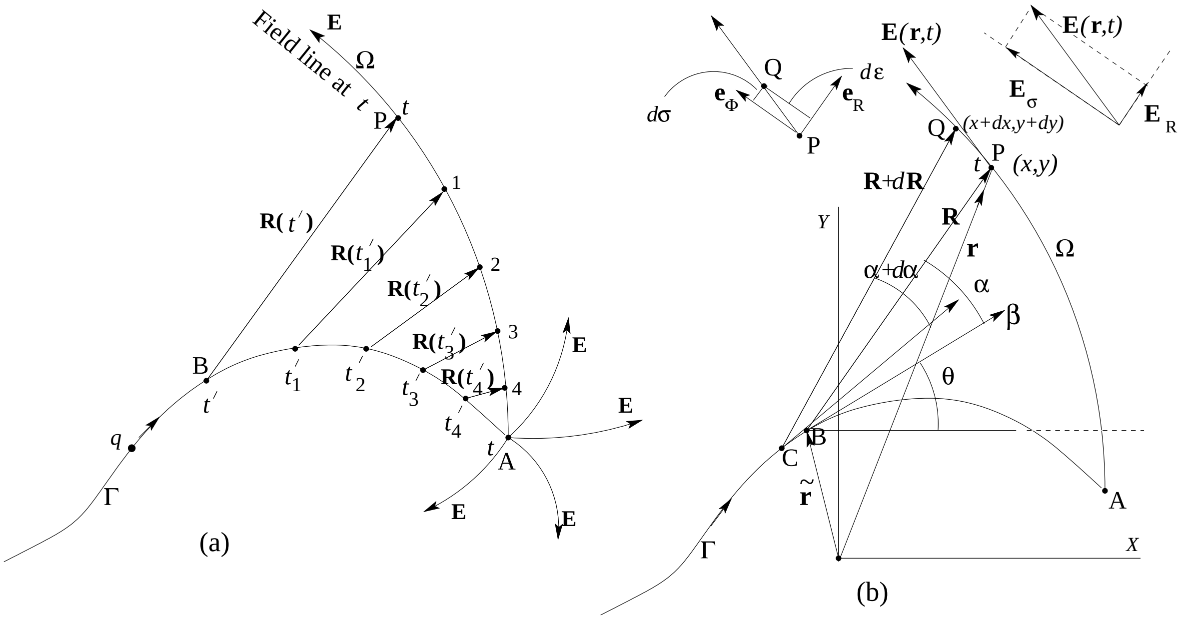


Figure 2: *Understanding the field lines. (a) Explaining the CtF vector* **R***; (b) the components of the displacement d***r** *from P to Q, and components of* **E** *at P for explaining the construction of the field equation.*

boundary condition for the field lines. Applying the boundary condition we should be able to identify Ω.

In Fig. 1b, and again in Fig.2b, we have shown the following vector and angles at the retarded location $\tilde{\mathbf{r}}(t')$: the velocity $c\boldsymbol{\beta}(t')$ of the particle; the angle $\theta(t')$ between $\boldsymbol{\beta}(t')$ and the X axis; and the angle $\alpha(t')$ between $\boldsymbol{\beta}(t')$ and $\mathbf{R}(t')$. It follows that

$$\begin{aligned} &c\beta_x(t') = \dot{\tilde{x}}; \quad c\beta_y(t') = \dot{\tilde{y}}; \\ &\beta(t') = \sqrt{{\beta_x}^2 + {\beta_y}^2}; \\ &\tan\theta(t') = \tfrac{\beta_y(t')}{\beta_x(t')}. \end{aligned} \tag{12}$$

Here $(\dot{\tilde{x}}, \dot{\tilde{y}})$ are derivatives of the functions $\xi(t'), \zeta(t')$ written in Eq. (4), with respect to $t'$. The $1 \leftrightarrow 1$ relation between $t$ and $t'$, and between $\mathbf{r}(x,y)$ and $\tilde{\mathbf{r}}(\tilde{x},\tilde{y})$, will permit us to mark a point $\mathbf{r}(x,y)$ on the field line Ω (at the present time $t$) in terms of its Curvilinear coordinates $(R,\alpha)$ (replacing $(R,\Phi)$). Noting that $\mathbf{r} = \overrightarrow{\mathrm{OP}}, \mathbf{R} = \overrightarrow{\mathrm{BP}}$, it is seen, either from Fig. 1(b), or from Fig.2(b), that

$$\begin{aligned} \mathbf{r} \quad &= \tilde{\mathbf{r}}(t') + \mathbf{R}(t'). \\ \mathbf{R}(t') \quad &= R(t')[\cos\Phi(t')\mathbf{i} + \sin\Phi(t')\mathbf{j}] \\ &= R(t')[\cos(\theta(t') + \alpha(t'))\mathbf{i} \\ &\quad + \sin(\theta(t') + \alpha(t'))\mathbf{j}]. \end{aligned} \tag{13}$$

The above equations, written in terms of Cartesian components, takes the form:

$$\begin{array}{rcll} x & = & \tilde{x}(t') + R\cos[\theta(t') + \alpha], & (a) \\ y & = & \tilde{y}(t') + R\sin[\theta(t') + \alpha], & (b) \\ \text{where} \quad t' & = & t - \frac{R}{c}, \quad t \text{ is fixed,} & (c) \end{array}$$
$$\tag{14}$$

and the variable $\theta(t')$ is given by Eq. (12). The task now boils down to finding the differential equation and the boundary conditions from which the functional relationship between $R$ and $\alpha$ is to be obtained.

At this point some explanatory notes can be useful. Consider how we determine the trajectory of a particle under inverse square law force[8]. First we set up polar coordinates $(r, \theta)$ and obtain a differential equation involving the second derivative of $r$ with respect to $\theta$, solve that equation to obtain $r = f(\theta)$ and then use this solution to get the trajectory in the parametric form: $x = r\cos\theta = f(\theta)\cos\theta; y = r\sin\theta = f(\theta)\sin\theta$, taking $\theta$ as the parameter.

In the present case we take $R$ as the parameter, obtain a differential equation that will involve derivative of $\alpha$ with respect to $R$, solve the differential equation to obtain $\alpha = f(R)$, and then the parametric equation of the field line in the form (14), in which $R$ is the parameter.

Referring to Fig. 2(b), consider two neighbouring points P and Q on the field line, at radius vectors $\mathbf{r}$ and $\mathbf{r} + d\mathbf{r}$ respectively. Suppose we write the line element $d\mathbf{r}$ as

$$\begin{aligned} d\mathbf{r} &= d\sigma\, \mathbf{e}_R + d\varepsilon\, \mathbf{e}_\Phi, \\ \text{where} \quad d\sigma &= d\mathbf{r}\cdot\mathbf{e}_R;\ d\varepsilon = d\mathbf{r}\cdot\mathbf{e}_\Phi. \end{aligned} \tag{15}$$

If the components of $\mathbf{E}(\mathbf{r}, t)$ in the directions of $\mathbf{e}_R$ and $\mathbf{e}_\Phi$ are $E_R$ and $E_\Phi$ respectively, then

$$\frac{d\sigma}{d\varepsilon} = \frac{E_R}{E_\Phi}. \tag{16}$$

The above equation would become the desired differential equation for the field lines after expressing the differentials $d\sigma, d\varepsilon$ in terms of the new coordinates $(R, \alpha)$ and their differentials. The steps are long and difficult. One reason is that the displacement $d\mathbf{r} = \overrightarrow{\mathrm{PQ}} = d\tilde{\mathbf{r}} + d\mathbf{R}$, and a look at Fig 2(b) will indicate that determination of the two vector differentials can be complicated. We have shifted this work to Sec. A.2 in Appendix A. What matters right now is the final differential equation for the field lines which we write as[5]:

$$\frac{d\alpha}{dR} = \frac{1}{c}\gamma^2[\dot{\theta} - \dot{\boldsymbol{\beta}}\cdot\mathbf{e}_\Phi]. \tag{17}$$

## 4 The Boundary Condition

As with all differential equations, the equation (17) alone will not lead to the particular solution desired by us, unless we know how to evaluate the arbitrary constant(s) that come from integration of the differential equation. To get *the* particular solution of a differential equation one has to specify the initial condition(s), or the boundary condition(s) and evaluate the constant(s).

In this case there are infinite number of field lines that will satisfy Eq. (17). We shall have to identify one of them as *the* line. For this purpose we have to go to the the point A, the present location of the charge $q$, the fountain head from where all field lines spring out, and identify the one of our choice by specifying the boundary condition.

It can be seen from Eq. (5) that *close to the location of the particle*, $R \to 0$, and the

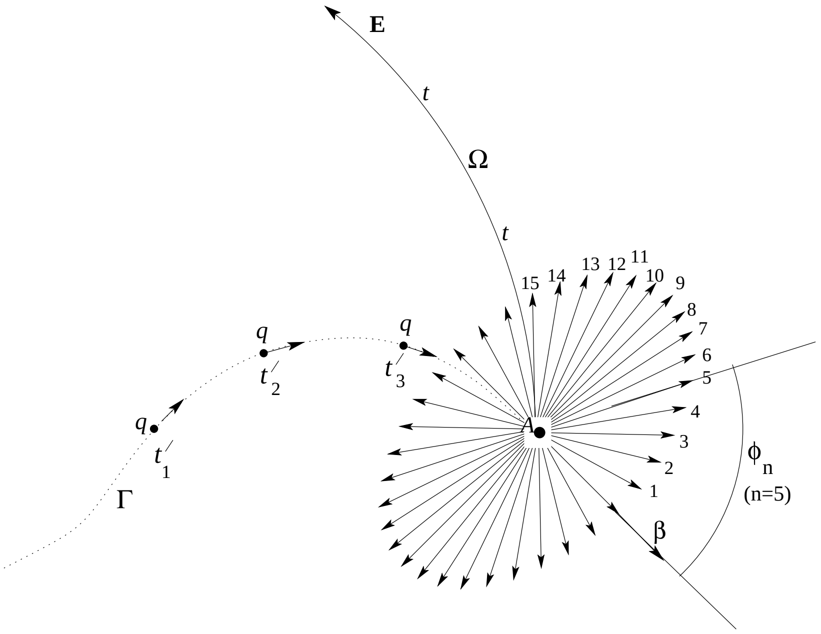


Figure 3: *Fountain lines at the current location of q, for understanding the boundary condition*

*field becomes a velocity field*, with field lines issuing out as straight lines, in all directions [4]. We have reproduced the same field lines in Fig. 3, after rotating them in such a way that the 0-th line aligns with the direction of the particle velocity $c\boldsymbol{\beta}$ at the time $t$. We shall call them *fountain lines.*

Let us imagine $N$ fountain lines, all of them confined to the $XY$ plane, issuing out from A, and tag them as 1,2,...,N. (This number is 36 in the figure.) When $\beta \rightarrow 0$ the angular intervals between successive fountain lines becomes the same and equal of $2\pi/N$, so that the fountain line $\#n$ will make angle $\theta_n = 2\pi n/N$ with the $\boldsymbol{\beta}$ vector. According to the Purcell condition[4], the angle $\phi_n$ that the same fountain line will make with the direction of $\boldsymbol{\beta}$, when it is relativistic, is

$$\tan\phi_n = \gamma\tan\theta_n = \gamma\tan\left(\frac{2\pi n}{N}\right). \qquad (18)$$

We have illustrated this in Fig. 3.

We would like to plot $N$ field lines, each of them issuing out of the point A (the present location of the particle) as fountain lines, such that the $n$-th field line ($n = 1, 2, ..., N$) is tangential to the fountain line $\#n$, making the angle $\phi_n$ with $\boldsymbol{\beta}$ at the source point A, as required by Eq. (18). *This is our boundary condition.*

An example is shown in Fig. 3, in which the field line $\Omega$ issues out from the source point as the fountain line # 15. This figure sheds further light on the difference between the path $\Gamma$ of the particle, shown as a dotted line, and the field line $\Omega$, shown as a solid line. The path $\Gamma$ does not exist as one piece at any given instant of time. Different points on this curve exist as dots at different times $t'_1, t'_2, t'_3, ....$ Also it is confined to the region in which the particle exists, either at rest or in motion. The field line $\Gamma$ exists as one piece at any given instant of time, issues out from the point A, and stretches out all the way to infinity, mainly due to Coulomb's law.

It is obvious that $R = 0$ at A, and the tangent to the field line is

$$\left.\frac{dy}{dx}\right|_{R=0} = \left.\frac{E_y}{E_x}\right|_{R=0} = \tan(\phi_n + \theta)|_{R=0}. \quad (19)$$

It is seen from Eq. (5) that $\mathbf{E} \to \mathbf{E}_v$ as $R \to 0$. Therefore,

$$\begin{aligned} \left.\frac{E_y}{E_x}\right|_{R=0} &= \left.\frac{E_{vy}}{E_{vx}}\right|_{R=0} = \left.\frac{(\mathbf{e}_R - \boldsymbol{\beta})\cdot\mathbf{j}}{(\mathbf{e}_R - \boldsymbol{\beta})\cdot\mathbf{i}}\right|_{R=0} \\ &= \left.\frac{\sin(\theta+\alpha) - \beta\sin\theta}{\cos(\theta+\alpha) - \beta\cos\theta}\right|_{R=0}. \end{aligned} \quad (20)$$

Let us write $\alpha = \alpha_n$ for the $n$-th line at $R = 0$. Then from (19) and (20),

$$\tan(\phi_n + \theta)|_{R=0} = \frac{\sin(\theta + \alpha_n) - \beta\sin\theta}{\cos(\theta + \alpha_n) - \beta\cos\theta}, \quad (21)$$

from which it follows that (Sec.A.3, Appendix A)

$$\tan\phi_n = \frac{\sin\alpha_n}{\cos\alpha_n - \beta}. \quad (22)$$

The boundary condition (18) now reduces to the form

$$\gamma\tan\left(\frac{2\pi n}{N}\right) = \frac{\sin\alpha_n}{\cos\alpha_n - \beta} \quad (23)$$

Some trigonometrical manipulations (Sec.A.3, Appendix A) will convert the above equation to the following form[5]

$$\tan\frac{\alpha_n}{2} = \sqrt{\frac{1-\beta}{1+\beta}}\tan\frac{\pi n}{N};\ n = 1, 2, \cdots N, \quad (24)$$

in which we shall apply the Boundary Condition..

We *summarize* as follows. A general field line $\Omega$, at a given instant $t$, will be expressed in the parametric form $\alpha = f(R)$. We shall obtain $N$ field lines $\{\Omega_n : \alpha_n = f_n(R);\ n = 1, 2, 3, \cdots, N\}$, springing out from the present location A of the point charge $q$, at the angles $\phi_n$ as given by Eq.(18). These $N$ lines $\{\Omega_n\}$ are to be obtained (and subsequently plotted) by the following procedure.

$$\boxed{\begin{array}{rrcll} \text{Solve the differential equation} & \dfrac{d\alpha}{dR} & = & \dfrac{1}{c}\gamma^2[\dot{\theta} - \dot{\boldsymbol{\beta}}\cdot\mathbf{e}_\Phi], & (a) \\ \text{subject to the boundary condition} & \tan\dfrac{\alpha}{2} & = & \sqrt{\dfrac{1-\beta}{1+\beta}}\tan\dfrac{\pi n}{N}, & (b) \\ \text{to obtain} & \alpha_n & = & f_n(R);\ n = 1, 2, 3, \cdots, N & (c) \\ \text{Insert this in Eq.(14):} & x & = & \tilde{x}(t') + R\cos[\theta(t') + \alpha_n], & (d) \\ & y & = & \tilde{y}(t') + R\sin[\theta(t') + \alpha_n], & (e) \\ \text{where} & t' & = & t - \dfrac{R}{c}, \quad \text{(see Eq. 1)} & (f) \end{array}} \quad (25)$$

to get the parametric equations for the $N$ field lines, remembering that Eqs.(a),(b),(d),(e) involve $\beta$ which is a function of $t'$, and hence through Eq.(f), a function of $R$.

# 5 Specific Example 1: Particle Accelerating along a Straight Line

## 5.1 General Formula

Fig. 4 shows the particle moving along the $X$ axis, with acceleration $a$. We shall specialize Eqs. (25), for this special case, by setting $\theta = 0; \dot{\theta} = 0$. Note from Eq. (8b) that $\dot{\boldsymbol{\beta}} \cdot \mathbf{e}_{\Phi} = -\dot{\beta} \sin\alpha$. The differential equation for the field lines, Eq. (25a), now reduces to:

$$\begin{aligned} & \tfrac{d\alpha}{dR} = \tfrac{1}{c}\gamma^2\dot{\beta}\sin\alpha \\ \text{Or,}\quad & \csc\alpha\, d\alpha = \tfrac{1}{c}\gamma^2\dot{\beta}\, dR \\ & = \tfrac{1}{c}\gamma^2 d\beta\, \tfrac{dR}{dt'} = -\tfrac{d\beta}{1-\beta^2}. \end{aligned} \tag{26}$$

For the last equality in the last line we used Eq. (1), to get $dR/dt' = -c$.

We now integrate the last line of Eq. (26), and get

$$\begin{aligned} \ln\tan\tfrac{\alpha}{2} &= \ln\sqrt{\tfrac{1-\beta}{1+\beta}} + \text{const} \\ \text{Or, } \tan\tfrac{\alpha}{2} &= K\sqrt{\tfrac{1-\beta}{1+\beta}} \end{aligned} \tag{27}$$

where $K$ is a constant. Going back to the boundary condition given in Eq. (25b) we evaluate this constant as $K = \tan\frac{\pi n}{N}$, leading to the solution given as

$$\boxed{\tan\frac{\alpha_n}{2} = \sqrt{\frac{1-\beta}{1+\beta}}\tan\frac{\pi n}{N};\ n = 1, 2, \cdots N.} \tag{28}$$

The above equation represents the desired relation between $\alpha$ and $R$ for the field line that starts from the source point as the fountain line $\#n$, making angle $\theta_n = 2\pi n/N$ with the $X$-axis. We shall illustrate this for the special case of a particle moving under a constant force along the $X$ axis.

## 5.2 Special Case: Particle moving under a constant accelerating force

Before going far, we shall write the expression for the $\mathbf{E}$ field from rectilinear motion of the charge $q$ specializing Eq. (11) to this case, by setting $\dot{\theta} = 0$.

$$\begin{aligned} E_{R\text{v}} &= \frac{q}{4\pi\varepsilon_0}\left[\frac{(1-\beta\,\cos\alpha)}{\kappa^3 R^2}\right]_{t'=t_r} && (a) \\ E_{\Phi\text{v}} &= \frac{q}{4\pi\varepsilon_0}\left[\frac{\beta\sin\alpha}{\gamma^2\kappa^3 R^2}\right]_{t'=t_r} && (b) \\ E_{\Phi\text{a}} &= \frac{q}{4\pi\varepsilon_0}\left[\frac{\dot{\beta}\,\sin\alpha}{\gamma^2\kappa^3 c R}\right]_{t'=t_r} && (b) \end{aligned} \tag{29}$$

We have broken up the field into its “velocity” components $E_{R\text{v}}, E_{\Phi\text{v}}$, and “acceler-

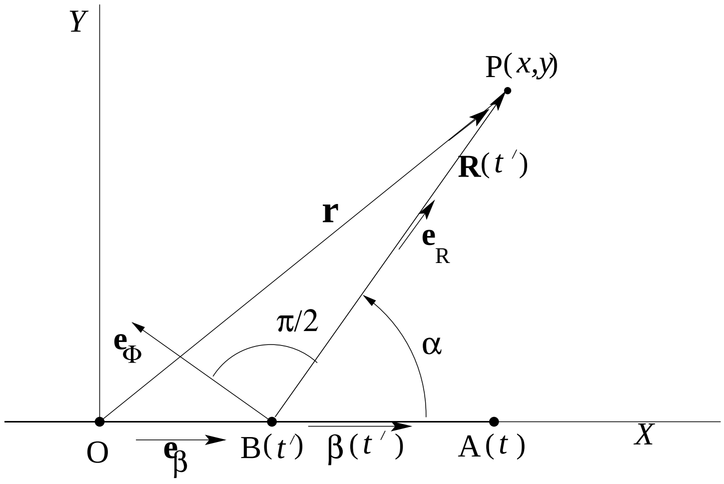


Figure 4: *Particle in linear acceleration*

ation" component $E_{\Phi\mathrm{a}}$. As emphasized following Eq. (11), the field is predominantly "acceleration" far from the charge (where our interest lies). Taking $\beta \sim 1$ the transverse components of the two field components are nearly equal at $R = c^2/a$. Hence the criterion is that

$$\mathbf{E} \approx \mathbf{E}_{\mathrm{a}}, \quad \text{for } R \gg c^2/a. \tag{30}$$

Page and Adams[9] have shown plots of the acceleration component $\mathbf{E}_{\mathrm{a}}$. which are circles passing through the present location of $q$, with centres on the the plane $x = -c^2/a$. In our case we shall plot the entire $\mathbf{E}$ field.

For some of the examples we are going to consider soon, (i) $a = 0.17 \times 10^{18}\,\mathrm{m/s^2}$, and $c^2/a = 0.529$ m, same as the unit of distance $\tau c$ we shall adopt, (ii) $a = 0.17 \times 10^{20}\,\mathrm{m/s^2}$, and $c^2/a = 0.005$ m, far less than the unit of distance we shall adopt.

Now, let us consider the rectilinear motion of particle of charge $q$ and *rest mass* $m_0$, under the influence of a uniform electric field $\mathbf{E} = E\mathbf{i}$. This electric field induces a constant acceleration $a = qE/m_0 = F/m_0$ in the $X$ direction, as measured in its *instantaneous rest frame.* The particle will ultimately achieve relativistic velocity.

The displacement $\tilde{x}(t')$ and the velocity $c\beta(t')$, both as functions of the retarded time $t'$, are given by the following formulas[10, 11, 12]:

$$\begin{aligned} \tilde{x}(t') &= c\tau\left[\sqrt{1+\left(\tfrac{t'}{\tau}\right)^2}-1\right] \\ \beta(t') &= \left[1+\left(\tfrac{\tau}{t'}\right)^2\right]^{-\frac{1}{2}}; \quad t' > 0. \end{aligned} \tag{31}$$

In the above equations we have defined a characteristic time $\tau$, such that the particle would reach the velocity $c$ in this time, if non-relativistic mechanics had been applicable. That is,

$$a\tau = c. \quad \text{Or, } \tau = c/a \tag{32}$$

Replacing $t'$ by $t - R/c$ in Eqs. (31) (see Eq. 1), we get expressions for $\tilde{x}$ and $\beta$ in terms of the parameter $R$, at the present time $t$.

The equation of the field line is now obtained by going back to Eqs. (25d,e), then setting $\tilde{y} = 0, \theta(t') = 0$:

$$\begin{aligned} x &= \tilde{x}(R) + R\cos\alpha_n. & (a) \\ y &= R\sin\alpha_n. & (b) \end{aligned} \tag{33}$$

Let us consider a charged particle, e.g., an electron, which is accelerated in a 30m long linear accelerator (e.g., pelletron) to 30 MeV. The electric field through which the particle is accelerated is assumed to be uniform, and equal to $E = 10^6$ V/m. Because of our preference for positive quantities, we shall assume the particle to be positively charged, but having the the same magnitude of charge as that of an electron or proton, viz. $e = 1.6 \times 10^{-19}$ C, and mass that of an electron or positron. viz., $m_0 = 9.11\times\times 10^{-31}$ kg. The acceleration is then

$$\begin{aligned} a &= \tfrac{eE}{m_0} = \tfrac{1.6\times10^{-19}\times10^6}{9.11\times\times10^{-31}} \\ &= 0.17 \times 10^{18}\text{m/s}^2 \\ \tau &= (3\times10^8)/(0.17\times10^{18}) \\ &= 17.6\times10^{-10}\,\text{s}. \\ \tau c &= 17.6\times10^{-10}\times3\times10^8 \\ &= 0.528\,\text{m}. \end{aligned} \tag{34}$$

The particle is at the point A at the current time $t$. At this point $t' = t$. Also note from Eq. (25f) that (i) $R = ct$ at $t' = 0$, when the particle is at O, and (ii) $R = 0$ at $t' = t$, when the particle is at A.

We shall consider two variations of this exercise. For each case we shall assume that the particle is originally sitting at rest at the origin O, untill the **E** field is turned on at the time $t = 0$.

***Case 1***. The **E** field continues for ever. It is at the origin at $t' = 0$ and reaches a point A on the $X$ axis at $t' = \tau$, which we set equal to the present time $t$ (in the Ex.I to follow). That is, $t = \tau$.

***Case 2***. The **E** field is switched off at $t = \tau$, when it is at A1, then continues to move with constant velocity and reaches the point A2 when $t' = 2\tau$, which we set equal to the present time $t$. That is, $t = 2\tau$.

### 5.2.1 Plotting the E field for Case 1

To avoid any confusion we shall remind the reader that we have been, and shall be, using two symbols for time, viz., $t'$, to mean time measured *along the track of the particle*, and $t$, to mean *present time*, i.e., the time when the entire **E** field, and the field lines are viewed. Both times are measured in the "Lab frame" (No relativity is involved.)

We shall take three successively increasing values of the present time $t$, resulting in successively increasing values of $\beta$ at $t$, and get the required values at A and O, using Eqs. (31).

Ex. I: $t = \tau = 17.6 \times 10^{-10}$ s.

$$\begin{aligned} \beta(\text{A}) &= 1/\sqrt{2} = 0.707. \\ \tilde{x}(\text{A}) &= c\tau\,[\sqrt{2}-1] = 0.218\text{ m}. \\ R(\text{O}) &= \tau c = 0.528\,\text{m}. \end{aligned} \tag{35}$$

Ex. II: $t = 2\tau = 35.2 \times 10^{-10}$ s.

$$\begin{aligned} \beta(\text{A}) &= 2/\sqrt{5} = 0.894. \\ \tilde{x}(\text{A}) &= c\tau\,[\sqrt{5}-1] = 0.653\text{ m}. \\ R(\text{O}) &= 2\tau c = 1.056\,\text{m}. \end{aligned} \tag{36}$$

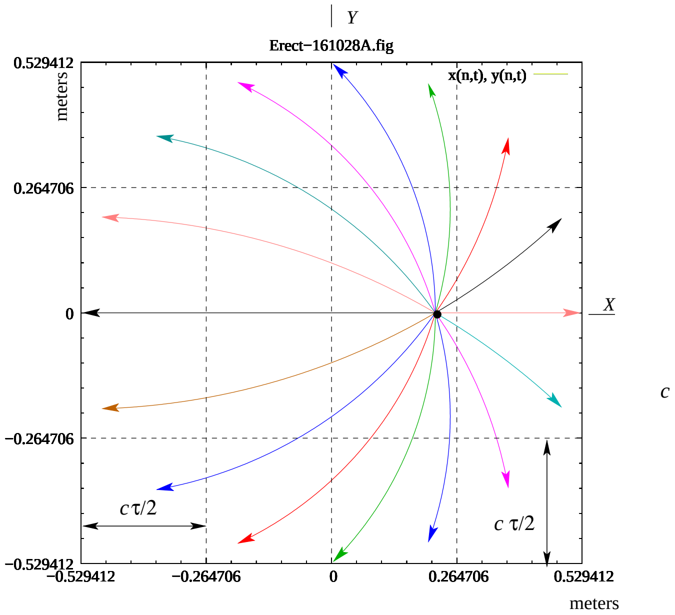
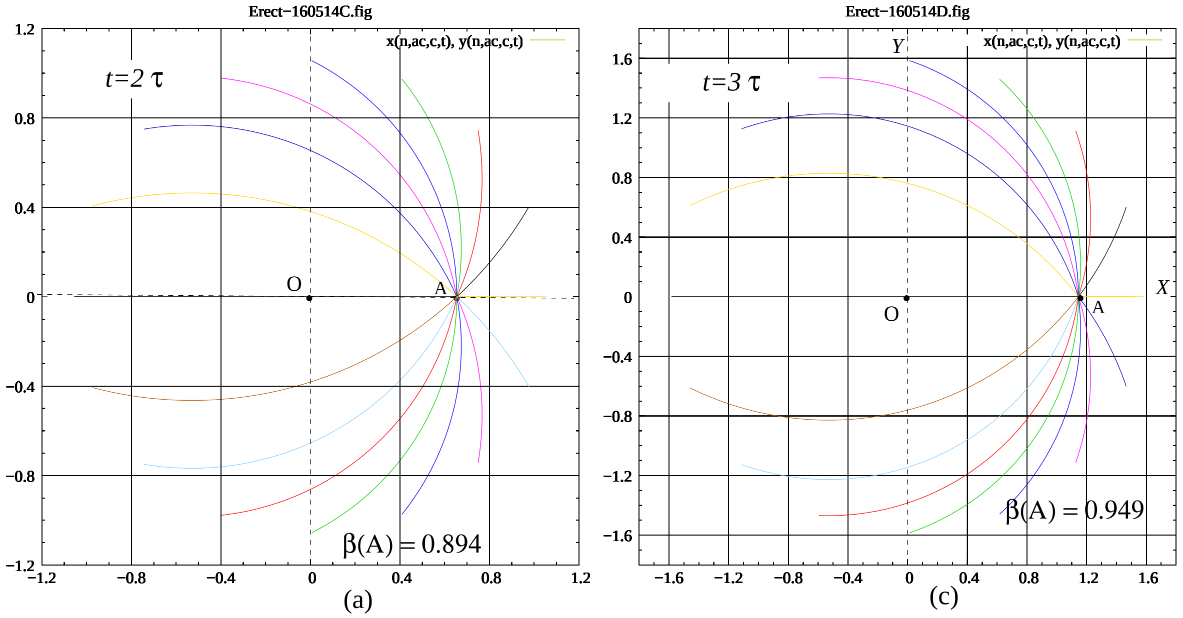


Figure 5: **E** *field line at the* $t = \tau$ *due to charge* $q$ *moving with constant acceleration*



(a) (c)

Figure 6: **E** *field line from a particle moving with constant acceleration at two different instants of time: (a)* $t = 2\tau$*, (b)* $t = 3\tau$*.*

Ex. III: $t = 3\tau = 52.8 \times 10^{-10}$ s.

$$\begin{aligned} \beta(\mathrm{A}) &= 3/\sqrt{10} = 0.949. \\ \tilde{x}(\mathrm{A}) &= c\tau\,[\sqrt{10} - 1] = 1.14 \text{ m.} \\ R(\mathrm{O}) &= 3\tau c = 1.584\,\text{m}. \end{aligned} \tag{37}$$

We shall plot the field at the present time, by the following steps.

- Insert the values of $a$ and $\tau$ as obtained in Eqs.(34) in Eqs. (31) to get $\tilde{x}(t'), \beta(t')$ as functions of $t'$ only;
- Obtain $\alpha(t')$ as a function of $t'$, from Eq.(28);
- Use the expressions of $\tilde{x}(t'), \alpha(t')$ in Eqs. (33), in which $x, y$ now look like functions of $t'$;
- Change the variable from $t'$ to $R$ using the relationship (25f), i.e., by setting $t' = t - R/c$, treating $t$ as a constant. We now get the equations of the field lines in the form of the parametric equations: $x = x(R); y = y(R)$, where $R$ is the parameter;
- $R = (0, ct)$, at (A,O). We shall therefore take the range of $R$ as $[0 : ct]$.

We have plotted three sets of field lines, taking $N = 16$ in all cases, and corresponding to $t = \tau$ in Fig.5, and corresponding to $t = 2\tau, 3\tau$ in Fig.6.

### 5.2.2 Plotting the E field for Case 2

We consider the following situation. The same particle, as described in Case 1, originally sitting at the origin O, undergoes the same acceleration $a$ from $t' = 0$ to $t' = \tau$, to reach the point A1, then moves with constant velocity. We would like to plot the **E** field emanating from the particle at the present time $t = t' = 2\tau$ when it is instantaneously located at the point A2, as shown in Fig. 7.

Here it should be noted that during the time $t' < 0$, when the particle was sitting at O, it was giving out the Coulomb field continuously from $t' = -\infty$ to $t' = 0$, spreading over all space, all the way to infinity. Our plots will also take a glimpse of a part of this Coulomb field. For this purpose we shall take the *retarded time* in the range $-\tau < t' \leq 2\tau$, corresponding to the CtF vector $3\tau > R \geq 0$. We shall therefore divide the time zone into 3 parts and write down expressions for velocity $c\beta$ and displacement $\tilde{x}$ and for these 3 parts. For this purpose we shall represent the the functions $\beta(t'), \tilde{x}(t')$ written in Eqs.(31) by new symbols $\bar{\beta}(t'), \bar{x}(t')$.

| Time range → | $-\tau \leq t' \leq 0$ | $0 < t' \leq \tau$ | $\tau < t' \leq 2\tau$ |
|---|---|---|---|
| $\beta(t') =$ | 0 | $\bar{\beta}(t')$ | $\bar{\beta}(\tau)$ |
| $\tilde{x}(t') =$ | 0 | $\bar{x}(t')$ | $\bar{x}(\tau) + c\bar{\beta}(\tau)(t' - \tau)$ |

(38)

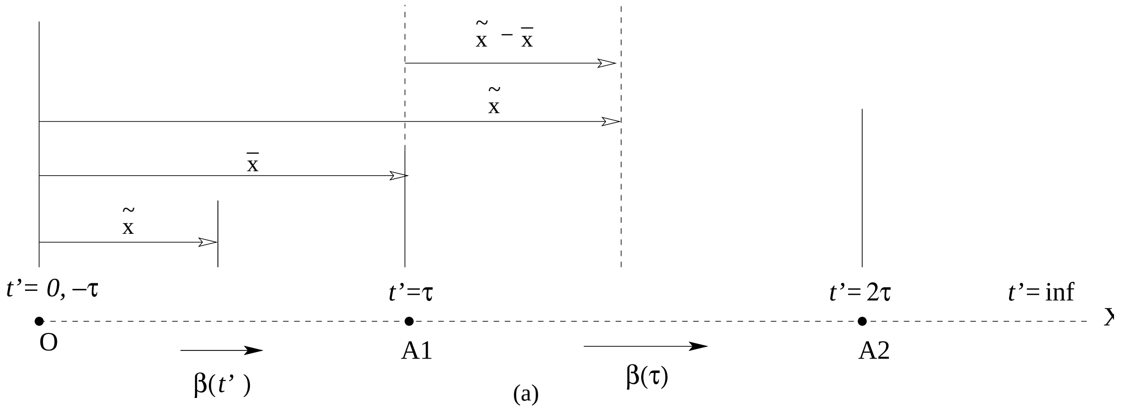


Figure 7: *Path of $q$ along O - A1 - A2 - $\infty$.*

Taking $N = 16$, we have plotted 16 field lines emanating from $q$ at the time $t' = t = 2\tau$, when it is located at A2, following the same steps as given for Case 1. The plots of the field lines are displayed in Fig.8(a).

## 5.3 Lessons from the field lines of accelerating charge

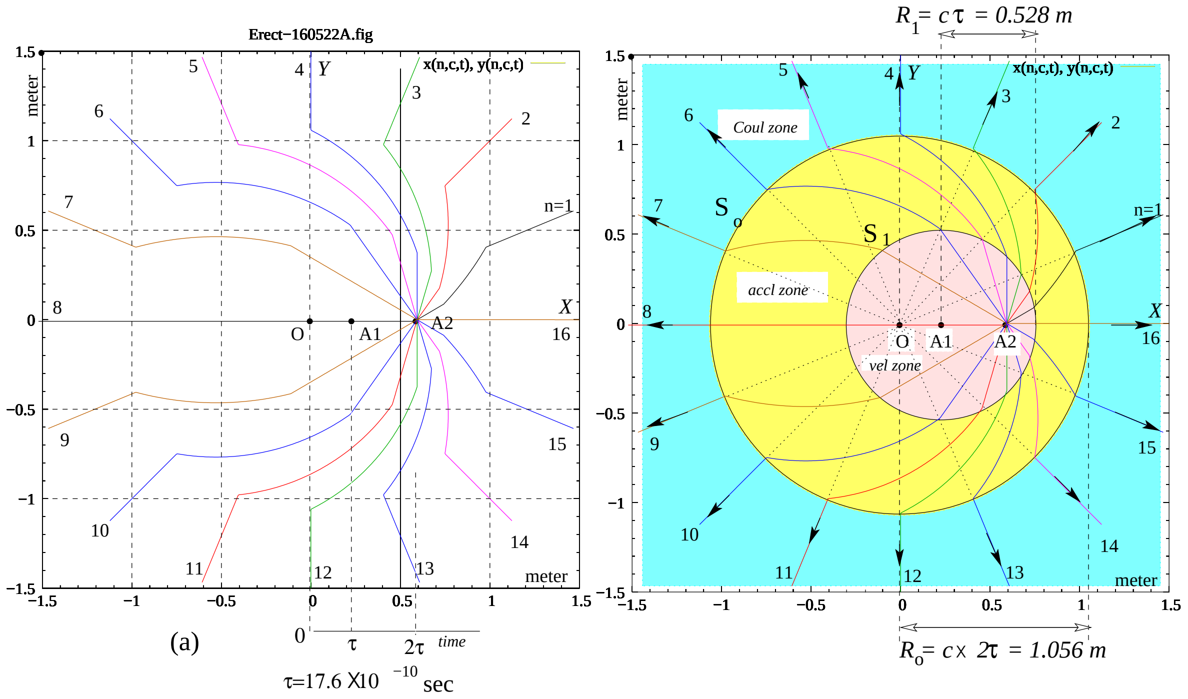
The plots for the Case 2 can give some valuable insight into the propagation of the electromagnetic field, if we examine the field lines with some interest and care. For this purpose we have reproduced the plots of Fig.8(a) in Fig.8(b), and divided the space into 3 zones: (i) Coulomb zone, (ii) acceleration zone, (iii) velocity zone.

The zone (i) is all space, all the way to infinity, but lying beyond a sphere of radius $R_0 = c\tau = 1.056$ m, centred at O, and labelled $S_0$. The field lines in this zone are straight lines, spreading out isotropically in all directions. If these lines are extended backward, they will all meet at the point O, the original rest house of the charge $Q$. These lines are the typical *Coulomb field lines.*

The sphere $S_0$ arises at $t = 0$ when the particle starts moving, and keeps expanding with the speed of light. In the short time $t = 2\tau = 35.2 \times 10^{-10}$ s, this sphere has expanded from zero radius to the radius $R_0 = 2c\tau = 1.056$ m, and keeps expanding forever. Observers outside this sphere, thousands of kilometers away, even light years away from the origin O, see the particle unmoved, see only its “static” Coulomb field, with the radial field lines passing by him.

The zone (iii) lies within another sphere labelled $S_1$ with centre at the fixed point A1. Its radius of has expanded from zero radius at $t = \tau$ to radius $R_1 = c\tau = 0.528$ m at $t = 2\tau$.

Observers within this sphere see the **E** field of the charge $q$ moving with constant velocity $0.707c$, and the field lines issuing out radially from the moving point A2, the instantaneous location of the charge $q$. These lines are similar to the lines shown in Fig. 3, bunched more



(a)

Figure 8: **E** *field lines from a particle accelerating from $t = 0$ to $t = \tau$, then moving with constant velocity.*

densely transverse to the direction of the velocity vector $c\boldsymbol{\beta}$, than in the longitudinal direction.

The zone (ii) lies between the two spheres $S_0$ and $S_1$. Observers in this zone see the particle moving in the $X$ direction with constant acceleration $a$.

Fig. 9 in which we have focussed on the time evolution of only one field line, corresponding to $n = 3$, may shed further light on the concept of field propagation. Fig (a) shows how this field line has evolved from $t = \tau$ to $2\tau$ to $3\tau$. The *Coulomb field* that had been emanating from O from $t' = -\infty$ to $t' = 0$, and propagating along the the radial line Oc, is plotted as the segment $b1 - b2 - b3 - c$. The *acceleration field*, the source of which moved from O to A1, is plotted at the present times $t = \tau, 2\tau, 3\tau$, as the shifting arcs $\Omega_1, \Omega_2, \Omega_3$ respectively, with the points O mapped onto $b1, b2, b3$, and A1 onto $a1, a2, a3$, at the same respective present times. The *velocity field* that ensued from the moving point A1 → A2 → A3 as $t' = \tau \to 2\tau \to 3\tau$, is shown as straight lines changing from A2-$a2$ to A3-$a3$ as the present time changed from $t = 2\tau$ to $t = 3\tau$.

Since the Coulomb lines for three lines are merged in Fig (a), we have separated them into three individual lines in Figs. (b), (c), (d) for a clearer look at them.

The boundary condition (18) is valid for all

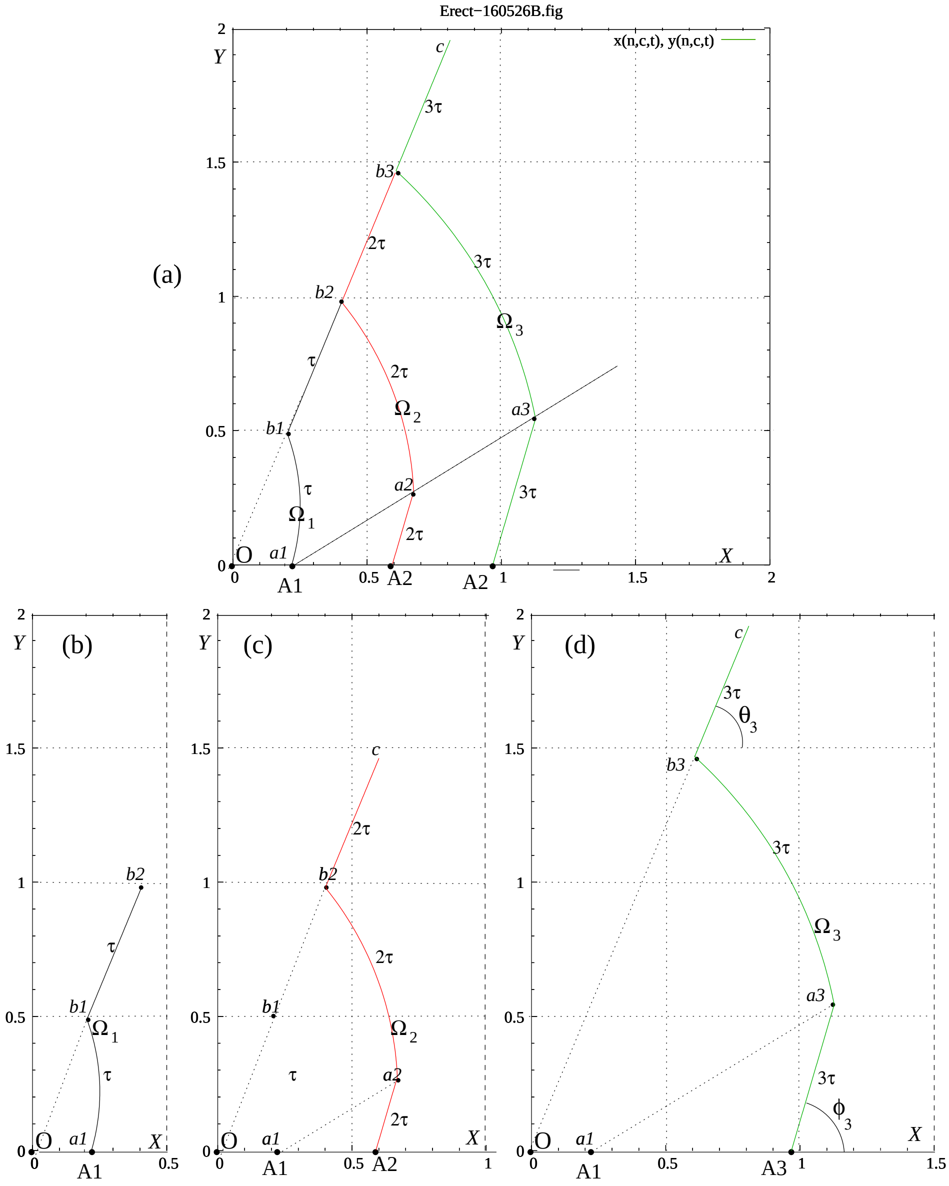


Figure 9: *Field lines corresponding to* $n = 3$*: at* $t = \tau, 2\tau, 3\tau$ *in (a), at* $t = \tau$ *in (b), at* $t = 2\tau$ *in (c), at* $t = 3\tau$ *in (d).*

the three field lines. The reader should verify this by setting $n = 3$, $\gamma = \sqrt{2}$, so that $\theta_3 = \frac{6\pi}{16} = 67.5^0$, $\phi_3 = 73.65^0$. Now if he makes a printout of this page and measures the angles $(\theta_3, \phi_3)$, with a protractor, he will verify these angles. This lends further confirmation of our faith in the amazing power of Gnuplot.

## 5.4 Special Case: Particle moving under a constant decelerating force

### 5.4.1 Plot of the Field Lines

We shall now consider the deceleration case. The same constant electric field **E**, resulting in the same acceleration vector $\boldsymbol{a}$, pointing in the x-direction will apply in this case as well. However, the particle will be moving in the *negative* x direction, approaching the origin from $+\infty$, during particle time $t' = -\infty$ to $t' = 0$, (instead of moving away from the origin, from $t' = 0$ to $t' = +\infty$ examined in the case of accelerating particle, considered in the previous section). The **E** field is acting like a break (as in bremsstrahlung radiation) bringing $q$ to complete stop at O. We shall modify the velocity and displacement functions of Eq. (31) to its new forms:

$$\begin{aligned} \tilde{x}(t') &= c\tau\left[\sqrt{1+\left(\tfrac{t'}{\tau}\right)^2} - 1\right]\ ; \quad t' \le 0 \quad (a) \\ \beta(t') &= -\left[1+\left(\tfrac{\tau}{t'}\right)^2\right]^{-\frac{1}{2}}\ ; \qquad t' < 0. \quad (b) \\ &= 0\ ; \qquad\qquad\qquad\qquad t' = 0 \quad (c) \end{aligned} \tag{39}$$

Note that the velocity function has undergone a change in sign, whereas the displacement function remains the same. We have presented a picture of the configuration in Fig.10.

Fig.10(a) shows the path of the particle, coming from $+\infty$ with velocity $c\beta_\infty = \beta(-\tau) = -\frac{c}{\sqrt{2}}$, undergoes deceleration from the point B1 to O, and stops at O. Fig.10(b) shows the configuration on the $t'$ scale.

Before plotting the **E** field lines for the decelerating charge, let us take stock of what we should expect, in comparison with the plots of the accelerating charge viewed at length in Sec. 5.2. For this purpose we go back to Eqs. (29).

The acceleration vector (in the guise of deceleration) continues to be in the positive x direction. Just as $\beta$ becomes negative, the angle $\alpha$ changes to angle $\alpha+\pi$, so that $\beta\sin\alpha$ does not change. Of the three components, $E_{\Phi\mathrm{v}}$ and $E_{\Phi\mathrm{a}}$ do not change at all. Only the component $E_{R\mathrm{v}}$ changes sign, but has negligible effect in regions perpendicular to the path. The net result is that the field lines for the deceleration case should not be drastically different from the acceleration case.

However, there is one important difference. The decelerating charge has stopped (momentarily) at the present time $t = 0$, whereas the accelerating charge is moving with speed $\beta = 0.707$ at the present time $t = \tau$, and the field lines are issuing out from $q$ at the present times, in all the examples. Therefore the field lines from the accelerating charge will make steeper angles with the x-axis at its source point, compared to the decelerating one, in accordance with the Purcell's condition.

Let us now turn to plotting. We shall con-

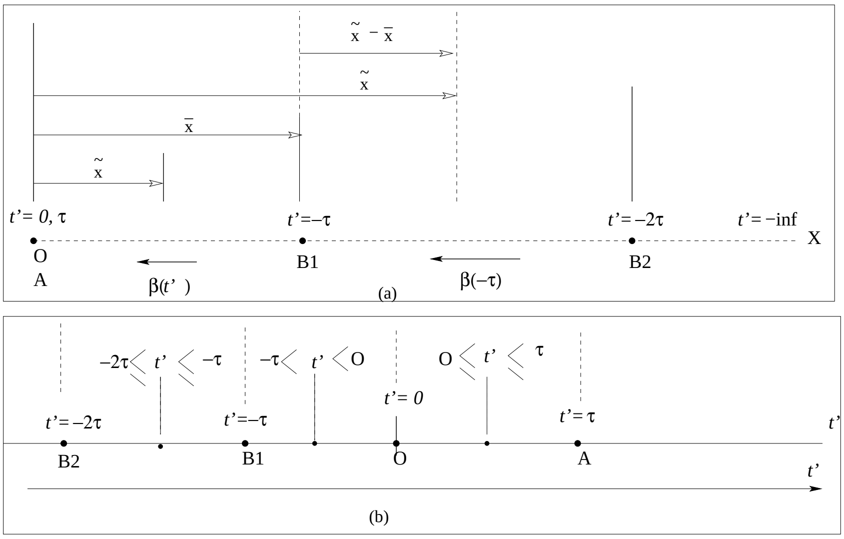


Figure 10: *path of charge $q$ moving from infinity towards the origin $O$ under deceleration. Fig (a) shows the direction of motion and the location of $q$ on the x-axis; Fig (b) shows the retarded time $t'$ axis, divided into three segments.*

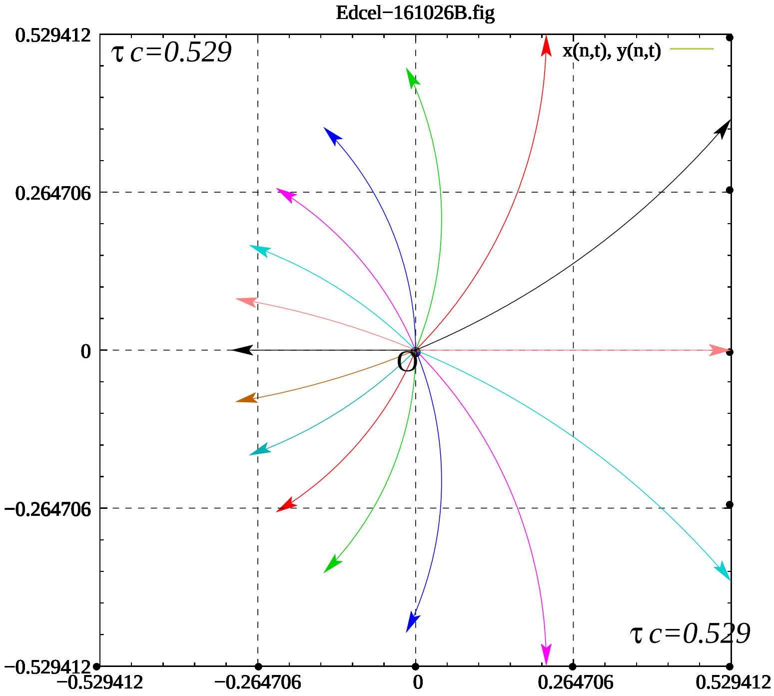


Figure 11: *Plots of* **E** *for case I of deceleration*

sider the following three cases.

***Case I***. Deceleration $a = a_o = 0.17 \times 10^{18}\,\mathrm{m/s^2}$ is forever, never stops. Present time is $t = t' = 0$ at the event "the particle reaches O". Plots of the field lines are relatively simple, and are shown in Fig.11.

***Case II***. The particle passes B2 at $t' = -2\tau$, reaches B1 at $t' = -\tau$, when deceleration $a = a_o$ starts, reaches O at $t' = 0$, when deceleration stops. The particle stays at O from $t' = 0$ to $t' = +\infty$. Present time is $t = t' = \tau$.

The velocity and displacement as functions of $t'$ are given in the table below, which is analogous to the equation table (40).

| Time range → | $-2\tau \le t' \le -\tau$ | $-\tau < t' < 0$ | $0 \le t' \le \tau$ |
|---|---|---|---|
| $\beta(t') =$ | $\bar{\beta}(-\tau)$ | $\bar{\beta}(t')$ | 0 |
| $\tilde{x}(t') =$ | $\bar{x}(-\tau) + c\bar{\beta}(-\tau)(t' + \tau)$ | $\bar{x}(t')$ | 0 |

(40)

We have plotted the $\beta(t')$ and $\tilde{x}(t')$ functions in Fig. 12. Plots for the field lines are shown in Fig.13.

***Case III***. The situation is the same as in Case 2, with the the following changes:

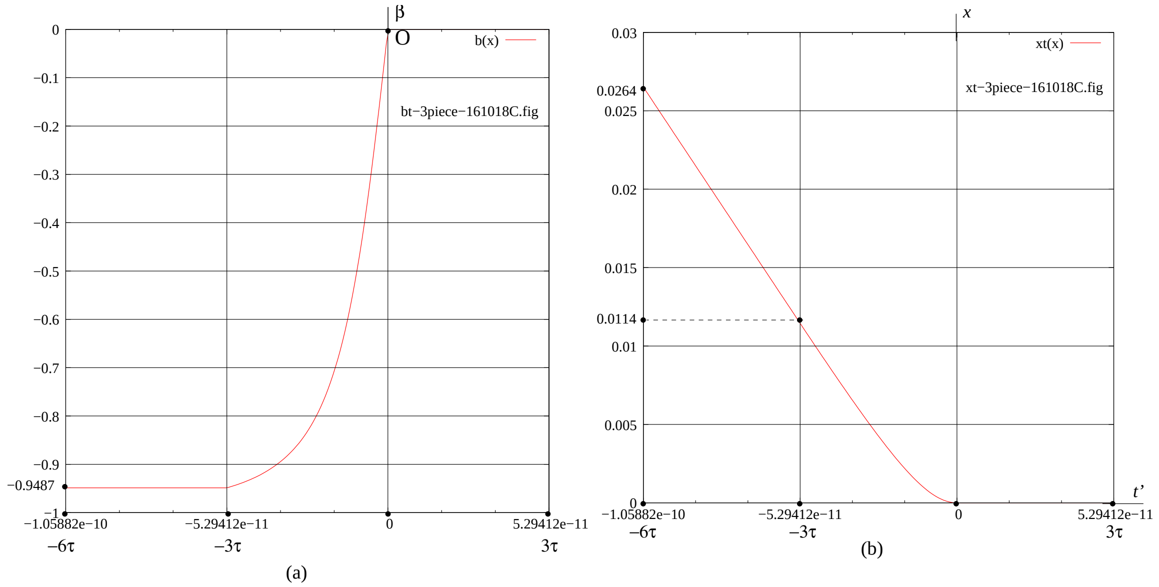
1. The velocity is now increased from $\beta_\infty = -0.707$ to $\beta_\infty = -0.949$, as in the last example in Sec.5.2.1. This value has been decided by setting $(t'/\tau) = 3$ in Eq.(39b). Hence, the previous time unit $\tau$, is now replaced by the new unit $3\tau$.
2. The acceleration is increased to 100 times of the previous value. That is $a_{\text{old}} = a_o = 0.17 \times 10^{18}\,\mathrm{m/s^2} \to a_{\text{new}} = 100a_o = 0.17 \times 10^{20}\,\mathrm{m/s^2}$.
3. The particle passes B2 at $t' = -6\tau$, reaches B1 at $t' = -3\tau$, when deceleration starts, reaches O at $t' = 0$, when deceleration stops. The particle stays at O from $t' = 0$ to $t' = +\infty$. Present time is $t = t' = 3\tau$.

Plots for the field lines are shown in Fig.14.

## 5.5 Lessons from the field lines of decelerating charge

We shall point out some important features of the field lines for the decelerating charge, with reference to Fig.15. Fig. (a) shows the bare field lines for the Case 2. Fig.(b) has two circles S1, S2 superimposed on the bare field lines, similar to what we had done in Sec. 5.3.

The circles represent two spherical surfaces S1 and S2, dividing the space into 3 zones: (i) *Coulomb* zone, i.e., the region *inside* the sphere S1; (ii) the *Deceleration* zone, i.e., the region *between* the spheres S1 and S2; (iii)



(a) (b)

Figure 12: *Plots of* $\beta(t')$ *and* $\tilde{x}(t')$ *corresponding to case II of deceleration*

the *Velocity* zone, i.e., the region *outside* the sphere S2.

1. The particle was moving with the constant velocity $c\beta_\infty = -1/\sqrt{2}\,c$ during its displacement B1 ← B2. The field lines are radial straight lines, making "steeper angles" $\phi_n$ with the X axis, as given by the Purcell condition. This uniform motion came to end at B1, at the time $t' = -\tau$, but the current time is $t' = \tau$. Hence there is a time gap of $2\tau$ between the end of uniform motion and now. Observers beyond the radius $2\tau c$ about B1 see the particle still in uniform motion with the constant velocity $c\beta_\infty = -1/\sqrt{2}\,c$. The sphere S2 of radius $R_2 = 2\tau c$ with centre at B1, marks the boundary beyond which the field is still the velocity field.

   However, the velocity field lines would emanate from the "present" location of the particle, had the particle been in constant uniform motion. Where is that "present" position? It is the point V, the virtual location of the particle, if the particle had been moving with the velocity $-c\beta_\infty$ unchanged.

   What will be the x-coordinate of V?

   The coordinate of the point B1 is $x(\mathrm{B1}) = (\sqrt{2} - 1)\tau$c. Let $\Delta x$ = distance the particle will move in the negative x direction in time $2\tau$, which equals $c\beta(-\tau) \times 2\tau = -1/\sqrt{2} \times 2\tau c = -\sqrt{2}\tau$. Hence, $x(V) = x(B1) + \Delta x = -\tau c$.

   Therefore, the velocity field lines seem to be emanating from V, at $x = -\tau c$, as shown in Fig 15.

2. The particle is sitting at O from $t' = 0$

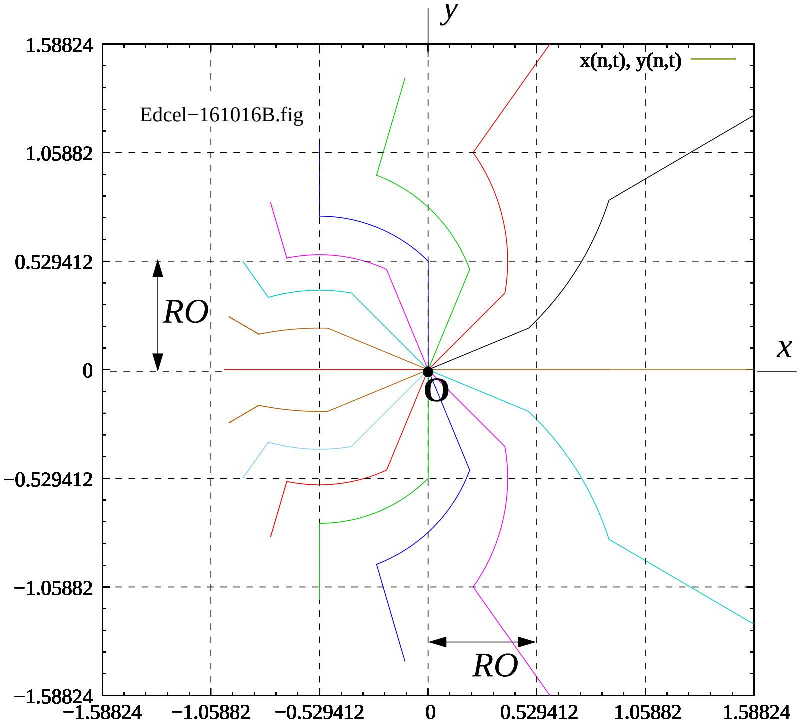


Figure 13: *Plots of* **E** *for case II of deceleration.*

to the present time $t = \tau$. The Coulomb field has, in this time, propagated up to a radius of R1 $= \tau c$. The sphere S1 marks that sphere. The field inside this is purely Coulomb.

## 5.6 Comparison between Acceleration field lines and Deceleration field lines

In Fig. 16 we have compared the plots of the Case I of *deceleration* with those of Case I of *acceleration* considered in Sec. 5.2.1, by superimposing the two sets of plots one upon another with a common origin O. The lines with arrowheads represent the decelerating case, and the ones without the accelerating case. They appear to match the predictions made in Sec.5.4.1.

However, there is one difference. The field lines for the deceleration case appear to be longer in the +x direction, and those for acceleration in the -x direction. This may lead to an erroneous conclusion that the field is stronger in the +x direction in the deceleration case and its opposite in the acceleration case.

To dispel this notion let us stress that the length of a field line has no relation to the field strength. Secondly, the furthest point on the field line receives its contribution from

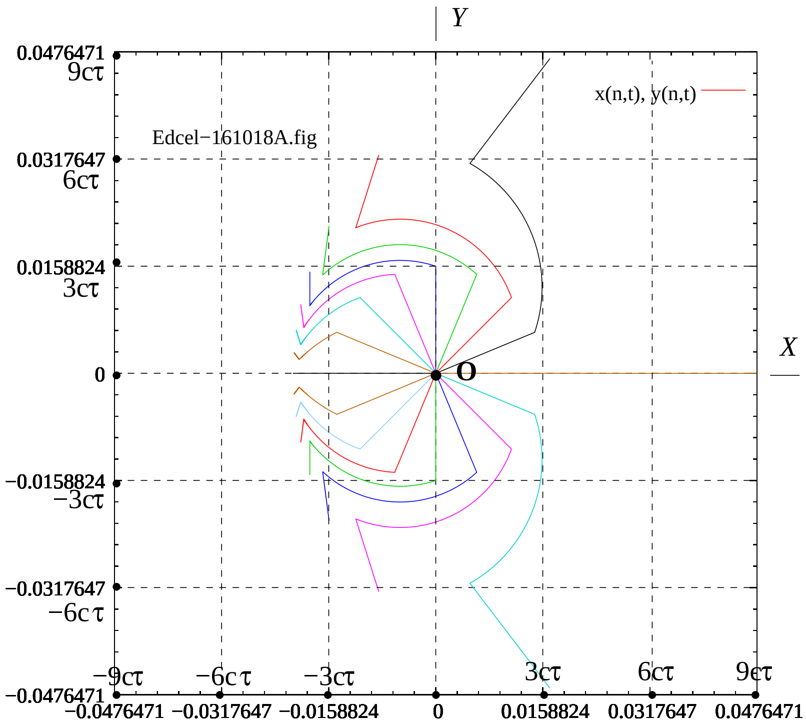


Figure 14: *Plots of* **E** *for case III of deceleration*

the corresponding retarded point, same as the furthest point B on the track of the particle for the given range of $R$ considered (see Fig.2 .)

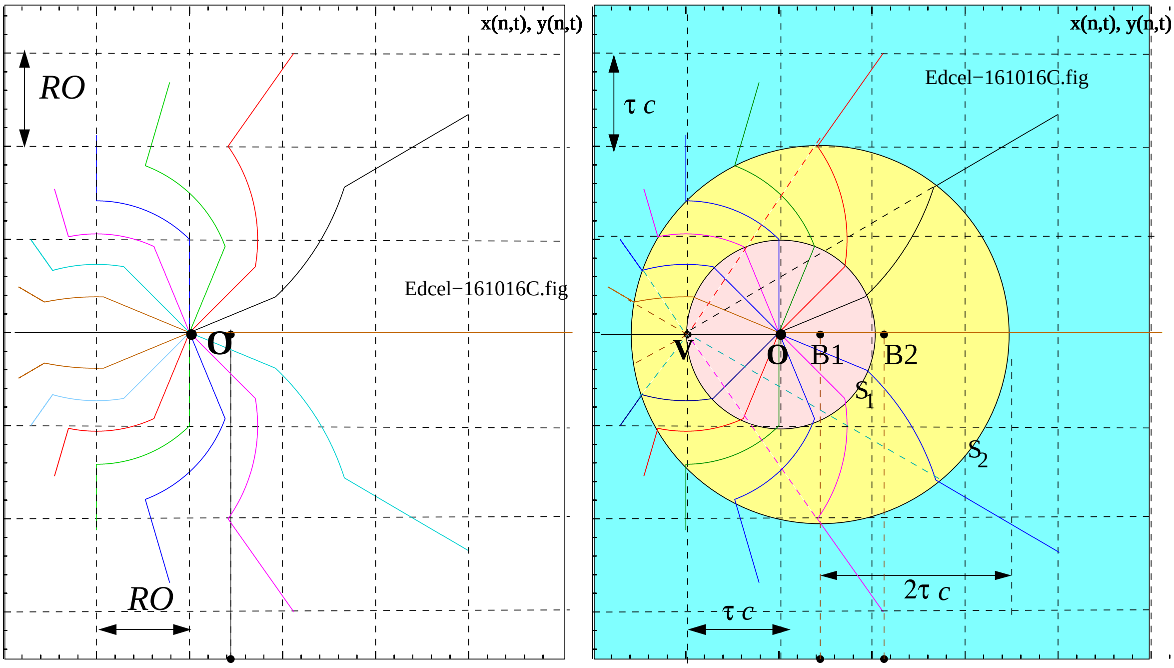
The track of the particle in the present case the $X$-axis. In the acceleration case the present location is at $x = c\tau$ and the furthest point is $x = 0$. In the deceleration case the present location is $x = 0$ and the furthest point is $x = c\tau$. Let us mark the furthest point on the field line as C. Going back to Eq. (13), we can now write the radius vector for the point C as

$$\begin{array}{lcll} \mathbf{r}_c & = & -\tau c\mathbf{i} + \mathbf{R}_c. & \text{acceleration} \\ \mathbf{r}_c & = & +\tau c\mathbf{i} + \mathbf{R}_c. & \text{deceleration} \end{array} \qquad (41)$$

The additive vectors $\mp\tau c\mathbf{i}$ can be considered as “bias vectors”. We have replotted the deceleration field lines by removing this bias, i.e., by plotting the bare CtF vector $\mathbf{R}(t')$, as shown in Fig.17. In this modified plot the feld lines on the x-axis have the same length on either side of the origin.

## 5.7 Gravity Effect?

According to Einstein’s Principle of Equivalence an accelerating frame of reference replicates gravity[11]. A particle in a frame of reference which is accelerating in the positive x direction feels the effect of gravity in the negative x direction. Gravity pulls down ev-



(a) (b)

Figure 15: *Features of the Field lines from decelerating charge*

erything, including light, forcing it to deviate from its straight line path. A cursory look at the plots of the **E** field may give the impression that the field lines are bending down, i.e., in the direction of the induced gravity, whenever the particle is in an accelerated state of motion[13] (for both acceleration and deceleration). In the absence of acceleration, the field lines betray their normal character, i.e., straight lines originating from the instantaneous location of the charge, whether at rest or in relativistic motion.

Whether the formulas for the EM fields from an accelerating charge, as obtained from Maxwell's equations, can be traced to the Principle of Equivalence is a question on which experts in General Relativity can enlighten us.

## 6 Specific Example 2: E field from Circular Motion

### 6.1 Expression for the E field

Referring to Fig. 18, we shall describe the configuration as follows. A charged particle carrying charge $q$ is moving in a circle of radius $a$ on the $XY$ plane with angular velocity $\omega$ in the *clockwise* direction. The observer P

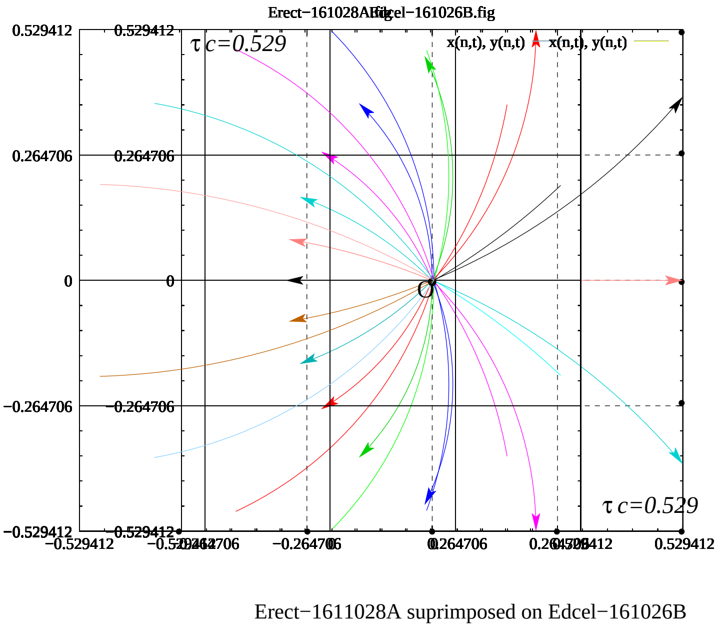


Figure 16: *Acceleration field lines (lines without arrows) superimposed on deceleration field field lines (lines with arrows) for comparison*

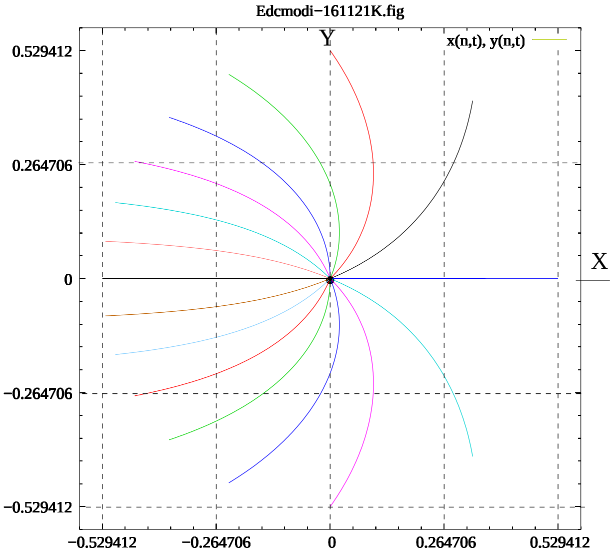


Figure 17: *Deceleration field lines without bias*

is sitting on this $XY$ plane, at a distance $r$ from the origin.

It should be noted that we have taken the rotational angular velocity negative, so that the angle $\theta$ that the velocity $\boldsymbol{\beta}$ makes with the $X$ axis, at the retarded point B is a positive acute angle. The same with the angle $\alpha$. We have chosen the coordinate axes in such a way that the observer A and the location of the particle at the present time $t$ (set equal to 0) lie on the $Y$ axis. With these choices of positive acute angles in the drawings derivations of trigonometrical relations becomes easier.

Our objective is to plot the $\mathbf{E}$ field as at the present time $t = 0$.

Following Eq. (4) the radius vectors $\tilde{\mathbf{r}} = \tilde{x}\mathbf{i} + \tilde{y}\mathbf{j}$ from the origin to the location Q of the particle at any arbitrary time $t$ is given by the equations

$$\tilde{x} = a \sin \omega t; \quad \tilde{y} = a \cos \omega t. \tag{42}$$

We need to evaluate the required quantities. Let $t'$ stand for the retarded time $t'$, B the location of the particle at $t'$, and $\mathbf{R}(t')$ the radius vector from B to P at the time $t$. Then

$$\left.\begin{array}{rcll}
\mathbf{R}(t,t') & = & r\,\mathbf{j} - \tilde{\mathbf{r}} = -a \sin \omega t' \mathbf{i} + (r - a \cos \omega t')\mathbf{j}. & (a) \\
R^2(t,t') & = & (a \sin \omega t')^2 + (r - a \cos \omega t')^2 = a^2 + r^2 - 2ra\, \cos \omega t'. & (b) \\
\text{where} \quad t & = & t' + \frac{R(t,t')}{c} = t' + \frac{1}{c}\sqrt{a^2 + r^2 - 2ra\, \cos \omega t'}, & (c) \\
c\boldsymbol{\beta} & = & \dot{\tilde{x}}\,\mathbf{i} + \dot{\tilde{y}}\,\mathbf{j} = \omega a(\cos \omega t'\,\mathbf{i} - \sin \omega t'\,\mathbf{j}) & (d) \\
c\beta & = & \omega a; \quad \dot{\beta} = 0; \quad \theta = -\omega t'; \quad \dot{\theta} = -\omega, & (e) \\
\mathbf{R} \cdot c\boldsymbol{\beta} & = & -r\omega a \sin \omega t', & (f) \\
\cos \alpha & = & \dfrac{\mathbf{R} \cdot c\boldsymbol{\beta}}{Rc\beta} = -\dfrac{r\, \sin \omega t'}{\sqrt{a^2 + r^2 - 2ra\, \cos \omega t'}} = \dfrac{r\, \sin \theta}{\sqrt{a^2 + r^2 - 2ra\, \cos \theta}}, & (g)
\end{array}\right\} \tag{43}$$

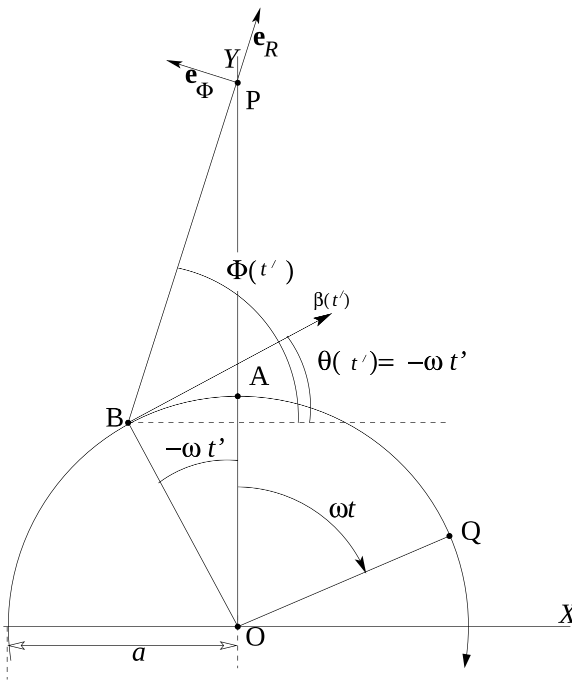


Figure 18: *Charge q in circular motion*

We now go back to Eqs.(11) and insert the values just determined: $\dot{\beta} = 0; (R/c)\beta\dot{\theta} = -(R/a)\beta^2$. Set $\frac{q}{4\pi\varepsilon_0} = 1$, $a = 1$.

$$\begin{aligned} E_R &= \frac{(1-\beta^2)(1-\beta\ \cos\alpha)}{\kappa^3 R^2} \\ E_\Phi &= \frac{\frac{R}{c}\{\dot{\beta}\ \sin\alpha+\beta\,\dot{\theta}(\beta-\cos\alpha)\}+(1-\beta^2)\beta\sin\alpha}{\kappa^3 R^2} \\ &= \frac{-R\beta^2(\beta-\cos\alpha)+(1-\beta^2)\beta\sin\alpha}{\kappa^3 R^2}. \end{aligned} \tag{44}$$

## 6.2 Equations for the Field Lines

We shall specialize the set of equations (25) for this special case. In this case $\dot{\beta} = 0$. Hence, from Eq. (8b):

$$\begin{aligned} &\dot{\boldsymbol{\beta}} \cdot \mathbf{e}_\Phi = \beta\dot{\theta}\cos\alpha \\ &\frac{d\alpha}{dR} = \frac{1}{c}\gamma^2\dot{\theta}(1-\beta\cos\alpha). \quad \text{[from (25 a)]} \\ &(1-\beta\cos\alpha)^{-1}\,d\alpha = \frac{1}{c}\gamma^2\dot{\theta}dR \\ &= \frac{1}{c}\gamma^2\,d\theta\,(dR/dt') \\ &= -\gamma^2\,d\theta. \quad \text{[using Eq. (1)],} \end{aligned} \tag{45}$$

Let us be reminded that the independent variables are $(R, \alpha)$. However, we have $\theta$ on the right side. It is a function of $R$ through the relation $\theta = \theta(t')$ and $t' = t - R/c =$

$-R/c$. From integral table

$$\int \frac{d\alpha}{1-\beta\cos\alpha} = \frac{2}{\sqrt{1-\beta^2}} \times \tan^{-1}\left(\sqrt{\frac{1+\beta}{1-\beta}}\tan\frac{\alpha}{2}\right). \quad (46)$$

Therefore, integrating both sides of the last line of Eq. (45)

$$\sqrt{\frac{1+\beta}{1-\beta}}\tan\frac{\alpha}{2} = \tan\left[-\frac{\gamma\theta}{2}+\frac{k}{2\gamma}\right]. \quad (47)$$

where $k$ is a constant. Therefore,

$$\begin{aligned}\tan\tfrac{\alpha}{2} &= \sqrt{\tfrac{1-\beta}{1+\beta}}\tan\left[-\tfrac{\gamma\theta(t')}{2}+\tfrac{k}{2\gamma}\right] \\ &= \sqrt{\tfrac{1-\beta}{1+\beta}}\tan\left[-\tfrac{\gamma\theta(t-R/c)}{2}+\tfrac{k}{2\gamma}\right].\end{aligned} \quad (48)$$

At the present location of the particle $R=0$. Setting this in (48), and going back to the boundary condition given in Eq. (25b), we get for the $n$-th field line,

$$\begin{aligned}-\tfrac{\gamma\theta(t)}{2}+\tfrac{k}{2\gamma} &= \tfrac{\pi n}{N}. \\ \text{Or,}\quad \tfrac{k}{2\gamma} &= \tfrac{\pi n}{N}+\tfrac{\gamma\theta(t)}{2}.\end{aligned} \quad (49)$$

Depositing this value in (48) we get the complete relation between $R$ and $\alpha$.

$$\tan\frac{\alpha}{2} = \sqrt{\frac{1-\beta}{1+\beta}} \times \tan\left[-\frac{\gamma\theta(t-R/c)}{2}+\frac{\pi n}{N}+\frac{\gamma\theta(t)}{2}\right]. \quad (50)$$

As seen from Fig. 18, $\theta(t-R/c) = -\omega(t-R/c) = -c\beta/a\times(t-R/c)$. Substituting this in (50) we get

$$\tan\frac{\alpha}{2} = \sqrt{\frac{1-\beta}{1+\beta}}\tan\left[-\frac{\gamma\beta R}{2a}+\frac{\pi n}{N}\right]. \quad (51)$$

Fig. 19 shows an arbitrary field line $\Omega$ coming from a Synchrotron accelerator[14] which accelerates charged particles, like electrons, protons, in circular motion. We shall obtain the parametric equation for the field line.

$$\begin{array}{ll}\tilde{x}(t') = a\,\sin\omega t';\quad \tilde{y}(t') = a\,\cos\omega t'. & (a)\\ c\beta_x(t') = \omega a\cos\omega t';\quad c\beta_y(t') = -\omega a\,\sin\omega t'. & (b)\\ \omega = \beta c/a;\quad \omega t' = (\beta c/a)(-R/c) = -\beta R/a. & (c)\\ \tan\theta(t') = \beta_y(t')/\beta_x(t') = -\tan\omega t' = \tan(\beta R/a). & (d)\\ \theta(t') = \beta R/a = -\omega t'. & (e)\end{array} \quad (52)$$

The required parametric equations are given in Eqs. (25d,e), which we adopt for the present case.

$$\begin{aligned}x &= -a\,\sin(\beta R/a) + R\cos[(\beta R/a)+\alpha],\\ y &= a\,\cos(\beta R/a) + R\sin[(\beta R/a)+\alpha],\end{aligned} \quad (53)$$

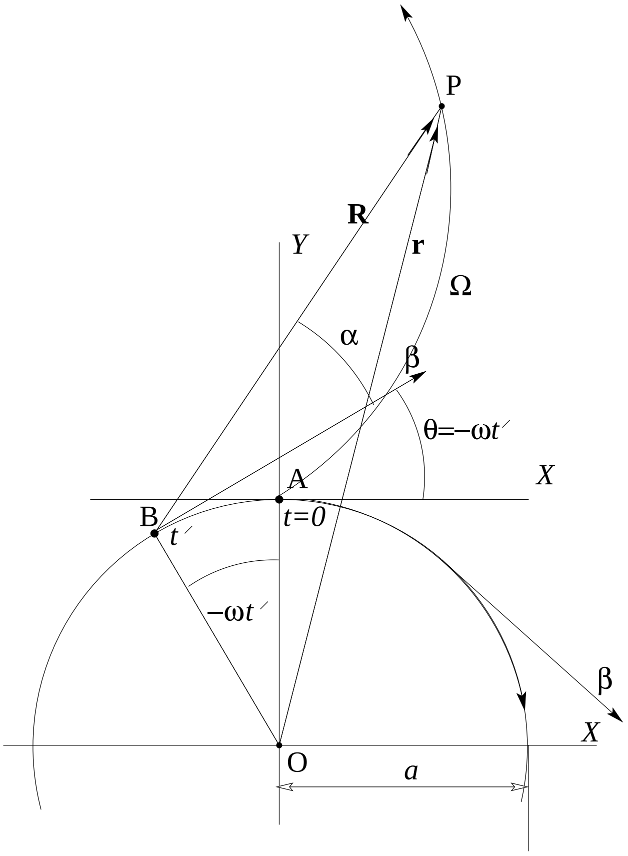


Figure 19: *Field line from a synchrotron accelerator*

with $\alpha$ determined from Eq. (50), and $R \geq 0$.

### 6.3 Plot of the Field Lines

We shall plot several examples of field lines from a particle moving in a circle of radius $a = 1$ meter, with relativistic velocity $c\beta$. For each example we shall plot $N$ field lines originating as spring lines from the instantaneous present location of the particle, marked A, at the present time which has been taken as $t = 0$. The values of $\beta, N$, and the range of the parameter $R$, for the 7 exercises we have undertaken, are specified in columns 2,3,4 of the table below. The plot label includes the values of $\beta$ (b5 for $\beta$=0.5, b95 for $\beta$=0.95, etc), N, the number of "samples" (400, 1000, 3000), and the date when the plot was made.

If we denote the angular coordinate of the particle at the retarded time $t'$ as $\phi = \omega t'$, measured from the Y axis in the direction of its motion, i.e., clockwise, then it is seen from (52e) that $R = -\phi/\beta$. Hence $\pi$ can be a convenient unit of the linear distance

$R$. In all plots we have use $\pi$ (in meters), as an alternative the unit of displacement, and set the tic marks along the $X$ and $Y$ axes in integral multiples of $\pi$. These plots are shown in Figs.20 - 26.

In the final plot, shown in Fig.27, $\beta \to 1$, and $R = -\phi$, so that positive range of $R$ will correspond with the same numerical range of $\phi$, but with negative values.

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Ex. | $\beta$ | $N$ | Range | Plot label | Figure label |
| 1 | 0.1 | 16 | [0:60$\pi$] | Qcir-b1-16-800-160720A.fig | Fig. 20 |
| 2 | 0.5 | 16 | [0:13$\pi$] | Qcir-b5-16-400-160706.fig | Fig. 21 |
| 3 | 0.5 | 8 | [0:6$\pi$] | Qcir-b5-16-400-160717.fig | Fig. 22 |
| 4 | 0.8 | 8 | [0:6$\pi$] | Qcir-b8-8-1000-160706.fig | Fig. 23 |
| 5 | 0.9 | 8 | [0:5$\pi$] | Qcir-b9-8-3000-160707.fig | Fig. 24 |
| 6 | 0.9 | 4 | [0:5$\pi$] | Qcir-b9-4-3000-160707.fig | Fig. 25 |
| 7 | 0.95 | 8 | [0:5$\pi$] | Qcir-b95-8-3000-160707.fig | Fig. 26 |

(54)

## 6.4 Notable features and Lessons from the Field Lines; Synchrotron Radiation

As we progress from Fig. 20 to 26, we witness changes in the field pattern as the velocity moves up from "slow" *non-relativistic* to "very fast" *ultra-relativistic*.

When a charged particle moves in a circle with non-relativistic velocity, it emits electromagnetic field, and hence radiation, which is nearly isotropic, but oscillates with the frequency of rotation. At ultra-relativistic velocity, it emits what is often referred to as *Synchrotron radiation*. Such radiation can be "seen" in a proton synchrotron in famous Nuclear Research Laboratories[14]. However, synchrotron radiation is also an Astronomical phenomenon, and has been observed in the light coming from pulsars and the Crab nebula[15].

Coming back to the plots of the **E** field, which is our current interest, we go to Fig. 20 to see the "slow example", the charged particle moving in a circle of radius $a$, with velocity $c/10$. We have shown 16 field lines issuing out from the present location A of the particle, at equal angular intervals of $2\pi/16$. The charge at the source being positive, **E** vector is pointing away from the source, as indicated with arrowheads. (We shall avoid arrows in subsequent plots.) As the field propagates, it oscillates with the same angular frequency

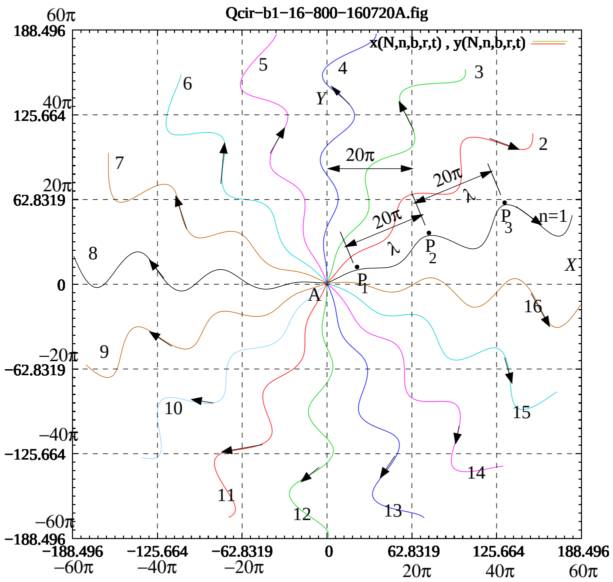


Figure 20: *Field line:* $\beta$*=0.1, N=16,* $R\,\epsilon\,[0:60\pi]$

$\omega$ as that of rotation and hence carries wavelength $\lambda = 2\pi c/\omega$. Now, $\omega = \beta c/r = \beta c$, since the radius of the circle is $a = 1$ m. Hence, $\lambda = 2\pi/\beta$. For $\beta = 1/10$, $\lambda = 20\pi$, also manifested in the figure. In one cycle the "crests" at $\mathrm{P}_1, \mathrm{P}_2$ are displaced to $\mathrm{P}_2, \mathrm{P}_3$, respectively, each displacement being equal to one wavelength $20\pi$.

Figs.21 and 22 correspond to the same values of $\beta, N$, equal to 0.5, 16 respectively. However, the range of $R$ used in the latter is 1/3 of the former and presents a magnified picture of the field lines close to the synchrotron.

In the same way, even though Figs.24 and 25 correspond to the same value of $\beta$, equal to 0.9, the former uses 8 field lines, whereas the latter only 4. As a consequence the latter presents a clearer picture of the formation of the sharp zig zag kinks. As we have shown on the field line $n = 2$, it encounters sharp reversals in its direction (kinks) at the points G and H, in quick succession, and then proceeds onward. If we watch the kink regions of $n = 1, \cdots, N$, in Figs.24 and 25, they appear to merge or smear into a spiral band.

Fig. 26 presents an ultra-relativistic picture: $\beta = 0.95$. The particle moves in the same circle with velocity $0.95c$. Here the the formation of the spiral band is conspicuous.

We have plotted the spiral band in Fig. 27(a), by first obtaining its parametric equation. First note that at the kink points $E_\Phi = 0$. Assuming that the observer is far away, the field varies as $1/R$. Hence from Eq. (44), $\beta - \cos\alpha = 0$. For the ultra-relativistic case $\beta \to 1 \Rightarrow \alpha = 0$.

Now we go back to Eq. (53), set $a = 1$, $\alpha =$

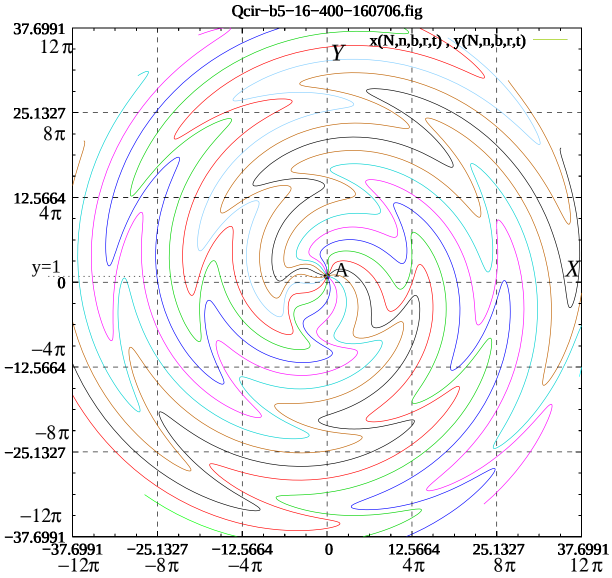


Figure 21: *Field line:* $\beta$*=0.5, N=16,* $R\,\epsilon\,[0:13\pi]$

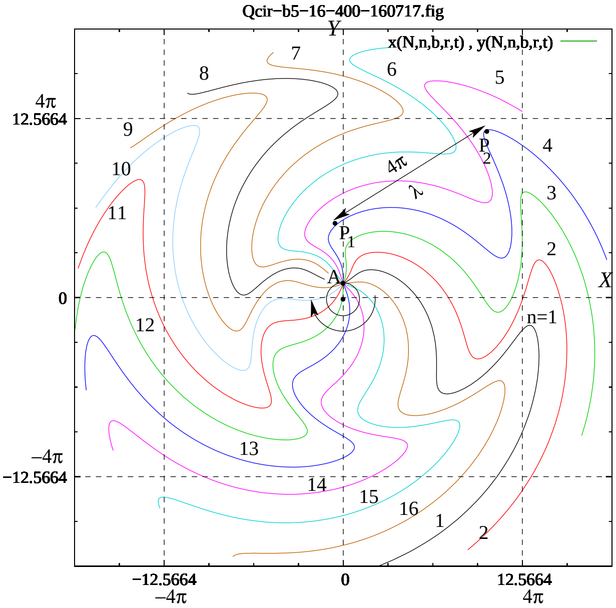


Figure 22: *Field line:* $\beta$*=0.5, N=16,* $R\,\epsilon\,[0:6\pi]$

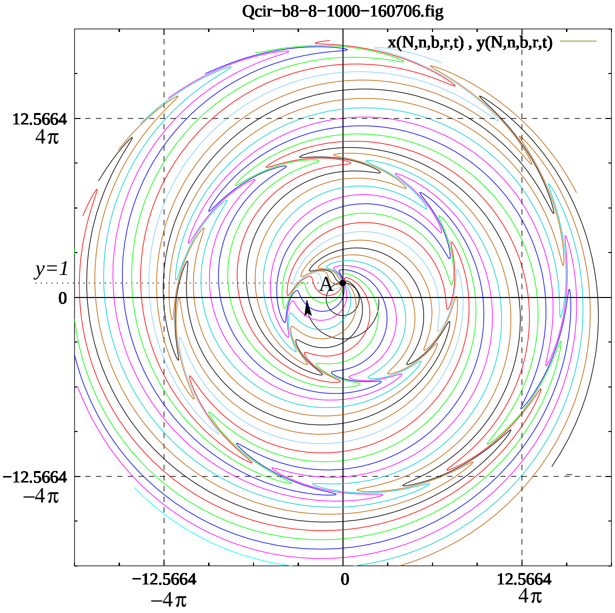


Figure 23: *Field line:* $\beta$*=0.8, N=8,* $R\,\epsilon\,[0:6\pi]$

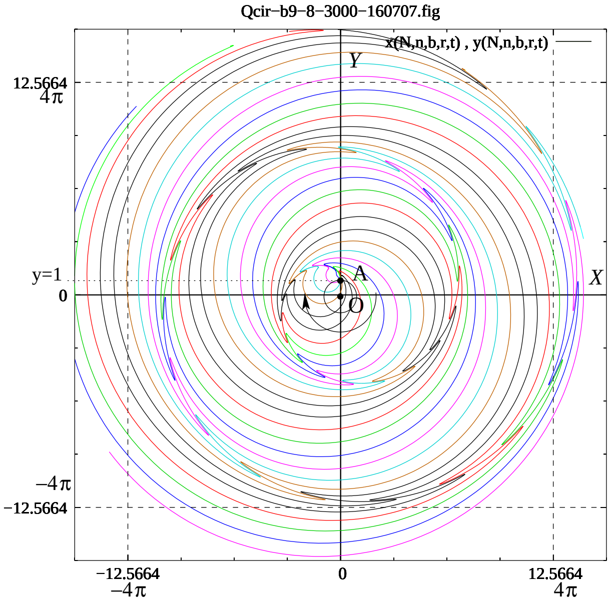


Figure 24: *Field line:* $\beta$*=0.9, N=8,* $R\,\epsilon\,[0:5\pi]$

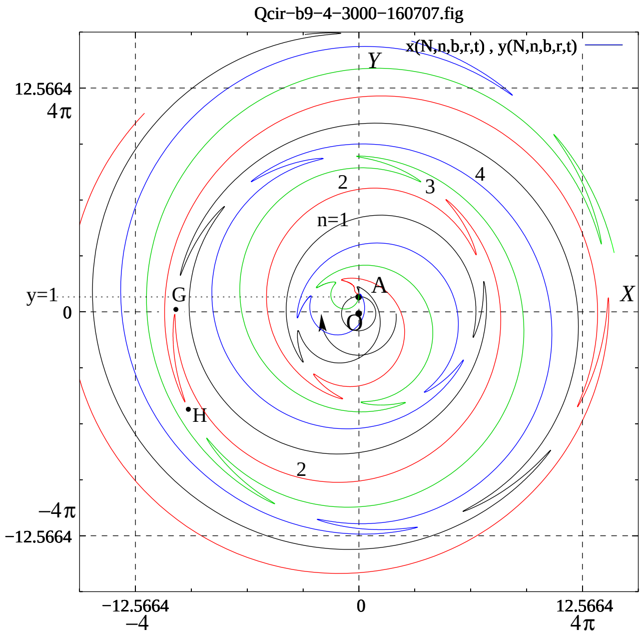


Figure 25: *Field line:* $\beta$*=0.9, N=4,* $R\,\epsilon\,[0:5\pi]$

0, $\beta = 1$. Adopt a new variable $\phi = \omega t'$. From Eq. (52c) $R = -\phi$. Change the variable from $R$ to $\phi$ to get the locus of the kink points as

$$x = \sin\phi - \phi\cos\phi;\ \ y = \cos\phi + \phi\sin\phi. \quad \phi \le 0. \tag{55}$$

This is the equation of the *involute* of a unit circle, which can be visualized as the path of the free end P of a rope which is unwound from a circular cylinder of unit radius to which it had been tightly wound[16], as illustrated in Fig. 27b. The locus of P can be easily obtained from the diagram as

$$x = \sin\theta - \theta\cos\theta;\ \ y = \cos\theta + \theta\sin\theta. \quad \theta \ge 0. \tag{56}$$

which is the same as (55), except that the two angles $\phi$ and $\theta$, both measured from the $Y$ axis, are clockwise and anti-clockwise respectively. The spiral shown in Fig. 27(a) is a plot of (55), but drawn with a thick line (line width = 4).

Fig. 27(a) can provide a valuable insight into the nature of synchrotron radiation. Let the particle be located at some point Q on the circle at some retarded time $-t'$, corresponding to the present time $t = 0$. Let there be observers $P_1, P_2$, K, all of them located on the plane of the synchrotron. Draw the directed straight line $QP_1P_2$ tangential to the circle, in the direction of the instantaneous velocity $\boldsymbol{\beta}c$ at $-t'$, such that the points $P_1, P_2$ lie on the two nearest arms of the spiral. These two points receive sharp pulses at the present time, but coming from the location Q at re-

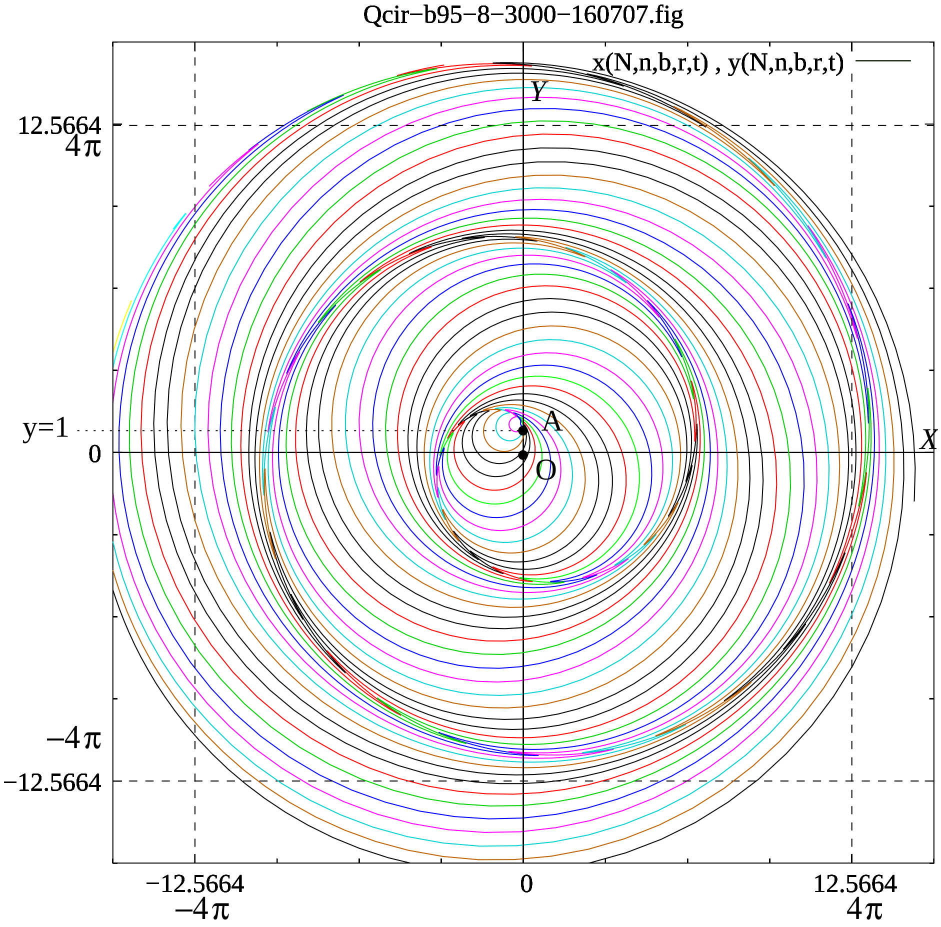


Figure 26: *Field line:* $\beta$*=0.95, N=8,* $R \epsilon [0 : 5\pi]$

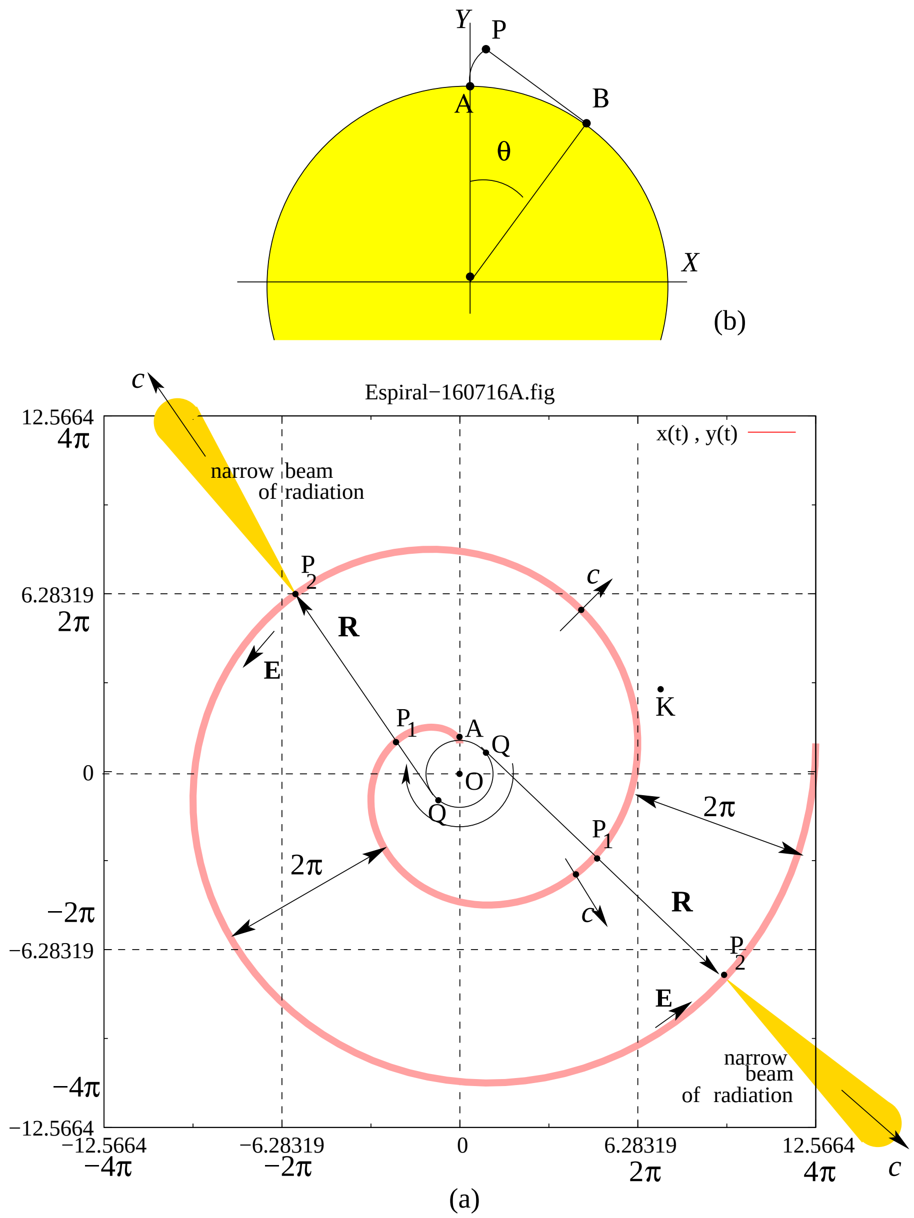


Figure 27: *Concentration of* **E** *field along a spiral, indicating synchrotron radiation*

tarded times $-t'$ and $-t' - T$ respectively, where $T$ is the period of circular motion.

Consider an arbitrary observer K sitting somewhere on this plane. Like the observers at $P_1, P_2$, he receives periodic radiation pulses at those moments when the spiral, turning with the same angular velocity $\omega$, intercepts the point K.

Jackson has shown the same spiral band with the following observation[17], “The form of the angular distribution of radiation ... corresponds to a narrow cone or search light beam of radiation directed along the instantaneous velocity vector of the charge.” He has presented a picture of this search light beam[18]. This picture of the searchlight beam can also be found in many other references[15].

We shall summarise some of the lessons we derive from our exercise for the special case $\beta \to 1$.

1. An observer located on the plane of the synchrotron at a point K will receive a burst of radiation pulse every time the spiral band intercepts the point K.
2. These radiation pulses will reach him at the regular time interval $T$, coming from the retarded locations of the particle corresponding to its velocity vector $c\boldsymbol{\beta}$ pointing towards him. A very important example of this is the radiation coming from astronomical objects, like pulsars, and the Crab Nebula.
3. The radius vector $\mathbf{R}$ from the retarded location Q to its corresponding field point P at the present time, has the following properties: (a) it is tangent to the synchrotron circle, (b) it is perpendicular to the spiral arc at the point P.
4. Hence, the Pointing’s vector, which represents the electromagnetic energy flux density, is perpendicular to every segment of the spiral arm at each of its points.
5. Hence, the arm of the spiral, at any point P, is expanding outward along the radius vector $\mathbf{R}$ with the velocity $c$, even as it rotates with the angular velocity $\omega$.

The reader should prove the statements made in items 3,4,5. He can prove the property 3, using the equation of the involute as given in (56). For proving properties 3 and 4, he can use theorems and formulas of classical electrodynamics.

# A Appendix A

## A.1 New Differentials in terms of Old Ones

We shall go back to the definition of the new coordinates $(R, \alpha)$, as given in Eqs.(14), and construct the required differentials:

$$\begin{aligned} dx &= \tfrac{\partial x}{\partial R} dR + \tfrac{\partial x}{\partial \alpha} d\alpha \\ dy &= \tfrac{\partial y}{\partial R} dR + \tfrac{\partial y}{\partial \alpha} d\alpha \end{aligned} \tag{A. 1}$$

Using Crammer’s formula:

$$dR = \frac{1}{J}\begin{vmatrix} dx & \frac{\partial x}{\partial \alpha} \\ dy & \frac{\partial y}{\partial \alpha} \end{vmatrix}; \quad d\alpha = \frac{1}{J}\begin{vmatrix} \frac{\partial x}{\partial R} & dx \\ \frac{\partial y}{\partial R} & dy \end{vmatrix} \tag{A. 2}$$

where $J$ is the Jacobian of the transformation. To find the derivatives and $J$, we proceed as follows, going back to Eqs. (14).

$$\begin{array}{rrcl} & \frac{\partial x}{\partial R} & = & \frac{\partial \tilde{x}}{\partial t'}\,\frac{\partial t'}{\partial R} + \cos(\theta+\alpha) - R\,\sin(\theta+\alpha)\,\dot{\theta}\,\frac{dt'}{dR} \\ & & = & c\beta_x(-1/c) + \cos(\theta+\alpha) - R\,\sin(\theta+\alpha)\,\dot{\theta}\,(-1/c) \\ & & = & -\beta_x + \cos(\theta+\alpha) + (R\dot{\theta}/c)\sin(\theta+\alpha). \\ & \frac{\partial x}{\partial \alpha} & = & -R\,\sin(\theta+\alpha). \\ \text{Similarly,} & \frac{\partial y}{\partial R} & = & -\beta_y + \sin(\theta+\alpha) - (R\dot{\theta}/c)\cos(\theta+\alpha). \\ & \frac{\partial y}{\partial \alpha} & = & R\,\cos(\theta+\alpha). \end{array} \tag{A. 3}$$

$$\begin{array}{l} J = \begin{vmatrix} \frac{\partial x}{\partial R} & \frac{\partial x}{\partial \alpha} \\ \frac{\partial y}{\partial R} & \frac{\partial y}{\partial \alpha} \end{vmatrix} \\ J_1 = \begin{vmatrix} -\beta_x & -R\sin(\theta+\alpha) \\ -\beta_y & R\cos(\theta+\alpha) \end{vmatrix} = -\beta_x R(\cos(\theta+\alpha) - \beta_y\sin(\theta+\alpha)) \\ = -\beta_x R(\cos\Phi + \beta_y\sin\Phi) = -R\,\boldsymbol{\beta}\cdot\mathbf{e}_R = \beta\,\mathbf{e}_\beta\cdot\mathbf{e}_R = -R\beta\cos\alpha. \quad \text{See Eq. } 7(a). \\ J_2 = \begin{vmatrix} \cos(\theta+\alpha) & -R\sin(\theta+\alpha) \\ \sin(\theta+\alpha) & R\cos(\theta+\alpha) \end{vmatrix} = R \\ J_3 = \begin{vmatrix} (R\dot{\theta}/c)\sin(\theta+\alpha) & -R\sin(\theta+\alpha) \\ -(R\dot{\theta}/c)\cos(\theta+\alpha) & R\cos(\theta+\alpha) \end{vmatrix} = 0 \\ J = (1-\beta\,\cos\alpha)\,R. \end{array} \tag{A. 4}$$

$$dR = \frac{1}{J}\left(\frac{\partial y}{\partial \alpha}dx - \frac{\partial x}{\partial \alpha}dy\right) = \frac{R\,\cos(\theta+\alpha)\,dx + R\,\sin(\theta+\alpha)dy}{(1-\beta\,\cos\alpha)\,R} = \frac{d\mathbf{r}\cdot\mathbf{e}_R}{1-\beta\,\cos\alpha}. \tag{A. 5}$$

$$\begin{array}{l} d\alpha = \frac{1}{J}\left(\frac{\partial x}{\partial R}dy - \frac{\partial y}{\partial R}dx\right) \\ = \frac{\{-\beta_x + \cos(\theta+\alpha) + (R\dot{\theta}/c)\sin(\theta+\alpha)\}dy - \{-\beta_y + \sin(\theta+\alpha) - (R\dot{\theta}/c)\cos(\theta+\alpha)\}dx}{(1-\beta\,\cos\alpha)\,R}. \end{array} \tag{A. 6}$$

We shall simplify Eq. (A. 6), using (6). A part of the numerator simplifies as follows.

$$\{-\beta_x\,dy + \beta_y\,dx\} + \{\cos(\theta+\alpha)\,dy - \sin(\theta+\alpha)dx\} = (1-\beta\cos\alpha)\,(d\mathbf{r}\cdot\mathbf{e}_\Phi) - (\beta\,\sin\alpha)\,(d\mathbf{r}\cdot\mathbf{e}_R). \tag{A. 7}$$

*Proof:* The RHS of (A. 7)

$$\begin{aligned}
&= d\mathbf{r}\cdot\mathbf{e}_\Phi - \beta[\cos\alpha\,(d\mathbf{r}\cdot\mathbf{e}_\Phi) + \sin\alpha\,(d\mathbf{r}\cdot\mathbf{e}_R)] \\
&= d\mathbf{r}\cdot\mathbf{e}_\Phi - \beta[\cos\alpha\,\{-\sin(\theta+\alpha)dx + \cos(\theta+\alpha)dy\} \quad \text{[from (6b) ]} \\
&\quad + \sin\alpha\,\{\cos(\theta+\alpha)dx + \sin(\theta+\alpha)dy\}] \quad \text{[from (6a) ]} \\
&= d\mathbf{r}\cdot\mathbf{e}_\Phi - \beta[\{-\cos\alpha\,\sin(\theta+\alpha) + \sin\alpha\,\cos(\theta+\alpha)\}dx \\
&\quad +\{\cos\alpha\,\cos(\theta+\alpha) + \sin\alpha\,\sin(\theta+\alpha)\}dy] \\
&= d\mathbf{r}\cdot\mathbf{e}_\Phi + \beta[\sin\theta\,dx - \cos\theta\,dy] \\
&= d\mathbf{r}\cdot\mathbf{e}_\Phi + \beta_y dx - \beta_x dy \\
&= -\sin(\theta+\alpha)dx + \cos(\theta+\alpha)dy + \beta_y dx - \beta_x dy
\end{aligned}$$

Q.E.D.

The remaining part of the numerator in Eq. (A. 6) is

$$\begin{aligned}
&(R\dot{\theta}/c)\{\sin(\theta+\alpha)dy + \cos(\theta+\alpha)dx\} \\
&= (R\dot{\theta}/c)\,(d\mathbf{r}\cdot\mathbf{e}_R).
\end{aligned} \tag{A. 8}$$

Hence, from (A. 6),

$$\begin{aligned}
&d\alpha = \tfrac{1}{R}\left[(d\mathbf{r}\cdot\mathbf{e}_\Phi) + (1-\beta\cos\alpha)^{-1}\right. \\
&\left.\times\{(R\dot{\theta}/c) - \beta\sin\alpha\}(d\mathbf{r}\cdot\mathbf{e}_R)\right].
\end{aligned} \tag{A. 9}$$

## A.2 Differential Equation for the Field Line

As shown in Fig. 2(b), $d\mathbf{r}$ is the displacement vector from P to Q, and $ds$ is the corresponding infinitesimal arc length. The differential equations we propose to write are based on the following equations.

$$\begin{aligned}
d\mathbf{r}\cdot\mathbf{e}_R &= \tfrac{\mathbf{E}\cdot\mathbf{e}_R}{E}ds = \tfrac{E_R}{E}ds; \\
d\mathbf{r}\cdot\mathbf{e}_\Phi &= \tfrac{\mathbf{E}\cdot\mathbf{e}_\Phi}{E}ds = \tfrac{E_\Phi}{E}ds.
\end{aligned} \tag{A. 10}$$

Inserting these in (A. 5) and (A. 9) we get

$$\begin{aligned}
&dR = \tfrac{d\mathbf{r}\cdot\mathbf{e}_R}{1-\beta\ \cos\alpha} = \tfrac{E_R}{E(1-\beta\ \cos\alpha)}ds. && (a) \\
&d\alpha = \tfrac{1}{R}\left[(d\mathbf{r}\cdot\mathbf{e}_\Phi) + (1-\beta\cos\alpha)^{-1}\right. \\
&\left.\times\{(R\dot{\theta}/c) - \beta\sin\alpha\}(d\mathbf{r}\cdot\mathbf{e}_R)\right] \\
&= \tfrac{1}{RE}\left[E_\Phi + (1-\beta\cos\alpha)^{-1}\right. \\
&\left.\times\{(R\dot{\theta}/c) - \beta\sin\alpha\}(E_R)\right] ds. && (b)
\end{aligned} \tag{A. 11}$$

We go back to the expression for $E_R$ ad $E_\Phi$ in Eq. (11), rewrite them in a reduced from,

$$\begin{aligned}
&\tfrac{\kappa^3 R^2}{\tilde{q}}E_R = (1-\beta^2)(1-\beta\ \cos\alpha) && (a) \\
&\tfrac{\kappa^3 R^2}{\tilde{q}}E_\Phi = \frac{R}{c}\{\dot{\beta}\ \sin\alpha + \beta\,\dot{\theta}(\beta - \cos\alpha)\} \\
&+(1-\beta^2)\beta\sin\alpha && (b)
\end{aligned} \tag{A. 12}$$

We go back to (A. 11a), and using (A. 12a) get

$$\tfrac{\kappa^3 R^2 E}{\tilde{q}}\tfrac{dR}{ds} = \tfrac{(1-\beta^2)(1-\beta\ \cos\alpha)}{(1-\beta\ \cos\alpha)} = 1-\beta^2. \tag{A. 13}$$

Similarly, going back to (A. 11b), and using (A. 12a,b) we get

$$\frac{\kappa^3 R^2 E}{\tilde{q}}\frac{d\alpha}{ds} = \frac{1}{R}\left[\frac{R}{c}\{\dot{\beta}\,\sin\alpha + \beta\,\dot{\theta}(\beta - \cos\alpha)\} + (1-\beta^2)\beta\sin\alpha\right.$$

$$\left. +(1-\beta\cos\alpha)^{-1}\{(R\dot{\theta}/c) - \beta\sin\alpha\}(1-\beta^2)(1-\beta\,\cos\alpha)\right] \qquad \text{(A. 14)}$$

$$= \frac{1}{c}\left[\dot{\beta}\,\sin\alpha + \beta\,\dot{\theta}(\beta - \cos\alpha) + \dot{\theta}(1-\beta^2)\right] = \frac{1}{c}\left[\dot{\beta}\,\sin\alpha - \beta\,\dot{\theta}\cos\alpha + \dot{\theta}\right]$$

Finally, dividing the Eq. (A. 14) with (A. 13) we get the required differential equation:

$$\frac{d\alpha}{dR} = \frac{1}{c}\gamma^2[\dot{\beta}\,\sin\alpha - \beta\,\dot{\theta}\cos\alpha + \dot{\theta}], \quad \text{(A. 15)}$$

where $\gamma^2 = 1/(1-\beta^2)$. See the line following Eq. (10).

It is seen from (8) that

$$\dot{\boldsymbol{\beta}}\cdot\mathbf{e}_{\Phi} = -\dot{\beta}\,\sin\alpha + \beta\,\dot{\theta}\,\cos\alpha. \qquad \text{(A. 16)}$$

Hence, the differential equation for the field lines takes the following form, written by Tsien:

$$\frac{d\alpha}{dR} = \frac{1}{c}\gamma^2[\dot{\theta} - \dot{\boldsymbol{\beta}}\cdot\mathbf{e}_{\Phi}]. \qquad \text{(A. 17)}$$

## A.3 Boundary Condition

From Eq. (21) we shall prove the identity:

$$\begin{array}{lrcll} \text{If,} & \tan(\phi_n+\theta) & = & \frac{\sin(\theta+\alpha_n)-\beta\sin\theta}{\cos(\theta+\alpha_n)-\beta\cos\theta}, & (a) \\ \text{then,} & \tan\phi_n & = & \frac{\sin\alpha_n}{\cos\alpha_n-\beta}. & (b) \end{array} \qquad \text{(A. 18)}$$

*Proof*: Let us write

$\mu = \phi_n + \theta.$ Then from (21)

$\tan\mu = \frac{\sin(\theta+\alpha_n)-\beta\sin\theta}{\cos(\theta+\alpha_n)-\beta\cos\theta}$

$\tan\phi_n = \tan(\mu-\theta) = \frac{\tan\mu-\tan\theta}{1+\tan\mu\tan\theta} = \frac{\frac{\sin(\theta+\alpha_n)-\beta\sin\theta}{\cos(\theta+\alpha_n)-\beta\cos\theta}-\tan\theta}{1+\frac{\sin(\theta+\alpha_n)-\beta\sin\theta}{\cos(\theta+\alpha_n)-\beta\cos\theta}\times\tan\theta} = \frac{\text{num}}{\text{den}}$

$\text{num}\times[\cos(\theta+\alpha_n)-\beta\cos\theta] = \sin(\theta+\alpha_n)-\beta\sin\theta-\tan\theta[\cos(\theta+\alpha_n)+\beta\cos\theta]$

$= \sin\theta\cos\alpha_n+\cos\theta\sin\alpha_n-\tan\theta[\cos\theta\cos\alpha_n-\sin\theta\sin\alpha_n]$

$= \cos\theta\sin\alpha_n+\frac{\sin^2\theta}{\cos\theta}\sin\alpha_n = \frac{\sin\alpha_n}{\cos\theta}$

$\text{den}\times[\cos(\theta+\alpha_n)-\beta\cos\theta] = \cos(\theta+\alpha_n)-\beta\cos\theta+[\sin(\theta+\alpha_n)-\beta\sin\theta]\tan\theta$

$= \cos\theta\cos\alpha_n-\sin\theta\sin\alpha_n-\beta\cos\theta+[\sin\theta\cos\alpha_n+\cos\theta\sin\alpha_n-\beta\sin\theta]\tan\theta$

$= \frac{\cos^2\theta\cos\alpha_n-\cos\theta\sin\theta\sin\alpha_n-\beta\cos^2\theta+\sin^2\theta\cos\alpha_n+\sin\theta\cos\theta\sin\alpha_n-\beta\sin^2\theta}{\cos\theta}$

$= \frac{\cos\alpha_n-\beta}{\cos\theta}.$

Hence, $\tan\phi_n = \frac{\text{num}}{\text{den}} = \frac{\sin\alpha_n}{\cos\alpha_n-\beta}.$

Q.E.D.

Combining Eqs. (18), (A. 18) we get the boundary condition as

$$\gamma \tan\left(\frac{2\pi n}{N}\right) = \frac{\sin\alpha_n}{\cos\alpha_n - \beta} \qquad \text{(A. 19)}$$

Let us now prove a useful inversion formula:

$$\begin{array}{lrcll} \text{If,} & \tan\theta & = & \sqrt{\frac{1-\beta}{1+\beta}}\tan\phi, & (a) \\ \text{Then,} & \tan 2\phi & = & \frac{1}{\gamma}\frac{\sin 2\theta}{\cos 2\theta - \beta}, & (b) \\ \text{where,} & \gamma & = & \frac{1}{\sqrt{1-\beta^2}}. & (c) \end{array} \qquad \text{(A. 20)}$$

*Proof:* From (A. 20a),

$$\tan\phi = \sqrt{\frac{1+\beta}{1-\beta}}\,\tan\theta.$$

$$\tan 2\phi = \frac{2\tan\phi}{1-\tan^2\phi} = \frac{2\sqrt{\frac{1+\beta}{1-\beta}}\,\tan\theta}{1-\frac{1+\beta}{1-\beta}\,\tan^2\theta}$$

$$= \frac{2\sqrt{\frac{1+\beta}{1-\beta}}\,\tan\theta}{(1-\beta)-(1+\beta)\tan^2\theta} \times (1-\beta)$$

$$= \frac{2\sqrt{1-\beta^2}\,\tan\theta}{(1-\tan^2\theta)-\beta(1+\tan^2\theta)}$$

$$= \frac{1}{\gamma}\frac{2\tan\theta}{\frac{\cos^2\theta-\sin^2\theta}{\cos^2\theta}-\beta\frac{\cos^2\theta+\sin^2\theta}{\cos^2\theta}}$$

$$= \frac{1}{\gamma}\frac{2\tan\theta}{\frac{\cos 2\theta-\beta}{\cos^2\theta}} = \frac{1}{\gamma}\frac{2\sin\theta\cos\theta}{\cos 2\theta-\beta} = \frac{1}{\gamma}\frac{\sin 2\theta}{\cos 2\theta-\beta}.$$

Q.E.D.

Setting $\theta \equiv \frac{\alpha_n}{2}$; $\phi \equiv \frac{\pi n}{N}$ in Eq. (A. 20), we get

$$\begin{array}{ll} \text{If,} & \tan\frac{\alpha_n}{2} = \sqrt{\frac{1-\beta}{1+\beta}}\tan\frac{\pi n}{N}, \\ \text{Then,} & \gamma\tan\frac{2\pi n}{N} = \frac{\sin\alpha_n}{\cos\alpha_n-\beta}, \\ & \text{and vice versa,} \end{array} \qquad \text{(A. 21)}$$

Q.E.D.

Due to Eq. (A. 19) and identity between the equations (A. 21a) and (A. 21b), the former will now serve as the reduced boundary condition.

# B Appendix B

## B.1 Plot Commands for Gnuplot

In our previous article we had written the commands used in Gnuplot in full, to encourage the reader to replicate all the plots in that article. In this article we shall be economical, avoid the Preamble parts and go straight into the main commands, copying only six examples from the Console. Using them as a guide, the reader should be able to replot all the field lines.

Note that all plot functions are parametric functions of the form: x=f(t), y= g(t), in which t is the parameter. In the first five examples this parameter t represents the CtF distance R. In the final example, t represents the retarded time.

### B.1.1 Example 1. E field line from a particle moving under constant acceleration. See Fig.5

```
ac = 0.17*10**18
c=3*10**8
tau= c/ac
```

```
print tau

  1.76470588235294e-09
set term fig color portrait size 16 20 metric pointsmax 1000 solid font ”Times-Roman, 12” depth 50
set size square
set grid
set parametric
t0=tau
tr(t) = t0-t/c
RO = tau*c
RO = tau*c
b(t)=1/sqrt( 1+( tau/tr(t) )**2 )
print b(0.001*RO), b(0)

  0.706752962498301 0.707106781186547
print b(0.999*RO), b(0)
0.000999999500000358 0.707106781186547
xt(t)=c*tau*( sqrt(1+( tr(t)/tau)**2 ) -1 )
print xt(RO), xt(0)

  0.0 0.21928953302105
a(n,t) =2*atan( sqrt( (1-b(t) )/(1+b(t) ) )*tan( pi*n/16) )
x(n,t) = xt(t) + t* cos(a(n,t) )
y(n,t) = t* sin(a(n,t) )
set xrange [-RO:RO]; set yrange [-RO:RO]
set trange [0:RO]
set samples 200
set xtics RO/2; set mxtics 5
set ytics RO/2; set mytics 5
set title ”Erect-161028A.fig”
set out ”Erect-161028A.fig”
do for [n=1:16] plot x(n,t), y(n,t) lt n-1
```

#### B.1.2 Example 2. E field lines from a particle accelerating from $t = 0$ to $t = \tau$, then moving with constant velocity. See Fig.8(a)

This exercise invokes piecewise functions $\beta(t'), \tilde{x}(t')$, each defined for 3 different ranges in Eq. (38). Page 29 of Gnuplot Manual[6] tells us how to do it. Notice how we have defined the functions b(c,t) and xt(c,t) to represent $\beta(t'), \tilde{x}(t')$.

```
ac = 0.17*10**18
c=3*10**8
tau = c/ac
t0=2*tau
set parametric
tr(c,t)=t0-t/c
RA2=0
RA1= c*tau
RO= 2*RA1
RB= 3*RA1
bet(c,t)=1/sqrt( 1+ ( tau/tr(c,t) )**2 )
b(c,t) = RA2 <= t && t < RA1 ? bet(c,RA1) : RA1 <= t && t < RO ? bet(c,t) : RO <= t && t <= RB ? 0 : 1/0
print b(c,RA2), b(c,RA1), b(c,RO), b(c,RB)
  0.707106781186547 0.707106781186547 0 0
xta(c,t)=c*tau*( sqrt( 1+( tr(c,t)/tau )**2 ) -1 )
xtb(c,t)= xta(c,RA1) + c* b(c,RA1)*(tr(c,t)-tau)
xt(c,t) = RA2 <= t && t < RA1 ? xtb(c,t) : RA1 <= t && t < RO ? xta(c,t) : RO <= t && t <= RB ? 0: 1/0
print xt(c,RA2), xt(c,RA1), xt(c,RO), xt(c,RB)
  0.593640181884517 0.21928953302105 0 0
```

```
a(n,c,t) =2*atan( sqrt( (1-b(c,t) )/(1+b(c,t) ) )*tan( pi*n/16) )
x(n,c,t) = xt(c,t) + t* cos(a(n,c,t) )
y(n,c,t) = t* sin(a(n,c,t) )
set term fig color portrait size 16 20 metric pointsmax 1000 solid font ”Times-Roman, 12” depth 50
set size square
set xrange [-1.5:1.5]; set yrange [-1.5:1.5]
set trange [0:RB]
set samples 300
set xtics 0.5; set mxtics 5
set ytics 0.5; set mytics 5
set grid
set title “Erect-160522A.fig”
set out “Erect-160522A.fig”
do for [n=1:16] plot x(n,c,t), y(n,c,t) lt n-1
```

### B.1.3 Example 3. E field line from $q$ moving under constant deceleration. See Fig.11

```
ac = 0.17*10**18
c=3*10**8
tau(ac,c) = c/ac
set parametric dummy variable is t for curves, u/v for surfaces
t0(ac,c)=0
tr(ac,c,t)=t0(ac,c)-t/c
b(ac,c,t)=1/sqrt( 1+( tau(ac,c)/tr(ac,c,t) )**2 )
xt(ac,c,t)=c*tau(ac,c)*( sqrt(1+( tr(ac,c,t)/tau(ac,c))**2 ) -1 )
print xt(ac,c,0), xt(ac,c,0.528) 0.0
0.218291931024583
a(n,ac,c,t) =2*atan( sqrt( (1-b(ac,c,t) )/(1+b(ac,c,t) ) )*tan( pi*n/16) )
x(n,ac,c,t) = xt(ac,c,t) + t* cos(a(n,ac,c,t) )
y(n,ac,c,t) = t* sin(a(n,ac,c,t) )
set term fig color portrait size 16 20 metric pointsmax 1000 solid font ”Times-Roman, 12” depth 50 Terminal type set to ’fig’ Options are ’color small pointsmax 1000 portrait metric solid textnormal font ”Times Roman,12” linewidth 1 depth 50 version 3.2 size 16 20’
set size square
set xrange [-0.8:0.8]; set yrange [-0.8:0.8]
set trange [0:0.528]
set samples 200
set xtics 0.4; set mxtics 4
set ytics 0.4; set mytics 4
set grid
set title ”Edcel-161003A.fig”
set out ”Edcel-161003A.fig”
do for [n=1:16] plot x(n,ac,c,t), y(n,ac,c,t) lt n-1
```

### B.1.4 Example 4. E field lines from $q$ decelerating from $\beta_\infty = -0.707$ during $t = -\tau$ to $t = 0$, then sitting still. See Fig.13

```
ac = 0.17*10**18
c=3*10**8
tau= c/ac
print tau
  1.76470588235294e-09
set parametric
tr(t) =tau - t/c
R2=3*c*tau; R1=2*c*tau; RO=c*tau
print R2, R1, RO
  1.58823529411765 1.05882352941176
0.529411764705882
bet(t) = -1/sqrt( 1+( tau/tr(t) )**2 )
```

```
b(t) = 0 <= t & & t <= RO ? 0 : RO < t & & t < R1 ? bet(t) : R1<= t & & t <= R2 ? bet(R1) : 1/0
print b(0), b(RO), b(R1), b(R2) 0 0 -0.707106781186547 -0.707106781186547
xta(t)=c*tau*( sqrt( 1+( tr(t)/tau )**2 ) -1 )
xtb(t)= xta(R1) + c* bet(R1)*(tr(t)+tau)
xt(t) = 0 <= t & & t <= RO ? 0 : RO < t & & t < R1 ? xta(t) : R1 <= t & & t <= R2 ? xtb(t) : 1/0
print xt(0), xt(RO), xt(R1), xt(R2)


 0 0 0.21928953302105 0.593640181884517
a(n,t) =2*atan( sqrt( (1-b(t) )/(1+b(t) ) )*tan( pi*n/16) )
x(n,t) = xt(t) + t* cos(a(n,t) )
y(n,t) = t* sin(a(n,t) )
set term fig color portrait size 16 20 metric pointsmax 1000 solid font ”Times-Roman, 12” depth 50
set size square
set grid
set xrange [-R2:R2]; set yrange [-R2:R2]
set trange [0:R2]
set samples 1000
set xtics RO; set mxtics 5
set ytics RO; set mytics 5
set title ”Edcel-161016B.fig”
set out ”Edcel-161016B.fig”
do for [n=1:16] plot x(n,t), y(n,t) lt n-1
set xrange [-2*RO:5*RO]; set yrange [-3.5*RO:3.5*RO]
set title ”Edcel-161016C.fig”
set out ”Edcel-161016C.fig”
do for [n=1:16] plot x(n,t), y(n,t) lt n-1
```

### B.1.5 Example 5. Field line from particle in circular motion with $\beta = 0.5, N = 16, R\,\epsilon\,[0 : 13\pi]$. See Fig. 21

```
g(b) = 1/sqrt(1-b*b) # defines γ
a(N,n,b,r,t)=2* atan( sqrt((1-b)/(1+b)) * tan( -(g(b)*b*t)/(2*r) + pi *n/N) ) # defines α
x(N,n,b,r,t) = -r* sin(b*t/r) + t*cos((b*t/r) +a(N,n,b,r,t) ) # x coordinate of field line
y(N,n,b,r,t) = r* cos(b*t/r) + t*sin((b*t/r) +a(N,n,b,r,t) ) # y coordinate of field line
set term fig color portrait size 16 20 metric pointsmax 1000 solid font ”Times-Roman, 12” depth 50
set size square
set grid
set parametric
set trange [0:6*pi]
set xrange [-6*pi:6*pi]; set yrange [-6*pi:6*pi]
set samples 400
set xtics 4*pi; set mxtics 4
set ytics 4*pi; set mytics 4
N=16; b=0.5; r=1
set title “Qcir-b5-16-400-160717.fig”
set out “Qcir-b5-16-400-160717.fig”
do for [n=1:N] plot x(N,n,b,r,t) , y(N,n,b,r,t) lt n-1
```

### B.1.6 Example 6. Concentration of E field along a spiral. See Fig. 27

```
set parametric
set term fig color portrait size 16 20 metric pointsmax 1000 solid font “Times-Roman, 12“ depth 50
```

```
x(t) = sin (t) - t* cos(t); y(t) = cos(t) + t * sin(t)
set size square
set grid
set samples 200
set trange [-4*pi:0]
set xrange [-4*pi:4*pi]; set yrange [-4*pi:4*pi]
set xtics 2*pi; set mxtics 2
set ytics 2*pi; set mytics 2
set title “Espiral-160716A.fig”
set out “Espiral-160716A.fig”
plot x(t) , y(t)
```

## Acknowledgement

I am grateful to Prof Rogers Fuller, Associate Director of Membership, *American Association of Physics Teachers*, and to Harold Q and Charolette Mae Fuller for granting me *Fuller Fund Membership of the American Association of Physics Teachers.* As a privilege of this membership I could get online access to all the past issues of the American Journal of Physics, which made it possible to write this article.

I thank Prof A.N.Maheshwari and Prof Sophocles J. Orfanidis (Rutgers University) for reading through my manuscript and giving valuable suggestions.

I thank the anonymous referee for helpful suggestions leading to an improved manuscript.

## References

[1] Somnath Datta, *Visualization of Electromagnetic Fields Using Gnuplot*, Physics Education, www.physed.in, Indian Association of Physics Teachers, Vol 31, No.4, (Oct-Dec 2015), Article Number 8, 39 pages.

[2] John David Jackson, *Classical Electrodynamics*, 3rd Ed, John Wiley & Sons, New York (2004). p. 664.

[3] David J. Griffiths, *Introduction to Classical Electrodynamics*, 3rd Ed., Pearson Education, Inc, (2006), p. 456.

[4] Edward M, Purcell, *Electricity and Magnetism, Berkeley Physics Course - Vol 2*, Tata McGraw Hill Publishing Co Ltd, New Delhi (1965), pp.161-167.

[5] Roger Y. Tsien, *Pictures of Electric Fields*, American Journal of Physics, **40**, 46 (1972).

[6] Thomas Williams & Colin Kelley, *Gnuplot 4.6*, Manual originally prepared by Dick Crawford, 2012, Version 4.6, http://sourceforge.net/projects/gnuplot (2012).

[7] Lee Phillips, *Gnuplot Cookbook*, Packt Publishing Ltd., Birmingham B3 2PB, UK..ISBN 978-1-84951-724-9 www.packtpub.com (2012).

[8] Somnath Datta, *Mechanics*, Pearson Education, Chennai (2013), pp. 355.

[9] L. Page and N.I.Adams, *Electrodynamics*, Dover, New York (1940), pp.166-167.

[10] Ref. 8, pp. 591-593

[11] Somnath Datta, *Introduction to Special Theory of Relativity*, Allied Publishers Ltd., New Delhi, (1998), p. 79

[12] Charles Kittel, Walter D. Night and Malvin A. Ruderman, *Mechanics, Berkeley Physics Course - Vol 1*, Tata McGraw Hill Publishing Co Ltd, New Delhi (1964), pp.407-408.

[13] Hint from the referee

[14] Emilio Segre, *Nuclei and Particles*, W.A.Benjamin, Inc (1965), pp.137-142.

[15] Wikipedia, *Synchrotron Radiation*, https://en.wikipedia.org/wiki/Synchrotron_radiatio (7/21/2016).

[16] G.B.Thomas and R.S.Finney, *Calculus and Analytic Geometry*, 6th Ed, pp. 713, 783.

[17] Ref. 3, p.672.

[18] John David Jackson, *Classical Electrodynamics*, 1st Ed, John Wiley & Sons, New York (1962). p.477.

# Lorentz transformations without light

Franco Battaglia[1], Thomas F. George[2]

1. *Università di Modena e Reggio Emilia, Italy*
2. *University of Missouri–St. Louis, St. Louis, MO*

*E -mail: tfgeorge@umsl.edu*

(*Submitted:19-07-2016*)

## Abstract:

A six-step derivation is given for the Lorentz transformation which, without any reference to light and without resorting to advanced group-theory arguments, should avoid any misunderstanding about the connection of *light* with relativity theory.

When presenting the special theory of relativity, very appropriately one usually starts with two postulates: (1) Galilean relativity, according to which fundamental physics laws must be invariant upon going from one inertial frame to another, and (2) the universality of the speed of light, according to which the speed of light is independent of the motion of its source, i.e., any inertial frame, upon measuring the speed of the *same* photon, reports the same value of $c = 3\times10^8$ m/s . Following these two postulates, the space-time coordinate transformations from one inertial frame to another – Lorentz transformations (LT) – are derived.

However, light *per se* does not have much to do with relativity, besides the fact that it just happens to travel at the limiting speed – $c$, for *celeritas* – prescribed by the theory. The second postulate is then a convenient way of positing the existence of such a limiting speed – convenient because we have at our disposal something that experimentally certifies that this is the way the world is constituted: the reference to light is then due to the historical role that electromagnetism has played in the discovery of special relativity.[1]

Here is shown a six-step derivation of the LT which, without any reference to light and without resorting to advanced group-theory arguments,[2] should avoid any misunderstanding about the connection of *light* with relativity theory. This holds not because the speed of light is invariant, but because space-time is so specially constituted: the experimentally verified invariance of the speed of light is just one of many experiments confirming that LT are the correct space-time coordinate transformations between two equivalent frames.

**Two options**

The framework on which physics is built is that of a time which is uniform and a space which is homogeneous and isotropic, i.e., the fundamental laws of physics are invariant under translation in time, and translation and rotation in space. We also require invariance upon going from one frame, say $F$, to another, say $F'$, moving with constant velocity, $\mathbf{V}$, with respect to $F$. This is the usual first postulate of relativity, Galileo's postulate. In particular, the space–time coordinate transformations from $F$ to $F'$ must have the same functional form of those from $F'$ to $F$, a circumstance which we shall refer to as *reciprocity*. For the time being, we do not require the speed of light to be constant. We now proceed to seeking the transformations between the space-time coordinates from $F$ to $F'$ and *vice versa*. This is done in six steps. In what follows, we shall put a prime to any quantity evaluated in the frame $F'$.

**First step.** From reciprocity, it is sufficient that the sought transformations be linear, which implies that the inverse transformations have the same functional form. The transformations are then of the type

$$x_{\mu'} = \sum_{\nu} L_{\mu'\nu} x_{\nu} + k_{\mu'} \ , \qquad (1)$$

where $x_{\mu}$ and $x_{\mu'}$ are the space-time coordinates in $F$ and $F'$, and $L_{\mu'\nu}$ and $k_{\mu'}$ are quantities independent of those coordinates. Later on, we shall specify on what they *may* depend. The indices vary from 1 to 4, with $x_1 = x$, $x_2 = y$, $x_3 = z$, $x_4 = t$.

**Second step.** The symmetry properties of space and time allow us to choose the space and time origin and the orientations of the coordinate axes at will. Designating a given set of the four space-time coordinates as an 'event,' we make the choice that the event labeled as $x = y = z = t = 0$ in $F$ is labeled as $x' = y' = z' = t' = 0$ in $F'$. Likewise, we may as well choose $y' = y$ and $z' = z$ at $t = 0$ and the positive direction of both the $x$-and $x'$-axes along the positive direction of $\mathbf{V}$. From $x_{\mu'} = 0 = x_{\mu}$ for all $\mu$ it follows that $k_{\mu'} = 0$ for all $\mu$.

**Third step.** We determine the transformation for the *longitudinal* space-coordinate component, i.e., with the choice made, the $x$- coordinate. Since $x' = 0 = t'$ when $x = 0 = t$, then $L_{1'2} = L_{1'3} = 0$, as can be seen from Eq. (1). The $x$- coordinate transformation equation must then be of the form $x' = \gamma x + \delta t = \gamma(x + \frac{\delta}{\gamma} t)$. At all times, the first spatial coordinate of the origin $O'$ are $x' = 0$ and $x = Vt$, i.e. $\delta/\gamma = -V$. The transformation equation for the $x$-coordinate is therefore

$$x' = \gamma(x - Vt) \ . \qquad (2a)$$

At this point we can say that $\gamma > 0$, from the choice we have made of the coordinate axes, according to which $\mathrm{sgn}(x') = \mathrm{sgn}(x)$ at $t = 0$. Also, $\gamma$ *may* depend on $V$ and, if so, the isotropy of space requires that $\gamma(-V)=\gamma(V)$, reciprocity requires that $|V'|=|V|$, and the chosen direction of the $x$- and $x'$- axes requires that $V'=-V$. Hence, the inverse transformation is obtained by the replacements $x'\leftrightarrow x$ and $-V\leftrightarrow V$:

$$x=\gamma(x'+Vt') \ . \tag{2b}$$

**Fourth step.** We now find the transformations for the time coordinate. From Eq. (2b), isolating $t'$ and inserting Eq. (2a), we get

$$t'=\gamma\left(t-\frac{\varepsilon}{V}x\right) , \tag{3}$$

where we have set

$$\varepsilon\equiv 1-\frac{1}{\gamma^2} , \tag{4a}$$

by which we must have $\varepsilon<1$. From Eq. (4a),

$$\gamma^2\equiv\frac{1}{1-\varepsilon} \ . \tag{4b}$$

**Fifth step.** We determine the transformations for the *transverse* space-coordinate components, i.e., the $y$- and $z$- coordinates. Having chosen $y'=y$ and $z'=z$ at $t=0$, we have $L_{2'1}=L_{2'3}=0=L_{3'1}=L_{3'2}$ and $L_{2'2}=1=L_{3'3}$. However, from Eq. (3) we see that the time-coordinate transformation does not involve the transverse spatial components, so that $L_{2'4}=0=L_{3'4}$. The transformations for the transverse spatial components are then

$$y'=y \qquad \text{and} \qquad z'=z \ . \tag{5}$$

**Sixth step.** We determine the dimensionless quantities $\gamma$ and $\varepsilon$. The crucial point is that there are two options here. Either they are constant, or they depend on $V$. The matter has to be resolved experimentally, unless some extra assumption comes into play.

In the former case, $\gamma$ may be taken equal to 1: any other value would simply imply, as can be seen from Eq. (2), a change in scale of the units chosen. With $\gamma=1$, we have $\varepsilon=0$, and

$$x'=x-Vt \qquad \text{and} \qquad t'=t \ . \tag{6}$$

Equations (5) and (6) are the first-option transformations,i.e., Galileo transformations. Notice that the Galilean velocity-composition rules readily follow from Eqs. (5) and (6):

$$u_{x'}\equiv\frac{dx'}{dt'}=\frac{dx'}{dt}=u_x-V \tag{7a}$$

$$u_{y'}\equiv\frac{dy'}{dt'}=\frac{dy}{dt}=u_y \qquad \text{and}$$

$$u_{z'}\equiv\frac{dz'}{dt'}=\frac{dz}{dt}=u_z \ . \tag{7b}$$

If the second option holds, being dimensionless quantities, $\gamma$ and $\varepsilon$ must rather depend on $V/c$, where $c$ is some $V$-independent quantity (i.e., some universal constant) with the dimensions of velocity, whose value must be determined from experiments. Let us then determine $\gamma$ and $\varepsilon$. The velocity-composition rules that follow from Eqs. (2) and (3) are

$$u_{x'}\equiv\frac{dx'}{dt'}=\frac{dx-Vdt}{dt-\frac{\varepsilon}{V}dx}=\frac{u_x-V}{1-\frac{\varepsilon}{V}u_x} \tag{8a}$$

$$u_{y'} \equiv \frac{dy'}{dt'} = \frac{dy}{\gamma\left(dt - \frac{\varepsilon}{V}dx\right)} = \frac{u_y}{\gamma\left(1 - \frac{\varepsilon}{V}u_x\right)} \qquad (8b)$$

And

$$u_{z'} \equiv \frac{dz'}{dt'} = \frac{u_z}{\gamma\left(1 - \frac{\varepsilon}{V}u_x\right)} . \qquad (8c)$$

Squaring and adding these velocity components, and making use of Eq. (4b), one obtains

$$(u')^2 = \frac{(u_x - V)^2 + (u^2 - u_x^2)(1-\varepsilon)}{\left(1 - \frac{\varepsilon}{V}u_x\right)^2} . \qquad (9)$$

With $c$ as an invariant velocity, setting $u^2 = c^2 = (u')^2$ , Eq. (9) becomes, after rearrangement,

$$c^2 = \frac{c^2\left(1 - 2\frac{V}{c^2}u_x + \frac{\varepsilon}{c^2}u_x^2\right) + V^2 - \varepsilon c^2}{\left(1 - \frac{\varepsilon}{V}u_x\right)^2} . \qquad (10)$$

With $\varepsilon < 1$, the only physically meaningful solution to this equation is

$$\varepsilon = \frac{V^2}{c^2} \equiv \beta^2 . \qquad (11)$$

The other solution, $\varepsilon = (V/u_x)(2 - V/u_x)$, has to be discarded on physical grounds, for $\varepsilon$ would otherwise depend on the orientation of frames in space.

We then see that $0 \le \varepsilon < 1$, whereby $V < c$ (no frame can travel at $V \ge c$ ),

$$\text{and } \gamma = \frac{1}{\sqrt{1-\beta^2}} > 1 . \qquad (12)$$

Equation (3) then becomes

$$t' = \gamma\left(t - \frac{V}{c^2}x\right) , \qquad (13a)$$

and its inverse is

$$t = \gamma\left(t' + \frac{V}{c^2}x'\right) . \qquad (13b)$$

Equations (2), (5), (12) and (13) are the sought second-option transformations, i.e., Lorentz transformations.

Notice, from Eq. (9), that $u < c \Rightarrow u' < c$ and $u = c \Rightarrow u' = c$ , where $u$ and $u'$ are the speeds of a particle in $F$ and $F'$: not only is $c$ an invariant, but it is also a limiting speed.

## Deciding between the options: The tragedy of a muon

Which one is the option to pick is an experimental matter (unless other assumptions are added). Of course, the fact that photons *do* travel at an invariant speed tells us that the world is constituted according to the second option, and the speed of light has to be identified with the constant $c$ . This is the path historically taken. However, if light *did not* travel with an invariant speed – or, for that matter, if *nothing* traveled at an invariant speed – the second option could not be discarded, and should have been (and it has been) answered by appropriate experimental results, as it should be recalled at this point.

There are, among others, two remarkable consequences of LT: length contraction and time dilation. According to *any* frame, length-of-a-stick means the distance between the stick end-points when their spatial coordinates are evaluated *at the same time*. A stick at rest in frame $F$, along the $x$- axis, and with end-point spatial coordinates $x_A$ and $x_B$, would have length

$\lambda \equiv \Delta x = |x_B - x_A|$ in frame $F$. From Eq. (2b), the end-points have coordinates $x_A = \gamma(x'_A + Vt'_A)$ and $x_B = \gamma(x'_B + Vt'_B)$. In order for $|x'_B - x'_A|$ to be the length $\Delta x'$ of the stick in frame $F'$, in the above equations $t'_A = t'_B$ should hold, whereby $\lambda = |x_B - x_A| = \gamma|x'_B - x_A| = \gamma\Delta x'$. Therefore,

$$\Delta x' = \frac{\lambda}{\gamma} \ , \tag{14}$$

i.e., a stick in $F'$ is shorter than in $F$ (where it is at rest and with length $\lambda$) by a factor $\gamma$, a circumstance called *length contraction*.

Similarly, if two events happen in $F$ at the same place $(x_A = x_B)$ and with a lag in time given by $\tau = |t_B - t_A|$, the lag in time $|t'_B - t'_A| \equiv \Delta t'$ between them, according to $F'$, is obtained from Eq. (13a): $t'_A = \gamma\left(t_A - \frac{V}{c^2}x_A\right)$ and $t'_B = \gamma\left(t_B - \frac{V}{c^2}x_B\right)$. With $x_A = x_B$, one obtains

$$\Delta t' = \gamma\tau \ , \tag{15}$$

i.e., in $F'$ the time interval between two events is longer by a factor $\gamma$ than in $F$ (where the two events happen at the same place), a circumstance called *time dilation*.

Now, we recall – in a simplified version, i.e., not realistic but adapted to our purposes – a remarkable experiment.[3] In a laboratory on Earth, it is possible to detect a muon, born at, say, $15\ \text{km}$ up in the troposphere, striving to reach us at the speed $V = 2.997\times10^8$ m/s, but dying just before touching the Earth's surface, after a life 50.05 $\mu$s long. From the reference frame of the muon, once born, she sees the planet Earth heading against her at the speed $V = 2.997\times10^8$ m/s; however, she dies after only 2.2 $\mu$s, just before experiencing the crush. The two events – birth and death of the muon – have a different time lag in the two frames: undoubtedly, the world must be constituted according to the second option. The value of the limiting invariant speed can now be evaluated. From Eq. (15), with $\tau = 2.2\ \mu$s and $\Delta t' = 50.05\ \mu$s, $\gamma$ turns out to be $\gamma = 22.75$, and from Eqs. (11) and (12), $\beta \equiv V/c = 0.999$, whereby $c = 3\times10^8$ m/s. (Once again we stress the simplified version given of the actual experiment.)

Of course, if the muon had lived long enough to survive the crush, she could have evaluated how far was our planet at the time she was born: $V\tau = 659$ m, a result in agreement with what would be obtained from Eq. (14), with $\lambda = 15$ km and $\gamma = 22.75$.

## Conclusions

We would like to stress once again that $c$, usually called the speed of light in vacuum, is rather the invariant (and limiting) speed built in the geometry obeyed by our space-time. This is a notion that could (and should) be conveyed right at the beginning when LT are derived. Due to reciprocity, they have to be of the form of Eq. (1), with the coefficients that, without any further assumption, either depend or do not depend on $V$: *tertium non datur*. In the former case, the invariant speed must also be a limiting speed, regardless of whether or not there exist particles travelling at that speed.

The development of the theory provides for a free particle with mass $m$ and linear momentum $p$, an energy $E$ given by $E^2 = m^2c^4 + p^2c^2$: unlike classical mechanics, relativity allows zero-mass particles, in which case $E = pc$. However, according to the theory, $E = mc^2\gamma_u$, where $\gamma_u \equiv \left(1-u^2/c^2\right)^{-1/2}$: the last two relations for the energy of a zero-mass particle are compatible only if $u = c$, i.e., massless particles must travel at speed $c$.

The fact that ubiquitous light travels at speed $c$ has been indeed a lucky occurrence, without which, everything else being equal, the realization of the space-time real structure might have waited some longer time.

**References**

1. P. B. Pal, “Nothing but relativity”, Eur. J. Phys. **24**, 315-319 (2003); A. Pelissetto and M. Testa, “Getting the Lorentz transformations without requiring an invariant speed”, Am. J. Phys. **83**, 338-340 (2015).
2. J-M. Lévy-Leblond, “One more derivation of the Lorentz transformation”, Am. J. Phys. **44**, 271-277 (1976).
3. D. H. Frisch and J. H. Smith, “Measurement of the relativistic time dilation using $\mu$-mesons”, Am. J. Phys. **31**, 342-355 (1963).

# Teaching Superfluidity at the Introductory Level

L. Ruzhitskaya[1] and W. Montfrooij[2]

[1]School of Sciences
Saint Francis University
Loretto, Pennsylvania 15940, United States.
lruzhitskaya@francis.edu

[2]Department of Physics and Astronomy
University of Missouri
Columbia, Missouri 65211, United States.
montfrooijw@missouri.edu

(Submitted 13-07-2016)

## Abstract

Standard introductory modern physics textbooks do not exactly dwell on superfluidity in $^4$He. Typically, Bose-Einstein condensation (BEC) is mentioned in the context of an ideal Bose gas, followed by the statement that BEC happens in $^4$He and that the ground state of $^4$He exhibits many interesting properties such as having zero viscosity. Not only does this approach not explain in any way why $^4$He becomes a superfluid, it deprives students of the opportunity to learn about how the symmetry requirements on the wavefunction for bosonic systems lead to the emergence of energy gaps, which in turn lead to superfluidity and superconductivity. We revisit superfluid $^4$He by starting with Feynman's explanation of superfluidity, and we present exercises for the students that will allow them to arrive at a very accurate estimate of the superfluid transition temperature. This paper represents a self-contained account of superfluidity, which can be covered in one or two lessons in an introduction to modern physics class.

# 1 Introduction

Kamerlingh Onnes liquefied helium in 1908 for the first time. He used it as a coolant in order to be able to study the low temperature properties of mercury and was awarded the Nobel Prize in physics in 1913 for his discovery of superconductivity. However, his coolant turned out to be a surprise in itself when Kapitza discovered in 1938 [1] that, when helium is cooled down below 2.17 K, it can flow through small openings without any friction. When the liquid reaches this superfluid state, it stops boiling in the sense that, in contrast to normal fluid helium, there are no bubbles visible. In addition, it becomes a conductor of heat orders of magnitude better than copper. We show this transition of liquid helium from a normal fluid phase into a superfluid phase in figure 1.

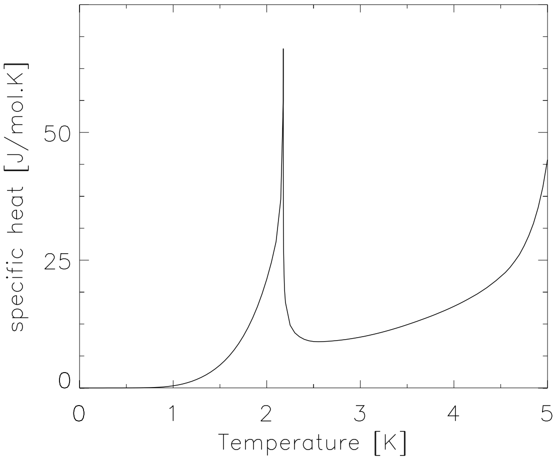


Figure 1: When liquid helium is cooled down, the specific heat (measured at saturated vapor pressure) displays a sharp increase peaking at 2.17 K. In general, such increases are associated with transitions from disordered to more ordered states; yet, in the case of liquid helium, the liquid remains a liquid. The shape of the specific heat curve has given this transition the name of $\lambda$-transition, the transition between a phase where helium behaves as an ordinary liquid, and a phase below 2.17 K where it can flow without friction, the so-called superfluid phase. The upturn at 5 K is caused by the approach to the critical point (5.2 K, 2.23 bar); the boiling point under atmospheric pressure is at 4.2 K. The data have been taken from reference [2].

Shortly after the discovery of superfluidity, London [3] suggested that the phenomenon is directly associated with the formation of a Bose-Einstein condensate, that is, a state of matter where a large fraction of the atoms condense in a state of zero momentum. In 1947, Bogoliubov showed [4] that a system of weakly interacting Bose atoms would indeed, upon forming a Bose-Einstein condensate, exhibit the property of superfluidity. However, it was also realized that this need not necessarily apply to superfluid helium since this is a system of strongly interacting bosons. Based on the temperature dependence of the specific heat of the superfluid phase, Landau deduced [5] what the elementary excitation spectrum of the superfluid would look like: he predicted how much energy it would cost to excite a density disturbance of a specific wavelength in the liquid. When the spectrum of these elementary excitations were measured in

1957 [6] by means of neutron scattering, it turned out that Landau's predictions were remarkably accurate. Note that Landau's predictions neither proved nor disproved Bogoliubov's work as the presence of a condensate did not need to be invoked to calculate how much energy it would cost to excite the system out of the groundstate.

Feynman showed [7] in arguments we will revisit later on in this paper that, by invoking the Bose statistical nature of helium-4 atoms, even short-wavelength excitations corresponding to the rearrangment of just a few atoms could not be created at a negligible energy cost. Thus, he explained in a qualitative sense one of the essential features of the excitation spectrum of superfluid helium-4, namely the feature that is associated with the critical velocity above which helium no longer behaves as a superfluid, as we shall see in the remainder of this paper. Feynman also was fairly successful in coming up with qualitative predictions for the excitation spectrum, but those calculations are beyond the scope of this paper [8, 9]. However, in here we present an alternative approach that builds upon the findings of Landau and Feynman, where rather than attempting to reproduce the entire excitation curve from scratch, we connect microscopic manifestations of Bose particles to macroscopic quantities, such as the superfluid transition temperature. In order to do so, we must first clarify exactly the difference between superfluidity and Bose-Einstein condensation.

Superfluidity is the property of a liquid to flow without friction through thin capillaries [1]. This property is manifest in $^4$He below $T_\lambda$= 2.17 K, the so-called superfluid or lambda-transition (named after the shape of the specific heat curve). Below 1 K, 100% of the liquid exhibits this property. Bose-Einstein condensation (BEC) on the other hand, is the property that a large fraction of the particles that make up a system condense into the same state. For instance, in an ideal Bose gas (a gas made up of bosons that do not interact with each other), 100% of the particles will condense into the state with the lowest available energy and form a Bose-Einstein condensate. However, an ideal Bose gas does not become superfluid. And conversely, in liquid $^4$He only about 7% [10] of the atoms actually form a condensate, even though essentially 100% of the atoms can flow without friction below 1 K.

In fact, there is no reason why a system could not become a superfluid even if only a very small fraction of the atoms were to form a condensate. All this nicely illustrates the fact that superfluidity and BEC are two different phenomena. However, the way these two phenomena are presented in introductory textbooks tends to give students the impression that the two are one and the same phenomenon.

The main difference between BEC and superfluidity is that BEC is a property of the ground state, while superfluidity is a property of the excited states. This is entirely analogous to standard superconductivity, where the electrons condense into Cooper pairs (ground state), and where the interaction between the Cooper pairs introduces a finite energy gap between the ground state and excited states. In turn, this energy gap is responsible for the system becoming a superconductor. Thus, in both systems, it is the interaction between the particles that is responsible for the exotic behaviors, not how they arrange themselves in the ground state.

In this paper we focus on the property of superfluidity rather than on BEC. We repeat Feynman's arguments that show that any Bose liquid that stays liquid down to low enough temperatures must become a superfluid because of the presence of an energy gap. We also derive a very accurate estimate of the superfluid transition temperature using basic conservation laws and some straightforward approximations. Altogether, this should give students a much better understanding of what superfluidity entails and why it necessarily must occur in $^4$He. In addition, our simple calculations should bestow upon them the idea that they have already learned enough physics to be able to come up with a very accurate estimate of something as complex as the superfluid transition temperature in $^4$He.

# 2 Superfluidity: qualitative understanding

## 2.1 Elementary excitations

First, the fact that helium does not solidify at any temperature is a pure quantum effect. The weak van der Waals forces between the atoms are not strong enough to overcome the zero point motion associated with trying to confine a helium atom to a lattice site. The second aspect that makes helium stand out from other liquids is that it takes a finite amount of energy to create a disturbance in the liquid. This is shown in Fig. 2. The actual amount of energy required depends on the wavelength $\lambda$ (or momentum $p = h/\lambda$) of this disturbance. The measured values [11] for the energy cost are shown in Fig. 2. At low momentum transfers (long wavelengths) the energy disturbance is just a run-of-the-mill sound wave (which also goes by the name of a phonon), and its energy is given by $E_{ex}(p) = cp$, the standard hydrodynamics result for any liquid, not just superfluids [12]. Here $c$ stands for the speed of sound, and the reader might be more familiar with this result when it is written in the familiar form $\lambda f = c$, with $E_{ex} = hf$ and $h$ being Planck's constant.

When we go to shorter wavelengths, such as the density disturbance pictured in Fig. 3, we can see from Fig. 2 that the energy cost starts to deviate from $E_{ex}(p) = cp$. For

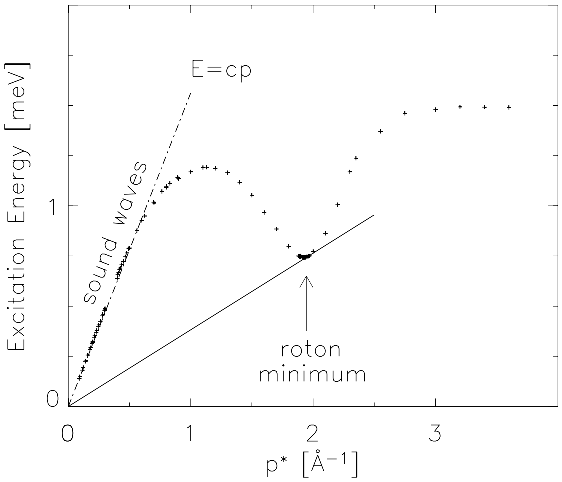


Figure 2: The measured excitation energies $E(p)$ [11], that is, the cost of creating an excitation out of the groundstate of $^4$He, as a function of momentum transfer $p^* = p/\hbar = 2\pi/\lambda$. The slope of the curve at small $p^*$ (dashed line) is given by the velocity of sound $c$= 237.4 m/s [13], the overall minimum slope (corresponding to a speed of 58 m/s) is given by the solid line which is tangent to the excitation curve near the so-called roton minimum [$p^* = 1.94$ Å$^{-1}$, $E(p)$= 0.743 meV].

wavelengths comparable to the interatomic spacing $d$, the energy cost goes through a minimum, after which it goes up again. This minimum of the energy gap between the ground state and the excited state is commonly referred to in the literature on superfluid helium as the roton minimum [14], or simply 'the roton', while the entire curve is referred to as the phonon-roton dispersion curve. The maximum of the dispersion in between the phonon and roton region is commonly referred to as 'the maxon'.

The roton turns out to be the determining feature of superfluids. As pointed out in the preceding, the presence of this roton gap is analogous to the presence of a similar gap in superconducting systems. The presence of a gap also firmly sets superfluid $^4$He apart from normal fluids where nothing resembling an energy gap exists. This is shown in Fig. 4 where we compare helium in the superfluid phase to helium in the normal fluid phase.

The curve in figure 2 that displays the energy cost of an elementary excitation as a function of wavelength can also be measured in solids. These curves are referred to as dispersion curves. The excitations represented by such curves are collective excitations that propagate from one place to another, with a distinct energy and momentum. It is for this reason that these excitations are often referred to as quasi-particles. When the system is not at absolute zero, a number of such quasi-particles will be present in the system. From the dispersion curve, the specific heat as a function of temperature can be determined, something which is routinely done for solids, but which works equally well for helium. Landau achieved the opposite by inferring the dispersion curve from specific heat data shown in figure 1.

## 2.2 Criterion for superfluidity

One can easily verify from Fig. 2 that the presence of a non-zero energy gap is synonymous with the property of superfluidity. The slope of a line that goes through the

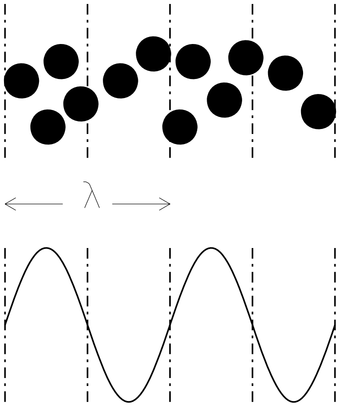


Figure 3: A real space visualization of a density disturbance (in this figure that would correspond to a departure from the average density of about 3 atoms in between the dashed lines) that resembles a sound wave with a wavelength $\lambda$ of about 3 times the average atomic separation. The roton minimum corresponds to a disturbance with a wavelength comparable to the interatomic separation [14].

origin and a point on the excitation curve gives the (group) velocity of the excitation. For instance, this slope at small momentum transfers is given by the speed of sound (see Fig. 2). Overall, the smallest slope is encountered near the roton minimum. The value of the slope at this point corresponds to the velocity below which the liquid can flow without friction (at least through small capillaries as we will discuss in the following): if the liquid is flowing at a lower speed, then the liquid cannot slow down because of restrictions due to the energy and momentum conservation laws, as detailed below.

Following standard arguments [15], we focus on a liquid mass $M$ that is flowing at speed $v$. For it to slow down to speed $v'$ by creating an excitation with energy $E_{ex}$ and momentum $p_{ex}$ we must have

$$\begin{aligned} \tfrac{1}{2}Mv^2 &= \tfrac{1}{2}Mv'^2 + E_{ex} \\ M\vec{v} &= M\vec{v}' + \vec{p}_{ex}. \end{aligned} \tag{1}$$

Eliminating $v'$ we obtain

$$\vec{v}.\vec{p}_{ex} - p_{ex}^2/2M = E_{ex}. \tag{2}$$

Even in the best case scenario where $\vec{v}$ and $\vec{p}_{ex}$ are parallel and in which $M$ is very large, we find that the minimum requirement on the flow velocity $v$ for the liquid to be able to slow down is given by:

$$v_{min} \geq E_{ex}/p_{ex} \tag{3}$$

In a normal liquid without an energy gap, liquid flow will always be viscous because the minimum slope would be zero. For instance, the spectra shown in Fig. 4 shows that a neutron can transfer any amount of energy to the fluid, no matter how small. Similarly, a mass $M$ of moving liquid can transfer any amount of energy to the rest of the liquid, there does not exist a minimum requirement.

Eqn 3 is known as the Landau criterion for superfluidity [15], and it corresponds to a flow velocity of 58 m/s. The actual critical velocity in bulk superfluid helium is much lower because in a bulk liquid large scale rotational excitations can form. These excitations are akin to the vortex

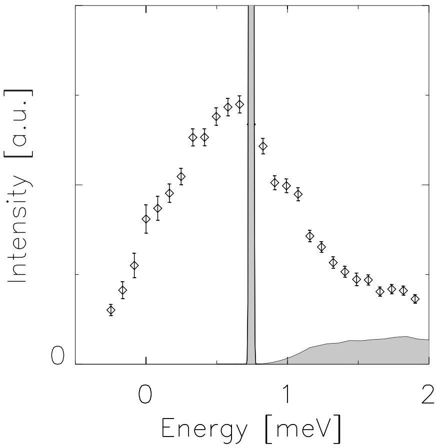


Figure 4: Detailed view of the excitations of $^4$He corresponding to the roton minimum in the superfluid phase (shaded) and in the normal fluid phase (points plus errorbars) [16]. The data are taken by means of neutron scattering experiments. The neutron transfers energy $E$ to the liquid and an amount of momentum corresponding to the roton minimum [see vertical arrow in Fig. 2]. When the amount of energy transferred exactly matches the energy difference between the ground state and the excited state, then a sharp resonance peak (at 0.743 meV) can be seen in the superfluid. Note that there is no signal below this peak. In the normal fluid the behavior is very different; even a small amount of energy is sufficient to excite the liquid, and there is a clear signal even at $E$ =0. The signal at $E <$ 0 implies that the liquid can transfer some of its energy to the neutron, which, of course, can only take place if the liquid is not at zero Kelvin.

that we can see when stirring a cup of coffee. However, these excitations can be suppressed by making the superfluid flow through very small orifices that inhibit the formation of such large scale vortices. When this is done, as demonstrated in experiments by Varoquaux *et al.* [17], then the critical velocity is indeed close to the one predicted by eqn 3. Experiments where negative ions are used to create excitations yield critical velocities as high as 60 m/s [18]. Thus, eqn 3 is a correct measure of the underlying mechanism behind superfluidity.

## 2.3 BEC and superfluidity

Eqn 3 also explains why an ideal Bose gas does not become superfluid. In an ideal Bose gas, consisting of non-interacting particles, the excitation energies are given by $E_{ex} = p_{ex}^2/2m$. Of course, a parabolic curve does not have a minimum slope, and therefore, no matter how slow an ideal Bose liquid is flowing, it is always possible to transfer energy to the rest of the liquid by creating an excitation, and the liquid will slow down. This is shown in Fig. 5a. Also, note that even though it requires less energy to create a sound wave than a roton excitation, the roton minimum actually determines the critical flow velocity (compare the two slopes in Fig. 2).

As an aside, this is a good point to elucidate why the presence of a Bose condensate is still required in order to obtain superfluidity, and what the difference is between dilute superfluid Bose Einstein condensates (BEC) and superfluid helium. This also explains

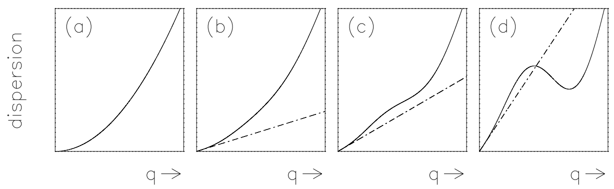


Figure 5: (a) An ideal Bose gas forms a BEC, but because the particles are non-interacting, the dispersion will be quadratic in momentum $q$. As a result, there is no minimum slope to the dispersion curve, and the system is not superfluid. (b) In a weakly interacting Bose system, the presence of the condensate renormalizes the low-q part of the dispersion [4]; the result is a minimum slope (dashed line), and the system is a superfluid. (c) Once the system becomes strongly interacting, the dispersion curve will also reflect the local structure of the liquid. This does not affect its superfluid properties, even though the condensate fraction can be depleted. (d) In a stongly interacting system, we still find a minimum slope to the dispersion, but unlike in weakly interacting Bose systems, the actual minimum value of the slope is no longer determined by the Bose condensate, that is, by the slope given by the low-q part of the dispersion curve (dashed line).

why the roton excitation is the determining factor in superfluidity in helium, whereas such an excitation is not even present in dilute BECs. Bogoliubov showed [4] that the low momentum part of the dispersion of a dilute Bose system renormalizes from a quadratic dependence $\sim q^2$ to a linear dependence $\sim q$ in the presence of a Bose condensate. This is shown in Fig. 5b. By and large, this is the situation in the experimentally realized BECs where superfluididty has been established [19]. There now exists a non-zero slope to the dispersion, and by virtue of the Landau criterion, such a system will exhibit superfluidity. This was verified in 2005 by Zwierlein *et al.* [19]

With increased ensity of the liquid (Fig. 5c and d), the dispersion curve will also reflect the local structure of the liquid. Compared to the (barely) interacting liquid, it will take more energy to create a hole in the liquid, and less energy to impose a fluctuation that already matches the natural separation between the atoms. This results in an increase in excitation energy in the maxon region, and a decrease in the roton region. Once the density is high enough, the slope that is tangent to the excitation energy of the roton will be lower than the slope given by the low momentum excitations. This is how the roton excitation energy becomes the determining factor for the critical velocity of superfluid helium; this effects is not present in dilute BECs.

## 2.4 Feynman's explanation for superfluidity

Next, we will explain why there is this roton energy gap in the first place, how the size of this gap relates to the superfluid transition temperature, and how actual values for all parameters involved can be estimated. To be clear, with a gap we mean that the minimum excitation energy (of the roton) is non-zero. We would also call the liquid shown in Fig. 5b a liquid with a gapped excitation spectrum, whereas we refer to Fig. 5a as a non-gapped liquid since this dispersion does not have a minimum slope associated with it.

Feynman explained in a beautiful argument why this energy gap is the unavoidable consequence of the fact that $^4$He atoms obey Bose statistics. We refer the reader to Feynman's 1955 account [7] and 1972 textbook [8] for details, but in a nutshell the argument is the following.

Assume that a certain configuration of the helium atoms represents the state with lowest energy, the ground state. The quantum mechanical wave function $\phi$ of this state depends on the positions of all atoms: $\phi(\overrightarrow{R}_1, \overrightarrow{R}_2, ..\overrightarrow{R}_N)$. The energy of this ground state consists of a kinetic energy term that depends on the gradient of the wave function $\sim |\nabla\phi|^2$, and of a potential term $V|\phi|^2$. The same holds for the wave function $\psi(\overrightarrow{R}_1, \overrightarrow{R}_2, ..\overrightarrow{R}_N)$ describing the excited state that is lowest in energy of all excited states. The potential operator has terms $\sim 1/|\overrightarrow{R}_i - \overrightarrow{R}_j|^n$ which tell us that the force between the atoms is strongly repulsive when they are too close together, which is another way of saying that atoms cannot occupy the same space.

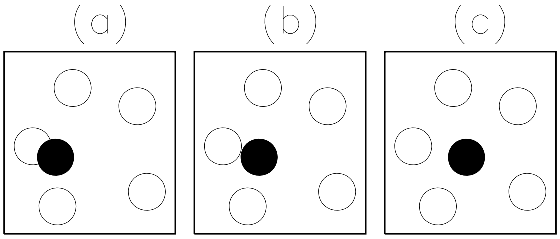


Figure 6: A depiction of various configurations representing different energies [8]. Part (a) shows an unlikely ground state configuration since two atoms being in the same spot implies a high potential energy. Similarly, part (b) shows an unlikely ground state because this configuration would represent a high kinetic energy (see text). By and large, in the ground state the atoms will be spread out as shown in part (c).

The exact details of the ground state are not important, but both the kinetic and potential term should be small. From this requirement we can expect that the atoms in a configuration that could represent the ground state are fairly well spread out (see Fig. 6c). After all, if they were to sit on top of each other (Fig. 6a), we would have to pay a high price in potential energy, and hence, we can assume the amplitude of the ground state wave function $\phi$ to be zero for those cases where $R_i \approx R_j$.

Also, the atoms will not be very close to each other (Fig. 6b), because this would correspond to a high gradient $\nabla\phi$, making it an unlikely choice of ground state. We can verify that atoms almost touching each other would correspond to a high gradient as follows: if the amplitude of $\phi$ would not be zero for a configuration where two atoms are very close, then we would have the situation that by slightly changing the coordinate of one atom to make it sit on top its neighbor (going from Fig. 6b to 6a), we would go from a non-zero to a zero amplitude for $\phi$. This, in turn, would imply a steep gradient $\nabla\phi$ (high kinetic energy), and therefore, the amplitude of $\phi$ must also be zero for configurations where atoms are too close.

Another way of saying the above is that if an atom actually were to move from Fig. 6b to 6a, it must have had a large kinetic energy in the first place in order to be able to approach the other atom as closely as shown in Fig. 6a. However, note that we do not actually 'move' atoms, we just compare the wave function for two different configurations $\overrightarrow{R}^N$. Thus, when we say 'move', we do not imply any dynamics. Rather, we are trying to get our mind around quantum mechanical manipulations by evoking classical images.

We also know that $\phi$ (Fig. 6c), being the ground state, will not have any nodes other than at the edge of the box confining the liquid. Hence, we can assume without loss of generality that the amplitude of $\phi$ is positive for all configurations. Also, the excited state wave function $\psi$ should have at least one more node. This is equivalent to the first harmonic of a guitar string not having a node other than where it is being held in place, while the second harmonic has a node in the middle of the string. Also note that the second harmonic is a higher note, representing higher energy, similar to the excited state in helium having a higher energy than the ground state.

The preceding statements imply that half of the configurations representing the excited state $\psi(\overrightarrow{R}_1, \overrightarrow{R}_2, ..\overrightarrow{R}_N)$ should correspond to a positive amplitude, and half of the configurations should correspond to a negative amplitude [20]. Next, we will try to create an excited state that barely differs in energy from the ground state. If we were to succeed, then the roton minimum would be close (in energy) to the ground state (yielding a very small minimum slope), and we would be able to stifle superfluidity. However, Feynman showed that this cannot be done [7], and that one always has to end up with a sizeable energy difference between the ground state and the excited states for excitations that are not phonons.

In Fig. 7 we have sketched a configuration that we arbitrarily will take to correspond to a maximum positive amplitude for $\psi$. We do not really know what the configuration should look like, but we tried to make it look like the ground state, with atoms not sitting on top of each other. Next, we will rearrange the atoms to end up with a configuration that would correspond to a maximum negative amplitude. To achieve this, we should rearrange the helium atoms over large distances. We are not interested in short distances, because this would imply that $\psi$ goes from a maximum to a minimum over short distances, which would correspond to a large gradient $\nabla\psi$, and therefore, to a

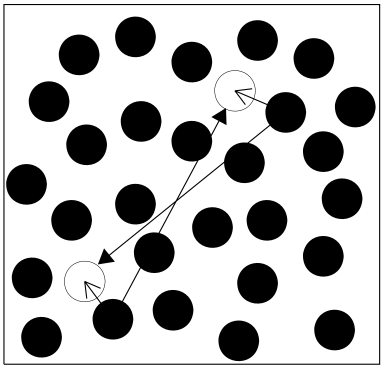


Figure 7: A reasonable guess of what a low energy state $\psi$ in helium could look like [8]. In an attempt to flip the sign of $\psi$, atoms are 'moved' over large distances (long arrows) while smoothing out any holes left in the liquid to minimize the energy cost. However, since Bose particles cannot be distinguished from each other and since permutation of particles does not affect the wave function, the outcome of the 'movements' indicated by the solid arrows are identical to the 'movements' indicated by the short arrows.

high energy. We will also smooth out any holes or bumps that may materialize, otherwise we would end up with an excitation that would look like a phonon (see Fig. 3) and we already know that a phonon does not represent the minimum slope of the excitation curve. The required changes to the configuration are shown in Fig. 7 by the long arrows, and we appear to have achieved our aim.

However, the above approach does not work in a Bose liquid since all the atoms are indistinguishable: interchanging two atoms does not lead to a change in the amplitude of the wave function. Thus, one could have gotten the same final configuration by simply 'moving' the atoms affected by the rearrangement over distances less than half the atomic separation (denoted by the short arrows in Fig. 7). In fact, half the atomic separation is the best that one could hope to achieve. However, such rapid variations (from maximum positive amplitude to maximum negative amplitude over half the atomic separation) would represent a large gradient and thus, it would signify a significant step up in energy. This is the exact opposite of our intended goal of creating a configuration for an excited state that would be very close in energy to the energy to the ground state.

In other words, because of the Bose nature of the atoms, it is not possible to make an excited state (which is not a phonon) that differs by a vanishingly small amount in energy from the ground state. Therefore, an energy gap must be present in a Bose liquid and provided the liquid does not freeze, it must become a superfluid. Whether a Bose-Einstein condensate forms or not is not relevant to this argument since it links the property of superfluidity to the scarcity of excited states. Even a Bose liquid where only a tiny fraction of the atoms condenses will (necessarily) become a superfluid when the temperature is low compared to the energy gap.

Thus, from a qualitative point of view, it is clear why a Bose liquid that remains liquid down to low enough temperatures has to become a superfluid: Bose statistics force the emergence of an energy gap. However, this argument does not provide us with a numerical estimate for the size of the gap. Moreover, it does not even tell us how the transition temperature is linked to the size of a gap. This may explain why introductory textbooks tend not to mention Feynman's arguments.

# 3 Superfluidity: quantitative understanding

So how low does the temperature have to be for the liquid to become a superfluid? Since we do not think that the temperature at which BEC occurs in an ideal Bose gas (for the record [8], at 3.13 K) has much to do with the magnitude of the energy gap in a real liquid, we must take a different approach. We first discuss the relationship between the size of the energy gap and the superfluid transition temperature, followed by a discussion on how to estimate the size of the gap based on the speed of sound and on the particle density of the liquid. Doing so, we will end up with an accurate estimate of the superfluid transition temperature. Our estimates apply to $^4$He at zero pressure, possible extensions to higher pressures are given in the references.

As a note of caution, while our estimates turn out to be remarkably accurate, the following should not be read as anything other than being a set of instructional exercises intended to help students in their understanding of liquid helium in particular, and in using their acquired knowledge from introductory physics to real world problems. Thus, the following is not a rigorous derivation of the transition temperature in superfluid helium, even though our estimate turns out to be very accurate.

## 3.1 Connection of $T_\lambda$ to the excitation gap

We first apply the same reasoning as in the preceding section to figure out what the minimum velocity requirement is for a single helium atom to be able to slow down. For this atom of mass $m$ moving with speed $v$ through the sea of other atoms to be able to slow down to speed $v'$ we have:

$$\begin{aligned} \tfrac{1}{2}mv^2 &= \tfrac{1}{2}mv'^2 + E_{ex}(p) \\ m\vec{v} &= m\vec{v}' + \vec{p}_{ex}. \end{aligned} \tag{4}$$

Since we are interested in the bare minimum, we assume that this atom will give up all of its energy [$v'$=0]. Dividing the above equations we get a minimum condition on the speed of the atom for it to be able to transfer energy to the rest of the liquid:

$$v_{min} \geq 2[E_{ex}/p_{ex}]_{min} \tag{5}$$

Compared to eqn 3 we have picked up a factor of 2, which is the result of dealing with a small mass $m$ instead of with the large mass $M$ of a moving liquid. Note that the actual mass of the atom does not play a role in this expression. This ensures the validity of eqn 5 in real liquids in which the actual movement of an atom is accompanied by a flow pattern where atoms are temporarily pushed out of the way, bestowing the moving atom with an effective mass which is larger than the mass of a non-interacting atom (about 2∼3 times larger, see ref [21] for details).

If an atom moves faster than this minimum velocity requirement, then it can create an excitation in the rest of the liquid and slow down; if it moves slower then it cannot and it will not be able to slow down (that is, it will not experience any friction). Thus, we should expect to see a qualitative difference in behavior of the liquid above and below the temperature at which the average thermal speed $v_{thermal}$ meets the minimum requirement contained in eqn 5.

In order to estimate this temperature, we assume that the classical equipartition of energy principle can be extended to quantum liquids at low temperature, namely

$$\frac{1}{2}mv^2_{thermal} = \frac{3}{2}k_B T, \tag{6}$$

with $k_B$ being Boltzmann's constant. Combining eqn 6 with eqn 5, we find that we can expect changes in liquid behavior at a temperature $T_\lambda$ when $v_{thermal} = v_{min}$:

$$T_\lambda = \frac{4m[E_{ex}/p_{ex}]^2_{min}}{3k_B}. \tag{7}$$

We read off the value for the minimum of the slope [21] from Fig. 2: $(E_{ex}/p_{ex})_{min}$= (58.05 ± 0.15) m/s, which combined with the mass of a helium atom of 6.646 x $10^{-27}$ kg yields a transition temperature of $T_\lambda = 2.162$ K ± 0.012 K, in good agreement with the actual transition temperature of 2.17 K. The fact that the agreement is essentially perfect might be fortuitous, however, it does show that our assumption of being able to use the equipartition of energy theorem to be not too far off the mark [22]. Also, eqn 7 tells us that the transition temperature is determined by an intrinsic microscopic velocity of the liquid, as we would have expected for superfluidity.

## 3.2 Connection between the excitation gap and the speed of sound

Now that we have made the connection between the macroscopic transition temperature and the microscopic parameters for the roton, we can try to estimate these roton parameters based on other macroscopic quantities. We start with the value of the energy gap at the roton minimum. As an aside, we note that the entire excitation curve shown in Fig. 2 can in principle be calculated with great precision from first principles [23], including the flat part at higher momentum values ($p^* \geq 3$ Å$^{-1}$) and its termination at twice the roton energy [16]. The procedure is straightforward, but cumbersome. Since we only need the value of the energy gap at the roton minimum, we use the following more instructive shortcut.

As Feynman pointed out in his argument

as to why a Bose liquid should exhibit an energy gap [7], one gets the lowest lying excited state when one 'moves' atoms by half the atomic separation. We can use this to calculate the value of the energy gap based on the macroscopic speed of sound in superfluid helium.

From thermodynamics we have that the speed of sound $c$ is given by

$$c^2 = \frac{\gamma}{m}\left(\frac{\partial P}{\partial n}\right)_T \tag{8}$$

with $P$ the pressure, $\gamma$ the ratio of specific heat at constant pressure $c_p$ and at constant volume $c_v$, and $n$ the number density for $N$ atoms in a volume $V$: $n = N/V$. For superfluid helium at low temperature we find that $\gamma = 1$ [13] as a direct consequence of the large zero-point motion of the atoms [24]. Thus, at constant volume $V$ we have

$$c^2 = \frac{1}{m}\left(\frac{\partial PV}{\partial N}\right)_{T,V}. \tag{9}$$

We can now calculate $c$ by adding one more atom to the liquid at constant volume. In this case, $\triangle N$ =1 and $\triangle PV$ is the amount of work we have to do to make room for this additional atom.

We calculate the amount of work by comparing the energy of a configuration with a hole in it to the energy of a configuration without such a hole. We can make a cubic hole in the liquid by 'moving' atoms along the positive x-direction by half the atomic separation $d/2$, and by doing the same thing along the negative x-direction, and by repeating the process in the y and z-directions. Each of these 6 configurational changes should increase the energy of our state, namely by $E_{roton}$ for every step. This can be seen as follows.

Provided we can 'move' atoms over a distance of at least $d/2$, and provided we have plenty of room to smooth out any variations in local particle density, then we should be able to do each of the six steps of the process at a minimum cost in energy. Thus, the total cost will be six times the minimum excitation energy in liquid helium, or $6E_{roton}$. Of course in doing this, we actually made the hole too big (volume $d^3$) since we only needed to make a sphere of diameter $d$ [volume $4\pi(d/2)^3/3$]. Thus, the hole was too big by a factor of $\pi/6$; we only needed to provide $6E_{roton}\pi/6 = \pi E_{roton}$ in work. Combining this with eqn 9 we find [25]

$$E_{roton} = mc^2/\pi. \tag{10}$$

To see how reliable an estimate this is, we compare this prediction to the measured quantities of superfluid helium at 1.2 K. Using c = 237.4 m/s [13] (Fig. 1), we obtain $E_{roton}$ = 11.92 x $10^{-23}$J = 0.744 meV. The value that has actually been measured by means of neutron scattering [11] is 0.743 meV. Thus, we have found a very accurate value for the energy gap based on the speed of sound. In essence, we have used the speed of sound to gauge the strength of the interatomic potential, which in turn determines the roton energy.

### 3.3 Connection of the roton excitation to the density of the liquid

So far we have connected the superfluid transition temperature to a minimum flow velocity determined by the excitation curve shown in Fig. 2, and we have connected the minimum excitation energy $E_{roton}$ to the macroscopic speed of sound. To complete a quantitative estimate of the superfluid transition temperature, we should also estimate the momentum value corresponding to this minimum excitation energy, so that we can get an overall estimate of the minimum value of $E_{ex}/p_{ex}$ for the excitations shown in Fig. 2.

As can be seen from Fig. 2, $E_{roton}/p_{roton}$ is an accurate estimate for the minimum slope, only fractionally too large. The roton momentum $p_{roton}$ is given by $p_{roton} = h/\lambda_{roton} = h/d$, with $h$ Planck's constant and $d$ the interatomic separation. In words, the energy cost to create an excitation of wavelength $\lambda$ is least when the wavelength of this disturbance matches the natural length scale in the liquid, namely the atomic separation. This can also be seen in Fig. 7 which contains a roton excitation, yet the average separation between the atoms is pretty much as it is in the ground state. Had it not been the case, then the energy of the excited state would have been much higher.

The separation $d$ depends on the number density $n$ of liquid helium as $d \sim n^{-1/3}$. For a solid consisting of spheres packed in a cubic structure the atomic packing factor $nd^3$ would be exactly 1 since one atom occupies a volume of $d^3$, so that $d = n^{-1/3}$. Estimating the proportionality factor in a liquid is a somewhat nebulous undertaking because unlike in a solid, we do not have a nice periodic arrangement. We do expect that the average separation in a liquid will be smaller than that in a simple cubic structure since the atoms are not packed as tightly as possible. For our estimate we will use that helium fairly accurately resembles a liquid of closely packed spheres. This is shown in Fig. 8.

The atomic packing factor for atoms in a close packed structure is $\pi/3\sqrt{2}$= 0.741 [26]. If we identify this close-packed packing factor with $nd^3$, then we obtain the folowing estimate for the average atomic separation in a liquid resembling a close packed structure [27]

$$d = [\pi/(3n\sqrt{2})]^{1/3}; p_{roton} = h[\pi/(3n\sqrt{2})]^{-1/3}. \tag{11}$$

Note that the above average separation is slightly lower than that for a simple cubic structure ($d = (1/n)^{1/3}$), something which simply tells us that the atoms are closer together in a more closely packed structure. Putting in the helium density ( n= 0.02183 atoms/Å$^3$), we find $p^*_{roton}$= 1.94 Å$^{-1}$. This number is spot on, so it would appear that eqn 11 provides us with a good estimate of the interatomic separation.

We are now in the position to combine all our estimates into one expression for the su-

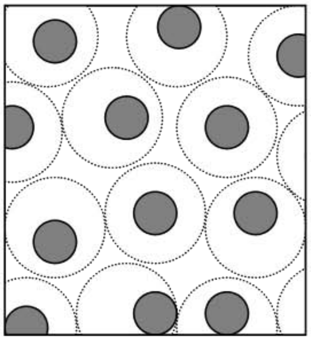


Figure 8: The average separation between atoms is, by and large, given by a closely packed structure. The atoms, depicted by the grey disks, cut out a volume for themselves (shown by the larger, transparent disks) by colliding with their neighbors and thereby keeping them at a distance. Since all atoms have very similar kinetic energies, we can expect the cages that they create to be of very similar sizes. These cages are stacked in a pattern that closely resembles a close packed structure.

perfluid transition temperature. Combining eqns 7, 10 and 11 we get

$$T_\lambda = \frac{4m}{3k_B}[\frac{mc^2(\pi/3\sqrt{2})^{1/3}}{h\pi n^{1/3}}]^2. \qquad (12)$$

When we plug in all the numbers (c= 237.4 ± 0.5 m/s, n= 0.02183 atoms/Å$^3$), we find $T_\lambda$= 2.17 ± 0.02 K. Thus, when we combine our quantitative approximation for the relationship between the minimum of the dispersion curve and $T_\lambda$ (eqn 5) with our approximation for the roton energy (eqn 10) and with our approximation for the roton position (eqn 11), we still find very good agreement with experiment.

While it is satisfying to have ended up with such a good agreement, we note that the agreement is probably better than we had reason to expect given the simplicity of our estimates. From an instructional point of view however, we consider the individual links that we have made between microscopic parameters and macroscopic quantities (that is, Eqs, 5, 10 and 11) to be the most important points of this section since it allows students to apply basic physics reasoning in order to arrive at predictions for measurable quantities.

In summary, we have shown that superfluidity can be explained to students without going into lengthy calculations, and without having to invoke details about a Bose condensate. We have tied Feynman's arguments about the origin of the energy gap to the actual superfluid transition temperature, and we have shown that very accurate estimates of all parameters involved can be obtained through straightforward reasoning. While we have included the actual numerical values for our calculations in this paper to make the discussion less abstract, and while we have even given an expression of the superfluid transition temperature in terms of macroscopic quantities such as the density and speed of sound, the real message of the paper is that students should be able to develop a better sense of what causes the property of superfluidity in terms of Bose statistics and energy gaps; that is, better compared to the standard Bose-Einstein condensation remarks that are normally encountered in textbooks. Finally, since a very similar relationship exists between the transition temperature and the energy gap in superconductors, this paper could also serve as an introduction to the physics of superconductivity.

# References

[1] P. L. Kapitza, Nature **141**, 74 (1938).

[2] J. S. Brooks and R. J. Donnelly, J. Phys. Chem. Ref. Data **6**, 51-104 (1977).

[3] F. London, Nature **141**, 643-644 (1938).

[4] N. N. Bogoliubov, J. Phys. (USSR) **11**, 23-32 (1947).

[5] L. D. Landau, J. Phys. (USSR) **11**, 91-92 (1947).

[6] H. Palevsky, K. Otnes, K.E. Larsson, R. Pauli, and R. Stedman, Phys. Rev. **108**, 1346-1347 (1957).

[7] R. P. Feynman, in *Progress in Low Temperature Physics*, Vol. I, edited by C.J. Gorter (North-Holland, Amsterdam, 1955).

[8] R. P. Feynman, *Statistical Mechanics: A Set of Lectures* (Benjamin, Reading, MA, 1972).

[9] Wouter Montfrooij and Ignatz de Schepper, *Excitations in simple liquids, liquid metals and superfluids* (Oxford university Press, Oxford, 2010).

[10] H.R. Glyde, R.T. Azuah and W.G. Stirling, Phys. Rev. B **62**, 14337-14349 (2000).

[11] R. J. Donnely, J. A. Donnely and R. N. Hills, J. Low Temp. Phys. **44**, 471-489 (1981); H. R. Glyde, M. R. Gibbs, W. G. Stirling and M. A. Adams, Europhys. Lett **43**, 422-426 (1998).

[12] The reader can verify that a linear relationship is common to all systems by listening to a piece of music and noticing that the high and the low tones travel just as fast, which can only happen provided $E/p$ is constant for small $p$.

[13] J. Maynard, Phys. Rev. B **14**, 3868-3891 (1976).

[14] Different visualizations of what exactly a roton is have been put forward over the years, although, of course, it is not really possible to visualize an wave-like particle. For instance, it can be described as a sound wave with a wavelength corresponding to the interatomic spacing, Onsager called it the 'ghost of a vanishing vortex ring', while Chester viewed it as 'one atom moving very fast through the sea of others'. Our arguments do not depend on exactly what a roton is.

[15] See for instance, C. Kittel, *Thermal Physics* (Wiley, New York, 1969).

[16] W. Montfrooij and E. C. Svensson, Journ. of Low. Temp. Phys. **121**, 293-302 (2000).

[17] E. Varoquaux, G.G. Ihas, O. Avenel, and R. Aarts, Phys. Rev. Lett. **70**, 2114-2117 (1993).

[18] P. C. Hendry, N. S. Lawson, P. V. E. McClintock, and C. H. D. Williams, Phys. Rev. Lett. **60**, 604-607 (1988).

[19] M. W. Zwierlein, J. R. Abo-Shaeer, A. Schirotzek, C. H. Schunck, and W. Ketterle, Nature **405**, 1047-1051 (2005).

[20] The ground state and excited state wave functions are orthogonal in the following sense: $\phi.\psi = 0 = \int \phi\psi d\overrightarrow{R}^N$. Since $\phi$ does not have any nodes (does not change sign), we must have that $\psi$ is positive for half of all possible configurations $\overrightarrow{R}^N$, and negative for the other half.

[21] W. Montfrooij and I. M. de Schepper, Phys. Rev. **B 51**, 15607-15609 (1995).

[22] The above arguments can be extended to helium under pressure; we refer the reader to reference [21] for details.

[23] Deok Kyo Lee and Felix J. Lee, Phys. Rev. B **11**, 4318-4330 (1975); V. Apaja, and M. Saarela, Phys. Rev. B **57**, 5358-5361 (1998).

[24] The difference in specific heats is proportional to the thermal expansion coefficient $\alpha$ as $c_p - c_v \sim \alpha^2$. At low temperatures, $\alpha = 0$ because the density of helium is determined by the large zero-point-motion of the particles; reducing the temperature will not lead to any increase in density.

[25] This relation [eqn 10] can be generalized to cover higher pressures by including a term $\sim PV$.

[26] This number can be verified by stacking marbles and comparing the volume of all the marbles to the total volume of the resulting structure.

[27] The inverse of this relation ($p^*_{roton} = 2\pi/d$) gives an accurate prediction for the position of the roton minimum for all pressures, which is known to vary as $\sim n^{1/3}$, see ref. [13].

# Studies on Newtonian chirp from inspiraling compact binaries

Saikruba .K

Department of Theoretical Physics
University of Madras
Chennai 600025, India.
krubakrish@gmail.com

(Submitted 24-06-2016)

---

**Abstract**

In this article we study the gravitational waves produced from inspiral phase of compact binaries in circular orbit using the Newtonian chirp of leading order post-Newtonian(PN) approximation. Expressions governing the Newtonian chirp are useful for obtaining the inspiral waveform, frequency variation plot, merging time value and to plot the amplitude spectral density. We start with the discussion of Newtonian chirp, where we assume the compact binaries to be located at galactic center to obtain the waveform and frequency variation plots. For the singular value of peak frequency seen at end of inspiral phase in Newtonian chirp, the concept of gravitational wave frequency corresponding to inner most stable circular orbit(ISCO) provided by general relativity is described. The qualitative differences observed in chirp duration and frequency evolution of these systems is explained. We also take the GW150914 event parameters reported by LIGO team to obtain the inspiral waveform from Newtonian chirp with $F_{ISCO}$ as inspiral limit and overlay on LIGO discovered full waveform to show the merger and ringdown phase of a real signal. We plot the amplitude spectral density of various compact binaries along with GW150914 and GW151226 events and overlay it on LIGO and aLIGO sensitivity curve; it establishes the merging events and possible distances which can be detected in detectors. We also give a brief discussion on the supermassive black hole binary and slowly orbiting binary systems. We also bring out the contribution of higher order PN terms to number of cycles for various compact binaries. Inferences are drawn from these studies which reveals useful insights and information.

---

## 1 Introduction

Einstein's theory of general relativity describes gravity as curvature of space time. Einstein's equation are non-linear whereas the linearised Einstein's equation in free space can be written down as a wave equation [1]. In leading order the gravitational waves are produced by a time varying mass-energy quadrupole moment. This is because for a given mass-energy configuration the monopole and dipole moment corresponds to total energy and total angular momentum of the system respectively. The laws of conservation of mass-energy and angular momentum prohibits any change in monopole and dipole moment. Binary stars are continuous sources of gravitational waves because its quadrupole moment when calculated for circular and elliptical orbit allows non-vanishing time derivatives [1].

The quadrupole moment is defined as [2],

$$I_{ij} = \int \rho(r) r_i r_j d^3 r \qquad (1)$$

where $\rho(r)$ is mass-energy density.

Consider a binary system where the masses $m_1$ and $m_2$ are moving in circular orbit about their common center of mass with an angular velocity $\omega$. We assume the masses to be confined in x-y plane and the distance between them to be $a$. In the center of mass frame this can be reduced to a one body problem, where a mass $\mu$ moves in a circular orbit of radius $a$. We have $\mu = m_1 m_2/(m_1 + m_2)$ which is the reduced mass and $M = m_1 + m_2$ is the total mass of the binary. The position coordinates are $x = acos(\theta)$, $y = asin(\theta)$ and $z = 0$.

$$I_{xx} = \mu xx = \mu a^2 \cos^2(\omega t) = \frac{\mu a^2}{2}(1 + \cos(2\omega t))$$

$$I_{yy} = \mu yy = \mu a^2 \sin^2(\omega t) = \frac{\mu a^2}{2}(1 - \cos(2\omega t))$$

$$I_{xy} = I_{yx} = \mu xy = \mu a^2 \cos(\omega t)\sin(\omega t) = \frac{\mu a^2}{2}\sin(2\omega t)$$

$$I_{xz} = I_{zx} = I_{yz} = I_{zy} = I_{zz} = 0$$

The presence of $r_i r_j$ in quadrupole moment $I_{ij}$ results in $\cos^2 \omega t$ function which gives $\cos(2\omega t)$ function. We know that the phase of $\cos(2\omega t)$ is twice than that of $\cos(\omega t)$. Hence the frequency of $\cos(2\omega t)$ is twice of $\cos(\omega t)$.

The time varying quadrupole moment govern the gravitational waves produced, thus for a circular binary in leading order the frequency of gravitational waves comes out to be twice the orbital frequency of binary [2].

## 2 Newtonian Chirp

Due to the emission of gravitational waves the stars inspiral and the orbital frequency of the system increases as per Kepler's third law. In the leading order post-Newtonian approximation we compute the GW signal. The emitted gravitational waveform is a "chirp" signal due to its increasing amplitude and frequency. We consider the masses to be point objects, hence their size and radius doesn't come into picture in any of the expressions. The essential expressions of this approximation presented in this section is based on Ref. [3] & [4].

The distance $D$ of the binary system from earth, the masses $m_1$ and $m_2$ in the binary and the frequency of orbital motion $F_{system}$ are the determining parameters of gravitational waves produced by the system. The gravitational wave parameters like amplitude, frequency and strain are closely related to $D$, $m_1$, $m_2$ and $F_{system}$ parameters of the binary. We usually consider the chirp mass of the system, which is

$$\mathscr{M} = \mu^{3/5} M^{2/5} \qquad (2)$$

The time varying strain of gravitational waves is given by

$$h(t) = A(t)\cos\phi(t) \qquad (3)$$

where $A(t)$ is the amplitude and $\phi(t)$ is the gravitational wave phase.

The strain is a time-varying signal due to the time dependence of amplitude and gravitational wave phase. As the binary system evolves with time its orbital radius decreases and frequency increases. The energy emitted due to gravitational waves is large in smaller radii orbit, therefore the amplitude and phase is expected to change with time. In the approximation considered the amplitude and phase formula are

$$A(t) = \frac{4\mathscr{M}^{5/3}\pi^{2/3}F(t)^{2/3}}{D} \tag{4}$$

$$\phi(t) = \phi_0 - 2\left[\frac{1}{256(\pi f_0 \mathscr{M})^{8/3}} - \frac{t}{5\mathscr{M}}\right]^{5/8} \tag{5}$$

where $F(t)$ is the instantaneous gravitational wave frequency and $f_0$ is the initial frequency of the wave received in detectors. The frequency variation with time is given by

$$F(t) = \frac{(\mathscr{M} f_0^9)^{1/8}}{\left[(\mathscr{M} f_0)^{1/3} - 256 f_0^3 \mathscr{M}^2 \pi^{8/3}(t/5)\right]^{3/8}} \tag{6}$$

By defining the term

$$\tau_m = \frac{5}{256(\pi f_0)^{8/3}\mathscr{M}^{5/3}} \tag{7}$$

We have the gravitational wave frequency rewritten as,

$$F(t) = f_0\left(1 - \frac{t}{\tau_m}\right)^{-3/8}$$

Starting from an initial frequency at $t = 0$, the frequency of gravitational wave increases with time. This is to be expected since the instantaneous frequency of gravitational wave is twice the instantaneous frequency of the binary system from which it is produced. The separation between the masses is governed by the equation

$$a = a_0\left(1 - \frac{t}{\tau_m}\right)^{1/4}$$

Where at $t = \tau_m$ the separation between the masses becomes zero. This is the merging time at which the corresponding binary frequency $\omega^2 = M/a^3$ goes to $\infty$, thus we can see that at this instant the frequency of gravitational waves also shoots to $\infty$. The singular value of frequency is due to the merging of two point masses in which inspiral continues till $a = 0$. But it has been found in general relativity that not all the orbits are stable till $a = 0$ and the inner most stable circular orbit (ISCO) corresponds to $a = 6M_\odot$ [5]. Hence in post-Newtonian approximation, the inspiral phase is taken to end at this orbital separation and the corresponding gravitational wave frequency at this separation is given by

$$F_{ISCO} \approx \frac{4400}{M} Hz \tag{8}$$

Inspiral phase is followed by a short lived-merger phase. The merger phase begins when the masses starts to merge into a single object. After this the ringdown phase starts where the new object formed radiates away the deformations resulted from merging [6]. Gravitational waves are produced in all these phases. The merger phase is studied by solving Einstein's equation numerically whereas perturbation theory is required for ringdown phase.

Gravitational waves are continuously produced by binary systems and many of them might reach earth but can be too weak to be noticed in the LASER interferometer detectors. As the inspiral phase of the binary system continues, the frequency of orbital motion increases and if it reaches a value at which the signal is strong enough to be detected, we start seeing the GW from this frequency value (called as initial frequency). Ground based gravitational wave detectors are capable of detecting signals of frequency greater than 10 Hz. At lower frequencies the sources from space is overcome by seismic and other earth-based environmental noise sources [5]. Hence in this study we are interested in the beginning of signal in the detector at $t = 0$ starting at $f_0 = 10$Hz. The period where frequency and amplitude is increasing considerably is called as the chirp duration.

## 3 Binaries at galactic center

We reside in the milky way galaxy. The galactic center is located at 8 Kpc distance from Earth. We consider Black Hole-Black Hole (BH-BH) binary, Black Hole-Neutron Star (BH-NS) binary, Black Hole-White dwarf (BH-WD) binary, Neutron Star-Neutron Star (NS-NS) binary, Neutron Star-White dwarf (NS-WD) binary and White dwarf-White dwarf (WD-WD) binary being at our galactic center. Chandrasekhar's limit gives the maximum mass of a stable white dwarf star which is about $1.39M_\odot$. Most white dwarf stars are less massive than this value, so we have taken the mass of white dwarf star for Newtonian chirp study to be $1M_\odot$. The theoretical value of maximum mass of a stable neutron star is uncertain [7]. Tolman-Opphenheimer-Volkoff (TOV) limit gives the maximum mass of stable neutron star, the modern estimates range from approximately $1.5M_\odot$-$3M_\odot$. Equation of state for extremely dense matter is not well known and hence there exists an uncertainty in the limit. The observational data suggests that mass of most of the pulsars lies between $1.30M_\odot$ to $1.50M_\odot$ [7]. With this regard we take the mass of neutron star to be $1.4M_\odot$ in this study. Since there exists no well defined limit on mass of neutron star for gravitational collapse we consider the black hole to be of $10M_\odot$.

We assume that at time $t = 0$ the frequency of gravitational wave received in detector to be $f_0 = 10$ Hz. We can find the corresponding orbital frequency of the masses in the binary system, $f_{system} = f_0/2 = 5$ Hz. Gravitational waves takes time to travel from binaries to earth. On receiving in detector at $t = 0$ if the gravitational wave has initial frequency of 10 Hz it doesn't imply that the binaries are currently orbiting at 5 Hz, but it should be thought of as the information encoded about binary system in gravitational waves signal at $t = 0$. The aim is to see for a binary system undergoing 5 orbital rotation per second whose corresponding gravitational wave signal arrives with frequency 10 Hz on Earth, what will be the waveform and frequency variation as they inspiral to merge. We use the binary parameters: masses of two stars, distance from the earth and frequency of orbiting in the Mathematica code to obtain the plots in figure[1,2,3]. The details of Mathematica code which can be downloaded is given in appendix.

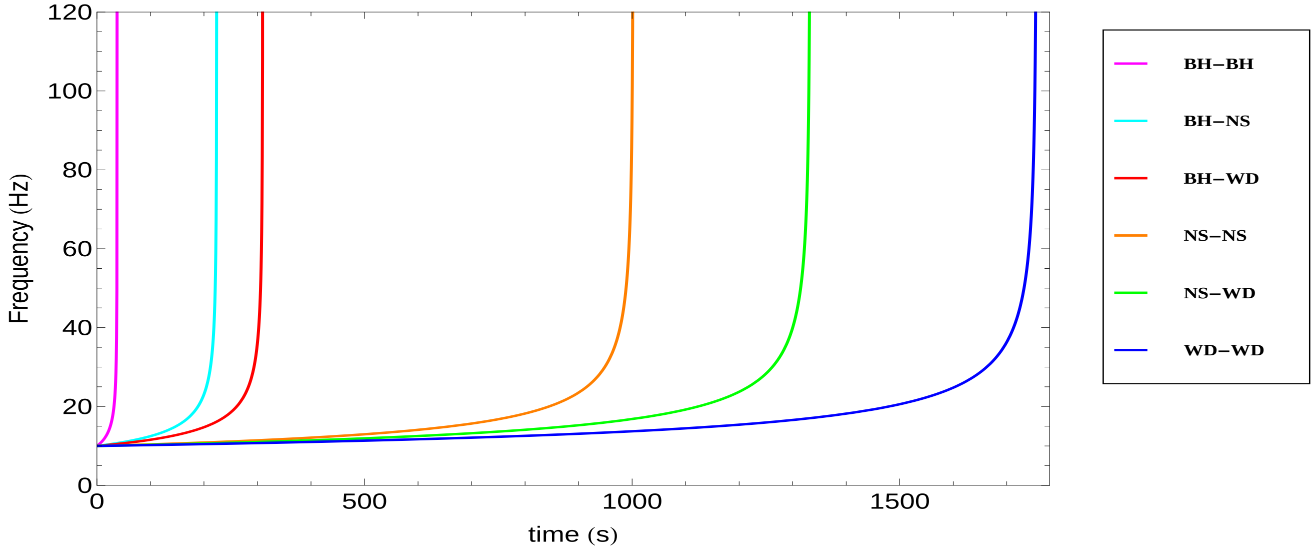


Figure 1 : Frequency variation plot of compact binaries

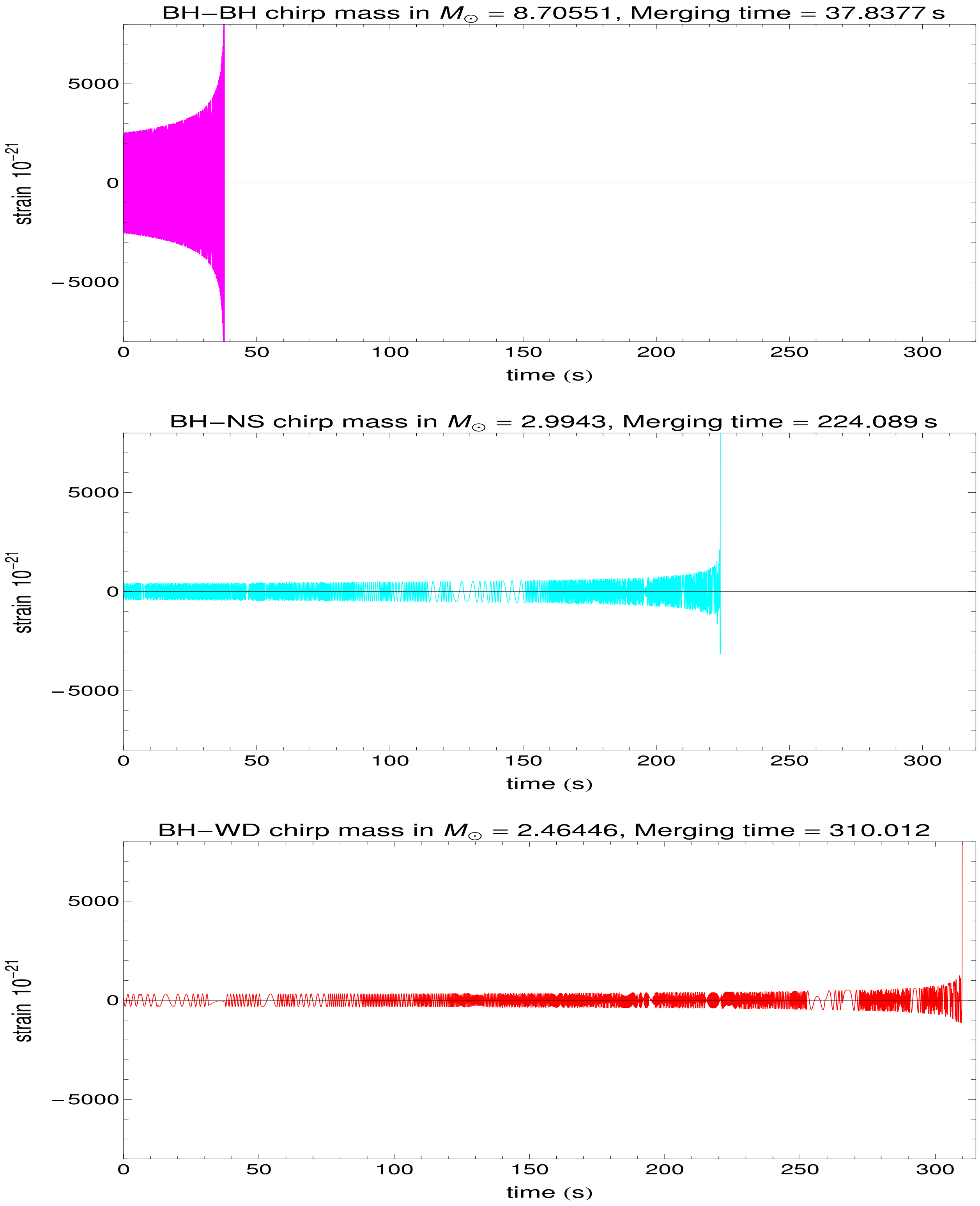
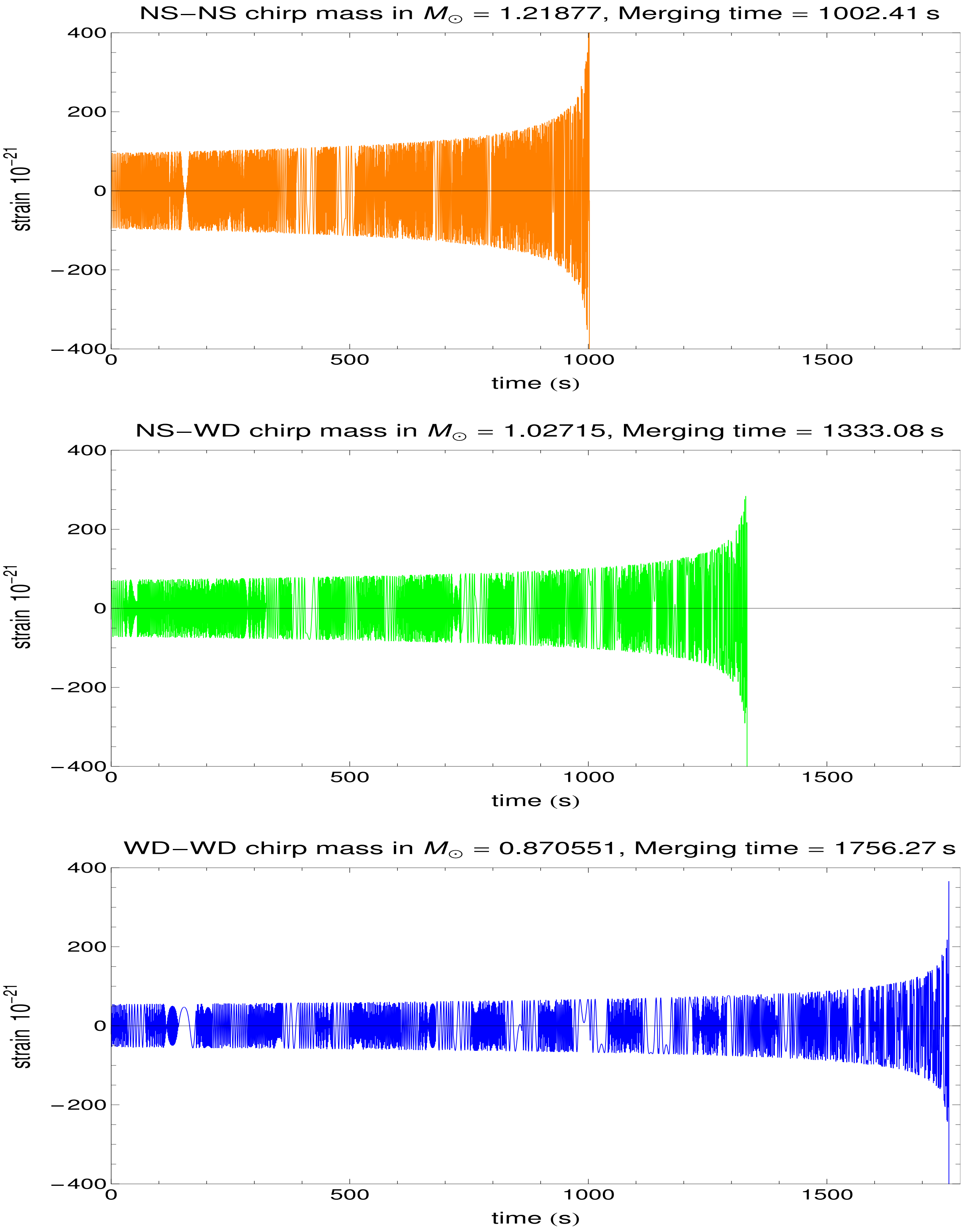


Figure 2: Waveforms of BH-BH, BH-NS, BH-WD at galactic center



NS−NS chirp mass in $M_\odot$ = 1.21877, Merging time = 1002.41 s

NS−WD chirp mass in $M_\odot$ = 1.02715, Merging time = 1333.08 s

WD−WD chirp mass in $M_\odot$ = 0.870551, Merging time = 1756.27 s

Figure 3: Waveforms of NS-NS, NS-WD, WD-WD at galactic center

It can be seen that the waveforms produced from Mathematica code starts at $t = 0$ and extends till the merging time $\tau_m$ which can be calculated from the formula. The frequency varies from the initial value starting at $t = 0$ and approaches to infinity at $\tau_m$.

Table 1: Compact binaries at galactic center

| Name | Masses($M_\odot$) | | | Peak strain(approx) | Merging time | Peak freq. |
|---|---|---|---|---|---|---|
| | $m_1$ | $m_2$ | $\mathscr{M}$ | $h_{peak}$ | $\tau_m$ (sec) | $F_{ISCO}$ (Hz) |
| BH-BH | 10 | 10 | 8.7005 | $8000 \times 10^{-21}$ | 37.8377 | 220 |
| BH-NS | 10 | 1.4 | 2.9943 | $2000 \times 10^{-21}$ | 224.089 | 385.96 |
| BH-WD | 10 | 1 | 2.46446 | $1000 \times 10^{-21}$ | 310.012 | 400 |
| NS-NS | 1.4 | 1.4 | 1.2187 | $400 \times 10^{-21}$ | 1002.41 | 1571.43 |
| NS-WD | 1.4 | 1 | 1.02715 | $250 \times 10^{-21}$ | 1333.08 | 1833.33 |
| WD-WD | 1 | 1 | 0.870551 | $200 \times 10^{-21}$ | 1756.27 | 2200 |

For binaries at same distance from Earth with same initial orbital frequency in terms of merging time $\tau_m$ we can arrange as-

$$\tau_m(BH-BH) < \tau_m(BH-NS) < \tau_m(BH-WD) < \tau_m(NS-NS) < \tau_m(NS-WD) < \tau_m(WD-WD)$$

For binaries at same distance from Earth with same initial orbital frequency in terms of peak strain $S_p$ we can arrange as-

$$S_p(BH-BH) > S_p(BH-NS) > S_p(BH-WD) > S_p(NS-NS) > S_p(NS-WD) > S_p(WD-WD)$$

With point mass consideration in terms of variation of frequency per unit time ($\delta V_f$) in 10 Hz to 40 Hz interval we can arrange as-

$$\delta V_f(BH-BH) > \delta V_f(BH-NS) > \delta V_f(BH-WD) > \delta V_f(NS-NS) > \delta V_f(NS-WD) > \delta V_f(WD-WD)$$

Massive binaries lose large amount of energy in the form of gravitational waves hence they inspiral more rapidly which can very well be seen in their frequency-time plot. They rapidly sweep through frequency interval 10Hz−40Hz compared to less massive binaries. They also undergo less number of cycles due to rapid inspiral which can be seen in their waveforms and they merge faster. In case of less massive binaries, since they emit small amount of gravitational they inspiral slowly, they sweep through frequency interval 10 Hz−40 Hz slowly undergoing large number of cycles and take comparatively longer time to merge. Since massive binaries emit more gravitational radiation during their inspiral phase than less massive ones their magnitude of strain is always more during the evolution stage and at the end of inspiral stage they attain peak strain larger than less massive ones.

Different binary systems end up their inspiral phase with final orbital frequency given by $F_{ISCO}/2$, which depends inversely upon the total mass of system. This is based upon the assumption that masses are treated as point objects hence in reality the binaries may end up with different orbital frequency before merging, as their size and radius may cause them to collide and merge prior to reaching peak frequency.

## 4 Supermassive black holes and slowly orbiting systems

Supermassive black holes can have masses ranging from few hundred solar masses to thousands of solar mass. We have considered a supermassive BH-BH binary with each BH having a mass of $100M_\odot$ and another such system with each BH having a mass of $200M_\odot$. The binaries are taken to be located outside our galaxy at 500 Mpc orbiting at 5 Hz. We have obtained its waveform and the frequency variation plot in figure[4] and the results tabulated in table 2.

Table 2 : SBH binary at 500 Mpc

| Masses($M_\odot$) | | | strain | $\tau_m$ | $F_{ISCO}$ |
|---|---|---|---|---|---|
| $m_1$ | $m_2$ | $\mathscr{M}$ | $h_{peak}$ | (sec) | (Hz) |
| 100 | 100 | 87.0051 | $10^{-21}$ | 0.815188 | 22 |
| 200 | 200 | 174.11 | $10^{-21}$ | 0.256768 | 11 |

For supermassive black hole binaries with masses $100M_\odot$ each and $200M_\odot$ each, orbiting at 5 Hz the merging happens within a second and their $F_{ISCO}$ values being 22 Hz and 11 Hz respectively lies very near to gravitational wave starting frequency 10 Hz. The study of these supermassive black holes reveal that by the time 5Hz is reached in their evolution stage which makes their gravitational signal cross the seismic noise frequency 10 Hz and detectable in ground based interferometer detectors they are at very end of their inspiral stage and merge within a second, hence rather than inspiral phase the merger and ringdown phases are more possible to be detected in aLIGO [8].

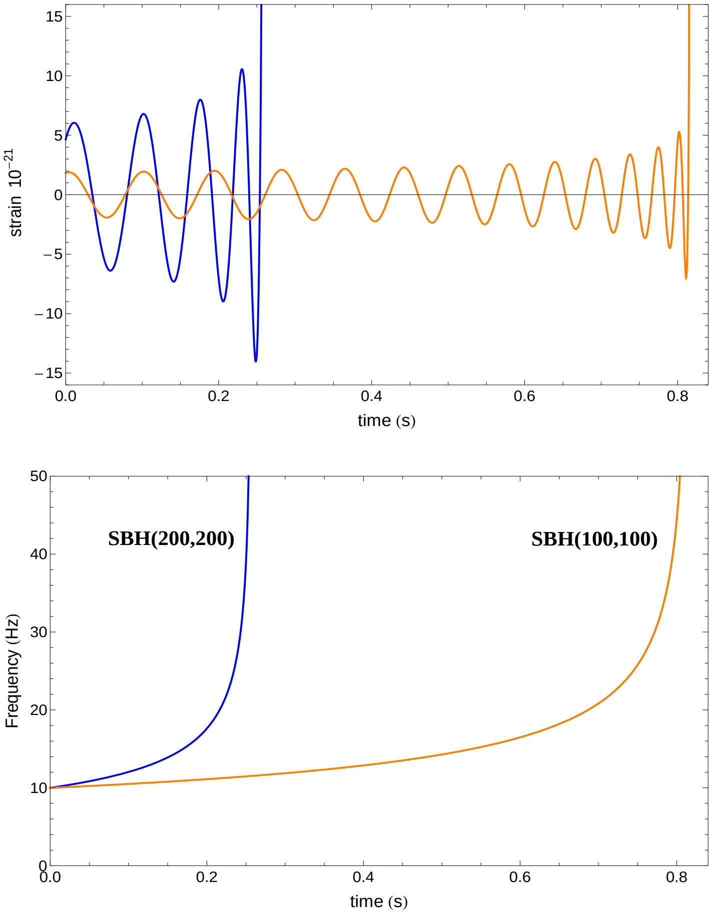


Figure 4: SBH binary at 500 Mpc

For supermassive black hole binary with mass $1000M_\odot$ the $F_{ISCO}$ comes out to be 4.4 Hz and for $10^6M_\odot$ the $F_{ISCO}$ comes out to be $4.4\times10^{-3}$, hence their gravitational waves in inspiral phase never reach 10 Hz to be observed in aLIGO. Considerable inspiral phase of GW from these supermassive black holes can be seen using Laser Interferometer Space Antenna(LISA) which has observable frequency in milli-Hertz [8].

The astrophysical process where merging of supermassive black holes can be seen is in merging of two galaxies. There exists observational evidence that every galaxy has a supermassive black hole at its center and hence the supermassive BH-BH merger is very likely to happen. Strong evidence of supermassive black hole binary exist in NGC 6240 [9], it is considered as a new galaxy formed by the merging of two different galaxies. The two black holes present in it are currently about 3000 light-years apart. The galaxy spans only 300000 light-years so it is expected that the black holes will merge to form a single black hole.

Not all binary systems orbit very rapidly to have a orbital frequency of 5 Hz, so we consider slowly orbiting binaries for our discussion. The orbital frequency of known double neutron star systems comes out to be about $10^{-6}$ Hz so we assume that initially the double neutron star system possesses this frequency. We also consider WD-WD binary orbiting at the same frequency. Hence, the arriving gravitational waves is seen with initial frequency of $2 \times 10^{-6}$ Hz. We assume both the systems to be located at 1pc distance. The waveform and frequency variation plot can be obtained for these systems, the results are tabulated in table 3.

Table 3 : Slowly orbiting binary system at 1 pc

| System | Masses($M_\odot$) | | | Peak strain order | Merging time | Peak freq. |
|---|---|---|---|---|---|---|
| | $m_1$ | $m_2$ | $\mathscr{M}$ | $h_{peak}$ | $\tau_m$ (sec) | $F_{ISCO}$ (Hz) |
| NS- NS | 1.4 | 1.4 | 1.21877 | $10^{-21}$ | $7.32765 \times 10^{20}$ | 1571.43 |
| WD-WD | 1 | 1 | 0.870551 | $10^{-21}$ | $1.28384 \times 10^{21}$ | 2200 |

The merging time of slowly orbiting NS-NS system comes out to be $7.32765 \times 10^{20}$ sec and WD-WD system comes out to be $1.28384 \times 10^{21}$ sec whereas the age of our universe is $13.799 \times 10^9$ yrs = $4.35165 \times 10^{17}$ sec. Hence compact binary systems which at present are slowly orbiting their merger is not likely to be seen in aLIGO anytime in future. The detections of NS-NS merger, NS-WD merger and WD-WD merger if is spotted in aLIGO then those systems must have evolved in cosmic times to reach this late inspiral stage or they had a very less orbital separation and relatively large orbital frequency at the time of formation or due to any other factor.

## 5 LIGO discovery

LIGO collaboration discovered gravitational waves produced from merging events. The events detected by LIGO are named as "GW" followed by date in YYMMDD format. Two of the events are the most promising merger events detected in year 2015 run of LIGO detectors. When GW150914 was received in detectors it had a strain which attained peak order of $10^{-21}$, this signal was seen to increase in frequency from 35Hz and reached 250 Hz in about 0.2 seconds. It was seen that in 8 cycles the signal increased from 35 Hz to 150 Hz [10]. The gravitational wave frequency of 150 Hz corresponds to the orbital frequency of 75 Hz for the binary system. For Newtonian point mass binary this gives 350 Km as the orbital separation. The WD-WD, NS-NS or NS-BH merger should occur at a much larger distance than this value hence LIGO collaboration suggested that the signal to be from a BH-BH binary. This was the first direct observation of black holes and their existence. The masses of binary system could be computed from data analysis and estimated to be $36 M_\odot$ and $29 M_\odot$ [10]. The luminous distance of this system was estimated to be 410 Mpc. The second event GW151226 when received in detectors it had a peak strain of the order $10^{-22}$, this signal was seen for about a second in which its frequency increased from 35 Hz to 450 Hz in about 55 cycles [11]. The binary parameters predicted from analysis was reported to be $14.2 M_\odot$ and $7.5 M_\odot$. The luminous distance of this system was estimated to be 440 Mpc. We take the prameters of GW150914 and GW151226 to obtain the Newtonian Chirp plots in figure[5], the results are tabulated in table 4.

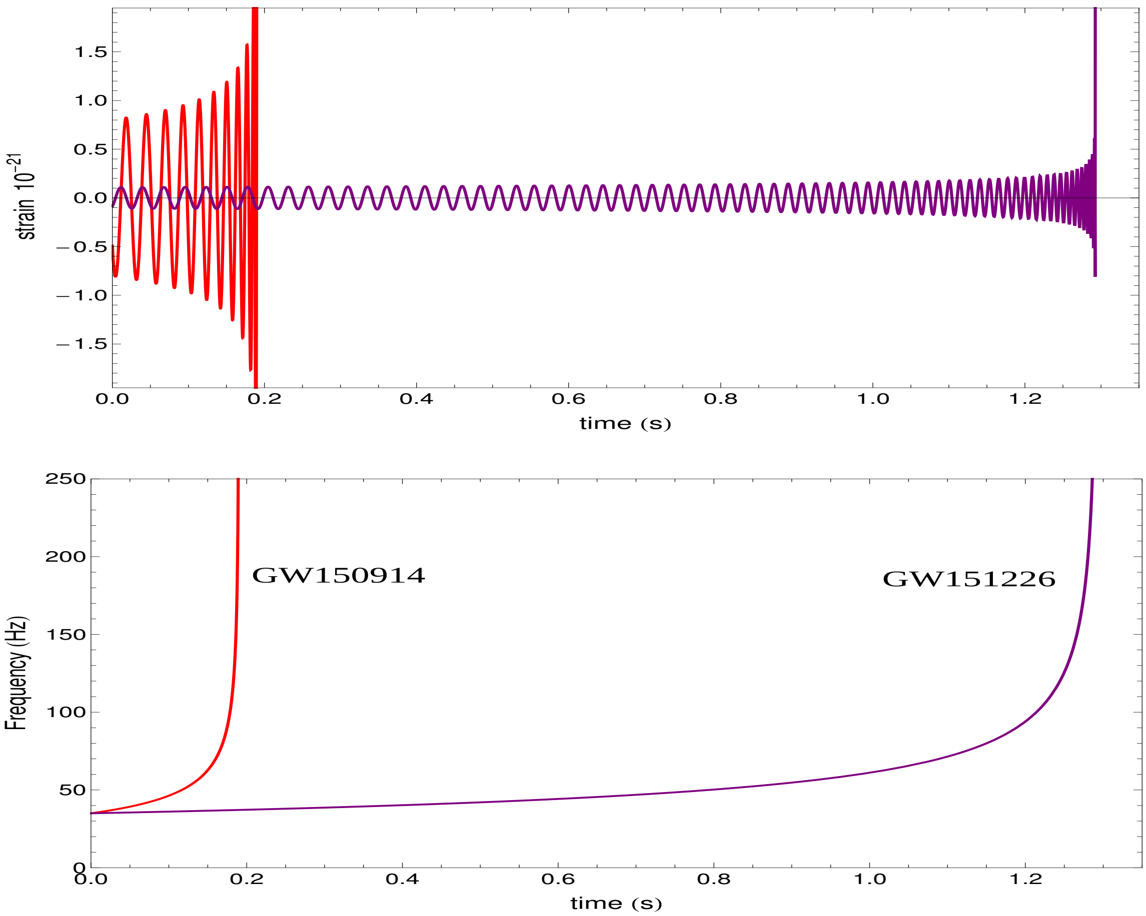


Figure 5: LIGO events Newtonian chirp

Table 4: Newtonian chirp of GW150914 and GW151226

| Data source | Masses($M_\odot$) | | | Peak strain order | Merging time | Peak frequency |
|---|---|---|---|---|---|---|
| | $m_1$ | $m_2$ | $\mathscr{M}$ | $h_{peak}$ | $\tau_m$ (sec) | $F_{ISCO}$ (Hz) |
| GW150914 | 36 | 29 | 28.0956 | $10^{-21}$ | 0.19 | 67.69 |
| GW151216 | 14.2 | 7.5 | 8.89442 | $10^{-22}$ | 1.29 | 202.76 |

GW151226 chirp has a smaller strain amplitude, undergoes large number of cycles and its signal energy is spread over a longer time interval than GW150914 chirp.

The strain recorded by LIGO's Livingston detector for GW150914 is plotted along with the inspiral part from Newtonian chirp with $F_{ISCO}$ as inspiral limit in figure[6]. The $F_{ISCO}$ comes out to be 67.69 Hz and the time corresponding to it can be found from frequency-time plot of

Newtonian chirp. For a signal starting at 0.25 seconds this corresponds to 0.407 seconds. The Newtonian chirp doesn't mimic exactly the real signal's amplitude and phase observed, but its worth appreciating that the strain order and evolution during inspiral phase has resemblance. This is to be expected as higher order terms of Post-Newtonian approximation if considered bring changes in phase and amplitude. We can also see that merger and ringdown part of the actual signal is not captured by the Newtonian chirp. The sources for LIGO detectors observed strain data and details of Mathematica code which can be downloaded is given in appendix.

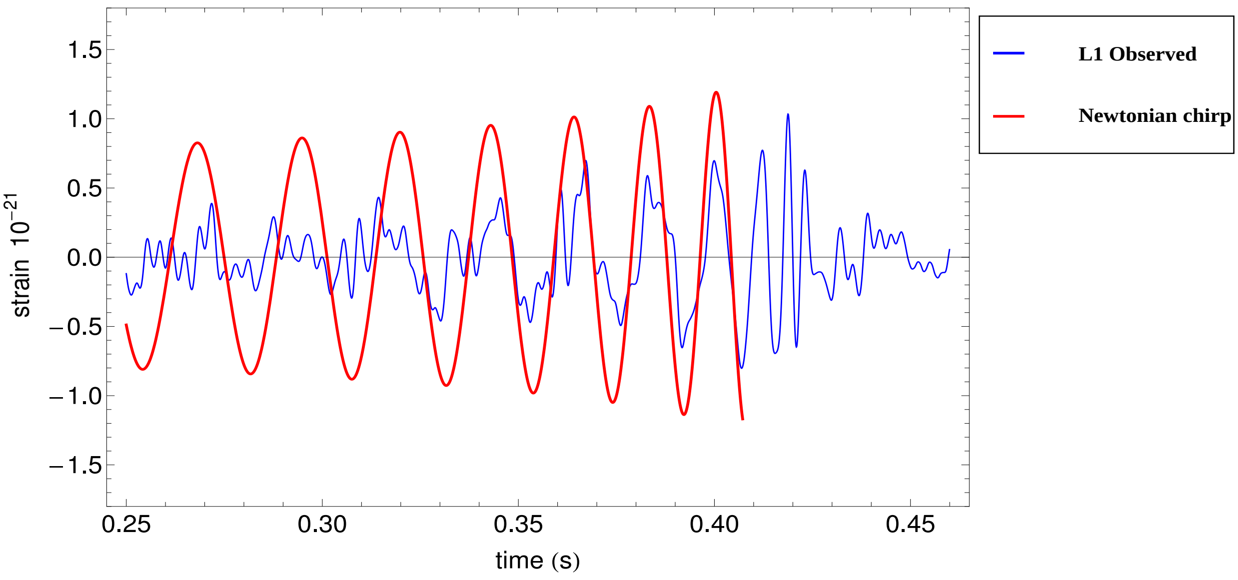


Figure 6: GW150914 strain observed in Livingston with Newtonian chirp

# 6 Sensitivity curve

Sensitivity curve is used to determine whether a source is detectable by the detector. The power spectral density $S_h$ is the power per unit frequency. The square root of the power spectral density gives the amplitude spectral density defined as $h_f = \sqrt{S_h}$ [2]. If $h_f$ of the source lies above the sensitivity curve of the detector then the source can be detected [2]. For a source the amplitude spectral density is given by [12],

$$\sqrt{S_h} = h_c f^{-1/2} = 2f^{1/2}\left|\tilde{h_c}\right| \qquad (9)$$

where $h_c$ is characteristic strain and $\tilde{h_c}$ is the Fourier transform of signal.

For inspiraling binaries in the leading order it has been found that [3],

$$\tilde{h_c} = A(f)e^{i\psi}$$

where

$$A(f) = \sqrt{\frac{5}{24}\frac{\mathscr{M}^{5/6}}{D\pi^{2/3}}}f^{-7/6} \qquad (10)$$

$$\psi = 2\pi f t_0 - \phi_0 - \pi/4 + \frac{3}{128}(\pi\mathscr{M}f)^{-5/3}$$

Using the above expressions we get,

$$\sqrt{S_h} = 2f^{1/2}\left|\tilde{h_c}\right| = \sqrt{\frac{5}{6}}\frac{\mathscr{M}^{5/6}}{D\pi^{2/3}}f^{-2/3} \qquad (11)$$

In inspiral a gravitational wave signal spends greater number of cycles in lower frequency whereas lesser number of cycles in higher frequency. Since the amplitude spectral density $h_f$ is inversely related to frequency, with increase in frequency $h_f$ decreases. Detectors sensitivity curve is governed by various noises present in the detector. We plot the $h_f$ vs frequency plot for the various chosen compact binaries and also for the two LIGO events along with the LIGO design sensitivity curve and aLIGO design sensitivity curve to see if the signal could be detected. The area between the signal curve and detector's noise curve indicates the signal to noise ratio [13].

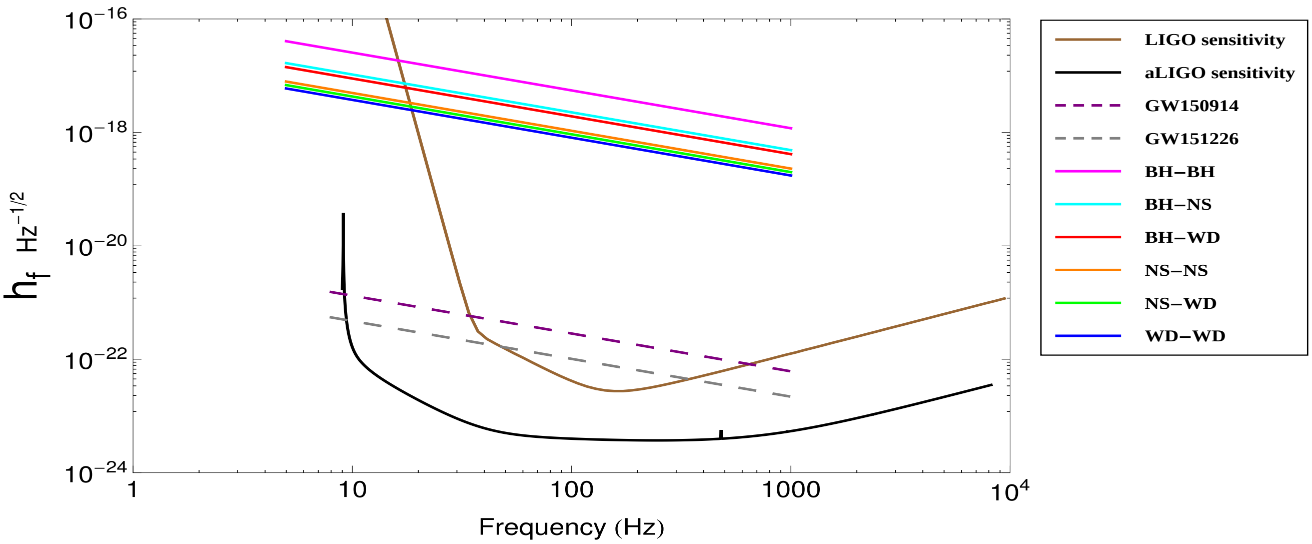


Figure 7: Compact binaries at 8kpc along with GW150914 and GW151226

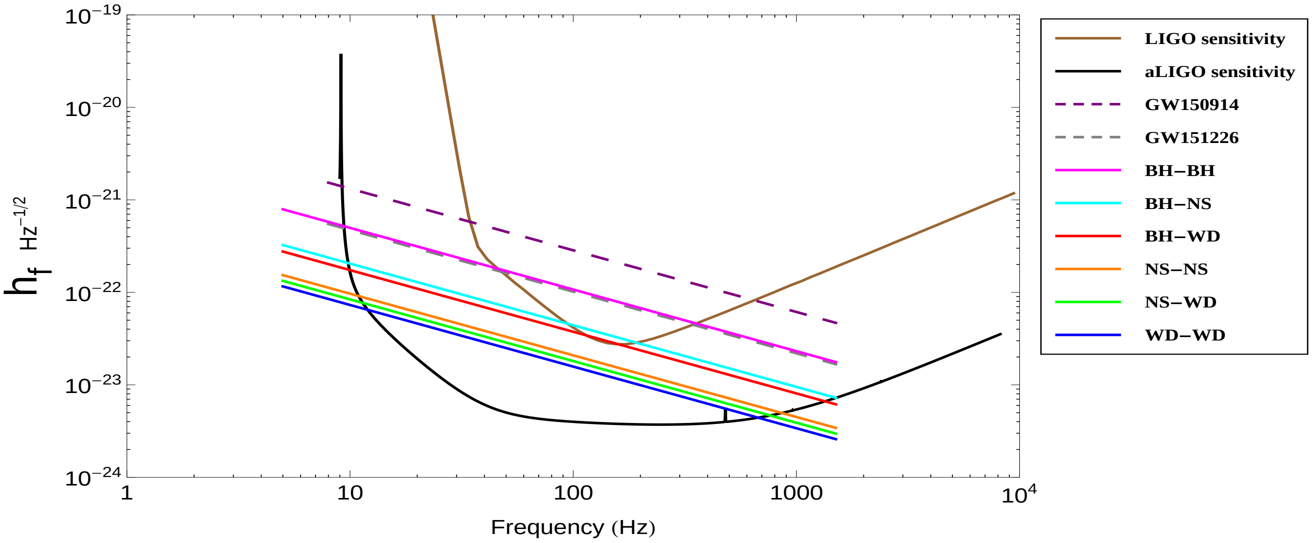


Figure 8: Compact binaries at 410kpc along with GW150914 and GW151226

We first consider the compact binaries in their location at galactic center and then at 410 Mega parsec (the distance from which GW150914 originated). We keep the distance of GW150914 event and GW151226 their respective ones to obtain the plots in figure[7,8]. The details of data obtained for LIGO and aLIGO sensitivity curve along with Mathematica code for the plot is given in appendix.

In the plot with Newtonian chirp context we have extended the frequency of these systems till 1000 Hz but the inspiral phase of each compact binary will end at their respective peak frequency $F_{ISCO}$. The amplitude spectral density curves of GW150914 and GW151226 during the inspiral, merger and ringdown phases reported by LIGO and VIRO collaboration can be seen from Ref [13].
The amplitude spectral density curve of compact binaries at galactic center (8kpc) lie above LIGO and aLIGO sensitivity curve, hence they are within the detector detection capabilities. When the same binaries are located at 410 Mpc only the BH-BH is well within the detection capability of LIGO whereas BH-NS and BH-WD systems lie very slightly above the LIGO curve in a small frequency range but extracting the signal from noise is hard. The NS-NS, NS-WD and WD-WD systems lie below the sensitivity curve of LIGO at all frequency values. Even when located at 410 Mpc NS-NS, NS-WD and WD-WD systems are in detection capabilities of aLIGO. Hence the merger of these systems can be expected to be observed in future. GW150914 signal lies well above the aLIGO detector sensitivity curve than GW151224; whereas the NS-NS, NS-WD and WD-WD binary curve located at 410 Mpc lies over the aLIGO curve lowest among the all. Hence sophisticated data analysis is required to pull out such signals from noise.

## 7 Higher order Post-Newtonian terms

In Post-Newtonian approximation the expressions for orbital frequency, phase of gravitational wave and waveform is in the form of expansion of PN order contributions. The 0 PN order term (neglecting higher order contributions) gives the orbital frequency, phase $\phi$ and $h$ describing the Newtonain chirp. The higher order terms gives a significant contribution to phase and amplitude of gravitational wave. The nomenclature of the $nPN$ order is for terms proportional to $x^n$ [6]. Where,

$$x = (M\omega)^{2/3} \tag{12}$$

and $\omega = \pi F$ is the angular velocity, $F$ is the gravitational wave frequency.

The expression for the phase as a function of x is given by the equation [6],

$$\phi = \frac{-x^{\frac{-5}{2}}}{32\eta}[C_0 + C_1 x + C_{1.5}\ x^{\frac{3}{2}} + C_2\ x^2 + C_{2.5}\ x^{\frac{5}{2}} + C_3 x^3 + C_{3.5} x^{\frac{7}{2}}]$$

The coefficients are-

$$C_0 = 1 \qquad C_1 = \left(\frac{3715}{1008} + \frac{55}{12}\eta\right) \qquad C_2 = \left(\frac{15293365}{1016064} + \frac{27145}{1008}\eta + \frac{3085}{144}\eta^2\right)$$

$$C_{1.5} = -10\pi \qquad C_{2.5} = \left(\frac{38645}{1344} - \frac{65}{16}\right)\pi\ln\left(\frac{x}{x_0}\right) \qquad C_{3.5} = \left(\frac{77096675}{2032128} + \frac{378515}{12096}\eta - \frac{74045}{6048}\eta^2\right)\pi$$

$$C_3 = \frac{12348611926451}{18776862720} - \frac{160}{3}\pi^2 - \frac{1712}{21}C - \frac{856}{21}\ln(16x) + \left(\frac{-15737765635}{12192768} + \frac{2255}{48}\pi^2\right)\eta + \frac{76055}{6912}\eta^2 - \frac{127825}{5184}\eta^3$$

C=0.577 called as the Euler-Mascheroni constant.

Table 5: Contribution of each PN term to the number of gravitational wave cycles

| PN order | BH-BH | BH-NS | BH-WD | NS-NS | NS-WD | WD-WD |
|---|---|---|---|---|---|---|
| 0 | 602.712 | 3582.14 | 4956.29 | 16056.7 | 21354.6 | 28134.8 |
| 1 | 59.3521 | 213.313 | 280.495 | 441.305 | 526.559 | 618.958 |
| 1.5 | -51.3811 | -181.362 | -242.645 | -210.884 | -241.22 | -265.799 |
| 2 | 4.06297 | 9.80028 | 12.4367 | 9.91219 | 10.733 | 11.3563 |
| 2.5 | -7.14497 | -20.0051 | -26.426 | -11.6897 | -12.4027 | -12.4674 |
| 3 | 2.17844 | 2.28362 | 2.26033 | 2.55632 | 2.56639 | 2.58872 |
| 3.5 | -0.81811 | -1.81626 | -2.33096 | -0.90533 | -0.93078 | -0.91179 |

We take $F_1 = f_0$ and $F_2 = F_{ISCO}$, the value of $f_0$ is related to lowest possible frequency observable by the detector. The contribution of nth PN order to number of cycles can be found using [14],

$$N_{cyles} = \frac{\phi(F_2) - \phi(F_1)}{\pi} \tag{13}$$

Using $\phi$ expression with $n = 0$ order term alone and neglecting the higher orders, we can find the contribution to number of cycle due to 0 PN order. The phase of 0 PN order is given by ,

$$\phi_{n=0} = \frac{-(x)^{-5/2}}{32\eta} = \frac{-(M\pi F)^{-5/3}}{32\eta}$$

$$N_{cycles} = \frac{-(M\pi)^{-5/3}}{32\eta}\left(\frac{F_2^{-5/3} - F_1^{-5/3}}{\pi}\right)$$

In similar way the contribution due to higher order terms can also be found. For the compact binaries taken in this study we have found contribution to number of cycles due to various PN orders. The results are tabulated table 5. The Mathematica code to find contributions due to various orders is given in Appendix.

## 8 Conclusion

Newtonian chirp in the leading order Post-Newtonian approximation is based upon the assumption that masses are point objects, which is not true for real systems but still is a good approximation. The amplitude spectral density of Newtonian chirp is used to determine whether the inspiral phase of a binary can be observed in ground base detectors and hence the study of it for binaries with the aLIGO and LIGO sensitivity curves helps us to understand the detector's capabilities and the possible developments that could be made to make the detectors more efficient.

The real gravitational wave signals obtained from the detectors are more complicated, when we try to match the Newtonian chirp waveform against the real waveform obtained from LIGO Livingston detector for GW150914 we could see that it didn't exactly match with the real signal's amplitude and phase observed. It needs the study of the higher order terms (Post-Newtonian terms) to describe the real wave's inspiral exactly.

We see that the formula for peak frequency of gravitational waves together with the phase formula having the higher order PN terms, could determine the number of cycles contributed by each higher order in addition to the leading order for the real gravitational wave signal from the binary.

This study brings out the essential features of GW produced by different compact binary systems and the possibilities of detecting them ; also helps us to understand the insights which Newtonian chirp convey regarding the actual gravitational wave signals which are complex in nature.

# Appendix

All expressions are taken in geometrized units where G=c=1. Mass and distance have units of seconds. In geometrical units, $1M_{\odot}$=4.92549095$\times 10^{-6}$sec and 1pc=1.0292712503$\times 10^{8}$sec. Physical units can be obtained by replacing a mass $\mathscr{M}$ by $G\mathscr{M}/c^3$ and a distance $D$ by $cD$.

The Newtonian chirp waveform, frequency-time plot and amplitude spectral density plot can be obtained using the Mathematica code (or any programming language). LIGO event data and sensitivity curves can be plotted in 2-D using GNU, R program, Mathematica or any other tool. We provide the Mathematica code nb file and pdf file along with required txt files for the plots used in the paper.

### 1. Waveform and frequency variation

Mathematica code for waveform and frequency-time plot be downloaded as
nb file : https://drive.google.com/open?id=0B1PM7VcZPwY2eWZwQWdSRk1mREU
pdf file : https://drive.google.com/open?id=0B1PM7VcZPwY2X0NzVWxHZkxiQXM

Mathematica code for frequency-time plot for all compact binaries plotted together can be downloaded as
nb file : https://drive.google.com/open?id=0B1PM7VcZPwY2bVRNWi1EeVNaN3c
pdf file : https://drive.google.com/open?id=0B1PM7VcZPwY2T1ljM2o0T2FxR0k

Mathematica code for single column waveform plot can be obtained as nb file : https://drive.google.com/open?id=0B1PM7VcZPwY2QUUzUFB2LXZxbUE .

### 2. LIGO events

GW150914 event strain data observed in Livingston detector can be obtained from the LIGO website : https://losc.ligo.org/events/GW150914/ . Alternatively can also be downloaded as txt file from : https://drive.google.com/open?id=0B1PM7VcZPwY2WFh0VkFzS0M0T0U .

Mathematica code to plot the strain of GW150914 and overlay the inspiral part of Newtonian chirp can be downloaded as
nb file : https://drive.google.com/open?id=0B1PM7VcZPwY2cU9WXzh5M2lWX1E
pdf file : https://drive.google.com/open?id=0B1PM7VcZPwY2RVJ1VXJmT2N0ZjQ

### 3. Strain spectral amplitude

The LIGO design sensitivity curve data can be obtained from LIGO website : https://dcc.ligo.org/LIGO-E950018/public , where the file named "SRD-strain-4k.txt" is the required txt file. Alternatively as txt file from : https://drive.google.com/open?id=0B1PM7VcZPwY2VUhYd1dsQzBUVDA.

The aLIGO design sensitivity curve data can be obtained from LIGO website : https://dcc.ligo.org/LIGO-T0900288/public , where the file named "ZERO-DET-high-P.txt" is the required txt file. Alternately as txt file from : https://drive.google.com/open?id=0B1PM7VcZPwY2SlFPa3JrWWRmd00 .

Mathematica code for plotting sensitivity curves and amplitude spectral density of compact binaries(at galactic centre) can be downloaded as
nb file : https://drive.google.com/open?id=0B1PM7VcZPwY2TlhJTUpKRzAwUDQ
pdf file : https://drive.google.com/open?id=0B1PM7VcZPwY2alNyQzBjdEVOeWs

Mathematica code for plotting LIGO sensitivity curves and amplitude spectral density of compact binaries binaries(at 410 Mpc) can be downloaded as
nb file : https://drive.google.com/open?id=0B1PM7VcZPwY2SnJFTkkwQ1FDRDA
pdf file : https://drive.google.com/open?id=0B1PM7VcZPwY2M0lNZllJMGZBOVE

4. Higher order PN terms

The code to calculate the contribution of higher order PN terms to number of cycles can be obtained as
nb file : https://drive.google.com/open?id=0B1PM7VcZPwY2ZWhnNUo1cDl6dzg
pdf file : https://drive.google.com/open?id=0B1PM7VcZPwY2U0tsYkFFbEM3aEU

All the Mathematica nb files, pdf files and txt files can be downloaded in single folder as zip file from : https://drive.google.com/open?id=0B1PM7VcZPwY2UlNpR3ZENHJtTm8 .

# References

[1] M.P. Hobson, G.P. Efstathiou and A.N. Lasenby, *General Relativity an Introduction for Physicists*, (Cambridge University Press 2006).

[2] General Relativity course lecture notes, Utah State State University (2013). www.physics.usu.edu/Wheeler/GenRel2013/Notes/GravitationalWaves.pdf

[3] P. Ajith, S. Mohapatra, and A. Pai, *A Gravitational Wave Data Analysis Primer for the IndiGO Mock Data Challenge.* http://www.gw-indigo.org/tiki-download_wiki_attachment.php?attId=22

[4] P. Ajith, A. Ghosh, *A Tutorial on GW Data Analysis.* http://www.nikhef.nl/~vdbroeck/GW_cursus/Werkcollege_5.pdf

[5] Katherine Grover, *Physics and Astrophysics with Gravitational Waves from Compact Binary Coalescence in Ground Based Interferometers*, PhD thesis, University of Birmingham, October 2015. http://etheses.bham.ac.uk/6410/1/Grover15PhD.pdf

[6] Tjonnie G.F.Li ,*Extracting Physics from Gravitational Waves*, PhD thesis, VU University Amsterdam, October 2013. http://www.nikhef.nl/pub/services/biblio/theses_pdf/thesis_T_G_F_Li.pdf

[7] G. Srinivasan, *The Maximum Mass of Neutron Stars*, Bull.Astr.Soc.India (2002) **30**, 523-547.

[8] Cutler and Thorne, *An overview of Gravitational wave sources*, arXiv:gr-qc/0204090.

[9] Hubble Revisits Tangled NGC 6240. www.nasa.gov/image-feature/goddard/hubble-revisits-tangled-ngc-6240

[10] B.P. Abbott et al., *Observation of Gravitational Waves from a Binary Black Hole Merger*, Physics Review Letters **116**, 061102 (2016).

[11] B.P. Abbott et al., *GW151226: Observation of Gravitational Waves from a 22-Solar-Mass Binary Black Hole Coalescence*, Physics Review Letters **116**, 241103 (2016).

[12] C.J. Moore, R.H. Cole and C.P.L. Berry, *Gravitational-wave sensitivity curves*, Classical & Quantum Gravity, 32(1):015014 (2015), arXiv:1408.0740v2 [gr-qc].

[13] The LIGO Scientific Collaboration and The Virgo Collaboration, *Binary Black Hole Mergers in the first Advanced LIGO Observing Run*, arXiv:1606.04856v2 [gr-qc].

[14] B. Moore, M. Favata, K.G. Arun and C.K. Mishra, *Gravitational-wave phasing for low-eccentricity inspiralling binaries to 3PN order*, Phys. Rev. D 93, 124061 (2016), arXiv:1605.00304v2 [gr-qc].

# A New Tool to Study Universe: Gravitational Waves

Jyoti kapil

Maitreyi college,

University of Delhi
*E -mail: jyoti.physics@gmail.com*

(*Submitted:04-03-2016*)

## Abstract

The most elusive Gravitational waves have finally been detected by the Laser Interferometer Gravitational waves Observatory (LIGO) located in United States, produced by the collision of two massive black holes almost 1.3 billion years ago. The discovery has confirmed the predictions made by Albert Einstein almost a century back in 1916 about the existence of gravitational waves.

**Introduction:**

11th Feb, 2016 was a remarkable day not just for the scientific community but for the human kind, a day when an international team of scientists, including Indian scientists, announced the first direct detection of the most elusive waves: the gravitational waves. The LIGO (Laser Interferometer Gravitational waves Observatory) observatory in United States heard and recorded the sound of gravitational waves for the first time ever in history which were produced by the collision of two massive black holes 1.3 billion years ago.

The signals of this event were first recorded on 14$^{th}$ Sep, 2015. It was a century back, in 1916, when the world famous physicist, Albert Einstein, first hypothesized the existence of gravitational waves. According to him, the whole Universe is permeated by a flexible space-time fabric and the sheer presence of matter or energy results in warping or curving of the space-time around it which gives rise to gravity and whenever a mass is accelerated on the surface of this fabric, gravitational waves are produced.

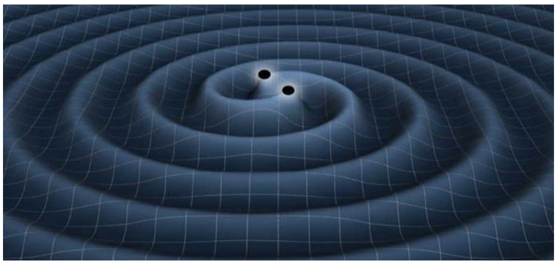


**FIG. 1The binary system of black holes orbiting each other and producing gravitational waves; source: www.space.com**

## WHAT ARE GRAVITATIONAL WAVES?

But exactly what are gravitational waves and how are they produced? According to Einstein, the three dimensions of space and one dimension of time are unified in a single 4 dimensional space-time. This unified space-time fabric is warped by the presence of heavy objects like planets and stars, akin to the curvature produced in the surface of a rubber sheet or a trampoline by a marble or a heavy iron ball. Due to this depression produced in the space-time fabric a nearby object experiences an attractive force towards it which we know as “gravity”, similar to a marble rolling down on the surface of the trampoline towards a depression produced by the heavy iron ball. Heavier the object larger is the curvature produced in the space-time fabric and larger is the force of attraction (gravitational pull) experienced by a body in its vicinity.

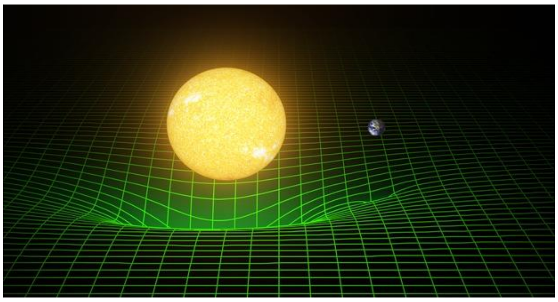


**FIG. 2 warping of space-time; source: www.ligo.caitech.edu**

This is the reason why earth and other planets move around Sun. Whenever a cataclysmic event happens in the outer space, such as neutron stars or black holes orbiting each other at great speeds, gravitational waves are produced as ripples in the fabric of space-time and travel outward from the source with the speed of light, similar to the ripples produced on the surface of water in a pond when a pebble is dropped in it. In general, anything having mass when accelerated, produces gravitational waves.

Any accelerated mass which we observe around us; a moving car, a rotating top or a planet moving around sun, is a source of gravitational waves, but the strength of these waves is quite weak rendering them very difficult to be detected. Strong gravitational waves are produced by violent or catastrophic events taking place in the outer space such as coalescing neutron stars or white dwarf stars, supernovae, colliding black holes or the Big Bang itself. The gravitational waves which have been detected by the LIGO detector were produced by two massive black holes having masses 36 times and 29 times the mass of the sun, orbiting about each other at tremendous speed, continuously losing their energy and finally colliding and forming a single black hole having mass 62 times the mass of the sun; the remaining mass (3 times the mass of the sun) being radiated as energy in the form of gravitational radiation (energy carried by gravitational waves).

## DETECTION:

Was it easy to detect these waves? Certainly not! From the day when the gravitational waves were first predicted by Einstein, scientists and engineers all over the world made efforts to devise some method to detect them directly, but till now they had been able to get only some indirect proof of their existence. Like in 1974 the strong evidence of existence of gravitational was found by Joseph Taylor and Russell Hulse from the radiation emitted by a binary system of stars, one of them was a pulsar, revolving about each other and observing how their orbital period was continuously reducing, losing energy as a result, and giving off this lost energy as gravitational waves. But the idea of making use of a laser interferometer for direct detection of gravitational waves had already been taken up by then, the fundamental principle employed by LIGO. LIGO is an observatory having world's most sensitive experimental setup for detection of gravitational waves. It is so sensitive that it can detect vibrations from its nearby regions to as far as deep regions of the outer space.

**FIG.3 LIGO at Livingston and Hanford; Source: www.ligo.caltech.edu**

It is not a single observatory; there are two observatories working in unison, one located at Hanford, Washington and the other 3002 k.m. away from it, at Livingston, Louisiana. The reason for the "twin observatory" and their being so far apart is to make the observations of gravitational waves more prominent over any local noise by filtering out the noise detected by the detectors due to a number of sources like vehicles moving nearby, any internal fluctuations in the laboratory or instruments, earthquakes etc.

## DESCRIPTION OF THE EQUIPMENT:

The heart of this experimental setup is a large "laser interferometer" which works on the principle of "interference of waves". It is the largest interferometer in the world consisting of two 4k.m. long steel vacuum tubes, 1.2m in diameter arranged perpendicular to each other, and having mirrors at their ends to reflect light. It consists of a beam splitter which splits the laser beam coming from a laser source into two beams of equal intensity, travelling down the two arms of the interferometer. To make the effective distance travelled by the laser beam larger in an arm, two more mirrors are arranged near the beam splitter, one in each arm. This is done to increase the sensitivity of interferometer because larger the distance travelled by the beam before interference, more sensitive it becomes to vibrations. The laser beam travelling down the arm suffers multiple reflections between the end mirror and the mirror near the beam splitter, making the effective distance travelled by the beam much larger than the physical distance of 4k.m. before superimposing with the laser beam coming from the other arm.

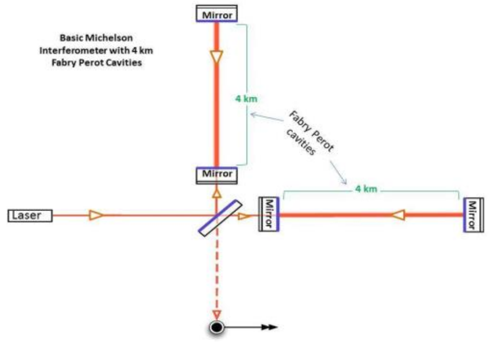


**FIG. 4 The Interferometer (modified version of a basic Michelson's Interferometer) used in LIGO; source: www.ligo.caltech.edu**

The advanced LIGO can detect change in its arm's length to 10,000 times smaller than the diameter of a proton, making it most sensitive instrument till date. In the absence of any gravitational waves, the beams reaching the detector superimpose on each other destructively producing no resultant light. But in the presence of gravitational waves, the arm length of the interferometer change; if one arm stretches, the other is contracted and vice-versa because the space-time is distorted by the gravitational waves. Because of this contraction and expansion of the arms, one laser beam travels larger distance as compared to the other at a given time. As a result the two beams do not interfere destructively and some light is detected by the detector, which carries information about the incoming gravitational waves.

## SIGNIFICANCE OF THE DISCOVERY:

Why is the scientific community so excited and thrilled about this discovery? Why is it so significant? The answer is simple- the detection of gravitational waves has given us a new vision to look and explore the universe from a completely new perspective, revealing the secrets of the deepest region of space. The information carried by gravitational waves is different from that carried by light waves, x-rays, gamma rays etc.; they can pass through matter unchanged unlike electromagnetic waves which get absorbed or reflected when interact with matter. This will make us understand events in the outer space which do not radiate in electromagnetic spectrum and therefore go unnoticed. With the help of these waves we would be able to track supernovae, understand life cycles of black holes and neutron stars. It would also make it possible to get more information about dark matter and the Big Bang itself. The discovery has also provided an evidence for the existence of black holes. The contribution of Indian scientists has also been significant in this discovery. And it can be said that they have been well rewarded by the Indian government by approving the construction of a LIGO observatory in India; definitely beginning of a new era for science and humanity.

## REFERENCES:

[1] Concepts of Modern Physics, 6th Edition, by Arthur Beiser

[2] https://www.ligo.caltech.edu

[3] https://www.google.co.in

# DARK MATTER AND NEUTRINOS

Gazal Sharma[1], Anu[2] and B. C. Chauhan[3]

Department of Physics & Astronomical Science
School of Physical & Material Sciences
Central University of Himachal Pradesh (CUHP)
Dharamshala, Kangra (HP) INDIA-176215.
[1]gazzal.sharma555@gmail.com [2]3anoman7@gmail.com
[3]chauhan@associate.iucaa.in

(submitted 03-08-2015)

---

**Abstract**

The Keplerian distribution of velocities is not observed in the rotation of large scale structures, such as found in the rotation of spiral galaxies. The deviation from Keplerian distribution provides compelling evidence of the presence of non-luminous matter i.e. called dark matter. There are several astrophysical motivations for investigating the dark matter in and around the galaxy as halo. In this work we address various theoretical and experimental indications pointing towards the existence of this unknown form of matter. Amongst its constituents neutrino is one of the most prospective candidates. We know the neutrinos oscillate and have tiny masses, but there are also signatures for existence of heavy and light sterile neutrinos and possibility of their mixing. Altogether, the role of neutrinos is of great interests in cosmology and understanding dark matter.

---

## 1 Introduction

As a human being the biggest surprise for us was, that the Universe in which we live in is mostly dark. The NASA's Plank Mission revealed in 2013 that our Universe contains 68.3% of dark energy, 26.8% of dark matter, and only 4.9% of the Universe is known matter which includes all the stars, planetary systems, galaxies, and interstellar gas etc.. This raises a number of questions in our minds; e.g. how much and how well we know about our Universe? What are dark matter and dark energy? What are they made up of? The very first suggestion of dark matter in our galaxy was made by Kapteyn and Jeans in 1922 and then by Lindblad in 1926. They proposed the existence of dark matter while observing the motions of nearby stars at right angle to the plane of our Milky way galaxy. Oort in 1932 claimed that there exists substantial amount of dark matter near the sun by observing the vertical motions of stars. However, in 1991, Kuijken and Gilmore argued that there were no significant evidence for dark matter with in the galactic disk near the sun.

Sinclair Smith and Fritz Zwicky in 1933, studied the large clusters of galaxies and found that galaxies were on average moving too fast for the cluster to be held together only by the mass of the visible matter. They concluded that in rich clusters of galaxies, a large portion of the matter is not visible i.e. the dark matter. The idea of dark matter in galactic halo was given by Freeman in 1970, while studying the rotation curve for NGC 300 and M33 by using the 21cm-Line of neutral hydrogen did not show the expected Keplerian decline beyond the optical radii. Then in 1979, Vera Rubin proposed that normal spiral galaxies contain substantial amount of dark matter present at great distances from the central regions. An influential model was proposed by Caldwell and Ostriker in 1981 for the density of core-halo type model of dark matter. The halo model is valid till now but the exact distribution of dark matter is still a mystery.

The next question to be addressed is about the constituents of dark matter. One of the biggest discoveries made by Hubble Space Telescope (HST) of NASA was the confirmation of invisible matter in the Universe. A 3D map of dark matter was derived from largest survey of the Universe made by the HST, the Cosmic Evolution Survey (COSMOS). The COSMOS survey covers a sufficiently wide area of sky - nine times the area of the full Moon (1.6 square degrees) - for the large-scale filamentary structure of an invisible form of matter that makes up most of the mass of Universe i.e. dark matter to be clearly evident [1]. The theory of Big-Bang nucleosynthesis (BBN), i.e. formation of light nuclei just after Big-Bang, as well as experimental evidences from anisotropies in Cosmic Microwave Background Radiation (CMBR) observed by NASA's Wilkinson Microwave Anisotropy Probe (WMAP) indicate that most of the dark matter stuff is non-baryonic (which is not made up of regular matter).

Many experiments has been performed in search of dark matter candidates. Neutrinos, which are electrically neutral and tiny particles, seem potential candidates for dark matter, as they are long-lived and almost non-interacting with other particles. However, the three known types of neutrinos, called active neutrinos, are not massive enough to account for all of the dark matter of Universe. So, theorists proposed another type of neutrinos that would not interact at all with the regular matter, but are massive. If the sterile neutrino is heavy enough about $\sim 10\ keV$, it could account for the substantial amount of dark matter. The present article aims to introduce reader about the dark matter, its evidences, possible constituents and the potential candidature of neutrinos in the composition of dark matter.

## 2 Dark Matter

As discussed above the dark matter is the matter, which does not interact with light at all or may interact very poorly that it remains dark and unseen. As such, a question arises in ones mind; how one can detect something which does not interact with light. The answer may be 'gravity'; such that there are many astrophysical motivations for the detection of dark matter. There have been obtained a number of observational evidences for the existence of dark matter because of its gravitational effects, like galactic rotation curves of galaxies measured by Vera Rubin, confinement of hot gas in the galaxies, measurement on the basis of gravitational lensing [2], etc...

### 2.1 Flattened Orbital Velocity Curves

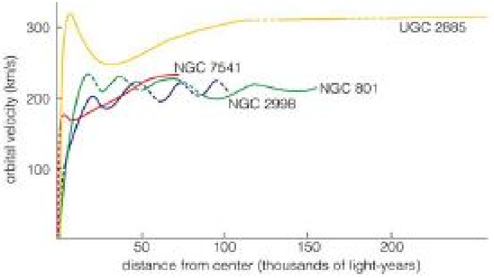
Before describing observations let us see how celestial objects respond to the gravitational force acting on them and how that response can reveal the large scale distribution of matter. For the planets in orbit around the sun which embodies essentially all the mass of the solar system, the decrease in gravitational attraction with distance is given by Newton law of gravitation. It has been found that orbital velocities of planets decreases with distance from the centre of the sun. In spiral galaxy the gas, dust and stars in the disk of the galaxy all orbit around a common centre. Like planets in solar system, the gas and stars move in response to the combined gravitational attraction of all other mass. If the galaxy is visualized as a spheroid, we can calculate the gravitational attraction due to mass $M_r$ lying between the centre and an object of mass $m$ in an equatorial orbit at a distance $r$ from the centre. If the galaxy is neither contracting nor expanding then the gravitational force is exactly equal to the centrifugal force on the mass at distance $r$ is given by the equation

$$GmM_r/r^2 = mv_r^2/r, \tag{1}$$

where $v_r$ is the orbital velocity. When the equation is solved for $v_r$ , the value of $m$ drops out and the velocity of a body at a distance $r$ from the centre is determined only by the mass $M_r$ inward from its position. In the solar system, virtually all the mass is concentrated near the centre and the orbital velocity decrease as $1/\sqrt{r}$, that is called Keplerian decline.

In a galaxy the brightness is strongly peaked near the centre and falls off rapidly with distance. As per the distribution of luminosity, it was expected that stars at increasing distance from the centre would have decreasing Keplerian orbital velocities. When the orbital velocity of different stars present at different distances in a galaxy was studied by Vera Rubin, the unexpected results came

Figure 1: *Variation of orbital velocity with radius* [3]



out. This observation has been made for different spiral galaxies like Sa, Sb, Sc etc... Although each galaxy exhibits distinctive feature in its rotational pattern, the systematic trends that emerge are impressive. With increasing luminosity galaxies are bigger, orbital velocities are higher and the velocity gradient across the nuclear bulge is steeper. Moreover, each type of galaxy displays characteristic rotational properties.

Therefore we can draw conclusion from our observation that all the rotation curves are either flat or rising out to the visible limits of the galaxy. There are no extensive regions where the velocities fall off with distance from the centre, as would be predicted if mass were centrally concentrated. The conclusion is inescapable- mass unlike luminosity is not concentrated near the centre of spiral galaxies. Thus the light distribution in a galaxy is not at all a guide to mass distribution. Instead the mass inside any given radial distance is increasing linearly with distance and contrary to what one might expect, is not converging to a limiting mass at the edge of the visible disk. The linear increase in mass with radius indicates that each successive shell of matter in the galaxy must contain just as much mass as every other shell of the same thickness. Since the volume $V$ of each successive shell increases as the square of the radius, the density $\rho$ of matter in successive shells must decrease in order to keep $\rho V$ constant [4].

The widely accepted idea about the dark matter is that each spiral galaxy is embedded into a halo of dark matter. The gravitational attraction of the unseen mass keeps the orbital velocities high at larger distance from galactic centre. Till now we are not able to find the exact distribution of dark matter but we can say that it is strongly clumped around the galaxies. The density of dark halo decreases with distance from galactic centre as given by Caldwell and Ostriker

$$\rho_d = \frac{\rho_0}{1+\frac{r^2}{a^2}}. \tag{2}$$

They found a fit for the data with $\rho_0 = 1.37 \times 10^{-2} M_{\odot} pc^{-3}$, and $a = 7.8kpc$.

If we consider a different distribution for dark matter in which we put all the unseen matter in a disk, the disk will quickly become unstable. Therefore P. Ostriker and Peeble suggest that the halos are important for stabilizing the disk. Additional evidence on the high rotational velocity was provided by the 21-centimetre radio waves emitted by the neutral hydrogen in the galactic disk. From the above discussion we can draw a conclusion that the density of dark matter halo surrounding the visible matter decreases slowly outwards.

## 2.2 Gravitational lensing

According to Einstein's theory of general relativity large objects with their immense masses can distort space-time therefore large massive objects such as galaxy clusters bend light from distant sources, creating distorted images that we can see here on earth. This is called gravitational lensing. This technique is especially useful for detecting dark matter. Since dark matter doesn't interact with light, it can't be seen directly. However, since dark matter is very massive, it can be detected indirectly by the distorted images it creates of normal matter through gravitational lensing. By measuring the angle of bending, the mass of the gravitational lens can be calculated-greater the bend, more massive the lens is. Therefore the angle of deflection is given by [5]

$$\alpha = \frac{4GM}{c^2 b}, \tag{3}$$

where $b$ is the impact parameter. Using this method, astronomers have confirmed that the galactic clusters indeed have high masses exceeding those measured by the luminous matter. There have been several positive reports on the observation of such microlensing, even though typically only one in a million stars examined is expected to show such an effect. The bending of light by a massive object, a general relativity effect has been verified to extreme accuracy (better than 1%) by studying radar echoes from the planets when they are in conjunction. Experiments like the Large Synoptic Survey Telescope (LSST), under construction in Chile, aim to take advantage of gravitational lensing to map the dark matter in the Universe and provide clues to its nature. MOA (Micro-lensing Observations in Astrophysics) is a Japan/NZ collaboration that makes observations on dark matter, extra-solar planets and stellar atmospheres using the gravitational micro-lensing technique at the Mt John Observatory in New Zealand. HST of NASA recently produces several images of gravitational lensed objects. Therefore finding enough gravitational lenses to constrain the properties of dark matter structures requires a powerful telescope with a huge field of view like LSST.

## 2.3 Fluctuations in CMBR

The Cosmic Microwave Background (CMB) is the earliest photograph of our Universe. The patterns that we see in observations of the CMB were set up by competition between two forces acting on matter; the force of gravity causing matter to fall inward and an outward pressure exerted by photons (or particles of light). This competition caused the photons and matter to oscillate into-and-out-of dense regions. If the Universe consisted partially of dark matter in addition to normal matter, that pattern would be affected dramatically. The existence of dark matter leaves a characteristic imprint on CMB observations, as it clumps into dense regions and contributes to the gravitational collapse of matter, but is unaffected by the pressure from photons. We can predict these oscillations in the CMB with and without dark matter, which often present in the form of a power spectrum.

The power spectrum of the CMB shows us the strength of photon-matter oscillations at different parts of sky. The Far-Infrared Absolute Spectrophotometer (FIRAS) instrument has measured the spectrum of the cosmic background radiation, making it the most precisely measured black body spectrum in nature. The Cosmic Background Explorer (COBE) was launched in 1989 in search of temperature anisotropies; frequency power spectrum; solar system and galactic dust foregrounds. The WMAP in 2010 was the first instrument to measure the CMB power spectrum through the first peak of oscillations, and showed that the existence of dark matter is favoured. Comparison of such calculations with the observations of CMB Radiation by Plank mission team in 2013 have shown that the total mass energy of the known Universe contains 4.9% ordinary matter, 26.8% dark matter and 68.3% dark energy. Thus dark Universe constitutes 95.1% of the total matter energy content of the Universe [6].

### 2.4 X- Ray Studies

The observational evidences from X- ray studies also supports the existence of dark matter. The basic technique is to estimate the temperature and density of the gas from the energy and flux of the X-rays using X-ray telescopes, which would further enable the mass of the galactic cluster to be derived. The measurements of hot gas pressure in galactic clusters by X-ray telescopes, such as CHANDRA X-ray observatory by NASA, have shown that the amount of superheated gas is not enough to account for the discrepancies in mass and that the visible matter approximately constitutes only $12 - 15\%$ of the mass of the cluster. Otherwise, there won't be sufficient gravity in the cluster to prevent the hot gas from escaping [6].

Recently in 2014, data came from the European Space Agency's (ESA's) XMM-Newton spacecraft, which was analysed by an international team of researchers. After scouring through thousands of signals, they spotted a weird spike in X-ray emissions coming from two different spots in the Universe: the Andromeda galaxy and the Perseus galaxy cluster. The signal doesn't correspond to any known particle or atom, and is unlikely to be the result of a measurement or instrument error hence it could have been produced by a dark matter particle. The signal's distribution within the galaxy corresponds exactly to what we expects with dark matter i.e. concentrated and intense in the centre of objects and weaker and diffuse on the edges. Scientists believe that there is a possibility that it could come from dark matter candidate i.e. possibly the hypothetical heavy sterile neutrinos; as it is believed the decay of these particles could produce X-rays [7].

## 3 Baryonic & Non-Baryonic Dark Matter

On the basis of observed orbital velocity curves, and other evidences we can say that dark matter exists. Baryonic dark matter is non-luminous matter in which most of the mass is attributed to baryons, most probably neutrons and protons. Candidates for baryonic dark matter include non-luminous gas, Massive Astrophysical Compact Halo Objects (MACHOs). These MACHOs may include condensed objects such as black holes, neutron stars, white dwarf, very faint stars, or non-luminous objects like planets and brown dwarfs. Baryonic dark matter cannot be detected by its emitted radiation because these objects have very low luminosity, but the presence of these objects can be inferred from their gravitational effects on visible matter [6].

The nucleosynthesis of the elements and observations of the Cosmic Microwave Background Radiations (CMBR) puts constraints on the density of baryonic matter. No more than 15% of the matter in the Universe can be baryonic but most of dark matter stuff is non-baryonic. Non-baryonic dark matter (NBDM) is non-luminous matter made from non-baryonic stuff (other than protons, neutrons etc.). Recent measurements of the matter density $\Omega_m^0$ and the energy density $\Omega_\Lambda^0$ comes from three types of observations: 1) supernova measurements of the recent expansion history of the Universe; 2) cosmic microwave background measurements of the degree of spatial flatness, and 3) measurements of the amount of matter in galaxy structures obtained through big galaxy redshift surveys agree with each other in a region around the best current values of the matter and energy densities $\Omega_m^0 \simeq 0.27$ and $\Omega_\Lambda^0 \simeq 0.73$. Where $\Omega$ is the energy density of Universe defined by

$$\Omega = \frac{\rho}{\rho_c}, \tag{4}$$

where $\rho_c$ is the critical density (average density of Universe to halt its expansion) of the Universe and $\Omega^0$ represents present energy density of Universe. Measurements of the baryon density in the Universe using CMBR spectrum and primordial nucleosynthesis (i.e. BBN) constrain the baryon density $\Omega_b^0$ to a value less than 0.05. The difference $\Omega_m^0 - \Omega_b^0 \simeq 0.22$ must be in form of non-baryonic dark matter [8]. The value of total matter density

$$\Omega_m^0 h^2 = 0.135^{+0.008}_{-0.009}, \tag{5}$$

out of which the baryonic matter is

$$\Omega_b^0 h^2 = 0.0224^{+0.0009}_{-0.0009}, \tag{6}$$

in the form of neutrinos

$$\Omega_\nu^0 h^2 < 0.0076, \tag{7}$$

and the matter in the form of Cold Dark Matter (CDM) is

$$\Omega_{CDM}^0 h^2 = 0.113^{+0.008}_{-0.009}. \tag{8}$$

The results of BBN that tell that $\Omega_B \sim 0.01$ and therefore if $\Omega$ total is truly unity, then the bulk of the mass of the Universe must be in the form of some sort of non-baryonic matter. From baryon to photon ratio i.e. $\eta = \eta_B/\eta_\gamma$, one can find the range for $\eta$ as given by [20]

$$4.7 \times 10^{-10} \leq \eta \leq 6.5 \times 10^{-10}. \tag{9}$$

We can find relative baryon density $\Omega_B$ as

$$0.017 \leq \Omega_B^0 h^2 \leq 0.024. \qquad (10)$$

This shows that Universe is not closed by baryonic matter and this gives the indication of existence of dark matter. From the analysis of the existing data follows that

$$\Omega_{DM} \simeq 0.20. \qquad (11)$$

The non-baryonic dark matter is classified in terms of the mass of the particle that is assumed to make it up, and the typical velocity dispersion of those particles (since more massive particles move more slowly). There are three prominent hypotheses on non-baryonic dark matter, called Hot Dark Matter (HDM), Warm Dark Matter (WDM), and Cold Dark Matter (CDM); some combination of these is also possible. CDM is composed of substantially massive particles ($\sim GeV$) expected to be moving with non-relativistic speeds. The leading candidates for CDM called WIMPs (Weakly Interacting Massive Particles). WIMPs could include large number of exotic particles such as neutralinos, axions, photinos etc. These particles forms dark matter, because they have too much mass to move at high speeds and that they are the best candidates for dark matter. As WIMPs can interact through gravitational and weak forces only, they are extremely difficult to detect. There are several experiments setup for detection of WIMPs such as SuperCDMS, NASA's Fermi Gamma-Ray Space telescope, Large Hydron Collider (LHC) at Geneva etc... Experimental efforts to detect WIMPs include the search for products of WIMP annihilation, including gamma rays, neutrinos and cosmic rays in nearby galaxies and galaxy clusters; direct detection experiments designed to measure the collision of WIMPs with nuclei in the laboratory, as well as attempts to directly produce WIMPs in colliders, such as the LHC. However all the efforts in this direction has been fruitless so far.

The HDM consists of particles to be moving nearly at the speed of light, when the pre-galactic clumps began to form. HDM includes massive ($\sim eV$) neutrinos. The neutrinos are the only hot dark matter candidate as they are light enough to move with the speed of light. The Universe is full of neutrinos left over from just after the Big-Bang, when matter and anti-matter were formed. There are huge amount of neutrinos, that if they have just a tiny mass, then they can significantly account for the dark matter. The dark matter that has properties intermediate between those of hot dark matter and cold dark matter named as Warm Dark Matter (WDM). WDM is composed of sub-relativistic particles having masses ($\sim keV$) causing structure formation to occur bottom-up (micro to macro scale)from above their free-streaming scale, and top-down (macro to micro scale) below their free streaming scale. The most common WDM candidates are considered to be sterile neutrinos and gravitinos. The WIMPs when produced non-thermally could be candidates for WDM [9].

The most widely discussed models for non-baryonic dark matter is based on the CDM hypothesis. CDM leads to bottom-up formation of structure in the Universe i.e. small scale structures led to the formation of large scale structures. On the other hand, the HDM results in top-down formation scenario i.e. first super-cluster formed and then galaxies and then the formation of small structure takes place. However, WDM has intermediate role in large scale structure formation.

# 4 Neutrino Dark Matter

Neutrinos are most abundant particles in the Universe. They are electrically neutral and have tiny mass. Out of four interactions in nature neutrinos interact only via the weak interaction and feebly via gravitational force. They rarely interact with any material, which makes experimental detection of these particles extremely challenging. There are three types of neutrinos so far detected, which are denoted as electronic ($\nu_e$), muonic ($\nu_\mu$), and tauonic ($\nu_\tau$) flavour eigenstates. In fact, in the Standard Model of particle physics, neutrinos are massless. However, in the late 90's and beginning of $21^{st}$ century, physicists observed neutrino oscillation, a quantum mechanical effect which would not occur unless neutrinos have mass. The theory of neutrino oscillation describes the flavor eigenstates as the mixing or linear superposition of mass eigenstates $\nu_1$, $\nu_2$, $\nu_3$. For two flavour case the mixing is shown as

$$\begin{pmatrix} \nu_e \\ \nu_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \tag{12}$$

where $\theta$ is a mixing angle. From the observation of the neutrino oscillations phenomenon, it is confirmed that neutrinos have mass. The nature of neutrinos is not yet understood i.e. whether they are Dirac or Majorana particles. In case of Dirac nature neutrino and antineutrino are different, while in the Majorana nature they are the same particle. Despite the tininess, the neutrino mass has far-reaching implications in astrophysics and cosmology.

Neutrinos are considered to be the constituent of dark matter via thermal mechanism. As discussed above the hot dark matter is the matter that was relativistic until just before the epoch of galaxy formation, neutrinos of very low mass are strongest candidates for hot dark matter. It is believed that neutrinos were in thermal equilibrium with the hot plasma which filled the early Universe. As the Universe expanded and cooled, the rates of weak interaction processes decreases and neutrino decoupled when these rates became smaller than the Hubble expansion rate. Since for the three known light neutrinos with masses smaller than $1eV$, the decoupling occurred when they were relativistic called hot relics. As their interaction of cross section with matter is very small therefore, the direct detection of these relativistic neutrinos is an extremely difficult task. In early Universe, when $1MeV \leq T_\gamma \leq 100MeV$, neutrinos were kept in equilibrium with primordial plasma by the weak interactions. The reactions of neutrinos with nucleons were negligible, because the number density of the non-relativistic nucleons was much smaller than the density of relativistic electrons and positrons. The interaction rate for each neutrino is given by[13]

$$\Gamma\ = n < \sigma v >, \tag{13}$$

where $n$ is the number density of target particles, $\sigma$ is the cross-section and $v$ is the neutrino velocity. The bracket denote the thermal averaging. For weak interaction processes

$$< \sigma v > = G_F^2 T_\gamma^2, \tag{14}$$

where the temperature ($T_\gamma$) gives the order of magnitude of the energies of the relativistic particles participating in the reactions. As the number density of relativistic particles is given as $n \sim T_\gamma^3$, the interaction rate for each neutrino became

$$\Gamma \sim G_F^2 T_\gamma^5. \tag{15}$$

So we can say that interaction rate decreases rapidly with the decrease of the temperature due to expansion of the Universe and we obtain the decoupling temperature for neutrinos $T_\gamma^\nu \sim 1MeV$.

If the active neutrinos have a non-zero mass, as indicated by several neutrino oscillation experiments, the sterile neutrinos will take part in the neutrino oscillations. The sterile neutrinos are 'sterile' as they practically inactive, and they don't interact via any other interactions with active neutrino except by mixing [21]. This allows a possibility for a radiative decay under emission of an X-ray photon with energy of half the sterile neutrino mass. However, it needs much more confirmation before one accepts this as the explanation.

The sterile neutrino was originally proposed as a dark matter candidate by Dodelson and Widrow in 1993 to solve the discrepancies between the CDM predicted structure formation and observations [20]. Since neutrinos were relativistic at the time of decoupling, the number density of relic neutrinos is given by the relativistic expression independent from the values of their masses. In other words, light neutrinos are hot relics and contribute to the hot dark matter in the Universe. Sterile neutrinos have been invoked to generate masses for light neutrinos; as such the mix with light neutrinos and hence can be produced via oscillations [20]. With this mechanism, their relic density is estimated to be

$$\Omega_N \approx \left(\frac{\sin^2\theta}{3\times10^{-9}}\right)\left(\frac{M_N}{3keV}\right)^{1.8}. \tag{16}$$

Here, $\theta$ is the mixing angle between the sterile neutrinos $N$ with mass $M_N$ and the active neutrinos. It has been seen that a viable sterile neutrino to be the dark matter candidate requires a mass of $keV$ and a very small mixing angle. It is a WDM candidate and its interactions are dominated by gravity, as preferred by the structure formation [22].

Neutrinos with masses much smaller than the effective neutrino temperature are still relativistic and have negligible contribution toward the energy density of the Universe. Despite the second most abundant particles after the photons, neutrinos fail to accommodate the observed abundance of dark matter. The relic density of light neutrinos is fixed as

[13]

$$\Omega_\nu^0 h^2 = \frac{\sum_i m_i}{94.14eV}. \tag{17}$$

Thus, the neutrino energy density is proportional to the sum of neutrino masses. This value is relevant for the present energy balance if there are neutrinos with masses of the order of $1eV$ or more. Before the neutrino decoupling around $T_\gamma \simeq 1MeV$, the weak processes were in equilibrium.

$\Omega > 1$, implies a closed Universe, which means that at some time the gravitation attraction will stop the expansion and Universe will collapse again. An $\Omega < 1$, means a Universe which expands forever. However $\Omega = 1$ means a flat Universe. At present time, $\Omega$ is changing on time scale of seconds. Since our existence is not compatible with the Universe which is either closed or continuously expanding, the only long term value that $\Omega$ is close to unity. Although the detailed physical mechanism for driving the expansion is not well determined and differs in different grand unified theories.

The phenomenon of sudden and fast expansion of Universe caused by a scalar field present in the nascent stage is known as 'inflation'. Inflation provides a possible mechanism to set the initial conditions. From the inflation paradigm, it is the argument that the only long lived natural value for $\Omega$ is unity and that inflation provided the early Universe with the mechanism to achieve that value and thereby solve some of the main problems of standard model of cosmology; e.g. the flatness and smoothness problems.

The WMAP-7 data provides a quite stringent constraint on the sum of neutrino masses of $\sum m_\nu < 1.3\ eV$ at 95% c.l. [10], which is more constrained than $\approx 2.1eV$, that is the first releases [11]. However, the most recent and sophisticated analysis of Lyman-a data gives an upper bound of 0.9 eV for the sum of neutrino masses. In summary, at present the bound on the sum of neutrino masses can be in the range between 0.3 and more than 2 eV, depending on the data and parameters used. The bound can be relaxed somewhat when more parameters, such as sterile neutrinos ($\nu_s$) are included. In the most conservative case the bound is above 2.5 eV if only CMB data is used. When CMB data is combined with LSS data in the linear or almost linear regime, combined with a prior on the Hubble parameter the upper bound is robustly below 1 eV. This is true even for extended models. Here it should perhaps also be noted that the bound on neutrino mass from cosmic structure formation applies to any other, hypothetical particle species which decouples while still relativistic. This could for example be low mass sterile neutrinos. It could also be relatively high mass axions which decouple after the QCD phase transition.

Neutrinos have a kinematical advantages over the dark matter candidates is that they cluster on large scales, where the dark matter is needed to hold the large clusters of galaxies. In HDM, since they decoupled at a temperature of the order of $1MeV$ when they were relativistic and formed relativistic HDM gas. The HDM perturbations within the horizon are erased by free streaming (i.e. the random particle velocities close to the velocity of light disperse all HDM over-densities). Free streaming ceases when the HDM gas becomes non-relativistic at some red-shift $Z_{nr}$. Thus, only the HDM perturbations with wavelength larger than the horizon distance at $Z_{nr}$ survive and can take part in the generation of structure in the Universe. Since the horizon distance at $Z_{nr}$ is typically much larger than the volume corresponding to the galactic size masses, so in a Universe dominated by HDM, the formation of structures must proceed according to top-down mechanism. However the observed statistical properties favours bottom-up mechanism i.e. small structures leading to the formation of large scale structures. Hence the HDM contribute to the formation of small scale structures while CDM is responsible for binding of large scale structures [13].

The standard model does not predict any masses for the active neutrinos, but as stated above the masses are required by the experimentally verified neutrino oscillations. A simple way to incorporate the neutrino masses is to extend the model with the right-handed neutrinos just as done for the other elementary particles of SM. It is possible to add an arbitrary number of sterile neutrinos, but at least three sterile neutrinos are needed to explain the neutrino oscillations, the baryon asymmetry, and the dark matter [14]. The successful 'three sterile neutrinos' extension of the standard model is called the (Neutrino Minimal Standard Model)($\nu$MSM). It is re-normalisable and in agreement with most particle physics experiments [15]. The Big-Bang production of $^4He$ increases with $\eta$. Thus upper limit to $^4He$ abundance and a lower limit to baryon density lead to an upper limit to number of neutrino species $N_\nu$ (i.e. so called BBN bound). The lower limit to baryon density is based on the Big-Bang production of deuterium $^2H$, which rises rapidly with decreasing baryon density. Since all the neutrons end up in forming $^4He$, which is the most tightly bound stable light nucleus, the mass fraction of $^4He$ is denoted as $Y_p$, and is given by [13]

$$Y_p \simeq \left( \frac{2n_n}{n_n + n_p} \right) \simeq 0.25. \qquad (18)$$

As per recent estimates of $Y_p$ with conservative assumptions - for $^3He$ chemical evolution and $Y_p = 0.252$, less than four neutrino species are possible; however, for extreme assumption- no limit to primeval deuterium- less than five neutrino species are allowed which implies there exist fourth neutrino flavor that is sterile neutrino. In summary, there are healthy signatures for additional degrees of freedom $N_\nu > 3$ i.e. the species of sterile neutrinos from various studies, which are given below [16]: $2.98 < N_\nu < 4.48$ [BBN](68% CL); $3.03 < N_\nu < 7.59$ [WMAP5+SDSS-DR7+Ho ] (95% C.L.); $3.46 < N_\nu < 5.20$ [WMAP7+BAO+New Ho](68% CL); $4.0 < N_\nu < 6.6$ [WMAP7+ACT data](68% C.L.); $2.22 < N_\nu < 9.66$ [WMAP3](68% CL).

Using recombination-era observables including the CMB, the shift parameter RCMB and the sound horizon from Baryon Acoustic Oscillations (BAO) severely constrain the sterile neutrino $\sin 2\theta < 0.026(m_s/eV)^{-}2.$[17]. Recent bounds on the mixing between the active and the sterile neutrinos have been derived from the combination of neutrino oscillation data and direct experimental searches for sterile neutrinos.[18] Electron neutrino-sterile neutrino mixing bound [19] from joint fits of solar, KamLAND, Daya-Bay and Reno experiments is $\sin 2\theta_{es} < 0.2$. and the analysis of cosmological data in terms of $\Lambda CDM$ constrains the mass square difference with one sterile family $\Delta m^2_{41} < 0.25 eV^2$.

# 5 Conclusions

Ever-since the dawn of civilisation man has been fascinated by the stars, planets and other heavenly objects, wondering what essentially the magnificent Universe was made up of. We learnt that our Universe is almost completely dark. To understand this mystery was the main thrust to know more about the invisible Universe. The story of dark matter began nearly a century ago, when Kapteyn and Jeans propounded of existing a such kind of weird matter. Later, Smith and Zwicky discovered that in some large clusters of galaxies the individual members are moving so rapidly that their mutual gravitational attraction is insufficient to keep the clusters from flying apart. Either such clusters should be dissolving or there must be enough dark matter present to hold them together. Since, almost all the evidences suggest that clusters of galaxies are stable configuration. Hence it was concluded that the clusters consist of both luminous and non-luminous matter, which was termed as dark matter.

In this paper we have discussed about the dark matter, various experimental hints and evidences for dark matter, its composition, the role of neutrinos in dark matter formation and understanding of its dynamics. Baryon to photon ratio shows that our Universe is not closed by baryonic matter, which gives a clear indication of the existence of dark matter. Given the properties, neutrinos fit to be a strong candidate constituent for the dark matter as they have an advantage over other dark matter candidates, e.g. they cluster on large scale where the dark matter is needed to hold the large clusters of galaxies. Despite the weakness of interactions and smallness of masses, they can play an important role in cosmology.

In addition to three active flavours of neutrinos, there could also exist extra massive neutrino states that are sterile, i.e. they are singlets of the Standard Model gauge group and thus insensitive to weak interactions. Most of the current data on neutrino oscillations can perfectly be explained with only three active species, but there exist a few experimental results [23]-[28] that cannot be explained in this framework. If neutrino oscillations are responsible for all the experimental data, a solution might require additional (sterile) neutrino species. These kind of particles are predicted by many theoretical models beyond the SM [29]. Their masses are usually heavy, while lighter sterile neutrinos are rarer but possible. Recent studies propose sterile neutrino with a mass of the order of a few keV's and a very small mixing with the active neutrinos. Such heavy neutrinos could be produced by active-sterile oscillations but not fully thermalized, so that they could play the role of dark matter and replace the usual CDM component. But due to their large thermal velocity (smaller than that of active neutrinos), they would behave as WDM and erase small-scale cosmological structures. At present the neutrino physics and neutrino astrophysics and cosmology are at the cross roads. On the one hand, it is impossible to deny that neutrinos oscillate and thus presumably have small masses, and on the other unless a sterile neutrino truly exists, there is a sense that neutrino masses are too small to be of very much cosmological interests.

In the galaxy formation scenario, galaxies can only form by the collapse of super-clusters. The detailed study shows that the collapse of super-clusters only happens very late and may be in contradiction with the existence of quasars of large red shift. Although the evidences for dark matter is wide and deep and existence of dark matter is based on the assumption that the laws of motion and gravity as formulated by Newton and extended by Einstein apply. On the other hand the modification in the theory of gravity can explain the effects attributed to dark matter and some scientists have proposed MOND (Modified Newtonian Dynamics). According to this theory at very low acceleration, corresponding to large distances, the usual law of gravitation is modified. Although MOND has had some success in explaining observations of galaxies, but failed to explain the observation of Bullet Clusters. So we need more experimental evidences to give a conclusive theory of dark matter.

# Acknowledgments

We thank Debasish Majumdar, SINP, for giving useful inputs. The Inter-University Centre for Astronomy & Astrophysics (IUCAA), Pune is also acknowledged for providing research facilities during the completion of this work.

# References

[1] Massey, R. et al., Dark matter maps reveal cosmic scaffolding, doi:10.1038/nature05497, 7 January 2007 as the related references.

[2] E. Copeland, *Dark Matter in the Universe*, Contemporary Physics, 35(1), 1994.

[3] www.google.co.in/imghp [image taken from google].

[4] V. Rubin, *Dark Matter in Spiral Galaxies*, Sc. Am. 248(6), 1983.

[5] A.C. Melissinos, Lecture Notes on Particle Astrophysics, Physics 593- Spring 1995.

[6] S. Sharma, *Bulletin of the Indian Association Of Physics Teachers*, 6(9), 2014.

[7] A. Boyarsky, O. Ruchayskiy, D. Iakubovskyi and J. Franse, *Unidentified Line in X-Ray Spectra of the Andromeda Galaxy and Perseus Galaxy Cluster*, Phys. Rev. Lett. 113, 251301, 2014.

[8] P. Gondolo, *Introduction to Non-Baryonic Dark Matter*, Lecture delivered at the NATO Advanced Study Institute, France [astro-ph/0403064].

[9] S. R. Sorensen, *Sterile neutrinos as a dark matter candidate*, Master Thesis in Physics, Niels Bohr Institute, Dark Cosmology Centre, August 25, 2006.

[10] E. Komatsu et al., *Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation*, WMAP Collaboration, Astrophys.J.Suppl. 192 18 (2011).

[11] K. Ichikawa, M. Fukugita and M. Kawasaki, Phys. Rev. D 71, 043001 (2005).

[12] S. Dodelson and L. M. Widrow, *Sterile-neutrinos as dark matter*, Phys. Rev. Lett. 72 (1994) 17?20.

[13] C. Giunti, C. W. Kim, *Fundamental of Neutrino Physics and Astrophysics*, Oxford University Press, New York, 2007.

[14] T. Asaka and M. Shaposhnikov, , Phys. Lett. B 620, 17, 2005.

[15] T. Araki and Y. F. Li, $Q_6$ *Flavour Symetry Model for the Extension of the MSM by the three Right Handed Sterile Neutrinos*, Phys. Rev. D85, 065016, 2012.

[16] P. C. de Holanda and A. Yu. Smirnov, hep-ph/1012.5627 [references contained therein].

[17] A. C. Vincent, E. Fernandez Martinez, P. Hernandez, M. Lattanzi, O. Mena, JCAP 04 (2015)006.

[18] O. Ruchayskiy, A. Ivashko, JHEP 1206 (2012) 100.

[19] A. Yu. Samirnov, Nucl. Phys. Proc. Suppl. 235, 431 (2013) [references contained therein].

[20] S. Bilenky, *Introduction to the Physics of Massive and Mixed Neutrinos*, Springer, Lecture Notes in Physics, 817, 2010.

[21] Signe Riemer-Srensen, *Sterile neutrinos as a dark matter candidate*, Master Thesis in Physics Niels Bohr Institute, 2006.

[22] C. J. Copi, D. N. Schramm, and M. S. Turner, *Big-bang Nucleosynthesis limit to the number of neutrino species*, Phys. Rev. D 55, 3389, 1997.

[23] A. Aguilar-Arevalo, *et al.*, (LSND Collaboration), *Evidence for neutrino oscillations from the observation of anti-neutrino(electron) appearance in a anti-neutrino(muon) beam*, Phys.Rev. D 64, 112007, 2001.

[24] A. Aguilar-Arevalo, *et al.*, (MiniBooNE Collaboration), Phys, Rev, Lett. 102, 101802, 2009.

[25] A. Aguilar-Arevalo, *et al.*, (MiniBooNE Collaboration), Phys. Rev. Lett. 110, 161801, 2013.

[26] G. Mention *et al.*, Phys. Rev. D 83, 073006, 2011.

[27] T. Mueller *et al.*, Phys. Rev. C 83, 054615, 2011.

[28] D. Frekers *et al.*, Phys. Lett. B 706, 134, 2011.

[29] S. Antusch *et al.*, *Unitarity of the Leptonic Mixing Matrix*, JHEP 0610, 084, 2006.

[30] D. O. Caldwell, *Neutrino Dark Matter*, University of California, Santa Barbara, CA 93106-9530, USA, hep-ph/9902219.

[31] G. G. Raffelt, *Neutrino Astrophysics At The cross roads*, Proc. of Summer School in High Energy Physics & Cosmology at ICTP Italy, World Sc., 1998 [ hep-ph/9902271].

[32] K. Zuber, *Neutrino Physics*, CRC Press, first edition, 2013.