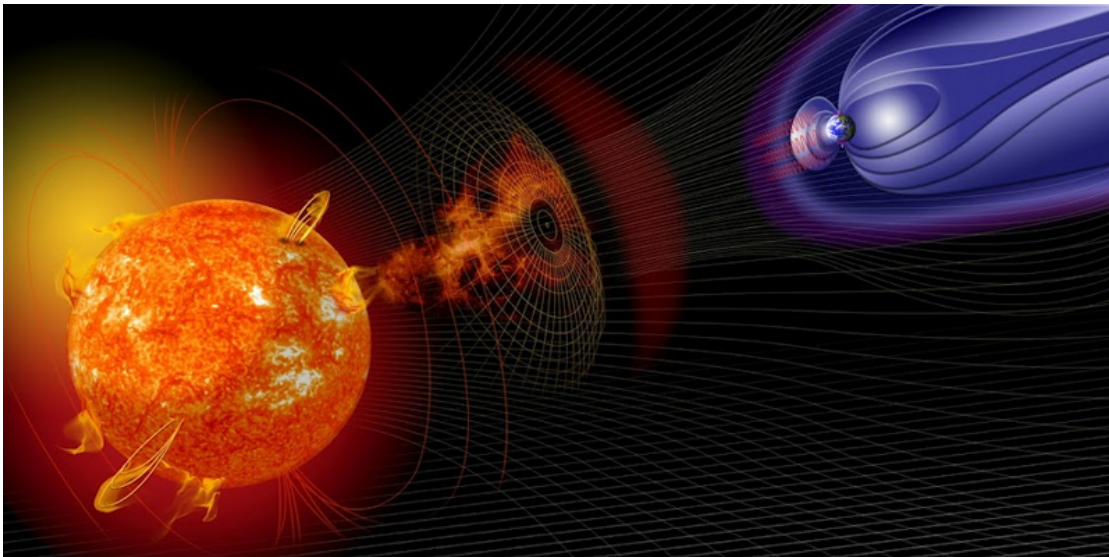


Space Physics and Measurement Techniques

Lecture Notes

Written by
Kjell Rönmark, Asta Pellinen-Wannberg,
Eva Krämer, Patrik Norqvist, Anders Eriksson

With contributions from
David Andrews, Maria Hamrin, and Sara Östmann.



Umeå University
Department of Physics
September 3, 2026

Release notes for version 2.0.0

This compendium have been used in different forms for many years. Recently we have decided to restructure and rework parts of this compendium. This means that this compendium is still ongoing work and we constantly try to improve on it to make it more comprehensible and more complete. As we add new things or rewrite old things, we sometimes introduce errors. Therefore we need YOUR help to improve on it for future generations of students. Please feel free to contact your supervisor at any time with suggestions and corrections! We appreciate all feedback.

Contents

1. Introduction	7
I. Basic Plasma Physics	9
2. Charged Particles in an Electromagnetic Field	11
2.1. Introduction	11
2.2. Constant magnetic field, $\mathbf{B} = B_0 \hat{z}$	11
2.2.1. Magnetic moment	13
2.3. Constant $\mathbf{B}$ and $\mathbf{E}$ fields	13
2.4. Slowly varying magnetic fields	14
2.4.1. Variations perpendicular to $\mathbf{B}$	15
2.4.2. Variations parallel to $\mathbf{B}$	17
2.4.3. Conservation of magnetic moment	18
2.5. Adiabatic invariants	19
2.5.1. Pitch angles and the loss cone	20
2.6. Conclusion	20
3. Plasma Physics	21
3.1. Basic plasma behaviour	21
3.2. Definition of a plasma	21
3.3. Kinetic plasma theory	24
3.3.1. The distribution function	24
3.3.2. Reminder: Total/Convective derivatives	25
3.3.3. The Vlasov equation	25
3.4. Magnetohydrodynamic theory	26
3.4.1. Fluid equations	27
3.4.2. Magnetohydrodynamics - single fluid equations	27
3.4.3. Ohm's law	28
3.4.4. Frozen-in magnetic fields	29
3.4.5. Magnetic pressure	31
3.4.6. MHD generator	32
3.4.7. Flux ropes	33
3.4.8. Break down of MHD - magnetic reconnection	34
4. Waves in Plasma	37
4.1. Introduction	37
4.2. The Basic Equations	37
4.3. Basic classification of waves	38
4.4. Linearisation	40
4.5. Electrostatic waves	40
4.5.1. Plasma oscillations/ Langmuir waves	41
4.5.2. Ion acoustic waves	42

4.6.	Magnetohydrodynamic waves	44
4.6.1.	Magnetohydrodynamic waves in cold plasmas	44
4.6.2.	Magnetohydrodynamic waves in warm plasmas	49
II.	The space environment	51
5.	The Sun and the Solar Wind	53
5.1.	Introduction	53
5.2.	The Sun's interior	53
5.2.1.	Energy production and transport in the sun	53
5.2.2.	The Solar Cycle and Sun's magnetic field	55
5.3.	The Sun's atmosphere	55
5.3.1.	The photosphere	55
5.3.2.	The chromosphere	56
5.3.3.	The corona	58
5.4.	The Solar Wind	58
6.	The Magnetosphere	63
6.1.	Introduction	63
6.2.	Coordinate systems	63
6.3.	The Earth magnetic field	64
6.3.1.	Particle motion in the Earth magnetic field	66
6.3.2.	Summary of particle motion in the magnetosphere	67
6.4.	The bow shock and the magnetosheath	68
6.5.	The magnetopause	70
6.6.	The magnetotail and magnetospheric convection	72
6.7.	The inner magnetosphere	74
6.8.	Ring current	75
6.9.	The radiation belts	76
6.10.	The high latitude magnetosphere	76
6.11.	Current systems in the ionosphere	77
6.12.	Magnetospheres around other bodies	77
7.	The upper Atmosphere and Ionosphere	83
7.1.	Introduction	83
7.2.	The atmosphere	83
7.2.1.	Origin of the atmosphere	83
7.2.2.	Height profile of the atmosphere	83
7.2.3.	Ozone layer	85
7.3.	The ionosphere	88
7.3.1.	History	88
7.3.2.	Formation of the ionosphere	88
7.3.3.	The ionospheric layers	89
7.3.4.	Collisions and conductivity	91
7.3.5.	Height integrated conductivity	94
7.3.6.	Ionospheric currents	95
8.	The Aurora	97
8.1.	Introduction	97
8.2.	Auroral light emission	97

8.3. Diffuse aurora	99
8.4. Auroral waves and ion heating	100
8.5. Substorms and discrete aurora	101
8.6. Short summary	105
9. Space Weather	107
9.1. Introduction	107
9.2. Geomagnetic storms	108
9.3. Space Weather effects	110
9.3.1. Geomagnetically induced currents	110
9.3.2. Navigation	110
9.3.3. Increased radiation	111
9.4. Short summary	111
III. Space engineering	113
10. Mechanics of Rockets and Satellites	115
10.1. Introduction	115
10.2. Rockets in Free Space	115
10.3. Launch Mechanics	117
10.4. Stability of a rocket	118
10.5. Motion Under a Central Force	119
10.6. Elliptic Orbits	122
10.7. Hyperbolic trajectories	124
10.8. Determination of satellite orbits from initial conditions.	125
10.9. Coordinate systems and orbital elements	127
10.10 Non-Keplerian Effects	129
10.11 Example of Orbits around Earth	132
10.12 Lagrange points	134
10.13 Transfer between circular orbits	137
10.14 Gravity Assist	138
11. Space Measurements	143
11.1. Introduction	143
11.2. Magnetic fields	143
11.2.1. Search Coil magnetometers	144
11.2.2. Flux gate magnetometer	144
11.2.3. Ground based magnetometers	145
11.3. Electric fields	146
11.4. Particle distributions	146
12. Ground-based space measurements	151
12.1. Introduction	151
12.2. Solar activity indices	152
12.3. Magnetometers	153
12.4. Geomagnetic activity indices	155
12.5. Riometers	155
12.6. Optical methods	157
12.7. Radars	158
12.7.1. The radar equation	159

Contents

12.7.2. Ionosonds	160
12.7.3. Coherent radars	161
12.7.4. Incoherent scatter	163
12.7.5. EISCAT - The European Incoherent SCATter radars	164
12.7.6. EISCAT_3D	166

1. Introduction

Northern lights - most of you have probably seen (or at least heard) about the phenomena that can be observed during dark nights. The Northern lights, scientifically known as the Aurora Borealis, results from the complicated interaction between particles coming from the Sun and the Earth's magnetic field. Northern lights is just one of the many phenomena that space physicists have been studying for numerous years.

Space physics, a relatively young branch of physics, revolves around understanding the dynamics of space and utilizes *in situ* ("at the place") measurements obtained from spacecraft. The dawn of space physics can be traced back to the historic launch of the first Sputnik satellite on October 4, 1957, which opened the door to a whole new realm of scientific exploration. Nevertheless, prior to that milestone, significant studies of physics in space were already underway, particularly through the investigation of radio wave propagation.

Space also serves as a laboratory of large scale phenomena that can not be reproduced on Earth. Space physics has shown that the space between planets and stars is not empty but filled with particles. The particles are both neutral and charged, have wide ranges in their energies, and interact with one another (mostly) via electromagnetic fields, and are capable of causing a range of disturbances and phenomena. As humanity continues to explore and use space for satellites it is important to understand this environment. The space environment continuously challenges scientists and engineers with high energy events that can possibly harm human infrastructure.

This compendium provides an introduction to space physics, divided into three main parts: Basic plasma physics, the space environment, and Space engineering.

The first part introduces plasma, a state of matter describing the collective behavior of charged particles. We will mainly consist plasma that consists of negatively charged electrons and positively charged protons. Plasma constitutes about 99% of the baryonic matter in the universe. Understanding and probing plasmas does therefore not only give rise to processes in our solar system but also in the remaining Universe. First we discuss the motion of a single charged particle in an electromagnetic field in chapter 2. The motion of a plasma is dominated by electromagnetic forces and dominates its collective behaviour. In order to understand plasma we also need to understand and describe the collective particle behaviour which will be covered in chapter 3. In order to understand energy and momentum transfer in plasma environments a good understanding of different wave phenomena in a plasma is needed. A brief introduction on different waves in plasmas will be given in chapter 4.

In Part II we will move on the plasma environment in our solar system. Our space environment is driven by supersonic plasma flow originating from the sun, the solar wind. Chapter 5 gives an introduction to the fundamental processes in our sun and typical characteristics of the solar wind. The plasma environment close to planets results

1. Introduction

from the interaction of the solar wind with the planets magnetic field or it's atmosphere. Chapter 6 presents the near Earth space environment and its most prominent features. The ionosphere is the natural boundary between space and our atmosphere. It is heavily affected by extreme space conditions and is the place where aurora can be observed, these phenomena will be covered in chapter 7 and chapter 8.

The last part, Part III, of this compendium covers the basics of space engineering. Space physicists probe the space environment with artificial satellites. The motion of these satellites is governed by the laws of celestial mechanics, which have been developed over centuries by astronomers to describe the motion of celestial bodies in a gravitational field. Rocket propulsion and celestial mechanics are also briefly introduced in chapter 10. In order to interpret and make use of the data collected by artificial satellites it is crucial to understand how different scientific instruments work. Ground-based measurements complement spaceborne missions, providing a cost-effective way to monitor the near-Earth space environment throughout the year. In chapter 12 we will give a brief introduction of the ground-based instrumentation used to monitor space.

Part I.

Basic Plasma Physics

2. Charged Particles in an Electromagnetic Field

2.1. Introduction

Space, from the ionosphere to far away galaxies, is filled with ionized matter in the form of free electrons and positively charged ions. Gravitational forces typically play only a minor role for these charged particles. Instead their motion is determined almost exclusively by electromagnetic forces. Before we describe the collective motion of many charged particles (a plasma), we will first consider the motion of a single charged particle in an electromagnetic field. Even though this approach is impractical to describe the collective motion, the single particle motion is important for some effects that we will consider in this compendium. In chapter 3 we will then consider the collective behavior of charged particles using the *fluid* and *kinetic* approach.

In general, the electromagnetic field in space can be divided into two components. One, large scale, slowly varying component that is generated by large charge and current distributions far away. Another, small scale, rapidly fluctuating component is generated by a few particles nearby. When two particles come close to each other, this second component becomes very strong, and we have a collision. However, except in stars, planets, and other celestial bodies, the density is very low in space. In most cases, the density is so low that particles can travel long distances without colliding. As a first approximation, we can then try to understand the physics of charged particles in space by studying the motion of a single particle, neglecting the interaction with other nearby charges. Often, the theory presented below can be used to describe the motion of particles with energy W in the range $1 \text{ eV} < W < 1 \text{ MeV}$, but the validity of the approximations must be verified in each case.

The force on a charged particle in an electromagnetic field is given by the Lorentz force equation

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \quad (2.1)$$

which when inserted into Newton's equation $m \frac{d\mathbf{v}}{dt} = \mathbf{F}$ yields the equation of motion

$$\frac{d\mathbf{v}}{dt} = \frac{q}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \quad (2.2)$$

As discussed above, we assume that there are no other particles in the vicinity, and that the fields $\mathbf{E}$ and $\mathbf{B}$ are smooth and slowly varying.

2.2. Constant magnetic field, $\mathbf{B} = B_0 \hat{z}$

Let us first consider the simplest case, which is motion in a homogeneous magnetic field $\mathbf{B} = B_0 \hat{z}$ with no electric field, $\mathbf{E} = 0$. Writing the equation of motion in Cartesian

2. Charged Particles in an Electromagnetic Field

components when the magnetic field is in the $\hat{\mathbf{z}}$ direction we find

$$\frac{dv_x}{dt} = \frac{q}{m}v_y B_0 \quad (2.3a)$$

$$\frac{dv_y}{dt} = -\frac{q}{m}v_x B_0 \quad (2.3b)$$

$$\frac{dv_z}{dt} = 0. \quad (2.3c)$$

The last of these equations simply tells us that v_z will remain constant, unaffected by the magnetic field. To simplify the first two equations we introduce the gyrofrequency

$$\omega_c = \frac{qB_0}{m}, \quad (2.4)$$

and take the time derivative of Equation 2.3a and 2.3b to find

$$\frac{d^2v_x}{dt^2} = \omega_c \frac{dv_y}{dt} = -\omega_c^2 v_x \quad (2.5a)$$

$$\frac{d^2v_y}{dt^2} = -\omega_c \frac{dv_x}{dt} = -\omega_c^2 v_y. \quad (2.5b)$$

The two velocity components are now separated, and they both satisfy the equation $\ddot{v}_{x,y} = -\omega_c^2 v_{x,y}$. The general solution of these differential equations may be written

$$v_{x,y} = A_{x,y} \cos \omega_c t + B_{x,y} \sin \omega_c t. \quad (2.6)$$

The integration constants $A_{x,y}$ and $B_{x,y}$ must be chosen so that Equation 2.3 are satisfied at all times. This constraint implies $A_x = -B_y$, $A_y = B_x$. The remaining two constants, A_x, A_y , are determined from the assumed initial conditions. As an example we choose $A_x = 0$ and $A_y = v_\perp$ which gives

$$v_x = v_\perp \sin \omega_c t \quad (2.7a)$$

$$v_y = v_\perp \cos \omega_c t. \quad (2.7b)$$

We find that these values of the constants correspond to the initial conditions $v_x = 0$ and $v_y = v_\perp$ at $t = 0$. Adding the squares of the two equations we obtain $v_x^2 + v_y^2 = v_\perp^2$, which shows that the velocity perpendicular to the magnetic field, $v_\perp$, is constant in magnitude. The particle will move in a circle with radius $r_c = v_\perp/\omega_c$, the cyclotron radius (or gyro radius), and the angular frequency of this motion is ω_c , the cyclotron frequency (or gyro frequency). Notice that the cyclotron radius depends on the particle velocity in addition to the magnetic field strength B and the charge and mass of the particle.

From Equation 2.3c we concluded that the velocity along the magnetic field, v_z , is constant, and entirely determined by the initial conditions. To really obtain a circular motion, we must assume $v_z = 0$. In the general case, with $v_z \neq 0$, the particle will move in a spiral around the magnetic field line. This spiraling motion is very important in space physics, since it means that space plasmas cannot move freely across magnetic field lines. As an example, the Earth's magnetic field prevents the solar wind, the plasma ejected from the Sun, from entering freely into the magnetosphere, which is the region that is dominated by Earth's magnetic field. This will be discussed in detail later.

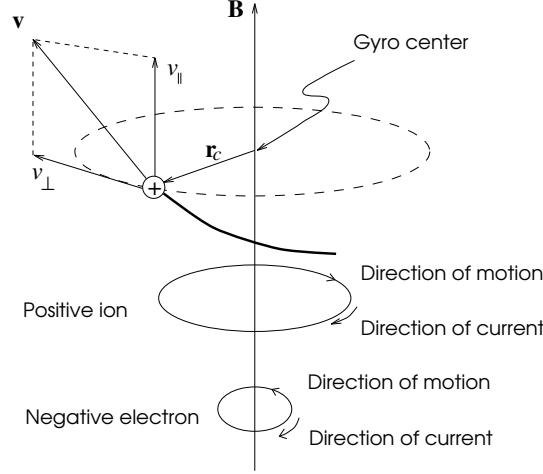


Figure 2.1.: Motion in a magnetic field. Figure from Fälthammar [2001].

2.2.1. Magnetic moment

The circular motion of a particle in a magnetic field may be looked upon as a small current loop, which the charge q travels around $\omega_c/2\pi$ times per second. This corresponds to a current $I = q\omega_c/2\pi$ in the loop. The magnetic moment of such a loop is the current times the area A of the loop. Since the radius of the particle orbit is $v_\perp/\omega_c$ we find the magnetic moment μ from

$$\mu = I\mathbf{A} = -I\pi r_c^2 \hat{\mathbf{b}} = -q \frac{\omega_c}{2\pi} \frac{\pi v_\perp^2}{\omega_c^2} \hat{\mathbf{b}} = -\frac{mv_\perp^2}{2B_0} \hat{\mathbf{b}}. \quad (2.8)$$

The sense of the normal that defines the vector area $\mathbf{A}$ must come from the application of the right hand rule to the current loop I , hence opposite to the magnetic field direction, $\hat{\mathbf{b}} = \mathbf{B}/B$. The magnetic moment depends only on the perpendicular kinetic energy $mv_\perp^2/2$ and on the magnetic field. The direction of the magnetic moment is opposite to $\mathbf{B}$, independent of the sign of the charge. Hence, the gyro motion of the plasma particles tends to weaken the magnetic field—the plasma is diamagnetic.

2.3. Constant $\mathbf{B}$ and $\mathbf{E}$ fields

We now consider the effects of an electric field in addition to the constant $\mathbf{B} = B_0 \hat{\mathbf{z}}$. The force is still given by Equation 2.1. If the electric field has a component parallel to $\mathbf{B}_0$, it will appear on the right hand side of Equation 2.3c as

$$\frac{dv_z}{dt} = \frac{q}{m} E_z. \quad (2.9)$$

Just as in the absence of a magnetic field, the particle will be accelerated and the velocity will increase linearly with time, $v_z = v_0 + q/mE_z t$. This is well known, but the effect of a perpendicular electric field may be more surprising. Choosing $\mathbf{E}_\perp = E_y \hat{\mathbf{y}}$, Equation 2.3a and 2.3b are modified into

$$\frac{dv_x}{dt} = \omega_c v_y \quad (2.10a)$$

$$\frac{dv_y}{dt} = \frac{q}{m} E_y - \omega_c v_x. \quad (2.10b)$$

$$(2.10c)$$

2. Charged Particles in an Electromagnetic Field

Differentiating these equations once more, we find

$$\frac{d^2 v_x}{dt^2} = \omega_c^2 (E_y/B_0 - v_x) \quad (2.11a)$$

$$\frac{d^2 v_y}{dt^2} = -\omega_c^2 v_y. \quad (2.11b)$$

The equation for v_y is the same as Equation 2.5b, and the solution is thus given by Equation 2.6. However, the equation for v_x is now inhomogeneous, and to the solution Equation 2.6 of the homogeneous equation we must add a particular integral. It is easy to verify that

$$v_x = v_\perp \sin \omega_c t + E_y/B_0 \quad (2.12a)$$

$$v_y = v_\perp \cos \omega_c t, \quad (2.12b)$$

satisfies Equation 2.10b and Equation 2.10c if we recall that $\omega_c = qB_0/m$. The new term in Equation 2.12a is a constant drift velocity $v_d = E_y/B_0$. An electric field perpendicular to $\mathbf{B}$ does not accelerate the particle but gives it a constant velocity, and even more surprisingly, the drift velocity is perpendicular to both $\mathbf{B}$ and the electric field $\mathbf{E}$! Furthermore, the magnitude and direction of v_d are independent of the mass and charge of the particle. Light, negative electrons and heavy, positive ions will drift together with the same velocity v_d . In vector notation, the drift velocity can be written

$$\mathbf{v}_d = \frac{\mathbf{E} \times \mathbf{B}}{B^2}. \quad (2.13)$$

This formula is correct even if $\mathbf{E}$ is not perpendicular to $\mathbf{B}$, but the parallel acceleration must then be considered separately.

We can also generalize this result to arbitrary forces $\mathbf{F}$ besides the electric field, by simply substituting $\mathbf{E} = \mathbf{F}/q$ to obtain

$$\mathbf{v}_d = \frac{\mathbf{F} \times \mathbf{B}}{qB^2}. \quad (2.14)$$

Here, $\mathbf{F}$ can be, for example a gravitational force.

2.4. Slowly varying magnetic fields

Above, we have assumed the magnetic fields to be homogeneous. In practice, this is always an approximation—over sufficiently large distances there will surely be some variation. Considering slow variations of the magnetic field, i.e. no sharp boundaries, we can discern three ways to specify what we mean by slow variations:

$$|\frac{\partial}{\partial x} B| + |\frac{\partial}{\partial y} B| \ll \frac{B}{r_c} \quad (2.15a)$$

$$|\frac{\partial}{\partial z} B| \ll \frac{B\omega_c}{v_z} \quad (2.15b)$$

$$|\frac{\partial}{\partial t} B| \ll \omega_c B. \quad (2.15c)$$

In all three cases above, we have assumed the magnetic field to be mainly in the z direction, i.e. $B \approx B_z$.

2.4.1. Variations perpendicular to $\mathbf{B}$

Case Equation 2.15a applies to variations perpendicular to the main direction of the field. We can for example assume $\mathbf{B} \approx (B_0 + y \frac{\partial}{\partial y} B_0) \hat{\mathbf{z}}$, where B_0 and $\frac{\partial}{\partial y} B_0$ are constants. The equation of motion can then be written

$$\frac{dv_x}{dt} = \omega_c v_y + \omega_c v_y \frac{y}{B_0} \frac{\partial B_0}{\partial y} \quad (2.16a)$$

$$\frac{dv_y}{dt} = -\omega_c v_x - \omega_c v_x \frac{y}{B_0} \frac{\partial B_0}{\partial y}. \quad (2.16b)$$

Since $\frac{\partial}{\partial y} B_0$ is small, we can assume that the particle will move almost as when $\mathbf{B}$ is homogeneous, and in the small second term we can replace v_x by $v_\perp \sin(\omega_c t)$ and v_y by $v_\perp \cos(\omega_c t)$ from Equation 2.7a, 2.7b and y by $r_c \sin \omega_c t$. We then find

$$\frac{dv_x}{dt} = \omega_c v_y + \omega_c v_\perp \cos(\omega_c t) r_c \sin(\omega_c t) \frac{\frac{\partial}{\partial y} B_0}{B_0} \quad (2.17a)$$

$$\frac{dv_y}{dt} = -\omega_c v_x - \omega_c v_\perp \sin(\omega_c t) r_c \sin(\omega_c t) \frac{\frac{\partial}{\partial y} B_0}{B_0}. \quad (2.17b)$$

Notice that the second term is guaranteed to be small by the assumption Equation 2.15a, and on a short time scale we expect the particle to move almost in a circle. To find the long term effects of the magnetic field gradient, we average Equation 2.17 over the cyclotron period $T_c = 2\pi/\omega_c$ according to

$$\langle \dots \rangle = \frac{1}{T_c} \int_0^{T_c} \dots dt. \quad (2.18)$$

Since $\langle \cos(\omega_c t) \rangle = \langle \sin(\omega_c t) \rangle = \langle \cos \omega_c t \sin(\omega_c t) \rangle = 0$ and $\langle \sin^2(\omega_c t) \rangle = 1/2$, this averaging leads to

$$\left\langle \frac{dv_x}{dt} \right\rangle = 0 \quad (2.19a)$$

$$\left\langle \frac{dv_y}{dt} \right\rangle = -\frac{1}{2} \omega_c v_\perp r_c \frac{\frac{\partial B_0}{\partial y}}{B_0} = -\frac{v_\perp^2}{2B_0} \frac{\partial}{\partial y} B_0. \quad (2.19b)$$

The averaged Equation 2.19b corresponds to a force $F_y = m \langle \frac{dv_y}{dt} \rangle = -\mu \frac{\partial}{\partial y} B_0$, where we have used the magnetic moment introduced in Equation 2.8. We can now use Equation 2.14 to find the drift velocity as

$$\mathbf{v}_d = \frac{\mathbf{F} \times \mathbf{B}}{qB^2} = \frac{\mu}{q} \frac{\mathbf{B} \times \nabla B}{B^2}. \quad (2.20)$$

Earth's magnetic field is roughly dipolar, and B increases as we approach the planet. A charged particle high above the equatorial ionosphere will see a magnetic field coming up perpendicular to the equatorial plane and increasing in strength towards Earth. This is similar to the situation discussed above, and we conclude that particles in the vicinity of Earth will drift around the planet. Positive and negative particles will drift in opposite directions, and the gradient drift will hence produce a current in the plasma. The drift velocity is proportional to $mv_\perp^2/2$, which tells us that high energy particles will drift

2. Charged Particles in an Electromagnetic Field

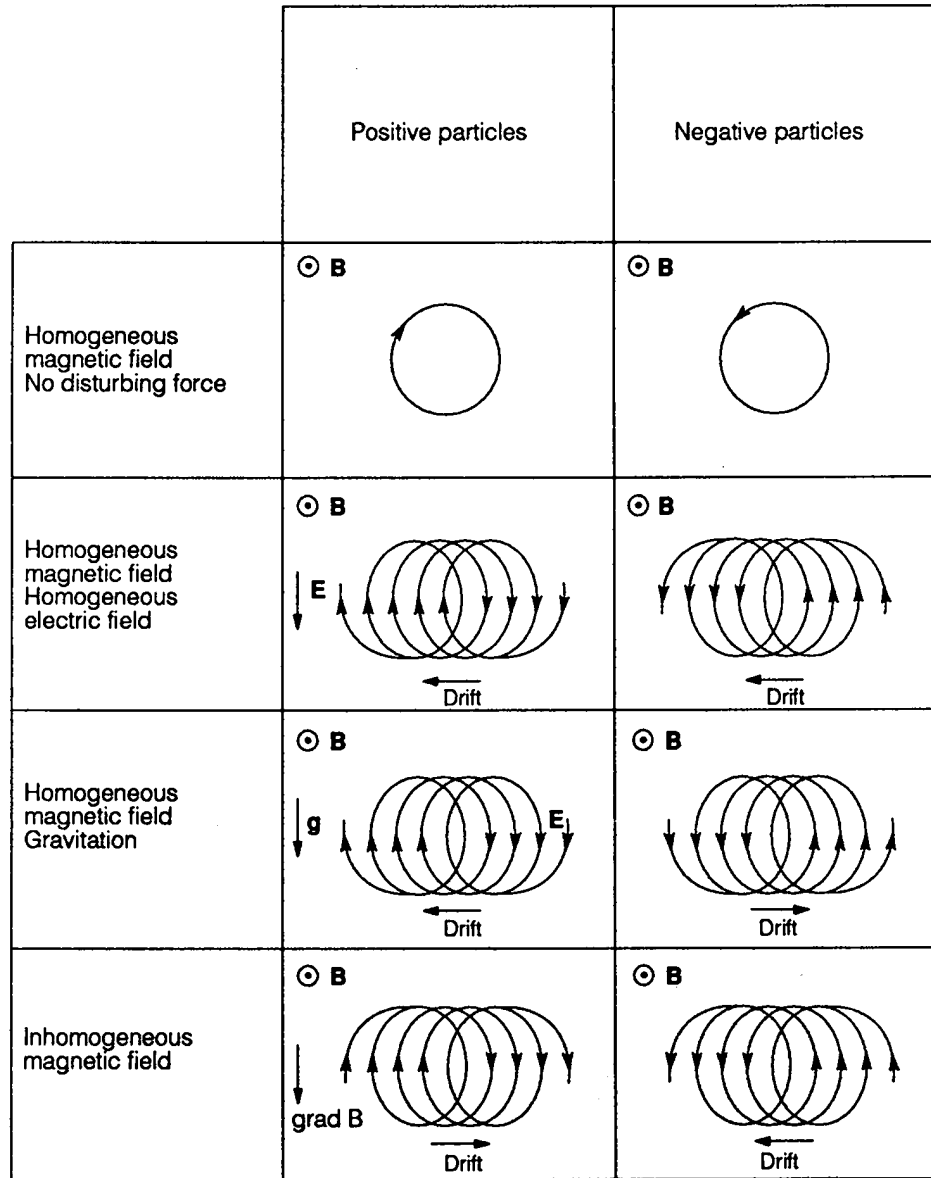


Figure 2.2.: Drift motions of charged particles in a magnetic field under the influence of forces or gradients transverse to the magnetic field lines. Figure from Fälthammar [2001].

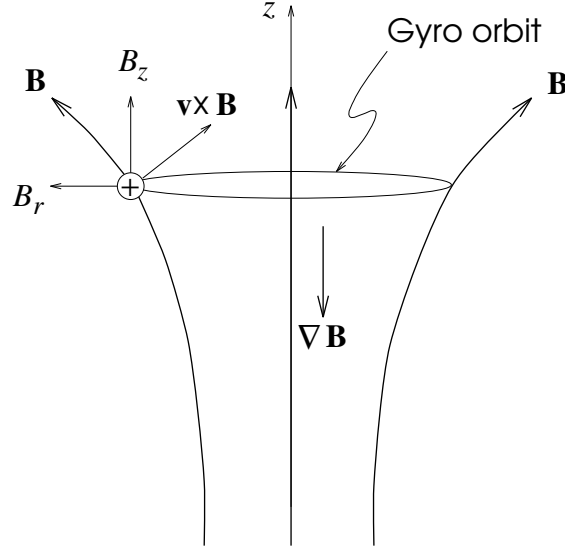


Figure 2.3.: The force on a gyrating particle resulting from the convergence of the field lines. Figure from Fälthammar [2001].

faster. These properties makes the gradient drift qualitatively different from the $\mathbf{E} \times \mathbf{B}$ drift.

A summary of particle motions perpendicular to the magnetic field is given in Figure 2.2.

2.4.2. Variations parallel to $\mathbf{B}$

From Maxwell's equations we know that

$$\nabla \cdot \mathbf{B} = \frac{\partial}{\partial x} B_x + \frac{\partial}{\partial y} B_y + \frac{\partial}{\partial z} B_z = 0. \quad (2.21)$$

This equation tells us that we cannot have a magnetic field of the form $\mathbf{B} = B_z \hat{\mathbf{z}}$, with $B_x = B_y = 0$, unless $\frac{\partial}{\partial z} B_z = 0$. A variation of the magnetic field along the field line is always accompanied by some variation in the transverse direction. As illustrated in Figure 2.3, the increase in the field strength corresponds to converging field lines.

If we assume cylindrical symmetry and introduce cylindrical coordinates, the convergence corresponds to a small radial field component B_r , directed outwards. In cylindrical coordinates, Equation 2.21 is transformed into

$$\nabla \cdot \mathbf{B} = \frac{1}{r} \frac{\partial}{\partial r} r B_r + \frac{\partial}{\partial z} B_z = 0. \quad (2.22)$$

During one gyro period, the particle travels the distance $v_z T_c$ along the field line. Condition Equation 2.15b implies that $v_z T_c \frac{\partial}{\partial z} B_z \ll B_z$, or that the particle only will see a small variation in B_z while going once around the field line. Assuming $\frac{\partial}{\partial z} B_z$ to be constant and $B_r(r=0, z) = 0$, we can integrate Equation 2.22 with respect to r and obtain the radial field component at the distance r from the symmetry axis along a reference field line as

$$B_r = -\frac{r}{2} \frac{\partial}{\partial z} B_z. \quad (2.23)$$

2. Charged Particles in an Electromagnetic Field

This component will give rise to a force in the z direction. The z component of the equation of motion of a particle gyrating around the reference field line is

$$m \frac{dv_z}{dt} = q \mathbf{v}_\perp \times \mathbf{B}_r \cdot \hat{\mathbf{z}} = -qv_\perp \frac{r_c}{2} \frac{\partial}{\partial z} B_z = -\frac{mv_\perp^2}{2B_z} \frac{\partial}{\partial z} B_z = -\mu \frac{\partial}{\partial z} B_z. \quad (2.24)$$

We see that the magnetic field gradient gives rise to a force that tends to push the particles in a direction opposite to the gradient, that is in the direction from stronger to weaker magnetic field.

2.4.3. Conservation of magnetic moment

Let us analyze what happens to the magnetic moment while the particle moves. The total rate of change of the magnetic moment is

$$\frac{d\mu}{dt} = \frac{d}{dt} \left(\frac{mv_\perp^2}{2B_z} \right) = \frac{1}{B_z} \frac{d}{dt} \left(\frac{mv^2}{2} - \frac{mv_z^2}{2} \right) - \frac{\mu}{B_z} \frac{d}{dt} B_z. \quad (2.25)$$

Magnetic forces are always perpendicular to the velocity, so that the total kinetic energy is conserved, i.e. $\frac{d}{dt}(mv^2/2) = 0$. Since the magnetic field is assumed stationary, the change of B_z along the particle orbit is entirely caused by the motion of the particle, which gives us $\frac{d}{dt} B_z \equiv \frac{\partial}{\partial t} B_z + v_z \frac{\partial}{\partial z} B_z = v_z \frac{\partial}{\partial z} B_z$. If we also use $\frac{d}{dt} v_z^2 = 2v_z \frac{dv_z}{dt}$, we find

$$\frac{d\mu}{dt} = -\frac{v_z}{B_z} \left[m \frac{dv_z}{dt} + \mu \frac{\partial}{\partial z} B_z \right] = 0, \quad (2.26)$$

where it follows from Equation 2.24 that the expression in brackets is zero. This means that the magnetic moment of particles moving in an inhomogeneous magnetic field is constant.

It can be shown that the magnetic moment is conserved also when the magnetic field is weakly time dependent, as assumed in Equation 2.15c. The change in μ due to a varying magnetic field strength is

$$\frac{\partial}{\partial t} \mu = \frac{1}{B} \frac{\partial}{\partial t} \left(\frac{mv_\perp^2}{2} \right) - \frac{\mu}{B} \frac{\partial}{\partial t} B, \quad (2.27)$$

where the second term gives the change in the magnetic moment if the velocity were unaffected by the change in B , while the first term gives the change in μ due to the change in $v_\perp$. From Faraday's law, $\nabla \times \mathbf{E} = -\frac{\partial}{\partial t} \mathbf{B}$, we know that a time dependent magnetic field will induce an electric field. The work done by this field during one gyration of the particle is found by integrating the force around the circular orbit, $\oint q \mathbf{E} \cdot d\mathbf{l}$. Dividing by the gyroperiod we find that the perpendicular energy changes at the rate

$$\frac{\partial}{\partial t} \left(\frac{mv_\perp^2}{2} \right) \approx \frac{1}{T_c} \oint q \mathbf{E} \cdot d\mathbf{l} = \frac{q}{T_c} \int_S \nabla \times \mathbf{E} \cdot d\mathbf{S} = -\frac{q}{T_c} \int_S \frac{\partial}{\partial t} \mathbf{B} \cdot d\mathbf{S}. \quad (2.28)$$

Here, we have first used Stokes' theorem to convert the line integral to an integral over the area S encircled by the orbit, and then applied Faraday's law.

From Figure 2.4, we see that with $d\mathbf{l} = \mathbf{v} dt$, the surface element $d\mathbf{s} = \mathbf{r}_c \times d\mathbf{l}$ will be directed opposite to $\mathbf{B}$. Assuming that $\frac{\partial}{\partial t} B$ is constant on the surface area $S = \pi r_c^2$ we find

$$\frac{\partial}{\partial t} \left(\frac{mv_\perp^2}{2} \right) \approx \frac{q}{T_c} \pi r_c^2 \frac{\partial}{\partial t} B = \mu \frac{\partial}{\partial t} B. \quad (2.29)$$

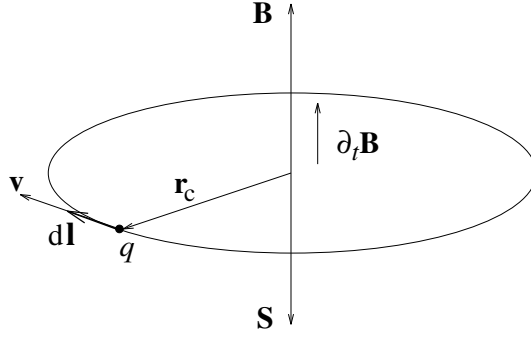


Figure 2.4.: A particle in a time dependent magnetic field. Figure from Fälthammar [2001].

Substituting this result into Equation 2.27 we see that μ is therefore constant even if the magnetic field varies with time. This conservation of magnetic moment has important consequences for the dynamics of the magnetospheric plasma.

2.5. Adiabatic invariants

The concept of adiabatic invariants is extremely powerful in describing the motion of particles in magnetic fields. These invariants are not totally constant such as the energy or momentum. But they vary very slowly compared to some other typical timescales of the charged particle. Therefore, they can be assumed to be constants over these smaller timescales.

More mathematically this can be described as follows. Consider the motion of a particle described by a pair of variables (p_i, q_i) that are generalized momenta and coordinates. Then, for each coordinate q_i that is periodic, the action integral J_i

$$J_i = \oint p_i dq_i, \quad (2.30)$$

integrated over a complete period of oscillation of q_i is an invariant (constant) of motion. This action integral will remain invariant even if some property of the system is allowed to change. However, it is required that the change be slow (adiabatic change) as compared to the period of q_i and the change must not be related to the period. For particles in a dipole magnetic field, three distinct periodic motions are possible, namely gyration around a field line, magnetic “mirroring” or bouncing along magnetic field lines, and a drift perpendicular to the field. An adiabatic invariant is associated with each of these periodic motions.

The first adiabatic invariant is associated with the cyclotron motion of the particle. Choosing $p_i = mv_\perp$ and $dq_i = r_c \omega_c dt$, the integral corresponding to Equation 2.30 is

$$J_1 = \int_0^{T_c} mv_\perp r_c \omega_c dt = 2\pi mv_\perp r_c = 2\pi \frac{mv_\perp^2}{qB/m} \propto \mu. \quad (2.31)$$

Hence, the first adiabatic invariant is proportional to the magnetic moment, which we already have shown to be invariant.

2. Charged Particles in an Electromagnetic Field

The second adiabatic invariant is given by the integral

$$\oint mv_{\parallel} ds, \quad (2.32)$$

and the third invariant is related to conservation of the magnetic flux enclosed by the drift orbit (c.f., Figure 2.4). These invariants are generally more complicated to evaluate analytically, and more easily violated by fast changes in the magnetospheric conditions, and not quite as useful as the magnetic moment.

2.5.1. Pitch angles and the loss cone

Let us introduce the pitch angle α of a particle as the angle between the velocity and the magnetic field, i.e.

$$v_{\perp} = v \sin \alpha \quad (2.33a)$$

$$v_{\parallel} = v \cos \alpha. \quad (2.33b)$$

Conservation of the magnetic moment can then be expressed as

$$\mu = \frac{mv^2}{2} \frac{\sin^2 \alpha}{B} = \text{constant}. \quad (2.34)$$

When the forces felt by the particle are purely magnetic, the kinetic energy will be constant, and hence the ratio $\sin^2 \alpha/B$ is also constant. If a particle which has pitch angle α_1 at the point $\mathbf{r}_1$ moves in a magnetic field to another point $\mathbf{r}_2$, it will there have a pitch angle α_2 that can be calculated from

$$\frac{\sin^2 \alpha_1}{B(\mathbf{r}_1)} = \frac{\sin^2 \alpha_2}{B(\mathbf{r}_2)}. \quad (2.35)$$

2.6. Conclusion

In this section we have considered the motion of a charged particle in an electromagnetic field. Charged particles have a gyro motion in a constant magnetic field. A gradient in the magnetic field or an electric field causes the particles to drift. Furthermore, adiabatic invariants, such as the magnetic moment, can be defined which can be assumed to be constant if there are only slow changes in the system.

In section 6.3 we will discuss how charged particles move in the Earth magnetic field.

3. Plasma Physics

3.1. Basic plasma behaviour

In the previous chapter we considered the motion of single charged particles in prescribed electric and magnetic fields. We did not consider that these fields may be due, in whole or in part, to other nearby charged particles. In the following chapter we will investigate the collective motion of many charged particles, and the resulting phenomena, including their mutual interactions. This is the beginning of *plasma physics*. In principle, the complete behaviour of a plasma could be defined by knowing the location and velocity of all particles. With this, one could calculate all forces between all particles, and their resulting motion. However, this is generally very impractical due to the large number of particles present in any relevant volume of space. There are two common simplifying approaches employed in plasma physics; *fluid* models or *kinetic* models. Fluid models describe the plasma solely in terms of its macroscopic (“bulk”) quantities, treating the plasma as an overall charge neutral fluid, that nevertheless responds to the electromagnetic field. Kinetic models, on the other hand, describe the plasma in terms of a particle velocity distribution at each point in space. All in all, there are three possible ways to describe a plasma: the single particle description, the fluid description, or the kinetic description.

The fluid description is a powerful approach to the study of large scale plasma phenomena, such as the flow of the solar wind and the shaping of the magnetosphere. However, the fluid description is itself obtained from the kinetic plasma theory. This is a statistical treatment of the assembly of plasma particles, from which macroscopic variables such as densities, currents, and temperatures can be derived by averaging over the particle distribution.

In the following these two approaches of describing a plasma will be introduced in greater detail. We will first discuss the kinetic plasma theory and then the fluid approach.

3.2. Definition of a plasma

Before diving into the theoretical descriptions of a plasma we need to define what a plasma actually is. Plasma is often referred to as the fourth state of matter after solids, liquids, and gases. For a gas to make a transition to the plasma state, it is not simply enough that it contains some proportion of free electrical charge carriers (typically ions and electrons). There must instead be a sufficient number of these free charges that they start to control the behaviour of the whole medium, through the collective interactions between their own electromagnetic fields. The distances over which a given charge particle in a plasma can “see” others is much larger than that relevant for simple collisions between them, such as dominates in a neutral gas at high temperatures. This also allows a wide variety of waves to propagate through a plasma, as explored in the next chapter.

3. Plasma Physics

One important measure is the Debye length which describes over what distances a test charge is shielded by the rearrangement of the surrounding plasma. The length scale of interest should always be larger than the Debye length so we can apply the methods defined in this chapter. The Debye length, it will be shown, is fundamental in that it also defines the length scale within which thermal fluctuations in the net charge density can be expected. When considering volumes much larger than a sphere of diameter equivalent to a Debye length, the plasma will be “quasi-neutral”, meaning there will be an equivalent number of positive and negative charges on these large scales.

We again consider a homogeneous plasma with density n_0 , but we now take into account that the particles are moving around randomly, with average, thermal, velocities v_{te} and v_{ti} for electrons and ions respectively. In space physics, temperatures are often expressed in energy units, where the electron Volt $1 \text{ eV} = 1.602 \cdot 10^{-19} \text{ J}$ is a natural choice of scale, equivalent to a temperature of $11\,605 \text{ K}$ (recalling that in SI units $1 \text{ eV} = k_B T$). Assuming that electrons and ions have the same temperature T (in eV), we then have $T = m_e v_{te}^2 / 2 = m_i v_{ti}^2 / 2$. A test charge $Q > 0$ would in vacuum be surrounded by a potential $\phi_v(r) = Q(4\pi\epsilon_0 r)^{-1}$, but in the plasma the electrons that are attracted and the ions that are repelled from the test charge will modify the potential. Notice that the test charge may well be a randomly chosen plasma ion. To calculate the self consistent potential $\phi(r)$, including all the plasma particles, we start from Boltzmann’s equation for the density of particles with a given temperature in a potential $\phi(\mathbf{r})$

$$n(\mathbf{r}) = n_0 \exp\left(-\frac{q\phi(\mathbf{r})}{T}\right), \quad (3.1)$$

where $n = n_i$ or n_e and $q = \pm e$. This equation can be justified from the equation of motion if the force from the pressure $P = nT$ (n in eV) is included as

$$nm \frac{d}{dt} \mathbf{v} = nq\mathbf{E} - \nabla P. \quad (3.2)$$

In the stationary equilibrium, $\frac{d}{dt} \mathbf{v} = 0$ and the temperature is constant. Using $\nabla P = T\nabla n$ and letting $\mathbf{E} = -\nabla\phi$ we find

$$T\nabla n + nq\nabla\phi = 0, \quad (3.3)$$

or

$$\frac{1}{n} \nabla n = \nabla \ln n = -\nabla \frac{q\phi}{T}, \quad (3.4)$$

which can be integrated to give Equation 3.1. Applying Boltzmann’s equation to the ions and electrons in the plasma we have

$$n_i(r) = n_0 \exp\left(-\frac{e\phi(r)}{T}\right) \quad (3.5a)$$

$$n_e(r) = n_0 \exp\left(\frac{e\phi(r)}{T}\right). \quad (3.5b)$$

From these equations we can calculate the charge density in the plasma as $\rho(r) = e(n_i - n_e) = n_0 e [\exp(-e\phi/T) - \exp(e\phi/T)]$. Assuming that the perturbation due to the test charge is small in the sense that $e\phi \ll T$, we use a Taylor expansion to simplify the charge density as

$$\rho(r) = n_0 e [1 - e\phi/T + \dots - 1 - e\phi/T - \dots] \approx -2 \frac{n_0 e^2}{T} \phi. \quad (3.6)$$

The potential is determined by inserting this charge density in Poisson's equation

$$\nabla^2 \phi = -\frac{\rho}{\epsilon_0} = 2\frac{n_0 e^2}{\epsilon_0 T} \phi. \quad (3.7)$$

We now introduce the *Debye length*

$$\lambda_D = \sqrt{\frac{\epsilon_0 T}{n_0 e^2}}, \quad (3.8)$$

and also spherical coordinates in which $\nabla^2 = r^{-1} \nabla^2 r$. Poisson's equation is then rewritten as

$$\frac{1}{r} \nabla^2 r \phi(r) = \frac{2}{\lambda_D^2} \phi(r), \quad (3.9)$$

which can be integrated to give $r\phi = \text{const.} \exp(-\sqrt{2} r/\lambda_D)$. The integration constant is determined from the condition that the potential at short distances, $r \ll \lambda_D$, should be the same as in vacuum. The potential of a charge Q in a plasma is thus

$$\phi(r) = \frac{Q}{4\pi\epsilon_0 r} \exp\left(-\sqrt{2} \frac{r}{\lambda_D}\right). \quad (3.10)$$

At distances beyond λ_D , the potential from a charged particle in a plasma falls off exponentially. The test charge is surrounded by a cloud of particles with the opposite charge, which shields the test charge from the rest of the plasma. However, the test charge is not different from any other charged particle, and each particle in the plasma is shielded in the same way. The interaction between individual particles in a plasma is thus limited to distances shorter than the Debye length.

In a cold and dense plasma the Debye length is much shorter than in a hot and tenuous plasma. From Equation 3.8 we see that the Debye length may be expressed as

$$\lambda_D = \sqrt{\frac{\epsilon_0 m_e v_{te}^2}{2n_0 e^2}} = \frac{v_{te}}{\sqrt{2}\omega_{pe}}, \quad (3.11)$$

using the thermal velocity defined with

$$E = k_B T = \frac{m v_{te}^2}{2}. \quad (3.12)$$

Hence, λ_D is roughly the distance a thermal electron moves during one plasma oscillation. This is consistent with our interpretation of ω_{pe}^{-1} as the time scale on which the plasma responds to a perturbing electric field. More on this later.

In general, plasma physics concerns phenomena with a characteristic length scale $\gg \lambda_D$. Looking at distances shorter than λ_D (or volumes $\lesssim \lambda_D^3$), parameters such as the density n and the electric field $\mathbf{E}$ will be fluctuating violently. These microscopic fluctuations are normally not included in theories describing plasma physics. The fields ($n(\mathbf{r})$, $\mathbf{E}(\mathbf{r})$, ...) we consider in plasma physics are the average fields in a volume $\gtrsim \lambda_D^3$ around $\mathbf{r}$. If this volume contains many particles, the fluctuations will be small and the fields will be nice and smooth. Hence, a condition for the applicability of plasma physics theories is that there should be many particles in a Debye sphere. The average number of particles in a Debye sphere is the Debye number defined as

$$N_D = \frac{4}{3} \pi \lambda_D^3 n. \quad (3.13)$$

Only if $N_D \gg 1$ we can apply the approximations we make in this Chapter to describe the collection of charged particles.

3. Plasma Physics

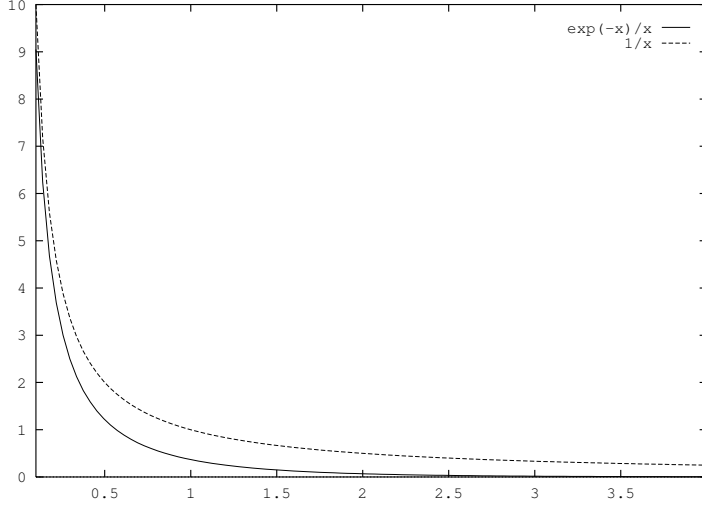


Figure 3.1.: The full line shows the potential of a charge in a plasma, and the dashed line the vacuum potential.

3.3. Kinetic plasma theory

The kinetic plasma theory is based on the velocity distribution function. Macroscopic quantities such as density and velocity can be derived from taking moments of the velocity distribution function.

3.3.1. The distribution function

We start by defining a *particle distribution function* $f(\mathbf{r}, \mathbf{v}, t)$ so that the number N of particles in a volume element $d\mathbf{r}$ around $\mathbf{r}$ in space and in $d\mathbf{v}$ around $\mathbf{v}$ in velocity space at time t is

$$N = f(\mathbf{r}, \mathbf{v}, t) d\mathbf{r} d\mathbf{v}. \quad (3.14)$$

The distribution function has dimension $\text{m}^{-3}(\text{m/s})^{-3}$, and gives the density of particles in the six-dimensional phase space spanned by $(\mathbf{r}, \mathbf{v})$. We consider $f(\mathbf{r}, \mathbf{v}, t)$ to be a smooth, continuous, function, which means that even if $d\mathbf{r} d\mathbf{v}$ is infinitesimal from a macroscopic point of view, it still contains many particles. We need one distribution function for each species of particles—the electrons are described by an electron distribution and the ions by one or several ion distributions. The *particle density* $n(\mathbf{r}, t)$, in particles/ m^3 , is obtained by integrating the distribution function over velocities

$$n(\mathbf{r}, t) = \int f(\mathbf{r}, \mathbf{v}, t) d\mathbf{v}. \quad (3.15)$$

We can also define an average particle velocity $\mathbf{u}$ (sometimes also called bulk velocity) as

$$\mathbf{u}(\mathbf{r}, t) = \frac{\int \mathbf{v} f(\mathbf{r}, \mathbf{v}, t) d\mathbf{v}}{\int f(\mathbf{r}, \mathbf{v}, t) d\mathbf{v}}. \quad (3.16)$$

In a plasma consisting of electrons with a distribution f_e and a single species of ions with distribution f_i , these equations give an electron density n_e and velocity $\mathbf{u}_e$ as well as ion density n_i and velocity $\mathbf{u}_i$. We can then, in an obvious way, introduce the mass

density $\rho_m = m_i n_i + m_e n_e$, the momentum density $\mathbf{p} = m_i n_i \mathbf{u}_i + m_e n_e \mathbf{u}_e$, the charge density $\rho_q = en_i - en_e$, and current density $\mathbf{j} = en_i \mathbf{u}_i - en_e \mathbf{u}_e$. These quantities can be used in a macroscopic, fluid-like description of the plasma.

Thermalisation, i.e., equipartition of energy by collisions, will eventually cause any collection of particles to move towards a Maxwellian distribution

$$f_M(\mathbf{r}, \mathbf{v}, t) = n \left(\frac{m}{\pi 2T} \right)^{3/2} \exp \left[-\frac{v^2}{v_t^2} \right], \quad (3.17)$$

with temperature $T = mv_t^2/2$. This is the only distribution for which a temperature is properly defined, but we shall often use effective temperatures and thermal velocities also for other, non-Maxwellian distributions. We will then use the average random kinetic energy and define the temperature in energy units by

$$\frac{3}{2}T = \frac{\int \frac{1}{2}m(v^2 - u^2)f(\mathbf{r}, \mathbf{v}, t) d\mathbf{v}}{\int f(\mathbf{r}, \mathbf{v}, t) d\mathbf{v}}, \quad (3.18)$$

which is the true temperature when the distribution is Maxwellian. The temperature thus defined is related to the trace (sum of diagonal elements) of the *pressure tensor*

$$\mathbf{P} = \int m(\mathbf{v}\mathbf{v} - \mathbf{u}\mathbf{u})f(\mathbf{r}, \mathbf{v}, t) d\mathbf{v}, \quad (3.19)$$

divided by the density n . When written out in Cartesian coordinates, the tensor $\mathbf{v}\mathbf{v}$ has elements such as $v_x v_x, v_x v_y, v_x v_z, \dots$, so the trace will be an average over $m(v_x^2 + v_y^2 + v_z^2)$, and $nT = 1/3 \text{ trace } \mathbf{P}$.

Equations 3.15, 3.16, and 3.19 are often referred to as the zeroth, first, and second *moment* of the distribution function.

3.3.2. Reminder: Total/Convective derivatives

The time variation of a fluid property, such as the density n , may be given in two different ways. First, we may place ourselves at a fixed point and observe the variation of the fluid as it flows past us. The *explicit* time variation we then observe corresponds to the partial derivative $\frac{\partial}{\partial t}n(\mathbf{r}, t)$, where ∇ is the fixed point we have chosen. Secondly, we may observe the density variations as we allow ourselves to flow along with the fluid, and obtain the total, or *convective derivative*

$$\frac{d}{dt}n(\mathbf{r}, t) = \frac{\partial}{\partial t}n(\mathbf{r}, t) + \mathbf{u} \cdot \nabla n(\mathbf{r}, t). \quad (3.20)$$

Essentially this is just the use of the chain rules for derivatives. The reason for the second term is that the point of observation (r) moves with the fluid velocity $\frac{d}{dt}\mathbf{r}(t) = \mathbf{u}$, and according to the usual chain rule for derivatives we have $\frac{d}{dt}n(\mathbf{r}(t)) = \nabla n(\mathbf{r}) \cdot \frac{d}{dt}\mathbf{r}(t)$.

3.3.3. The Vlasov equation

Let us consider a *collisionless plasma*. Here, the term collisionless means that the fluctuations in the electromagnetic field on scales less than the Debye length are assumed to have negligible effect on the plasma behaviour, compared to the large scale fields. This means that in the following we consider $\mathbf{E}$ and $\mathbf{B}$ to be the average fields in a

3. Plasma Physics

volume larger than the Debye sphere, and the interaction between individual particles is ignored. With these assumptions, we can calculate the rate of change of the number N of particles in a volume $\Delta\mathbf{r}\Delta\mathbf{v}$ in phase space. We do this calculation to find the equation that describes the time evolution of the velocity distribution function. If S_r is the surface of $\Delta\mathbf{r}$, the flux of electrons out of $\Delta\mathbf{r}$ in r space is

$$\int_{\Delta\mathbf{v}} d\mathbf{v} \int_{S_r} f \frac{d}{dt} \mathbf{r} \cdot d\mathbf{S}_r = \int_{\Delta\mathbf{v}} d\mathbf{v} \int_{S_r} f \mathbf{v} \cdot d\mathbf{S}_r, \quad (3.21)$$

and in $\mathbf{v}$ space the flux is

$$\int_{\Delta\mathbf{r}} d\mathbf{r} \int_{S_v} f \frac{d}{dt} \mathbf{v} \cdot d\mathbf{S}_v = \int_{\Delta\mathbf{r}} d\mathbf{r} \int_{S_v} f \mathbf{a} \cdot d\mathbf{S}_v, \quad (3.22)$$

where $\mathbf{a}$ is the acceleration and S_v is the surface of $\Delta\mathbf{v}$. Assuming conservation of particles, the change of N is entirely due to the flux of particles and

$$\frac{\partial}{\partial t} N = \int_{\Delta\mathbf{r}\Delta\mathbf{v}} \frac{\partial}{\partial t} f d\mathbf{r} d\mathbf{v} = - \int_{\Delta\mathbf{v}} d\mathbf{v} \int_{S_r} f \mathbf{v} \cdot d\mathbf{S}_r - \int_{\Delta\mathbf{r}} d\mathbf{r} \int_{S_v} f \mathbf{a} \cdot d\mathbf{S}_v, \quad (3.23)$$

Using Gauss' theorem, we can convert the surface integrals to volume integrals as

$$\frac{\partial}{\partial t} N = \int_{\Delta\mathbf{r}\Delta\mathbf{v}} \frac{\partial}{\partial t} f d\mathbf{r} d\mathbf{v} = - \int_{\Delta\mathbf{r}\Delta\mathbf{v}} \nabla \cdot (\mathbf{v}f) d\mathbf{r} d\mathbf{v} - \int_{\Delta\mathbf{r}\Delta\mathbf{v}} \frac{\partial}{\partial \mathbf{v}} \cdot (\mathbf{a}f) d\mathbf{r} d\mathbf{v}, \quad (3.24)$$

Since this must be true for all choices of $\Delta\mathbf{r}\Delta\mathbf{v}$, we have

$$\frac{\partial}{\partial t} f + \nabla \cdot (\mathbf{v}f) + \frac{\partial}{\partial \mathbf{v}} \cdot (\mathbf{a}f) = 0. \quad (3.25)$$

When the acceleration is given by the Lorentz force, $\mathbf{a} = q/m(\mathbf{E} + \mathbf{v} \times \mathbf{B})$, it can be shown that the last term can be written as $q/m(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \frac{\partial}{\partial \mathbf{v}} f$, and recalling that $\mathbf{r}$ and $\mathbf{v}$ are independent variables we can write the second term as $\mathbf{v} \cdot \nabla f$. We then obtain the fundamental equation of kinetic plasma theory, the *Vlasov equation*

$$\frac{\partial}{\partial t} f + \mathbf{v} \cdot \nabla f + \frac{q}{m}(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \frac{\partial}{\partial \mathbf{v}} f = 0. \quad (3.26)$$

This equation can also be written $\frac{d}{dt} f(\mathbf{r}, \mathbf{v}, t) = 0$, which states that the convective derivative of f in phase space is zero. The equation describes the time evolution of the particle distribution function. Another interpretation of this is that f is constant along the particle orbits $(\mathbf{r}(t), \mathbf{v}(t))$ determined by the large scale electromagnetic fields. Turning this around, we may also say that the particles move on surfaces of constant f in phase space.

In general, a Vlasov equation is needed for the distribution of each particle species in the plasma. These equations are then coupled by the electromagnetic field, which of course is the same for all species.

3.4. Magnetohydrodynamic theory

Magnetohydrodynamic theory describes the plasma as a conducting fluid interacting with the electromagnetic field. Either one fluid equation is used or several fluids representing different species. Here, we will first indicate how fluid equations for a one component plasma can be derived from kinetic theory. Then the magnetohydrodynamics (MHD) equations will be presented, and applied to some phenomena in space physics.

3.4.1. Fluid equations

We have in subsection 3.3.1 seen that quantities like fluid density and velocity can be defined as moments of the distribution function. We can therefore expect to obtain equations relating such macroscopic variables by taking moments of the Vlasov equation. The zeroth moment is simply the integral over velocity space of the Vlasov equation.

$$\int \frac{\partial}{\partial t} f d\mathbf{v} + \int \mathbf{v} \cdot \nabla f d\mathbf{v} + \int \frac{q}{m} (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot \frac{\partial}{\partial \mathbf{v}} f d\mathbf{v} = 0. \quad (3.27)$$

In the two first terms, we reverse the order of integration and differentiation, which is allowed since v is independent of t and r . In the third term we integrate partially, noticing that $f(\mathbf{r}, \mathbf{v}, t) \rightarrow 0$ as $|\mathbf{v}| \rightarrow \infty$, and obtain

$$\frac{\partial}{\partial t} \int f d\mathbf{v} + \nabla \cdot \int \mathbf{v} f d\mathbf{v} - \int f \frac{\partial}{\partial v} \cdot \frac{q}{m} (\mathbf{E} + \mathbf{v} \times \mathbf{B}) d\mathbf{v} = 0. \quad (3.28)$$

It is easily shown that $\frac{\partial}{\partial \mathbf{v}} \cdot \mathbf{v} \times \mathbf{B} = 0$, and hence the third term is zero. Using the definitions 3.15 and 3.16, we then find the *continuity equation*

$$\frac{\partial}{\partial t} n + \nabla \cdot (n\mathbf{u}) = 0, \quad (3.29)$$

which expresses the conservation of particles. It is only a matter of multiplying with m or q to convert this into expressing conservation of mass or charge.

Multiplying the Vlasov equation with v and integrating, we find after similar but more complicated calculations a *fluid equation of motion* for the plasma

$$mn \frac{d}{dt} \mathbf{u} = nq(\mathbf{E} + \mathbf{u} \times \mathbf{B}) - \nabla \cdot \mathbf{P}, \quad (3.30)$$

Equation 3.29 and 3.30 illustrate a problem with this procedure. The zeroth moment of the Vlasov equation contains the first velocity moment of f , which is $\mathbf{u}$. To determine $\mathbf{u}$, we need the first moment of the Vlasov equation, that is Equation 3.30, which in turn contains $\mathbf{P}$, the second moment of f . Clearly, this sequence will have to be terminated by some justifiable procedure. One possibility is to ignore viscosity, assume the pressure to be isotropic so that $\mathbf{P} = P\mathbf{I}$, and introduce an *equation of state*

$$Pn^{-\gamma} = \text{constant}. \quad (3.31)$$

For adiabatic processes, γ is the ratio C_P/C_V of specific heats at constant pressure and volume, and for isothermal processes $\gamma = 1$.

3.4.2. Magnetohydrodynamics - single fluid equations

We have above indicated how a set of fluid equations for each plasma component can be derived. A plasma consisting of electrons and ions with equal densities is then described as two fluids. However, these fluids are coupled together by electromagnetic forces, and it turns out that we can describe the plasma as a single fluid. The difference between the electron and ion flows will then show up as a current density. Since the plasma is a good conductor, any large charge density in the plasma will rapidly be neutralized. Usually, it is safe to assume that the plasma is *quasineutral*, which means that ion and electron densities are almost equal, $n_i \approx n_e \approx n$.

3. Plasma Physics

We first define the appropriate parameters for a neutral plasma:

$$\text{mass density } \rho_m = n_i m_i + n_e m_e \approx n m_i \quad (3.32a)$$

$$\text{mass velocity } \mathbf{U} = \frac{1}{\rho_m} (n_i m_i \mathbf{u}_i + n_e m_e \mathbf{u}_e) \approx \mathbf{u}_i \quad (3.32b)$$

$$\text{current density } \mathbf{j} = e(n_i \mathbf{u}_i - n_e \mathbf{u}_e) \approx n e (\mathbf{u}_i - \mathbf{u}_e). \quad (3.32c)$$

Adding the continuity Equation 3.29 for electrons and ions, multiplied by the appropriate mass, we obtain a single continuity equation

$$\frac{\partial}{\partial t} \rho_m + \nabla \cdot (\rho_m \mathbf{U}) = 0, \quad (3.33)$$

which tells us that the mass of the plasma is conserved.

From Equation 3.30 we can write the equations of motion as

$$m_i n \frac{d}{dt} \mathbf{u}_i = n e (\mathbf{E} + \mathbf{u}_i \times \mathbf{B}) - \nabla P_i \quad (3.34a)$$

$$m_e n \frac{d}{dt} \mathbf{u}_e = -n e (\mathbf{E} + \mathbf{u}_e \times \mathbf{B}) - \nabla P_e, \quad (3.34b)$$

where we have used $n_i \approx n_e \approx n$ and assumed a scalar pressure. Adding these equations and writing the total pressure $P = P_i + P_e$, we find

$$\rho_m \frac{d}{dt} \mathbf{u} = \mathbf{j} \times \mathbf{B} - \nabla P, \quad (3.35)$$

which is called the *momentum equation* in MHD theory.

3.4.3. Ohm's law

We can also derive a *generalized Ohm's law* for a plasma by subtracting the equations of motion. Since the transition between a collision dominated material and a collisionless plasma can be rather tricky, we will here make it explicit by including a collision frequency ν_c that produces some friction between electrons and ions. For convenience, we also divide by the mass and rewrite Equation 3.34a and Equation 3.34b as

$$n \frac{d}{dt} \mathbf{u}_i = \frac{n e}{m_i} (\mathbf{E} + \mathbf{u}_i \times \mathbf{B}) - \nabla P_i / m_i - \frac{m_e}{m_i} \nu_c n (\mathbf{u}_i - \mathbf{u}_e) \quad (3.36a)$$

$$n \frac{d}{dt} \mathbf{u}_e = -\frac{n e}{m_e} (\mathbf{E} + \mathbf{u}_e \times \mathbf{B}) - \nabla P_e / m_e + \nu_c n (\mathbf{u}_i - \mathbf{u}_e). \quad (3.36b)$$

When subtracting these equations, we make some further simplifications. We neglect terms that are small by the factor m_e/m_i , and replace $\frac{d}{dt}$ by $\frac{\partial}{\partial t}$ assuming that the squares of the velocities will be small. We also use that $n e \mathbf{u}_e \approx n e \mathbf{U} - \mathbf{j}$ to find

$$\frac{\partial}{\partial t} \mathbf{J} = \epsilon_0 \omega_{pe}^2 (\mathbf{E} + \mathbf{U} \times \mathbf{B}) - \frac{e}{m_e} (\mathbf{J} \times \mathbf{B} - \nabla P_e) - \nu_c \mathbf{J}. \quad (3.37)$$

This equation is the generalized Ohm's law and can be used to obtain the electric field. In a familiar laboratory conductor such as a metal wire, this equation is dominated by the first and last terms on the right hand side. The relation between the current density and the electric field is then given by the so called ideal Ohm's law

$$\mathbf{j} = \sigma \mathbf{E}. \quad (3.38)$$

Here, the conductivity $\sigma = \epsilon_0 \omega_{pe} / u_c$ is a constant characterizing the material. In a collisionless plasma where $u_c \rightarrow 0$, the last term of Equation 3.37 is negligible. When time variations are slow and the plasma flow rather homogeneous, both $\frac{\partial}{\partial t} \mathbf{j}$ and $\mathbf{j} \times \mathbf{B} - \nabla P_e$ will be small. It turns out that in many cases Ohm's law for space plasmas can be written

$$\mathbf{E} + \mathbf{U} \times \mathbf{B} = 0. \quad (3.39)$$

The theory resulting from this approximation is known as *ideal* MHD, and is often applicable to the description of large scale plasma phenomena in space. Equation 3.39 is also known as the *ideal* Ohm's law. However, in places where the acceleration of the plasma flow is strong (cf. Equation 3.60), the ideal MHD approximation may become invalid.

In the laboratory, we usually consider the current to be related to the electric field by a relation such as Equation 3.38. In a collisionless, magnetized plasma, on the other hand, there is no simple relation between the current and the electric field. Currents are related to gradients in the pressure and flow velocity, and to the magnetic field by Ampère's law, while the electric field is simply $\mathbf{E} = -\mathbf{U} \times \mathbf{B}$. A perpendicular electric field will not cause a current but a plasma flow.

In addition to these equations and the equation of state Equation 3.31, we will need two of Maxwell's equations, namely Faraday's law

$$\nabla \times \mathbf{E} = -\frac{\partial}{\partial t} \mathbf{B}, \quad (3.40)$$

and Ampère's law

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j} + c^{-2} \frac{\partial}{\partial t} \mathbf{E}. \quad (3.41)$$

The displacement current $\frac{\partial}{\partial t} \mathbf{E}$ can often be neglected from Ampère's law, since the fluid equations usually are applied to slow processes.

Together, Equation 3.31, 3.33, 3.60, 3.39, 3.38, 3.40, and 3.41 constitute a set of fourteen scalar equations for the fourteen unknown n , P , $\mathbf{j}$, $\mathbf{U}$, $\mathbf{E}$, and $\mathbf{B}$. This closed system is a simple set of MHD equations. Slightly different MHD models are found if the equation of state 3.31 is replaced by some other method for closing the system, but these will not be treated here.

3.4.4. Frozen-in magnetic fields

In general, the concept of magnetic field lines is ambiguous if the magnetic field is time dependent. However, in a medium of infinite conductivity, such as an ideal MHD plasma, individual field lines can be uniquely defined even in non-stationary situations, as we shall show shortly. In ideal MHD, the motion of the plasma and the change in the $\mathbf{B}$ field are coupled so that the magnetic field lines appear to follow the plasma motion. This was first recognized by Hannes Alfvén, who described it as *frozen-in magnetic field lines*. His results can be summarized in the following two statements:

1. The flux of magnetic field through any closed contour moving with the plasma is constant.
2. Fluid elements that lie on a magnetic field line remain on the same field line.

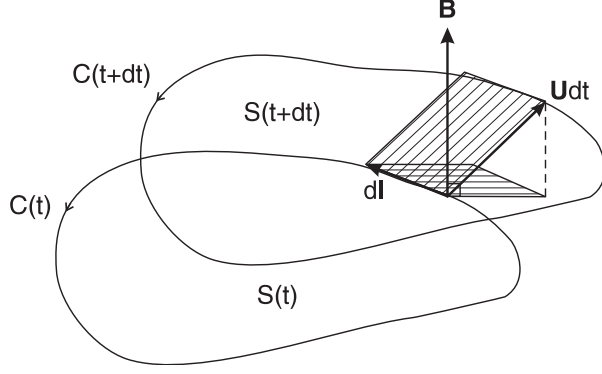


Figure 3.2.: Change in magnetic flux due to the motion of the surface S .

To prove these statements, we first notice that in ideal MHD, we can eliminate $\mathbf{E}$ from Faraday's law 3.40 by means of 3.39 to obtain

$$\frac{\partial}{\partial t} \mathbf{B} = \nabla \times (\mathbf{U} \times \mathbf{B}). \quad (3.42)$$

We now consider the change in the magnetic flux Φ through a surface S enclosed by a contour C . Each point on the surface moves with the plasma velocity $\mathbf{U}$, which may be different on different parts of the surface. Let $d\mathbf{l}$ be an infinitesimal vector along C , moving with velocity $\mathbf{U}$. The area swept over by $d\mathbf{l}$ in time dt is then $dt \mathbf{U} \times d\mathbf{v}$, as indicated in Figure 3.2. The change in Φ is partly caused by the change in the magnetic field strength, partly due to the motion of the surface perpendicular to the magnetic field, and can be written

$$\frac{d}{dt} \Phi = \int_S \frac{\partial}{\partial t} \mathbf{B} \cdot d\mathbf{S} + \int_C \mathbf{B} \cdot \mathbf{U} \times d\mathbf{l} \quad (3.43)$$

$$= \int_S \frac{\partial}{\partial t} \mathbf{B} \cdot d\mathbf{S} - \int_C \mathbf{U} \times \mathbf{B} \cdot d\mathbf{l} \quad (3.44)$$

$$= \int_S \left(\frac{\partial}{\partial t} \mathbf{B} - \nabla \times (\mathbf{U} \times \mathbf{B}) \right) \cdot d\mathbf{S}. \quad (3.45)$$

Here, we have first used the vector identity $\mathbf{B} \cdot \mathbf{U} \times d\mathbf{l} = -\mathbf{U} \times \mathbf{B} \cdot d\mathbf{l}$, and then Stoke's theorem to change from a contour integral to a surface integral. Comparing the integrand on the third line with Equation 3.42, we find $\frac{d}{dt} \Phi = 0$. Hence, Φ is constant, which proves statement 1 above.

A magnetic field line is composed of line elements $d\mathbf{r}$ which everywhere are parallel to the local magnetic field $\mathbf{B}$. Consider two neighboring fluid elements that at time t are located at $\mathbf{r}$ and $\mathbf{r} + d\mathbf{r}$ and initially on the same field line. The vector $d\mathbf{r}$ must then satisfy

$$\mathbf{B}(\mathbf{r}, t) \times d\mathbf{r} = 0. \quad (3.46)$$

After a short time dt , the plasma elements have moved so that their separation is

$$d\mathbf{r}' = d\mathbf{r} + (\mathbf{U}(\mathbf{r} + d\mathbf{r}, t) - \mathbf{U}(\mathbf{r}, t))dt, \quad (3.47)$$

and they will remain on the same field line if the magnetic field has changed so that

$$\mathbf{B}(\mathbf{r} + \mathbf{U}dt, t + dt) \times d\mathbf{r}' = 0 \quad (3.48)$$

To prove this, it is sufficient to carry out the calculations to first order in the infinitesimals dt $d\mathbf{r}$. Then, we first notice that the magnetic field can be written

$$\mathbf{B}(\mathbf{r} + \mathbf{U}dt, t + dt) = \mathbf{B}(\mathbf{r}, t) + \frac{\partial}{\partial t}\mathbf{B}dt + \mathbf{U} \cdot \nabla \mathbf{B}dt \equiv \mathbf{B}', \quad (3.49)$$

and similarly

$$d\mathbf{r}' = d\mathbf{r} + d\mathbf{r} \cdot \nabla \mathbf{U}dt. \quad (3.50)$$

Using these two relations and writing $\mathbf{B} = \mathbf{B}(\mathbf{r}, t)$ and $\mathbf{U} = \mathbf{U}(\mathbf{r}, t)$ for short, we find

$$\mathbf{B}' \times d\mathbf{r}' = (\mathbf{B} + \mathbf{U} \cdot \nabla \mathbf{B}dt + \frac{\partial}{\partial t}\mathbf{B}dt) \times (d\mathbf{r} + d\mathbf{r} \cdot \nabla \mathbf{U}dt) \quad (3.51)$$

$$\begin{aligned} &= \mathbf{B} \times (d\mathbf{r} \cdot \nabla) \mathbf{U}dt + (\mathbf{U} \cdot \nabla \mathbf{B}) \times d\mathbf{r}dt + \frac{\partial}{\partial t}\mathbf{B} \times d\mathbf{r}dt \\ &= \left[-(\mathbf{B} \cdot \nabla) \mathbf{U} + (\mathbf{U} \cdot \nabla) \mathbf{B} + \frac{\partial}{\partial t}\mathbf{B} \right] \times d\mathbf{r}dt \\ &= \left[\frac{\partial}{\partial t}\mathbf{B} - \nabla \times (\mathbf{U} \times \mathbf{B}) \right] \times d\mathbf{r}dt. \end{aligned} \quad (3.52)$$

To get from the second line to the third, we have used that $d\mathbf{r}$ is parallel to $\mathbf{B}$, and the calculations leading to the last line are left to the exercises, where some hints are given. Comparing the last line of Equation 3.51 with Equation 3.42, we find $\mathbf{B}' \times d\mathbf{r}' = 0$. This proves that two plasma elements that initially were on the same field line will remain on the same field line.

These results show that the magnetic field lines will be frozen in and move with the perfectly conducting plasma. However, when interpreting this result we must be careful and remember that a magnetic field line is an abstract concept, not a physical entity. Notice, in particular, that the magnetic field $\mathbf{B}$ is never moving!

3.4.5. Magnetic pressure

When we consider slow processes or stationary situations, we can often neglect the displacement current $\epsilon_0 \frac{\partial}{\partial t}\mathbf{E}$ in Ampère's law 3.41, which then becomes

$$\mathbf{j} = \frac{1}{\mu_0}(\nabla \times \mathbf{B}). \quad (3.53)$$

Inserting this in the momentum Equation 3.60 we find

$$\rho_m \frac{d}{dt}\mathbf{U} = \frac{1}{\mu_0}(\nabla \times \mathbf{B}) \times \mathbf{B} - \nabla P \quad (3.54)$$

$$= \frac{1}{\mu_0}(\mathbf{B} \cdot \nabla) \mathbf{B} - \nabla \left(P + \frac{B^2}{2\mu_0} \right). \quad (3.55)$$

Here, we see that $B^2/2\mu_0$ appears as a *magnetic pressure*, similar to the kinetic pressure P . Often, the plasma β (beta) parameter defined as

$$\beta = \frac{P}{B^2/2\mu_0} = \frac{\text{particle pressure}}{\text{magnetic pressure}}, \quad (3.56)$$

is used to characterize a plasma. The β parameter measures the relative importance of particle and magnetic pressures. Roughly speaking, we may say that in a high-beta plasma, when $\beta \gtrsim 1$, the magnetic field lines will follow the motion of the plasma particles, but in a low-beta plasma ($\beta \ll 1$) the plasma particles will follow the motion of the field lines. In space, both high- and low-beta plasmas are encountered.

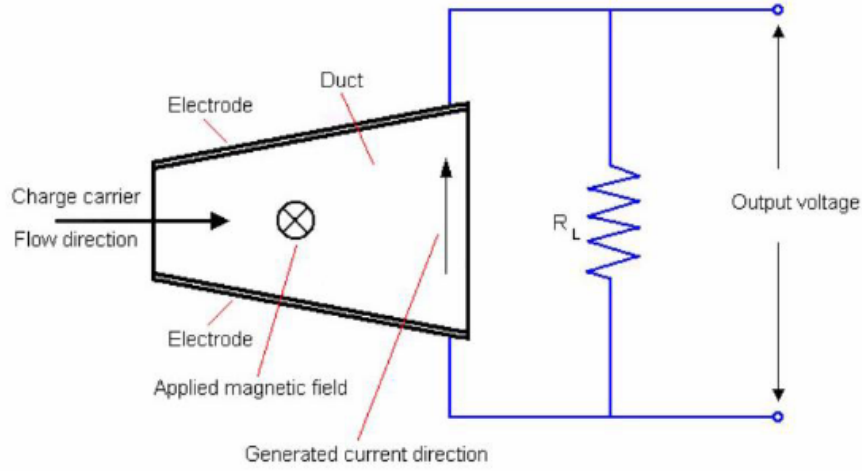


Figure 3.3.: Schematic of an MHD generator. Figure from Krishnan and Kalluvila [2013].

3.4.6. MHD generator

A generator is a device that converts kinetic energy into electromagnetic power. Such generators operate at several places in space. However, it is often difficult to identify the different parts of a generator in space, and hence we will first discuss the laboratory version of an MHD *generator*.

A sketch of an MHD generator is shown in Figure 3.3. To the left in the figure, a heat source produces a plasma and forces it to flow through the tube. An external magnet produces a magnetic field, going vertically into the tube. When the plasma is forced to flow across the magnetic field lines, the particles feel a force $q(\mathbf{U} \times \mathbf{B})$. The ions will then be forced upwards, and the electrons downwards. In the absence of load on the generator, this charge separation would create an electric field $\mathbf{E}$ from top to the down. Such that in equilibrium the forces balance, $\mathbf{E} + \mathbf{U} \times \mathbf{B} = 0$, and the bulk of the plasma can then flow freely. If we now connect a load to electrodes on the sides of the generator, the current through the load will reduce the charge on the sides of the generator tube. The $\mathbf{E}$ field in the plasma then becomes weaker, and can no longer balance the $\mathbf{U} \times \mathbf{B}$ force. The charges in the plasma will then be forced to move in the directions of the $q(\mathbf{U} \times \mathbf{B})$ force, trying to restore the charges on the tube walls. The other possibility is to achieve balance by reducing $\mathbf{U}$. In practice, we will have both a current through the load and a reduced flow velocity when we draw power from the generator.

Notice that inside a generator, the current flows in the direction opposite to the electric field, since the $q(\mathbf{U} \times \mathbf{B})$ force is larger than $q\mathbf{E}$. The electromagnetic power $\mathbf{E} \cdot \mathbf{J}$ is negative in a generator, and the produced electromagnetic energy comes from the kinetic energy $\rho_m U^2/2$ of the plasma flow driven by the heat source. If we increase the load (reduce the resistance), we need a larger pressure drop to maintain the flow velocity.

By connecting a current source to the electrodes, we can convert an MHD generator into a *plasma engine*. Consider a plasma initially at rest. Charge added to the electrodes by the external source will create an electric field. As long as $\mathbf{E}$ is larger than $\mathbf{U} \times \mathbf{B}$ the plasma will accelerate and a current will flow parallel to the $\mathbf{E}$ -field. The plasma is accelerated to the $\mathbf{U} \times \mathbf{B}/B^2$ velocity, and the kinetic energy comes from the power $\mathbf{E} \cdot \mathbf{j} > 0$ fed into the plasma from the external power supply.

Some further insights into the physics of MHD generators can be gained from the momentum Equation 3.60. We assume that the flow is stationary ($\frac{\partial}{\partial t}\mathbf{U} = 0$), and choose the x axis along the tube. Allowing for variations in the width W of the tube, and hence in the flow velocity $U(x)$, we find

$$\rho_m \mathbf{U} \cdot \nabla \mathbf{U} = \nabla \frac{\rho_m}{2} U^2 = \mathbf{j} \times \mathbf{B} - \nabla P. \quad (3.57)$$

Integrating over the length L of the generator we obtain

$$\int_0^L |\mathbf{j} \times \mathbf{B}| dx = \frac{\rho_m}{2} (U_I^2 - U_O^2) + P_I - P_O, \quad (3.58)$$

where U_I and P_I are the velocity and pressure at the inlet to the generator and U_O and P_O are measured at the outlet. The current I_L through the load is found by integrating over the height H of the generator. Noticing that the voltage $V_L = UWB$ will be constant if the flow is incompressible so that $UW \approx \text{constant}$, we find that the power produced by the generator and dissipated in the load is

$$P_L = V_L I_L \approx HWU \left(\frac{\rho_m}{2} (U_I^2 - U_O^2) + P_I - P_O \right). \quad (3.59)$$

For maximum efficiency, the generator should be designed to convert all the available heat power to electric power, which implies $U_O \approx P_O \approx 0$.

In the laboratory device discussed above, the boundaries are clearly defined and the load is clearly separated from the generator. In space, there are no walls that guide the flow, and no electrodes and wires that conduct the current from the generator to the load. However, some parts of an MHD generator can still be identified. The Sun is an obvious source of heat, producing the solar wind. At the magnetopause, the solar wind encounters a strong magnetic field perpendicular to the flow. Most of the solar wind flow will be diverted around the magnetosphere, but there will be a boundary layer that works as an MHD generator, driving a magnetopause current in the East-West direction. At least parts of this current will flow down along the dipolar magnetic field lines to the ionosphere on the evening side. Collisions in the ionosphere make the conductivity finite, and while the current flows across to the morning side the ionosphere acts as a load where power is dissipated. The circuit is then closed along the dipolar field lines on the morning side.

The current systems in the magnetosphere are in reality much more complicated, and much is still unknown about their structure and dynamical behaviour. As long as we consider currents generated directly by the solar wind dynamo, the causal relationship is at least clear—it is obvious that the solar wind flow drives the current. Inside the magnetosphere, we may have a self-consistent model including a current and a plasma flow, but it may not be obvious whether the flow is driving the current or the current is causing the flow.

3.4.7. Flux ropes

Let's consider a static force balance where the plasma is stationary and no pressure gradient force or gravitation is acting on the plasma ($u \approx 0$ and $\nabla P \approx 0$). Then Equation 3.57 simplifies to

$$\rho_m \frac{d}{dt} \mathbf{u} + \nabla P = \mathbf{j} \times \mathbf{B} = 0. \quad (3.60)$$

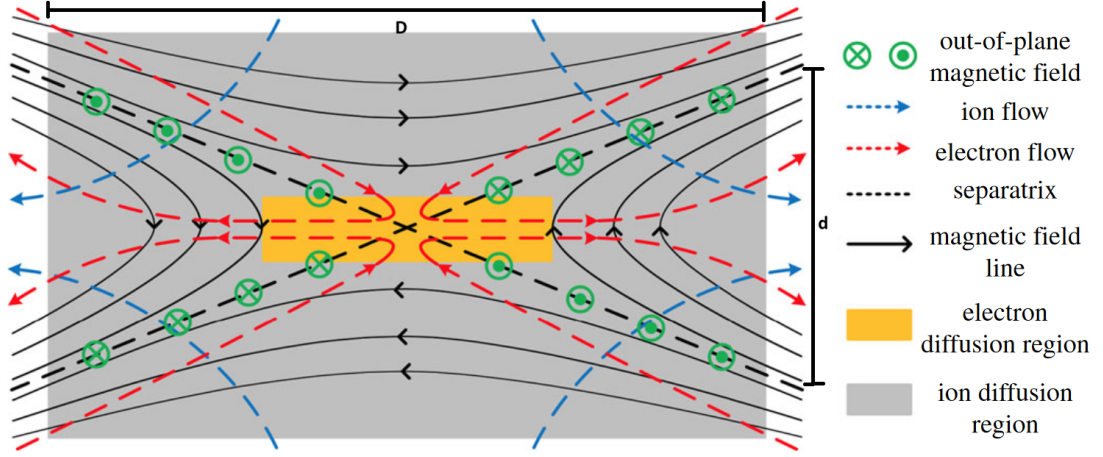


Figure 3.4.: Schematics of magnetic reconnection. Adapted from Zweibel and Yamada [2016].

This equation has the trivial solution $\mathbf{j} = 0$ which is not of special interest in space physics. Another solution is given when $\mathbf{j} \parallel \mathbf{B}$ which gives rise to a complex and force-free structure, a flux rope. According to Ampère's law, the magnetic field is twisted in a way that the current flows parallel to the magnetic field: $\mathbf{j} \times \mathbf{B} = (\nabla \times \mathbf{B}) \times \mathbf{B} = 0$.

3.4.8. Break down of MHD - magnetic reconnection

The fluid description of a plasma is powerful and is commonly used in space physics. However, there are regions with plasma where the fluid description is not appropriate to describe the plasma and the ideal MHD condition $\mathbf{E} + \mathbf{v} \times \mathbf{B} = 0$ cannot be applied anymore. Under such circumstances, Alfvén's *frozen-in* theorem does not hold anymore and the magnetic field can diffuse through the plasma. This process is known as magnetic reconnection where the magnetic topology is re-arranged. Ultimately, magnetic energy is converted into kinetic energy by accelerating plasma that flows away from the reconnection region.

Generally, magnetic reconnection occurs where the Ohmic term in the generalized Ohm's law (Equation 3.37) becomes relevant such that

$$\mathbf{E} = -\mathbf{u} \times \mathbf{B} + \eta \mathbf{J}, \quad (3.61)$$

with the resistivity η given by

$$\eta = \frac{\nu_e}{\epsilon_0 \omega_{pe}} = \frac{\nu_e m_e}{n_e e^2}. \quad (3.62)$$

In space plasma physics collisionless plasmas are often considered, so the collisions frequency $\nu_e = 0$, but inside the reconnection region we can assign an *anomalous resistivity* which stems from wave-particle interactions. These anomalous resistivity can be considered to be an effective resistivity. Combining the Ohm's law (Equation 3.61) with the Faraday's law which describes the time evolution of a magnetic field yields

$$\frac{\partial \mathbf{B}}{\partial t} = -\nabla \times \mathbf{E} = \underbrace{\nabla \times (\mathbf{u} \times \mathbf{B})}_{\text{magnetic convection}} - \underbrace{\nabla \times (D_B \nabla \times \mathbf{B})}_{\text{magnetic diffusion}}, \quad (3.63)$$

with the magnetic diffusion coefficient $D_B = \eta / \mu_0$.

Magnetic reconnection takes place in regions where the diffusion terms becomes important. The Reynolds number R_m can be used to determine which term dominates in a certain plasma region and is defined as

$$R_m = \frac{\text{magnetic convection term}}{\text{magnetic diffusion term}} \approx \frac{Lu}{D_B}, \quad (3.64)$$

with u the convection speed and L the length scale over which the magnetic field varies significantly (for example the thickness of the current sheet). If $R_m \gg 1$, then the convection term is dominant and the diffusion term negligible ($\eta = 0$). However, if $R_m \ll 1$, then anomalous resistivity becomes important. In space plasma we often have situations where L or u is small in a boundary region even though D_B is still large. This happens for example in thin current sheets which are only a few gyroradii thick.

Let's consider such a current sheet which forms at a boundary regions between two regions of a uniform plasma with opposite directed magnetic fields, see Figure 3.4. According to Ampère's law, a change in magnetic field is consistent a current to flow, a current sheet. In Figure 3.4 the current is flowing into the paper. In addition, there is a neutral line where the magnetic field is very weak. The ion diffusion region (gray region in Figure 3.4) is the region where ions are not *frozen-in* anymore, however, due to the smaller gyroradius, electrons are still *frozen-in*. In this region we also find an anomalous resistivity. The electron diffusion region is much smaller (yellow region in Figure 3.4) and defines the region where the electrons are not *frozen-in*.

In a simplified picture, magnetic reconnection occurs in the diffusion region. At the center of the diffusion region lies the so-called X-point. Due to plasma inflow from both sides (above and below in Figure 3.4), magnetic flux is continuously carried towards the X-point. Inside the electron diffusion region the magnetic field lines loose their identity. Field lines are merged from opposite sides of the current sheet and convected away from the X-point. This causes the magnetic field topology to change. The regions containing field lines with different magnetic topology are separated by the separatrix (dashed black line). The continuous plasma inflow yields a plasma outflow (according to the continuity equation) and plasma flows away from the reconnection region. The inflow region is much larger (D in Figure 3.4) than the outflow region (d in Figure 3.4) which indicates that the plasma is accelerated. Typical outflow speeds are in the order of the Alfvén speed in the inflow region.

In the ion diffusion region, the average path of electrons (red line in Figure 3.4) and average path of ions (blue line) differ. The electrons are still frozen in while the ions do not. This gives rise to a current directed in the opposite direction as the electron flow. These currents then give rise to an additional magnetic field (depicted in green in Figure 3.4), the Hall magnetic field. This Hall magnetic field is in the quadrupole configuration. Due to the motion of the ions which is not bounded to the magnetic field in the ion diffusion region, ions can cross the boundary between the two plasma regions. If the two plasma regions contain different particle populations, these particles can be observed close to the boundary in both region after reconnection onset.

4. Waves in Plasma

4.1. Introduction

We are always surrounded by two types of waves—light and sound. In a molecular gas such as air, electromagnetic and acoustic waves are not coupled. The pressure variations we hear as sound do not cause any electromagnetic phenomena, and light, or radio waves, will make no noise as they propagate through the air. In a plasma, electromagnetic and acoustic phenomena can no longer be treated separately. The electric and magnetic field of an electromagnetic wave penetrating into a plasma will act on the charged particles, and the free electrons and ions will start to oscillate in the force field of the wave. These vibrations of the plasma particles will also cause pressure fluctuations that are reminiscent of an acoustic wave. Similarly, a sound wave in a plasma creates charge densities and currents that are sources of fluctuating electromagnetic fields. Hence, a plasma can support many kinds of waves that are mixtures between the purely electromagnetic and purely acoustic modes that propagate in the neutral atmosphere.

Why is it important to study waves? In the atmosphere, energy between particles is exchanged through collisions. In space we mostly work with collisionless plasma which means energy cannot be transferred by collisions. Instead, the energy can be transferred via wave-particle interactions. Waves can grow and decay and therefore decrease and increase the energy in particles. Furthermore, waves transport energy and information across large distances. To understand energy transfer and transport in our magnetosphere, we need to study waves.

4.2. The Basic Equations

Waves in plasmas can be described by several different mathematical models at various levels of sophistication. All treatments of plasma waves are based on Maxwell's equations

$$\nabla \cdot \mathbf{B}(t, \mathbf{r}) = 0 \quad (4.1a)$$

$$\nabla \cdot \mathbf{E}(t, \mathbf{r}) = \frac{\rho}{\epsilon_0} \quad (4.1b)$$

$$\nabla \times \mathbf{E}(t, \mathbf{r}) = -\frac{\partial \mathbf{B}(t, \mathbf{r})}{\partial t} \quad (4.1c)$$

$$\nabla \times \mathbf{B}(t, \mathbf{r}) = \frac{1}{c^2} \frac{\partial \mathbf{E}(t, \mathbf{r})}{\partial t} + \mu_0 \mathbf{j}(t, \mathbf{r}). \quad (4.1d)$$

These equations describe how the electric field $\mathbf{E}$ and the magnetic field $\mathbf{B}$ are related to each other and to the charge density ρ and current density $\mathbf{j}$. Sometimes Equation 4.1b is called Gauss equation, Equation 4.1c is Faraday's law, and Equation 4.1d is Ampère's

4. Waves in Plasma

law. The mathematical models differ mainly in the way the source terms ρ and $\mathbf{j}$ are calculated from the plasma properties. In vacuum the charge and current density vanish. The solution to the Maxwell equations in vacuum are electromagnetic (light) waves travelling at the speed of light. On the other hand, the free electrons and ions in a plasma can support strong currents, and these currents can make plasma waves very different from waves in vacuum.

A well defined solution of the Maxwell's equations is ensured by the Helmholtz's theorem (for the proof and boundary conditions, check your electromagnetics course). The Helmholtz theorem states that a vector field $\mathbf{F}(\mathbf{r})$ can be divided into a curl-free part $\mathbf{G}(\mathbf{r})$, completely determined by its divergence sources, and a divergence-free part $\mathbf{R}(\mathbf{r})$, determined by its curl sources such that

$$\mathbf{F}(\mathbf{r}) = \mathbf{G}(\mathbf{r}) + \mathbf{R}(\mathbf{r}). \quad (4.2)$$

The curl-free part needs to satisfy $\mathbf{G}(\mathbf{r}) = -\nabla\Phi(\mathbf{r})$ and the divergence-free part needs to satisfy $\nabla \cdot \mathbf{R}(\mathbf{r}) = 0$.

We may thus divide the electric field $\mathbf{E}$ into a curl-free part $\mathbf{E}_S$ and a divergence-free part $\mathbf{E}_I$. The curl-free part is known as the *electrostatic field*, satisfying

$$\nabla \cdot \mathbf{E}_S(t, \mathbf{r}) = \rho(t, \mathbf{r})/\epsilon_0 \quad (4.3)$$

$$\nabla \times \mathbf{E}_S(t, \mathbf{r}) = 0, \quad (4.4)$$

and may thus be written in terms of a scalar potential as

$$\mathbf{E}_S(t, \mathbf{r}) = -\nabla\Phi(t, \mathbf{r}). \quad (4.5)$$

Note that even though the field is called *electrostatic*, it does not have to be static at all: if the charge density ρ is varying in time, so will $\mathbf{E}_S$. In fact, we will find that a plasma supports electrostatic waves – a phenomenon unknown in vacuum and neutral gases.

The divergence-free part is known as the *induced electric field*, and obeys

$$\nabla \cdot \mathbf{E}_I(t, \mathbf{r}) = 0 \quad (4.6)$$

$$\nabla \times \mathbf{E}_I(t, \mathbf{r}) = -\frac{\partial \mathbf{B}(t, \mathbf{r})}{\partial t}. \quad (4.7)$$

The total electric field is

$$\mathbf{E} = \mathbf{E}_S + \mathbf{E}_I. \quad (4.8)$$

According to Gauss' law (4.1a), there are no divergence sources for the magnetic field, and therefore no magnetic analogy of the electrostatic field. When working with waves, this concept can be used to separate a wave mode in an electrostatic and an electromagnetic part since they can be described separately from each other.

4.3. Basic classification of waves

Before we start discussing a range of plasma waves there are a few things to keep in mind when working with plasma waves.

In the following we will only work with linear waves. We will use the approach of *linearisation* as described in section 4.4 which only applies to small perturbations.

Once the perturbations by the wave itself is of the same size as the background field, the perturbation is not *small* anymore and the background field cannot be considered constant anymore. This is the so called *non-linear* regime.

There are a lot of plasma waves that have been described and derived in the literature. Often there are even multiple ways to derive the same wave equation (the so called dispersion relation). That makes naming conventions often confusing as the same wave can have several names depending on the context. All these waves are derived under certain assumption, however space plasma can be more complicated. While the observed wave might have slightly different properties compared to the theory, researchers often try to put names on the waves.

Waves are described by their dispersion relation. A dispersion relation $\omega(\mathbf{k})$ relates the frequency of a wave ω with the wavenumber $\mathbf{k}$. From the dispersion relation we can derive the phase velocity $\mathbf{v}_{ph}$

$$\mathbf{v}_{ph} = \frac{\omega}{k} \hat{k} \quad (4.9)$$

and the group velocity $\mathbf{v}_g$

$$\mathbf{v}_g = \frac{\partial \omega}{\partial \mathbf{k}}. \quad (4.10)$$

Dispersion relations are in general complex functions. The real part gives information about propagation properties and the imaginary part gives information about wave growth or decay.

In the following we name some useful classification schemes for waves. Here, ξ_0 refers to the time independent (constant) part of a variable (*the background*), ξ_1 is smaller but time and space dependent, that means it is oscillating and can be considered the *wave part*. This ansatz will be discussed in detail in the next section. These classification schemes can help to order the *zoo of waves*.

Waves can be categorized as either *electromagnetic* waves or *electrostatic* waves. An electromagnetic wave comprises both electric and magnetic components that oscillate ($\mathbf{B}_1 \neq 0$, $\mathbf{E}_1 \neq 0$). On the other hand, electrostatic waves lack the magnetic component and solely exhibit variations in the electric field ($\mathbf{B}_1 = 0$, $\mathbf{E}_1 \neq 0$).

Alternatively, waves can be categorized as *electron* waves or *ion* waves. In the context of electron waves, ions are assumed to remain stationary, which can be mathematically expressed as $m_i = \infty$. In contrast, ion waves involve the motion of ions. Typically, ion waves exhibit lower frequencies compared to electron waves, primarily due to the significantly higher mass of ions in comparison to electrons.

In the context of magnetized plasmas ($B_0 \neq 0$), we can use the propagation direction relative to background magnetic field B_0 to classify waves. Waves can propagate *parallel* to the magnetic field ($\mathbf{k} \parallel \mathbf{B}_0$) or *perpendicular* ($\mathbf{k} \perp \mathbf{B}_0$). If a wave's direction of propagation does not coincide with either parallel or perpendicular configurations, it is considered to propagate at an oblique angle.

Finally, waves can be categorized as *longitudinal* waves or *transversal* waves. In a longitudinal wave, the electric field oscillates parallel to the direction of propagation ($\mathbf{k} \parallel \mathbf{E}_1$ or $\mathbf{k} \times \mathbf{E}_1 = 0$). On the other hand, in a transverse wave, the electric field oscillates perpendicular to the direction of propagation ($\mathbf{k} \perp \mathbf{E}_1$ $\mathbf{k} \cdot \mathbf{E}_1 = 0$). From Farady's law Equation 4.1c we know that $\mathbf{k} \times \mathbf{E}_1 = 0$ which implies that electrostatic waves are always

4. Waves in Plasma

longitudinal waves. A transversal wave will always have a varying magnetic field and is therefore an electromagnetic wave.

4.4. Linearisation

Linearisation is a fundamental method for studying the response of an equilibrium situation to a **small perturbation**, and is used in all parts of physics. We summarize the method here.

Assume we have a non-linear system of equations $F(\xi) = 0$ for the field variables ξ . For two-fluid plasma theory we have 14 unknown field components ($\mathbf{E}$, $\mathbf{B}$, $\mathbf{v}_i$, $\mathbf{v}_e$, n_i och n_e), so ξ is a vector in 14 dimensions.

1. **Ansatz:** $\xi(t, \mathbf{r}) = \xi_0 + \xi_1(t, \mathbf{r})$, where $\xi_1 \ll \xi_0$.
2. **Use in field equations:** $F(\xi_0 + \xi_1) = 0$
3. **Derivatives of background values are zero.**
4. **Neglect higher terms:** This gives linearised equations $G(\xi_0)\xi_1 = 0$.
5. This homogeneous system of linear equations has non-trivial solutions only if the determinant is zero, $\det(G(\xi_0)) = 0$. This equation is the **dispersion relation**.

Looking at our process from a different angle, we can think of it as expanding the non-linear system using a Taylor series around the point ξ_0 . The background fields ξ_0 fulfill the field equations $F(\xi_0) = 0$. We then get

$$0 = F(\xi) = F(\xi_0 + \xi_1) = F(\xi_0) + F'(\xi_0)\xi_1 + \dots = F'(\xi_0)\xi_1 + \dots \quad (4.11)$$

Neglecting higher order terms, we have

$$F'(\xi_0)\xi_1 = 0 \quad (4.12)$$

as our linearised system. Thus, $G(\xi_0)$ above is nothing else than $F'(\xi_0)$. This is a theoretically elegant way of summarizing the linearisation process, but in practice, this Taylor expansion formalism is cumbersome to handle. For the 14 dimensional state vector of two-fluid plasma theory, for instance, we get $F'(\xi_0)$ as a 14×14 matrix so the dispersion relation $\det(F'(\xi_0)) = 0$ will be rather bulky.

In the following we will derive some wave dispersion relation based on linearization, but instead of linearising all variables we will just linearise some variables based on the procedure explained above.

4.5. Electrostatic waves

Electrostatic waves are, as previously discussed, waves with no varying magnetic field. In the following we will consider two types of electrostatic waves, plasma oscillations and ion acoustic waves.

4.5.1. Plasma oscillations/ Langmuir waves

We will start by deriving the dispersion relation for plasma oscillations, which is an electrostatic "wave". Ions are much heavier than electrons, and at very high frequencies, the ion inertia is so large such that ions will not react to the electron motion. Therefore, we want to consider a wave where we can neglect the ion motion ($m_i \sim \infty$), and we assume that all ions are uniformly distributed ($n_0 = \text{constant}$). In addition, we assume that there is no background magnetic field ($B_0 = 0$). To simplify our calculation, we assume that all variations are only along x . This leaves us with three relevant equations, namely the continuity equation 3.29, the momentum equation 3.30, and Gauss' equation 4.1b. For Gauss' equation we used that the charge density $\rho = e(n_0 - n_e)$. All this gives

$$\frac{\partial}{\partial t} n_e + \nabla \cdot (n_e \mathbf{u}_e) = 0 \quad (4.13)$$

$$m_e n_e \left(\frac{\partial}{\partial t} u_e + u_e \frac{\partial}{\partial x} u_e \right) = - \frac{\partial}{\partial x} p_e - e n_e E_x \quad (4.14)$$

$$\frac{\partial}{\partial x} E_x = \frac{e(n_0 - n_e)}{\epsilon_0}, \quad (4.15)$$

where u_e is the electron velocity, E is the electric field, and p_e is the electron pressure. These equations cannot be solved exactly. Instead we will linearize them and we assume that we have cold electrons such that $p_e = 0$.

1. Ansatz.

$$n_e(x, t) = n_0 + n_1(x, t) \quad (4.16)$$

$$u_e(x, t) = 0 + u_1(x, t) \quad (4.17)$$

$$E_x(x, t) = 0 + E_1(x, t) \quad (4.18)$$

$$p_e(x, t) = p_0 + p_1 = 0 \quad (4.19)$$

2. Insert the ansatz into the field equations (4.13) – (4.15).

3. Derivatives of background values disappear.

4. Neglect quadratic and higher order terms. This procedure provides us with a set of linearized MHD equations:

$$\begin{aligned} \frac{\partial}{\partial t} n_1 + n_0 \frac{\partial u_1}{\partial x} &= 0 \\ m_e n_0 \frac{\partial u_1}{\partial t} &= -e n_0 E_1 - \frac{\partial p_1}{\partial x} \\ \frac{\partial}{\partial x} E_1 &= -\frac{e n_1}{\epsilon_0} \end{aligned} \quad (4.20)$$

We can obtain the solution by assuming plane wave in the form of

$$E_1(x, t) = \tilde{E}_1 \exp(-i\omega t + ikx) \quad (4.21)$$

$$n_1(x, t) = \tilde{n}_1 \exp(-i\omega t + ikx) \quad (4.22)$$

$$u_1(x, t) = \tilde{u}_1 \exp(-i\omega t + ikx) \quad (4.23)$$

4. Waves in Plasma

with $\tilde{E}_1$, $\tilde{n}_1$, and $\tilde{u}_1$ are the complex amplitudes. This ansatz we can now insert into the equations

$$-i\omega\tilde{n}_1 + ikn_0\tilde{u}_1 = 0 \quad (4.24)$$

$$-i\omega\tilde{u}_1 + \frac{e}{m_e}\tilde{E}_1 = 0 \quad (4.25)$$

$$\frac{e\tilde{n}_1}{\epsilon_0} + ik\tilde{E}_1 = 0 \quad (4.26)$$

$$(4.27)$$

This system of linear equations has the solution (solve by substitution or set the determinant of the coefficient matrix to 0)

$$\omega^2 = \frac{n_0 e^2}{m_e \epsilon_0} \equiv \omega_{pe}^2. \quad (4.28)$$

The solution is independent of the wavenumber k . Therefore this "wave" cannot propagate since $v_g = \frac{\partial}{\partial k}\omega = 0$ but only oscillate, which means no information can be transferred. These oscillations are also called *plasma oscillations*. Instrumentation, such as electric field instruments, that can measure these oscillations can be used to determine the plasma density.

In the derivation we assumed that the electrons are cold so their thermal energy is 0. If we drop this assumption (calculation not shown here) we need to include the perturbation of the pressure gradient in Equation 4.14. The resulting dispersion relation

$$\omega^2 = \omega_{pe}^2 + 3k^2 v_{th}^2, \quad (4.29)$$

with the thermal speed for electrons $v_{th} = \sqrt{\frac{2T_0}{m_e}}$, is indeed a propagating wave. These waves are called *Langmuir waves*.

4.5.2. Ion acoustic waves

For the Langmuir waves, we explicitly assumed high frequency in the derivation. We now do the opposite assumption: assuming waves of so low frequency that the electrons have plenty of time to find an equilibrium. They will then be distributed in space according to the Boltzmann relation

$$n_e(t, \mathbf{r}) = n_0 \exp\left(\frac{e\Phi(t, \mathbf{r})}{KT_e}\right) \quad (4.30)$$

since their potential energy is $-e\Phi$. As usual, we linearize the equation for waves of small amplitude. Then Φ is assumed small, and we only keep the first term in the Taylor expansion of the exponential,

$$n_e(t, \mathbf{r}) \approx n_0 \left(1 + \frac{e\Phi(t, \mathbf{r})}{KT_e}\right) \quad (4.31)$$

Electrons are much lighter than the ions, therefore they will essentially short-circuit any charge imbalance caused by the ion motion such that

$$n_i = n_e. \quad (4.32)$$

Now we can apply our method of linearization for small perturbations.

1. **Ansatz.**

$$n_e(x, t) = n_0 + n_1(\mathbf{r}, t) = n_0 + n_0 \frac{e\Phi(t, \mathbf{r})}{KT_e} \quad (4.33)$$

2. **Insert the ansatz into the field equations**

We will use the ion equation of motion and the continuity equation. We also assume isothermal¹ conditions, so the temperature of the system stays constant. Then, the ion equation of motion is

$$m_i n_0 \frac{\partial v_{1i}(t, \mathbf{r})}{\partial t} = -n_0 e \nabla \Phi_1(t, \mathbf{r}) + KT_i \nabla n_{1i}(t, \mathbf{r}) \quad (4.34)$$

and their equation of continuity is

$$\frac{\partial n_i(t, \mathbf{r})}{\partial t} + \nabla \cdot [n_i(t, \mathbf{r}) \mathbf{v}_i(t, \mathbf{r})] = 0. \quad (4.35)$$

3. **Derivatives of background values disappear.**4. **Neglect quadratic and higher order terms.**

After the usual linearization and the plane wave ansatz procedure, we get

$$-i\omega m_i n_0 \mathbf{v}_i = -i\mathbf{k} [n_0 e \Phi + n_{1i} KT_i] \quad (4.36)$$

$$-i\omega n_{1i} + i n_0 \mathbf{k} \cdot \mathbf{v}_i = 0. \quad (4.37)$$

Combining these two equations with (4.33) and (4.32) yields

$$\omega^2 n_{1i} = \omega n_0 \mathbf{k} \cdot \mathbf{v}_i = \frac{1}{m_i} \mathbf{k} \cdot [\mathbf{k} (n_{1i} KT_i + n_0 e \Phi)] = \frac{k^2}{m_i} (KT_i + KT_e) n_{1i}, \quad (4.38)$$

giving the dispersion relation

$$\omega^2 = c_{ia}^2 k^2$$

(4.39)

where the ion acoustic speed c_{ia} is defined by ²

$$c_{ia}^2 = \frac{KT_i + KT_e}{m_i}. \quad (4.40)$$

This wave mode is called the *ion acoustic wave*. We note that in the dispersion relation, we find the mass of the ions, but the temperature for ions as well as electrons. A common situation in space plasmas is to have $T_i \ll T_e$, giving a dispersion relation $\omega^2 = (KT_e/m_i)k^2$. The properties of the wave are then determined by the ion inertia and the electron temperature. This may be understood as follows. The oscillations are an ongoing energy transformation between particle kinetic energy and potential energy in the wave electrostatic field. As the ions are much heavier than the electrons, their kinetic energy is dominating, and thus the ion mass is important. If the electrons as well as the ions were cold, they would quickly short-circuit the wave electric field, as it has

¹As usual, an adiabatic approximation would be more physical, but we stick to the isothermal approximation for simplicity.

²Assuming adiabatic ions with $\gamma = 3$ would have resulted in the same dispersion relation (4.39) with $c_{ia}^2 = (3KT_i + KT_e)/m_i$.

4. Waves in Plasma

a frequency far below the plasma frequency, which is the frequency where the electron inertia becomes important, and there would be no wave. However, since they have a certain temperature, the screening of the electric field is not perfect. Compare to how the electrons cannot completely neutralize a charge imbalance in Debye screening of a stationary charge. Since in the Debye case, the screening is less efficient the higher the electron temperature is, making it easier for the wave to propagate, i.e. increasing its speed.

Thus, when considering the ion motion in a thermal plasma, we find an electrostatic mode propagating at frequencies below the plasma frequency.

4.6. Magnetohydrodynamic waves

In the context of magnetohydrodynamic (MHD) theory, we can derive electromagnetic waves with low frequencies in comparison to the ion gyrofrequency ($f \ll f_{gi}$). These waves possess the remarkable ability to induce disturbances that propagate over large distances and transfer energy across significant scales.

In this chapter, we will derive the dispersion relation specifically for cold plasmas. Subsequently, we will provide a brief description of magnetohydrodynamic waves in warm plasmas, broadening our understanding of these captivating phenomena.

4.6.1. Magnetohydrodynamic waves in cold plasmas

First, we will focus on cold plasma magnetohydrodynamics, where $p = T = 0$, i.e. the thermal pressure p is unimportant. The waves we will find in this model are known as magnetohydrodynamic (MHD) waves. Collecting the equations from the preceding section, we describe the plasma dynamics by the following system:

$$\rho_m(t, \mathbf{r}) \frac{d}{dt} \mathbf{v}(t, \mathbf{r}) = \mathbf{j}(t, \mathbf{r}) \times \mathbf{B}(t, \mathbf{r}) \quad (4.41)$$

$$\nabla \times \mathbf{E}(t, \mathbf{r}) = -\frac{\partial}{\partial t} \mathbf{B} \quad (4.42)$$

$$\nabla \times \mathbf{B}(t, \mathbf{r}) = \mu_0 \mathbf{j}(t, \mathbf{r}) \quad (4.43)$$

$$\mathbf{E}(t, \mathbf{r}) + \mathbf{v}(t, \mathbf{r}) \times \mathbf{B}(t, \mathbf{r}) = 0. \quad (4.44)$$

These equations are non-linear, so we apply our usual method of linearization for small perturbations denoted by the subscript 1 (section 4.4):

1. Ansatz.

$$\begin{aligned} \rho_m(t, \mathbf{r}) &= \rho_0 + \rho_1(t, \mathbf{r}) \\ \mathbf{B}(t, \mathbf{r}) &= \mathbf{B}_0 + \mathbf{B}_1(t, \mathbf{r}) \\ \mathbf{E}(t, \mathbf{r}) &= \mathbf{E}_1(t, \mathbf{r}) \\ \mathbf{v}(t, \mathbf{r}) &= \mathbf{v}_1(t, \mathbf{r}) \\ \mathbf{j}(t, \mathbf{r}) &= \mathbf{j}_1(t, \mathbf{r}) \end{aligned} \quad (4.45)$$

2. Insert the ansatz into the field equations (4.41) – (4.44).

3. **Derivatives of background values disappear.**

4. **Neglect quadratic and higher order terms.** This procedure provides us with a set of linearized MHD equations:

$$\rho_{m0} \frac{\partial \mathbf{v}_1(t, \mathbf{r})}{\partial t} = \mathbf{j}_1(\mathbf{r}, t) \times \mathbf{B}_0 \quad (4.46)$$

$$\nabla \times \mathbf{E}_1(t, \mathbf{r}) = - \frac{\partial \mathbf{B}_1(t, \mathbf{r})}{\partial t} \quad (4.47)$$

$$\nabla \times \mathbf{B}_1(t, \mathbf{r}) = \mu_0 \mathbf{j}_1(t, \mathbf{r}) \quad (4.48)$$

$$\mathbf{E}_1(t, \mathbf{r}) + \mathbf{v}_1(t, \mathbf{r}) \times \mathbf{B}_0 = 0. \quad (4.49)$$

As this is a linear system, we may confine our interest to plane waves. The linearized equations transform to

$$-i\omega \rho_0 \mathbf{v}_1 = \mathbf{j}_1 \times \mathbf{B}_0 \quad (4.50)$$

$$i\mathbf{k} \times \mathbf{E}_1 = i\omega \mathbf{B}_1 \quad (4.51)$$

$$i\mathbf{k} \times \mathbf{B}_1 = \mu_0 \mathbf{j}_1. \quad (4.52)$$

$$\mathbf{E}_1 + \mathbf{v}_1 \times \mathbf{B}_0 = 0. \quad (4.53)$$

We choose the z -axis to lie along $\mathbf{B}_0$:

$$\mathbf{B}_0 = B_0 \hat{\mathbf{z}}. \quad (4.54)$$

From now on, we will drop the subscript “1” in the wave fields for $\mathbf{E}$, $\mathbf{v}$ and $\mathbf{j}$, for which there are no background fields to cause confusion. In component form, the equation of motion then writes as

$$-i\omega \rho_{m0} v_x = j_y B_0 \quad (4.55)$$

$$-i\omega \rho_{m0} v_y = -j_x B_0 \quad (4.56)$$

$$v_z = 0, \quad (4.57)$$

while (4.53) appear as

$$E_x = -v_y B_0 \quad (4.58)$$

$$E_y = v_x B_0 \quad (4.59)$$

$$E_z = 0. \quad (4.60)$$

Multiplying (4.55) and (4.56) by $i\omega \mu_0 / B_0$ and using the velocity components from (4.58) and (4.59) yields

$$i\omega \mu_0 j_x = \omega^2 E_x / c_A^2 \quad (4.61)$$

$$i\omega \mu_0 j_y = \omega^2 E_y / c_A^2 \quad (4.62)$$

where we have introduced *the Alfvén speed*

$$c_A = \frac{B_0}{\sqrt{\mu_0 \rho_{m0}}} \quad (4.63)$$

The physical interpretation of c_A will be unveiled below. The components of Ampère’s law (4.52) are

$$\mu_0 j_x = i(k_y B_{1z} - k_z B_{1y}) \quad (4.64)$$

4. Waves in Plasma

$$\mu_0 j_y = i(k_z B_{1x} - k_x B_{1z}) \quad (4.65)$$

$$\mu_0 j_z = i(k_x B_{1y} - k_y B_{1x}), \quad (4.66)$$

while Faraday-Henry's law (4.51) looks like

$$\omega B_{1x} = -k_z E_y \quad (4.67)$$

$$\omega B_{1y} = k_z E_x \quad (4.68)$$

$$\omega B_{1z} = k_x E_y - k_y E_x \quad (4.69)$$

where we used $E_z = 0$, which follows from (4.60). We now eliminate $\mathbf{B}_1$ from (4.64) - (4.66) by use of (4.67) - (4.69), and insert the resulting expressions for $\mathbf{j}$ in terms of $\mathbf{E}$ in (4.61) and (4.62), whereby we find that

$$(k_y^2 + k_z^2 - \omega^2/c_A^2)E_x = k_x k_y E_y \quad (4.70)$$

and

$$(k_x^2 + k_z^2 - \omega^2/c_A^2)E_y = k_x k_y E_x. \quad (4.71)$$

Multiplying these equations, we get

$$\begin{aligned} 0 &= (k_y^2 + k_z^2 - \omega^2/c_A^2)(k_x^2 + k_z^2 - \omega^2/c_A^2) - k_x^2 k_y^2 \\ &= (k_z^2 - \omega^2/c_A^2)(k_x^2 + k_z^2 - \omega^2/c_A^2) + \\ &\quad + k_y^2 k_x^2 + k_y^2(k_z^2 - \omega^2/c_A^2) - k_x^2 k_y^2 = \\ &= (k_z^2 - \omega^2/c_A^2)(k_x^2 + k_y^2 + k_z^2 - \omega^2/c_A^2), \end{aligned} \quad (4.72)$$

so one of the parentheses must be zero. We thus get two dispersion relations representing independent wave modes in the plasma. One of them is known as the *Alfvén mode*,

$$\omega^2 = k_z^2 c_A^2 = k_{\parallel}^2 c_A^2 \quad (4.73)$$

while the other mode here will be known as the *compressional mode*³,

$$\omega^2 = (k_x^2 + k_y^2 + k_z^2)c_A^2 = k^2 c_A^2 \quad (4.74)$$

By neglecting of the displacement current in the Ampère-Maxwell's law, our results are non-relativistic, valid only for $c_A \ll c$, i.e. for sufficiently weak ambient magnetic fields and high plasma densities (see the definition (4.63) of the Alfvén speed).

How do these waves behave? Let's first consider the compressional mode. This mode propagates equally in all directions, meaning it's isotropic. This is related to the fact that only the magnitude k of the wave vector $\mathbf{k}$ is included in the dispersion relation (4.74), not the direction. Since the dispersion relation is a linear relation between k and ω , all frequencies propagate at the same phase velocity, therefore the mode is called non-dispersive. The propagation characteristics of the compressional wave therefore are

³These wave modes both have other names too and there is no universal nomenclature. *Alfvén wave* sometimes refer to the mode we call the Alfvén mode, but can also denote all MHD waves. Then the modes are called *compressional Alfvén wave* and *shear* or *torsional Alfvén wave*. The compressional mode can sometimes be known as *magnetosonic waves* even in the cold case with no thermal corrections.

similar to light wave in vacuum or a sound wave in air. Just the speed of propagation differs, for the compressional mode it is c_A rather than c or c_s . As is the case for all non-dispersive waves, the group velocity $\mathbf{v}_g$

$$\frac{\partial \omega}{\partial k} = c_A \hat{k}$$

is parallel the wave vector $\mathbf{k}$, and its magnitude is c_A in all directions. A perturbation from a point source thus propagates uniformly in all directions, like the rings on the water surface when a stone is dropped in a pond. To make sure energy is conserved, the amount of wave energy passing through any sphere with radius r around that point stays the same. This means that as the distance from the source increases, the wave energy density decreases, since the energy spreads out over a larger area. Mathematically, this translates to the wave energy density decreasing as $1/r^2$ and the wave fields (like $(\mathbf{E}_1, \mathbf{B}_1$ etc.)) decreasing as $1/r$.

The Alfvén mode is also characterized by c_A , but otherwise it behaves quite differently. From the dispersion relation (4.73),

$$\mathbf{v}_{ph} = \frac{\omega}{k} = \frac{k_{\parallel}}{k} c_A = c_A \cos \theta, \quad (4.75)$$

where θ is the angle between $\mathbf{B}_0$ and $\mathbf{k}$, and

$$\mathbf{v}_g = \frac{\partial \omega}{\partial \mathbf{k}} = \pm c_A \hat{\mathbf{z}} \quad (4.76)$$

where z as usual is chosen along the ambient magnetic field $\mathbf{B}_0$. The result for the group velocity $\mathbf{v}_g$ is of particular interest: the Alfvén wave transports energy exclusively along the magnetic field!

Note that this is independent of wave vector $\mathbf{k}$ —the wave will propagate along the direction of the magnetic field $\mathbf{B}_0$ only. This is important because it allows the Alfvén wave to transport energy over long distances without the usual decrease in its intensity that occurs with radiation from a point source (which follows a $1/r^2$ dependence). In other words, the Alfvén wave can travel far without losing its energy rapidly.

What do the actual fields from these wave modes look like? Let us choose coordinates so that $\mathbf{B}_0 = B_0 \hat{\mathbf{z}}$ and $\mathbf{k} = k_x \hat{\mathbf{x}} + k_z \hat{\mathbf{z}}$: as there only are two preferred directions in the problem, defined by $\mathbf{B}_0$ and $\mathbf{k}$, this implies no loss of generality. We thus have $k_y = 0$, $k_{\parallel} = k_z$ and $k^2 = k_x^2 + k_z^2$. Using the dispersion relations for Alfvénic (4.73) and compressional (4.74) waves to eliminate ω^2 from equations (4.70) and (4.71), we find that for the compressional mode, $E_x = 0$, while $E_y = 0$ for the Alfvén mode. By going backwards through equations (4.69) to (4.58), we can find all other field components as well.

The result is presented in Table 4.1. Equation (4.44) implies that the magnetic field lines may be considered to be “frozen” in the plasma as has been discussed previously in the compendium. The compressional wave behaves more like a sound wave, with compressions and decompressions of the plasma. For compressional waves, one may envisage a sort of wave motion in a drapery made up of hanging ropes, connected to each other by horizontal springs. The Alfvén wave bends the magnetic field without compressing it. The two wave modes are illustrated in Figure 4.1.

4. Waves in Plasma

Alfvén mode	Compressional mode
$E_x \neq 0$ $E_y = 0$ $E_z = 0$	$E_x = 0$ $E_y \neq 0$ $E_z = 0$
$B_{1x} = 0$ $B_{1y} = E_x/c_A$ $B_{1z} = 0$	$B_{1x} = -k_z E_y/\omega$ $B_{1y} = 0$ $B_{1z} = k_x E_y/\omega$
$v_x = 0$ $v_y = -E_x/B_0$ $v_z = 0$	$v_x = E_y/B_0$ $v_y = 0$ $v_z = 0$
$j_x = -i\omega E_x/\mu_0 c_A^2$ $j_y = 0$ $j_z = ik_x E_x/\mu_0 c_A$	$j_x = 0$ $j_y = -i\omega E_y/\mu_0 c_A^2$ $j_z = 0$
$\mathbf{v}_g = c_A \hat{z}$ $\mathbf{v}_{ph} = \cos \theta c_A \hat{k}$	$\mathbf{v}_g = c_A \hat{k}$ $\mathbf{v}_{ph} = c_A \hat{k}$

Table 4.1.: The wave fields of hydromagnetic waves. The coordinates are chosen so as to have $\mathbf{B}_0 = B_0 \hat{z}$ and $\mathbf{k} = k_x \hat{x} + k_z \hat{z}$.

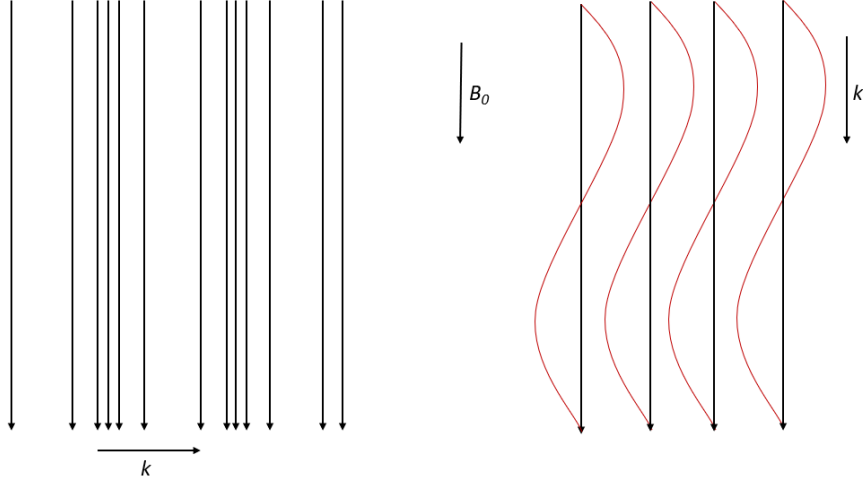


Figure 4.1.: An illustration of the magnetic field of the Alfvén mode (right) and the compressional mode (left).

Alfvén waves are very important in many cosmic contexts. As an example, we may consider the terrestrial magnetosphere-ionosphere system. The Alfvén wave is responsible for transport of information from the magnetopause to the ionosphere. Changes in the field-aligned currents in the magnetosphere are transmitted by Alfvén waves. This is due to the fact that Alfvén waves have a current component along $\mathbf{B}_0 = B_0 \hat{\mathbf{z}}$ (compare Table 4.1). When using a plane wave ansatz for a time-dependent field aligned current, the components at non-zero frequency will be Alfvén waves⁴. Because of their relation to field-aligned currents, Alfvén waves are important in auroral processes. Looking at Table 4.1, we find that the magnetohydrodynamic waves do not carry any electric field component along $\mathbf{B}_0$, so at first glance they should not be able to accelerate particles in parallel to the magnetic field. However, in auroral regions the perpendicular scale sizes (of for example auroral arcs) often are of the order of the ion gyroradius, where the simple MHD theory we have used here is no longer applicable. More elaborate descriptions in fact show that Alfvén waves in fact carry a parallel electric field if $k_\perp \neq 0$. This field is normally much smaller than the perpendicular electric field, but nevertheless very important.

4.6.2. Magnetohydrodynamic waves in warm plasmas

We will briefly discuss the effects of a warm plasma where the thermal pressure cannot be neglected anymore ($\nabla p \neq 0$). We will not derive the resulting dispersion relation in this context, but we will briefly discuss the solution. In the cold plasma derivation, we found two different wave modes, the compressional mode and the Alfvén mode. In the warm plasma case we can find three different modes, the Shear Alfvén wave, the fast and the slow mode. The dispersion relation of the shear mode is given by

$$\left(\frac{\omega}{k}\right)^2 = c_A^2 \cos^2 \theta. \quad (4.77)$$

The dispersion relation of the fast (+) and the slow (-) mode are governed by

$$\left(\frac{\omega}{k}\right)^2 = 0.5 \left(c_s^2 + c_A^2 \pm \sqrt{(c_s^2 + c_A^2)^2 - 4c_s c_A \cos^2 \theta} \right), \quad (4.78)$$

with c_s as the speed of sound and c_A the Alfvén speed.

⁴Assuming the frequency to be well below ω_{ci} .

4. Waves in Plasma

Example

Assume that a cold plasma with electron density n_0 and equal density of infinitely heavy ions is perturbed by an electric field of the form $\mathbf{E}(\mathbf{r}, t) = \mathbf{E}_0 \exp[i(\omega t - \mathbf{k} \cdot \mathbf{r})]$. This will create a charge density $\rho = q(n(\mathbf{r}, t) - n_0)$, where q is the electron charge and $n(\mathbf{r}, t)$ is the perturbed density. Use the equation of motion $m\dot{\mathbf{v}} = q\mathbf{E}$, Poisson's equation $\epsilon_0 \nabla \cdot \mathbf{E} = \rho$, and the continuity equation relating electric charge and current densities, $\partial_t \rho + \nabla \cdot \mathbf{j} = 0$, to find an equation for electrostatic plasma oscillations. Must the frequency ω always be equal to the plasma frequency ω_p , or are there other possibilities?

Solution

Since the electrons are cold and the perturbation is small ($|\mathbf{v}| \ll \omega/k$) we can in the equation of motion neglect the convective term

$$m\dot{\mathbf{v}} = m\left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v}\right) \approx m\frac{\partial \mathbf{v}}{\partial t} = q\mathbf{E}$$

The time derivative of Poisson's equation gives

$$\frac{\partial \rho}{\partial t} = \epsilon_0 \frac{\partial}{\partial t} \nabla \cdot \mathbf{E}$$

which can be used to eliminate $\partial_t \rho$ from the continuity equation

$$\epsilon_0 \nabla \cdot \frac{\partial}{\partial t} \mathbf{E} = -\nabla \cdot \mathbf{j}$$

Only the electrons are moving, so the current is $\mathbf{j} = qn\mathbf{v} \approx qn_0\mathbf{v}$. Remember that $(n - n_0)$ and $\mathbf{v}$ are small perturbations, which means that $(n - n_0) \cdot \mathbf{v}$ can be neglected. Taking the time derivative once more we find

$$\epsilon_0 \nabla \cdot \frac{\partial^2}{\partial t^2} \mathbf{E} = -\nabla \cdot \frac{\partial}{\partial t} \mathbf{j} = -qn_0 \nabla \cdot \frac{\partial}{\partial t} \mathbf{v}$$

Using the equation of motion we then obtain

$$\epsilon_0 \nabla \cdot \frac{\partial^2}{\partial t^2} \mathbf{E} = -\frac{q^2 n_0}{m} \nabla \cdot \mathbf{E}$$

which can be rewritten as

$$\nabla \cdot \left(\frac{\partial^2}{\partial t^2} \mathbf{E} + \frac{q^2 n_0}{\epsilon_0 m} \mathbf{E} \right) = 0$$

With the assumed form of the electric field we have $\frac{\partial}{\partial t} \mathbf{E} = i\omega \mathbf{E}$ and $\nabla \cdot \mathbf{E} = -i\mathbf{k} \cdot \mathbf{E}$. We also introduce the plasma frequency by $\omega_p^2 = \frac{q^2 n_0}{\epsilon_0 m}$ which simplifies the equation to

$$\mathbf{k} \cdot \mathbf{E} (\omega^2 - \omega_p^2) = 0$$

This equation is satisfied with $\mathbf{E} \neq 0$ if $\omega = \omega_p$, which corresponds to longitudinal ($\mathbf{k} \perp \mathbf{E}$) plasma oscillations. The other possibility is $\mathbf{k} \cdot \mathbf{E} = 0$ ($\mathbf{k} \parallel \mathbf{E}$), which corresponds to transverse electromagnetic waves which can propagate with $\omega \neq \omega_p$.

Part II.

The space environment

5. The Sun and the Solar Wind

5.1. Introduction

From an astronomer's point of view, the Sun might be considered a fairly ordinary star among 10^{11} other stars in our galaxy. However, it is also apparently unique in many ways, and since the Sun is so close to us and its importance for us and our environment can hardly be overstated. Obviously, life on Earth could not exist without the Sun. The light and particles emitted from the Sun are also shaping our space environment, and this makes it appropriate to begin our study of the space environment with the Sun.

The Sun is also the source of the solar wind, which is plasma continuously flowing away from the Sun at supersonic speeds. Embedded in this plasma is a magnetic field, the interplanetary magnetic field (IMF). The solar wind interacts with all planetary bodies and is therefore the driver for the plasma environment that will be discussed in detail in chapter 6.

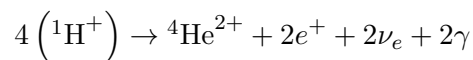
The overall structure of the Sun is shown in Figure 5.1. Characteristic densities and temperatures are indicated in this figure, as well as some important regions that are discussed on the following pages.

5.2. The Sun's interior

The Sun has a mass $M_{\odot} = 1.99 \cdot 10^{30}$ kg and a radius $R_{\odot} = 6.96 \cdot 10^8$ m, which means that the average density is 1410 kg/m^3 . This is denser than water, and the average density of the Earth is only about four times higher. About 90% of the solar mass is protons ($^1\text{H}^+$) and the remaining 10% is mostly α -particles ($^4\text{He}^{2+}$). There is of course also a charge-neutralizing background of free electrons. The temperature in the Sun is so high that electrons are knocked out of the atoms in frequent and energetic collisions – the gas is almost completely ionized, and in the plasma state.

5.2.1. Energy production and transport in the sun

The Sun is heated by fusion reactions in the *core* at its center. In the core, where the temperature is $1.5 \cdot 10^7$ K and the density $1.6 \cdot 10^5 \text{ kg/m}^3$, about $5.7 \cdot 10^{11}$ kg of hydrogen is converted to helium with a mass loss of $4.3 \cdot 10^9$ kg each second, releasing a power of $3.86 \cdot 10^{26}$ W. The fusion reaction, which in reality proceeds through several intermediate steps, can be summed up as



The two positrons (e^+) will also produce energy in the form of γ -rays after being annihilated by ambient plasma electrons, and the total energy released in the formation

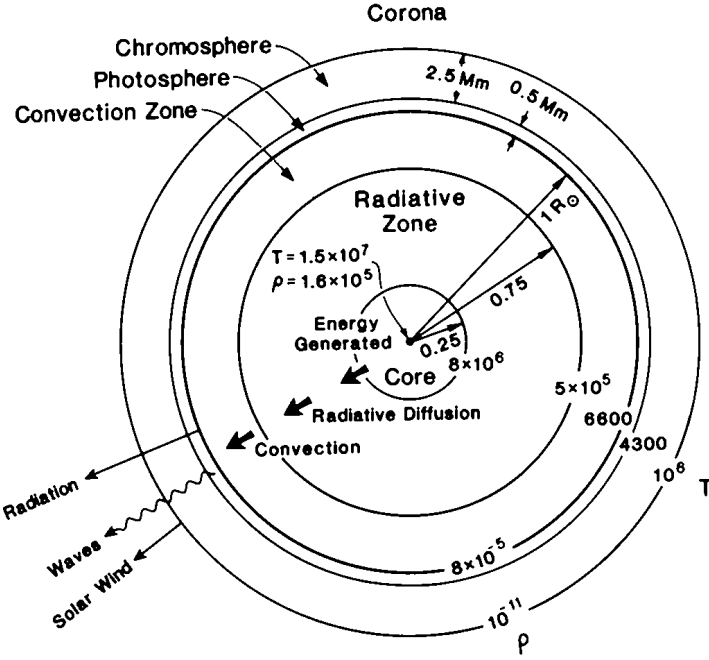


Figure 5.1.: Overall structure of the Sun. Figure from Kivelson and Russell [1995b]

of a single helium nucleus is then 26.7 MeV. In the dense solar core, the photons will lose energy by collisions with electrons until they are in thermal equilibrium with the particles. The γ -rays have then been converted to X-rays, which slowly leak out of the core. The core extends out to approximately $0.25 R_{\odot}$, where the temperature has gone down to $8 \cdot 10^6$ K.

The region outside the core, between $0.25 R_{\odot}$ and $0.8 R_{\odot}$, is called the *radiative zone*. The energy produced in the Sun's core is transferred through the radiative zone by photons. The interior of the Sun is very opaque to X-rays, and the photons are constantly scattered by electrons. If the photons moved freely they would reach the surface in ~ 2 s, but in reality there are so many scatterings that it takes them millions of years. As we move away from the core the temperature decreases, and the character of the radiation changes from X-rays to ultraviolet (UV). At the outer boundary of the radiation zone, the temperature has gone down to $5 \cdot 10^5$ K.

In the *convection zone*, outside $0.8 R_{\odot}$, the temperature is sufficiently low to allow a significant number of electrons to attach to the positive ions. Although these electrons rapidly will be torn off again by the radiation, they lead to increased energy absorption and steepening of the temperature gradient. Scattering of radiation by bound electrons is much less elastic than scattering by free electrons. As a result of the strong heating near the bottom of the convection zone, the plasma/gas mixture starts to move, churning and turning over much like water boiling in a pan. Volumes of plasma/gas will circulate between the bottom of the convection zone, where they are heated, and the solar surface where they are cooled. The primary mode of energy transport through the convection zone is thus the rising of hot gas elements and the downfall of cooler gas. These motions give rise to the distinctive pattern seen on the solar surface, referred to as 'supergranulation'. Near the surface of the Sun, the temperature is of the order $5 \cdot 10^3$ K.

5.2.2. The Solar Cycle and Sun's magnetic field

The Sun's magnetic field is much more complicated and variable than Earth's. In general, the Sun is observed to have a large scale dipolar field resembling the terrestrial field. An average magnetic field of $100 \mu\text{T}$ compared with $60 \mu\text{T}$ on Earth, dominates the northern polar region, and a similar field of opposite polarity dominates the corresponding southern region. However, the magnetic poles and the symmetry axis are not well defined on the Sun. A remarkable property is that the polarity of the solar magnetic field is reversed every eleven years, approximately. This, and other features like sunspots, is connected to the peculiar rotation of the Sun. The Sun's deeper layers are believed to rotate like a rigid body and significantly faster than the upper layers. At the surface, the equatorial region rotates faster than the poles. A full revolution near the equator takes 26 days, but near the poles it requires 37 days. The shear forces that appear near the bottom of the convection zone at the boundary between the fast rotating interior and the slower outer parts are believed to drive a generator that produces the Sun's magnetic fields.

As was shown in the chapter on plasma physics, the magnetic field lines in a highly conducting plasma can be regarded as *frozen in*, meaning that two plasma volumes that are initially on the same field line will remain on a common field line. If the magnetic energy density, or *magnetic pressure*, $B^2/2\mu_0$ is larger than the pressure nT of the particles, the magnetic field will control the flow of the plasma, but if the particle pressure is larger the plasma motion will determine the evolution of the magnetic field.

The eleven-years *solar cycle* was first observed as a variation in the number of *sunspots*. Sunspots are regions with a very strong vertical magnetic field, up to 0.3 T. They appear as dark spots on the Sun's surface because their temperature is lower than the surroundings. Usually, sunspots appear in pairs or in larger groups. The magnetic field of the leading spot then has the same polarity as the high latitude field in the same hemisphere, while the trailing spot in the pair has a field of opposite polarity. Early in the solar cycle the sunspots appear at high latitudes. The typical lifetime of sunspots is a few solar rotations. New active regions with sunspots and other phenomena will appear equatorward of the older ones. Late in the solar cycle the activity is mainly at low latitudes, but in the middle of the cycle sunspots are found over a wide range of latitudes. During the period of maximum solar activity the large scale magnetic fields in the polar regions become very weak. When they reappear late in the solar cycle, the magnetic poles have changed places with each other, and at the start of the next cycle the polarity of sunspots will be reversed. The variation of the magnetic fields in sunspots and around the solar poles during a solar cycle is shown schematically in Figure 5.2. Notice that the full period of the magnetic field reversal is 22 years, or two solar cycles.

5.3. The Sun's atmosphere

5.3.1. The photosphere

The visible surface of the Sun is called the *photosphere*. It is a rather thin layer, about 500 km, where the Sun becomes opaque to visible light. This is also the layer where the sunlight that we see comes from. In the visible range, the solar spectrum is approximately that of a black-body at a temperature of 5800 K. The continuous spectrum of sunlight is

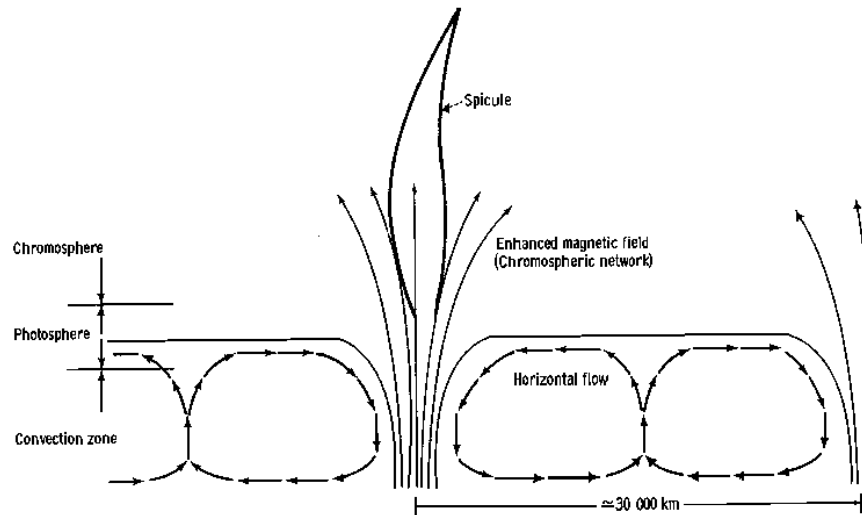


Figure 5.2.: Supergranulation and spicules. The flow of plasma from the center to the edges in supergranulation cells carries magnetic fields to the cell boundaries and a network of enhanced field result. The high magnetic pressure ejects plasma in the form of spicules. Figure from Russell et al. [2016].

partly caused by interaction with free electrons, but particularly at shorter wavelengths the photons are also emitted and absorbed in the formation and dissociation of negative hydrogen ions, H^- . A second electron may loosely attach itself to a hydrogen atom, forming a negative ion with an ionization potential of 0.75 eV.

A fine network of darker matter surrounding brighter patches can be seen in the photosphere. This pattern is called *granulation*. The characteristic size of the granulation pattern is around 1000 km, and the temperature difference between bright and dark areas is typically 100 K. Granulation is created by convection cells that reach about 500 km below the photosphere. These shallow cells may be due to strong absorption by neutral hydrogen near the top of the convection zone. Neutral hydrogen, which has an ionization potential of 13.6 eV, exists only at lower temperatures than neutral helium (25.4 eV) or singly ionized helium (54.1 eV). The heating associated with recombination to ^4He and $^4\text{He}^+$ takes place at lower levels where the temperature is higher. The convection cells driven by this heating reach deep into the convection zone and have a characteristic size of 30 000 km. These cells generate a pattern called *supergranulation*, but this larger pattern is not seen as brightness variations in the photosphere.

5.3.2. The chromosphere

For a few seconds before and after a total eclipse, a bright red flash can be seen from the *chromosphere*, a 2500 km thick layer above the photosphere. This red flash has given the chromosphere its name—chroma is Greek for colour. The base of the chromosphere is defined as the height of minimum temperature, about 4300 K, in the solar atmosphere. Across the chromosphere, the density decreases rapidly from 10^{-5} kg/m^3 to 10^{-10} kg/m^3 while the particle temperature increases to 40 000 K. In the upper part of the chromosphere, the magnetic forces dominate the particle pressure forces. Since the chromosphere is transparent to light except at certain wavelengths, the particles and radiation are not in thermal equilibrium above the photosphere. The red

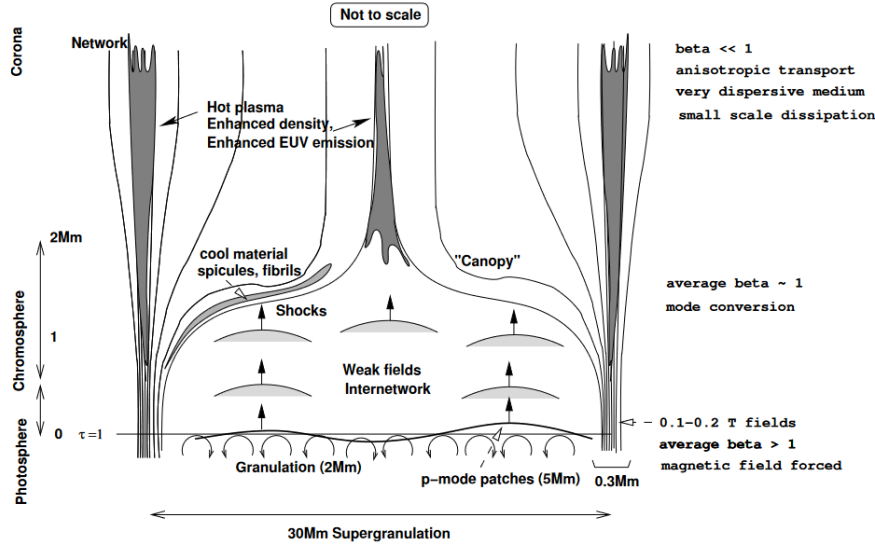


Figure 5.3.: A sketch of the structure of the solar chromosphere showing supergranulation and spicules. Figure from Judge [2006].

colour of the chromosphere comes from the hydrogen $H\alpha$ line at 656.28 nm, which is absorbed and re-emitted in this layer. At this wavelength, the supergranulation pattern and related phenomena can be seen. Using the Doppler effect, a horizontal motion from the center to the edge of a supergranular cell can be detected, and *spicules* are seen at the boundaries between cells. Spicules are jets of outstreaming gas that rises to around 10 000 km with a velocity of about 25 km/s. Since the magnetic field is frozen into the plasma, the outward flow in the supergranular cells will build up strong magnetic fields between the cells in the regions where spicules are formed. While a supergranular cell has a lifetime of around 24 hours, a spicule lasts for only about 10 minutes. The structure of supergranulation and spicules is illustrated in Figure 5.3.

Several phenomena related to solar activity are also seen in emissions from the chromosphere. An *active region*, sometimes called a solar storm, contains sunspots but also chromospheric features called *plages* (the French word for “beaches”) and *prominences*. All these phenomena are related to strong and dynamic magnetic fields. Plages are regions of higher density and temperature in the chromosphere. Plages first appear in the space between several supergranular cells. As the plage expands, sunspots appear and the arch-shaped structures called prominences are formed.

Prominences are really huge vertical sheets of plasma that is much cooler and denser than the surrounding corona. Typical dimensions are 200 Mm long, 50 Mm high, and 6 Mm wide, and they can remain stable for months. Prominences lie across a reversal of the vertical photospheric magnetic field, and are supported by a strong (~ 1 mT) horizontal magnetic field that forms an arch between regions of opposite magnetic polarity. Plage (or active region) prominences are smaller than their quiescent cousins and have stronger magnetic fields ($\lesssim 20$ mT). Sometimes, when the uplifting magnetic pressure becomes stronger than the gravitational force, the prominence becomes unstable. This results in a *coronal mass ejection*, when a magnetic flux tube rises until it erupts, and releases chromospheric plasma into the corona. Erupting prominences are also associated with the very rapid brightening of a spot on the Sun known as a *solar flare*.

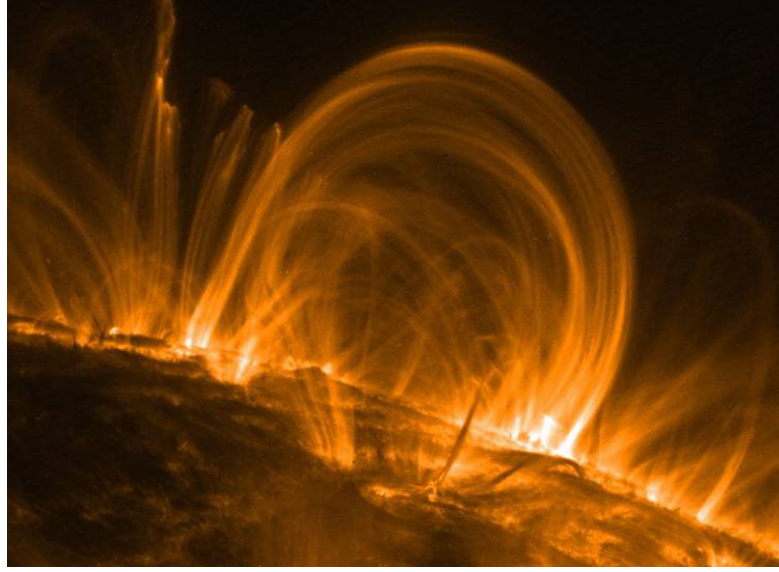


Figure 5.4.: Coronal loops observed at the Sun. Credit: NASA/TRACE

5.3.3. The corona

The *corona* is separated from the chromosphere by a very strong temperature gradient. About 3000 km above the solar surface, the temperature suddenly rises to about 10^6 K. It is still not known how the much cooler layers below can heat the corona to such high temperatures. Current theories involve low frequency magnetic waves, which dissipate their energy in the corona when they are damped, or the action of magnetic reconnection or turbulent heating of the plasma. An optical image of the corona is shown in Figure 5.5. The ray structures seen in this figure reflect the complexity of the coronal magnetic fields. A magnetic field line interpretation of this image is sketched in Figure 5.6. In the corona, the magnetic forces dominate the dynamics, but the magnetic field decreases faster with altitude than the plasma density. At a distance of a few $R_\odot$ the plasma pressure becomes comparable to the magnetic pressure, and the solar wind then takes over. Since the corona may be very inhomogeneous, the altitude at which the transition from corona to solar wind takes place can vary considerably.

5.4. The Solar Wind

The solar atmosphere is different from planetary atmospheres in the sense that it is not in a static equilibrium. From the atmosphere of a planet like the Earth there is no (significant) steady outflow of gas into space. It is in a static equilibrium, where the average vertical gas velocity $u = 0$. Hence, a gradient in the pressure P balances the gravitational force $-\rho MGr^{-2}$ of a planet with mass M . We can eliminate the mass density $\rho = nm$, where n is the number density of ions (or electrons) and m is the average ion mass, from the force balance equation

$$\frac{\partial}{\partial r}P + \frac{\rho MG}{r^2} = 0. \quad (5.1)$$

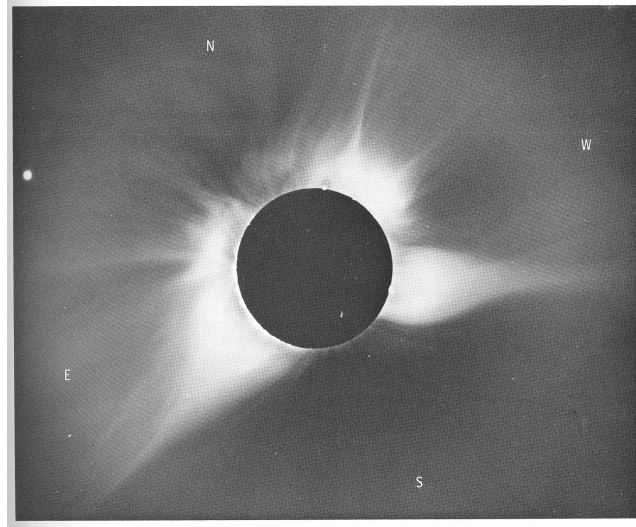


Figure 5.5.: The solar corona observed by G. Newkirk, Jr., at Pulacayo, Bolivia, during an eclipse in 1966. The bright spot to the left is Venus. Figure from Kivelson and Russell [1995b].

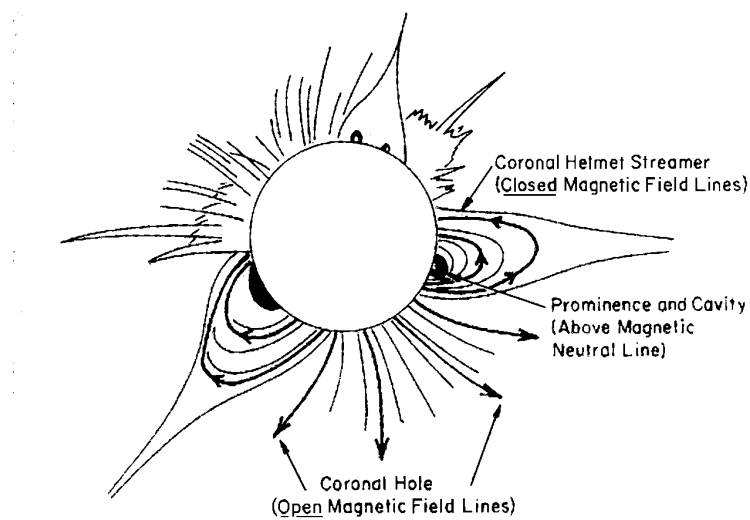


Figure 5.6.: A sketch of a magnetic field configuration consistent with the image in Figure 5.5. Figure from Kivelson and Russell [1995b].

5. The Sun and the Solar Wind

Since ions and electrons both contribute to the pressure we have $P = 2nk_B T$, and inserting $\rho = mP/2k_B T$ we find

$$\frac{1}{P} \frac{\partial}{\partial r} P = -\frac{mMG}{2k_B T r^2}. \quad (5.2)$$

This equation has solutions of the form

$$P(r) = P_0 \exp \left[\frac{mMG}{2k_B T} \left(\frac{1}{r} - \frac{1}{R} \right) \right], \quad (5.3)$$

where we have adopted the definition that the pressure is P_0 when $r = R$ at the bottom of the atmosphere. Very far from the planet, when $r \gg R$, the pressure approaches $P_\infty = P_0 \exp[-mMG/2k_B TR]$, which for a small and cold planet is low enough to match the interplanetary pressure. However, using parameters characteristic of the solar corona one finds $P_\infty \approx 2 \cdot 10^{-6}$ Pa, but the interstellar pressure is less than 10^{-13} Pa. Hence, a static ($u = 0$) model does not work for the solar atmosphere.

To allow the coronal pressure of 5 mPa at $r \approx R_\odot$ to fall to 10^{-13} Pa far from the Sun we must include a finite radial velocity u , a *solar wind*, in our model. Assuming that the flow is stationary in time ($\frac{\partial}{\partial t} u = 0$), the acceleration of the radial flow is given by

$$\rho \frac{d}{dt} u(r(t)) = \rho \frac{\partial r}{\partial t} \frac{\partial}{\partial r} u = \rho u \frac{\partial}{\partial r} u = -\frac{\partial}{\partial r} P - \frac{\rho M_\odot G}{r^2}. \quad (5.4)$$

For simplicity we will assume that the temperature is constant, and after dividing by the density we rewrite Equation 5.4 as

$$u \frac{\partial}{\partial r} u = -\frac{2k_B T}{m} \frac{\frac{\partial}{\partial r} n}{n} - \frac{M_\odot G}{r^2}. \quad (5.5)$$

Noticing that in a stationary state the same number N of particles per second must flow across each spherical surface around the Sun, we conclude that $N = 4\pi r^2 n u$ is a constant, independent of r . Using that $\frac{\partial}{\partial r} N = 0$ and $n = N/(4\pi r^2 u)$ to evaluate $\frac{\partial}{\partial r} n$ we find

$$\frac{1}{n} \frac{\partial}{\partial r} n = -\frac{2}{r} - \frac{1}{u} \frac{\partial}{\partial r} u. \quad (5.6)$$

Inserting this in Equation 5.5, we obtain

$$(u^2 - v_t^2) \frac{1}{u} \frac{\partial}{\partial r} u = \frac{2v_t^2}{r} - \frac{M_\odot G}{r^2}, \quad (5.7)$$

where $v_t^2 = 2k_B T/m$ is the thermal velocity of the ions. At the base of the corona, the flow velocity will be less than the thermal velocity, i.e. $u^2 < v_t^2$. Inserting typical numbers one finds $2v_t^2 < M_\odot G/R_\odot$, and the right hand side is thus negative in the corona. This implies that we must have $u^{-1} \frac{\partial}{\partial r} u > 0$, and the velocity increases as we go further from the Sun. At a *critical radius* $r_c = M_\odot G/2v_t^2$, the right hand side of Equation 5.7 changes sign. Then, the left hand side must also be zero when $r = r_c$. One possibility is that u remains smaller than v_t and reaches a maximum so that $\frac{\partial}{\partial r} u = 0$ at $r = r_c$. The velocity then starts to decrease, and at large r where the velocity goes to zero the solution is similar to the static case. The other possibility is that $u = v_t$ at $r = r_c$, while $\frac{\partial}{\partial r} u$ remains positive, and the velocity continues to grow. The sound speed in an isothermal plasma equals the thermal speed, so at this point the flow becomes *supersonic*, and propagates faster than sound. Since $n = N/(4\pi r^2 u)$, it follows that in this case the

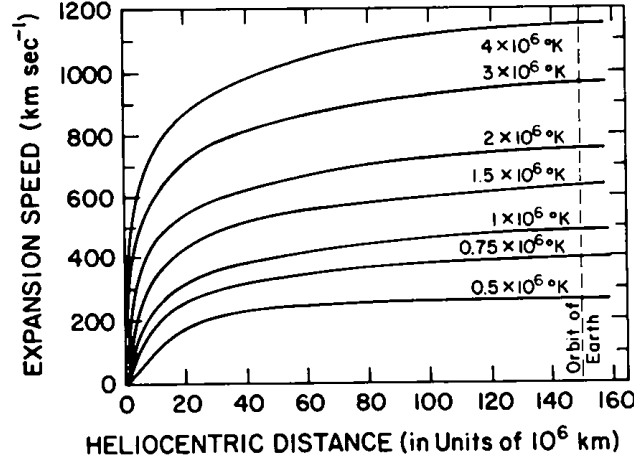


Figure 5.7.: Radial expansion speed $u(r)$ derived from isothermal solar wind expansion models with varying coronal temperatures. Figure from Kivelson and Russell [1995b].

density, and the pressure, will approach zero as $r \rightarrow \infty$. Solutions of this type, with a solar wind that becomes faster far from the Sun, can represent an equilibrium state connecting the high pressure corona with the very low pressure interstellar medium.

The nonlinear differential equation Equation 5.7 cannot be solved explicitly by elementary methods, but it can easily be integrated term by term to obtain

$$\frac{u^2}{2} - v_t^2 \ln u = 2v_t^2 \ln r + \frac{M_\odot G}{r} + K. \quad (5.8)$$

When the integration constant K is chosen so that $u(r_c) = v_t$, we obtain the solar wind solution

$$\frac{u^2}{v_t^2} - \ln \left(\frac{u^2}{v_t^2} \right) = 4 \left(\frac{r_c}{r} - \ln \left(\frac{r_c}{r} \right) \right) - 3, \quad (5.9)$$

which determines the solar wind speed for a given coronal temperature. Notice that for large r , the right hand side goes to infinity as $\ln(r_c/r) \rightarrow -\infty$. If $u > v_t$ the first term on the left hand side goes to infinity as u increases for large r , but if $u < v_t$ the second term dominates and $-\ln(u^2/v_t^2)$ goes to infinity as u decreases with r . Figure 5.7 shows the numerical solution $u(r)$ of Equation 5.9 for a range of coronal temperatures.

In the real solar wind, the temperature is not constant but decreases with r . This leads to somewhat lower velocities at large distances from the Sun, but at least as long as $T(r)$ decreases slower than $1/r$ the basic nature of the solution is not changed. The observed solar wind speed at 1 AU ($= 1.496 \cdot 10^{11}$ m) is typically 450 km/s, but varies between 200 km/s and 1000 km/s. The fast solar wind comes from *coronal holes*, or regions with open magnetic field lines (see Figure 5.6). The plasma has a temperature of about 10 eV (10^5 K), and an electron density of $7 \cdot 10^6 \text{ m}^{-3}$. Most of the ions are protons, mixed with about 5 % He^{2+} . As might be expected, the solar wind properties are strongly affected by various forms of solar activity.

The Sun's magnetic field extends also into the solar wind. As discussed earlier, the magnetic field can be regarded as frozen into the plasma. Since the Sun rotates, plasma that has been ejected radially outwards from a certain point on the equator of the Sun

5. The Sun and the Solar Wind

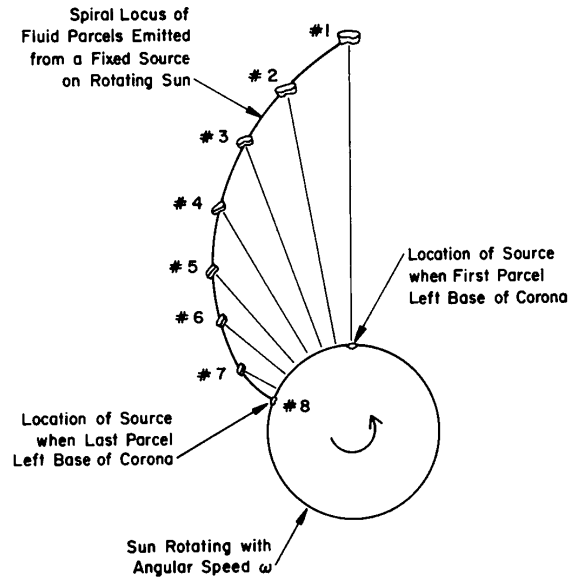


Figure 5.8.: Position of a succession of plasma parcels emitted at constant speed from a source fixed on the rotating Sun. Figure from Kivelson and Russell [1995b].

will later be found along a spiral as illustrated in Figure 5.8. At the Earth's orbit, the angle between the magnetic field line and the direction to the Sun is about 45° . The magnetic field strength at 1 AU is about 7 nT, but the polarity depends on the polarity of the coronal field. The density at 1 AU is about 5 cm^{-3} and the Temperature is 10 eV. Since the solar magnetic field is highly variable, there will significant variations superimposed on this basic magnetic structure of the solar wind.

6. The Magnetosphere

6.1. Introduction

The field lines of a magnetic dipole in vacuum extend out to infinite distances. The geomagnetic field is very similar to a dipolar field out to a few R_E from Earth, and it may not be obvious why the field lines further out are confined almost completely to the region called the magnetosphere. The reason is that electric currents flow at the magnetopause – the boundary of the magnetosphere. These currents create magnetic fields that bend the field lines, and keep the geomagnetic field separated from the interplanetary magnetic field. Since the magnetic fields are frozen-in to the plasma, the magnetopause is also a boundary between the magnetospheric plasma and the solar wind. Although the solar wind cannot (generally) flow directly into the magnetosphere, the energy and momentum supplied by the solar wind is nevertheless the main source of magnetospheric dynamics.

A schematic diagram of Earth’s magnetosphere in the noon–midnight plane is shown in Figure 6.1. The basic particle and magnetic field features in this figure are representative also of other planetary magnetospheres, although the details can be different. In the following sections, we will discuss how the features shown in Figure 6.1 result from the interaction of the solar wind with the geomagnetic field.

6.2. Coordinate systems

Space physicists have developed a range of coordinate systems that are used to analyze data and describe different regions in the magnetosphere. To be able to orient ourselves for the following discussion, we will briefly introduce two of these coordinate systems; the Geocentric Solar Ecliptic (GSE) coordinate system and the Geocentric Solar Magnetospheric (GSM) coordinate system.

The GSE coordinate system is defined as follows. Firstly, the x -axis is directed from Earth’s center towards the sun. The z -axis is perpendicular to the plane of Earth’s orbit around the sun and its positive towards north. The y -axis finally completes a right-handed coordinate system and is therefore pointing towards dusk. The GSE coordinate system is illustrated in 6.2. The GSE coordinate system is often used to describe the magnetosphere and its boundaries.

The GSM coordinate meanwhile has its x -axis also towards the Sun, as for the GSE coordinate system. The z -axis is the projection of the Earth’s magnetic dipole axis on to the plane perpendicular to x -axis. The y -axis completes the right hand coordinate system. The conversion from the GSM into the GSE coordinate system is therefore a rotation around the x -axis. Which coordinate system is most appropriate to use will depend on the specifics of the problem in question, i.e. whether it is more valuable to have a ‘solid’ reference for the flow of the solar wind, or for the orientation of the Earth’s

6. The Magnetosphere

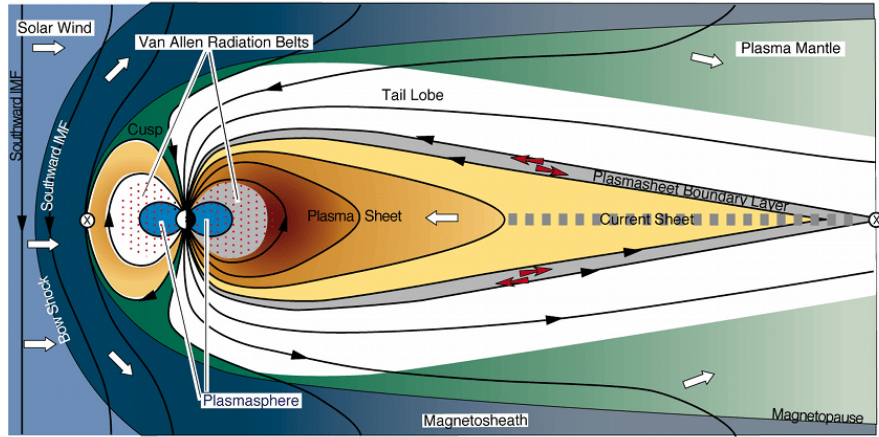


Figure 6.1.: A schematic of Earth's magnetosphere. The solar wind flowing in from the right side interacts with the Earth's magnetic field. The intensity of the color illustrated the density and the warmth of the color illustrates the temperature of the respective region. Credits: T. W. Hill.

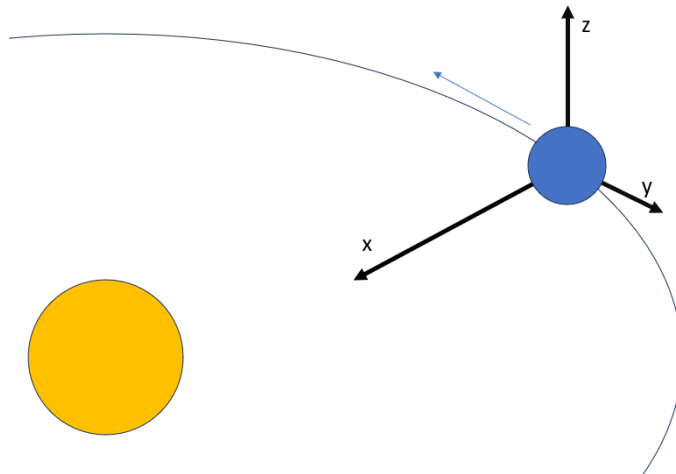


Figure 6.2.: The GSE coordinate system.

dipolar magnetic field.

6.3. The Earth magnetic field

Near the surface of Earth, the magnetic field resembles the field generated by a magnetic dipole pointing to the south at the center of Earth. More accurately, the dipole should be tilted by 11° and displaced from the Earth's center by 342 km, so that the dipole axis cuts the surface of Earth at 80.1°N , 161.8°E , but here we will neglect these details.

The magnetic field $\mathbf{B}$ from a dipole can be obtained from a scalar magnetic potential

$$\Psi(\mathbf{r}) = \frac{\mu_0}{4\pi} \frac{\mathbf{m} \cdot \mathbf{r}}{r^3}, \quad (6.1)$$

where $\mathbf{m}$ is the magnetic moment of the dipole. Introducing spherical coordinates (r, θ, ϕ)

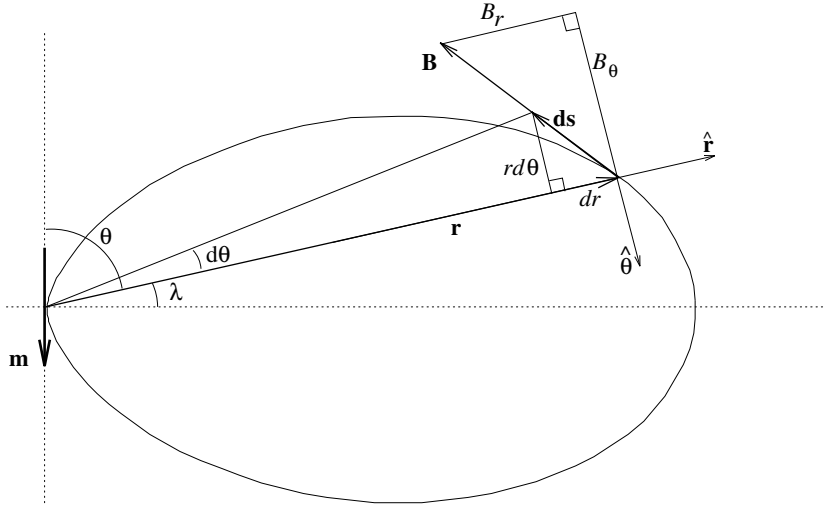


Figure 6.3.: Geometry of a geomagnetic field line. Figure from Fälthammar [2001].

and taking $\mathbf{m} = -m\hat{\mathbf{z}}$ (since the North pole really is a magnetic south pole), we can from $\mathbf{B} = \nabla\Psi$ obtain the magnetic field as

$$\begin{aligned}\mathbf{B} &= \frac{\mu_0 m}{4\pi} \left(\hat{\mathbf{r}} \frac{\partial}{\partial r} + \hat{\theta} r^{-1} \frac{\partial}{\partial \theta} + \hat{\phi} r^{-1} \frac{\partial}{\partial \phi} \right) r^{-2} \cos \theta \\ &= -\frac{\mu_0 m}{4\pi r^3} (2\hat{\mathbf{r}} \cos \theta + \hat{\theta} \sin \theta).\end{aligned}\quad (6.2)$$

We know that at the equator, where $r = R_E$ and $\theta = \pi/2$, the magnetic field is $B_0 = 31\mu\text{T} = \frac{\mu_0 m}{4\pi R_E^3}$. Using this we can eliminate $\mathbf{m}$ to find

$$\mathbf{B}(r, \theta) = -B_0 \left(\frac{R_E}{r} \right)^3 (2\hat{\mathbf{r}} \cos \theta + \hat{\theta} \sin \theta). \quad (6.3)$$

The shape of the field line is illustrated in Figure 6.3, where the polar coordinates are also shown. The magnetic field lines are curves that at every point are tangent to the local magnetic field vector. If $d\mathbf{s}$ is an infinitesimal vector along the field line we thus demand $d\mathbf{s} \propto \mathbf{B}$. In polar coordinates we have $d\mathbf{s} = \hat{\mathbf{r}} dr + \hat{\theta} r d\theta$, and $d\mathbf{s}$ will, as shown in Figure 6.3, be parallel to $\mathbf{B}$ if

$$\frac{dr}{r d\theta} = \frac{B_r}{B_\theta} = \frac{2 \cos \theta}{\sin \theta}. \quad (6.4)$$

This differential equation may be written

$$\frac{1}{r} \frac{d}{d\theta} r = \frac{1}{\sin^2 \theta} 2 \sin \theta \cos \theta = \frac{1}{\sin^2 \theta} \frac{d}{d\theta} \sin^2 \theta. \quad (6.5)$$

We can easily integrate this to find that r and θ along a field line are related by

$$r = r(\theta) = r_0 \sin^2 \theta, \quad (6.6)$$

where the constant of integration r_0 , which corresponds to the distance at which the field line crosses the equatorial plane, often is written as $r_0 = LR_E$. (This dimensionless

6. The Magnetosphere

quantity L is used to define or specify the “L-shell” wholly containing particular dipolar field line, or particle moving upon this surface.) We can also replace the colatitude θ by the latitude $\lambda = \pi/2 - \theta$ in the formulas above. The equation for the field line is then

$$r(\lambda) = LR_E \cos^2 \lambda. \quad (6.7)$$

A magnetic field line that crosses the equator at $r = LR_E$ will reach the surface of Earth ($r = R_E$) at a latitude determined by $R_E = LR_E \cos^2 \lambda$, or $\cos^2 \lambda = L^{-1}$. When relating satellite measurements in the magnetosphere to ground observations, it is often convenient to use the invariant (magnetic) latitude Λ , defined by

$$\Lambda = \arccos(L^{-1/2}). \quad (6.8)$$

An auroral phenomenon at latitude Λ in the ionosphere is then by a magnetic field line connected to the equatorial plane at a distance LR_E . Notice that the invariant latitude is the same all along a field line. Since the particles travel along magnetic field lines (*frozen-in* theorem), it is important to know if two observations are on the same field line.

Substituting $\theta = \pi/2 - \lambda$ and Equation 6.7 in Equation 6.2, we can express the $\mathbf{B}$ -field in terms of L and λ as

$$\mathbf{B}(L, \lambda) = -B_0 \frac{\hat{\mathbf{r}} 2 \sin \lambda + \hat{\boldsymbol{\theta}} \cos \lambda}{L^3 \cos^6 \lambda}. \quad (6.9)$$

The magnitude of the magnetic field is $B = \sqrt{B_r^2 + B_\lambda^2}$ ($B_\phi = 0$ because of symmetry), and the magnetic field strength is thus

$$B(L, \lambda) = B_0 \frac{(4 \sin^2 \lambda + \cos^2 \lambda)^{1/2}}{L^3 \cos^6 \lambda} = B_0 \frac{(1 + 3 \sin^2 \lambda)^{1/2}}{L^3 \cos^6 \lambda}. \quad (6.10)$$

When this equation is combined with Equation 2.35 we can calculate how the pitch angle of a particle changes as λ varies along the orbit (Remember that L is constant along the field line!), and from Equation 6.11 we can find the λ dependence of the loss cone.

The real geomagnetic field is not a pure dipole field, even if we allow the dipole to be tilted and displaced from the center of Earth. However, the dipole model is often useful as a first approximation out to distances of six or seven R_E . As we go to higher altitudes, the influence of magnetospheric currents becomes more important, and the deviations from a dipole field increase. In the outer magnetosphere, near the magnetopause and in the magnetic tail, the topology may even be different from the dipole model, since some field lines connect to the interplanetary magnetic field.

6.3.1. Particle motion in the Earth magnetic field

The following discussion is based on the particle motion that was discussed in chapter 2. In Earth’s magnetic field, a particle moving in a spiral along a field line will encounter a stronger field when it approaches one of the poles. As B increases, the pitch angle must also increase until $\alpha = 90^\circ$. At this point the particle is reflected, and moves up again, away from the pole to weaker fields. A sufficiently strong field acts as a magnetic mirror for the particle. After passing through the equatorial region, the particle will be reflected in the same way in the other hemisphere.

The orbit resulting from this behaviour is illustrated in Figure 6.4. However, if the reflection point is too close to Earth, below about 100 km altitude, the particle will be

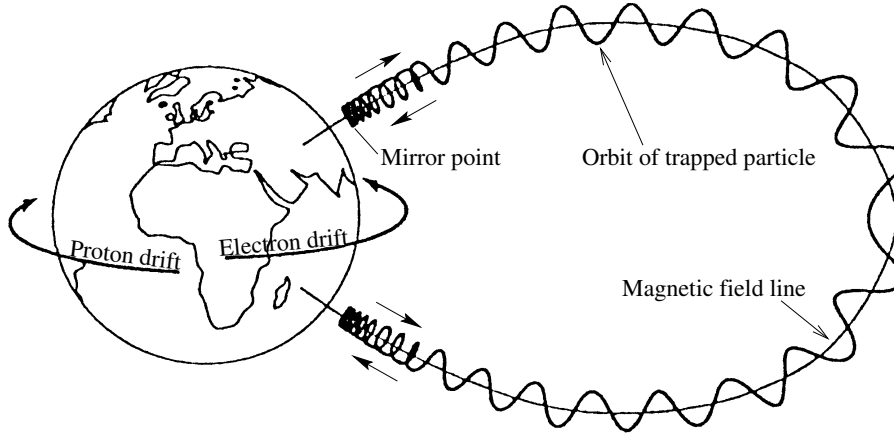


Figure 6.4.: The orbit of a magnetically trapped particle in the geomagnetic field. Figure from Fälthammar [2001].

knocked out of its orbit by collisions with atmospheric molecules and atoms. Energetic (energy $\gtrsim 10$ eV) charged particles that go down into the dense atmosphere below 100 km will normally not come back. Their energy goes mainly into heating and ionizing the topside atmosphere, but some is also converted to visible light in auroras. The upgoing particles at 100 km have an energy distribution determined by the temperature at this altitude, which typically is ~ 0.1 eV ($1 \text{ eV} = 11\,605 \text{ K}$). Hence, there will be very few particles with energy ≥ 10 eV coming up at altitudes just above 100 km—all the energetic particles are either going down or are very close to their reflection point.

Consider a particle at an altitude of 100 km which just has been reflected and has a pitch angle $\alpha = 90^\circ$. If the energy of this particle is W , its magnetic moment is $\mu = W/B_I$, where B_I is the magnetic field strength in the ionosphere. The particle will be pushed up along the field line by the force given in Equation 2.24. The orbit will change from a circle to an upward spiral as the parallel velocity increases. From Equation 2.33b we see that this corresponds to a decrease in the pitch angle. (We assume that the particle started above the southern hemisphere where the magnetic field vector is upward.) At a point $\mathbf{r}$ along the orbit where the magnetic field is $\mathbf{B}(\mathbf{r})$ we can calculate the pitch angle α_{LC} of a particle that is reflected at 100 km altitude from Equation 2.34 as

$$\sin^2 \alpha_{LC} = \frac{B(\mathbf{r})}{B_I}. \quad (6.11)$$

The pitch angle α_{LC} is the smallest pitch angle particles can have if they are reflected above the atmosphere. Particles with velocity vectors within the *loss cone* defined by $\alpha < \alpha_{LC}$ will be lost into the atmosphere, and under stationary conditions we expect to find only a few low energy particles in the loss cone. On the other hand, particles outside the loss cone ($\alpha > \alpha_{LC}$) will be trapped within the magnetosphere.

6.3.2. Summary of particle motion in the magnetosphere

In Earth's magnetosphere, we can find examples of all the effects discussed in chapter 2.

In a dipole field, we have three types of periodic motion, as illustrated in Figure 6.5. All particles will gyrate around the field lines on a time scale T_c that typically is in the range 10^{-6} s to 1 s. Electrons in the ionosphere have the shortest gyroperiods, and ions

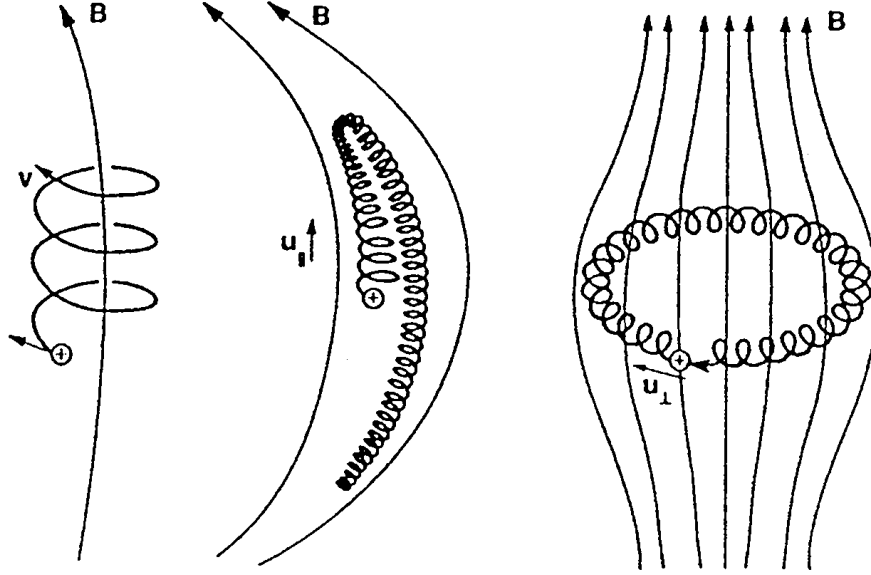


Figure 6.5.: Three types of periodic motion: Gyration, magnetic mirroring and azimuthal drift. Figure from Fälthammar [2001].

in the outer magnetosphere have the longest. Trapped particles, for example those that are outside the loss cone, will also bounce back and forth between the magnetic mirrors created by the strong fields near the north and south pole. The bounce period is roughly given by $T_b \approx 4R/v$, where R is the distance to where the orbit crosses the equatorial plane, and v is the velocity of the particle. An electron with an energy 300 eV will, for example, have $v \sim 10^7$ m/s and if it crosses the equatorial plane at $R = 6R_E \sim 4 \cdot 10^7$ m we find $T_b \sim 16$ s. The bounce period of an ion with the same energy will be more than 10 minutes.

The azimuthal drift resulting from the effects discussed in section 2.3 and subsection 2.4.1 can be written

$$\mathbf{v}_d = \frac{\mathbf{E} \times \mathbf{B}}{B^2} + \frac{\mathbf{F}_g \times \mathbf{B}}{qB^2} - \mu \frac{\nabla B \times \mathbf{B}}{qB^2}. \quad (6.12)$$

The relative importance of the three terms in Equation 6.12 varies from case to case, as illustrated in some of the exercises. The second term, where $\mathbf{F}_g$ denotes the gravitational force, is usually small in the magnetosphere. The first, $\mathbf{E} \times \mathbf{B}$ term, is responsible for much of the plasma transport in the magnetosphere. Currents driven by the third, $\nabla B \times \mathbf{B}$ term create magnetic field perturbations that cause "magnetic storms" in the magnetosphere. In some regions of the magnetosphere, the particles may drift all around the Earth. The time scale, T_d , associated with this drift may be from hours to days.

6.4. The bow shock and the magnetosheath

After discussing the Earth magnetic field and particle motion in such, we will consider the interaction of the solar wind and the Earth magnetic field.

The speed of the solar wind at Earth's orbit is about 450 km/s, which is higher than the the speed of sound in the medium. When an obstacle such as Earth's magnetosphere is placed in a flowing gas, the flow is diverted around the obstacle by changes in the pressure

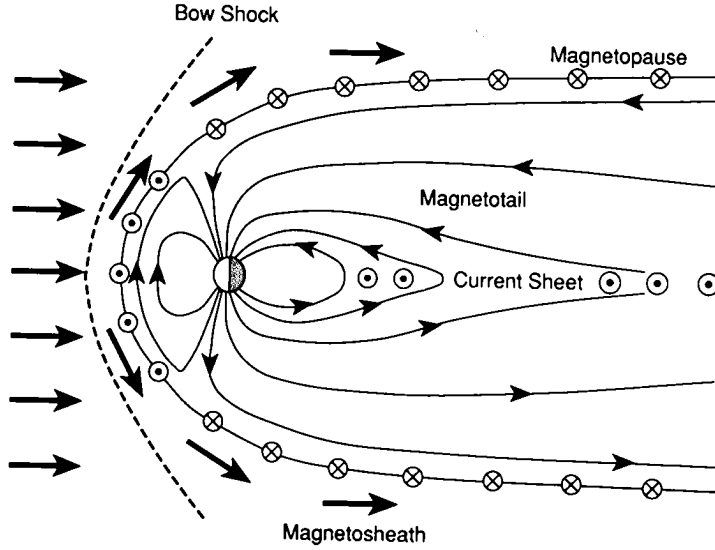


Figure 6.6.: Cross section of the simplest model of the magnetosphere in the noon-midnight meridian. The solar wind flow (thick arrows) is deflected at the bow shock and flows around the magnetosphere, forming the magnetosheath. The sheet currents at the magnetopause separate the interplanetary magnetic field from the magnetosphere. Figure from Kivelson and Russell [1995b].

and velocity, which are propagated with the sound speed. When the flow is supersonic, these changes cannot propagate upstream, and the solar wind gets no “warning” that the magnetosphere is in its way. Solar wind plasma that has been stopped will start to pile up against the magnetopause. When the density increases, more matter can be removed by subsonic flows. The compressed plasma will also be heated, leading to a higher sound velocity. In a sufficiently thick layer with enhanced density and temperature, subsonic flows can remove all the inflowing solar wind plasma. At the boundary between the supersonic solar wind and the subsonic flows of compressed plasma, a shock wave, called the *bow shock*, is formed. The region between the bow shock and the magnetopause is called the *magnetosheath*, and it is characterized by irregular subsonic flows and enhanced densities, temperatures, and magnetic fields when compared to the solar wind. Near the sunward nose the magnetosheath is just a few R_E thick, the density is high, and the flows are mostly towards the flanks. As we approach the tail, the magnetosheath widens, the density drops, and the velocity becomes more aligned with the solar wind as the shock becomes weaker. A rough sketch of the solar wind flow in the magnetosheath is given in Figure 6.6.

At the bow shock, the flow velocity jumps from U_{SW} in the solar wind to U_{MS} in the magnetosheath, and similarly the density jumps from ρ_{SW} to ρ_{MS} . If the shock is stationary the flux on either side must be the same, which requires (assuming that U is along the shock normal) $\rho_{SW}U_{SW} = \rho_{MS}U_{MS}$. Other similar relations may be derived for the conservation of momentum and energy at the shock, and these relations are the basis of a substantial theory about shock waves. Although shocks are common and important in space plasmas, we will not develop this theory further here.

The bow shock presents distinct characteristics depending on the angle θ_{Bn} between the local bow shock normal and the Interplanetary Magnetic Field (IMF), often leading to its

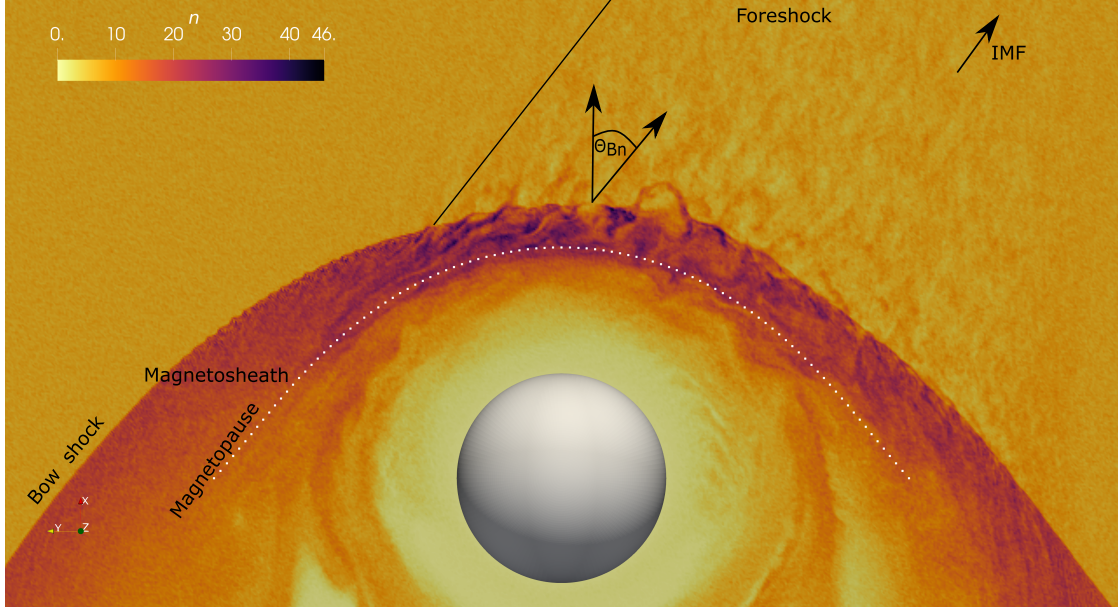


Figure 6.7.: Illustration of the quasi-perpendicular bow shock and the quasi-parallel bow shock. Particles are reflected at the quasi-parallel bow shock forming the foreshock, seen as a turbulent region upstream of the bow shock. Figure created with AMITIS simulation data [Fatemi et al., 2017].

division into two regions. The bow shock is considered quasi-parallel when $\theta_{Bn} < 45$ deg and quasi-perpendicular when $\theta_{Bn} > 45$ deg. The two different bow shock regions the depicted in Figure 6.7. When the bow shock is quasi-parallel, incoming solar wind particles undergo reflection and subsequently gyrate back into the solar wind. This process gives rise to the growth of instabilities and waves. The region upstream of the bow shock is known as the foreshock. As a result, the foreshock becomes a relatively turbulent area, comprising two distinct particle populations: the backstreaming particles and the incoming solar wind. Conversely, when the bow shock is quasi-perpendicular, particles can still be reflected, but they do not propagate as far upstream into the solar wind, leading to differing behavior compared to the quasi-parallel scenario.

Transients, i.e. short-duration and small-scale features arising at the bow shock, do have the capability to traverse the magnetosheath and interact with the magnetopause. As a result of these interactions, both local and global disruptions in the magnetopause occur, leading to observable disturbances at the ground and ionosphere.

6.5. The magnetopause

When we traverse the bow shock, there is a distinct change in the properties of the plasma and magnetic field, but the flow of solar wind plasma and the interplanetary magnetic field lines continue through the bow shock. The *magnetopause* is a rather different kind of boundary in space. The flow of solar wind plasma through the magnetopause is very limited, and the field lines on the magnetospheric side are mainly connected to the geomagnetic field.

We can roughly estimate the distance to the magnetopause by using the magnetic pressure concept. Near the Sun-Earth line in the equatorial plane, we can from Equation 6.10

estimate the geomagnetic field by $B_E(L) = B_0 L^{-3}$ where $B_0 = 31 \mu\text{T}$ is the equatorial field at $1 R_E$, and the distance to Earth is $r = LR_E$. The solar wind velocity is about 450 km/s, which is more than ten times the thermal speed of a proton with a temperature of 10 eV, characteristic of the solar wind. In this rough estimate we neglect the thermal pressure, as well as the interplanetary magnetic field. Using Equation 3.54, we can neglect the variation of the field strength along the field line near the equatorial plane, i.e. $(\mathbf{B} \cdot \nabla)\mathbf{B} \approx 0$. If we assume a stationary situation, and replace the convective derivative $\frac{d}{dt}\mathbf{U}$ by $\mathbf{U} \cdot \nabla\mathbf{U}$, we find the magnetic pressure needed to stop the solar wind flow by integrating the equation

$$\rho_{sw} \mathbf{U} \cdot \nabla\mathbf{U} = -\nabla \frac{B^2}{2\mu_0}, \quad (6.13)$$

from a point beyond the bow shock where the solar wind is unperturbed and $\mathbf{B} = 0$ to the stagnation point at r_s where $\mathbf{U}(r_s) = 0$ and the flow stops. This leads to

$$\frac{\rho_{sw} U_{sw}^2}{2} = \frac{B_E^2(r_s)}{2\mu_0} = \frac{B_0^2}{2\mu_0 L_s^6}, \quad (6.14)$$

which we can solve for $L_s = r_s/R_E$ and obtain a rough estimate of the magnetopause distance in Earth radii as

$$L_s = (\mu_0 \rho_{sw})^{-1/6} \left(\frac{B_0}{U_{sw}} \right)^{1/3} \sim 8, \quad (6.15)$$

where we have used $\rho_{sw} = 1.27 \cdot 10^{-20} \text{ kg/m}^3$ for the solar wind density. This estimate is a little less than the normally observed value, mainly because we have neglected the magnetopause current sheet. Neglecting any currents in space, the geomagnetic field at $8 R_E$ would be $B_0/8^3$, which is an order of magnitude larger than the interplanetary magnetic field. A current sheet that cancels this field outside the magnetopause will double the magnetic field strength on the inside. Inclusion of this effect adds a factor of $2^{1/3}$ to Equation 6.15 and gives $r_s \sim 10 R_E$, which is close to the distance typically observed. This rough calculation can be refined by adding the effects of kinetic pressure, interplanetary magnetic field, geometry of the magnetopause, diversion of the flow at the bow shock, *etc*, but the estimated distance to the magnetopause does not change much. The variations in the magnetopause distance due to changing solar wind conditions are much more significant.

The mechanism generating the *magnetopause currents* shown in Figure 6.6 is essentially an MHD generator. To see the similarities between the above discussion about pressures and the MHD generator, we may use Ampère's law. Inserting $\mu_0 \mathbf{j} = \nabla \times \mathbf{B}$ in Equation 3.57 and neglecting ∇P , we can recover Equation 6.13 since $(\nabla \times \mathbf{B}) \times \mathbf{B} = -\nabla B^2/2$ at the nose of the magnetopause. Integrating Ampère's law across the magnetopause, which we for simplicity assume to be an infinite plane sheet, we can also estimate the current required to create the magnetic field of $B_0/10^3 = 31 \text{ nT}$ which cancels the geomagnetic field outside the magnetopause. We then find that the current per unit length along the meridian is $I_{MP} = B_E/\mu_0 \sim 0.025 \text{ A/m}$, flowing from the morning to the evening side. Looking at Figure 6.6, we see that the region of dawn to dusk current on the sunward side has a width of at least $10 R_E$, so we can estimate the total dayside magnetopause current to be about 1.6 MA.

Observations have shown that the magnetopause is less than a thousand kilometers thick, and the flow velocity changes by several hundred km/s over this distance. Recalling

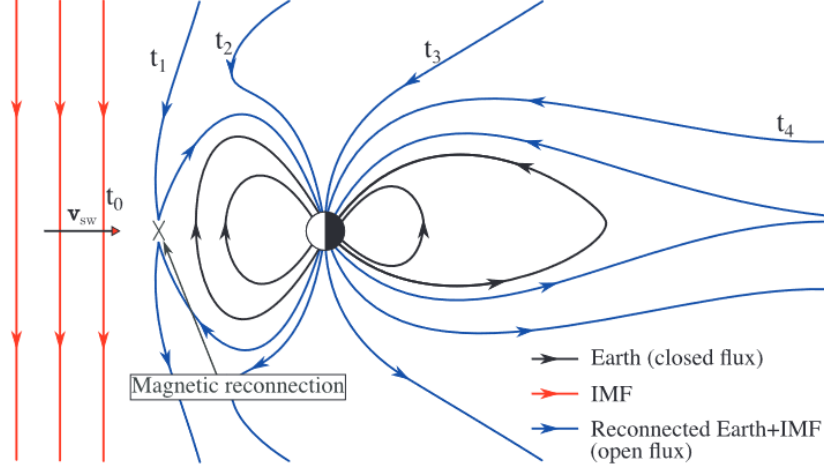


Figure 6.8.: Magnetic field lines in the reconnection (merging) region at the magnetopause. The apparent motion of the reconnected field lines (blue) are dragged from the magnetopause (t_0, t_1) into the magnetotail $t_2 - t_4$. Figure from Steinvall [2022].

the discussion about the validity of Equation 3.39, we may suspect that the ideal MHD approximation breaks down at the magnetopause. Considering a stationary flow that is stopped by the magnetopause, we can combine equations Equation 3.60 and Equation 3.37 to find

$$\mathbf{E} + \mathbf{U} \times \mathbf{B} = \frac{m_i}{e} \mathbf{U} \cdot \nabla \mathbf{U} + \frac{\nabla P_i}{ne}, \quad (6.16)$$

which indicates that strong braking of the flow causes deviations from the ideal MHD condition $\mathbf{E} + \mathbf{U} \times \mathbf{B} = 0$. Indeed, it is found that the magnetic field is not perfectly frozen-in under these conditions, and there is a plasma flux through the magnetopause. This process is known as *magnetic reconnection* (for more details see subsection 3.4.8). Magnetic reconnection is most effective if the IMF is oriented southward. This process transfers momentum and energy from the solar wind into the magnetosphere. The process is sketched in Figure 6.8.

6.6. The magnetotail and magnetospheric convection

As a result of the interaction with the solar wind, the magnetosphere is compressed on the dayside and an extended *magnetotail* is formed on the nightside. The three-dimensional geometry of the magnetopause currents is shown in Figure 6.9. In the tail, the magnetopause currents are flowing from the dusk to the dawn side, both above and below the magnetosphere. Close to Earth these currents are connected to magnetopause currents on the dayside, but beyond about $10 R_E$ downtail from Earth the return current flows from dawn to dusk in the equatorial plane. When combined with the magnetopause currents, this *crosstail current sheet* creates an Earthward magnetic field in the north *tail lobe* and a tailward field in the south. Near the magnetic equator the field is very weak, and the plane where only a small northward magnetic field remains is called the *neutral sheet*. The radius of the distant tail is about $30 R_E$, and the length of the magnetotail has been estimated to $600 R_E$.

The solar wind will, through magnetic reconnection, cause the plasma to circulate

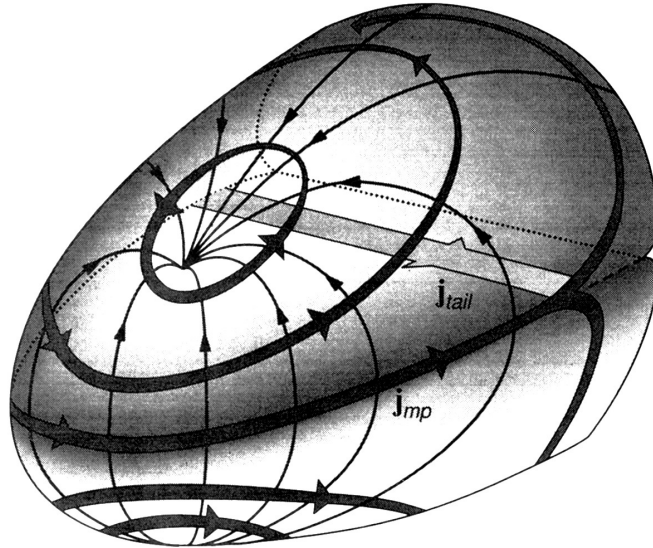


Figure 6.9.: Sketch showing the geometry of the magnetopause currents. Figure from Baumjohann and Treumann [1996].

within the magnetosphere. This circulation is called *magnetospheric convection*. In the *boundary layer* just inside the magnetopause the flow is mainly tailward, as indicated in Figure 6.10. This layer is also known as the *plasma mantle*. The plasma then flows equatorward through the tail lobes, and forms a rather hot and dense *plasma sheet* around the magnetic neutral sheet. This equatorward transport of plasma and its frozen-in magnetic field will amplify the fields on either side of the neutral sheet, and the density of the neutral sheet current must increase. The neutral sheet current and the northward component of $\mathbf{B}$ provide a force that pushes the plasma towards Earth. At some stage ideal MHD breaks down, and the magnetic field lines start to reconnect through the neutral sheet, as indicated in Figure 6.10. The reconnection is bursty and local, but in general reconnection in the tail is much more common during periods of southward IMF when the magnetospheric convection is enhanced. From the neutral line in the tail, the equatorial plasma flows both tailward and Earthward with speeds of several hundred km/s. At the same time, the magnetic field in the near tail becomes more dipolar. The onset of reconnection in the tail is closely related to magnetic substorms, which will be discussed in greater detail later.

Notice that reconnection of the anti-parallel magnetic fields at the neutral sheet converts magnetic field energy into kinetic energy. In effect it is a plasma engine that drives bursts of fast flows in the plasma sheet. The Earthward flowing plasma will eventually be diverted towards the outer parts of the magnetosphere, where it enters the boundary layer and circulates back to the dayside.

The Sunward convection of plasma inside the boundary layer is, for an observer in the Earth's reference frame, equivalent to an electric field $\mathbf{E} = -\mathbf{U} \times \mathbf{B}$, directed from dawn to dusk, parallel to the neutral sheet current. Integrating this electric field across the tail, the potential difference is about 50–100 kV.

6. The Magnetosphere

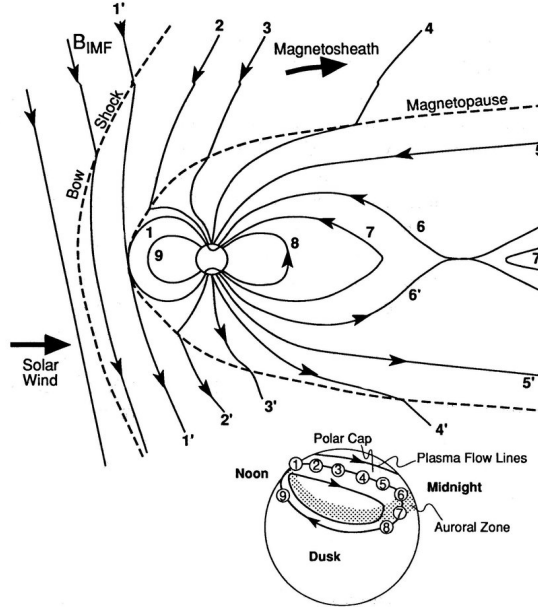


Figure 6.10.: The flow of plasma (convection pattern) in the magnetosphere, driven by magnetic reconnection as well as the movement in the polar ionosphere. The plasma is flowing antisunward in the polar cap and returning at lower latitudes. Figure from Kivelson and Russell [1995b]

6.7. The inner magnetosphere

In the outer magnetosphere the magnetic fields are weak, and the plasma $\beta = 2\mu_0 P/B^2$ is high. The plasma flows can then shape the magnetic field lines. Closer to Earth the field strength grows rapidly, and we have $\beta \ll 1$. The magnetic field then retains its dipolar shape, and takes control over the plasma flows. The atmosphere and the ionosphere are carried along with the rotation of the Earth. This creates an electric field in a non-rotating frame of reference, given by

$$\mathbf{E}_{cr} = -(\boldsymbol{\Omega}_E \times \mathbf{r}) \times \mathbf{B}, \quad (6.17)$$

where $\Omega_E = 7.27 \cdot 10^{-5}$ rad/s is the angular velocity of Earth's rotation. Using Equation 6.10 to express the magnetic field, we obtain the *corotational electric field* strength in the equatorial plane as

$$E_{cr} = \Omega_E B_0 R_E / L^2. \quad (6.18)$$

This electric field is directed radially inward towards Earth. Close to Earth it is stronger than the dawn to dusk electric field responsible for sunward convection, but it decreases with the square of the L -value. When the corotational electric field is added to a uniform dawn to dusk field, we obtain the structure shown in Figure 6.11

Since $\mathbf{U}$ is perpendicular to $\mathbf{E}$, the plasma flow will follow equipotential contours. Plasma from the tail will flow along the equipotentials or flow lines in Figure 6.11, until they enter the boundary layer where the assumed uniform electric field model is no longer valid. The density of this sunward convecting plasma is related to the density in the tail, and it is typically less than 10^7 m^{-3} . Closer to Earth, there is a region of closed equipotential contours marked by shading. This region, where the corotational electric field dominates, is called the *plasmasphere*. The magnetic field lines in the plasmasphere are never

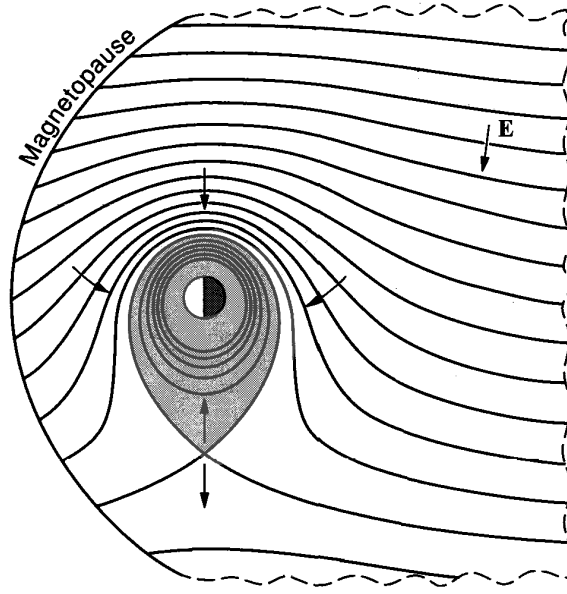


Figure 6.11.: Electric fields and equipotential contours in the equatorial magnetosphere. Figure from Baumjohann and Treumann [1996].

carried far out by the plasma flow. A plasmaspheric flux tube remains connected to the ionosphere at both ends, and it is filled with more than 10^9 m^{-3} ionospheric particles with energies in the range of 1 eV. At the *plasmopause*, the outer boundary of the plasmasphere, the density rapidly drops by more than two orders of magnitude.

6.8. Ring current

In the inner magnetosphere, where the magnetic field is rather dipolar and strong, the magnetic gradient drift given by equation (3.20) will have a significant effect on the motion of the plasma particles. In contrast to the $\mathbf{E} \times \mathbf{B}$ -drift, the magnetic gradient drift depends on the sign of the charge and causes a current in the plasma. Considering equatorial particles and assuming a dipolar field, we have $B(r) = B_0 R_E^3 / r^3$, $\frac{\partial}{\partial r} B = -3B/r$, and from (3.20) the magnitude of the gradient drift becomes

$$v_{gd} = \frac{3\mu}{qr} = \frac{3L^2}{qB_0 R_E} \frac{mv^2}{2}, \quad (6.19)$$

where we have used (3.8) for the magnetic moment μ . Keeping track of directions, we find that the ion drift, and hence the current, is in the westward direction, opposite to Earth's rotation. Comparing the magnetic gradient drift, which scales as L^2 , with the homogeneous electric field drift $\mathbf{E} \times \mathbf{B} / B^2$ that scales as L^3 and the corotation speed which is linear in L , we can suspect magnetic gradients to play a role at intermediate distances. The drift velocity given in Equation 6.19 is proportional to the particle energy, and it turns out that energetic ($\sim 100 \text{ keV}$) ions in the vicinity of the plasmopause can be carried westward around the Earth by the magnetic gradient drift. These ions form a *ring current*. When the magnetosphere is compressed by a strong solar wind, the ring current is intensified. On the surface of Earth, this is observed at low latitudes as a significant (several hundred nT) reduction of the magnetic field, which lasts a few days. These magnetic field reductions were observed long before the ring current was known, and they became known as *magnetic storms*.

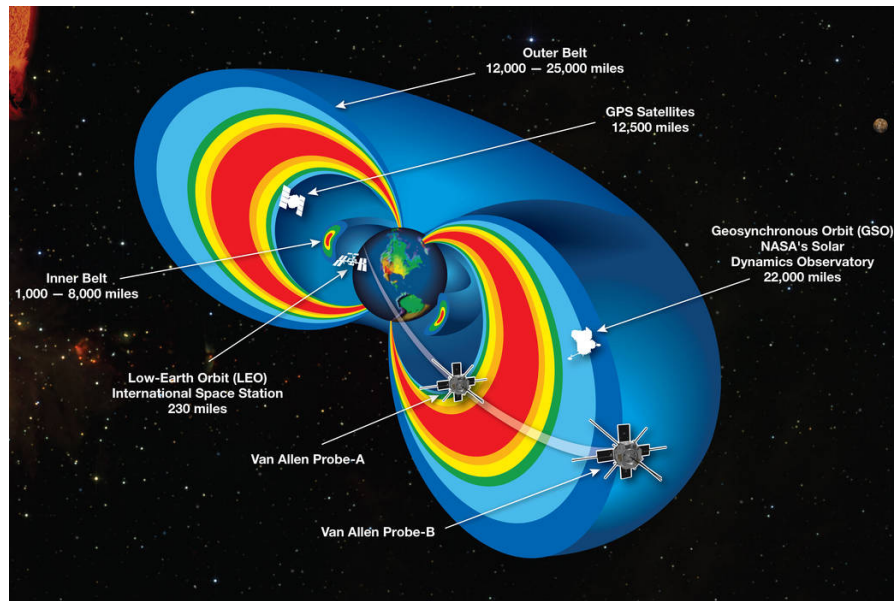


Figure 6.12.: A schematics of the two radiation belts forming a donut-shaped zone around Earth with high energy particles. Credits: T. W. Hill.

6.9. The radiation belts

The radiation belts, also called Van Allen radiation belts, are regions with trapped high-energy electrons and protons. Earth has two radiation belts and sometimes transient additional belts. Figure 6.12 illustrates Earth's radiation belts. There is an overlap between the radiation belt particles and ring-current particles. The ring current is carried by trapped particles as well as the particles in the radiation belts. The term radiation belt is often used to refer to high energy particles that can harm spacecraft and humans. Extra safety measures are needed for space missions crossing the radiation belts.

The inner belt ($1.5 - 3 R_E$ from Earth's center) is fairly stable, however during strong solar wind activity the inner belt can increase in size such that it can be detected 200 km above ground. Particles can fall out of the radiation belts and interact with Earth's atmosphere, changing its chemistry. The outer belt ($3 - 10 R_E$) varies significantly in size over time and consist mostly of energetic electrons.

6.10. The high latitude magnetosphere

Magnetic field lines that are connected to Earth's ionosphere at high latitudes do not close inside the magnetosphere. Instead they are considered to be *open field lines* that connect to the IMF. This ionospheric region with open field lines is the polar cap. The flow of solar wind plasma away from the Sun will also pull the ionospheric footprint of the flux tubes across the polar cap from the dayside to the nightside, as shown in Figure 6.10 (upper Figure). Field lines from the sunward convecting plasma sheet have their footprints in the auroral zone at slightly lower latitudes than the polar cap. The resulting convection pattern above the high-latitude ionosphere is shown in Figure 6.10 (lower Figure), where the numbers indicate the footprints of the numbered field lines in Figure 6.10 (upper Figure). Field lines 1 and 6 pass through the reconnection regions

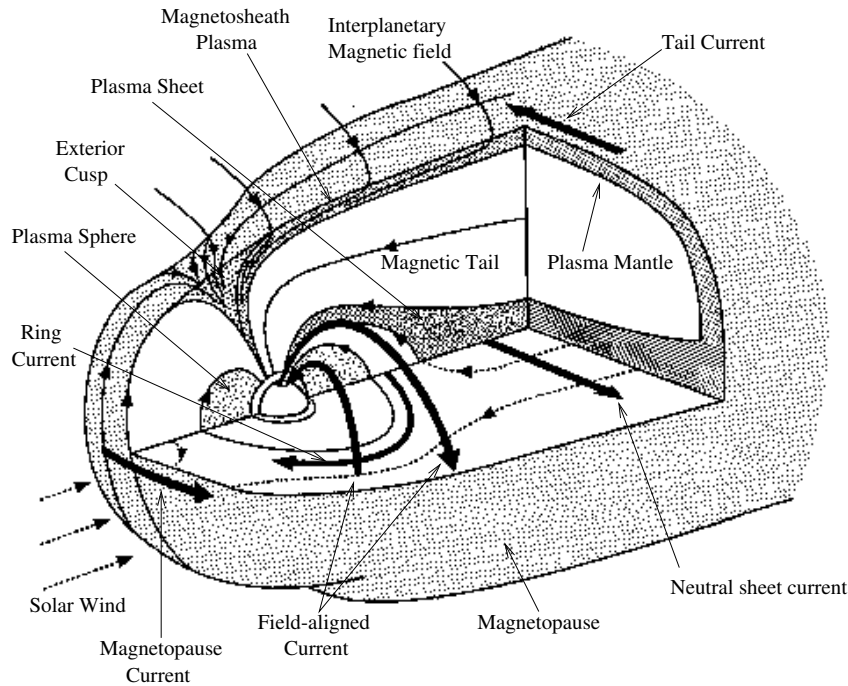


Figure 6.13.: Three-dimensional view of magnetospheric current systems and plasma regions. Modified after Kivelson and Russell [1995b].

where the frozen-in conditions are violated. The ionospheric flow is not well predicted by field-line mapping to these regions.

6.11. Current systems in the ionosphere

A summary of the various current systems in the magnetosphere is given in Figure 6.13. The ring current flows in the inner magnetosphere and is carried by the magnetic drift of energetic ions. The magnetopause current and the tail current are directly driven by the solar wind. The neutral sheet current, which is required to close the tail current system, powers a plasma engine that accelerates sunward plasma flows through reconnection. Breaking of the sunward flows generates field-aligned currents that close through the ionosphere, see subsection 7.3.6. The poleward field-aligned current (labeled Region 1 current in Figure 7.8) in the ionosphere connects to the magnetopause current. The equatorward field-aligned current (labeled Region 2 current in Figure 7.8) connects to the partial ring current. The partial ring current is an enhancement of the ring current on the nightside. The field aligned currents close as pedersen currents inside the ionosphere. This figure also shows the magnetospheric cusps, which are clefts in the high-latitude magnetopause on the dayside. Through the cusps, magnetosheath plasma can approach the ionosphere along the magnetic field lines.

6.12. Magnetospheres around other bodies

We have discussed in detail the magnetosphere around Earth, but other bodies, (for example other planets, comets, and moons) also form magnetospheres of different shapes and sizes. Generally, we can classify the solar wind interaction with celestial bodies into

6. The Magnetosphere

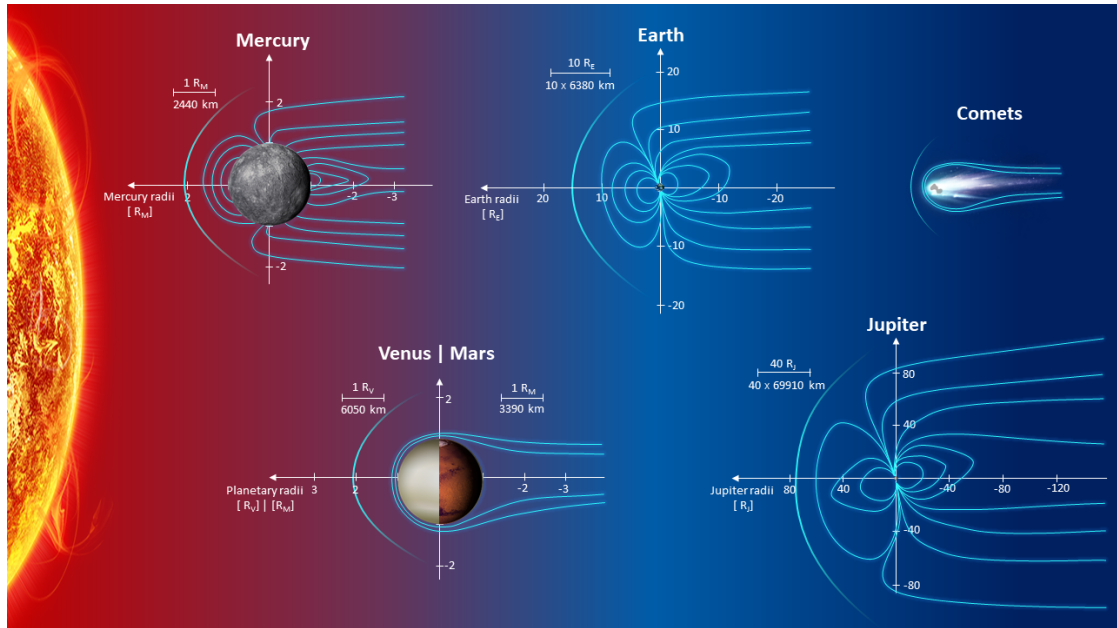


Figure 6.14.: The solar wind interaction with Mercury, Earth, Venus, Mars, Jupiter, and Comets forming magnetospheres of different sizes. Reprinted with permission of the Austrian Academy of Sciences, Space Research Institute and Daniel Schmid.

different types. The Lunar type describes the solar wind interaction with a body that has no atmosphere and no internal magnetic field. The Earth type describes the interaction with a body that possesses an atmosphere and an internal magnetic field, while the Venus type describes it the interaction with a body that has an atmosphere but not internal magnetic field. The Comet type describes the solar wind interaction with a Comet and is similar to the Venus type. An overview of the different interaction types is given in Figure 6.14.

The Lunar type is the solar wind interaction with an unmagnetized planet lacking an atmosphere. This type of interaction does not form a magnetosphere since the solar wind directly impacts the bodies surface and is absorbed. Due to the absorption of the solar wind, no bow shock forms and the solar wind is not deflected. Behind the body there is a solar wind cavity.

The Earth type is the solar wind interacting with a magnetized planet having an atmosphere. A self-consistent current system is created deforming the dipolar magnetic field. The magnetic field deflects the incoming solar wind plasma and there is little solar wind plasma entering the magnetosphere. The size of the magnetosphere scales with the strength of the magnetic dipole. For example, Jupiter and Saturn have a stronger magnetic field and have a therefore a larger magnetosphere. Mercury on the other hand has a weaker magnetic field and the scale size of the magnetosphere is therefore smaller.

The Venus type refers to the solar wind interaction with an unmagnetized planet that has an atmosphere. The upper atmosphere is ionized forming an ionosphere due to the solar radiation. The ionosphere is conducting and a current flows due to the $v \times B$ electric field. As a result, an induced magnetosphere forms which acts as an obstacle to the incoming solar wind. The interplanetary magnetic field drapes around the body.

Similar to Earth, a bow shock forms but with a smaller stand-off distance compared to Earth type magnetospheres. Mars has a similar interaction with the solar wind, however, Mars also has crustal magnetic fields in the southern hemisphere introducing an additional asymmetry in the system.

The comet type interaction describes the solar wind interaction with comets. Comets consist of ice, rocks, and dust. Comets often have more than one visible tail. The dust tail consists of dust grains that are pushed away from the Sun by the photon pressure of the sunlight. The ion tail is caused by the interaction with the solar wind, and that varies significantly along the comet's trajectory. Far away from the Sun, the comet is inactive. As the comet comes closer the Sun, it starts outgassing due to solar radiation sublimating material. This gas expands away from the comet since the comet's gravity is too weak. The comet atmosphere that is formed in this way is called the coma. The gas can be ionized by solar radiation and further interaction with the solar wind. Close to the Sun, when outgassing is strongest, the interaction with the solar wind becomes similar to the Venus-type interaction. The magnetosphere extends away from the Sun, and this what forms the ion tail. However, of all the planets, Mars is the most comet-like. This is because its small size, and hence weaker gravitational pull, lets its atmosphere extend further out into space, although not as far as for a high activity comet. A peculiar phenomenon, occurring at high activity comets, is the formation of a diamagnetic cavity around the nucleus. In the diamagnetic cavity the magnetic field is close to zero, because, the interplanetary magnetic field cannot penetrate the inner part of the coma, where the expanding plasma pushes it outward. Additionally, a bow shock forms farther upstream, and its shape and size is also highly dependent on the outgassing rate.

Some of the interaction are more complicated. For example Jupiter's moon Ganymede has its own magnetic dipole and the orbit is inside Jupiter's magnetosphere. Ganymede does build its magnetosphere but is lacking the bow shock since the plasma moves with a speed lower than the local Alfvén speed.

6. The Magnetosphere

Example

Jupiter's magnetopause location is determined by the pressure balance condition

$$(P_B + P_p)_{SW} = (P_B + P_p)_{JM}$$

where the subscript SW denotes the solar wind and JM denotes Jupiter's magnetosphere, $P_B = B^2/2\mu_0$ is the magnetic field pressure, and $P_p = \rho V^2/2$ is the dynamical pressure of a plasma with mass density ρ and velocity V .

The solar wind magnetic pressure $(P_B)_{SW}$ can be neglected, but we take into account that Jupiter has an intrinsic "Jupiter wind" of protons expanding radially outward with a velocity of 100 km/s and a number density of 10^6 m^{-3} near the magnetopause. We assume that Jupiter has a dipole magnetic field with strength $4.2 \times 10^{-4} \text{ T}$ at Jupiter's equatorial surface. Finally, we assume that the solar wind (consisting mainly of protons) has a number density of $5 \times 10^6 \text{ m}^{-3}$ and a velocity of 400 km/s at the Earth's orbit.

Using these assumptions, compute the distance, in Jupiter radii, from the center of Jupiter to Jupiter's magnetopause.

(Try before you look at the solution on the next page!)

Answer

Since the plasma pressure P_{SW} is (approximately) independent of the magnetopause distance and P_{JW} is given at the magnetopause, we write the pressure balance equation as

$$P_{BJ} = P_{SW} - P_{JW}, \quad (6.20)$$

where the magnetic field pressure is

$$P_{BJ} = \frac{(B_0(R_J/r)^3)^2}{2\mu_0}. \quad (6.21)$$

Here, $B_0 = 4.2 \cdot 10^{-4}$ T is the surface magnetic field. Solving for the magnetopause distance we find

$$r = \left(\frac{B_0^2}{2\mu_0(P_{SW} - P_{JW})} \right)^{1/6} R_J. \quad (6.22)$$

At the Earth's orbit, the solar wind velocity $v_E = 400$ km/s and density $n_E = 5 \cdot 10^6$ m⁻³ are given. Conservation of energy of a solar wind proton gives

$$\frac{mv_E^2}{2} + \frac{mM_\odot G}{r_E} = \frac{mv_J^2}{2} + \frac{mM_\odot G}{r_J}. \quad (6.23)$$

A rough estimate shows that the gravitational potential energy ($\sim 10^{-18}$ J/proton) is much less than the kinetic energy ($\sim 10^{-16}$ J/proton), and hence we can assume that the velocity is constant. Conservation of particles implies that the particles crossing a sphere of radius r_E must also cross the sphere at Jupiter's distance r_J

$$n_E v_E 4\pi r_E^2 = n_J v_J 4\pi r_J^2, \quad (6.24)$$

or, since $v_E = v_J$, that $n_J = n_E(r_E/r_J)^2$. The solar wind pressure is thus

$$P_{SW} = \frac{m}{2} v_E^2 n_E (r_E/r_J)^2, \quad (6.25)$$

where $m = 1.67 \cdot 10^{-27}$ kg is the proton mass, and the pressure from the "Jupiter wind" is

$$P_{JW} = \frac{m}{2} v_{JW}^2 n_{JW}. \quad (6.26)$$

Putting it all together, we find that the magnetopause distance of Jupiter should be

$$r = \left(\frac{B_0^2}{m\mu_0(v_E^2 n_E (r_E/r_J)^2 - v_{JW}^2 n_{JW})} \right)^{1/6} R_J. \quad (6.27)$$

Inserting the numerical values we find

$$r = (4.29 \cdot 10^9)^{1/6} R_J \approx 40.3 R_J. \quad (6.28)$$

Answer: The magnetopause distance is 40.3 Jupiter radii.

7. The upper Atmosphere and Ionosphere

7.1. Introduction

The interface between the neutral atmosphere and outer space is the ionosphere, a partially ionized gas that envelops the Earth. Owing to the strong gravitational field near Earth, the atmosphere and ionosphere are to first order horizontally stratified. Atmospheric structure is best described by a representative temperature profile, while the ionosphere is more sensibly organized by the density of free electrons. After a brief description of these structures, we shall consider some of the physical processes behind the creation and dynamics of the ionosphere.

7.2. The atmosphere

7.2.1. Origin of the atmosphere

The main ingredient of the atmosphere is about 78% nitrogen (N_2), which originates from volcanic activity and denitrification of decaying organic substances. Nitrogen is removed from the atmosphere by photochemical processes, lightning discharges, and the life processes of microbes. The oxygen (O_2) constitutes about 21% of the atmosphere, and is produced mainly by the photosynthesis of plants. The removal of oxygen is due to oxidation processes such as the decomposition of minerals, combustion, and the respiration of living organisms. The percentages given above apply to dry air—in addition the atmosphere also contains 1–5 % water vapour. Carbon dioxide (CO_2), which dominates the atmospheres of other planets ($\sim 95\%$ on Venus and Mars) is only about 0.04% of Earth's atmosphere. This composition of the atmosphere, which is very far from a normal chemical equilibrium, is maintained by the presence of life on Earth. The carbon in CO_2 is by photosynthesis built into organic molecules, and the oxygen is released into the atmosphere.

Most of the atmospheric constituents undergo a circulation. The residence time in the atmosphere can be very different for different components. For example, water vapour stays in the air a few weeks while argon (0.9%) remains for 100 000 years.

7.2.2. Height profile of the atmosphere

From sea level, the atmospheric temperature initially decreases with altitude by about 7 K/km in the *troposphere*. At the *tropopause*, at about 10 km altitude, this temperature trend reverses and the *stratosphere* begins. The temperature increase here is due primarily to the absorption, by ozone, of part of the ultraviolet portion of the sunlight. This effect maximizes at 50 km, where the temperature trend again reverses at the *stratopause*. Radiative cooling creates a very sharp temperature decrease to a minimum in the range 130–190 K at about 90 km. For heights above the temperature minimum (the

7. The upper Atmosphere and Ionosphere

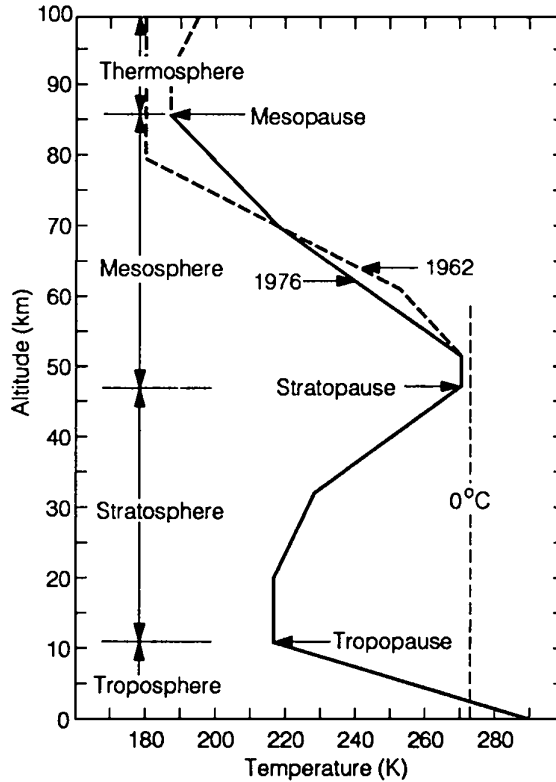


Figure 7.1.: Altitude variation of temperature in the Standard Atmosphere in the altitude interval 0–100 km. Figure from Fälthammar [2001].

mesopause), the temperature increases dramatically to values that are often well above 1000 K. Not surprisingly, this region is termed the *thermosphere*. The heating of the thermosphere is at least partly explained by absorption of another portion of the solar UV spectrum.

The World Meteorological Organization (WMO) has defined a "*Standard Atmosphere*", which defines the distribution of temperature, pressure, and density representative of year-round, mid-latitude, atmospheric conditions. The temperature variation in the Standard Atmosphere is illustrated in Figure 7.1. Obviously, the real atmosphere varies with time and place, and it is important to remember that it always deviates more or less from the Standard Atmosphere.

The composition of the atmosphere is relatively uniform below about 100 km due to a variety of turbulent mixing phenomena. Above this altitude, the *turbopause*, the constituents begin to separate according to their various masses. The region with homogeneous composition below the turbopause is referred to as the *homosphere*. Correspondingly, the region above the turbopause is called the *heterosphere*. The height distribution of various constituents in the Standard Atmosphere is shown in Figure 7.2.

Climbing a sufficiently high mountain, we can notice a decrease in the air pressure as we struggle against the gravitational force to higher altitudes. The gravitational force acts on air as well as on climbers, and on a volume element of density ρ_m it equals $g\rho_m$ (where $g = 9.81 \text{ m/s}^2$) and is directed downward. In hydrostatic equilibrium, the gravitational force is balanced by the pressure gradient force $-t_z p$, where z is the altitude coordinate,

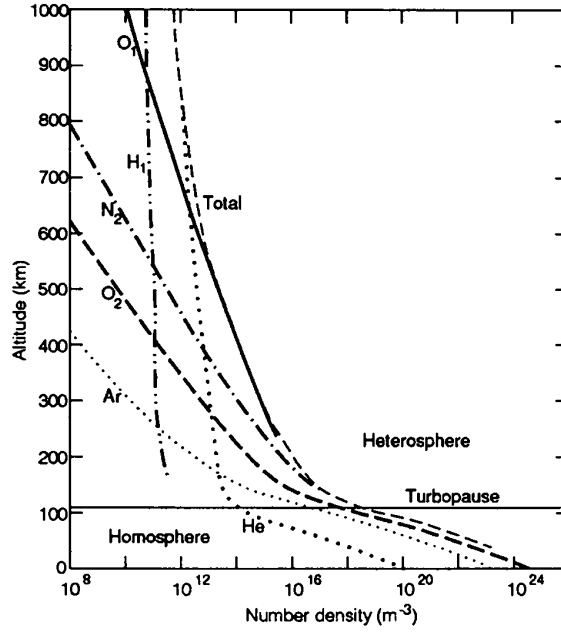


Figure 7.2.: Altitude distribution of the main components of the Standard Atmosphere. Figure from Fälthammar [2001].

positive upward. Thus

$$-\frac{d}{dz}p = g\rho_m \quad (7.1)$$

As outlined in subsection 3.3.1, the density and pressure are related by $p = nT = T\rho_m/m$, where n is the density of molecules of mass m . (Remember that we use energy units for temperature—if T is given in Kelvin, it should be multiplied by Boltzmann's constant $k = 1.38 \times 10^{-23}$ J/K.) If the temperature were constant, we would have $-\frac{d}{dz}p = -T/m \frac{d}{dz}\rho_m = g\rho_m$, which easily can be integrated to give

$$\rho_m(z) = \rho_{m0}e^{-zmg/T} = \rho_{m0}e^{-z/H} \quad (7.2)$$

where the parameter $H = \frac{T}{mg}$ has the dimension length and is called *scale height*. The mass of an N_2 -molecule is 4.7×10^{-26} kg. When the temperature is $20^\circ \text{C} = 293 \text{ K} \sim 4 \times 10^{-21} \text{ J}$, the scale height is thus about 10 km. This means that the air pressure at the peak of Mount Everest is about a third of its sea level value.

Since, as discussed above, the temperature really varies with altitude, the true large scale distribution is far from exponential, and the concept of a scale height can only be applied locally. In Figure 7.3, the pressure and mass density variation in the Standard Atmosphere are shown on logarithmic scales against altitude. If the temperature in the atmosphere really were constant, the curves in this figure would have been straight lines. Now, we can see that in particular the strong increase in temperature as we go from the mesopause into the thermosphere causes a fairly sharp bend in the curves.

7.2.3. Ozone layer

An interesting and fashionable example of photochemical processes in the upper atmosphere is provided by the ozone layer. In the stratosphere (altitudes 10–50 km), UV-light with

7. The upper Atmosphere and Ionosphere

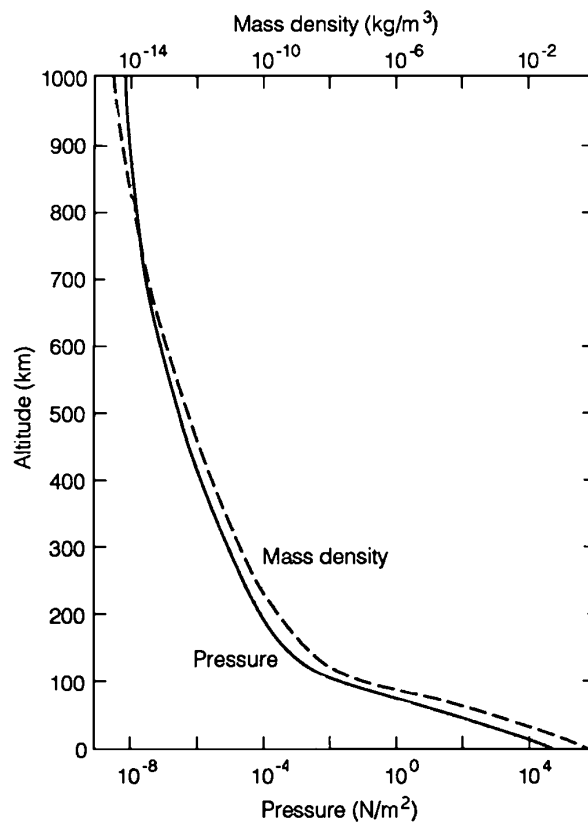


Figure 7.3.: Altitude distribution of the pressure and mass density in the Standard Atmosphere. Figure from Fälthammar [2001].

wavelength $\lambda < 242$ nm will dissociate oxygen molecules into two atoms



These free oxygen atoms will become attached to another O_2 molecule, forming ozone (O_3)

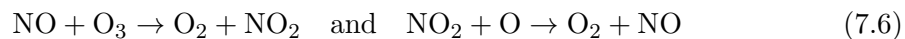


Here, M denotes a non-reacting molecule that helps to conserve energy and momentum. The ozone molecules will very efficiently absorb light in the important UV-B range, 280–310 nm. Although only one atmospheric molecule in a million is ozone, they can absorb almost all of the UV-B light before it reaches the ground, and this absorption is the main reason for the temperature increase in the stratosphere. Ozone molecules will rapidly be dissociated by photons with $\lambda < 210$ nm into O_2 and atomic oxygen. However, the oxygen atom will then quickly combine with another O_2 molecule to form ozone again, so there is no net destruction by this process. It turns out that the important quantity is the amount of "odd oxygen", O and O_3 together.

The simplest way of removing odd oxygen would be the reaction



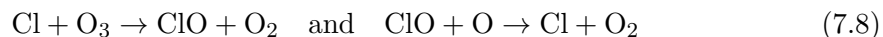
but this process turns out to be very slow. Molecules such as OH and NO , which occur naturally in the atmosphere, can speed up the process through catalytic reactions such as



These molecules are removed by the formation of nitric acid,



which is washed down into the troposphere. Unfortunately, another catalyst, chlorine (Cl), is not so easily removed. Chlorine in the stratosphere may come from man-made chloroflourocarbon (CFC) gases, which are marketed under names like Freon. CFC is chemically very stable and can slowly diffuse up into the stratosphere, where the UV-light can break it apart and release the Cl atom. The chlorine can then catalyze reactions very similar to Equation 7.6



The loss of chlorine is very slow, and a single Cl atom may remove 100 000 ozone molecules before it is removed. It seems likely that Cl atoms play a role in the creation of the ozone hole over Antarctica, but some of this chlorine may also come from natural sources such as volcanic eruptions. The long term effects of CFC on the ozone layer are still hard to predict, and we can only hope that the trend is reversed before the damage has become too severe.

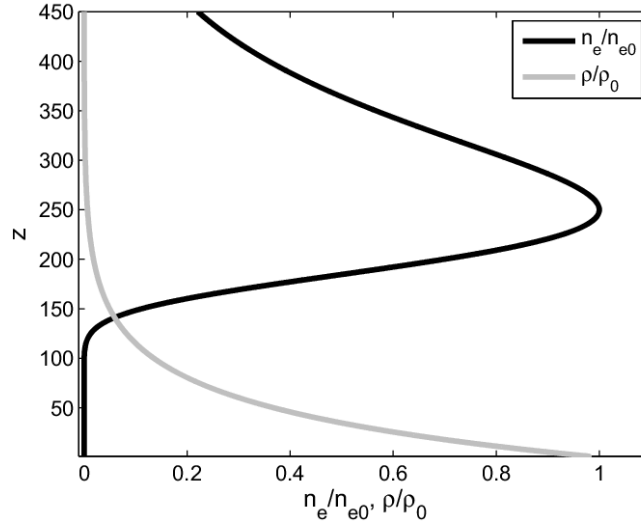


Figure 7.4.: The Chapman profile for the electron density. Figure from Norin [2008].

7.3. The ionosphere

7.3.1. History

In order to explain the daily variations in the Earth's magnetic field, Balfour Stewart suggested in 1882 that currents should be flowing in an electrically conducting layer in the upper atmosphere. The daily temperature variations should cause motions across the magnetic field lines in this layer, and this induces currents that perturb the geomagnetic field. In 1902, Kennely and Heaviside used a similar hypothesis to explain the Marconi's successful transatlantic radio communication. They proposed that the radio waves were reflected from a conducting layer at high altitudes, which for a long time was called the *Kennely-Heaviside-layer*. The validity of this hypothesis was proven experimentally in 1925 by Appleton and Barnett, and shortly thereafter by Breit and Tuve. These experiments also allowed the altitude of the conducting layer, the ionosphere, to be estimated to about 100 km.

7.3.2. Formation of the ionosphere

The air we breathe consists of electrically neutral molecules. Why, then, is there an ionosphere with free electrons and ions at high altitudes? The main reasons are the intense X-ray and UV radiation from the Sun, which ionizes the atoms, and the low density, which prevents the electrons and ions from recombining.

To illustrate the physics and overall structure of the ionosphere we shall analyze an atmosphere with exponentially decreasing density under the influence of ionizing radiation. We assume that the density n_n of neutral gas as in equation Equation 7.2 varies as

$$n_n(z) = n_o e^{-z/H} \quad (7.9)$$

where z is the altitude above the (plane) surface of the earth, and the scale height H is constant. High above the atmosphere ($z \rightarrow \infty$), the incident ionizing X-ray and UV radiation has intensity I_o . Due to absorption the radiation intensity $I(z)$ diminishes as

we approach the Earth according to the equation

$$\frac{d}{dz}I(z) = \sigma_a n_n(z)I(z) \quad (7.10)$$

where σ_a is the absorption coefficient. Inserting Equation 7.9 and integrating we find

$$I(z) = I_o \exp[-H\sigma_a n_o e^{-z/H}] \quad (7.11)$$

The number of ionizations will be proportional to the density of neutral atoms and to the radiation intensity and can be written $a_i n_n(z)I(z)$, where a_i is the ionization coefficient. Assuming singly ionized ions, the ion density will equal the electron density $n_e(z)$, and if the recombination coefficient is a_r the recombination rate is $a_r n_e^2(z)$. In equilibrium, we must have balance between the local ionization and recombination

$$a_i n_n(z)I(z) = a_r n_e^2(z) \quad (7.12)$$

Solving for the electron density and inserting Equation 7.9 and Equation 7.11 we the find

$$n_e(z) = \sqrt{\frac{a_i}{a_r} n_o e^{-z/H} I_o \exp[-H\sigma_a n_o e^{-z/H}]} \quad (7.13)$$

The maximum electron density is found by taking the derivative and occurs at the level

$$z_{max} = H \ln(\sigma_a n_o H) \quad (7.14)$$

By introducing the reduced height $\xi = (z - z_{max})/H$ from the maximum we can rewrite Equation 7.13 in the somewhat simpler form

$$n_e(\xi) = n_e^{max} \exp[(1 - \xi - e^{-\xi})/2] \quad (7.15)$$

which is the classical formula for the so-called Chapman profile, derived by Chapman in 1931. The shape of this profile is shown in Figure 7.4. The mechanism derived above can explain the existence of a ionosphere, but it cannot account for the fine structure and variations of the real ionosphere.

7.3.3. The ionospheric layers

The actual structure of the ionosphere is the result of an interplay between the spectral distribution of the incident radiation and the altitude variation of the chemical composition of the upper atmosphere. It is customary to distinguish four different ionospheric layers, called D, E, F₁, and F₂. The typical altitudes, electron densities, ion contents, and dominating causes of ionization are summarized in Table (5.3.1).

As seen from the table, and also from Figure 7.5 a, b, the electron density in the lower part of the ionosphere is much higher during the day than at night.

After sunset, the molecular ions in the D, E, and F₁ layers disappear rather quickly through dissociative recombination processes of the form



7. The upper Atmosphere and Ionosphere

Layer	D	E	F ₁	F ₂
Altitude (km)	60–85	85–140	140–200	200–ca 1500
Nighttime electron density (cm ⁻³)	$< 10^2$	2×10^3	—	$2 - 5 \times 10^5$
Daytime electron density (cm ⁻³)	10^3	$1 - 2 \times 10^5$	$2 - 5 \times 10^5$	$0.5 - 2 \times 10^6$
Ion species	NO ⁺ O ₂ ⁺	NO ⁺ O ₂ ⁺	NO ⁺ O ₂ ⁺ O ⁺	O ⁺ He ⁺ H ⁺
Cause of ionization	Lyman α (121.5 nm)	Lyman β (102.5 nm) X-rays	UV	UV

Table 7.1.: The ionospheric layers.

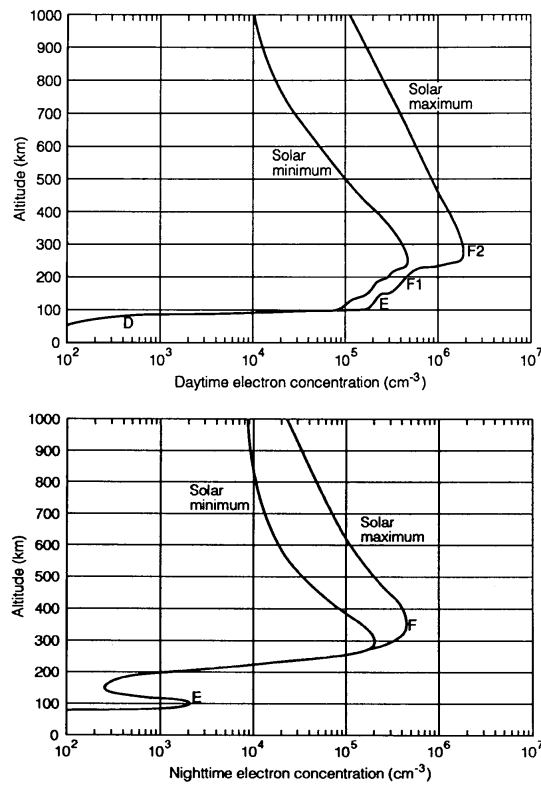


Figure 7.5.: Electron densities in the ionosphere at the extremes of the sunspot cycle for daytime (a) and nighttime (b) conditions. Figure from Fälthammar [2001].

Since this results in two atoms it is easy to conserve energy and momentum, and these processes have a high cross section. The F₂ layer, on the other hand, consists of atomic ions and dissociative recombination is thus forbidden. Processes like



where $\hbar\omega$ is a photon, and



where M is a third particle, are much slower, and hence the F₂ layer decays much less during the night. The ionosphere is also affected by the solar activity. During sunspot maximum the short-wave radiation from the Sun is intensified, and the electron density in the ionosphere then increases as shown in Figure 7.5.

The ionosphere merges without a sharp boundary into the magnetosphere. It is therefore a matter of definition where to consider the ionosphere to end. A common definition, used in Table (5.3.1), is that it goes over into the magnetosphere at a level of 1 000–2 000 km.

7.3.4. Collisions and conductivity

The ionosphere is a very good conductor, although the conductivity is highly variable. At the bottom of the ionosphere the density of neutrals is high, and the electron density is relatively low. Here, the current flow is impeded by collisions with the neutral gas molecules, and the conductivity is low. At around 100 km altitude, the electron density has increased and the collision frequency has decreased by orders of magnitude. This is where the conductivity is highest. Moving further up, collisions are no longer important and the conductivity decreases in proportion to the decreasing plasma density. The variation of the electron collision frequency is shown in Figure 7.6.

However, the conductivity does not only vary with altitude—it varies also with the direction in relation to the magnetic field. Below 100 km the current is limited by collisions and the magnetic field is not very important. The conductivity is then *isotropic* (the same in all directions). However, above this level, where the gyrofrequency becomes higher than the collision frequency, the electrons can move freely along the magnetic field but not across it. As we know from chapter 2, an electric field applied perpendicular to the magnetic field will make the charges move perpendicular to both the magnetic and electric fields. The ionospheric plasma at high altitudes has a very *anisotropic* conductivity. An electric field in one direction usually drives a current in a different direction. We must therefore distinguish between three different electrical conductivities, namely the

1. *Parallel conductivity* $\sigma_{\parallel}$, i.e., the conductivity parallel to the magnetic field.
2. *Pedersen conductivity* σ_P , i.e., the conductivity parallel to the component of the electric field that is perpendicular to the magnetic field.

7. The upper Atmosphere and Ionosphere

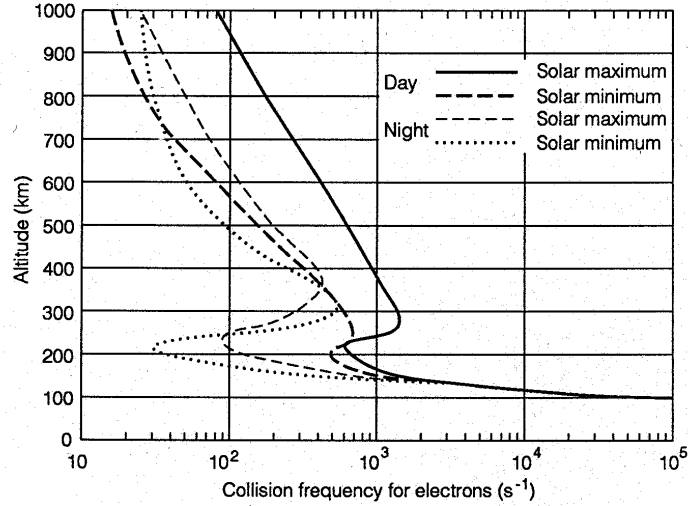


Figure 7.6.: Altitude variation of the electron collision frequency in the ionosphere. Figure from Fälthammar [2001].

3. *Hall conductivity* σ_H , i.e. the conductivity perpendicular to both the electric and the magnetic fields.

These three conductivities depend on both electron and ion collision frequencies. In the E and F layers, $\sigma_{\parallel}$ is several orders of magnitude greater than the others, which are of mutually comparable magnitude.

Let us see in some detail how collisions determine the conductivity of a magnetized plasma. If the collision frequency is ν , the probability that a particle collides during a time Δt is $\nu\Delta t$. To be specific, we will here by “collision” mean a deflection by 90° , even if this deflection really is the result of several smaller deflections. The probability that the time since a particle last collided is between t and $t + \Delta t$ is then

$$p(t)\Delta t = \nu e^{-\nu t} \Delta t \quad (7.19)$$

The probability density $p(t)$ decreases exponentially, since the number of particles that have not collided decreases. We consider a magnetic field $\mathbf{B} = B_0\hat{z}$ in the z -direction and an electric field $\mathbf{E} = E_y\hat{y} + E_z\hat{z}$ in the (y, z) -plane. As shown in chapter 2, the velocity of the particles is between collisions given by

$$v_x(t) = \frac{E_y}{B_0} - \frac{E_y}{B_0} \cos \omega_c t \quad (7.20a)$$

$$\frac{d}{dt}v_y(t) = \frac{E_y}{B_0} \sin \omega_c t \quad (7.20b)$$

$$\frac{d}{dt}v_z(t) = \frac{q}{m} E_z t \quad (7.20c)$$

where the gyrofrequency $\omega_c = qB_0/m$ includes the sign of the charge. To be consistent with our definition of a collision, we here assume the initial condition that the particles are restarted with velocity $\mathbf{v} = 0$ at $t = 0$ after each collision. The average velocity is given by the integral of $\mathbf{v}(t)$ weighted by the probability density as

$$\langle \mathbf{v} \rangle = \int_0^\infty \mathbf{v}(t)p(t)dt \quad (7.21)$$

Starting with the z -component we obtain

$$\langle v_z \rangle = \int_0^\infty \frac{q}{m} E_z t \nu e^{-\nu t} dt = \frac{q E_z}{m \nu} \quad (7.22)$$

In the ionosphere collisions with neutrals are important, and the electron collision frequency ν_e is not simply related to the ion collision frequency ν_i . In a plasma with electron density n , we obtain the current density as

$$j_z = ne(\langle v_{zi} \rangle - \langle v_{ze} \rangle) = \left(\frac{ne^2}{m_i \nu_i} + \frac{ne^2}{m_e \nu_e} \right) E_z \approx \varepsilon_0 \left(\frac{\omega_{pi}^2}{\nu_i} + \frac{\omega_{pe}^2}{\nu_e} \right) E_z \quad (7.23)$$

where we can identify the parallel conductivity $\sigma_{\parallel} = \varepsilon_0(\omega_{pi}^2/\nu_i + \omega_{pe}^2/\nu_e)$. Clearly, the parallel conductivity becomes infinite when the collision frequencies approach zero.

The Pedersen conductivity can similarly be calculated from the average v_y , which is

$$\langle v_y \rangle = \frac{E_y}{B_0} \int_0^\infty \sin \omega_c t \nu e^{-\nu t} dt = \frac{\nu \omega_c}{\omega_c^2 + \nu^2} \frac{E_y}{B_0} \quad (7.24)$$

Including the contribution from ions and electrons when calculating j_y we find that the Pedersen conductivity is given by

$$\sigma_P = \varepsilon_0 \left[\frac{\omega_{pi}^2 \nu_i}{\omega_{ci}^2 + \nu_i^2} + \frac{\omega_{pe}^2 \nu_e}{\omega_{ce}^2 + \nu_e^2} \right] \quad (7.25)$$

From this expression we see that both very high and very low collision frequencies lead to a vanishing Pedersen conductivity.

A similar calculation for the x -component leads to

$$\langle v_x \rangle = \frac{\omega_c^2}{\omega_c^2 + \nu^2} \frac{E_y}{B_0} \quad (7.26)$$

At first sight one may expect that this should lead to a finite Hall conductivity when $\nu \rightarrow 0$. However, adding the ion and electron contributions we find

$$\sigma_H = \varepsilon_0 \left[\frac{\omega_{pi}^2 \omega_{ci}}{\omega_{ci}^2 + \nu_i^2} - \frac{\omega_{pe}^2 \omega_{ce}}{\omega_{ce}^2 + \nu_e^2} \right] \quad (7.27)$$

which goes to zero as $\nu_{i,e} \rightarrow 0$ since $\omega_{pi}^2/\omega_{ci} = \omega_{pe}^2/\omega_{ce}$.

For an arbitrary direction of $\mathbf{E}$ we can sum up Ohm's law in a more compact notation as

$$\mathbf{j} = \boldsymbol{\sigma} \cdot \mathbf{E} \quad (7.28)$$

where the conductivity tensor $\boldsymbol{\sigma}$ in the coordinate system with $\hat{z}$ along $\mathbf{B}$ is given by

$$\boldsymbol{\sigma} = \begin{pmatrix} \sigma_P & \sigma_H & 0 \\ -\sigma_H & \sigma_P & 0 \\ 0 & 0 & \sigma_{\parallel} \end{pmatrix} \quad (7.29)$$

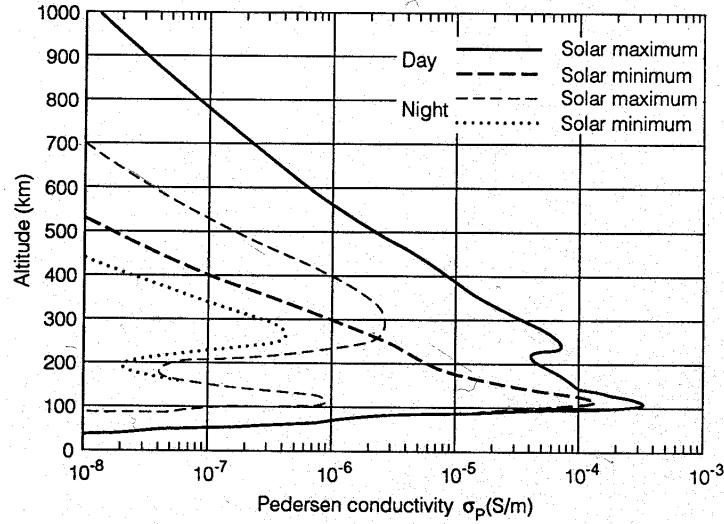


Figure 7.7.: Altitude variation of the Pedersen conductivity in the ionosphere.

As shown in Figure 7.6, the electron collision frequency rapidly drops at altitudes above 100 km. The parallel conductivity rapidly increases to about $10 (\Omega\text{m})^{-1}$ at 200 km, and remains high as long as the electron density is high. The variation of the Pedersen conductivity is shown in Figure 7.7. It has a clear maximum just above 100 km, where it during daytime reaches more than $10^{-4}(\Omega\text{m})^{-1}$. The variation of the Hall conductivity is similar to Figure 7.7, although the maximum around 300 km at night is not found in the Hall conductivity.

7.3.5. Height integrated conductivity

When the ionosphere via field aligned currents is made part of a magnetospheric current circuit, the altitude variation of the horizontal conductivities is often not very important. Most of the current will flow in a layer at altitudes just above 100 km, and in a magnetospheric perspective the thickness of this layer is negligible. Rather than considering the horizontal current density $\mathbf{j}_\perp$ in A/m^2 we then use the height integrated current $\mathbf{I}_\perp$ in A/m , defined by

$$\mathbf{I}_\perp = \int_0^\infty \mathbf{j}_\perp(z) dz \quad (7.30)$$

where z is vertically up through the ionosphere. Integrating Ohm's law over the height of the ionosphere we find

$$\mathbf{I}_\perp = \int_0^\infty \boldsymbol{\sigma}_\perp(z) \cdot \mathbf{E}_\perp(z) dz \approx \mathbf{E}_\perp \cdot \int_0^\infty \boldsymbol{\sigma}_\perp(z) dz = \boldsymbol{\Sigma} \cdot \mathbf{E}_\perp \quad (7.31)$$

where the height integrated conductivity tensor $\boldsymbol{\Sigma}$ is a two by two matrix consisting of Σ_P and Σ_H , and we have assumed that $\mathbf{E}_\perp$ is essentially independent of altitude.

Integrating the current continuity equation over the height of the ionosphere under stationary conditions ($\frac{\partial}{\partial t}\rho = 0$) and using the height integrated conductivity we find

$$\int_0^\infty \frac{\partial}{\partial z} j_z + \frac{\partial}{\partial \perp} \cdot \mathbf{j}_\perp dz = j_\parallel + \frac{\partial}{\partial \perp} \cdot \boldsymbol{\Sigma} \cdot \mathbf{E}_\perp \approx j_\parallel + \Sigma_P \frac{\partial}{\partial \perp} \cdot \mathbf{E}_\perp = 0 \quad (7.32)$$

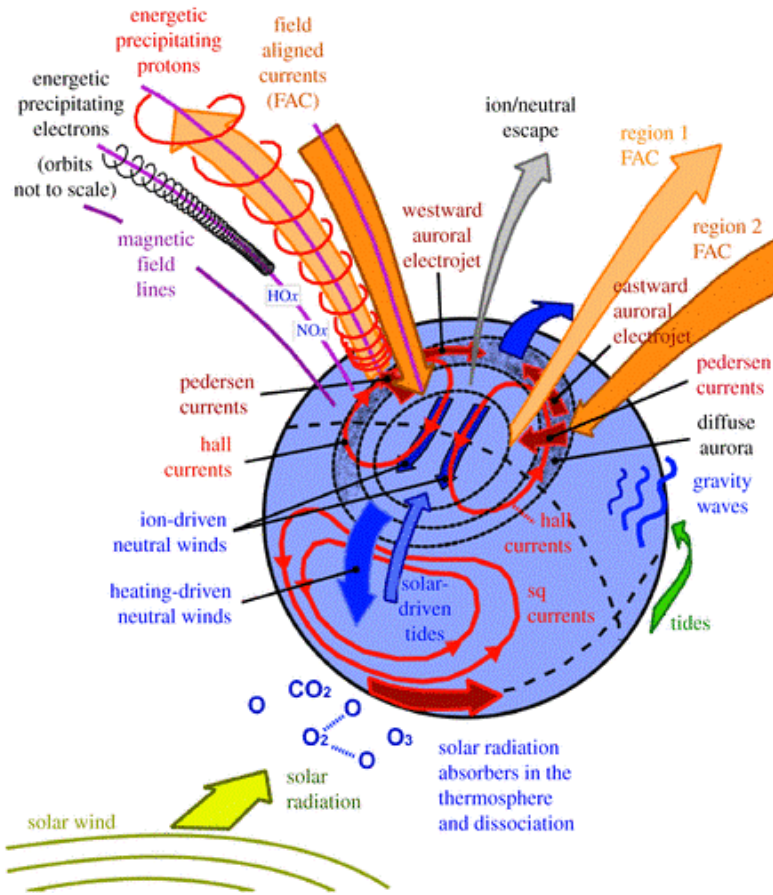


Figure 7.8.: Ionospheric current systems. Figure from Sarris [2019].

where we in the last step have neglected derivatives of Σ_P and Σ_H . Notice that the Hall current does not contribute to the field aligned current, since it has zero divergence in a homogeneous medium. There is no parallel current at the bottom of the ionosphere, and $j_{\parallel}$ denotes the current density at the top of the ionosphere. This equation determines the variation of the ionospheric electric field in terms of the field aligned current out of the ionosphere and the height integrated Pedersen conductivity. If we from Figure 7.7 take σ_P to be of the order $10^{-5} (\Omega \text{ m})^{-1}$ over a few hundred km, we can roughly estimate the height integrated conductivity to be $1\text{-}10 \Omega^{-1}$.

7.3.6. Ionospheric currents

As mentioned in subsection 7.3.1, the observations that first indicated the existence of an ionosphere were magnetic disturbances, noticed on ground and explained by electric currents high above. We now know that the hypothesis that magnetic disturbances could be caused by ionospheric currents was correct.

Some of the ionospheric current systems are closed within the ionosphere, but in most cases they are part of much larger current loops that extend far out in the magnetosphere. A schematic drawing of some of these currents is shown in Figure 7.8, which should be compared with Figure 6.13. The currents in the auroral zone, such as the auroral electrojet, are much intensified during auroral activity, and this can easily be observed by ground based magnetometers. Much is still unknown about the role played by the

7. The upper Atmosphere and Ionosphere

field aligned currents, but as we shall see they are certainly important for the interaction between the ionosphere and the magnetosphere.

8. The Aurora

8.1. Introduction

The aurora is a visible and fascinating consequence of the complex magnetospheric processes that are driven by the interaction between the solar wind and the geomagnetic field. Some regard the aurora as the most beautiful of all natural phenomena that can be observed in the sky, and it is certainly the most dynamic and variable. The aurora can appear as a diffuse pale band crossing the sky from east to west, or as rapidly moving bright and colourful curtains and rays covering a large fraction of the sky. We will not attempt to describe all the various forms that auroras may appear in, but only distinguish between diffuse and discrete auroras. The particles that cause these two main types are precipitated into the atmosphere by rather different physical processes, as described in this chapter. Magnetospheric substorms and the mechanisms leading to the formation of discrete auroras are still at the frontier of magnetospheric research. The picture presented here is my personal view, and other authors may describe these phenomena rather differently. However, first we shall consider how a flux of energetic electrons into the atmosphere gives rise to the emission of auroral light.

Generally, aurora can be observed both on the dayside and nightside, therefore termed nightside aurora and dayside aurora. Dayside aurora is observed between 6-18 MLT (magnetic local time), while nightside aurora is observed between 18-6 MLT. Nightside aurora is stronger and can be observed at a large range of latitudes. Dayside aurora gives weaker emissions compared to the nightside aurora and can only be observed during the polar night.

8.2. Auroral light emission

There are numerous ancient legends and beliefs about the aurora from various parts of the world. Many of the mythological ideas, as well as more scientific theories from the eighteenth century, explained the aurora as sunlight reflected, refracted, or scattered by various divine or natural processes. The first observations of the auroral spectrum by Ångström revealed that it consists of discrete spectral lines. This proved that auroral light is emitted by a gas, and ruled out reflected or scattered sunlight. In 1868 Ångström also measured the wavelength of the strongest line in the auroral spectrum and found 5567 Å. The correct value in SI-units is 557.7 nm.

Auroras are caused by charged particles with energy in the range from 10 eV to 30 keV. These particles collide in the upper atmosphere with atoms and molecules that are left in an excited state after the collision. About 90% of the aurora is caused by electrons, and the rest by protons. At altitudes about 100 km where most of the auroral light is emitted, the atmosphere consists mainly of nitrogen and oxygen as shown in Figure 7.2. In these gases the quantum mechanically allowed transitions that produce strong spectral lines are

8. The Aurora



Figure 8.1.: The aurora. Photo by Torbjörn Lövgren

all outside the visible part of the spectrum—otherwise air would not be transparent and colourless. The visible auroral emissions are due to transitions from forbidden metastable states, with low transition probabilities and long lifetimes. If the gas pressure is too high, these metastable states will relax to the ground state through collisions long before they decay by emitting light. However, at altitudes above 100 km collisions are sufficiently rare to allow emissions from these states. The lifetime of these metastable states determines how structured the aurora appears. A short lifetime allows more structured features while a long lifetime causes the aurora to appear more diffuse.

The strong green line at 557.7 nm is emitted in the decay of an excited state of atomic oxygen. This state has an excitation energy of 4.17 eV and a lifetime of 0.8 s. The peak height of green emission is around 115 km. It decays in a two step process. The green line is emitted in the first step, which takes the atom to a state at 1.96 eV. This intermediate state has a very long lifetime, 110 s, and the red line emission at 630.0 nm that takes it to the ground state mainly occurs at altitudes above 200 km. This red line can also be produced by direct excitation to the state at 1.96 eV by collisions with low energy electrons. Other strong lines in the auroral spectrum are emitted by nitrogen. Molecular N_2^+ ions are created in an excited state by collisions, and they decay to their ground state by emitting a bluish-violet line at 427.8 nm. This state has a lifetime of 70 ns and a peak height at 115 km. Sometimes the lowest part of strong auroras are coloured red by emissions near 600 nm from neutral N_2 molecules.

Figure 8.2 shows the typical altitude distribution of the auroral emissions. The cross section for excitation of the different lines depends on the energy of the incoming auroral particles. Usually, the precipitating particles have higher energy at night than during the day. Low energy electrons give up their energy at higher altitudes, and produce more red emissions. Another difference between day and night is that resonant scattering by sunlight enhances the N_2^+ line at 427.8 nm relative to the green line.

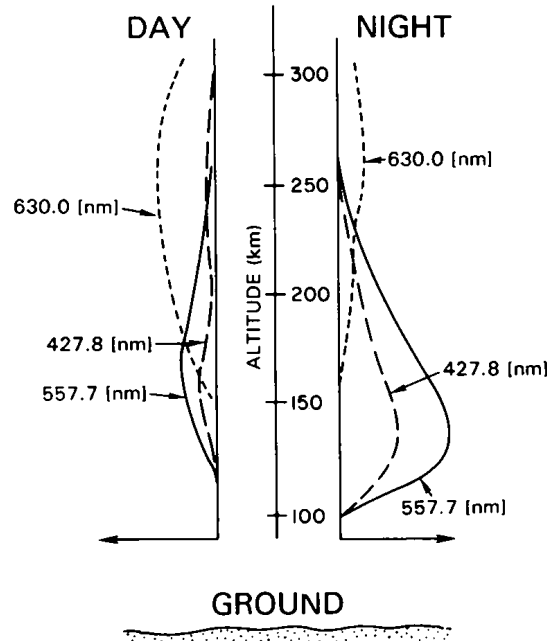


Figure 8.2.: The altitude distribution of auroral emissions at day and night. Figure from Kivelson and Russell [1995a].

8.3. Diffuse aurora

Some diffuse aurora is nearly always found in the auroral zone. On a clear and dark night it is often seen as a diffuse band, which may appear gray if the intensity is below the colour threshold of the eye. Even if no aurora can be seen from the ground, we know from satellite observations of precipitating particles and auroral spectral lines in the UV-range that the diffuse aurora forms a rather continuous band around the auroral zone.

Magnetosheath particles that enter through the magnetopause cusps are a source of diffuse aurora on the dayside. In the cusps there are field lines that connect the ionosphere to the magnetosheath, and magnetosheath particles with sufficiently small pitch angles will precipitate along these field lines. Around local noon, there will be a continuous flux of low energy ($\lesssim 100$ eV) particles, which at altitudes above 200 km cause a diffuse band of emissions at 630 nm.

On the nightside, and far into the evening and morning, diffuse auroras are caused by particles of plasmasheet origin. The energy spectrum of these particles extends from below 100 eV to above 20 keV. The continuous precipitation of electrons from the plasmasheet presents a problem, which has not been completely solved yet. Most of the plasmasheet is on closed field lines, and according to the discussion following Fig. 3.5, the loss cone in a stationary magnetosphere should be empty. To explain the diffuse aurora, we need some mechanism that can scatter the particles into the loss cone.

Plasma waves in the equatorial magnetosphere can cause strong pitch-angle scattering if they are resonant with the particles. Resonant in this context means that the parallel velocity of the particle and the phase velocity of the wave are related so that the particle feels an electric field that oscillates at the gyrofrequency. Significant changes in the velocities of the particles result since they systematically gain or lose energy while they

are resonant. At least for electrons with energies higher than a few keV, the required pitch-angle scattering can be provided by a type of waves in the whistler mode, known as magnetospheric hiss. However, the phase velocity of whistler mode waves may be too high to give resonance with lower energy electrons. Low energy electrons can be scattered efficiently by waves in the upper hybrid mode. Upper hybrid waves are electrostatic waves with frequency $\omega_{UH} \approx (\omega_{pe}^2 + \omega_{ce}^2)^{1/2}$, but it is not clear that these waves are sufficiently common to explain the diffuse precipitation of low energy electrons.

8.4. Auroral waves and ion heating

Plasma waves play several roles on the auroral stage. As we have seen, waves near the equator provide the pitch angle scattering that causes diffuse aurora. Closer to Earth there are several types of plasma waves that are generated by the intense flux of particles precipitating into the aurora.

Discrete auroras are strong sources of whistler mode waves that appear rather different from whistlers generated by lightning. The auroral emissions are characterized by a broadband spectrum covering frequencies from a few kHz to hundreds of kHz, and they are continuously generated with only slow amplitude variations. If the signal is amplified and connected to a loudspeaker a hissing noise is heard, and hence these waves are called auroral hiss. The auroral hiss propagates from the upper ionosphere to the equatorial magnetosphere where it contributes to the magnetospheric hiss. Auroral hiss also propagates to the ground, and at auroral latitudes it will cause noise in radio receivers tuned to low frequencies.

An interesting wave emission connected to discrete auroral arcs is called Auroral Kilometric Radiation (AKR). This is an intense radio wave radiation with peak intensity at frequencies just below 300 kHz, corresponding to a kilometric wavelength. However, the frequency can vary from 50 kHz to 1 MHz. The AKR is generated by a mechanism involving subtle relativistic effects in a low density plasma at a frequency very close to the local electron gyrofrequency. The highest frequencies are generated closest to Earth, where the magnetic field is strongest. The maximum power transmitted as AKR has been estimated to 1 GW, which makes auroras the strongest sources of radio waves from Earth. Electromagnetic radiation at these frequencies can propagate freely out of the magnetosphere where the density is low, but it will be reflected from the ionosphere at the level where the wave frequency equals the plasma frequency. Hence, AKR signals cannot be received on ground, and although these emissions are very strong they were not discovered until 1965. In fact, the corresponding auroral emission from Jupiter, which is known as Jovian decametric radiation and has higher frequency, was observed by radio astronomers ten years earlier. Radio emissions from other planets that show a strong time dependence, such that sometimes the radio source is observed and sometimes not, has been interpreted a sign of AKR and has been used to determine whether a planet has a magnetic field or not.

Electrostatic plasma oscillations and other electrostatic waves are also generated above auroral arcs. Electrostatic waves with low frequencies, in the vicinity of the ion gyrofrequencies, have important effects on the magnetospheric dynamics. These waves are generated by the auroral electrons, but they are damped by the ions. The wave energy absorbed by the ions goes mainly into their perpendicular velocity. The magnetic moment then increases, and the ions are pushed out of the ionosphere by the magnetic mirror force. This process

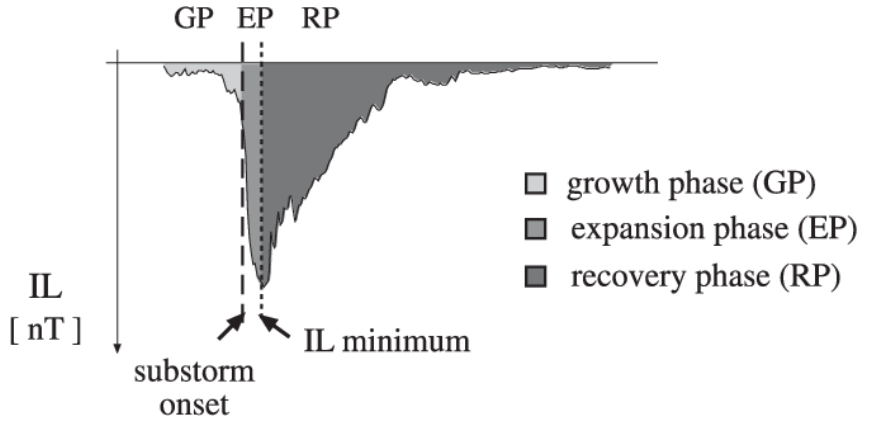


Figure 8.3.: The average magnetic field disturbance at auroral latitudes during a substorm caused by the enhanced westward electrojet. Figure from Tanskanen [2009].

can cause a flow of at least 10^{25} oxygen ions per second from the auroral ionosphere during times of auroral activity. This flow gives the magnetosphere a significant content of ionospheric particles, and reduces the plasma density in the upper ionosphere above auroral arcs. In the auroral cavity dug out by ion heating the electron density is often about $3 \cdot 10^5 \text{ m}^{-3}$ down to altitudes below $1 R_E$, at least when the ionosphere is in darkness. This density reduction is essential for the generation of AKR, and as we shall see also for the generation of discrete auroras.

8.5. Substorms and discrete aurora

As discussed in Chapter 6, a southward component in the interplanetary magnetic field leads to an increased reconnection rate at the magnetopause. More plasma and more magnetic flux is then transported to the magnetotail, and the pressure in the tail lobes increases. The plasma sheet is compressed, which on the ground can be seen as an equatorward motion and brightening of a pre-existing diffuse arc. The increased precipitation from the squeezed plasmasheet will also intensify the ion heating, which leads to lower densities in the upper ionosphere above the auroral arc. As the density of the cross tail current increases and the neutral sheet becomes thinner, an unstable situation builds up. At this stage, reconnection starts in some part of the tail Figure 6.10. This marks the onset of a magnetospheric substorm, and the processes described above are often called the growth phase.

When the magnetic field lines in the tail start to reconnect, part of the plasmasheet will be ejected from the tail and flow Earthward with a speed that often exceeds 400 km/s. The braking of this flow at the inner edge of the plasmasheet requires a substantial $\mathbf{j} \times \mathbf{B}$ -force, and hence a substantial dusk to dawn current. At the edges of the injection region, this current is diverted into field-aligned currents that drive the auroral electrojet, a strong westward current in the ionosphere. The geometry of this circuit is sketched in Figure 8.4. Notice that the auroral electrojet via Ohm's law implies a westward electric field and an equatorward flow of the ionospheric plasma.

A westward flowing current in the ionosphere causes a magnetic field disturbance towards

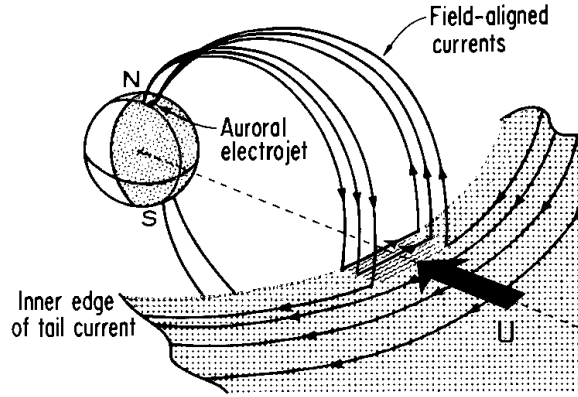


Figure 8.4.: The auroral electrojet current system that forms during substorms. Figure from Kivelson and Russell [1995a].

south. In Figure 8.3, the IL represents the minimum variation from the model magnetic field (negative towards south). This disturbance is typically divided into three phases: the growth phase, the expansion phase, and the recovery phase. The growth phase is identified by a slow decrease in IL. The expansion phase is then marked by a sudden decrease in IL until the maximum negative variation. The expansion is also associated with the auroral break-up when there is a sudden strong increase in aurora. Finally, during the recovery phase IL approaches pre-substorm values again.

The fast flow injected from the tail by reconnection must be stopped by a strong force, corresponding to a substantial $\mathbf{j} \times \mathbf{B}$. As illustrated in Figure 8.4, the perpendicular current associated with this force will be closed by field-aligned currents to the ionosphere. The field-aligned current is carried by electrons that are accelerated to keV energies. These electrons will cause aurora, and the sudden buildup of this current leads to a breakup of the quiet auroral arc that existed during the growth phase of the substorm.

The auroral breakup occurs in the region of upward field-aligned current, at the western end of the electrojet. The upward current is carried by downgoing electrons that have been accelerated by parallel electric fields at altitudes around $1R_E$. When these electrons start to precipitate, a very bright and dynamic auroral display begins. A suitably located observer may see a large part of the sky filled with rapidly moving auroral forms. As illustrated in the classical drawings of an auroral substorm shown in Figure 8.5, the aurora spreads poleward, and after a few minutes more stable discrete auroral arcs start to form.

The inertia of the plasma flowing in from the tail with high speed will carry it into a region where the ambient, mainly magnetic, pressure is higher than the pressure in the injected plasma. As the inflow continues the pressure increases, and the injected plasma will start to expand towards the evening and morning side of the magnetosphere. This happens somewhat further from Earth where the ambient magnetic pressure is lower, on field lines that reach the ionosphere poleward of the initial breakup. The azimuthal plasma flows associated with this expansion cause the discrete arcs that form poleward of the breakup.

The current circuit connected with a discrete auroral arc is shown in Figure 8.6. Comparing with Figure 8.4 we see that the flows have been rotated from radial to azimuthal. The

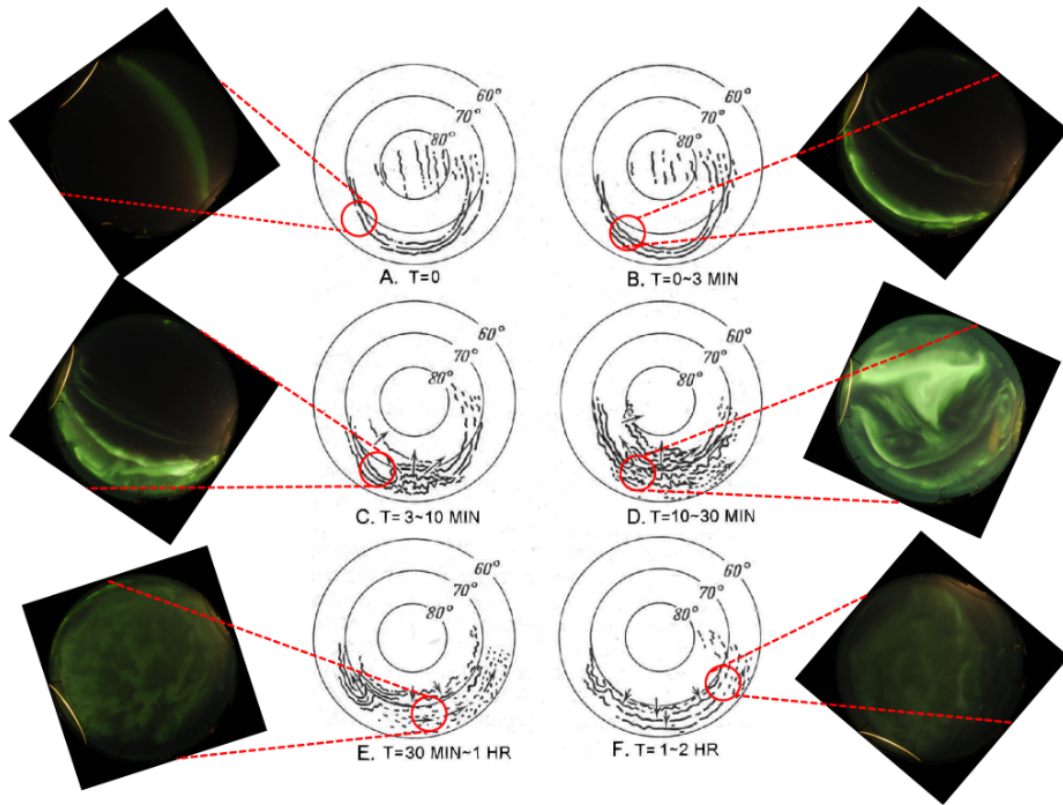


Figure 8.5.: A sketch of the auroral substorm, i.e the auroral forms seen during a magnetospheric substorm. The time $T = 0$ is arbitrary, but roughly at the start of the expansion phase. Figure from Kvammen [2021] (adapted from Akasofu and Meng [1968]).

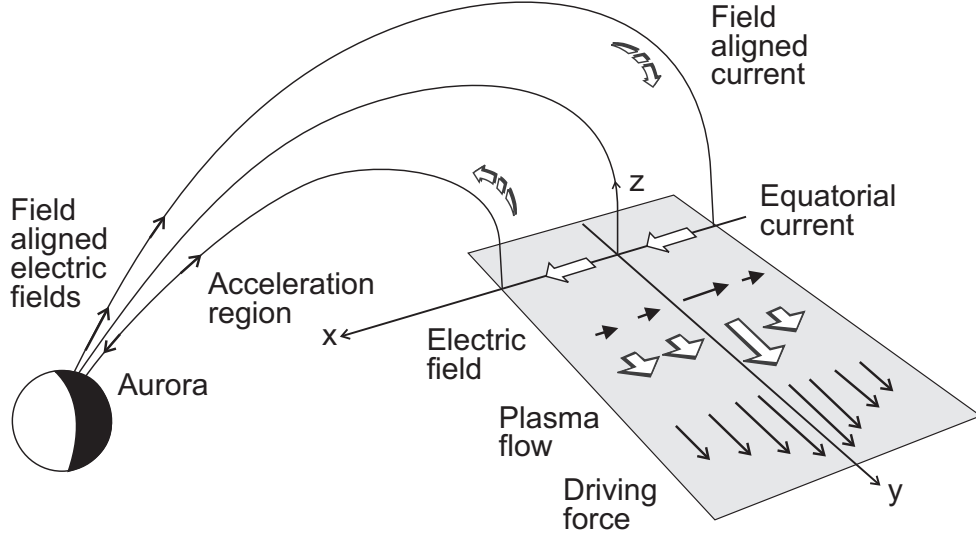


Figure 8.6.: Geometry of the current circuit in a discrete auroral arc. Figure from Rönmark [2002].

length scales are also different, since the auroral electrojet is hundreds of kilometers long and the ionospheric current in Figure 8.6 which runs across an auroral arc is at most a few km. Still the physics is very similar. When pointing out some details, we will here use the geometry shown in Figure 8.6.

Pressure forces in the equatorial plane try to establish a strong velocity shear $\frac{\partial}{\partial x}U_y$, which implies a strong $\frac{\partial}{\partial x}E_x$ in Figure 8.6. Recalling Equation 7.32 we find that the gradient of the ionospheric electric field is determined by the field-aligned current as

$$\frac{\partial}{\partial x}E_x = \frac{j_z}{\Sigma_P} \quad (8.1)$$

As long as there is no potential drop along the field lines, the magnetospheric and ionospheric electric fields are simply related, and the mapping of a strong velocity shear in the equatorial plane to the ionosphere demands a strong field-aligned current. However, a strong current through the auroral cavity means that the electrons must be accelerated by a potential drop

$$\Delta\phi = \frac{m_e}{2e^3} \frac{j_z^2}{n^2} \quad (8.2)$$

The appearance of this potential drop on field lines that carry currents up from the ionosphere means that the ionospheric and magnetospheric electric fields become decoupled. The magnetic field is not frozen-in in the acceleration region, and this breakdown of ideal MHD also means that the equatorial plasma is free to flow without dragging the ionosphere along. The potential drop $\Delta\phi$ in Equation 8.2 limits the current density, and hence the braking $\mathbf{j} \times \mathbf{B}$ -force. Notice that it is the low altitude acceleration in this potential drop, which in the energy spectrum of the precipitating electrons is observed as a peak at the energy $e\Delta\phi$, that distinguishes the discrete from the diffuse aurora.

The field lines carrying current up from the ionosphere connect to regions with a strong inward gradient of the driving force, which is the outer part of the volume with enhanced pressure. When the brakes in this region are released, the flow will at larger y build up a

narrower region of high pressure. Only the outer part of this smaller volume will continue to expand, and this process quickly creates a narrow flow channel in the y direction. The outer edge of this channel maps to a discrete auroral arc in the ionosphere.

In the model described above, the flows are constrained to the y -direction. In the real magnetosphere, the flows can be deviated in the radial, x , direction to form curls and vortices. As illustrated in Figure 8.5, the geometry of real auroral arcs is very complicated. The patterns are similar to the turbulence seen when water flows from a river into a lake. Although the complex pattern formed by auroral arcs cannot be predicted in detail, I hope that you now have gained some understanding of how an individual discrete arc is formed.

8.6. Short summary

The next time you watch an aurora, you should think about how the electrons must be accelerated in the auroral cavity in order to carry the current that stops the plasma injected from the tail where it has been accelerated by the reconnection that starts when the plasma sheet is compressed due to the increased convection driven by enhanced dayside reconnection caused by a southward component in the interplanetary magnetic field carried by a supersonic solar wind from the corona which is heated by oscillations of the field lines extending from the convection zone where they are rocked by the plasma that boils when it absorbs the heat radiated by fusion processes in the solar core.

9. Space Weather

Space weather is the physical and phenomenological state of natural space environments. The associated discipline aims, through observation, monitoring, analysis and modeling, at understanding and predicting the state of the Sun, the interplanetary and planetary environments, and the solar and non-solar driven perturbations that affect them, and also at forecasting and nowcasting the potential impacts on biological and technological systems. – COST Action 724

9.1. Introduction

Space weather describes the current conditions in space which are primarily influenced by the Sun's activity (see the official description above). The Sun's activity comprises the solar wind, coronal mass ejections (CMEs), and solar flares (sudden bursts of electromagnetic radiation). Especially, coronal mass ejections and solar flares are bursty events releasing large amounts of energy into space. When such energetic events interact with the Earth's magnetic field they cause abrupt changes in the Earth's magnetosphere and ionosphere which can affect various aspects of our daily lives, see Figure 9.1. Such events can, for example, disturb radio waves used for navigation, decrease the accuracy of GPS signals, induce currents in power lines, and cause radiation damage on spacecraft and satellites. A less threatening aspect of space weather is an increase in auroral activity.

One of the most extreme space weather events known was the **Carrington Event** on 1-2 September 1859. British astronomers recorded a strong and intense flare at the Sun's surface about a day prior to the event. During the event, aurora was visible at very low latitudes such as Mexico and southern Japan. Additionally, geomagnetically induced currents affected the telegraph system in Europe and North America. Telegraph operators reported getting electric shocks from the equipment and the network continued to operate even without connected batteries.

Even though the Carrington Event was a long time ago, similar extreme space weather events happened more recently, for example the Halloween Storm in 2003 and the Mother's Day Storm in 2024. The Halloween Storm in 2003 caused a power outage in southern Sweden, one of the spacecraft at L1 was damaged, and astronauts on board the ISS needed to stay in parts of the space station that are more shielded to be protected from radiation. During the Mother's Day storm, svenska kraftnät, the state authority responsible for Sweden's transmission system for electricity, was on high alert. The GPS signal was compromised which caused some drones to crash and guiding systems in tractors to not function properly so agricultural activities had to be delayed.

Due to these various effects of extreme space weather events, understanding and forecasting solar activity to predict space weather has become of increasing interest. Events like the Carrington Event are rare but would significantly impact today's technology in space

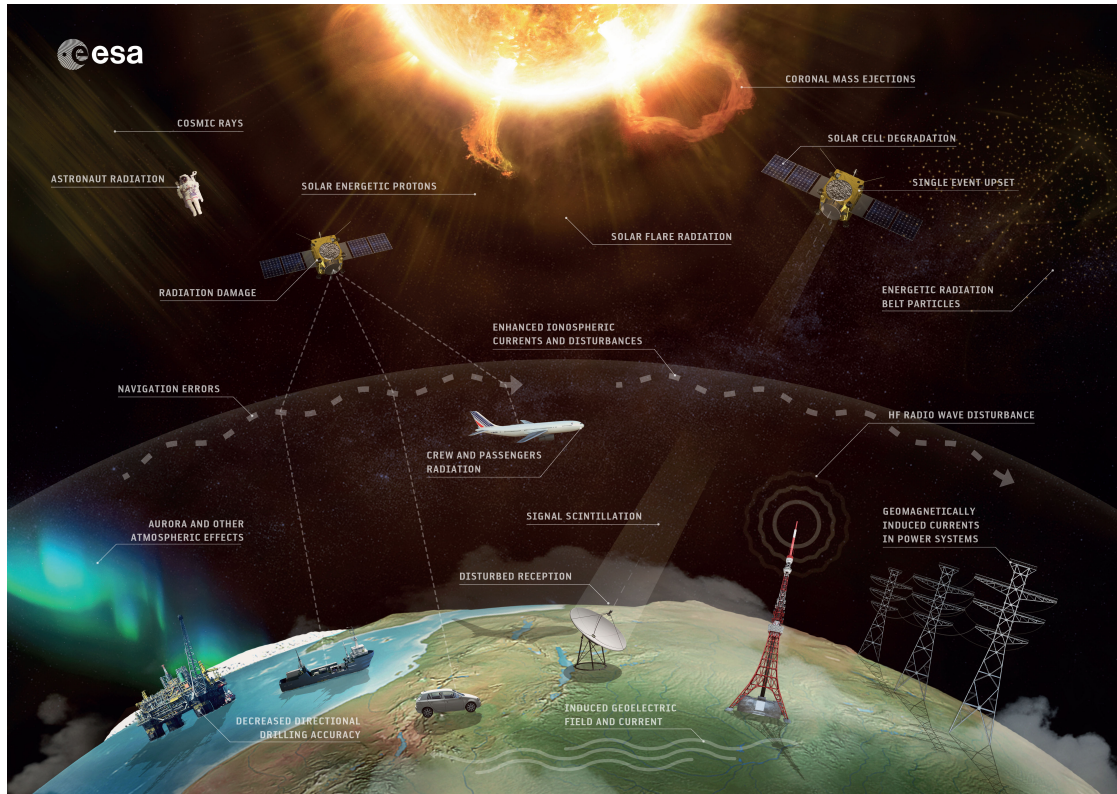


Figure 9.1.: Space Weather effects. Credits: ESA/Science Center

and on the ground. Especially, higher latitudes are usually more affected since they are connected via magnetic field lines to the outermost regions of the magnetosphere.

Even though the terminology used in space weather is similar to weather as we know it, the causes and effects have little in common. Extreme space weather events are rare, just like extreme weather events, but if they happen, they can cause a lot of damage.

9.2. Geomagnetic storms

The most severe space weather events are geomagnetic storms. Geomagnetic storms occur when CMEs interact with the Earth magnetic field. Due to their high velocity and density they compress the Earth's magnetosphere. In addition they carry strong magnetic fields that can transfer more energy than usual into the magnetosphere through magnetic reconnection on the dayside. The occurrence of geomagnetic storms is strongly correlated with the solar cycle (see chapter 5 for details). The solar activity is high during the maximum of a solar cycle, when there are many sunspots on the Sun's surface. That means that the occurrence rate of extreme events such as CMEs and solar flares increase, which in return increases the number of geomagnetic storms.

The intensity of a geomagnetic storm is measured with the K_p index (see section 12.4 for details). The index measures the geomagnetic activity and can take values between 0-9, and 5-9 is considered to be a geomagnetic storm. Figure 9.2 gives the different levels of geomagnetic storms, their effects, and the average occurrence rate as defined by the Space Weather Prediction Center (part of the National Oceanic and Atmospheric Administration in the USA).



NOAA Space Weather Scales



Category		Effect	Physical measure	Average Frequency (1 cycle = 11 years)
Scale	Descriptor	Duration of event will influence severity of effects		
Geomagnetic Storms			Kp values* determined every 3 hours	Number of storm events when Kp level was met; (number of storm days)
G 5	Extreme	Power systems: widespread voltage control problems and protective system problems can occur, some grid systems may experience complete collapse or blackouts. Transformers may experience damage. Spacecraft operations: may experience extensive surface charging, problems with orientation, uplink/downlink and tracking satellites. Other systems: pipeline currents can reach hundreds of amps, HF (high frequency) radio propagation may be impossible in many areas for one to two days, satellite navigation may be degraded for days, low-frequency radio navigation can be out for hours, and aurora has been seen as low as Florida and southern Texas (typically 40° geomagnetic lat.).**	Kp=9	4 per cycle (4 days per cycle)
G 4	Severe	Power systems: possible widespread voltage control problems and some protective systems will mistakenly trip out key assets from the grid. Spacecraft operations: may experience surface charging and tracking problems, corrections may be needed for orientation problems. Other systems: induced pipeline currents affect preventive measures, HF radio propagation sporadic, satellite navigation degraded for hours, low-frequency radio navigation disrupted, and aurora has been seen as low as Alabama and northern California (typically 45° geomagnetic lat.).**	Kp=8, including a 9-	100 per cycle (60 days per cycle)
G 3	Strong	Power systems: voltage corrections may be required, false alarms triggered on some protection devices. Spacecraft operations: surface charging may occur on satellite components, drag may increase on low-Earth-orbit satellites, and corrections may be needed for orientation problems. Other systems: intermittent satellite navigation and low-frequency radio navigation problems may occur, HF radio may be intermittent, and aurora has been seen as low as Illinois and Oregon (typically 50° geomagnetic lat.).**	Kp=7	200 per cycle (130 days per cycle)
G 2	Moderate	Power systems: high-latitude power systems may experience voltage alarms, long-duration storms may cause transformer damage. Spacecraft operations: corrective actions to orientation may be required by ground control; possible changes in drag affect orbit predictions. Other systems: HF radio propagation can fade at higher latitudes, and aurora has been seen as low as New York and Idaho (typically 55° geomagnetic lat.).**	Kp=6	600 per cycle (360 days per cycle)
G 1	Minor	Power systems: weak power grid fluctuations can occur. Spacecraft operations: minor impact on satellite operations possible. Other systems: migratory animals are affected at this and higher levels; aurora is commonly visible at high latitudes (northern Michigan and Maine).**	Kp=5	1700 per cycle (900 days per cycle)

* Based on this measure, but other physical measures are also considered.

** For specific locations around the globe, use geomagnetic latitude to determine likely sightings (see www.swpc.noaa.gov/Aurora)

Figure 9.2.: Classification of geomagnetic storms, their effects, and their occurrence rate. Credits: NOAA.

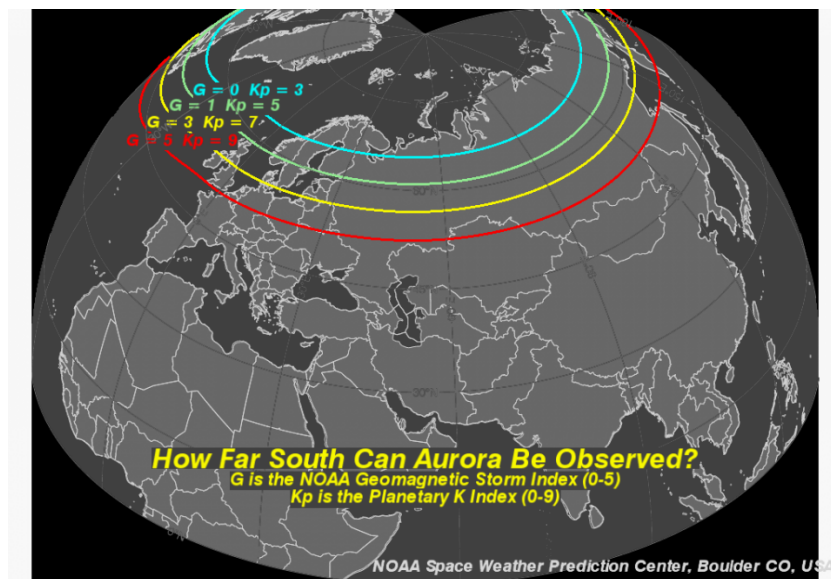


Figure 9.3.: Auroral latitudes depending on the geomagnetic activity. Credits: NOAA.

Where you can see aurora in Europe depends on the K_P index, see Figure 9.3. In northern Sweden you can see aurora already at very low geomagnetic activity (if there are clear skies), but only during geomagnetic storms, aurora can be observed in southern Sweden and mainland Europe.

9.3. Space Weather effects

Space weather has a large range of effects. We will limit our more detailed discussion to geomagnetically induced currents, compromises in the GPS signal, and increased radiation. Other effects are, for example, an increased drag on spacecraft causing them to lose altitude, or in the worst case to re-enter the atmosphere (this happened to 38 Starlink satellites in 2022).

9.3.1. Geomagnetically induced currents

Faraday's law of induction states that a time-varying magnetic field induces an electric field causing a current to flow in a conducting wire. Similarly, an electric current produces a magnetic field. In chapter 6 the current systems in the polar ionosphere were described. During geomagnetic storms these current systems exhibit large temporal and spatial variations causing rapid changes in the magnetic field created by field-aligned currents. These rapid changes induce currents in long metal structures, such as the electrical power grid. Compared to the 50 Hz frequency, the changes in the ionosphere are slow, adding a DC current to the AC current in the power grid. This additional current in the power grid can cause transformers (not the robots transforming into cars, the electrical component transforming voltages from one value to another), to operate outside the intended range which can damage them. Due to the additional DC current, the transformer can go into saturation during half of a cycle which produces harmonics in the AC waveform and make power transmission more inefficient. For example, during the Halloween storm in 2003, Sweden experiences a power black out because of geomagnetically induced currents.

9.3.2. Navigation

For navigation we are strongly relying on radio navigation systems, like the American GPS system or the European Galileo system. Usually we can determine our position with a few meters accuracy, however, during geomagnetic storms the accuracy decreases to 10s of meters. What is the physical reason behind it?

Navigation satellites send out radio signals that travel through the ionosphere to the receiver. As the signal passes through the ionosphere, it gets refracted changing the speed of the signal. To which degree the speed of the signal changes depends on the Total Electron Count in the ionosphere (this is a measure how strongly ionized the ionosphere is). During times without any major geomagnetic disturbance, the radio signal is compensated for this with models for the Total Electron Count. However, during times with major geomagnetic disturbance, the charge density in the ionosphere increases dynamically changing the Total Electron Count. The radio signal cannot be corrected for these dynamic changes and it will therefore cause uncertainties in the position.

9.3.3. Increased radiation

Some CMEs and solar flares can also accelerate particles to very high velocities close to the speed of light. These high energy particles are solar energetic particles and increased fluxes of these particles lead to a solar radiation storm. Upon reaching the Earth magnetic field, these high-energy particles will be guided along the magnetic field into the atmosphere. These high energy particles are effectively absorbed by the atmosphere so they do not cause harm to human live on Earth. However, they can cause damage to electronics and humans in space. This radiation can penetrate deep and cause damage to DNA and electronics on board of spacecraft. For very intense radiation events some satellites might be turned of to increase the live span of instruments. Passengers and crew on an airplane might also be exposed to higher radiation levels, especially for flights in high latitudes.

Another effect of intense radiation is blackouts in radio communication. High-frequency radio waves (defined as electromagnetic waves with frequency 3-30 MHz) are reflected and refracted in the ionosphere (in the F-region) under normal conditions. Due to refraction in the ionosphere it is possible to communicate over long-distances where line-of-sight communication is not possible. There are two mechanisms that can disturb high-frequency radio communication. One of the mechanisms are the high-energy particles which can penetrate the ionosphere. In the ionosphere they interact with neutral particles creating free electrons and charged ions, creating a plasma at lower altitudes than normal (in the D-region of the ionosphere). The second mechanisms is the radiation from a solar flare itself, which also causes increased ionization at lower altitudes of the ionosphere (the electromagnetic radiation impacts first compared to energized particles). If the lower altitudes of the ionosphere (the D-region) are stronger ionized, this region will absorb high-frequency radio waves instead of deflecting or transmitting the waves. This absorption causes a blackout in radio communication.

Blackouts in radio communication affects the sunlit side of Earth stronger since it is directly exposed to incoming radiation from the sun. Similar effects can occur on the nightside during auroral display when precipitating electrons cause enhanced ionization in the ionosphere.

9.4. Short summary

Considering the previously mentioned effects of severe space weather events on our daily lives, it is of no surprise that space weather forecasting has become of great importance. Spacecraft positioned at the Lagrange 1 point have been of great advantage as they can give us a approximately 30 minute warning for any incoming disturbances. There is also an increasing effort in analyzing images of the sun to make predictions 1-3 days ahead.

Part III.

Space engineering

10. Mechanics of Rockets and Satellites

10.1. Introduction

Most of our knowledge of space physics derives from measurements by instruments carried on spacecraft around the Earth and to various parts of the solar system. This chapter therefore intends to convey some basics of spaceflight mechanics.

The number density of the gas and plasma in space is low compared to the density of the air on Earth. The drag on a moving object relative to the background plasma is therefore small and can often be neglected. In addition, the interaction between charged particles and electromagnetic fields which is at the focus of all other chapters is usually unimportant for the motion of macroscopic bodies like spacecraft even if they acquire some electric charge. In consequence, spacecraft in the solar system move mainly under the influence of the gravitation from the Sun and nearby planets. The principles of spaceflight in this environment is the main topic of this chapter.

However, in Sections 10.2 – 10.4, we first consider how to control the motion of a spacecraft by application of some force on it at our will. The most important means to do so is by the rocket principle. The term "rocket" includes small *thrusters* on a spacecraft already in space, applying a usually not very high force to change the spacecraft trajectory, as well as large *launchers*, lifting spacecraft from the ground into space.

10.2. Rockets in Free Space

The reaction principle is the key to modern space technology. Experiments with rockets, driven by black powder, were probably started in China almost a thousand years ago, and shortly after 1200 they were an important weapon in the wars against Genghis Khan and the Mongols. In Europe, rockets have been used in wars as well as in fireworks for celebrations since the middle ages. Hence, practical experience with rockets existed long before Newton formulated the underlying theory. Still, when Jules Verne wrote *From the Earth to the Moon* in 1865, he did not realize that a rocket was more practical than a gun to launch a space vehicle.

A rocket works by ejecting mass at a high exhaust velocity v_e . By conservation of momentum, the rocket body itself must accelerate in the opposite direction. Typically, v_e is around 3000 m s^{-1} but this depends on the fuel and the engine. To see how this works, we consider the system illustrated in Figure 10.1. Adopting a reference frame in which the rocket velocity v is antiparallel to the direction of the gas exhaust, the motion is in one dimension only. We may then write the momentum of the rocket at time t as

$$p(t) = m(t) v(t). \quad (10.1)$$

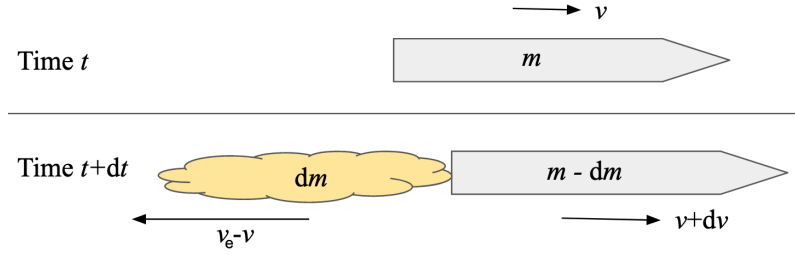


Figure 10.1.: Between times t and $t + dt$, the mass dm is ejected from the rocket. The ejecta leave the rocket at relative speed v_e .

A short time dt later, the rocket will have used a small amount of fuel dm and gained some velocity dv in the opposite direction. The ejected fuel moves at speed v_e with respect to the rocket and therefore at velocity $v - v_e$ in the chosen reference frame. By momentum conservation,

$$mv = (m - dm)(v + dv) + dm(v - v_e), \quad (10.2)$$

and therefore

$$mv = mv + m dv - v dm - dm dv + v dm - v_e dm. \quad (10.3)$$

Neglecting the second order term $dm dv$ the equation above simplifies to

$$m dv = v_e dm. \quad (10.4)$$

Up to now, we have considered dm to be positive. For dm to represent the change of the rocket mass $m(t)$, which decreases as the rocket burns its fuel, it must be negative. We therefore replace $dm \rightarrow -dm$ and so (10.4) becomes

$$m dv = -v_e dm. \quad (10.5)$$

We may use this in a couple of ways. First, by dividing by dt we find the equation of motion for the rocket,

$$m \frac{dv}{dt} = -v_e \frac{dm}{dt}. \quad (10.6)$$

The right hand side represents a force known as the rocket *thrust*

$$T = \left| v_e \frac{dm}{dt} \right|. \quad (10.7)$$

Second, Equation 10.5 can be rearranged for integration as

$$\frac{1}{v_e} \int_{v_0}^{v_1} dv = - \int_{m_0}^{m_1} \frac{dm}{m} \quad (10.8)$$

with indices 0 and 1 indicating values before and after the rocket burn. The integrals evaluate as

$$\frac{v_1 - v_0}{v_e} = - \ln \frac{m_1}{m_0} = \ln \frac{m_0}{m_1}. \quad (10.9)$$

Putting $\Delta v = v_1 - v_0$ as the total change of rocket speed for the use of fuel mass $\Delta m = m_0 - m_1$, we can rewrite (10.9) as

$$\Delta v = v_e \ln \left(1 + \frac{\Delta m}{m_1} \right). \quad (10.10)$$

This equation is often referred to as *the rocket equation*, sometimes with the name Tsiolkovsky attached to it in honour of its originator. It exposes one of the fundamental problems of rocket propulsion, more clearly seen if we solve it to provide the fuel mass Δm you need as function of the velocity boost Δv you want:

$$\Delta m = m_1 \left[\exp \left(\frac{\Delta v}{v_e} \right) - 1 \right]. \quad (10.11)$$

For a small change of spacecraft speed $\Delta v \ll v_e$, Taylor expansion of the right hand side yields $\Delta m/m_1 \approx \Delta v/v_e$. So for such small changes, the fuel you need increases linearly with the speed change you wish. This is not so bad: to get double speed you are probably prepared to pay double fuel. However, if you need a large boost of speed $\Delta v \gg v_e$, your fuel needs raise exponentially. The physics behind this unfortunate result is clear: the extra amount of fuel needed for later velocity changes needs to be carried during previous velocity changes. So if you need a fast spacecraft, adding more fuel is an expensive and often practically impossible solution. In consequence, interplanetary spacecraft often leave Earth at quite modest velocity and then use some tricks to gain speed en route to their final target, as will be discussed in Section 10.14.

If we let Δm be the full fuel supply carried onboard your spacecraft, the final mass m_1 in (10.10) becomes the mass of the spacecraft itself without fuel, known as the *dry mass*. The rocket equation (10.10) can be used to translate Δm to a total Δv available for all manoeuvres you need to do with the spacecraft during its mission. The total Δv is fully specified by v_e , m_{dry} and Δm which are properties known before launch. When defining a spacecraft mission to some part of the solar system, the total Δv is a crucial parameter.

10.3. Launch Mechanics

When launching rockets from Earth to space, gravity and air friction will complicate the situation. Neglecting air friction, as would be justified for launches from the Moon or other atmosphereless bodies, and considering a strictly vertical launch, we must add a gravity term to the equation of motion (10.6). Counting velocity positive upwards, we then have

$$m \frac{dv}{dt} = T - mg. \quad (10.12)$$

where T is the rocket thrust (10.7). A first and very obvious conclusion from this equation is that for the rocket to get off the ground, its thrust must overcome gravity: $T > mg$. For constant v_e and g , this equation would just lead to the addition of a term $-gt_{\text{burn}}$ on the right hand side of the rocket equation (10.10). More generally, we can write

$$\Delta v = \int_{v_0}^{v_1} dv = \int_{t_0}^{t_1} \frac{T}{m} dt - \int_{t_0}^{t_1} g dt \quad (10.13)$$

where t_0 and t_1 are the start and end times of the rocket burn. So not only must $T > mg$, it should preferably be very high so the time $t_{\text{burn}} = t_1 - t_0$ is kept as short as possible, to minimize the value of the last term. However, when air friction is taken into account, it turns out to be best not to achieve too high velocity early on when still in the dense lower atmosphere. Finding the optimal variation of launcher thrust with height requires numerical simulations including a model for the air friction at various altitudes.

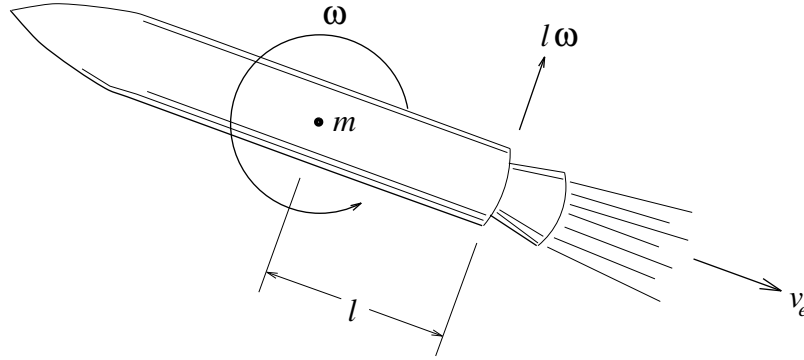


Figure 10.2.: Stability of a rocket in vacuum. Figure after Thomson [1986].

The rocket equation (10.10) tells us fuel economy is essential. Assuming we want to launch into an orbit in or near the equatorial plane, we can get some speed for free by using the eastward velocity any object on Earth has because of the rotation of the planet around its axis. This is highest at the equator, about 0.5 km s^{-1} , which is a main reason why many launch sites are located at low latitudes and also explains why rockets often can be seen to turn eastward soon after launch. For example, ESA's spaceport is located in French Guiana roughly 500 km north of the equator while the main NASA launch site is in Florida, the southernmost state of the mainland USA. Rockets launched eastward from such spaceports gain a natural additional launch velocity.

10.4. Stability of a rocket

In the air, the stability of a rocket is mainly determined by aerodynamics. When the axis of the rocket is tilted from the direction of motion, the pressure increases on the forward side. The resulting force acts along a line that crosses the axis of the rocket at the center of pressure. If the center of pressure is in front of the center of mass, the resulting torque will increase the angle between the rocket and the velocity. In this case the rocket is unstable and its flight is unpredictable, except that it will probably be a failure. Hence, a rocket should be designed to have its center of pressure behind its center of mass. The torque created by the pressure then quickly realigns the rocket with the velocity, and the flight of the rocket is stable. The remedy for instability has been known since ancient times by, for example, archers who want their arrows to fly straight. In practice, the center of pressure is moved sufficiently far back to achieve stability by adding wings near the rear of the rocket.

In space, above the atmosphere, wings have no effect and the stability is determined by other factors. From intuition based on trying to balance a pencil on a fingertip while accelerating it upwards, one may think a narrow rocket with the nozzle at the end would be inherently unstable. As we will now see, its ejection of mass in fact makes it stable.

Let the moment of inertia about a transverse axis through the center of mass be I . We can write the moment of inertia as $I = mk^2$, where k is an effective radius of gyration. For example, approximating the rocket in Figure 10.2 with a narrow cylinder with uniform mass distribution would mean $k \approx l/\sqrt{3}$. If the rocket rotates about the transverse axis with angular velocity ω , as shown in Figure 10.2, its angular momentum is $I\omega$. If the distance from the center of mass to the nozzle is l , the nozzle acquires a

transverse velocity ωl . As in section 10.2 we write the mass emitted during a short time dt as $-dm$. The exhaust gas then carries away an angular momentum $\omega l^2 dm$ from the rocket during this time, so by conservation of angular momentum,

$$\frac{d(I\omega)}{dt} - \frac{dm}{dt} \omega l^2 = 0. \quad (10.14)$$

Substituting for I and carrying out the differentiation we obtain

$$I \dot{\omega} + \omega \left[\dot{m} (k^2 - l^2) + 2 m k \dot{k} \right] = 0 \quad (10.15)$$

where the dots as usual are shorthand for the time derivative d/dt . Because the mass distribution changes slowly compared to the total mass we can neglect the $\dot{k}$ term and simplify to

$$\frac{\dot{\omega}}{\omega} = \frac{\dot{m}}{m} \left(\frac{l^2}{k^2} - 1 \right). \quad (10.16)$$

The rotation will slow down if $\dot{\omega} < 0$, and in this case the rocket is stable. Since the rocket is losing mass $\dot{m} < 0$, so the condition for stability becomes $l > k$. This is easily fulfilled in conventional designs with the motor at the rear of the rocket, but a space vehicle with the nozzle near the center of mass may become unstable.

Notice that there is an important difference between the two stability criteria discussed in this section. In air, the pressure forces will push the rocket back to an equilibrium position aligned with the velocity. If the rocket has an angular momentum when it comes back to its proper position, it may start to oscillate and we have not checked that these oscillations would be damped. In vacuum, on the other hand, we have found a damping mechanism, but if the rocket is tilted it will by itself realign with the velocity vector.

10.5. Motion Under a Central Force

After burnout of the rocket, the motion of a man made space vehicle follows the same laws of celestial mechanics as any other object in space. In most contexts, the most useful results still are the three laws of planetary motion deduced by Johannes Kepler in the early 17th century from analysis of observations of the planets. In the last decades of the same century, Newton could deduce Kepler's laws from his own general laws of motion and gravitation. This made clear that physics on Earth and in the heavens followed the same laws, provided insight into the range of validity of Kepler's results, and enabled the detailed calculations of modern celestial mechanics.

Kepler's laws apply strictly only to two-body problems, like a planet moving around the Sun or a satellite around the Earth. For most problems in spaceflight mechanics, this is an excellent approximation. There are situations when three-body interactions have to be taken into account, like the influence of the Moon on satellites around the Earth. Nevertheless, these can often be handled as perturbation to Kepler orbits. Detailed numerical calculations of many-body problems, for which no exact complete solutions exist, rarely needs to enter the early stages of planning of a space mission, but an understanding of Kepler orbits is helpful from the very start.

A force always directed towards a fixed point is called a central force. Notice that two-body problems can be reduced to this form by choosing the origin at the center of mass. For problems involving a spacecraft and a celestial body like the Sun or a planet, the

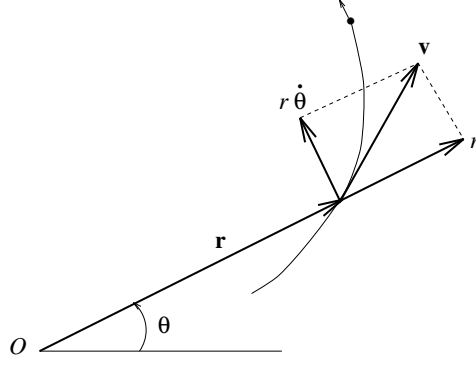


Figure 10.3.: Radial and transverse components of the velocity. Figure after Thomson [1986].

motion of the latter can be neglected and the centre of mass taken to be at the centre of the body. An obvious but important property of central forces is that the motion of any particle is confined to a constant plane spanned by the radius vector $\mathbf{r}$ from the central body and the particle velocity $\mathbf{v}$ at any arbitrary time. We can therefore analyze central force motion using the standard plane polar coordinates r and θ as in Figure 10.3.

Since the force acts along a line through the origin, the torque

$$\mathbf{M} = \mathbf{r} \times \mathbf{F} = 0. \quad (10.17)$$

Changes in the angular momentum are caused by a torque, so from Equation 10.17 it follows that the angular momentum is constant:

$$\mathbf{L} = m\mathbf{r} \times \mathbf{v} = \text{const.} \quad (10.18)$$

We introduce the angular momentum per unit mass,

$$h = L/m = |\mathbf{r} \times \mathbf{v}| = r^2\dot{\theta} \quad (10.19)$$

often referred to as the *specific angular momentum*. As the area dA swept out by the radius vector of the particle during a short time dt is a very narrow triangle with long sides r and short side $r\dot{\theta}dt$ we find that $dA/dt = h/2 = \text{constant}$. This is the exact content of *Kepler's second law* which therefore is equivalent to conservation of angular momentum.

The gravitational force from a mass M on a particle of mass m is

$$\mathbf{F} = -\frac{GMm}{r^2} \hat{\mathbf{r}}, \quad (10.20)$$

so the equation of motion of the particle is

$$m\dot{\mathbf{v}} + \frac{GMm}{r^2} \hat{\mathbf{r}} = 0. \quad (10.21)$$

In polar coordinates (Figure 10.3), we can write the velocity vector as $\mathbf{v} = \dot{r}\hat{\mathbf{r}} + r\dot{\theta}\hat{\boldsymbol{\theta}}$. Recalling the unit vectors $\hat{\mathbf{r}}$ and $\hat{\boldsymbol{\theta}}$ depend on t , you can verify that $d\hat{\mathbf{r}}/dt = \dot{\theta}\hat{\boldsymbol{\theta}}$ and $d\hat{\boldsymbol{\theta}}/dt = -\dot{\theta}\hat{\mathbf{r}}$. Equation 10.21 can then be split into radial and transverse components as

$$\ddot{r} - r\dot{\theta}^2 + \frac{GM}{r^2} = 0 \quad (10.22)$$

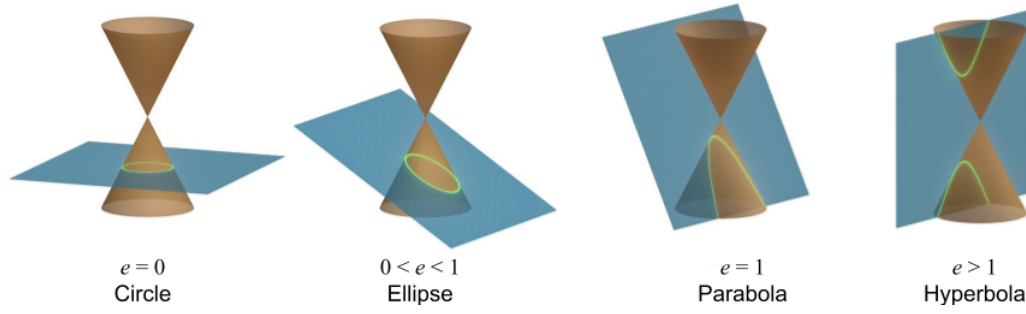


Figure 10.4.: Conic sections of increasing eccentricity from left to right. The two left cases describe bound periodic orbits, the two at right open trajectories. Adapted from work by user Seahen at Wikimedia Commons.

$$r\ddot{\theta} + 2\dot{r}\dot{\theta} = \frac{1}{r} \frac{d}{dt}(r^2\dot{\theta}) = 0. \quad (10.23)$$

Equation 10.23 is yet another statement of Kepler's second law, or the conservation of angular momentum. We will not attempt to solve Equation 10.22 as a function of time, but relations between r and θ , that is the geometry of the trajectory, can be found if we introduce the new variable $u = 1/r$ and rewrite some of the expressions above in terms of u rather than r :

$$\dot{r} = \frac{d}{dt} \left(\frac{1}{u} \right) = -\frac{1}{u^2} \frac{du}{dt} = -r^2 \dot{\theta} \frac{du}{d\theta} = -h \frac{du}{d\theta} \quad (10.24)$$

$$\ddot{r} = -h \frac{d}{dt} \left(\frac{du}{d\theta} \right) = -h \dot{\theta} \frac{d^2u}{d\theta^2} = -h^2 u^2 \dot{\theta} \frac{d^2u}{d\theta^2} \quad (10.25)$$

$$r\dot{\theta}^2 = \frac{(r^2\dot{\theta})^2}{r^3} = h^2 u^3. \quad (10.26)$$

By means of these relations, Equation 10.22 can be transformed to

$$\frac{d^2u}{d\theta^2} + u = \frac{GM}{h^2}. \quad (10.27)$$

This is a second order linear inhomogeneous ordinary differential equation. Its general solution is the sum of the general solution u_h to the corresponding homogeneous equation (10.27 with RHS = 0) and any particular solution u_p to the inhomogeneous equation. The former is $u_h = C \cos(\theta - \theta_0)$ where C and θ_0 are integration constants, and inspection of Equation 10.27 shows that $u_p = GM/h^2$ is a particular solution. Hence, we have $u = C \cos(\theta - \theta_0) + GM/h^2$ where the constants C and θ_0 can be determined from the initial conditions. The angle θ_0 just specifies the orientation of the curve in the plane so we can always choose the coordinate axes to get $\theta_0 = 0$. If we introduce the quantities

$$p = \frac{h^2}{GM} \quad (10.28)$$

$$e = pA = \frac{h^2}{GM} C \quad (10.29)$$

we can finally write the solution as

$$r = \frac{p}{1 + e \cos \theta}. \quad (10.30)$$

This is the equation of a *conic section*, a class of curves obtained by cutting a double cone by a plane at varying angle and position as shown in Figure 10.4. The shape of

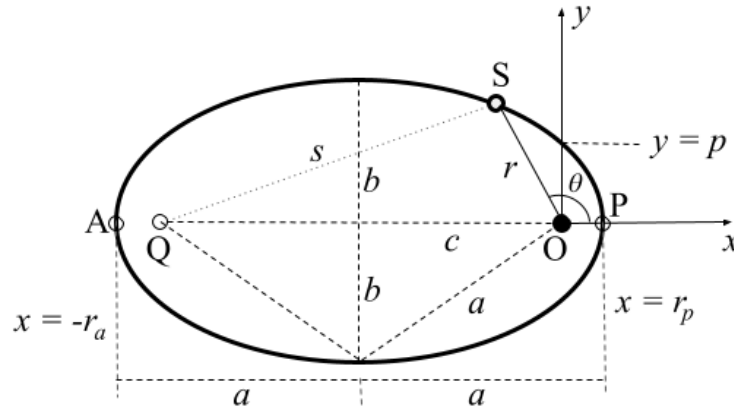


Figure 10.5.: Geometry of an elliptic orbit of a spacecraft S around a central object at the origin O which is placed in one of the focal points of the ellipse. The other focal point Q , separated from O by a distance $2c$, is empty. The sum of the distances from S to the focal points is always constant ($r + s = 2a$) so to draw an ellipse you can attach a string to the two focal points and stretch the string with a pencil you move around on a paper. The major and minor axes of the ellipse are $2a$ and $2b$, respectively. The points A and P are the apoapsis and periapsis of the orbit, at distances r_a and r_p from the central body, respectively. The semi-latus rectum $p = (1 - e^2)a$ of Equation 10.30 has geometrical interpretation as shown, but in practice the semi-major axis a is a more useful size parameter for an elliptic orbit.

such a curve is defined by the dimensionless *eccentricity* e and the size by the parameter p known as the semi-latus rectum. Equation 10.30 shows that for $0 \leq e < 1$ every θ gives a finite r , so in this case the trajectory described by this curve is a closed *orbit* in which the particle performs periodic motion. For the special case $e = 0$ we get a circular orbit as $r = p = \text{constant}$. For $0 < e < 1$, we get an ellipse (Figure 10.5). This means we have derived *Kepler's first law*, stating that a planetary (or other) orbit is an ellipse with the central body at one of its two focal points. We can note that this law is valid only for a $1/r^2$ central force¹, in contrast to the second law (see discussion around Equation 10.19) which holds for any central force.

For $e \geq 1$, Equation 10.30 shows that $r \rightarrow \infty$ for some value of θ . We have an open *hyperbolic trajectory*, leaving the vicinity of the central body to never come back again. The limiting case $e = 1$ is a parabola, also open.

10.6. Elliptic Orbits

As shown above, bound orbits in a Newtonian gravity field are ellipses, with the circle as a special case of the ellipse with $e = 0$. Elliptic orbits, or Kepler orbits, are therefore of highest interest for space applications.

The shape of the orbit is specified by the eccentricity e which measures how elongated

¹While our derivation only is valid for a $1/r^2$ force, we have actually not shown that other forces could not give the same result. However, one may show that the only other central force giving conical sections is a spring-like force $\mathbf{F} = -k\mathbf{r}$, of little interest for space applications.

the ellipse is. The geometric meaning of e is as the ratio of the distance $2c$ between the focal points to the major axis $2a$, i.e.

$$e = \frac{c}{a}. \quad (10.31)$$

The point of minimum distance to the central body is known as the *periapsis*, denoted by P in Figure 10.5. These are generic terms, valid in any context, while *perigee/apogee* and *perihelion/aphelion* are used for orbits around the Earth and Sun, respectively. Looking at Equation 10.30, we can see that this point is obtained for $\theta = 0$, yielding a periapsis distance $r_p = p/(1 + e)$. In the same way we find the apoapsis distance $r_a = p/(1 - e)$. A glance at Figure 10.5 shows that $r_p + r_a = 2a$, so

$$a = \frac{1}{2}(r_p + r_a) = \frac{1}{2} \left[\frac{p}{1+e} + \frac{p}{1-e} \right] = \frac{1}{2} \cdot \frac{p(1-e) + p(1+e)}{1-e^2} = \frac{p}{1-e^2}. \quad (10.32)$$

We can then write the distances of periapsis and apoapsis as

$$r_p = \frac{p}{1+e} = a(1-e) = \frac{h^2}{GM(1+e)} \quad (10.33)$$

$$r_a = \frac{p}{1-e} = a(1+e) = \frac{h^2}{GM(1-e)}. \quad (10.34)$$

where we also added the connection to the physical parameters provided by Equation 10.28. By any of these equations, we can express the specific angular momentum as

$$h = \sqrt{GMa(1-e^2)}. \quad (10.35)$$

The period T of an elliptic orbit can be found by considering the conserved angular momentum. Rewriting Equation 10.19 as $dt = r^2 d\theta/h$ we can calculate the orbit period by integrating over a full orbit ($0 \leq \theta < 2\pi$). As noted on page 120, the surface covered by the radius vector during motion from θ to $\theta + d\theta$ is a triangle of area $r^2 d\theta/2$, so the integral of $r^2 d\theta/2$ over a full orbit will give the area A of the ellipse. This is also given by the standard formula $A = \pi ab$. Here b is the semi-minor axis which by Figure 10.5 and Equation 10.31 can be calculated from $b^2 = a^2 - c^2 = a^2(1 - e^2)$. Thus

$$T = \oint dt = \frac{1}{h} \int_0^{2\pi} r^2 d\theta = \frac{2A}{h} = \frac{2\pi ab}{h} = \frac{2\pi a^2 \sqrt{1-e^2}}{\sqrt{GMa(1-e^2)}} = 2\pi \sqrt{\frac{a^3}{GM}}. \quad (10.36)$$

This is Kepler's third law, usually formulated as $T^2 \propto a^3$. We may note that the period is set by the semi-major axis a alone, with no reference to the eccentricity e .

For a constant gravitational field the sum of kinetic energy and potential energy is conserved. As was the case for the angular momentum, it is convenient to work with the energy per unit mass or *specific energy*,

$$\varepsilon = \frac{1}{2}v^2 - \frac{GM}{r} = \text{constant}. \quad (10.37)$$

As ε is constant we can find its value by calculating it at an arbitrary point on the ellipse. Looking at Figure 10.5, evaluating the velocity is particularly simple at P and

At where the velocity has no radial component and the speed therefore is $v = h/r$. At these two points, we have

$$\varepsilon = \frac{h^2}{2r_p^2} - \frac{GM}{r_p} \quad (10.38)$$

$$\varepsilon = \frac{h^2}{2r_a^2} - \frac{GM}{r_a}. \quad (10.39)$$

Eliminating h by taking $r_a^2 \cdot (10.39) - r_p^2 \cdot (10.38)$ results in

$$\varepsilon (r_a^2 - r_p^2) = -GM(r_a - r_p) \quad (10.40)$$

and so we arrive at a final expression for the specific orbit energy,

$$\varepsilon = -\frac{GM}{r_a + r_p} = -\frac{GM}{2a}. \quad (10.41)$$

Just as was the case for the period, the energy of an elliptic orbit is determined by only one geometrical parameter, the semi-major axis a . The semimajor axis is therefore a very important parameter of any orbit.

In Equation 10.37 we used a form of the potential energy which goes to zero at infinity. As a result, the energy of an object in a bound orbit is always negative, in accordance with Equation 10.41.

10.7. Hyperbolic trajectories

An open trajectory is not an orbit in the strict sense of closed periodic motion, but hyperbolic trajectories are sometimes grouped together with elliptic orbits as *Kepler orbits*. Whatever we call them, they are of obvious interest for interplanetary spaceflight. When launching a spacecraft from the Earth toward e.g. Jupiter we clearly wish to put it into a hyperbolic trajectory with respect to the Earth. However, as soon as it leaves the immediate vicinity of our planet the dominating gravitational force on the spacecraft will be the Sun, with respect to which it is most likely on an elliptic orbit (anything else would be very uneconomical). In other words, the spacecraft is launched with sufficient speed to leave the Earth but not the solar system. A planetary swing-by, where a spacecraft passes close to a planet on its way to some distant goal to change its velocity, is another example well described by a hyperbolic trajectory.

The geometry of a hyperbolic orbit is illustrated in Figure 10.6. We will here summarize some basic results without derivation, which proceeds along the same lines as for the elliptic case. The basic equations Equation 10.28 – 10.30 apply to hyperbolae just as well as to ellipses. There are different conventions for how to handle other aspects of the transition to the case $e > 1$. It can be seen that Equation 10.32 will give a negative value for a . Accepting this, we have

$$a = \frac{p}{1 - e^2} < 0. \quad (10.42)$$

With this convention, Equation 10.39 for the total specific energy is still valid, providing a positive value $\varepsilon > 0$ as appropriate for an open trajectory. The limiting case of $e = 1$, corresponding to a parabolic trajectory, has $\varepsilon = 0$. Equation 10.31 needs modification to $e = |c/a|$ as $a < 0$.

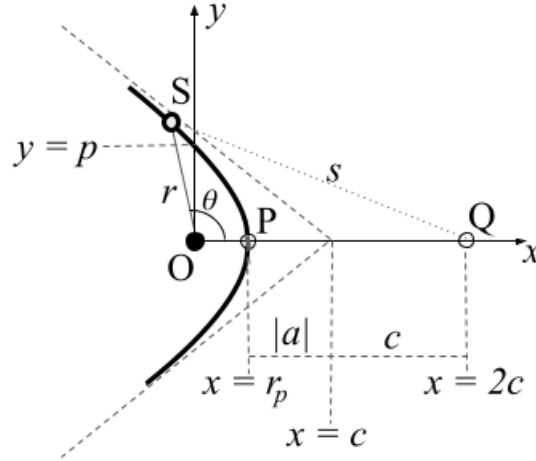


Figure 10.6.: Geometry of the hyperbolic trajectory of a spacecraft S near a central object at the origin O seated in one of the focal points of the hyperbola. The other focal point Q , separated from O by a distance $2c$, is empty. For a hyperbola, the difference between the distances from S to the focal points is always constant ($r - s = 2a < 0$). The point P is the periapsis of the orbit, at distance r_p from the central body. The semi-latus rectum p of Equation 10.30 has the same geometrical interpretation as in the elliptic case (Figure 10.5) As the spacecraft moves away from the central body, its trajectory asymptotically approaches the dashed line.

As the spacecraft moves away from the central body, its trajectory approaches the dashed asymptotes in Figure 10.6. As seen from the central body, the spacecraft approaches a constant value $\theta = \theta_a$ which we find by setting the denominator in Equation 10.28 to zero, yielding $\theta_a = \arccos(-1/e)$. An object arriving on a trajectory with eccentricity e just above 1 will therefore make a very large turn, going out on a course close to opposite to its incoming direction. Correspondingly, a trajectory with $e \gg 1$ will not change direction much during the flyby.

10.8. Determination of satellite orbits from initial conditions.

After the burnout of the launch rocket, we can measure the satellite position $\mathbf{r}_0$ and velocity $\mathbf{v}_0$. From this information we would like to determine the value of the eccentricity e , which determines the type of orbit.

Noticing that $\theta = 0$ at perigee, and using the angle β_0 introduced in Figure ??, we can split the velocity into components as

$$v_0 \cos \beta_0 = r_0 \dot{\theta}_0 = \frac{h}{r_0} \quad (10.43a)$$

$$v_0 \sin \beta_0 = \dot{r}_0 = \dot{\theta}_0 \frac{dr}{d\theta} \quad (10.43b)$$

The first of these equations yields the angular momentum per unit mass $h = r_0^2 \dot{\theta}_0 = r_0 v_0 \cos \beta_0$. In the second equation we use

$$r = \frac{(1 + e) r_p}{1 + e \cos \theta} \quad (10.44)$$

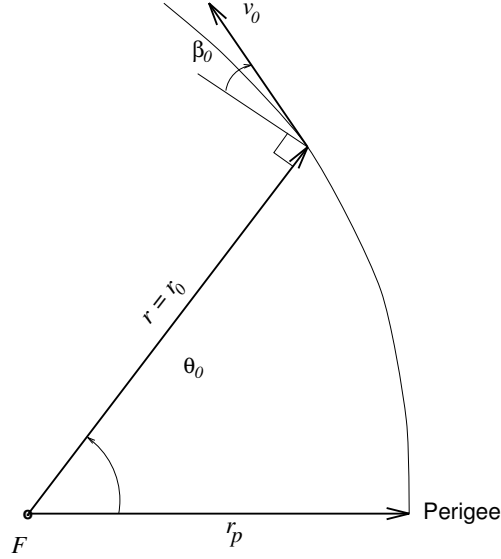


Figure 10.7.: Initial conditions at injection into orbit. Figure after Thomson [1986].

where r_p is the perigee distance, to calculate the derivative

$$\frac{dr}{d\theta} = \frac{(1+e)r_p}{(1+e\cos\theta_0)^2} e \sin\theta_0 = \frac{r_0^2}{(1+e)r_p} e \sin\theta_0 \quad (10.45)$$

which is inserted in Equation 10.43a. We notice that $\dot{\theta}_0 = h/r_0^2$ and obtain

$$v_0 \sin\beta_0 = \frac{h}{(1+e)r_p} e \sin\theta_0. \quad (10.46)$$

Solving Equation 10.44 for $e \cos\theta$ we find

$$e \cos\theta_0 = \frac{(1+e)r_p}{r_0} - 1 \quad (10.47)$$

and Equation 10.46 gives

$$e \sin\theta_0 = \frac{(1+e)r_p}{h} v_0 \sin\beta_0. \quad (10.48)$$

Adding the squares of these expressions and using that $(1+e)r_p = h^2/(GM)$ we obtain

$$e^2 = \left(1 - \frac{r_0 v_0^2}{GM}\right)^2 \cos^2\beta_0 + \sin^2\beta_0. \quad (10.49)$$

This equation allows us to calculate the eccentricity from the satellite position and velocity at burnout of the rocket.

The special case of a satellite launched with $\beta_0 = 0$ is instructive because of its simplicity of interpretation. From Equation 10.46 it is evident that $\theta_0 = 0$ in this case, which means that if the velocity is perpendicular to the vertical at rocket burnout, the launch point corresponds with perigee. Since a high perigee is important to avoid atmospheric friction, this type of launch is often used in practice. From Equation 10.49 we find that the eccentricity in this case is

$$e = \frac{r_0 v_0^2}{MG} - 1 \quad (10.50)$$

and the orbit will be circular if $r_0 v_0^2 = MG$.

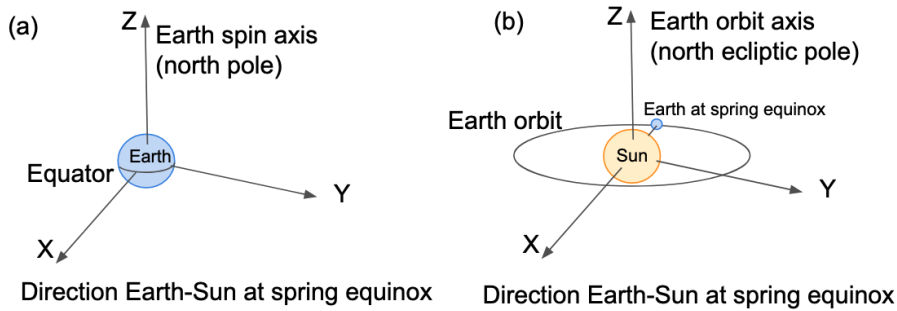


Figure 10.8.: Illustration of (a) geocentric equatorial inertial (GEQI or "celestial") and (b) heliocentric ecliptic inertial (HEI or just "ecliptic") coordinates.

10.9. Coordinate systems and orbital elements

In Section 10.6, we described the shape of an elliptic orbit by its semimajor axis a and eccentricity e . To describe how the ellipse is oriented in 3D space, we need also to specify the orientation of the plane of the ellipse with respect to some given coordinate axis as well as the direction of the major axis in this plane. Finally we need to describe where along the ellipse our spacecraft is at a particular time. The total number of parameters needed must be six, as the mechanical problem of orbital motion has six degrees of freedom (the initial values of the 3D position and velocity vectors). The usual choice is a set of six *orbital elements* in an order of refined detail, which to some extent also is a measure of their robustness against perturbations and measurement errors (see section 10.10 below). The first two, a and e , do not depend on any coordinate system, but the parameters describing directions must do so. We therefore need to define reference directions and a coordinate system.

We must specify the origin and two coordinate axes (the third follows by the cross product). For orbital problems, inertial systems are clearly preferable, to avoid complicating Newton's laws by the appearance of inertial forces. For the same reason of simplicity of the equations, the origin is usually placed at the centre of mass of the gravitating body as in Figure 10.5 (and Figure 10.6). For satellites orbiting the Earth, *celestial coordinates*, also known as Geocentric Equatorial Inertial (GEQI, see Figure 10.8(a)) coordinates, are commonly used, with the Z axis along the Earth's spin axis and X toward the "first point of Aries", the direction of the Sun at the vernal (spring) equinox². Celestial coordinates form an example of an *equatorial* system, with the equator defining the XY plane. As Earth spin axis is tilted by some 23° from the plane in which the Earth (and approximately the other planets) move around the Sun, interplanetary trajectories are better described in *ecliptic coordinates* also known as *Heliocentric Ecliptic Inertial (HEI)* coordinates. Compared to GEQI, the HEI X direction is unchanged but the Z axis points to the ecliptic north pole, i.e. along the angular momentum vector of the Earth in its orbit around the Sun, and of course the origin is shifted to the Sun (Figure 10.8(b)).

These and other inertial coordinate systems are useful for describing orbits and trajectories, while other problems require other coordinates. For describing details of the Earth's intrinsic magnetic field, a system rotating with the Earth will be convenient, and for

²None of these reference directions is exactly stationary, but the time scales of their variations (hundreds or thousands of years) are much longer than the typical orbit period of a spacecraft (hours or days) so celestial coordinates are still very useful. As an example, the vernal equinox point was in the constellation Aries in antiquity but has drifted into Pisces today, a couple of millennia later.

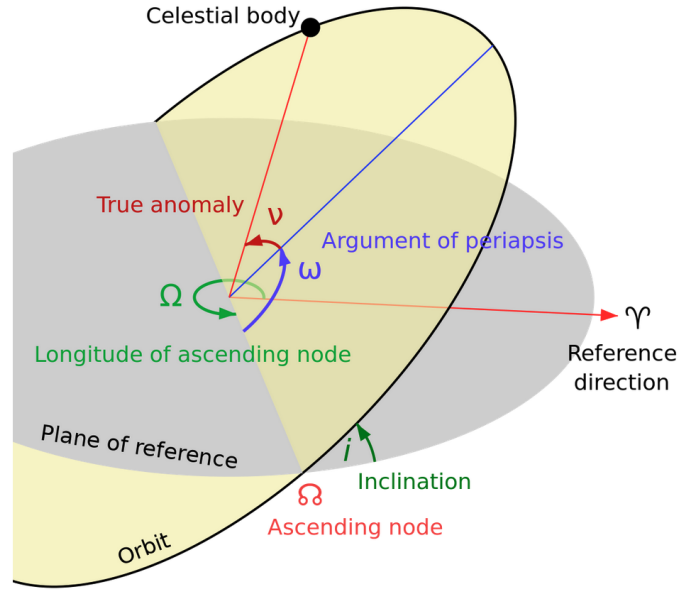


Figure 10.9.: Orbital elements in addition to a and e . The angles i , Ω , and ω define the orientation of the orbital ellipse while ν gives the position on the ellipse. The reference plane and direction are the XY plane and the X axis of the chosen coordinate system, usually celestial (GEQI) for orbits around the Earth and ecliptic (HEI) for interplanetary trajectories. Image credit: user Lasunncty at Wikimedia Commons, shared under CC BY-SA 3.0.

describing the interaction of the flowing solar wind with the environment of a planet a sun-pointing axis is more practical than an inertial or Earth-rotating one (c.f. section 6.2). We should note that whatever system we chose, we can let the mathematical representation of coordinates be Cartesian, spherical or cylindrical as we find convenient.

The remaining four orbital elements are all angles, illustrated in Figure 10.9. Two are needed to describe the orientation of a plane in 3D space. The *inclination* i is the angle between the angular momentum vector L of the orbit and the Z axis of the chosen coordinate system, $\mathbf{L} \cdot \hat{\mathbf{Z}} = L \cos i$. This can be seen as the angle between the orbital and the XY planes, but one may note that by the scalar product definition, i can reach values also above 90° . Such values correspond to *retrograde* orbits, in which the angular motion of the satellite is opposite to the one of the reference body (assuming this has been used to define the Z axis, as is the case for equatorial systems, and also for ecliptic systems if the reference body is taken to be the Earth in its orbit around the Sun). Inclinations up to 90° define *prograde* orbits.

A second angle, the *longitude* (or right ascension) *of the ascending node* Ω describes how the orbital plane is oriented with respect to the X axis. The ascending node is the point where the spacecraft passes through the equatorial (XY) plane moving from south to north, i.e. where $Z = 0$ and $dZ/dt > 0$. With these two angles, the orbital plane is fully specified. A third angle, the *argument of perigee* ω , describes how the orbital ellipse is oriented in this plane. Finally, the polar angle of the spacecraft in the orbital plane, known as the *true anomaly* ν and usually counted from the ascending node, defines its position on the ellipse. For a perfect Kepler orbit, only ν changes in time, evolving from 0 to 360° every orbit.

10.10. Non-Keplerian Effects

Sections 10.5 – 10.9 considered motion in a central force field. This is often a good approximation for most trajectories, like interplanetary trajectories with the Sun as gravitating central body or satellite orbits around the Earth, resulting in elliptic Kepler orbits or hyperbolic trajectories as discussed above. However, the central force condition is never perfect. If the non-central parts of the force are small, the orbits are close to Keplerian and can be described by allowing the orbital elements of Section 10.9 to slowly (compared to the orbit period) change in time. Such changes can be short-lived, averaging to zero over some number of periods, but some accumulate over time to cause a *secular change* to the orbit. We will here briefly discuss a few of these.

Non-spherical mass distribution of the central body results in a not perfectly radial gravitational force. For the Earth, the most well known non-spherical aspect is the "flattening at the poles", with an oblate spheroid better describing the shape of the planet than a perfect sphere, but all deviations from perfect spherical symmetry have an impact. The gravitational potential of the Earth can be expanded in spherical harmonics $Y_{lm}(\theta, \phi)$ as

$$\Phi(r, \theta, \phi) = -\frac{GM}{r} \left[1 + \sum_{l=1}^{\infty} \left(\frac{R_E}{r} \right)^l \sum_{m=-l}^l J_{lm} Y_{lm}(\theta, \phi) \right] \quad (10.51)$$

where (r, θ, ϕ) are the usual spherical coordinates, $G = 6.67 \cdot 10^{-11} \text{ m}^3 \text{ s}^{-1} \text{ kg}^{-1}$ is the gravitational constant, M is the total mass and $R_E = 6371.2 \text{ km}$ is the nominal radius of the Earth. The expansion coefficients J_{lm} define the relative strength of each term: if all $J_{lm} = 0$, we recover the central force case of a spherically symmetric mass distribution. Except for particular orbits, effects of variations with ϕ tend to average out with time as the satellite moves over regions with different ϕ on every orbit as the Earth rotates beneath it. One can then use the simplified expression

$$\Phi(r, \lambda) = -\frac{GM}{r} \left[1 + \sum_{l=2}^{\infty} \left(\frac{R_E}{r} \right)^l J_l P_l(\sin \lambda) \right] \quad (10.52)$$

where P_l denote Legendre polynomials and we have replaced the polar angle θ by the latitude $\lambda = 90^\circ - \theta$ ³. The dimensionless coefficients J_l are moments of the mass distribution,

$$J_l = \frac{1}{Ma^l} \int r^l \rho(r, \lambda) P_l(\sin \lambda) dV. \quad (10.53)$$

There is no $l = 1$ term as J_1 is zero when the origin is at the centre of mass. For Earth, $J_2 \approx -0.001$ and $J_2 \approx 2.5 \cdot 10^{-6}$ so the non-spherical contributions to the field are small. What effects will they have? When we do not consider longitudinal variations, the gravitational field is constant in time because of axial symmetry with respect to the Earth rotation axis. The specific energy (10.37) of the orbit must then be conserved, so, by (10.41), the first orbital element, the semimajor axis a , will stay constant. All other elements will see changes, with secular changes mostly turn up in two of them, Ω and ω (see section 10.9). One may (tediously) calculate the changes (in radians) per orbital

³The latitude λ better illustrates symmetry around the equatorial plane than does the polar angle θ and is essentially always used for giving positions on or near the Earth or around any other body in space. Spherical coordinates (r, θ, ϕ) can just as well be expressed as (r, ϕ, λ) (in that order, to have a right handed system).

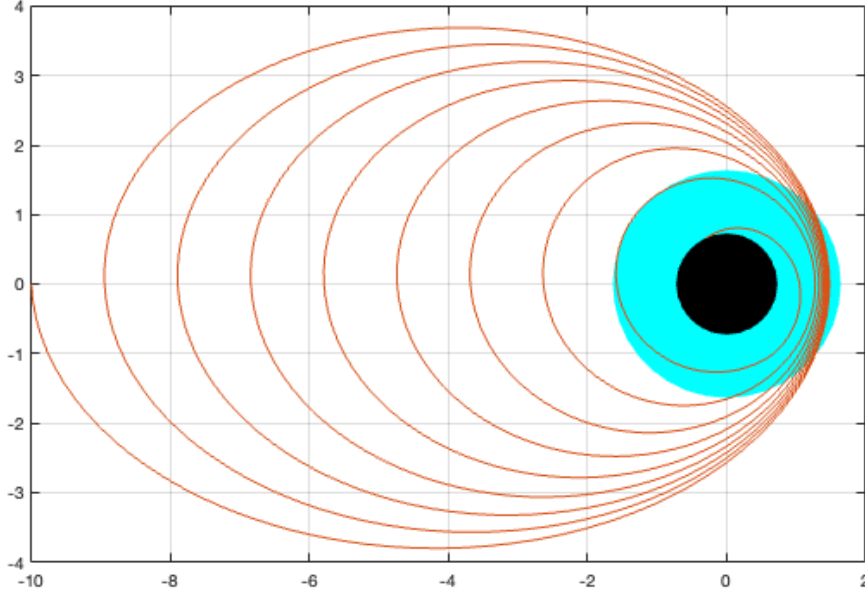


Figure 10.10.: Sketch of the impact of atmospheric drag on an elliptic orbit around a body (black) with an extended atmosphere (cyan). Initially the main effect is decay of the apogee with only slight change of perigee. As the orbit becomes more circular the spacecraft spends more of its time inside the atmosphere, increasing the rate of orbital decay.

period to be

$$\Delta\Omega = 3\pi \left(\frac{R_E}{a}\right)^2 \frac{J_2}{(1-e^2)^2} \cos i \quad (10.54)$$

$$\Delta\omega = -\frac{3\pi}{2} \left(\frac{R_E}{a}\right)^2 \frac{J_2}{(1-e^2)^2} (4 - 5\sin^2 i). \quad (10.55)$$

The precession of the orbital plane $\Delta\Omega$ becomes zero only for $i = 90^\circ$, i.e. a *polar orbit*, but the major axis will still turn in this plane ($\Delta\omega \neq 0$). For the opposite case of an *equatorial orbit* ($i = 0$), $\Delta\Omega$ and $\Delta\omega$ are both non-zero even though the motion now is in a plane with circular symmetry of the mass distribution⁴. The 3D mass distribution matters even in this case as Kepler's first and third laws only apply to a $1/r^2$ gravity force. When $J_2 \neq 0$ (like for a slightly flattened Earth) the presence of a $1/r^4$ term means the orbit cannot be a perfect ellipse. Interestingly, the precession of the major axis in the orbital plane, $\Delta\omega$, is zero for a particular inclination $i = \arcsin \sqrt{4/5} \approx 63.4^\circ$ (c.f. section 10.11 below). The precession equations illustrate a useful property of the set of orbital elements we discussed in section 10.9: weak non-Keplerian effects can be modelled by allowing these elements to change in time, and some elements do not change much at all.

Atmospheric drag can be important for orbits with low perigee. At any point in its orbit, a spacecraft of cross-sectional area A moving at speed v through an atmosphere of number density n will hit and push away nAv molecules per unit time and transfer an average momentum $\sim mv$ to each of them, where m is the mean molecular mass. The

⁴For exactly zero inclination, precession of the orbital plane has little meaning. Nevertheless, by taking the limit of Ω and ω as $i \rightarrow 0$ we find that the major axis will still turn.

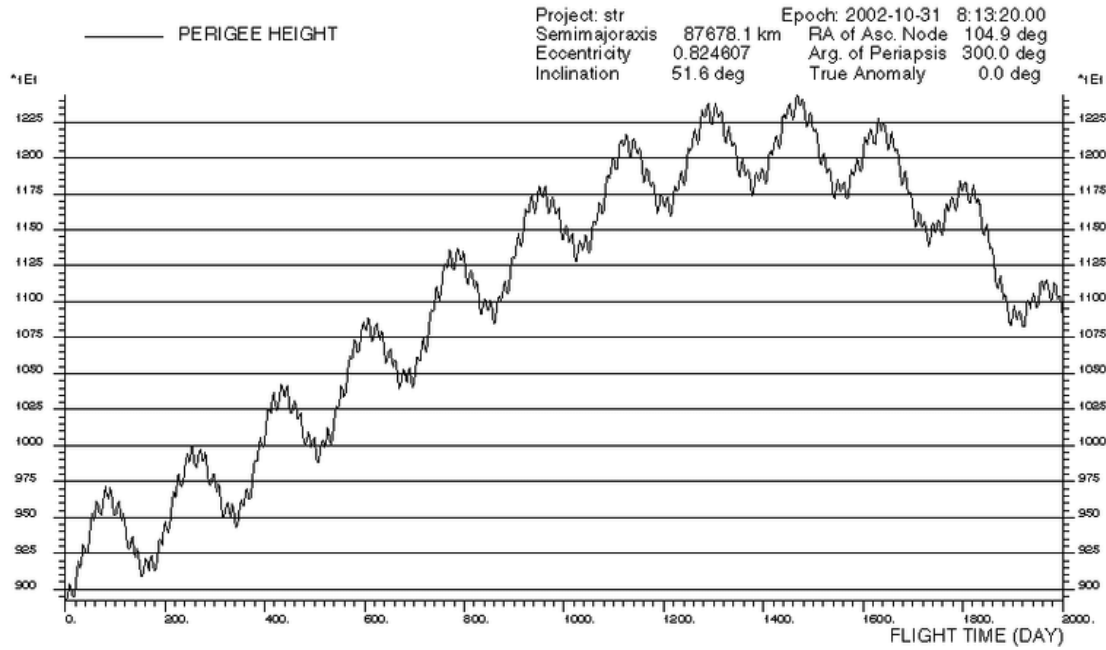


Figure 10.11.: Evolution of the perigee altitude of the spacecraft from the Integral mission during its first 2000 days in orbit. Figure from Eismont et al., A&A 2003, doi:10.1051/0004-6361:20031452.

satellite momentum loss per unit time, i.e. the drag force, therefore is $F_D = C_D m n A v^2$ where $C_D \sim 1$ is a numerical coefficient. As the atmospheric density changes rapidly with altitude, most of the momentum loss in an elliptic orbit occurs close to perigee. The main result of a weak drag therefore is the decline of the apogee altitude while the perigee altitude initially changes much less (Figure 10.10). The orbital energy decreases (becomes more negative) as a decreases, and the eccentricity e also goes down. The orbit becomes a little more circular with every orbit, so the drag force acts over a larger and larger part of the orbit. A circular orbit results, which then shrinks faster and faster as the satellite reaches denser and denser regions of the atmosphere. Rough order of magnitude lifetimes for circular Earth satellite orbits vary with altitude from thousands of years at 1,000 km to 10 years at 500 km, 1 year at 400 km, 1 month at 300 km and 1 day at 200 km.

Three-body or *many-body effects* often play a role. The Moon orbits the Earth at a distance of about 60 Earth radii. Satellite orbits with a sufficiently high apogee can therefore be subject to substantial forcing by the Moon. This force will vary over time in a complicated way as the spacecraft orbit elements change. As the force from the Moon is strongest when closest to the spacecraft, i.e. with the spacecraft at apogee when the moon is on the same side of the Earth, it is not surprising that one of the main effects is a change in perigee altitude. Figure 10.11 shows the variation of the perigee altitude of the ESA gamma ray observatory Integral during its first years in orbit. The fastest visible variations (small ripples) occur on a monthly timescale, due to the varying position of the Moon with respect to the orbital plane of Integral. More important are the stronger six-month cycle giving some 12 peaks in this diagram, and the even stronger and slower background variation, lifting the perigee altitude from 900 to 1200 km in the first four years and then bringing it down again. Another ESA project, the four-satellite Cluster mission, launched in 2000 into a $4 \times 19 R_E$ orbit, will come to an end in 2025

when the perigee of each spacecraft shrinks to atmospheric altitude and there no longer is any fuel left to raise it. In essence, the Moon finally kills the mission by throwing the spacecraft into the atmosphere.

10.11. Example of Orbits around Earth

There is of course an infinity of possible orbits, in the sense of different numerical values of each of the orbital elements, to chose from. However, some orbits tend to be more useful than others, and it is useful assign names to particularly useful classes of orbits. Before doing so, we should note that in everyday use, orbits are often refered to by altitude h rather than radial distance r , with

$$h = r - R_E \quad (10.56)$$

where the Earth radius is defined to be $R_E = 6371.2$ km. Instead of giving semimajor axis and eccentricity, orbits are often referred to by perigee and apogee: a "600 x 1,500 km orbit" thus is one with perigee and apogee at these altitudes. It is not always obvious if radial distance or altitude is intended. A common (but by no means universal) practice is to give altitudes in km and radial distances in R_E . Beware of confusion!

Using altitude for orbit classification, *Low Earth orbits* (LEO) are close to Earth with altitudes of less than 1000 km. For example, the International Space Station (ISS) is stationed in a LEO with about 400 km altitude. As the defined LEO altitude range is rather small compared to the Earth radius $R_E = 6371.2$ km, all satellites in LEO have periods of about one and a half hour (between 88 and 105 minutes) and move at speeds ~ 7.5 km/s. LEOs are cheaper to reach from Earth than orbits with higher altitude as one only needs to climb a small part of the potential well gravity creates around Earth. Particularly for ground imaging of various kinds, the low altitude also is an advantage. However, the high speed of objects in LEO makes tracking with ground stations more complicated. In addition, objects in LEO are affected by atmospheric drag which is the main cause of orbital decay (c.f. Section 10.10). Regular altitude boosts are therefore needed to compensate for the altitude loss for orbits in the lower LEO range.

Medium Earth orbits (MEO) stay at altitudes between 1,000 and 35,786 km (the GEO altitude, see below) and are suitable for e.g. navigation satellites such as the Galileo or GPS systems as the spacecraft is visible from a large part of the Earth's surface. *High Earth orbits* (HEO, above GEO) have less commercial use but are important for many scientific purposes. Sometimes, the abbreviation HEO is also used for Highly Elliptical Orbit. These elliptical orbits are for example used for scientific exploration of remote astronomical phenomena and in-situ measurements of the magnetosphere.

An orbit period of 23h 56min 4s, the Earth's rotation period in an inertial system, defines a *geosynchronous orbit* (GSO). A satellite in a GSO has the same *ground track* (Figure 10.12) on every revolution, i.e. it passes over the same positions on the ground every time. By (Equation 10.36), the semimajor axis of a GSO has to be $a = 42,164$ km = $6.62 R_E$. A particularly important GSO is the *geostationary orbit* (GEO), which is circular ($e = 0$) and equatorial ($i = 0$). For this orbit, the ground track becomes a point: as seen from ground, the satellite hangs over the same spot all the time. This orbit is obviously very convenient for communication purposes, allowing e.g. home television satellite dishes to be mounted in a fixed direction.

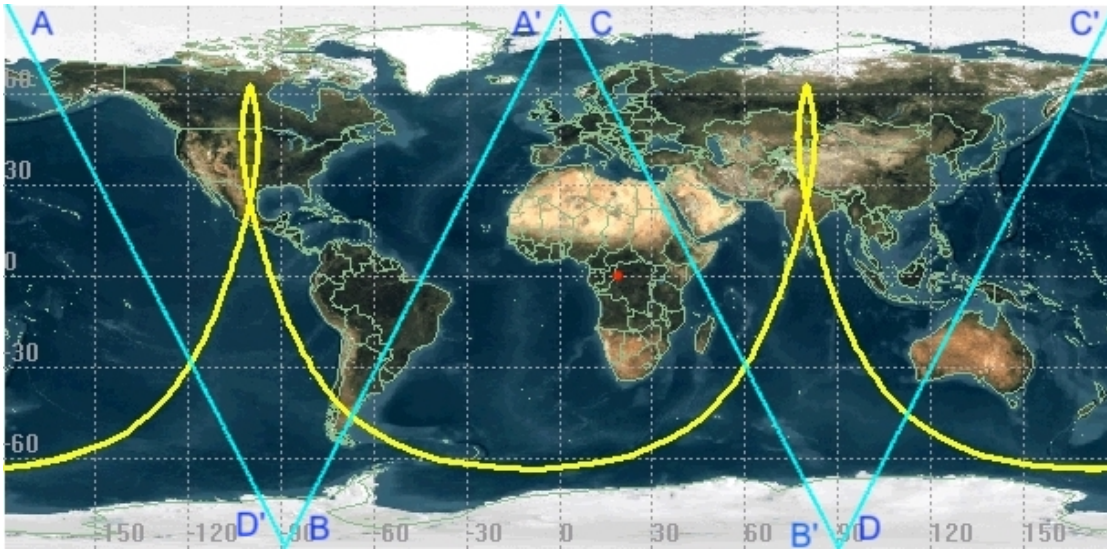


Figure 10.12.: Satellite ground track examples. The Earth rotates eastward, so satellites move westward on the map high above these tracks. Red: The track of any GEO is a single point somewhere on the equator. Blue: A polar satellite in circular orbit with 11 h 58 min period. Letters A-D denote the four polar passes during one Earth rotation (23 h 56 min). When crossing the pole the longitude changes by 180° , so on the map the track jumps from A' to A, from B' to B etc. The first orbit in this period thus is A'-B-B'-C, the second C'-D-D'-A. Yellow: The Molniya orbit.

If classified by inclination, the GEO is an example of an *equatorial orbit*. In contrast, *polar orbits* ($i \approx 90^\circ$) pass over (or close to) the north and south pole. If the period is chosen so as to not be a rational fraction of the Earth rotation period, a polar satellite will eventually pass over every single finite area on ground, so military observation satellites often are in polar orbits. For launches into polar orbits one cannot make use of the Earth's rotation (section 10.3) and therefore a higher Delta-v is required in comparison to equatorial orbits. One special type of polar orbit is the *sun-synchronous orbit* (SSO) with orbital elements tuned to make Equation 10.55 result in a full 360° turn of the orbital plane in one year, i.e. just below $1^\circ/\text{day}$. The angle of the orbital plane to the Sun then stays constant. An SSO example is the European MetOp meteorology satellites, observing Earth in similar illumination conditions every day.

By Equation 10.55, the inclination 63.4° has the particular property that there is no precession of the major axis of the orbital ellipse. This has been ingeniously used in the *Molniya orbit*. For this orbit, a is set to $4.18 R_E$ to get a period half that of a GSO, so the satellite will pass the same points on the Earth every second orbit it makes. With $e \approx 0.737$ and perigee in the southern hemisphere, this gives the yellow track in Figure 10.12. In accordance to Kepler's second law, the satellite spends most of its time near apogee, over Canada and Siberia in this case, with slow apparent motion from the ground. Communication satellites in Molniya orbits thus can provide almost the same service to high latitudes as GEO satellites can do over the equator, though 2-3 satellites are needed in each orbit to keep one over Siberia or Canada all the time.

It should be emphasized that the secular variations of the orbital plane and the major axis described by Equation 10.55 and Equation 10.54 are measured with respect to

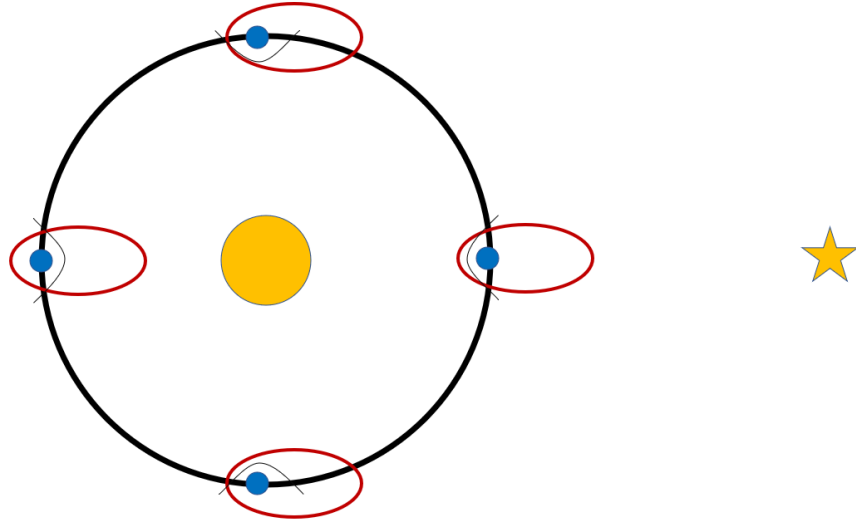


Figure 10.13.: Illustration of a spacecraft in orbit (red line) around Earth (blue circle) as Earth orbits the Sun (yellow circle). The black line close to Earth represents the bow shock which is part of the Earth's plasma environment. The orientation of the bow shock and the plasma environment is dependent on the Sun's relative position towards Earth. Therefore the spacecraft will probe different regions of the Earth's plasma environment throughout the year.

an inertial system. For spacecraft operations, the direction to the Sun can be more important, particularly as the main source of onboard electric power usually is solar panels and eclipses (passes through the shadow behind the Earth) therefore can be problematic. Spacecraft designed to probe the plasma environment around Earth are often in highly elliptical orbits to cover different regions is space out to several tens of Earth radii. As the precessions described by Equation 10.55 and Equation 10.54 decay as $1/a^2$, the orbital plane and major axis of such an orbit changes relatively slowly. The orbit therefore remains approximately the same over timescales of a year, so the orbit orientation with respect to the Sun changes as is illustrated in Figure 10.13 where the spacecraft is always at its perigee when facing a fixed star. However, the plasma environment around Earth is dependent on the position of the Sun. This is illustrated by the black curve in Figure 10.13 which represents the bow shock (c.f. section 6.4). This bow shock is always on the Earth's dayside which is facing the sun. The bow shock is changing its orientation relative to the fixed star over on Earth year. As the seasons evolve the spacecraft will therefore probe different plasma regions throughout the year. These seasonal variations need to be taken account when planning and designing a mission in an elliptical orbit around any planetary body.

10.12. Lagrange points

The so called Lagrange points are equilibrium positions for a small light body (a test particle) in the gravitational field of two massive bodies moving around their common centre of mass in Kepler orbits. For simplicity, we consider a case where the two massive bodies stay at constant distance from each other and where one is much more massive than the other, as is approximately true for e.g. the Earth-Moon or Earth-Sun systems.

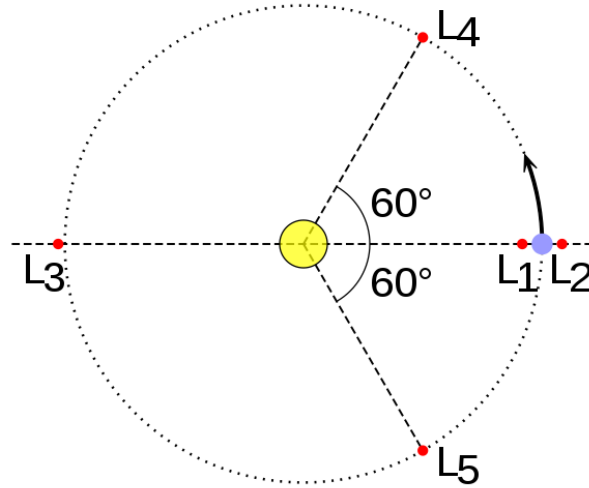


Figure 10.14.: Lagrange points in a system of two massive bodies at constant distance where one (the one at the centre) is much more massive than the other, like the Sun–Earth or Earth–Moon systems. Figure from Wikimedia commons, user EnEdC (public domain).

It is then possible to assume the lighter of the two bodies moves in a circle around the heavier one, and to introduce a reference frame in which both gravitating bodies are at rest. This frame obviously is rotating, so inertial forces will appear. We define the Lagrange points as positions where the centrifugal forces and the gravitational forces balance each other out, providing net zero acceleration in this rotating frame.

In total there are five Lagrange points (L1–L5) in any system of two massive bodies (Figure 10.14). These positions are commonly used for spacecraft as their orbits appear stationary in the two body system. The L1 of the Sun–Earth system is particularly valuable for monitoring the solar wind which will hit the Earth about an hour later, as done by the ACE and DSCOVER spacecraft, and for constant monitoring of the Sun as performed by for example SoHO. The L2 of the Sun–Earth system is useful for astronomical telescopes, for example the James Webb Space Telescope. L5 is exposed to the solar wind from the part of the solar surface which the rotation of the Sun will turn toward the Earth 4–5 days later. Solar wind measurements at L5 would therefore be very useful for space weather prediction purposes, as will be explored by the ESA mission Vigil.

In the following we will derive the position of the L1 point in the Sun–Earth system. The L2 and L3 positions can be derived by a similar procedure. We will then also discuss the perhaps less intuitive positions of L4 and L5 and the stability properties of these equilibrium positions. Possible influences from other solar system bodies are neglected and the Earth’s orbit is assumed to be circular. .

An object at any of the Lagrange points has a position fixed relative to Earth. The angular velocity ω of the rotating frame in which the Sun and Earth appear at fixed positions is therefore the same as the Earth’s angular velocity in its orbit, which we can find by equating the gravitational and centripetal forces in a non-rotating frame:

$$\frac{GMm}{d^2} = \frac{mv^2}{d} = m\omega^2 d, \quad (10.57)$$

with M the mass of the Sun, m the mass of Earth, and d the distance between the Sun and the Earth. In the rotating frame, the total force on an object will include inertial as well as gravitational forces. For an object at rest the Coriolis force is zero so the only inertial force is the centrifugal force. A satellite with mass m_s at distance r from the Sun will, in the rotating frame, experience a centrifugal force $m\omega^2 r$, directed away from the Sun. For a satellite a distance d_s from the Earth on the Sun-Earth line, the total force toward the Sun can then be written as

$$F = \frac{GMm_s}{(d-d_s)^2} - \frac{Gmm_s}{d_s^2} - m_s\omega^2(d-d_s). \quad (10.58)$$

We look for an equilibrium point (in the rotating system), where this force vanishes. From Equation 10.57 and Equation 10.58 we then get two expressions for the angular velocity, which must be equal:

$$\omega^2 = \frac{MG}{d^3} = \frac{MG}{(d-d_s)^3} - \frac{mG}{(d-d_s)d_s^2}. \quad (10.59)$$

Therefore

$$0 = \frac{MG}{d^3} - \frac{MG}{(d-d_s)^3} + \frac{mG}{(d-d_s)d_s^2}, \quad (10.60)$$

which can be rewritten as

$$0 = \frac{Md_s^2(d-d_s)^3 - Md^3d_s^2 + md^3(d-d_s)^2}{d^3d_s^2(d-d_s)^3}. \quad (10.61)$$

Only focusing on the nominator and setting $x = d_s/d$ we obtain

$$0 = M(x^2 - 3x^3 - 3x^4 + x^5) - Mx^2 + m(1 - 2x + x^2), \quad (10.62)$$

which results in

$$0 = Mx^5 - 3Mx^4 - 3Mx^3 + mx^2 - 2mx + m. \quad (10.63)$$

The above equation can be solved with numerical methods. Alternatively, the approximations $x \ll 1$ and $M \gg m$ can be used to simplify the equation to

$$-3Mx^3 + m = 0. \quad (10.64)$$

Thus, the distance of the L1 point from Earth toward the Sun is

$$x = \left(\frac{m}{3M}\right)^{1/3} \approx 0.01. \quad (10.65)$$

The positions of L2 and L3 can also be found by a similar procedure with different approximations for the roots of Equation 10.63. But how can the total force vanish on a test particle in L4 and L5? To see this, we leave the assumption of zero Earth mass and consider that in reality, the Sun and the Earth both orbit their common centre of mass. The centrifugal force therefore is directed radially away from this centre of mass, not from the Sun. For the total force on a test particle to be zero one must therefore be in a position such that the combined force from the Sun and the Earth points to the centre of mass. The forces are proportional to M and m , but so is also the position of the centre of mass. This means that the three forces cancel when the Sun, the Earth and the test particle forms an equilateral triangle⁵.

⁵This is most easily seen for two massive bodies have equal mass: the centrifugal force, directed away from their common centre of mass rather than from any single of them, would at L4 be directed directly upward in Figure 10.14.

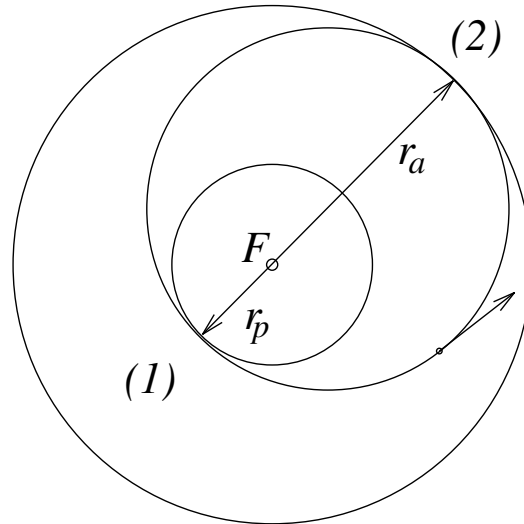


Figure 10.15.: Hohmann transfer orbit. Figure after Thomson [1986].

What about the stability of the Lagrange points? It turns out that L4 and L5 are stable, so if an object at e.g. L4 receives a small kick, perhaps by a dust impact, it will not leave L4 but stay in orbit around it. In contrast, L1-L3 are all unstable to perturbations in the transverse direction, so if perturbed, an object at e.g. L1 will start to drift away from it. For this reason, there is a large population of asteroids (known as Trojans) around the L4 and L5 Lagrange points of the Sun-Jupiter system, while the corresponding L1-L3 points are empty. Conversely, spacecraft positioned at L1 and L2 in the Sun-Earth system need active position control at regular intervals. However, there exist quasi-stable orbits around these points which allow the spacecraft to be left without any orbit adjustments for quite long times.

10.13. Transfer between circular orbits

A satellite is taken into orbit by a launch rocket. Many satellites then have their own smaller rockets, which allows the orbit to be modified and corrected. One simple maneuver of this kind is the transfer of a satellite from one Kepler orbit to another. For simplicity, we will consider the (often relevant) case of transfer between two circular orbits. This transfer can be effected by an elliptic orbit that is tangent to the circular orbits at perigee and apogee, known as the Hohmann transfer orbit. This means the thrust applied at the two manoeuvres will be parallel (or antiparallel) to the velocity. In this way, all the fuel goes to doing work, i.e. to change the energy of the orbit. There are many faster trajectories from e.g. Earth to Mars than a Hohmann transfer, but as they include a thrust component perpendicular to the velocity, part of the fuel is used without any change of orbital energy. Hohmann transfer therefore is the most economic way to go between two orbits.

Assuming transfer to take place from (1) to (2) in Figure 10.15, we can obtain the ratio r_a/r_p from Equation 10.43 by letting $\theta = 0$ (perigee) and $\theta = \pi$ (apogee)

$$\frac{r_a}{r_p} = \frac{1+e}{1-e}. \quad (10.66)$$

The eccentricity of the elliptic transfer orbit is thus

$$e = \frac{r_a - r_p}{r_a + r_p} \quad (10.67)$$

Comparing with Equation 10.50, we find that in the elliptic orbit after the rocket has been fired at (1), when $v_0 = v_p$,

$$\frac{r_p v_p^2}{MG} = 1 + \frac{r_a - r_p}{r_a + r_p} = \frac{2r_a}{r_a + r_p} \quad (10.68)$$

The velocity at this point is then

$$v_p = \sqrt{MG/r_p} \cdot \sqrt{\frac{2r_a}{r_a + r_p}} \quad (10.69)$$

In the initial circular orbit we have $e = 0$, and from Equation 10.50 we find that the velocity is $v_1 = \sqrt{MG/r_p}$. The increase in velocity effected by the rocket fired at (1) should thus be

$$\Delta v = v_p - v_1 = \sqrt{MG/r_p} \left[\sqrt{\frac{2r_a}{r_a + r_p}} - 1 \right] \quad (10.70)$$

Notice that we assume this increase to be instantaneous, and neglect the motion of the satellite during the acceleration.

The velocity is perpendicular to the radius vector at perigee and apogee, and from conservation of angular momentum we find $r_a v_a = r_p v_p$. At apogee in the elliptic orbit we then have, from Equation 10.68,

$$\frac{r_a v_a^2}{MG} = \frac{r_p}{r_a} \frac{r_p v_p^2}{MG} = \frac{2r_p}{r_a + r_p} \quad (10.71)$$

In the final circular orbit the velocity should be $v_2 = \sqrt{MG/r_a}$, and the velocity increase from the rocket at (2) required to go into the circular orbit is then

$$\Delta v_a = v_2 - v_a = \sqrt{MG/r_a} \left[1 - \sqrt{\frac{2r_p}{r_a + r_p}} \right] \quad (10.72)$$

10.14. Gravity Assist

Gravity assists (also called swing-by maneuvers) can be used to accelerate or change the direction of a spacecraft. Gravity assists make use of other astronomical objects, mostly planets, without using extra propellant. These types of maneuvers are important for missions that require large changes in velocity since rockets can not economically provide enough fuel to achieve these high changes in velocity.

The change of velocity is the result of the planet's (or other astronomical object's) motion around the Sun. Newtonian gravity is a conservative force, so the spacecraft speed relative to the planet must be the same before and after the gravity assist. But, during the gravity assist the planet moves on it's orbit around the sun. The result is a

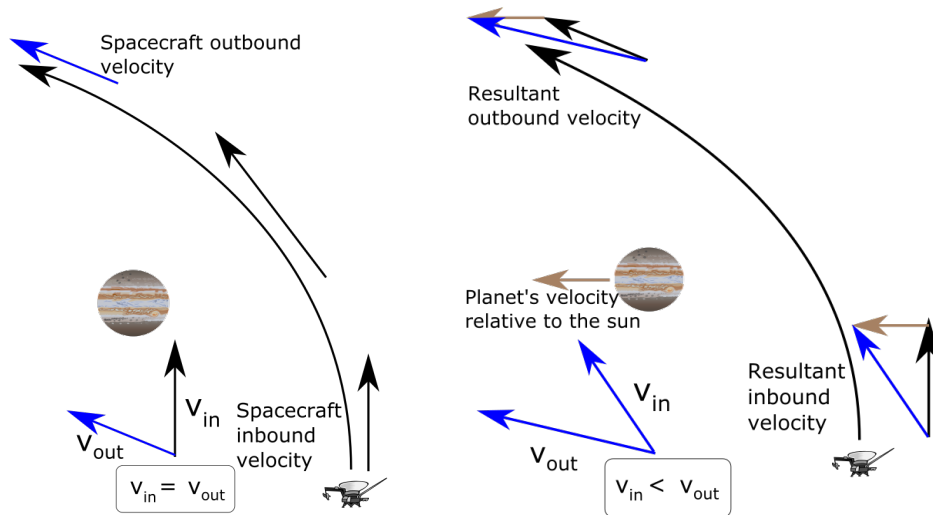


Figure 10.16.: A gravity assist as seen in the frame of the planet (left) and of the solar system (right). Image credit: user Zeimusu on Wikimedia commons, licence CC BY-SA 3.0.

change in velocity as illustrated in Figure 10.16. During this interaction the momentum of the planet-spacecraft system is conserved. The velocity loss of the planet is however negligible. Mathematically, this interaction resembles an elastic collision.

Gravity assists are nowadays routinely used to achieve economic trajectories in the solar system. Jupiter Icy Moons Explorer (Jupiter, launched 2023) is a good example, with one Venus and three Earth flybys on its 8-year route to Jupiter. The Ulysses spacecraft, as yet the only one to have passed over the poles of the Sun, was launched in 1990 into the the ecliptic plane along the direction of motion of the Earth, to maximize the orbit energy, on an orbit taking it out to Jupiter. Passing over the south pole of the giant planet, it used its gravity to achieve a 79° inclination (see page 128) over the ecliptic plane, at no fuel cost.

Example

Before launch a model rocket has a total mass of $m_0 = 50$ g. While the motor is burning, gases are ejected with velocity $v_e = 1000$ m/s. The rocket is launched vertically from the ground, and reaches the peak of its orbit at $t_p = 20$ s. How much fuel has been consumed by the rocket? (Aerodynamic drag and inhomogeneity of the gravitational field can be neglected.)

Solution

The motor burns during a time t_b . For $t \leq t_b$ the equation of motion is

$$\dot{\mathbf{v}}(t) = \frac{\dot{m}}{m(t)} \mathbf{v}_e + \mathbf{g} \quad (10.73)$$

where $m(t)$ is the instantaneous mass of the rocket. This can be integrated to give the velocity after burnout

$$\begin{aligned} \mathbf{v}(t_b) &= \mathbf{v}_e \int_0^{t_b} \frac{\dot{m}}{m(t)} dt + \mathbf{g} t_b \\ &= \mathbf{v}_e \int_0^{t_b} d(\ln m(t)) + \mathbf{g} t_b = \mathbf{v}_e \ln \left(\frac{m(t_b)}{m_0} \right) + \mathbf{g} t_b \end{aligned} \quad (10.74)$$

During the rest of the flight ($t > t_b$) we have $\dot{\mathbf{v}}(t) = \mathbf{g}$, which gives $\mathbf{v}(t) = \mathbf{v}(t_b) + \mathbf{g}(t - t_b)$. The velocity is zero at t_p when the peak is reached

$$\mathbf{v}(t_p) = \mathbf{v}(t_b) + (t_p - t_b)\mathbf{g} = 0 \quad (10.75)$$

or

$$\mathbf{v}_e \ln \left(\frac{m(t_b)}{m_0} \right) = -\mathbf{g} t_p \quad (10.76)$$

Denoting the mass of the fuel by $m_f = m_0 - m(t_b)$, and noticing that $\mathbf{v}_e$ is parallel to $\mathbf{g}$ since the launch is vertical, we find

$$\ln \frac{m_0 - m_f}{m_0} = -\frac{g t_p}{v_e} \quad (10.77)$$

or

$$m_f = m_0 \left(1 - e^{-g t_p / v_e} \right) \quad (10.78)$$

With the given numerical values $m_0 = 50$ g, $v_e = 1000$ m/s, $t_p = 20$ s, and $g = 9.81$ m/s⁻² we obtain

$$m_f = \left(1 - \exp \left[-\frac{9.81 \cdot 20}{1000} \right] \right) 50 \text{ g} = 8.9 \text{ g} \quad (10.79)$$

Answer: 8.9 g of fuel has been consumed.

Alternatively, the solution may be based on the rocket equation given in Physics Handbook on p 125. As the fuel burns, the mass changes from m_0 to $m(t_b)$. The corresponding velocity increment is

$$\Delta \mathbf{v}_f = \mathbf{v} - \mathbf{v}_0 = -\mathbf{v}_e \ln \left(\frac{m_0}{m(t_b)} \right) \quad (10.80)$$

Since gravitation acts all the time the rocket is in the air, the velocity increment from gravitation is $\Delta \mathbf{v}_g = \mathbf{g}t_p$. The final velocity at time t_p is zero, which implies $\Delta \mathbf{v}_f + \Delta \mathbf{v}_g = 0$, or

$$-\mathbf{v}_e \ln \left(\frac{m_0}{m(t_b)} \right) = \mathbf{v}_e \ln \left(\frac{m(t_b)}{m_0} \right) = -\mathbf{g}t_p \quad (10.81)$$

From here we continue as above.

Notice that the velocity and height reached by the rocket cannot be calculated from the given information. If the fuel burns slowly the rocket may hover for 20 s just over the ground, but if it burns instantaneously the rocket may reach a speed of 196.2 m/s and an altitude of 1962 m.

11. Space Measurements

11.1. Introduction

Although *in situ* measurements by instruments on spacecraft are essential in space physics, many important observations have been made, and are still made, from the ground. The earliest relevant observations were the auroral displays seen against the dark sky. One of the early truly scientific observations of importance for our present understanding of space physics was made by Olof Peter Hiorter. In 1741 he discovered that a magnetic needle was deflected by several degrees when an auroral arc moved over his head. This observation is today explained as the magnetic disturbance caused by the auroral electrojet. Optical and magnetic techniques were later supplemented by studies of radio wave propagation as important tools for mainly ionospheric observations. When it around 1960 became possible to launch scientific payloads into space, a rapid development of new types of instruments begun. Often the space environment demands and allows instrument designs that are very different from those used in the laboratory. Furthermore, the instruments need to be designed according to the plasma conditions in the respective environment and the main scientific objectives to be answered during a mission. The scientific instruments have to be designed in a way such that they have a long life-time without maintenance. Mass and power consumption are heavily restricted in space and by the spacecraft design.

Even though the working principles of such instruments are often rather simple, the resulting raw data needs careful processing and analysis. Even the final data products must be inspected carefully to separate real data from artifacts.

Many of the important parameters in space physics can be measured in several ways. Often, measurements from rockets and satellites give a very detailed snapshot of an event. Ground based methods continuously covering large areas provide an invaluable background and are often cheaper to operate and maintain. In the following we will introduce some commonly used space-borne instruments, chapter 12 will then introduce ground-based instrumentation.

11.2. Magnetic fields

In space physics, there are two types of magnetometers that are commonly in use. One is the search coil magnetometers that measure variations in the magnetic field. The other type are fluxgate magnetometers that provide the full magnetic field vector. On some spacecraft, both types of magnetometers can be found. One fluxgate magnetometer to measure the background magnetic field and a search coil magnetometer for variations in the magnetic field which provides data with a higher time resolution.

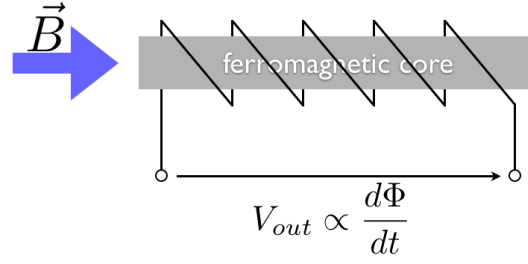


Figure 11.1.: Schematic of a search coil magnetometer. Figure from Ettelt [2012].

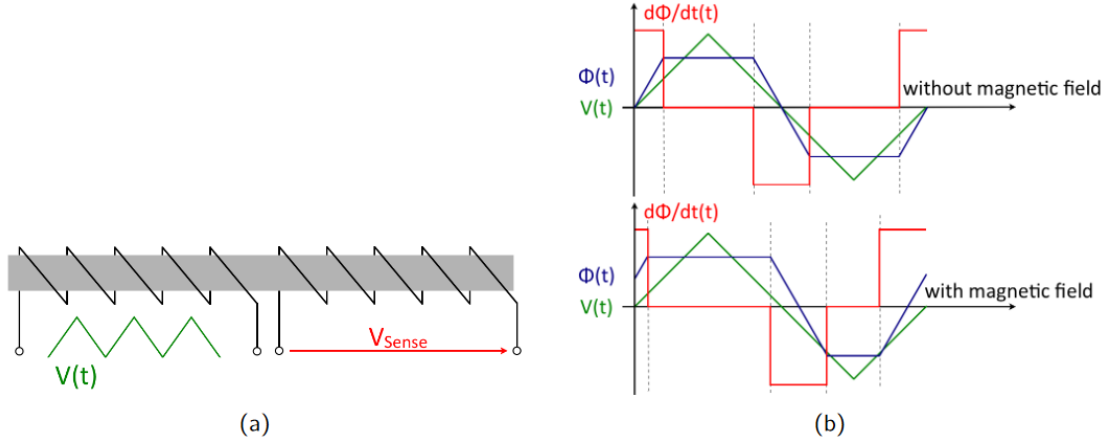


Figure 11.2.: Schematic of a flux gate magnetometer. a) A primary coil drives a ferromagnetic core into saturation. A secondary coil senses the changes in saturation. b) Induced signal with and without magnetic field. Figure from Ettelt [2012].

11.2.1. Search Coil magnetometers

Measuring variations in the magnetic field is much easier than measuring a constant magnetic field. Variations in the magnetic flux Φ induce a voltage u in a loop of wire according to Faraday's law,

$$u = -N \frac{d\Phi}{dt}, \quad (11.1)$$

with N the number of turns of a coil of wire. A search coil magnetometer simply measures the induced voltage which can then be used to derive the magnetic field in the direction that is parallel to the normal of the surface area of the coil. To get a three dimensional vector, three (perpendicular) directions need to be measured. However, a search coil magnetometer cannot measure a constant magnetic field or a slowly varying magnetic field. Figure 11.1 illustrates the working principle of a search coil magnetometer.

11.2.2. Flux gate magnetometer

A fluxgate magnetometer can measure the full three dimensional vector of the background magnetic field. The principles behind a fluxgate magnetometer are illustrated in Figure 11.2. A sinusoidal high-frequency current is driven through the primary winding on a ferromagnetic core. The current is strong enough to saturate the magnetic field in the core. A secondary coil senses the change of flux density in the core. If there is no external field applied, the

induced voltage in the secondary coil are equally spaced in time. With an external field, the magnetic flux in the core saturates earlier if the external field is aligned with the core magnetic field. If the external field and the magnetic field in the core are anti-parallel, the magnetic field in the core takes longer until it saturates. Therefore, the induced voltage in the secondary coil will not be equally spaced in time. The time differences can be used to calculate the magnetic field vector. As for the search coil magnetometer, three set-ups are needed to measure a three dimensional magnetic field.

A fluxgate magnetometer must be calibrated against a standard instrument, such as a proton-precession magnetometer. Since the ambient field is canceled by the offset winding, a fluxgate magnetometer can after calibration measure fields from well below 1 nT to nearly 1 mT, and its simple and robust design makes it easy to use on space vehicles. Usually, the signal from a fluxgate magnetometer is recorded a few hundred times per second, allowing waves with frequency less than about 100 Hz to be observed.

In order to get useful measurements from the magnetometers a spacecraft needs to be magnetically clean. A magnetically clean spacecraft should only generate magnetic field that will not effect the measurements of the magnetic field of the surrounding plasma. This can be partly achieved by increasing the distance of the magnetometers from the spacecraft body bu mounting the it on a deployable boom. Disturbance from artificially generated magnetic fields should still be small so the sizing of the boom remains small.

11.2.3. Ground based magnetometers

Ground based magnetometers are stationary magnetometers that measure for example the absolute magnetic field or short-time variations in the Earth's magnetic field due to current in the Ionosphere. For short-time variations the previously discussed fluxgate magnetometers can be used. For precis measurements a liquid magnetometer can be for example used.

Liquid magnetometers measure the magnetic field strength by determining the precession frequency of protons around the magnetic field lines. The magnetic moments of protons in a liquid, such as water, are first aligned by applying a strong magnetic field. When the applied field is turned off, the spin axis of the protons will all precess in phase around the geomagnetic field for a while, producing a macroscopic magnetic effect. The precession frequency f_p is related to the magnetic field B by

$$f_p = \frac{2\pi B}{\gamma_p}, \quad (11.2)$$

with γ_p the gyromagnetic ratio of protons. The precession frequency allows the magnetic field strength to be calculated in terms of fundamental constants, but it is rather complicated to find the direction of the field vector.

To get the direction of the magnetic field, D/I fluxgate theodolite are used. A theodolite is typically used to make precise measurements of angles in the horizontal and vertical plane. A D/I fluxgate theodolite has a fluxgate magnetometer mounted on top of it. The declination and inclination of the magnetic field can be found through null readings in the magnetic field measured by the fluxgate magnetometer. This type of instrument requires measurements by hand, therefore are these measurements taken less often.

11.3. Electric fields

From spacecraft, electric fields are usually measured by one or several pairs of double probes. These probes are spheres or cylinders with a diameter of about 10 cm. Each of the probes is biased to the *floating potential*, when the current to the surrounding plasma is zero. The potential difference between the probes is then measured, and the electric field component along the line between the probes is obtained by dividing by the probe distance. To obtain sufficient sensitivity, the probes must be separated by tens of meters. This is achieved by mounting the probes on wire booms that are suspended from the spinning satellite by the centrifugal force. The electric field component along the satellite spin axis is often measured with lower resolution, since shorter stiff booms must be used in this direction.

The double probe measurements give the electric field in the spacecraft frame of reference. To obtain the field in an Earth-fixed frame of reference we must subtract $\mathbf{v}_S \times \mathbf{B}$, and the accuracy of the measurement depends on an exact knowledge of the satellite velocity $\mathbf{v}_S$ and the magnetic field $\mathbf{B}$.

The double probe technique can be used to measure electric fields at all frequencies from DC to several MHz. Often, the spectrum of the electric field fluctuations has a well defined peak at the plasma frequency $\omega_{pe} = \sqrt{ne^2/\epsilon_0 m_e}$, which allows the electron density to be determined. The probes can also be operated as Langmuir probes. A Langmuir probe with a positive bias voltage relative to the plasma attracts electrons, while a negative probe attracts ions. The resulting current is a function of the bias, plasma density and temperature. By measuring the current at different bias voltages, the density and temperature of the plasma can be deduced. Density fluctuations can also be measured by keeping the probe at a fixed positive bias and measuring the current fluctuations. Hence, a carefully designed probe instrument provides not only the electric field but also other plasma parameters.

11.4. Particle distributions

Information about the particle distribution function in space plasmas can be obtained from particle spectrometers. These spectrometers consist of a energy filter, possibly a mass separator, a detector that registers the incoming particles, and a counter.

The energy filter exists of two conducting and charged spherical shells which separates particles of different energies. The energy filter only allows particles within a certain energy range to pass through. A sketch of such a energy filter is shown in Figure 11.3. In order for a particle to enter the instrument the resulting force needs to be

$$F_c = \frac{mv^2}{R} \quad (11.3)$$

with R the radius of the shell. An electric field E is created by a potential difference V between the spherical shells and exerts a force on the incoming particle according to

$$F = qE = q \frac{V}{dR}. \quad (11.4)$$

Particles with a certain kinetic energy W can pass the filter given an electric field and a radius

$$W = \frac{mv^2}{2} = \frac{mv^2}{R} \frac{R}{2} = qE \frac{R}{2}. \quad (11.5)$$

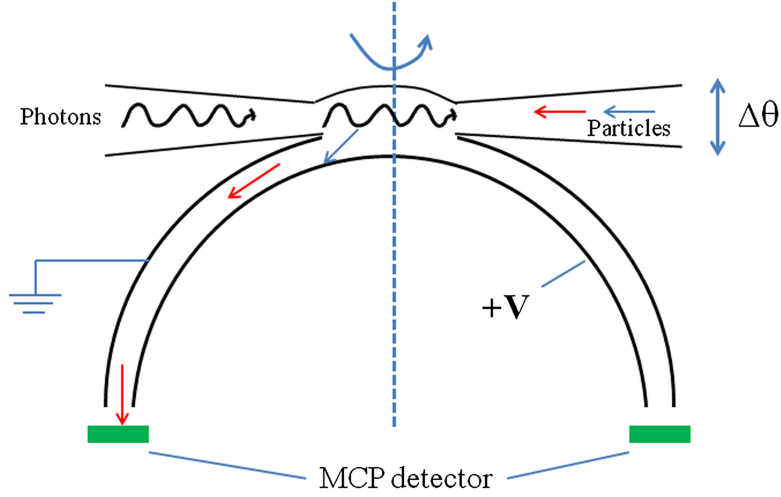


Figure 11.3.: Schematics of an electrostatic analyzer. Figure from Morel et al. [2017].

Therefore

$$\frac{W}{q} = E \frac{R}{2} \quad (11.6)$$

is constant. This filter is actually a W/q -filter and not a pure energy filter. The applied charges to the shells is then swept across the desired energy range to obtain the number of particles for each energy. Spectrometers that are mounted on three-axis stabilized spacecraft have often a limited field of view, not covering all directions. For spin stabilized spacecraft the rotation of the spacecraft can be used to obtain the full field of view of the velocity distribution function. An example of such a spectrometer is shown in Figure 11.4.

For electrons an energy filter is enough to determine the velocity distribution (assuming that no negatively charged ions are in the plasma which is usually a adequate assumption in space). For ions an extra filter for mass separation is often needed.

For mass separation a time-of-flight analyzer can be used. An example of a mass spectrometer is shown in Figure 11.5, For a time-of-flight analyzer the particles are injected in a field-free tube with length L where the particles travel with constant velocity v . With a measured time t that it takes for the particle to travel across the tube one can determine

$$\frac{W}{q} = \frac{mv^2}{2q} = \frac{mL^2}{2qt^2}. \quad (11.7)$$

The mass per charge m/q can be simply obtained by rewriting the previous equation to

$$\frac{m}{q} = 2 \frac{Wt^2}{qL^2}. \quad (11.8)$$

There are also instruments making use of deflection of charged particles in a magnetic field. However, this method requires the use of strong magnetic fields compared to the magnetic field in the plasma. It is desired to have spacecraft magnetically clean, which means that magnetic field should be local so they do not infer with other measurement devices.

The detector and counter then register incoming particles and generate an electric signal. Commonly, microchannel plates are used which generate secondary electrons

11. Space Measurements

by an incoming particle. This results in a charge amplification which is measurable by electronics. One problem can be caused by high plasma densities when the detector becomes saturated and will therefore not register all incoming signals. An example of an ion spectrometer without a mass filter is shown in Figure 11.4. An ion mass analyzer is shown in Figure 11.5.

The amount of detected counts are then converted to a distribution function $f(r, \mathbf{v})$ which gives the number of particles per unit volume from a certain direction that have a certain energy E . This is also called phase-space.

Consider a small sector at a energy $E \pm \Delta E$. How many particles are measured for these measurement assuming the same density of particles in phase-space? We assume to measure at a fixed energy during a time interval Δt . For particle instrument $\frac{\Delta E}{E}$ is kept constant (usually around 10%) such that

$$\Delta E = CE.$$

The energy of a particle is proportional to the particles velocity $E \sim v^2$, therefore

$$\Delta E \sim (v + dv)^2 - (v - dv)^2 = v^2 + 2dv v + dv^2 - (v^2 - 2dv v + dv^2) = Cv^2. \quad (11.9)$$

Assuming the dv^2 is small and dividing the above expression by v^2

$$1 + \frac{4dv}{v} - 1 = C, \quad (11.10)$$

which gives

$$dv = \frac{Cv}{4}. \quad (11.11)$$

Thus, dv is proportional to v . The total volume in velocity space dv^3 with velocities that can get through the energy filter is then proportional to v^3 . Since all detected particles have velocities more or less perpendicular to the sector opening ΔA , the volume in space entering ΔA is $v\Delta A\Delta t$, thus proportional to v . Combining this we see that the counts, N , we detect in a specific energy interval is

$$N(v, x) = k_0 v^4 f(v, x), \quad (11.12)$$

with k_0 a constant depending on the instrument and sector. The number of counts is therefore dependent on $v^4 E^2$

The resulting distribution function $f(v, x)$ has the following relation to the number of counts detected in a sector

$$f(v, x) = \frac{N(v, x)}{k_0 v^4} = \frac{N(v, x)m^2}{4k_0 E^2}. \quad (11.13)$$

Note that a oxygen ion is 16 times heavier a mass spectrometer we will detect $16^2 = 256$ times as many hydrogen atoms, at a given energy, if the particle densities and temperatures are the same.

One other unit often computed by space scientists is “Differential Energy flux”, $J(v, x)$ which is defined as

$$J(v, x) = \frac{kN}{\Delta E}, \quad (11.14)$$

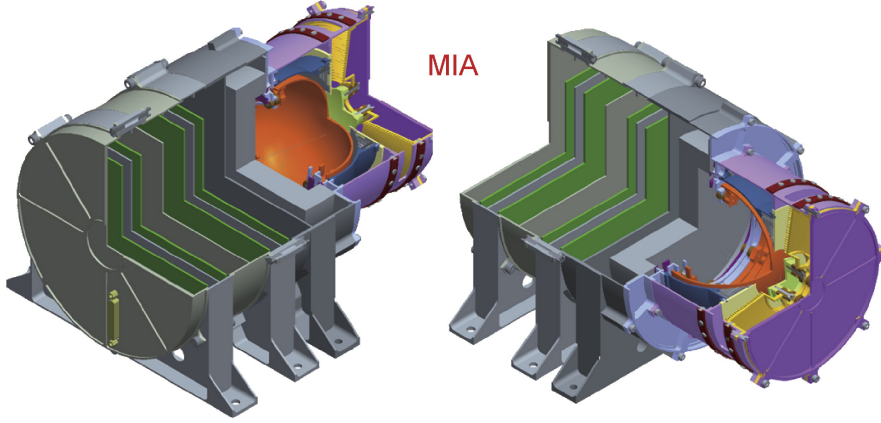


Figure 11.4.: Schematics of the ion spectrometer on board of the Mercury Magnetospheric Orbiter (MMO) of BepiColombo mission. This spectrometer has no TOF section and can therefore not resolve ion species. Figure from Saito et al. [2010]. The two spherical shells are marked in orange and blue.

where k is a constant depending on the sensitivity of the instrument, and ΔE the width of energy channel.

The particle distribution function can then be used to derive other physical constants such as the density n , the plasma velocity v , and the particle temperature T as described in subsection 3.3.1. Particle distribution can have different shapes depending on the plasma environment. One commonly used distribution is a Maxwellian distribution since thermalisation will cause any particle distribution to move towards a Maxwellian distribution. A 1-D Maxwellian distribution has the following form

$$f_1(v_x) = n_1 C \exp\left(-\frac{mv_x^2}{2k_B T}\right), \quad (11.15)$$

where

$$C = \sqrt{\frac{m}{2\pi k_B T}},$$

is a constant depending on mass m and temperature T . n_1 is the 1D particle density, and k_B is Boltzmann's constant. Thus, a function in 3D in Cartesian coordinates is

$$f_v(v_x, v_y, v_z) = f_1(v_x)f_1(v_y)f_1(v_z) = nC^3 \exp\left(-\frac{m(v_x^2 + v_y^2 + v_z^2)}{2k_B T}\right) \quad (11.16)$$

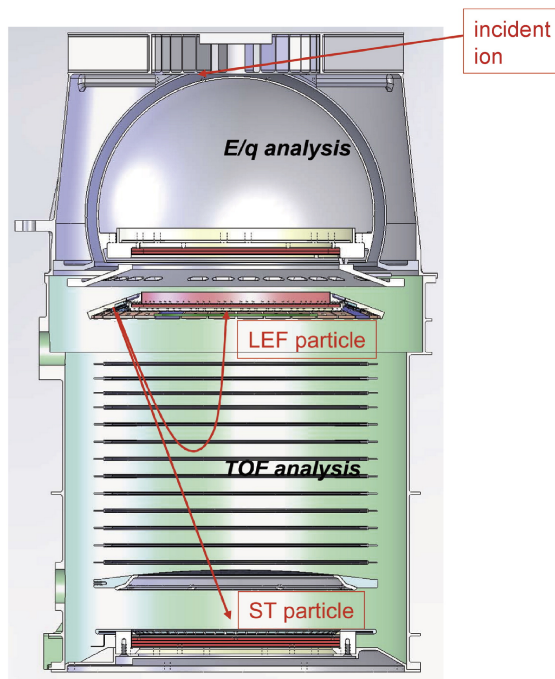


Figure 11.5.: Schematics of the ion msa spectrometer on board of the Mercury Magnetospheric Orbiter (MMO) of BepiColombo mission. This spectrometer has a section and can therefore resolve ion species. Figure from Saito et al. [2010].

12. Ground-based space measurements

12.1. Introduction

Astronomy is one of the oldest sciences. Systematic astronomical observations can be dated back to thousands of years in many ancient civilisations around the world. These observations were, however, limited to the part of the sky still visible today. The sky was, though, much easier to study even with the naked eye until quite recently due to the lack of contamination from electric light. The most important step to explore space further was the introduction of the telescope in the 1600s by Galileo. This technique opened a window into the solar system. This is also the region which today can be studied by active or in situ methods. The discipline is called *space physics* in contrast to astronomy, which monitors the sky by passive methods and today reaches far beyond our own galaxy out in the universe.

The development of photographic techniques in the late 1800s made it possible to use cameras and photographic emulsions to record various visible phenomena in the sky. The past century also opened a new realm in observations, this time at the non-visible wavelengths in the electromagnetic spectrum with an instrument called *RaDaR - Radio Detecting and Ranging*. After World War II and continuing into the early years of the space age with satellites in orbit, many redundant military radar systems were converted into research radars, and used in systematic space observing programs. Radars can measure continuously and they are not limited by daylight or bad weather. While spacecraft cover large areas along their orbit around the Earth, single satellites just receive information on the spatial variations along their path. By contrast, radars continuously monitor the same volume and can get very high time resolution on the processes occurring in the radar beam. These methods together have been very powerful for example in studying aurora in the polar regions.

Ground-based instrumentation has the advantages that is cheaper than instrumentation in space. Furthermore, the instrumentation can be continuously maintained and calibrated as needed. Disturbances and waves inside the magnetopause often travel along magnetic field lines which are connected to the ionosphere. Such a magnetic disturbance can be monitored from the ground with magnetometers. All closed magnetic field lines are connected to the ionosphere, therefore, ground-based instrumentation can be used to monitor the ionosphere and therefore the magnetosphere. Ground-based instrumentation provides a powerful tool to continuously monitor the magnetosphere and provide information about large parts of the magnetosphere.

In this chapter we present some basic characteristics of optical instruments and radars as well as how they are used for space observations. We start, however, by presenting some generally used geomagnetic indices and a set of basic observatory instruments useful to monitor the near space conditions. The *magnetometers* record continuously the state of the Earth's geomagnetic field and the *riometers* monitor the opacity of the ionosphere

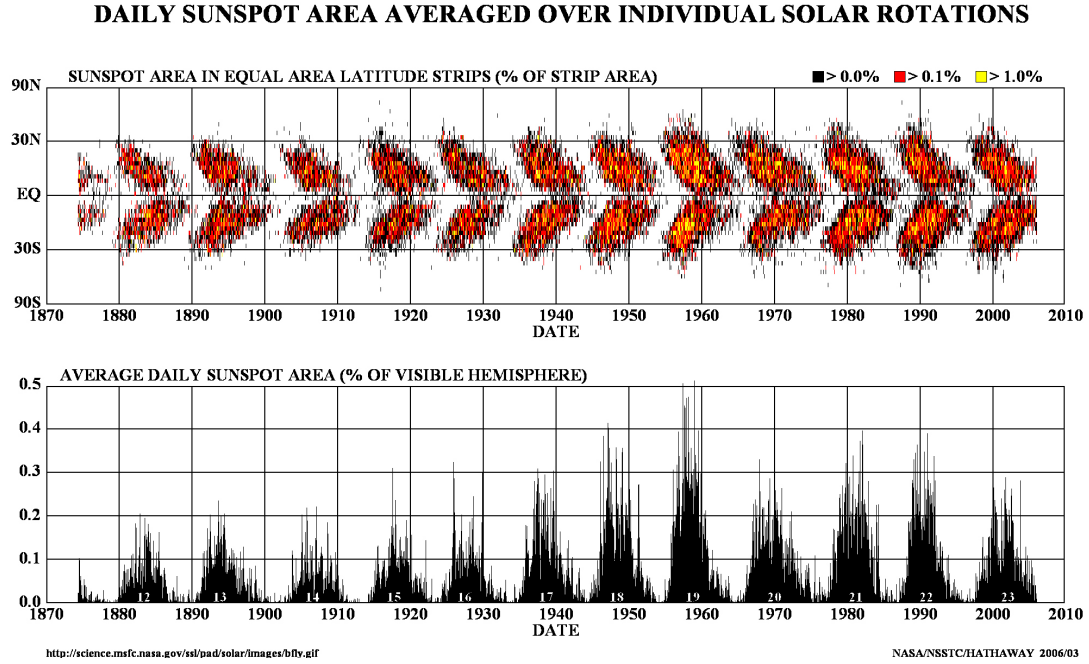


Figure 12.1.: Upper Figure: Sunspot area and latitude. Lower Figure: The variation of the sunspot numbers shows an approximately 11-year cycle. Credits: NASA/NSSTC/HATHAWAY

by recording cosmic noise transmission. The properties of the ionosphere can be probed by transmitting a radiowave and recording the signal that is echoed back to the ground with *ionosondes*. Using transmitters of increasing frequency and power, the echo can be produced by total reflection, coherent scattering against density fluctuations related to plasma waves in the ionosphere, or incoherent scattering against thermal fluctuations in the electron density.

12.2. Solar activity indices

Solar indices are indicators for the activity level on the sun and can give hints on upcoming disturbances in the solar-terrestrial interaction.

The *sunspot number* R gives the number of spots on the Sun. Figure 12.1 shows how R has varied over solar cycles since the start of observations. Sunspots are associated with active regions on the sun which are in return associated with solar flares and coronal mass injections. The number of sunspots is therefore an indicator for solar activity.

The *solar radio flux index* F10.7 is observed at a wavelength of 10.7 cm (2800 MHz) which corresponds to a radioemission line for iron and correlates with the solar extreme ultraviolet (EUV) flux as well as the sunspot number R . It is practical since it can be measured on the ground.

Furthermore, there are spacecraft at the L1 that continuously monitor the sun and take images at different wave length. Together these measurements give a good indication of the activity level of the Sun.

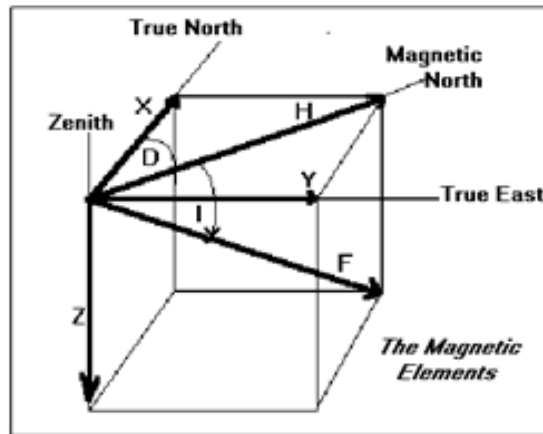


Figure 12.2.: The magnetic field at a point on the Earth's surface. The direction and strength of the magnetic field are usually given with the three orthogonal components, X (N), Y (E) and Z, or with one horizontal component and two direction angles (H, D, I). Credits: NOAA

12.3. Magnetometers

Magnetometers are versatile instruments for studying the local variations of the geomagnetic field. Short-term (a few seconds) variations are due to electric currents flowing in the E-region of the ionosphere (primarily at 100-130 km altitude) and induced in the surface layers of the solid Earth. Ground-based magnetometers function in the same way as the magnetometers described in subsection 11.2.3. However, care needs to be taken when positioning the magnetometers. Other sources of magnetic disturbances should not be placed nearby and the conductivity of the surrounding area influences the magnetometer.

Figure 12.2 shows how the geomagnetic field components are defined in the measurement process. Figure 12.3 illustrate the variation of the field components in so-called magnetograms. There are magnetometer networks all over the world and realtime magnetograms can be followed in all the geographical areas of interest. At low latitudes, an important source of magnetic disturbances is the magnetic storms caused by an intensified ring current (see section 6.8). In the polar regions the magnetic field variations are very well correlated with auroral activity.

A look at the local real-time magnetogram (and of course the cloud cover) can give an indication whether there is auroral activity or not. Figure 12.3 give examples of magnetograms during different geomagnetic activity. Auroral activity is often coupled to the substorm cycle (see section 8.5). Due to reconnection in the tail, there is a increased westward flowing current in the ionosphere on the nightside at auroral latitudes. A westward flowing current causes a southward disturbance in the magnetic field, and therefore $B_x < 0$. In the upper panel of Figure 12.3 such a substorm can be observed around 18-00 UT. From 18-19:30 UT there is a gradual decrease in B_x , which corresponds to the growth phase of a substorm. From 19:30-20 UT the B_x decreases drastically, which corresponds to the expansion phase. After 20 UT the recovery phase start where B_x recovers to pre-substorm values.

12. Ground-based space measurements

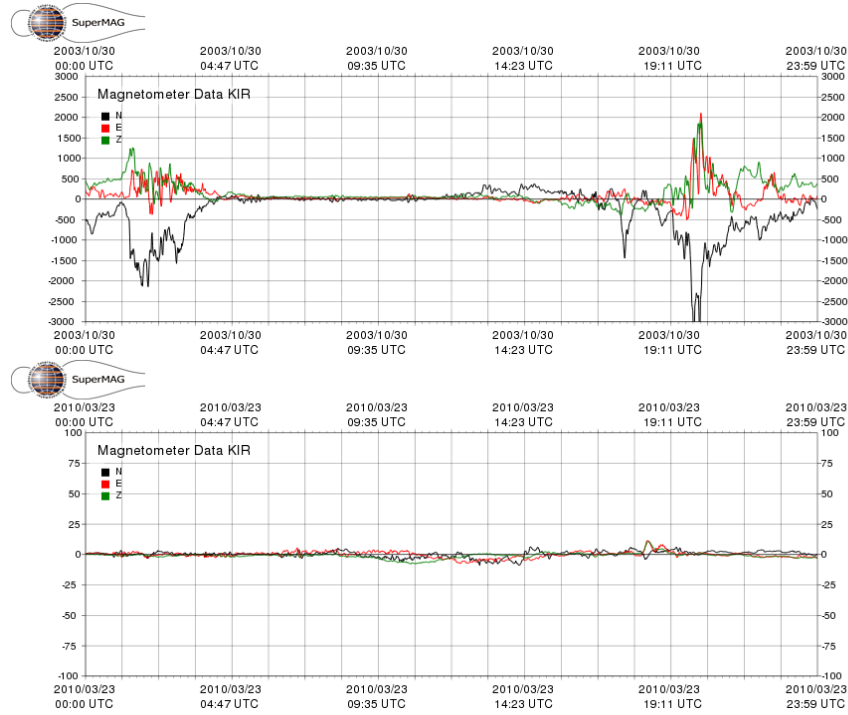


Figure 12.3.: Upper panel: The Kiruna magnetometer's three magnetogram components N, E, and Z under extreme high activity geomagnetic signature in October 2003, on a day which gave $K_p=8$. This solar storm period lasted for weeks and was called the Halloween Storm. Lower panel: The Kiruna magnetometer under very low geomagnetic activity. The global geomagnetic K_p index was 0. Note the different scaling of the y-axis compared to the upper panel. Quick look data from SuperMAG collection [Gjerloev, 2012] available under <https://supermag.jhuapl.edu/>.

12.4. Geomagnetic activity indices

Geomagnetic indices are useful to make a fast analysis of the degree of disturbances in the solar-terrestrial interaction. Three commonly used indices are the K_p -index, the auroral electrojet (AE) index, and the disturbance storm time index (Dst). There are some other indices that measure similar disturbances but for example use different magnetometer stations, these indices are then named differently. The indices are derived from magnetometers at different latitudes and are used to quantify magnetic disturbances.

The planetary K_p -index is derived from the *geomagnetic activity index* K -index at 13 magnetometer sites at mid-latitude locations. The K -index is defined by the most disturbed horizontal component of the geomagnetic field B_x during a 3 h period, see such variations in Figure 12.3. K varies between 0 and 9 and the conversion from the maximum fluctuation is determined individually for every magnetometer site. The K_p index can be used to quantify the level of geomagnetic disturbance and identify geomagnetic storms.

The AE-index quantifies magnetic disturbances in the horizontal component at higher latitudes in the auroral zone in the northern hemisphere. In total data from 10-13 observatories are used. For each station a monthly base value is determined from the five quietest days. Then, one-minute data from each observatory is collected and the base-value is subtracted. The data from all observatories is compared and the difference between the largest and smallest disturbance is the AE index. The AE index indicates the activity level of the electrojet, the horizontal currents flowing in the ionosphere. Due to the enhancement of the westward current during a magnetospheric substorm, the AE index will also be enhanced. The AE index can be used as an indicator of substorm activity which can cause auroral precipitation.

The Dst-index is derived from four magnetic observatories at low-latitudes and indicates the strength of the ring current. The ring current is enhanced during for examples magnetospheric storms. The index is derived from hourly average of disturbances in the horizontal magnetic field component compared to some baseline value. The Dst-index is another measure for geomagnetic storms.

12.5. Riometers

The riometers (Relative Ionospheric Opacity meter) measure the absorption of radio noise from the stars ('cosmic noise') in the ionosphere. They usually operate at frequencies between 25 MHz and 50 MHz where radio-waves are absorbed when there are significant numbers of free electrons in the height region from about 60 to 110 km. An increased ionisation in the lower ionosphere due to increased radiation (e.g. from a solar flare) causes an increase in absorption of the radio noise. Riometers make continuous observations of the noise level and observe decreases in noise relative to the quiet-day level. The riometers are not as sensitive instruments as magnetometers to show disturbances, but they show a dip in the absorption when intensive precipitation is associated with auroral activity. Data from riometers is shown in Figure 12.4 for different geomagnetic activity.

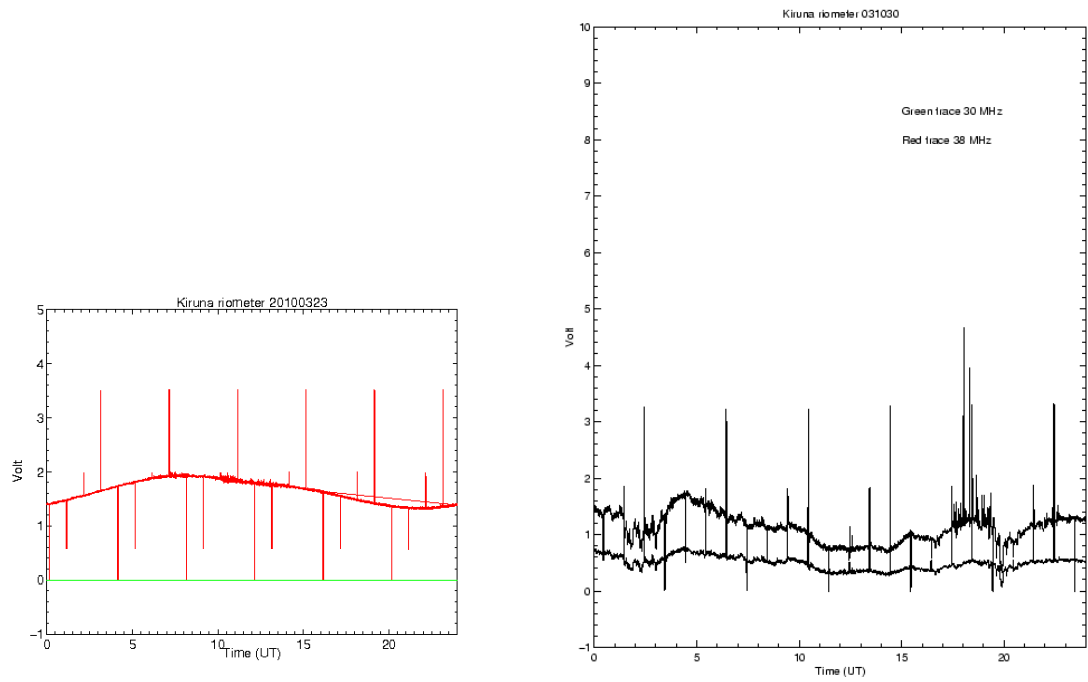


Figure 12.4.: Left: Typical undisturbed riometer record for the same date as the magnetogram in the lower panel of Figure 12.3. Right: A riometer record from the same date as the upper panel in Figure 12.3 showing substantial absorption around 02 UT and 18–20 UT. Quick look data from Kiruna Riometer operated by Kiruna Atmospheric and Geophysical Observatory (KAGO) at the Swedish Institute of Space Physics.

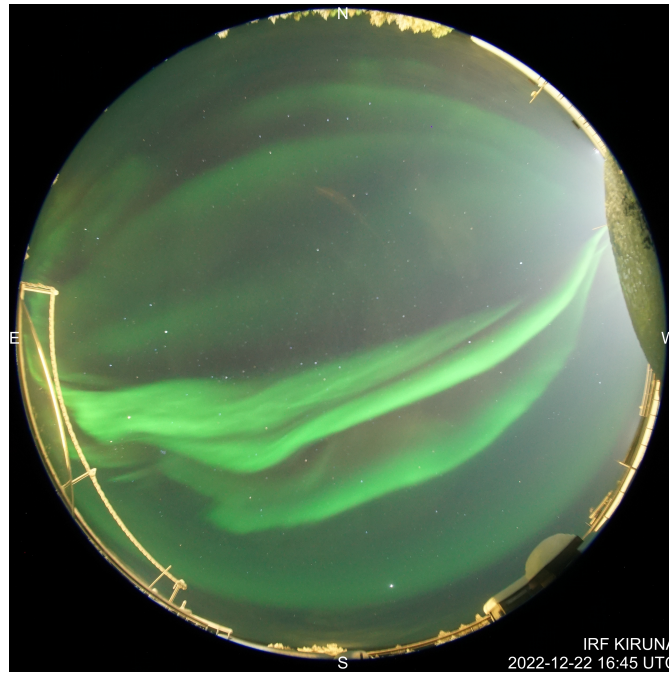


Figure 12.5.: Image from all-sky camera in Kiruna showing strong auroral activity. Data from the all-sky camera in Kiruna operated by Kiruna Atmospheric and Geophysical Observatory (KAGO) at the Swedish Institute of Space Physics.

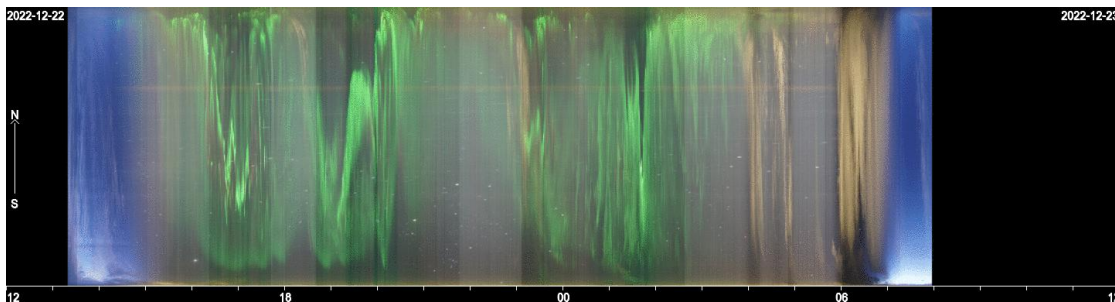


Figure 12.6.: Example of a keogram showing strong auroral activity throughout the day.

12.6. Optical methods

Aurora is the most apparent space phenomenon that can be observed on Earth. Since the northern lights are mostly visible in the sparsely-populated polar regions, they were a mythical phenomenon reported by sailors. That they rarely, maybe once in the 11-year solar cycle under exceptional solar activity, visit mid- or even southern latitudes made them even more mythical. It was not until Galileo's time that aurora were separated from the concept of meteors, both phenomena being something occurring "high in the air". Early visual observations of aurora were recorded by explorers and meteorologists on Arctic and Antarctic expeditions. With the introduction of photographic methods, systematic studies of auroras started. Their altitude, distribution over the Earth, time variations on scales from seconds to solar cycles, and different forms could be related to the geomagnetic field and solar conditions.

Spectroscopic studies showed that the observation of lines in the auroral spectrum by

12. Ground-based space measurements

Ångström showed that the light was emitted by the atoms or molecules in a gas. In the early 1900s, cameras that were sufficiently sensitive to take pictures of the aurora became available. Using cameras, the Norwegian physicist Störmer determined the altitude of auroras to about 100 km by comparing photographs taken simultaneously from different places. Today it is possible to obtain three-dimensional images of the aurora in different wavelengths by assembling data from arrays of digital cameras. Such pictures provide information about the fine structure of the aurora as well as the composition of the upper atmosphere and the flux of precipitating electrons at different energies.

All-sky cameras use special optical elements such as fish-eye lenses or spherical mirrors to acquire an image of the whole sky in one shot (hence the name). One ASC image covers a circular area with a diameter of about 600 km at 110 km altitude. An example of an image obtained by an all-sky camera is shown in Figure 12.5. All-sky cameras are operated in different band wavelength, even outside the visible spectrum. All-sky cameras can be used to determine if there is auroral activity and its evolution. When assuming a certain height of auroral emission one can map the obtained image onto a coordinate grid and determine the position of the auroral precipitation.

Photometers and spectrometers are two additional instruments that can provide useful information about auroral activity. A photometer cannot only take pictures but measures the amount of incoming photons and can therefore be used to describe the intensity of the aurora. This measurement can be used to estimate the energy flux of the particle precipitation that causes aurora. Spectrometers give information about spectral lines and therefore the atmospheric composition at the height of the emission.

The analysis of single can be tedious and therefore scientist often make use of keograms. Keograms show the intensity of the aurora in a time-versus-latitude dependent plot. Most often is a small band with north to south orientation from the all-sky camera are collected and added to one plot. An example of a keogram is shown in Figure 12.6.

Cloud coverage is often a problem for optical instruments, but images of the aurora can also be taken from above by cameras mounted on satellites. To reduce the background of light reflected from the atmosphere, satellite images of the aurora are usually made in UV-light with wavelengths less than 200 nm. This UV part of the sunlight is almost completely absorbed by ozone in the atmosphere, and very little is reflected back into space. Since auroral UV-light is emitted well above the ozone layer, a camera looking down from a satellite will see the aurora against a rather dark background, and good pictures can be obtained even if the atmosphere is sunlit. By comparing images at two or more wavelengths in the range 130–200 nm it is also possible to determine the characteristic energy of the precipitating electrons.

12.7. Radars

Radars were developed during the second world war to detect hostile approaching aircraft and to direct anti-aircraft weapons. Today radars are utilized for a variety of purposes. We discuss here the radars used for space research. For research applications, parameters such as frequency, power, beam width, steerability and location as well as the signal processing are essential. The frequencies range from MHz to GHz, corresponding to wavelengths of tens of meters to centimeters.

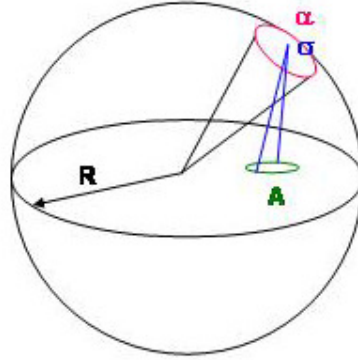


Figure 12.7.: A transmitting radar in the middle illuminating a target on the sphere, which scatters a part of the signal back to a receiver.

Radars can be classified as monostatic or multistatic. Monostatic systems use a single receiver, while multistatic systems use several receivers. Radars can be build using a radar dish or a phased array which uses several dipole antennas to generate the beam.

12.7.1. The radar equation

A radar comprises a *transmitter* and a *receiver* and a considerable amount of electronics and signal processing equipment. Here we derive a simple version of the *radar equation*. The transmitter generates pulses of electromagnetic energy with power P . The pulses are radiated from a gain antenna with gain G , that concentrates the transmitted power into a directed beam instead of spreading it equally in all directions. At range R_t from the transmitter, the antenna beam illuminates an area α , where the power density becomes $PG/4\pi R_t^2$ (see Figure 12.7).

A small target located within the border of α and having a radar cross-section σ scatters some of the power hitting it: $\sigma PG/4\pi R_t^2$. A fraction of the scattered power reaches the receiver after traveling back the distance R_r . The scattered power density is then $\sigma PG/(4\pi R_t^2 4\pi R_r^2)$.

If the receiving antenna has surface area A , it collects the scattered power

$$P_r = \frac{\sigma PGA}{4\pi R_t^2 4\pi R_r^2} \quad (12.1)$$

This is the fundamental form of the radar equation. The power P , the gain G and the aperture A are defined by the radar, while the radar cross-section σ contains the physics explained later in this chapter.

The raw data of a radar is therefore the received power, representing the power that was scatter in the ionosphere by different processes. Based on that raw data one can derive many useful physical parameters dependent on the radar system.

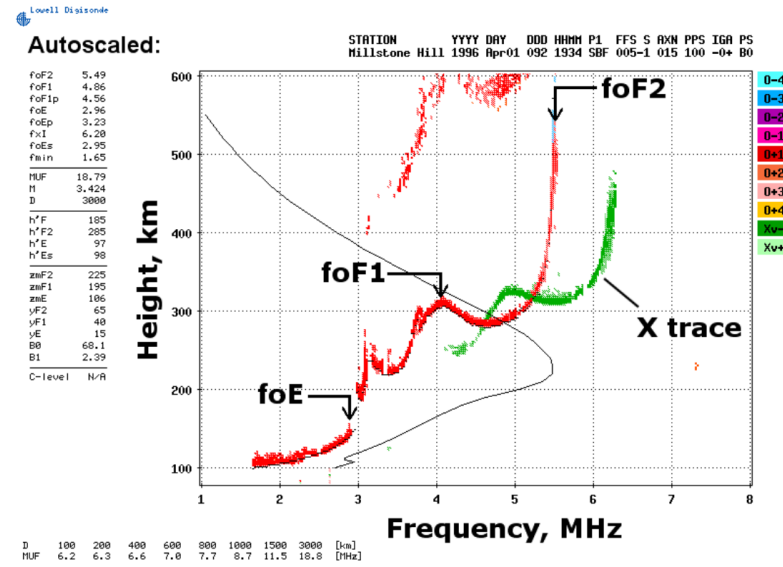


Figure 12.8.: Example of an ionogram. Figure from Reinisch et al. [2005].

12.7.2. Ionosonds

For many years, the ionosphere was mainly studied by analyzing the reflection of radio waves, using the physical principles discussed in section 7.3. A standard instrument was the ionosonde, a type of radar, which first constructed by Breit and Tuve in 1925. A vertically-transmitted wave with frequency f will be reflected at an altitude z where f equals the plasma frequency $f_{pe} = \omega_{pe}(z)/(2\pi)$. As long as f is less than the maximum plasma frequency in the ionosphere, the reflection is almost total and a clear echo is obtained even from a rather low power transmitter. By measuring the travel times of pulses with different frequencies, the variation of f_{pe} with altitude can be determined. Since

$$f_{pe} = \frac{\omega_{pe}}{2\pi} \approx 9 \sqrt{n_e [\text{m}^{-3}]} \text{ s}^{-1}, \quad (12.2)$$

the electron density profile $n(z)$ is then calculated from $n(z) \approx f_{pe}^2/81 \text{ m}^{-3}$.

Much of the terminology used in ionospheric physics is still influenced by these ionosonde observations. When probing the ionosphere with these instruments, it was often found that the travel times ($\propto$ height of reflection) increased in steps as the frequency was increased. These steps, which can be seen in the sample ionogram in Figure 12.8, were interpreted as layers with different electron densities, and they were named the D, E, F₁, and F₂ layers (see Figure 12.9). During nighttime only the E and F layer (which cannot be separated) is observed. Typically the E region peak is observed between 2–4 MHz and the F region peak between 3–12 MHz. From Equation 12.2 the corresponding electron density peak values can be estimated to $0.5\text{--}2.0 \cdot 10^5 \text{ cm}^{-3}$ in the ionospheric E region and about $0.1\text{--}1.8 \cdot 10^6 \text{ cm}^{-3}$ in the F region. The ionosonde can only provide rather indirect and crude information about the electron density in the ionosphere. However, the method is cheap and simple, and ionosondes with digital output that can be fed into computers are still used as standard instruments in research.

Ground-based ionosondes can only be used to probe the electron densities below the maximum. Information about higher altitudes became available after around 1960 when space technology had advanced so far that rockets could be launched into and above the ionosphere. Topside sounders, ionosondes mounted on satellites, were then used to

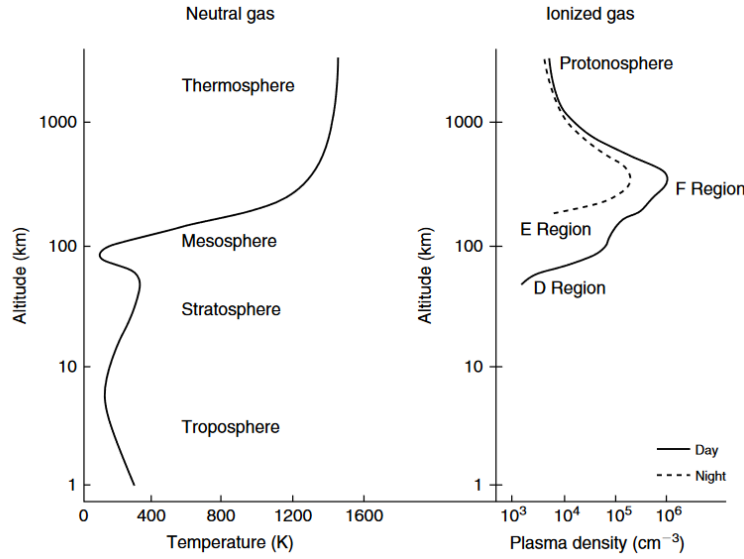


Figure 12.9.: The atmospheric and ionospheric layers. The curve on the right shows the ionization profile starting above 50 km in the D region, followed by the E region and F region divided into the two subregions F1 and F2. The curve on the left shows how the atmospheric temperature varies. It decreases in the troposphere, increases in the stratosphere and again cools in the mesosphere, reaching the coldest point in the mesopause and warming up again higher up with the decreasing density. Figure from Kelley [2009].

measure the electron densities above the maximum from above.

12.7.3. Coherent radars

There are always some fluctuations in the electron density in the ionosphere. When the ionosphere is disturbed, for example in connection with substorm activity, these fluctuations grow to form electrostatic waves. Low-frequency electrostatic waves are associated with field-aligned electron density variations, which act as a grating that efficiently scatters electromagnetic radiation. A coherent scatter radar transmits a strong radio wave, typically in the frequency range 8–20 MHz, and records the echo caused by scattering from density variations with scale lengths d of several meters. The term *coherent* indicates that the scattering is caused cooperatively by many electrons acting together within a large volume. In order to receive a signal the transmitted signal needs to be back scattered which is the case when the Bragg condition is satisfied

$$2d \sin \theta = n\lambda_r, \quad (12.3)$$

with the angle of incidence θ , the diffraction order n , and the wavelength of the radar λ_r .

From the Doppler shift of the received signal the line-of-sight velocity of the plasma can be determined. Multiple radar systems with an overlapping field-of-view then allow the determine the horizontal velocity. For example, a large part of the auroral region is covered by an international network of coherent radars called *Super Dual Auroral Radar Network* (SuperDARN). Based on the horizontal drift velocity the polar cap cross

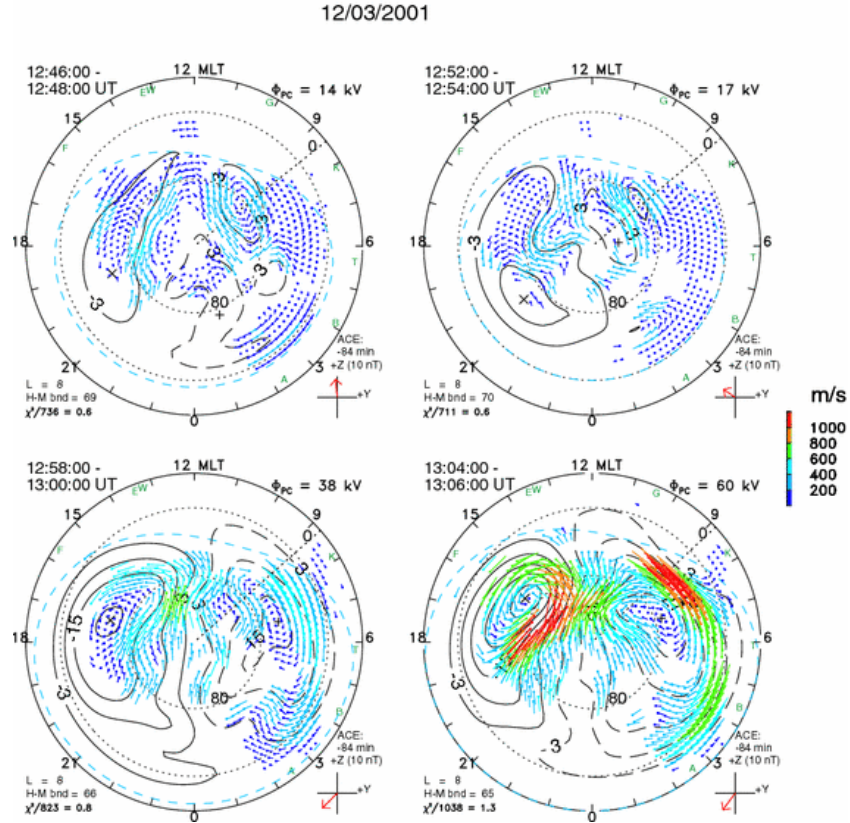


Figure 12.10.: Ionospheric convection pattern measured by SuperDARN. The dot represent the origin of the velocity vector, the line represents the direction and the color the speed. The dashed and dotted lines represent equipotential lines of the electric field derived from the flow velocities. The lower right convection pattern corresponds to the expected one from the Dungey cycle during southward IMF. The plasma/field lines are moving anti-sunward at high latitudes and sunward at low latitudes. Figure from Chisham et al. [2007].

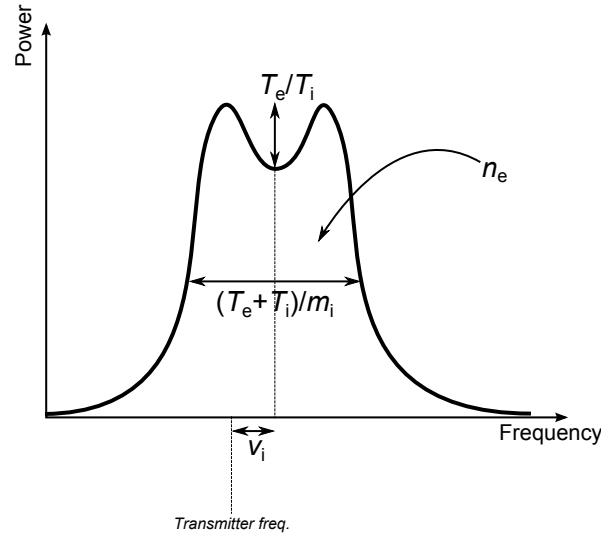


Figure 12.11.: The ion line in the incoherent scatter spectrum. Standard parameters that can be extracted from this are the electron density proportional to total area under the line, the ion temperature related to the width of the spectrum, the electron temperature related to the depth of the valley between the peaks of the spectrum and finally the ion velocity related to the Doppler shift of the whole spectrum. Credits: SRI International

potential $\mathbf{E} = -\nabla\phi$ can be determined using the drift velocity

$$v_d = \frac{\mathbf{E} \times \mathbf{B}}{B^2} \quad (12.4)$$

An example of the ionospheric convection and the derived large-scale electric field by SuperDARN is shown in Figure 12.10.

12.7.4. Incoherent scatter

The term "incoherent scatter" today is understood to refer to scattering from collective density fluctuations organized by waves in plasma created near thermal equilibrium conditions in the ionosphere. The dominant waves are ion-acoustic waves, corresponding to the collective motion of a cloud of electrons bound to ions, and Langmuir waves, corresponding to local longitudinal electron oscillations. For incoherent scatter radars the Bragg condition Equation 12.4 needs to be satisfied as well. The scale length of such density variations are less than a meter.

From the scattering from these waves, an incoherent scatter radar can monitor such instantaneous basic plasma parameters as *electron density*, *temperatures* and *plasma drifts*. Other quantities can be estimated, for example *collision frequency*, *mean ionic mass*, *electric field vector*, and some neutral atmosphere parameters. The cross-section for the incoherent scatter is very small; thus powerful transmitters, large antennas, and sophisticated signal-processing facilities are required to exploit this process as a useful ionospheric diagnostic. Due to the large costs there are only a few incoherent scatter radars in the world.

The incoherent scatter ion line spectrum is illustrated in Figure 12.11 showing how the four basic parameters n_e , T_e , T_i and v_i are acquired from the spectrum in an ideal case.

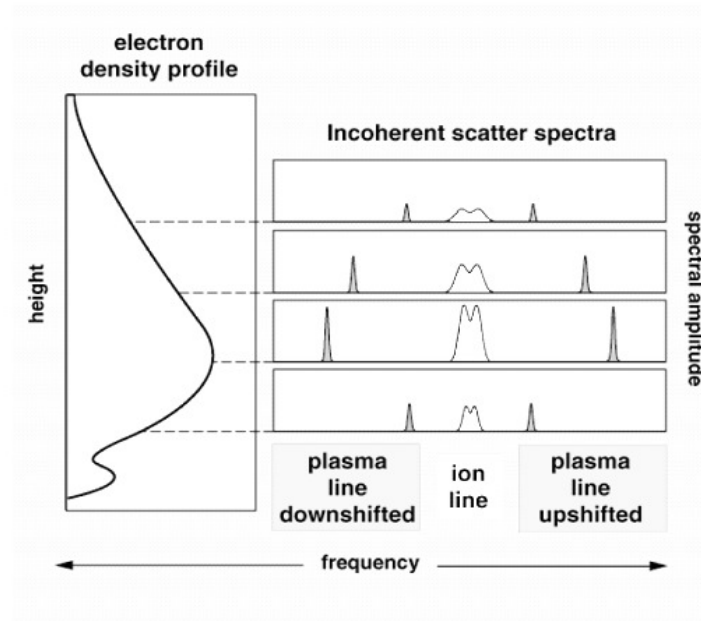


Figure 12.12.: The left panel shows the ionospheric profile schematically and the right panel how incoherent scatter spectrum typically looks like at the different altitudes. In addition to the ion line in the middle, the down- and upshifted plasma lines related to the Langmuir waves are included. Credits: SRI International

The spectrum results from a Fourier transform inverted from a measured autocorrelation function. The spectrum presents only a fraction of the complete measurement, a few seconds in time and hundreds of meters in altitude. However the actual resolution is very dependent on altitude range and physical processes of interest. The parameters are fitted with commonly accessible analysis programs, such as GUIDAP for EISCAT.

Figure 12.12 shows how the ion line and the plasma lines are related to the ionospheric density conditions and each other at different altitudes. Figure 12.13 shows the output of an analysis program for a four-day radar run extending from 80 to beyond 250 km. The four panels correspond to the four parameters and the interpretation of the data is given in the figure caption.

12.7.5. EISCAT - The European Incoherent SCATter radars

The European Incoherent SCATter Association (EISCAT) was established in 1976 to utilize the incoherent scatter technique in the polar regions, where auroral research had long traditions in Finland, Norway and Sweden. This was a collaborative effort between six European countries of which France and the UK had already used the new method to monitor the ionosphere with their domestic radar facilities St. Santin-Nançay in France and the Royal Radar Establishment at Malvern, UK. Germany was also one of the founding members and the three Nordic countries became the hosts for the organization. While the EISCAT Headquarters was located in Kiruna, Sweden, the two transmitters for the UHF and the VHF radars were placed at Ramfjordmoen close to Tromsø in Northern Norway. The 32 m parabolic dish receivers for the world-unique tristatic UHF radar were located in Kiruna, Sweden and Sodankylä, Finland. Later EISCAT received

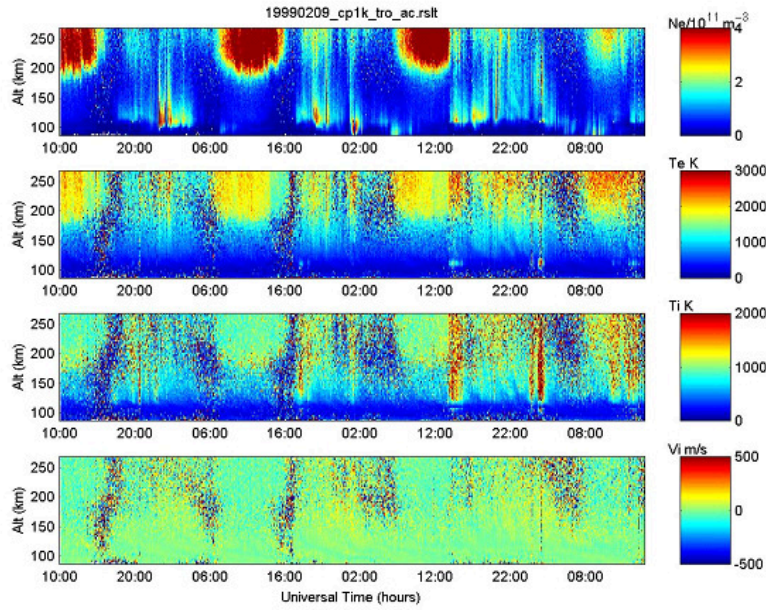


Figure 12.13.: EISCAT UHF data from the period 9-12 February 1999 showing the basic parameters of incoherent scatter resulting from a fit to the incoherent scatter spectra at every time and height. Panel 1 gives the electron density, panels 2 and 3 the electron and ion temperatures and panel 4 ion drift. The daytime maximum is seen both in the increase in the electron density as well as the electron temperature in the F region above 200 km. Typical auroral precipitation events are seen in the electron density panel during every night in the E region down to 100 km altitude. Some of these also result in the ion heating seen in the ion temperature panel down to 130 km altitude.



Figure 12.14.: The 42m and 32m 500 MHz EISCAT Svalbard Radar parabolic antennas. The one in the background is fully steerable, while the closer one points along the geomagnetic field line about 8.5 degrees from the zenith. Credits: EISCAT

a Heating Facility located at the Tromsø site from Germany.

In the 1990s the Association expanded northward to Svalbard as a common interest with Japan, which joined EISCAT (see the EISCAT Svalbard Radar antennas in Figure 12.14). Later China also joined the Association. The expansion of the mobile operator networks during the past two decades created competition with EISCAT for its bandwidth, especially for the UHF at 930 MHz, which was 16 MHz in the beginning. In 2011 the tristatic UHF system lost its capability in Finland.

12.7.6. EISCAT_3D

A lot of efforts have been made to replace the existing mainland EISCAT radars with a new, modern, next-generation incoherent scatter radar. The final EISCAT 3D system will consist of phased-array antenna fields located in northern Finland, Norway, and Sweden. Each site consist of about 20000 dipole antennas and sensitive receivers. The new system will allow the beam direction to change in milliseconds (compared to minutes in the traditional one-dish system) which wider spatial coverage and tracking of satellites and space debris. In addition, the sensitivity of the system is significantly improved and full drift velocity vectors can be measured. Due to the huge development in data storage and signal processing, phased array antennas (see Figure 12.15) have become popular with their interferometric capabilities. These can be electronically steered very quickly by changing the phase difference of the individual antennas or groups of them instead of moving an entire radar dish. They can also reach much better spacial resolution than traditional monostatic parabolic antennas.



Figure 12.15.: The EISCAT_3D demonstrator phased array in Kiruna. This antenna has received signals from the VHF transmitter in Tromsø. Credits: EISCAT

Bibliography

- S.-I. Akasofu and C.-I. Meng. Low latitude negative bays. *Journal of Atmospheric and Terrestrial Physics*, 30(2):227–241, 1968. ISSN 0021-9169. doi: [https://doi.org/10.1016/0021-9169\(68\)90078-0](https://doi.org/10.1016/0021-9169(68)90078-0). URL <https://www.sciencedirect.com/science/article/pii/0021916968900780>.
- W. Baumjohann and R. A. Treumann. Basic space plasma physics. *London: Imperial College Press*, 1996.
- G. Chisham, M. Lester, S. Milan, M. Freeman, W. Bristow, A. Grocott, K. McWilliams, J. Ruohoniemi, T. Yeoman, P. L. Dyson, et al. A decade of the super dual auroral radar network (superdarn): Scientific achievements, new techniques and future directions. *Surveys in geophysics*, 28:33–109, 2007.
- D. Ettelt. Conception and fabrication of a strain-gage magnetometer. 06 2012.
- S. Fatemi, A. R. Poppe, G. T. Delory, and W. M. Farrell. Amitis: A 3d gpu-based hybrid-pic model for space and plasma physics. In *Journal of Physics: Conference Series*, volume 837, page 012017. IOP Publishing, 2017.
- C.-G. Fälthammar. *Space Physics*. KTH, Stockholm, 2001.
- J. Gjerloev. The supermag data processing technique. *Journal of Geophysical Research: Space Physics*, 117(A9), 2012.
- P. Judge. Observations of the solar chromosphere. In *Solar MHD theory and observations: a high spatial resolution perspective*, volume 354, page 259, 2006.
- M. C. Kelley. *The Earth’s ionosphere: Plasma physics and electrodynamics*. Academic press, 2009.
- M. Kivelson and C. Russell. Introduction to space physics, 1995a.
- M. Kivelson and C. Russell. Introduction to space physics, 1995b.
- A. Krishnan and J. Kalluvila. Magnetohydrodynamic power generation. *Int J Sci Res Publ*, 3:1–2250, 06 2013.
- A. Kvammen. Auroral image processing techniques-machine learning classification and multi-viewpoint analysis. 2021.
- X. Morel, M. Berthomier, and J.-J. Berthelier. Electrostatic analyzer with a 3-d instantaneous field of view for fast measurements of plasma distribution functions in space. *Journal of Geophysical Research: Space Physics*, 122(3):3397–3410, 2017. doi: <https://doi.org/10.1002/2016JA023596>. URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JA023596>.
- L. Norin. *Secondary Electromagnetic Radiation Generated by HF Pumping of the Ionosphere*. PhD thesis, 12 2008.

Bibliography

- B. Reinisch, I. Galkin, G. Khmyrov, A. Kozlov, and D. Kitrosser. Automated collection and dissemination of ionospheric data from the digisonde network. *Advances in Radio Science*, 2(G. 1):241–247, 2005.
- C. T. Russell, J. G. Luhmann, and R. J. Strangeway. *Space physics: An introduction*. Cambridge University Press, 2016.
- K. Rönmark. Auroral current-voltage relation. *Journal of Geophysical Research: Space Physics*, 107(A12):SMP 11–1–SMP 11–8, 2002. doi: <https://doi.org/10.1029/2002JA009294>. URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JA009294>.
- Y. Saito, J. Sauvaud, M. Hirahara, S. Barabash, D. Delcourt, T. Takashima, K. Asamura, M. BepiColombo, M. Team, et al. Scientific objectives and instrumentation of mercury plasma particle experiment (mppe) onboard mmo. *Planetary and Space Science*, 58(1-2):182–200, 2010.
- T. E. Sarris. Understanding the ionosphere thermosphere response to solar and magnetospheric drivers: status, challenges and open issues. *Philosophical Transactions of the Royal Society A*, 377(2148):20180101, 2019.
- K. Steinvall. *Electrostatic plasma waves associated with collisionless magnetic reconnection: Spacecraft observations*. PhD thesis, Acta Universitatis Upsaliensis, 2022.
- E. Tanskanen. A comprehensive high-throughput analysis of substorms observed by image magnetometer network: Years 1993–2003 examined. *Journal of Geophysical Research: Space Physics*, 114(A5), 2009.
- W. T. Thomson. *Introduction to Space Dynamics*. Courier Corporation, 1986.
- E. G. Zweibel and M. Yamada. Perspectives on magnetic reconnection. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 472(2196):20160479, 2016.