PHYS 673: BEAM PHYSICS I

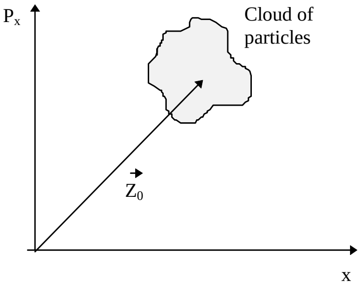
Lecture notes of
PHY861: Introductory Beam Physics
by Martin Berz (MSU)

I. What is Beam Physics

The field of Beam Physics deals with motion of **ensembles of particles** (usually charged) in electromagnetic **fields**. In the case of Beam Physics, all particles have **similar coordinates**. In many cases, the **positions** and **momenta** of the particles are sufficient to describe their motion. In this case, the particles are described by a state vector consisting of positions and momenta

$$\vec{Z} = (x, p_x, y, p_y, z, p_z)\,. \tag{1}$$

In other cases, additional coordinates may be needed; typical examples include the **mass**, sometimes the **charge**, or the **spin** vector of the particle. An ensemble of particles with such similar coordinates is called a **beam**, and the subfields concerned with the study of such beams is called **Beam Physics**. There are other subfields of Physics dealing with the study of the motion of such ensembles of particles; important examples are Plasma Physics and the Dynamics of Galaxies. These fields are different from Beam Physics in that in their case, the particles usually don't have rather similar coordinates, but occupy larger regions.



A Beam, an Ensemble of Particles

The space of state vectors $\vec{Z}$ is often called **phase space**, and a coordinate system showing $\vec{Z}$ is often called a **phase space diagram**diagram.. The **volume** of the cloud in phase space of particles has a special name, it is called **emittance.** As we shall see later, in many systems the emittance is conserved and hence plays a special role.

Because all particles are close together, it is often useful to pick one of these particles, typically one that is somewhere "in the middle", and describe the motion **relative** to this so-called **reference particle**. So if the reference particle has coordinate $\vec{Z}_0$,then the motion of the particles would be described in the relative coordinates $\Delta\vec{Z} = \vec{Z} - \vec{Z}_0$.

In many cases, the density of particles is so low that their **interaction** can be **neglected** or expressed by simple collective models.

If the fields are electromagnetic, then the motion is described by the **Lorentz Force Law** (Gaussian units)

$$\frac{d\vec{p}}{dt} = q\left(\vec{E} + \frac{1}{c}\vec{v}\times\vec{B}\right) \tag{2}$$

Here $\vec{E}$ and $\vec{B}$ are the electric and magnetic fields, respectively. These fields are connected to the scalar potential V and the vector potential $\vec{A}$ via the relations

$$\vec{B} = \vec{\nabla}\times\vec{A}\ ;\ \vec{E} = -\frac{1}{c}\frac{\partial\vec{A}}{\partial t} - \vec{\nabla}V \tag{3}$$

Although this may not be directly relevant now and only important later, we want to note here for the sake of completeness that the equations of motion in the form of the Lorentz force law can also be obtained from the Lagrangian

$$L = -mc^2\sqrt{1-\frac{v^2}{c^2}} + \frac{q}{c}\vec{v}\cdot\vec{A} - qV \tag{4}$$

From this Lagrangian, one can also obtain a Hamiltonian of the motion in a procedure that is standard for all Lagrangian systems.. One begins by defining the so-called canonical momentum:

$$\vec{p}_{can} = \frac{\partial L}{\partial\vec{v}} \tag{5}$$

which here has the form $\vec{p}_{can} = \gamma m\vec{v} + \frac{q}{c}\vec{A} = \vec{p}_{dyn} + \frac{q}{c}\vec{A}$; it is different from the relativistic dynamical momentum $p_{dyn} = \gamma m\vec{v}$. The Hamiltonian of the motion can then be found as $\overrightarrow{H} = \vec{p}_{can} \cdot \vec{v} - L$. This expression initially contains both $\vec{p}$ and $\vec{v}$, and it is necessary to eliminate $\vec{v}$ and express it in terms of $\vec{p}$. We find

$$\vec{v} = c \cdot \frac{\vec{p}_{can} - \frac{q}{c}\vec{A}}{\sqrt{\left(\vec{p}_{can} - \frac{q}{c}\vec{A}\right)^2 + m^2c^2}} \tag{6}$$

and then obtain for the Hamiltonian

$$H = \sqrt{\left(c\vec{p}_{can} - q\vec{A}\right)^2 + m^2c^4} + qV \tag{7}$$

When studying the fate of the beam from the time it is made until it is used, there are usually four steps involved. First, there must be a way for the **production** of the beam, and for the sake of efficiency if possible in such a way that its emittance is small. Next, in most cases the energy of the beam has to be increased; there has to be a mechanism of **acceleration**. Because of the outstanding importance of this process, the whole field is often called **Accelerator Physics.** Then it is necessary to **transport** the beam to where it is being used; and finally, there is often a need for **storage** of the beam for use at a later time or re-use. Lastly, often there is a need for **analysis** of the beam, in particular after the beam has been used for its purpose, which frequently is the facilitation of certain nuclear or high energy reactions.

II. Production of Beams

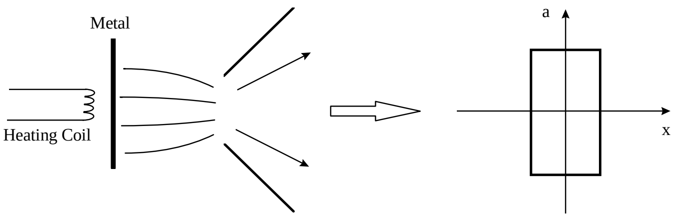
The mechanisms used for the production of the beam depend very much on the particular kind of beam that is needed, and they include atomic, nuclear, or high energy processes. In these notes, we want to be rather brief about the detailed processes and refer to the much more extensive lecture on ion sources.

A. Electron Sources

Electrons exist in abundance in metals, and to form them into beams requires their extraction from the metal. There are two main processes with which this can be achieved.

1. Heated Metal and Potential

The first method of extraction is based on **heating of metal**; by doing so, a small fraction of the electrons will achieve energies high enough to overcome the potential step that is necessary to leave the metal. Once outside the metal, they can be pulled away further by the application of strong fields. The basic principle is shown in the left part of the picture.



Electron Source

This method of acceleration usually leads to high currents. But because the area where the particles leave the metal is large, and they do so with a variety of different momenta, the emittances are usually rather large, as shown schematically on the right part of the above picture.

2. Field Emission

In the case of field emission, a **sharp needle** is brought into external electric fields. Because

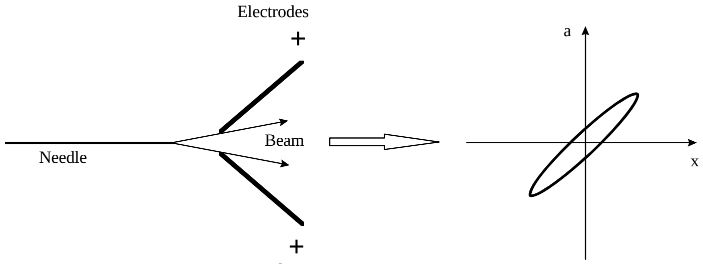


Figure 1: Electron Source

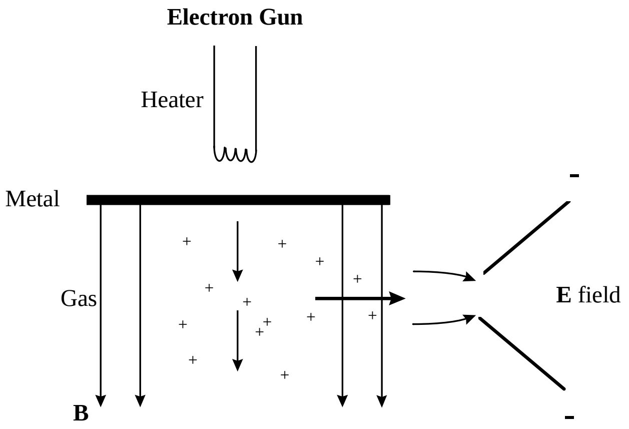
the needle is a conductor, it acts as an equipotential surface, and induces very strong electric fields near its tip. By choice of the right geometry, the fields can be made high enough to pull electrons out of the tip directly. All these electrons emerge approximately at one point, and usually their momenta are rather small; hence the emittance tends to be small.

B. Ion Sources

There are a large variety of different ion sources in existence, and they will be covered in detail in Richard Pardo's lecture. Two important basic mechanisms for the production of ion beams are as follows.

1. Bombarding Gas Atoms with Electrons

In this method which is useful for the extraction of beams from gases, a high-current electron source is used to send a stream of electrons through a gas. Due to collisions, ions are formed, which can then be extracted via some applied electric fields.

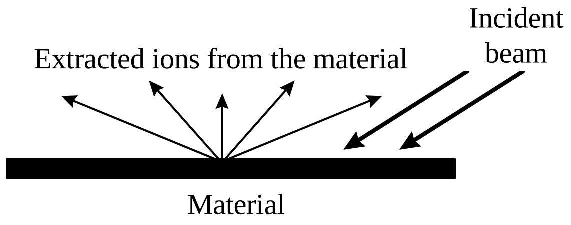


Ion Source

In order to not extract the electrons at the same time, magnetic fields are applied which restrict the movement of the relatively light electrons. Such ion sources can have quite high currents, but because the region of the space in which ions are produced is large and the ions can also have many different momenta, the emittance is often high.

2. Bombarding Surfaces Atoms with Projectiles

For non-gaseous atoms, one can often just hit a surface containing the atoms of interest with a beam of more easily produced particles. Due to the impact, often individual ions leave the surface, which can then be extracted with suitable electric fields.



Ion Source

Typically these sources have rather low current output.

3. Other Mechanisms

Depending on the kind of particles being desired, there are a large variety of other mechanisms to produce them. Important kinds of sources include positron sources (SLAC, LEP), antiproton sources (Fermilab), pions (LAMPF, Los Alamos), kaons, radioactive nuclei (MSU), etc.

III. Acceleration of Beams

We now assume that an ensemble of particles occupying a "small" volume of phase space has been created, and we thus have a "beam". In many if not most of the practical cases, the energy which the beam has after being produced by the source is not sufficient for the purpose it is to be used for, which frequently amounts to furnishing the energy necessary for atomic, nuclear, or particle processes of interest.

In most cases, the motion is best studied by first considering the motion of the **reference particle**, and once this motion is understood satisfactorily, to study the **relative** motion of the other particles. For a simple analysis of the relative motion, often a linear approximation with all the resulting simplifications is possible, but frequently a full understanding of the motion can only be achieved by considering the nonlinear effects.

Considering the special shape of the Lorentz force law, since $\vec{v}\times\vec{B}$ is perpendicular to the velocity $\vec{v}$, it is apparent that magnetic fields cannot be used for purposes of acceleration, which requires forces in the direction of the particle. Thus any **acceleration** has to be **provided by electric fields**. However, as we shall see, also magnetic fields have very good use in particle accelerators, as they can be employed to guide the beam to where it is needed. In particular, in the process of acceleration they are often used to guide the beam through the same region of electric field repeatedly and thus allow to maximize the use of the electric fields. Indeed, for this purpose of guiding the beam magnetic fields are usually even better suited than electric fields, since for the high velocities that beams usually have after even modest acceleration, the forces that can be attained with technologically available magnetic fields far exceed those that can be achieved with the respective electric fields.

Very generally, the amount of energy K a particle gains while travelling from time t_1 to time t_2 in an electric field $\vec{E}(\vec{r},t)$ that depends on position and time is given by the path integral

$$K = q \cdot \int_{t_1}^{t_2} \vec{E}(\vec{r}(t),t) \cdot \vec{v}(t)\; dt, \qquad (8)$$

where $\vec{r}(t)$ is the particle's position as a function of time and $\vec{v}(t)$ its velocity. In the special case that $\vec{E}$ is **time independent** and hence can be written in terms of a potential via $\vec{E} = -\vec{\nabla}V$, this path integral reduces in a natural way to the difference in potential as

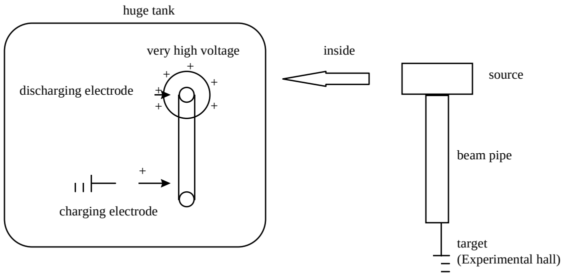
$$K = q \cdot (V(\vec{r}_1) - V(\vec{r}_2)) \qquad (9)$$

This simple fact implies a very important consequence for the design of electric accelerating fields: if there is to be any chance to utilize the same electric field **repeatedly** for the purpose of acceleration, then the electric field has to be **time dependent,** because otherwise repeated passing just results in a periodic increase and decrease of energy. In fact, the attempt to build an accelerator trying to create energy repeatedly by flying through the same time independent field is tantamount to the attempt to build a perpetual motion machine.

A. The Van de Graaff

The van de Graaff accelerator and several similar devices derived from it are the main representatives of the class of accelerators utilizing time independent fields. The voltage difference that the particles travel through is obtained with a **van de Graaff generator**, which consists of an endless non-conducting belt onto which charge is sprayed from a tip via field emission and then transported to the inside of a hollow metal sphere where it is deposited. Since any charges on a conducting object accumulate on the outside and create a field-free interior, new charge can be brought in from the belt on the inside of the sphere without experiencing any opposing fields, and thus large amount of charges can be accumulated on the sphere, resulting in a very high potentials.

In passing it is worth to remark that while the newly added charge does not experience a field when moving from the belt to the inside of the sphere, it certainly experiences a field while being approaching the sphere and being attached to the belt. Thus the potential energy contained on the charged sphere does not come for free, it is generated through the mechanical work that is necessary to move the belt and the attached charges towards the sphere.



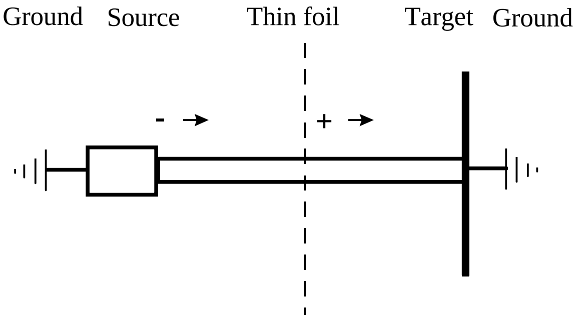
The Van de Graff Accelerator

The charged sphere is connected to a metal enclosing containing the ion source, thus elevating the source to a potential, which can then be utilized for the acceleration of the particles.

The main practical limitation of the van de Graaff Accelerator is the necessity to prevent sparks. This is achieved on the one hand by sheer **size**, because at the same potential difference, larger size means less electric field strength. On the other hand, it is important to **inhibit** the **spark formation** process. Microscopically, sparks form in a gas when small numbers of charged particles have a mean free path length that is long enough so they can attain energies sufficient to ionize other particles upon collision, resulting in an avalanche.. This can be avoided by choosing inert gases like He or SF_6, and on the other hand applying high pressure to reduce the mean free path length.

The van de Graaff accelerator has several desirable features, for example it can produce a fully continuous beam at high beam current. Its main limitation are the relatively low energies that it can produce, which seldom exceed about 20 MeV.

B. The Tandem Van de Graaff



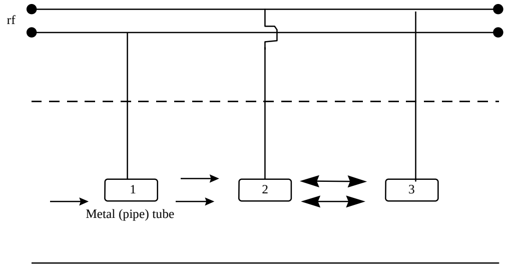
The Tandem Van de Graff Accelerator

The Tandem van de Graaff is an efficient modification of the van de Graaff concept, in which both the source and the target are kept at ground potential and which can efficiently **increase the energy** that can be obtained. For this purpose, a source is chosen that produces negatively charged ions, which are sent through a regular van de Graaff. At the end of the accelerating section, the ions are sent through a thin foil, in which many of them are **stripped** of some of their electrons, resulting in positive ions. Because the particles already have substantial energy when hitting the foil, often much higher charge states can be produced than in the ion source itself. These positive ions are sent through a second stage van de Graaff, which is essentially a reversion of the first stage, and by the time the target is reached, depending on their charge state after stripping, their energy is increased twofold or more. Having very similar characteristics to the original van de Graaff, the energies that can be achieved in this way are in the range of up to 60 MeV.

C. The Linac (Linear Accelerator)

It is an important observation that the field strength that can be obtained **in quickly oscillating** (rf, or radiofrequency) **electric fields** can be **substantially higher** than those that can be made statically in devices of similar size. This is partly due to reduced presence of spark formation, because the formation of an avalanche of charged particles requires time scales that are usually larger than the time the field is in one phase.

The use of an oscillating field, however, immediately entails that only half of the cycle can be used for acceleration, and thus different from static accelerators, the resulting beams always have a temporal **microstructure.** In practical use, usually several rf resonators are used sequentially, each one of which accelerating the particles, and it is very important that the phase relationship between the individual accelerating sections is correct. This is usually achieved by applying the fields between the edges of adjacent conducting tubes. The lengths of the tubes are chosen in such a way that the time the particles require to fly through them equals one half of the rf period, so that the particles never "see" an electric field of the wrong sign.



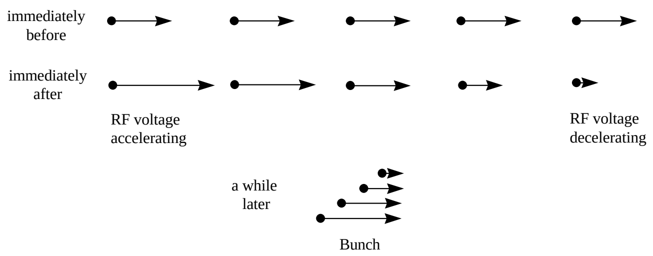
The Linac

So the length L_i of the ith tube has to be chosen so that it satisfies

$$L_i \approx \frac{1}{2} v_i T_{rf} \qquad (10)$$

where T_{rf} is the period of the rf frequency. Apparently this leads to a system of tubes of increasing length, i.e. $L_1 < L_2 < L_3 < \dots$ The exact lengths L_i of course depend on the relationship between the kinds of particles and the values of the accelerating voltages, and so often these designs are rather customized geometries. The geometric situation is much more straightforward for particles that already enter at speeds close to the speed of light, which allows the use of a purely repetitive geometry.

In order to maintain acceleration of all particles, it is important that the particles are injected into the linac with the right phase information; in particular, no particles should enter during the time the field points in the opposite direction. So the incoming beam has to consist of a uniformly arranged sequence of **bunches.** These bunches can be produced from a continuous beam of particles, which is what most ion sources deliver, by means of a **buncher.** This is merely an rf structure which accelerates the particles of a continuous beam different depending on the time at which the particles enter, as shown schematically in the picture. Because of the resulting different velocity profile, after a certain time the fast particles will eventually tend to catch up with the slow ones, resulting in packets of particles.



Schematic of a Buncher

An interesting **combination** of the need for **bunching, accelerating**, and focusing (which is discussed later in much greater detail) is the so-called Radio Frequency Quadrupole, or **RFQ**, accelerator.

In general, Linacs can provide beams of high current, and of higher energies than static accelerators, yet because of the single use of each electric field, they are still rather expensive per MeV. Linacs are frequently used as **pre-accelerators** for accelerators of higher energies. They also have the distinctive advantage that they **avoid synchrotron radiation**, which is often a limiting factor in circular accelerators for light particles such as electrons and positrons. This aspect is very important at SLAC and the Stanford Linear Collider, and is the main reason for the interest in a "Next Linear Collider", a pair of two Linacs shooting electrons and positrons at each other at high energy.

D. The Betatron

The betatron is arguably the simplest circular accelerator, and besides its practical use as a compact accelerator for lower energies, it is also a beautiful textbook-style application of principles of electrodynamics. In the case of the betatron, the orbit follows a circular shape, which is achieved by a magnetic field. If the motion is perpendicular to the magnetic field, then we have (SI units)

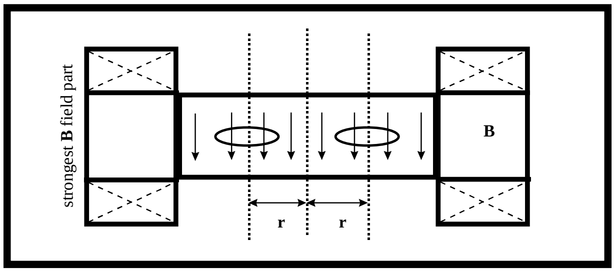
$$\frac{mv^2}{\rho} = qvB, \text{ and so } \rho = \frac{mv}{qB} = \frac{p}{qB}, \qquad (11)$$

and so the radius of motion depends only on the momentum and charge of the particle as well as the magnetic filed. Note that the equation is correct even in the relativistic case, if m is understood to mean the relativistic mass $m = \gamma \cdot m_0$. Commonly the ratio of momentum and charge p/q is denoted by χ_m and called **magnetic rigidity;** we apparently have

$$B\rho = \frac{p}{q} = \chi_m. \qquad (12)$$

Because $\chi_m = B\rho$, the magnetic rigidity has the unit Tesla meter, and is frequently simply referred to as "**B rho**".

In the case of the betatron, both **bending** and **acceleration** come from the same source, namely a **magnetic field** whose strength **increases with time** in such a way that its magnitude matches the increasing energy of the particles to keep them at nearly **constant radius**, and the circular **induced electric field** provides the **acceleration** for the particles.



The Betatron

In passing it is worthwhile to note that the basic idea of utilizing an electric field produced by a changing magnetic field also occurs in a much less mundane application from daily life: certain modern **cooking surfaces**. In this case, the electrons that are "accelerated" are not within the vacuum of a beam pipe, but merely in the metal that constitutes the bottom of the pot used for cooking; and of course since their mean free path is short, they don't attain high energies before colliding with either other electrons or the lattice atoms, thus transferring their whole kinetic energy to heat.

A quantitative understanding begins with Faraday's law of induction, now one of the Maxwell equations:

$$\begin{aligned}\vec{\nabla}\times\vec{E} &= -\frac{d\vec{B}}{dt}\Rightarrow && (13)\\ \int_A \vec{\nabla}\times\vec{E}ds &= -\frac{d}{dt}\int_A \vec{B}ds\Longrightarrow && (14)\\ \oint \vec{E}dl &= -\frac{d}{dt}\Phi\end{aligned}$$

where $\Phi = \int_A \vec{B}ds$ is the flux of the magnetic field through the surface. Here we restrict our interest to circular orbits with a radius called r, and the surface is the inside of the circle. Building the magnet rotationally symmetric entails a rotational symmetry of the fields, which simplifies the situation to

$$E = -\frac{1}{2\pi r}\frac{d}{dt}\Phi = -\frac{1}{2\pi r}\pi r^2\frac{d}{dt}\bar{B} = \frac{r}{2}\cdot\frac{d}{dt}\bar{B} \quad (15)$$

where $\overline{B}$ is the average magnetic field enclosed by the orbit. Thus we obtain for the momentum

$$\frac{d}{dt}(mv) = -qE = \frac{qr}{2}\cdot\frac{d}{dt}\bar{B}\Rightarrow mv = \frac{1}{2}qr\overline{B} \quad (16)$$

On the other hand, it must be true that the centrifugal force on the orbit with radius r is compensated by the Lorentz force there, which requires

$$\frac{mv^2}{r} = qvB(r)\Rightarrow mv = qrB(r) \qquad (17)$$

Thus altogether we obtain the following relationship between the field $B(r)$ at the orbit r and the average field:

$$B(r) = \frac{1}{2}\overline{B}; \qquad (18)$$

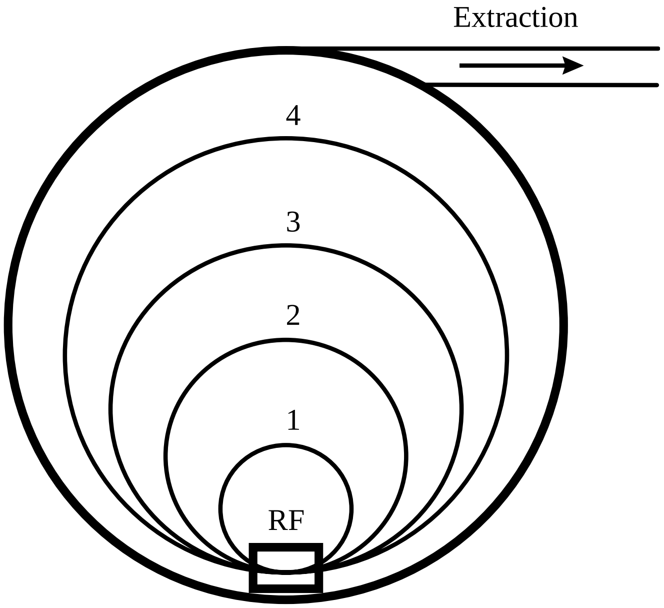
in order to achieve this, requires a magnetic field that is **stronger in the center** than where the particles move, which can be achieved by suitably **shaping the poles** of the magnet.

In principle the temporal behavior of $\overline{B}$ is irrelevant; in practice, one usually tries to "ram-p" it quickly, because the **pulsed beam** is only available at the end of ramping. This is usually achieved by making the magnet part of an LC circuit, which also conveniently allows to recover the energy stored in the magnetic field for the next ramping. For the practical use, it is important to try to limit Eddy currents in the iron, and in order to maintain the condition $B(r) = \overline{B}/2$, it is important to control saturation effects that may occur at any "edges" of the magnet.

The practical use of betatrons is nowadays mostly for electrons, where energies of about 300 MeV have been achieved; for protons, the values are about 50 MeV.

E. The Microtron

Also in the microtron, a magnet is used to bend the particles to let them pass through the same source of electric field repeatedly. Different from the betatron, the emphasis here lies on the production of a **continuous beam**. Since this requires that the whole acceleration process must be independent of the specific time of injection, this entails that that the **magnetic field is constant in time**. Thus an external voltage source is needed; as discussed above, if it is to be used repeatedly, it has to be time-dependent source, and in practice it is chosen to be an rf cavity. Altogether, the motion follows a sequence of **tangential circles** of increasing radius that touch at the location of the rf, as shown in the picture.



The Microtron

In order to **synchronize** the particle's motion and the momentary direction of the magnetic field, the revolution frequency of the rf has to be a multiple of the particle's revolution frequency, which can be obtained simply from

$$\frac{\gamma m_0 v^2}{r} = qvB \Rightarrow \omega = \frac{v}{r} = \frac{q}{\gamma m_0}B. \quad (19)$$

This means the motion has to be either such that $\gamma = 1$, which corresponds to nonrelativistic motion and hence severely limits the energy. The other possibility is to provide just enough acceleration in each turn that the revolution frequency decreases to the next multiple of the rf frequency. So the revolution frequencies would follow the pattern

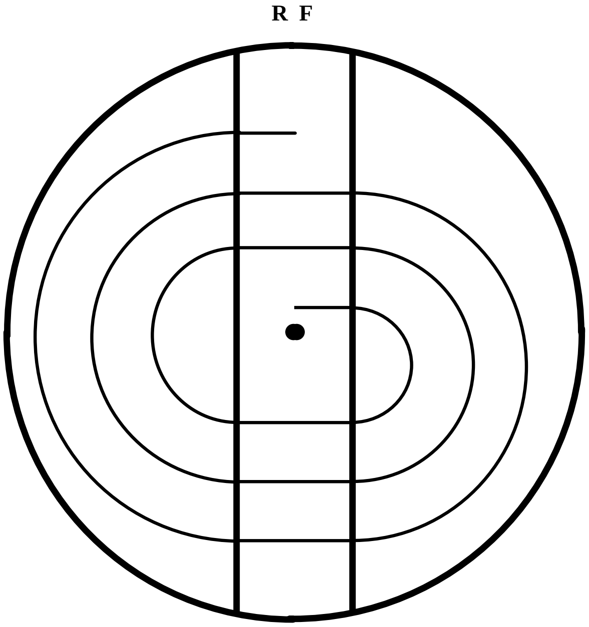
$$\omega = \omega_0, \frac{\omega_0}{2}, \frac{\omega_0}{3}, \frac{\omega_0}{4}, \frac{\omega_0}{5} \dots \quad (20)$$

This entails that the factor γ follows the sequence $\gamma = \gamma_0, 2\gamma_0, 3\gamma_0, 4\gamma_0, \dots$, which requires $\Delta\gamma = 1$ per turn. Since $E = mc^2 = \gamma m_0 c^2$, this means $\Delta E = m_0 c^2$, and thus the necessary energy gain per turn must equal the rest mass energy of the particle under consideration! For electrons, this means $\Delta E = 511$ keV and is thus possible, for protons and $\Delta E = 938$ MeV this is not easily possible within the the confines of a conventional magnet.

A very important further **development** of the concept of a microtron is based on the fact that if the orbits of the particles are far enough separated so that one can apply different magnetic fields for each orbit and can even change the shape of the orbit away from circular, then by careful choice of the orbit lengths, it is possible to maintain the synchronicity condition 20 while maintaining the freedom to have any amount of acceleration that is convenient. This is the basic idea behind the Continuous Electron Beam Accelerating Facility **CEBAF** at Thomas Jefferson National Laboratory, which will be covered in great detail in a future lecture.

F. The Cyclotron

The basic idea of the cyclotron is similar to that of the microtron, except that the rf is used more efficiently by providing acceleration twice or even more times per turn, and the orbits roughly follow **concentric circles**.



The Cyclotron

According to eq. (19), the revolution frequency is

$$\omega = \frac{1}{\gamma} \cdot \frac{q}{m_0} B \tag{21}$$

and the momentary radius of the orbit s

$$r = \frac{p}{qB} \tag{22}$$

This entails very similar restrictions regarding relativistic effects as in the case of the microtron; as before, any deviation from constancy of the magnetic field prevents continuous injection of beam and hence leads to a non-continuous outgoing beam. But because the orbits are nearly concentric, it is possible to at least partly compensate the relativistic effects by **increasing** B **radially** in such a way that the frequency in eq. (22) stays constant. If it is necessary to accelerate different particles in the same machine, then that entails that the actual field profile has to be adjustable, which is usually achieved by having one or several **trim coils**. The superconducting K1200 cyclotron located at NSCL on the MSU campus allows for such corrections of the profile of the magnetic field, and is currently the cyclotron achieving the highest energy.

If continuity of the beam is not of prime importance, it is possible to make the necessary relativistic corrections due to eq. (22) via a decrease of the rf frequency during the acceleration process, which is done in the case of the **synchrocyclotron.** This decrease obviously has to happen very quickly over the few hundred turns the particles stay within the accelerating structure, and thus the pulse frequency can still be rather high.

G. The Synchrotron

For any accelerator, the **ultimate energy limitation** comes from the strength of the magnetic field that is available via the **unavoidable restriction**

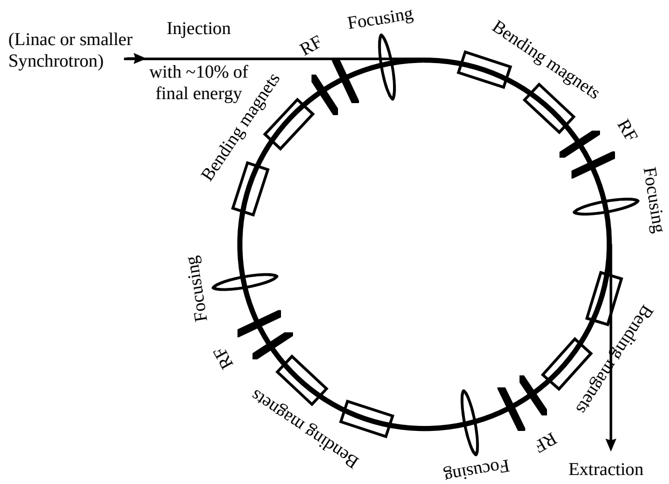
$$B\rho = \chi_m = \frac{p}{q} \tag{23}$$

But the range of available magnetic fields is rather limited; typical numbers are in the range of 1-2 T, the superconducting dipoles at SSC were planned to have 6.6 T, and the superconducting dipoles at LHC are planned to operate reliably at 8T. The highest magnetic fields that can be achieved are currently in the order of 30 T (Florida State University), but at a rather inconsistent field quality and usually not over extended times. So the only way to achieve high energies is to increase the deflection radius ρ.

This represents a significant **practical limitation to continuous beam accelerators**, in which B must be time independent and the size of the orbits increases in the acceleration process, since any region in which the beam may come has to be covered by magnetic fields. So for **really high energies**, the only realistic option is to have the particles follow the **same orbit** all the time by **ramping** the magnetic field during acceleration, and thus limit the region that has to be covered by magnetic field.

Of course this ongoing adjustment of the magnetic field during the acceleration process according to eq. (23) to maintain constancy of ρ prevents continous injection and hence continuous beams. Furthermore, since electric field strengths are comparatively limited, the fields of the cavities have to be re-utilized many thousands of times, resulting in a rather stretched-out acceleration process, and thus a rather **low repetition rate** of beam pulses.

All these thoughts lead to the concept of the **synchrotron**, in which the magnetic field strength is synchronized with the current energy or momentum of the particle so as to maintain a constant orbit. The figure shows a very schematic view of a synchrotron, consisting of a long nearly circular beam pipe, many bending magnets, and rf cavities. As we will see later, in the synchrotron it is particularly important to take proper care of the details of the relative motion, the most important aspect of which is the use of special "focusing" devices.



The Synchrotron

The table below shows a small number of hadron synchrotrons, their approximate dimensions, and the maximum energies for which they are designed.

IUCF	$\rho \approx 10m$	
Tevatron	$\rho \approx 1km$	$E \approx 1Tev$
SSC	$\rho \approx 15km$	$E \approx 20Tev$
LHC	$\rho \approx 5km$	$E \approx 8Tev$

H. The Storage Ring

The storage ring is not really an accelerator, it is a device to store the beam that is produced once so that it can be re-used; essentially it is a **synchrotron with rf turned off**. In many cases, particles orbit for minutes or days. In the case of the SSC, the desired time was about 8 hours, resulting in

$$n = \frac{3 \cdot 10^8 \mathrm{m/\,sec} \cdot 28800\,\mathrm{sec}}{8 \cdot 10^4 \mathrm{m}} \approx 10^8 \mathrm{turns} \tag{24}$$

Even more so than in the case of the synchrotron, one of the main design problems and physically perhaps the greatest challenge is to try to assure that particles actually stay contained over this large number of turns. Becaue the motion is nonlinear, this immediately leads to questions of nonlinear dynamics with all their interesting aspects.

One of the main applications of storage rings is in the **collider**, where counterrotating beams are brought to collision at various points around the ring. At very high energies, colliders have a significant **energy advantage** over fixed-target machines because a very large fraction of the beams' energies can be converted to reaction energy. As a detailed study of the relativistic dynamics shows, this is not at all the case for fixed target machines; in fact, conservation of energy and momentum severely limits the energy that can be set free. Important colliders are the Tevatron $(p, \overline{p})$, the now defunct SSC as well as the LHC (p, p), HERA (p, e^-),and LEP

(e^+, e^-).Besides the energy advantage, storage rings also limit the disadvantage of the slow ramping times typical for synchrotrons in that once the beam is stored, it is essentially **continuous** again.

But also for situation that require the beam to hit a **fixed target,** storage rings often offer an advantage over the use of synchrotrons by themselves, because it is often possible to extract the beam much more slowly than in the case of the synchrotron, resulting in a more easily managable duty cycle and reducing the problem of overflowing the electronics in the detectors. In this method of **ultraslow extraction**, the nonlinear dynamics of the device is adjusted very carefully and gently in such a way that over time a larger and larger part of the originally stored emittance becomes unstable. If it is possible to control the location around the ring where the spilling occurs, then the spilled particles can be direct towards the fixed target as needed. One storage where this approach is utilized is COSY at KFA in Jülich, Germany, which will be discussed in detail in a later lecture.

I. Summary and Comparison of Various Accelerators

After having discussed the various types of accelerators, it is useful to summarize their characteristic features side by side. The main physical criteria of performance are the energy that can be achieved, the (average) current, as well as the repetition frequency. For practical considerations, the size expressed in terms of a characteristic radius as well as the magnetic field in Tesla are also important. See table at the end of document.

IV. Linear Beam Theory

In the discussion of the basic physical principles of the above accelerators, we have casually neglected the fact that it is necessary to take care of more than one particle. In fact, all the above accelerators have to be able to simultaneously deal with an ensemble of particles with similar phase space coordinates, which is what the sources deliver, and hence with a **beam.** As outlined above, a detailed understanding of the motion of the beam requires the study of the motion of the **reference particle** as well as the motion of the **relative coordinates.**

In the case of **accelerators**, our demands on the relative motion are mostly that they beam does not become unreasonably large, and hence that the motion is somehow **bounded** within a suitable volume of phase space. While this appears to be a modest wish, for long single pass accelerators, and more so for repetitive systems, this problem actually turns out to be nontrivial. For other types of systems, more specific requirements have to be made for the beam; for example, to maximize the number of interactions at an interaction region of a collider, it is important to "squeeze" together the spatial coordinates of the beam, which under conservation of phase space volume then requires to make the momentum coordinates large. Devices like particle spectrographs or electron microscopes have different and usually more requirements yet.

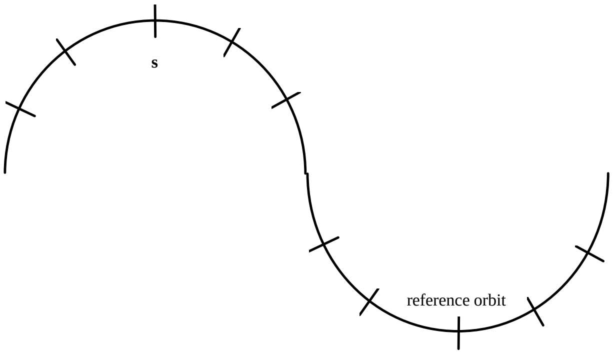
In all of these cases, it is important to study the relative motion carefully; as a first step, the motion is linearized, and for higher precision, the nonlinear effects of the motion have to be studied. Because the volume in phase space occupied by a beam is small, these nonlinear effects are often treated in a (later more precisely defined) **perturbative way**, in which the first order corresponds to linear motion, and nonlinear motion appears as higher order. Altogether, we have the following table:

0th order	motion of ref. particle
1st order	linear motion
2nd+ orders	nonlinear motion

A. Coordinates and Maps

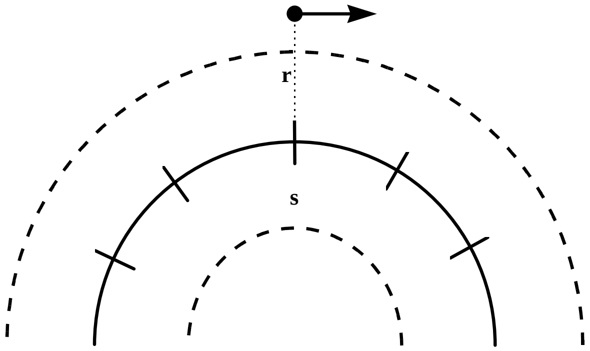
Usually when studying dynamics, the time t plays the role of the independent variable, and we study the motion of positions $\vec{x}$ and velocities $\vec{v}$ or momenta $\vec{p}$ as coordinates. Using the Lagrange mechanism, it is easy to transfer to new coordinates, in particular the coordinates that describe the **relative motion** around the reference orbit. Furthermore, instead of using t , we usually use the **arc length** s along the reference orbit as independent variable.

For the understanding of the motion in relative coordinates, let us assume we have studied and understood the motion of the reference orbit. In the case there is no field at all, this reference orbit will merely follow a **straight line**. Furthermore, there are a host of devices used in accelerators that have fields, but along one given straight line, all the fields vanish, and the device is lined up in such a way that the reference particle follows this line. Another important device uses magnetic fields, and along the reference orbit one tries to hold them constant, in which case the reference orbit is **circular**, at least within the element. In all other cases, it is usually necessary to **numerically integrate** the reference orbit.



The Reference Orbit

Assuming the position and momenta of the reference particle are $\vec{r}_{ref}(s), \vec{p}_{ref}(s)$ are known. As a technical detail, let us also assume that for all points s, we have $\vec{p}_{ref}(s) \not\parallel \vec{e}_z$, i.e. the motion is never pointing straight up (which for most real accelerators is no limitation whatsoever). Let furthermore ρ_t be smaller than the minimum radius of curvature that the reference orbit experiences in the section of the machine that we want to study. We now consider a "**flexible tube**" of radius ρ_t centered around the reference orbit, and restrict the particles that we want to describe to only those within the tube. Again, for practical devices this represents hardly a limitation; in the SSC, for example, the "tube" would be more than 20 km wide, much larger (hopefully) than the region required by the beam particles.



The Tube of Relative Coordinates

For any particle within the tube, there is now a **closest point** on the reference orbit; because only particles within the tube are allowed, this point is indeed **unique**. Let s be the arc length at this point, and $\vec{r}_{ref}(s)$ the position of the reference particle on the reference orbit. Then the relative coordinates of the point $\vec{r}$ are obviously $\vec{r} - \vec{r}_{ref}(s)$.

Let now $\vec{e}_s$ be a unit vector in the direction of $\vec{p}_{ref}$. Consider now the plane perpendicular to $\vec{e}_s$. Of all the unit vectors in this plane, let $\vec{e}_y$ be the one with the largest "upward" component; because $\vec{p}_{ref}$ and hence $\vec{e}_s$ are not allowed to go straight up, this vector is well defined. Finally choose a third vector $\vec{e}_x$ as $\vec{e}_x = \vec{e}_y \times \vec{e}_s$. Because $\vec{e}_y$ has maximum "upward" component, $\vec{e}_x$ has vanishing upward component and hence lies in the horizontal plane.

Denote now by $\mathbf{x}$ the component of $\vec{r}-\vec{r}_{ref}(s)$ in the direction of $\vec{e}_x$, and by $\mathbf{y}$ the component of $\vec{r}-\vec{r}_{ref}(s)$ in the direction of $\vec{e}_y$. Similarly, define p_x and p_y to be the momentum components in the directions $\vec{e}_x$ and $\vec{e}_y$.

Furthermore, denote by δ the relative difference between the total (kinetic plus potential) energy E of the particle under consideration and the reference energy E_0, i.e. $\delta = (E - E_0)/E_0$. Finally, introduce a space-like variable l to be the time of flight t minus the time of flight t_0 of the reference particle, multiplied by a constant k of dimension "velocity", i.e. $l = k\,(t - t_0)$. Then we form the **vector $\vec{Z}$ of particle optical coordinates** as

$$\vec{Z} = \left(\begin{array}{c} x \\ y \\ l = k\,(t-t_0) \\ a = p_x/p_0 \\ b = p_y/p_0 \\ \delta = (E-E_0)\,/E_0 \end{array} \right) \tag{25}$$

where p_0 is some previously chosen scaling momentum; a natural choice may be to select the momentum of the reference particle at the beginning.

Note that due to the definition of $\vec{Z}$, the reference particle itself corresponds to $\vec{Z} = 0$, and hence the vector $\vec{Z}$ does indeed describe the relative motion. In a seemingly simple way, most of the problems of beam physics now revolve around the question as to how $\vec{Z}$ evolves as a function of s.

The entire action of a beam physics device can now be expressed by how it manipulates the coordinates in the vector $\vec{Z}$. In fact, usually a set of initial conditions $\vec{Z}_0$ at position s_0 uniquely determines the future evolution and hence $\vec{Z}$ at any later position s. While a common notion, mathematically this **determinism** of classical mechanics rests on some subtle assumptions about the details of the fields that are allowed in the motion; but this course is not the place to be concerned about such issues. Assuming that indeed $\vec{Z}_0$ at s_0 uniquely determines the future evolution, we can define a function relating the initial conditions at s_0 to the conditions at s via

$$\vec{Z}\,(s) = \mathcal{M}\,(s_0, s)\left(\vec{Z}\,(s)\right) \tag{26}$$

The function $\mathcal{M}\,(s_0, s)$, which formally summarizes the entire action of the system, is of great importance for the description and analysis of beam physics systems. It is often called the **transfer function,** the **transfer map,** or simply the **map** of the system. Note that the transfer functions satisfy the relationship

$$\mathcal{M}\,(s_1, s_2) \circ \mathcal{M}\,(s_0, s_1) = \mathcal{M}\,(s_0, s_2)\,, \tag{27}$$

which merely says that transfer maps of systems can be built up from the transfer maps of the pieces.

Since $\mathcal{M}$ describes the motion in relative coordinates, we always have

$$\mathcal{M}(\vec{0}) = \vec{0}. \tag{28}$$

Furthermore, since by the very definition of a beam, the coordinates of $\vec{Z}$ are "small", $\mathcal{M}$ is usually only **weakly nonlinear;** because of this, its determination and analysis is very amenable to **perturbative techniques**. The very first step in this process is to consider only the linearization M of $\mathcal{M}$, the so-called **linear map**. Let $\mathcal{N} = \mathcal{M} - M$ be the remaining purely nonlinear part, so that we have

$$\mathcal{M} = M + \mathcal{N} \tag{29}$$

The linear map M is simultaneously the **most important** and the **easiest** to study, and a great deal of these lectures is related to it. The treatment of the nonlinear part $\mathcal{N}$ is much more complicated, and only later in the course will we address a small part of the problematic associated with its treatment.

In the next section, we will make a short excursion to a field that is at first sight disconnected from beam physics, namely the field of glass optics. However, a more close look shows that glass optics, which has existed long before the name beam physics has been introduced, certainly belongs to this field: the ensembles of light particles or rays typically associated with questions of glass optics form a beam not only in the conventional meaning of the word, but also under our stronger more formal definition.

B. Glass Optics

As one may recall from a basic course in optics, a distinction is made between so-called "**Gaussian optics**", which indeed turns out to just mean linear motion, and "**aberrations**" that describe nonlinear effects. Optics has developed its very own jargons and techniques, some of which are connected to complicated geometric ideas, and it is historically unfortunate that optics has not been treated with the methods of the **transfer map.** We shall remedy this situation here by simultaneously providing a short course on Gaussian optics in an appealing and unified way, and also develop our skills in dealing with linear maps.

For simplicity, let us restrict ourselves to systems that are rotationally symmetric, like most glass optical systems; it will be quite clear as we go what has to be done to treat non-rotationally symmetric systems. In this rotationally symmetric case, two variables are enough to study the motion; we here choose them as the position x and the slope m of a ray. The transfer map of an optical system than expresses how (x, m) behave as they transfer a system, and we have

$$\begin{pmatrix} x_2 \\ m_2 \end{pmatrix} = \mathcal{M} \begin{pmatrix} x_1 \\ m_1 \end{pmatrix}. \tag{30}$$

In fact, if we restrict ourselves to linear motion, then this can be expressed in terms of a **transfer matrix**

$$M = \begin{pmatrix} (x,x) & (x,m) \\ (m,x) & (m,m) \end{pmatrix}. \tag{31}$$

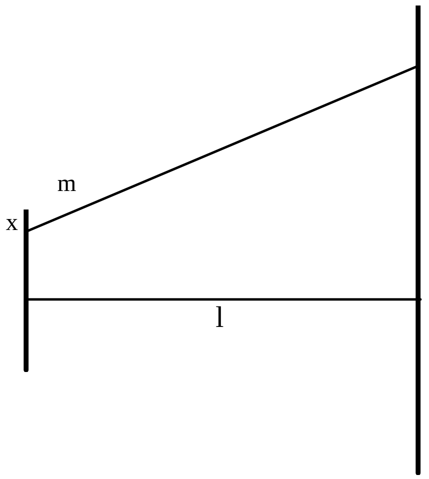
Note that the notation for the matrix elements is such that the quantity **before** the comma describes the **row,** and that **after** the comma describes the column. We remind again that knowing matrices of pieces allows the computation of matrices of more complicated systems, which is here achieved by mere matrix multiplication. Indeed, if M_1 through M_n are the matrices for the subsystems, then because of the associativity of matrix multiplication, we obtain for the ray after the last subsystem:

$$\begin{aligned} \begin{pmatrix} x_{n+1} \\ m_{n+1} \end{pmatrix} &= M_n \left(\cdots \left(M_1 \begin{pmatrix} x_1 \\ m_1 \end{pmatrix} \right) \right) \cdots \tag{32} \\ &= (M_n \cdot \cdots \cdot M_1) \begin{pmatrix} x_1 \\ m_1 \end{pmatrix} \tag{33} \end{aligned}$$

So we have shown that the matrix of a combined system equals to product of matrices of subsystems. Since especially on computers it is very simple to multiply matrices, this is the method of choice for the basic design of optical systems. In the following, we hence derive the forms of the matrices of common optical elements.

1. The Drift

The simplest part of glass optical elements is a region which doesn't contain any material, the drift. If we denote by x the position of a ray and by m its slope, then the final values x_2 and m_2 after a drift of length l can be connected very simply to the initial values x_1 and m_1:



$$x_2 = x_1 + m_1 \cdot l \tag{34}$$
$$m_2 = m_1 \tag{35}$$

This can obviously be written in a matrix form:

$$\begin{pmatrix} x_2 \\ m_2 \end{pmatrix} = \begin{pmatrix} 1 & l \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ m_1 \end{pmatrix} \tag{36}$$

For the later discussion it will prove important to note that the matrix $\begin{pmatrix} 1 & l \\ 0 & 1 \end{pmatrix}$ depends only on the characteristic properties of the element, which here is the length l. On the other hand, the vector $\begin{pmatrix} x_1 \\ m_1 \end{pmatrix}$ depends only on the parameters of the ray. Altogether, a drift performs a linear transformation in x, m space. Note that the determinant of the drift matrix is unity.

As a small exercise, let us now consider a combination of two drifts of lengths $l_1$and l_2, and let us ask ourselves for the value of the coordinates (x_3, m_3) after the combination of the two drifts. We obviously have

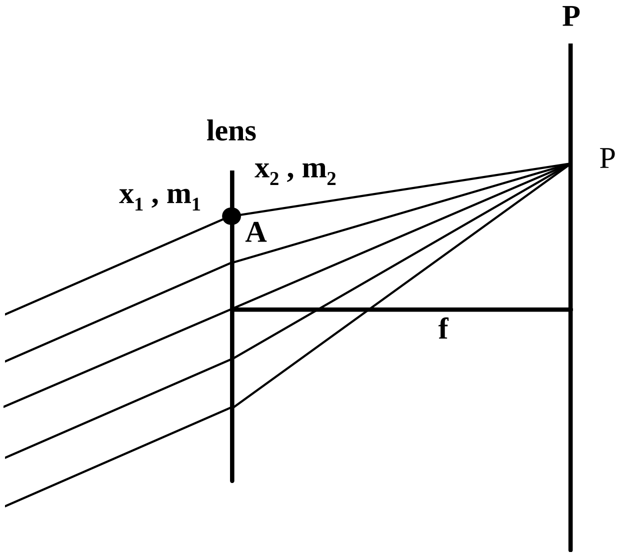
$$\begin{aligned} \begin{pmatrix} x_3 \\ m_3 \end{pmatrix} &= \begin{pmatrix} 1 & l_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_2 \\ m_2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & l_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & l_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ m_1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & l_1 + l_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ m_1 \end{pmatrix} \qquad (37) \end{aligned}$$

Here the necessary **composition** of maps just reduces to a common **multiplication** of transfer matrices. And the result is not surprising, the effect of two subsequent drifts is just the same as that of a drift of the combined length.

2. The Thin Lens

Besides empty space, glass optical devices contain lenses that change the direction of the light ray. We are here primarily interested in the thin lens, a somewhat idealized device without any length, which is characterized by the following facts that are also illustrated in the picture below:

1. Positions are not changed, but directions are
2. Any bundle of parallel light is unified in one point a distance f after the lens
3. A ray lighting the center of the lens goes straight through



The quantity f that describes the lens is called the focal length. Let us now consider a ray going through the lens; from the picture we read

$$\begin{aligned} x_2 &= x_1 && (38)\\ p &= f \cdot m_1 && (39)\\ x_1 + m_2 \cdot f &= p && (40) \end{aligned}$$

From which we infer

$$\begin{aligned} x_2 &= x_1 && (41)\\ m_2 &= -\frac{x_1}{f} + m_1 && (42) \end{aligned}$$

This relationship can again be written in matrix form:

$$\begin{pmatrix} x_2 \\ m_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ m_1 \end{pmatrix} \tag{43}$$

As in the case of the drift, the matrix $\begin{pmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{pmatrix}$ depends only on the lens, whereas the vector $\begin{pmatrix} x_1 \\ m_1 \end{pmatrix}$ depends on the ray.

The simple thin lens we have discussed here, the so-called "Gaussian" lens, represents quite an approximation for several reasons. First of all, any real lens performs a refraction at two different surfaces, so positions do change as one goes through the lens. Furthermore, for most lenses it is not really true that parallel rays all meet at a point a distance f behind the lens. This is connected to the fact that lenses are usually ground with spherical surfaces because anything else is technically difficult. Furthermore, the glass has dispersion and so different colors are affected differently. We note however that Snell's law still allows to determine the true transfer map of a thick, spherical lens in a rather straightforward way. It is important to note, however, that this transfer map will no longer be linear.

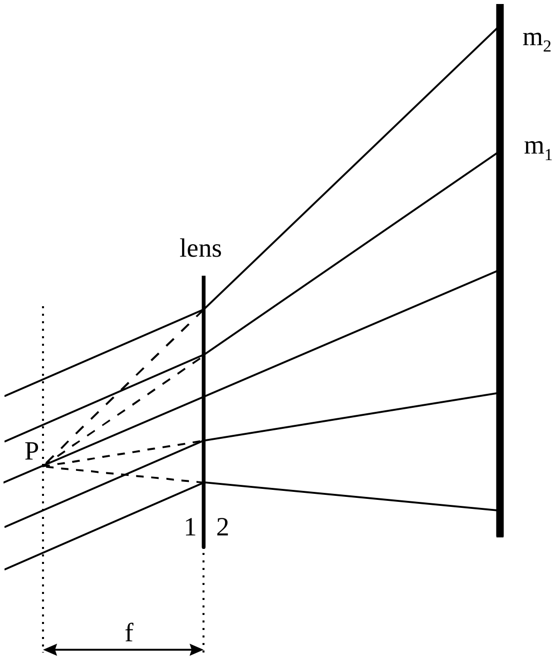
Quite interesting is the combination of two glass lenses, which can apparently be described by multiplying their matrices (note that always, the matrix of the **first element** is on the **right**). We obtain

$$\begin{pmatrix} 1 & 0 \\ -\frac{1}{f_2} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\frac{1}{f_1} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{1}{f_1} - \frac{1}{f_2} & 1 \end{pmatrix} \tag{44}$$

So the combination of two lenses provides the same effect as one lens with focus length f, where $1/f = 1/f_1 + 1/f_2$. This is of course a famous law of optics, the derivation of which is trivial is all but trivial in the matrix concept. Some of the power of the matrix approach becomes clear how powerful it is to prove this law using the standard geometric method of optics text books.

In a similar way as the focusing thin lens we can also treat the defocusing thin lens. In this case, the basic properties are

1. Positions are not changed, but directions are
2. Any bundle of parallel light exits the lens in such a way that it appears to come from a point a distance f in front of the lens
3. A ray lighting the center of the lens goes straight through



In a similar way as before, we can use basic geometry to determine the action of the lens; we read from the picture

$$
\begin{aligned}
x_2 &= x_1 & (45)\\
p &= -f m_1 & (46)\\
p &= x_2 - f m_2 & (47)
\end{aligned}
$$

Similar to before, we get $m_2 = x_2/f + m_1$,and altogether in matrix form

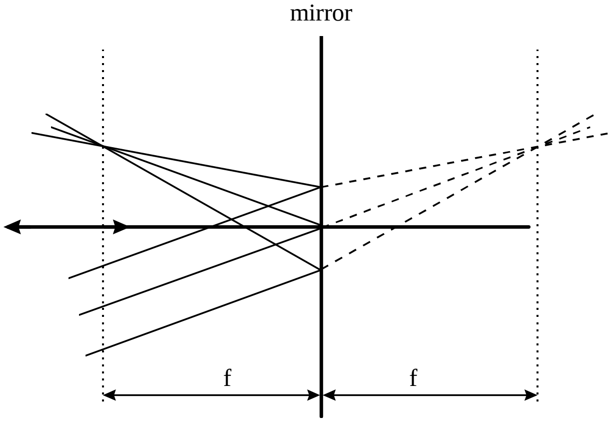
$$\begin{pmatrix} x_2 \\ m_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{1}{f} & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ m_1 \end{pmatrix}. \qquad (48)$$

This is essentially the same matrix as before, except that now the sign of the matrix element (m, x) has changed. Indeed, using the standard convention to count defocusing lenses with a negative focal length, the matrix has even exactly the same form as before.

3. The Thin Mirror

Besides lenses, mirrors are probably the second most important optical device, and similar to before, there are focusing and defocusing mirrors. Different from the lens, in the case of the mirror the reference orbit flips direction when hitting the mirror. A thin focusing mirror is defined by what it does to an ensemble of parallel light via the three conditions

1. Positions are not changed, but directions are
2. Any bundle of parallel light that is reflected by the mirror will meet in a point a distance f in front of the mirror
3. A ray hitting the center of the mirror is reflected such that its outgoing angle equals its incoming angle



A similar argument as in the case of the focusing lens shows that the transfer matrix of the focusing mirror is

$$M = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \qquad (49)$$

There is also a defocusing mirror, defined by the three conditions

1. Positions are not changed, but directions are
2. Any bundle of parallel light that is reflected by the mirror seems to emerge from a point a distance f behind the mirror

3. A ray hitting the center of the mirror is reflected such that its outgoing angle equals its incoming angle

A similar argument to before shows that also in this case, we have the transfer matrix

$$M = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix}, \tag{50}$$

where the convention to count the focal length f of a defocusing element negative is used.

So apparently mathematically, lenses and mirrors behave the same, aside from the fact that they reverse the reference orbit. The choice which to use in practice depends on a variety of practical factors. For situations requiring only small apertures like in most camera lenses, glass lenses are easily made, and have an advantage because of the straight beam path. For situations requiring large apertures, like in big telescopes, mirrors are the primary choice because it is much easier to manufacture and support large mirrors than large lenses. It is also easier to produce non-spherical shapes for mirrors than for lenses. Finally, mirrors have the additional advantage that they treat light of different colors equally, they don't show the dispersion commonly observed in glass lenses.

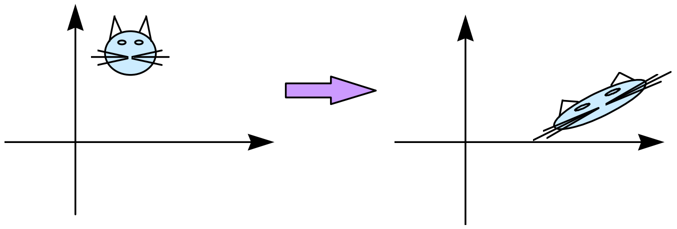
4. Liouville's Theorem for Glass Optics

As a direct consequence of the matrix notation for glass optics introduced above, for any combination of lenses, drifts, and mirrors, we can prove a special case of **Liouville's Theorem:** The volume of phase space occupied by the beam is conserved.

Indeed, let us assume we have an optical system consisting of n elements with matrices M_i. Then we have

$$\begin{aligned} \begin{pmatrix} x_{n+1} \\ a_{n+1} \end{pmatrix} &= M_n \left(M_{n-1} \left(\cdots M_1 \begin{pmatrix} x_1 \\ a_1 \end{pmatrix} \cdots \right)\right) \\ &= (M_n \cdot M_{n-1} \cdots M_1) \begin{pmatrix} x_1 \\ a_1 \end{pmatrix}; \end{aligned} \tag{51}$$

but the determinants of each of the matrices M_i are just unity, as they are all either drifts, lenses or mirrors, and so the determinant of the product is unity. But since under linear transformations, volumes in space transform with the size of the determinant, the volume is indeed conserved.

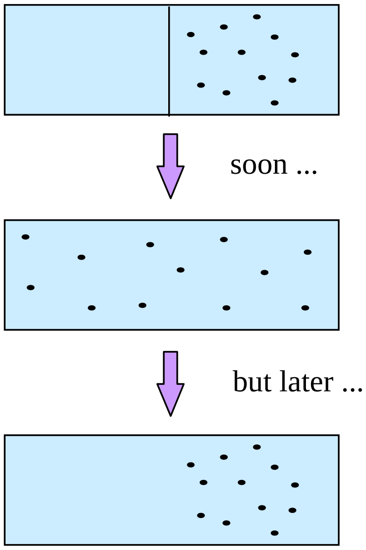


Liouville's Theorem

An interesting and remarkable consequence of Liouville's Theorem is the famous repetition theorem of Poincare. Let us assume we have some motion in n-dimensional phase space, and let us assume that we know that the motion is bounded. Let us further assume that the motion obeys Liouville's Theorem, which as we shall later see is the case for all Hamiltonian systems, and let the motion be deterministic. Then Poincare's repetition theorem states that for any given ε, the system after sufficient time comes back to its original state within a tolerance of at most ε.

Before we illustrate the proof of Poincare's theorem, let us illustrate some of its consequences. Consider for example a box with classical gas particles that are initially all located in one side of the box and kept there by a wall. After the wall is removed, the gas particles will distribute in the box evenly, as we expect from classical statistical mechanics, increasing their entropy. But their phase space is bounded: the positions cannot leave the box, and each particle's momentum is limited by the total heat energy contained in the box.

But as time progresses, according to Poincare, they will at one time in the future just recollect on one side of the box; and by reinserting the wall, they will be caught again on one side, in crass contradiction to the entropy principle.



Poincare's Repetition Theorem

Other examples abound: if we have a particle beam in an accelerator that we know is stable, it will eventually come back as close as we want in phase space - an effect that is actually observed somewhat routinely in tracking pictures. Even for our daily life, there are important consequences: if the universe is Hamiltonian and doesn't expand indefinitely, then up to minute details, history will keep repeating itself; we will all be born again, and we will all make the same mistakes all over; but since now we can't remember anything about our past life, also next time we won't remember our current life ...

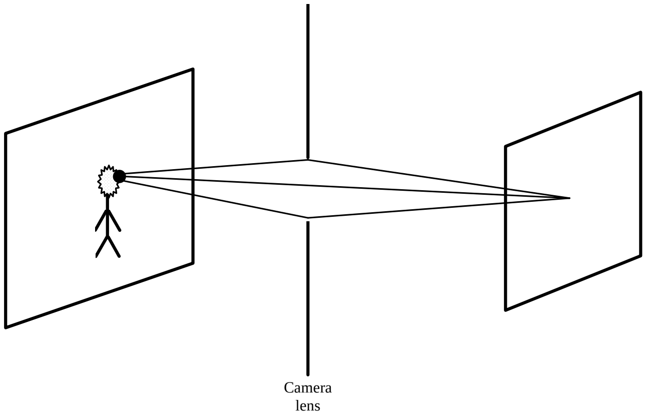
Now for the sketch of the proof of the repetition theorem: Let an ε be given, and consider an ε-ball with volume $V\varepsilon$ in phase space. Consider its motion by regular time steps Δt. Since the total available phase space volume is finite, say V_p, after at most $V_p/V\varepsilon$ time steps, the image of the ball must reach a part of phase space it has touched before, i.e. it must overlap a previous image of the ball. Let us assume this happens after N steps, and let us assume that the previous image is that after n steps, with $n < N$. But if the images after n steps I_n and after N steps I_N overlap, so must the images after $(n-1)$ and $(N-1)$ steps, respectively; and continuing backwards, so must the images after 0 and $(N-n)$ steps; hence after $(N-n)$ steps, we touch the original ε-ball again.

C. Special Optical Systems

In this section we want to apply the matrix techniques to the study of certain special categories of systems. In particular, we associate certain fundamental properties of systems with properties of the matrix. We begin with the imaging systems.

1. Imaging (Point-to-Point, $\cdot\,\cdot$) Systems

Imaging systems are perhaps the most important systems in optics, and they deserve some special attention. Suppose we study the action of a slide projector. At one end of the projector, light is sent through the slide. Suppose the slide shows a tree in the fall with one last green leaf. The image of this tree is to appear on the screen, and the green leaf is to appear at one particular location. This requires that all light going through the green spot on the slide in various directions has to be re-united at one spot on the screen.



This means that the final position of a ray is independent of its initial angle and only depends on the initial position. In terms of transfer matrices

$$M = \begin{pmatrix} (x \mid x) & (x \mid m) \\ (m \mid x) & (m \mid m) \end{pmatrix} \tag{52}$$

this means that the element $(x \mid m)$ has to vanish. Obviously the element (x, x) also has an important interpretation: it is the magnification of the system.

Besides the case of the slide projector, many other devices use imaging. They include the camera, the overhead projector, the eye, the photographic microscope, the electron microscope, as well as particle spectrographs.

It is worthwhile to study how imaging systems can be made. First of all we observe that a drift is imaging if and only if $l = 0$, a rather boring choice. A single lens is also always imaging as long as there are no drifts before and after, but that is another boring choice. The first interesting imaging system is the DLD system, consisting of a drift, a lens, and another drift. The transfer matrix of the DLD system is given by

$$\begin{aligned} M &= \begin{pmatrix} 1 & l_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{pmatrix} \begin{pmatrix} 1 & l_1 \\ 0 & 1 \end{pmatrix} && (53) \\ &= \begin{pmatrix} 1 & l_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & l_1 \\ -\frac{1}{f} & 1 - \frac{l_1}{f} \end{pmatrix} && (54) \\ &= \begin{pmatrix} 1 - \frac{l_2}{f} & l_1 + l_2 - \frac{l_1 l_2}{f} \\ -\frac{1}{f} & 1 - \frac{l_1}{f} \end{pmatrix} && (55) \end{aligned}$$

If such a system is supposed to be imaging, we have to satisfy $(x, m) = 0$, or $l_1 + l_2 - l_1 l_2 / f = 0$, which is equivalent to

$$\frac{1}{l_1} + \frac{1}{l_2} = \frac{1}{f} \tag{56}$$

This is another important result of conventional optics, which here is obtained in an almost trivial way. If the DLD system is actually imaging. In this case, the magnification is given by (x, x) and hence has the value $1 - l_2/f = -l_2/l_1$.

This principle is used in several different devices. In the slide projector, l_1 is very small and l_2 is very large, providing a large magnification. Probably the most important imaging system is the eye. Here the situation is just the opposite: l_1 is large and l_2 is small, allowing for large things to be mapped on the small retina of the eye.

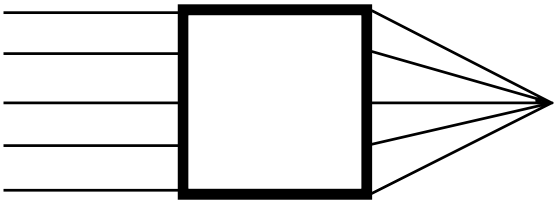
It is interesting to study the combination of two imaging systems:

$$\begin{aligned} & \begin{pmatrix} (x \mid x)_2 & 0 \\ (m \mid x)_2 & (m \mid m)_2 \end{pmatrix} \\ & \cdot \begin{pmatrix} (x \mid x)_1 & 0 \\ (m \mid x)_1 & (m \mid m)_1 \end{pmatrix} \\ = & \begin{pmatrix} (x \mid x)_2 (x \mid x)_2 & 0 \\ (m \mid x)_2 (x \mid x)_1 + (m \mid m)_2 (m \mid x)_1 & (m \mid m)_2 (m \mid m)_1 \end{pmatrix} \end{aligned} \tag{57}$$

As is to be expected, the total system is again imaging, and the magnification is just the product of the individual magnifications.

2. Parallel-to-Point ($\parallel \cdot$) Systems

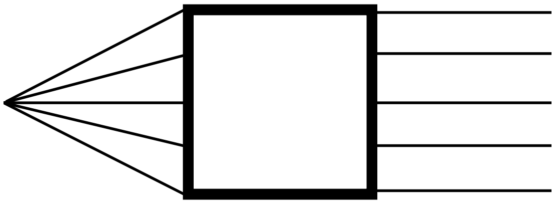
As we saw above, the human eye observing a nearby object is one of the prime examples of an imaging system. But what happens if the eye looks at things farther and farther away, in particular at the stars, a pastime of the human race and scientists for eternity? The length of the first drift l_1 becomes larger and larger, and for all practical purposes the light coming from one star reaches the eye as a parallel bundle. So what the eye is to interpret now is the angle under which the light comes in, and hence the position on the retina should depend only on the initial angle, but not on the initial position at which the light strikes the eye.



This requires that $(x|x) = 0$. If we look at the eye as a DLD system, this requires $1 - l_2/f = 0 \Leftrightarrow l_2 = f$, l_1 arbitrary. Thus the retina has to be exactly at the focal length; almost as important is that the distance to the object is arbitrary since we cannot change our distance to the stars significantly. Another important parallel-to-point system is the photographic camera.

3. Point-to-Parallel ($\cdot \,\|$) Systems

Another important class of systems is the point to parallel systems. In point to parallel systems, the final slope does depend only on the initial position, but not on the initial slope. So we have $(m|m) = 0$.



Examples include the flashlight, the microscope, and the SDI Transport over long distances. As an example, let us try to achieve a point-to-parallel system with a DLD combination. We obtain

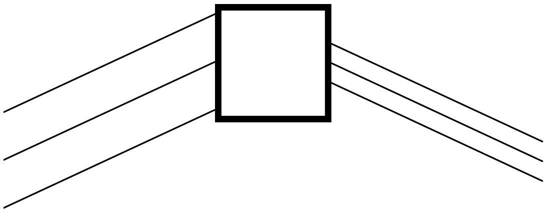
$$1 - l_1/f = 0 \text{ or } l_1 = f, \tag{58}$$

as we may have expected. Note there is no condition on l_2,which is fairly important for the operation of a microscope as it allows the observer to move his eyes.

From the transfer matrices, it follows rather directly that the combination of a point to parallel and a parallel to point system forms a point to point system. Using the relaxed eye as the parallel to point system, we can thus build a microscope by putting a suitable point to parallel system in front of the eye. It is interesting to see how the lengths in a point to parallel system have to be chosen; we obtain $(m|m) = 0 \Leftrightarrow 1 - l_1/f = 0 \Leftrightarrow l_1 = f$, l_2 arbitrary. The first part is as expected; the latter part is helpful because it allows the eye to move with respect to the microscope.

4. Parallel-to-Parallel ($\| \, \|$) Systems

The final important system is the parallel to parallel system. By putting it between the eye and the stars, a magnification of angles can be achieved. This is the principle of the telescope.



The system has to be such that the final slope depends on the initial slope, but not on the initial x, which requires $(m|x) = 0$. The magnification is given by (m,m). If we try to achieve this with a DLD system, then we have to satisfy $-1/f = 0$, which is impossible. This entails that a telescope has to contain at least two lenses.

So let us consider an LDL system; we have

$$\begin{aligned} M &= \begin{pmatrix} 1 & 0 \\ -\frac{1}{f_2} & 1 \end{pmatrix} \begin{pmatrix} 1 & l \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\frac{1}{f_1} & 1 \end{pmatrix} && (59) \\ &= \begin{pmatrix} 1 - \frac{l}{f_1} & l \\ -\frac{1}{f_2} - \frac{1}{f_1} + \frac{l}{f_1 f_2} & 1 - \frac{l}{f_2} \end{pmatrix} && (60) \end{aligned}$$

and thus we have to satisfy

$$-\frac{1}{f_2} - \frac{1}{f_1} + \frac{l}{f_1 f_2} = 0 \text{ or } l = f_1 + f_2, \tag{61}$$

which is actually a well-known condition for Newtonian or Galilean telescopes. The magnification of the telescope is given by

$$(m, m) = 1 - \frac{l}{f_2} = 1 - \frac{1}{f_2}\left(f_1 + f_2\right) = -\frac{f_1}{f_2};$$

thus to obtain large magnification requires $f_1 \gg f_2$. Since there is a limit on how short f_2 can be, it is thus necessary to make f_1 large, which entails the rather large size telescopes usually have.

5. Combination Systems

Often the question arises to what extent it is possible to simultaneously satisfy the requirements for the above systems. To some extent this is possible, but the fact that the determinant of the total system has to be unity imposes some restrictions. A closer look shows that

1. $\bullet\bullet$ and $\|\|$ is possible : $(x \mid m) = (m \mid x) = 0$
2. $\|\bullet$ and $\bullet\|$ is possible : $(x \mid x) = (m \mid m) = 0$

All other cases are impossible because they would require a zero determinant.

Another important question is what happens when two systems satisfying certain properties are combined into one system; for example, we already saw in eq. (57) that two point-to-point systems placed behind each other again produce a point-to-point system. A more detailed analysis shows that of the sixteen cases describing combinations of two systems, eight lead to another special system

$$\begin{array}{ccccccccccc}
\bullet\bullet & + & \bullet\bullet & = & \bullet\bullet & , & \|\,\| & + & \|\,\| & = & \|\,\| \\
\bullet\bullet & + & \bullet\,\| & = & \bullet\,\| & , & \|\,\| & + & \|\,\bullet & = & \|\,\bullet \\
\bullet\,\| & + & \|\,\bullet & = & \bullet\bullet & , & \|\,\bullet & + & \bullet\,\| & = & \|\,\| \\
\bullet\,\| & + & \|\,\| & = & \bullet\,\| & , & \|\,\bullet & + & \bullet\bullet & = & \|\,\bullet
\end{array} \tag{62}$$

The entries in the table are easy to memorize because it contains just those combinations for which the second symbol of the first system equals the first symbol of the second system, and the final result is obtained by "dropping" the two identical symbols. So in compact notation, we have:

$$\begin{aligned} \text{If } A, B, C &\in \{\bullet,\ \|\},\ \text{then} \\ AB + BC &= AC. \end{aligned} \tag{63}$$

V. Fields and Potentials

For the study of transfer maps of particle optical systems, it is first necessary to undertake a classification of the possible fields that can occur. All fields are governed by **Maxwell's equations**, which in SI units have the form

$$\operatorname{div} \vec{B} = 0 \tag{64}$$

$$\operatorname{curl} \vec{H} = \vec{j} + \frac{\partial \vec{D}}{\partial t} \tag{65}$$

$$\operatorname{div} \vec{D} = \rho \tag{66}$$

$$\operatorname{curl} \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{67}$$

In the case of particle optics, we are mostly interested in cases in which there are **no sources** of the fields in the region where the beam is located, and so in this region we have $\rho = 0$ and $\vec{j} = \vec{0}$. Of course any beam that is present would represent a ρ and a $\vec{j}$, but these effects are usually considered separately.

In the following, we want to restrict ourselves to **time-independent** situations, and neglect the treatment of elements with quickly varying fields including cavities. This limitation in very good approximation also includes slowly time-varying fields like the magnetic fields that are increased during the ramping of a synchrotron. In this case, Maxwell's equations simplify to

$$\begin{aligned} \operatorname{div} \vec{B} &= 0\ ,\ \operatorname{curl} \vec{H} = \vec{0} \\ \operatorname{div} \vec{D} &= 0\ ,\ \operatorname{curl} \vec{E} = \vec{0} \end{aligned} \tag{68}$$

where $\vec{B} = \mu_0 H$ and $\vec{D} = \varepsilon_0 \vec{E}$. Because of the vanishing curl, we infer that $\vec{B}$ and $\vec{E}$ have

scalar potentials V_E **and** V_B such that

$$\vec{E} = -\vec{\nabla} V_E \text{ and } \vec{B} = -\vec{\nabla} V_B \tag{69}$$

Note that here even the magnetic field is described by a scalar potential, and not by the vector potential $\vec{A}$ that always exist. From the first and third equations, we infer that both scalar potentials V_E and V_B satisfy Laplace's equation, and we thus have

$$\Delta V_{E,B} = 0. \tag{70}$$

In order to study the solutions of Laplace's equations for the electric and magnetic scalar potentials, we will proceed for two special cases, each of which will be treated in a coordinate system most suitable for the problem.

A. Fields with Straight Reference Orbit

The first major case of systems is those that have a straight reference orbit. In this case, there is no need to distinguish between particle optical coordinates and Cartesian coordinates, and in particular there is no need to transform Laplace's equation to a new set of coordinates. Many elements with a straight reference orbit possess a certain rotational symmetry around the axis of the reference orbit, and it is most advantageous to describe the potential in **cylindrical coordinates** with a "z axis" that coincides with the reference orbit. We first begin by expanding the r and ϕ components of the potential in Taylor- and Fourier series, respectively; the dependence on the cylindrical "z" coordinate, which here coincides with the particle optical coordinate s, is not expanded. So we have

$$V = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} M_{k,l}\left(s\right)\cos\left(l\phi + \theta_{k,l}\right) r^k \tag{71}$$

In cylindrical coordinates, the Laplacian has the form

$$\Delta V = \frac{1}{r}\frac{\partial}{\partial r}\left(\frac{r\partial V}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 V}{\partial \phi^2} + \frac{\partial^2 V}{\partial s^2} = 0; \tag{72}$$

inserting the Fourier-Taylor expansion of the potential, we obtain

$$\begin{aligned}
\Delta V &= \frac{1}{r}\frac{\partial}{\partial r}\left(\frac{r\partial V}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 V}{\partial \phi^2} + \frac{\partial^2 V}{\partial s^2} = \\
&= \frac{1}{r}\frac{\partial}{\partial r}\left\{\sum_{k=1}^{\infty}\sum_{l=0}^{\infty} M_{k,l}\cos\left(l\phi+\theta_{k,l}\right) k r^k\right\} \\
&\quad + \frac{1}{r^2}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} M_{k,l}\cos\left(l\phi+\theta_{k,l}\right)\left(-l^2\right) r^k \\
&\quad + \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} M_{k,l}^{''}\left(s\right)\cos\left(l\phi+\theta_{k,l}\right) r^k \\
&= \sum_{k=1}^{\infty}\sum_{l=0}^{\infty} M_{k,l}\cos\left(l\phi+\theta_{k,l}\right) k^2 r^{k-2} \\
&\quad - \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} M_{k,l}\cos\left(l\phi+\theta_{k,l}\right) l^2 r^{k-2} + \\
&\quad + \sum_{k=2}^{\infty}\sum_{l=0}^{\infty} M_{k-2,l}^{''}\left(s\right)\cos\left(l\phi+\theta_{k-2,l}\right) r^{k-2}
\end{aligned}$$

Now we note that in the first term, it is possible to let the sum start at $k = 0$ since there is no contribution anyway because of the factor k^2. Furthermore, using the convention that the coefficient $M_{k,l}(s)$ vanish for negative indices, we obtain

$$\Delta V = \sum_{k,l=0}^{\infty}\left\{\begin{array}{c} M_{k,l}\left(s\right)\cos\left(l\phi+\theta_{k,l}\right)\left(k^2-l^2\right) \\ +M_{k-2,l}^{''}\left(s\right)\cos\left(l\phi+\theta_{k-2,l}\right)\end{array}\right\} r^{k-2} \tag{73}$$

We begin the analysis by studying the case $k = 0$. Apparently $M_{0,0}$ and $\theta_{0,0}$ can be chosen freely because the factor $(k^2 - l^2)$ vanishes. Furthermore, since $M_{k-2,l}^{''}(s)$ vanishes for all l because of the convention regarding negative indices, we infer $M_{0,l} = 0$ for $l \geq 1$.

By induction over k, we now show that $M_{k,l}(s) \equiv 0$ for all cases where $k < l$. Apparently the statement is true for $k = 0$. Now let us assume that the statement is true up to $k-1$. If $k < l$, also $k - 2 < l$, and thus $M''_{k-2,l}(s) = 0$. Since $(k^2 - l^2) \neq 0$ and $\cos(l\phi + \theta_{k,l}) \neq 0$ for some ϕ because $l \neq 0$, this requires $M_{k,l}(s) \equiv 0$ for $k < l$. Thus the infinite matrix $M_{k,l}$ is strictly lower triangular.

We now study the situation for different values of l. We first notice that for all l, the choices of

$$M_{l,l}(s) \text{ and } \theta_{l,l} \text{ are free} \tag{74}$$

because $M''_{l-2,l}(s) = 0$ by the previous observation, and $(k^2 - l^2) = 0$ because $k = l$. Next we observe that the value $M_{l+1,l}(s)$ must vanish, because $(k^2 - l^2) \neq 0$, but $M''_{l-1,l}(s) \equiv 0$ because of the lower triangularity. Recursively we even obtain that

$$M_{l+1,l}(s), M_{l+3,l}(s), \ldots \text{ vanish.} \tag{75}$$

On the other hand, for $k = l + 2$, we obtain that $\theta_{l+2,l} = \theta_{l,l}$, and $M_{l+2,l}(s)$ is uniquely specified by $M_{l,l}(s)$. Applying recursion, we see that in general

$$\begin{aligned} \theta_{l,l} &= \theta_{l+2,l}, \theta_{l+4,l}, \ldots \text{ and} \\ M_{l+2n,l}(s) &= \frac{M_{l,l}^{(2n)}(s)}{\prod_{\nu=1}^{n}\left((l)^2 - (l+2\nu)^2\right)}. \end{aligned} \tag{76}$$

Let us now proceed with the physical interpretation of the result. The number l is called the **multipole order**, as it describes how many oscillations the field will experience in one 2π sweep of ϕ. The free term $M_{l,l}(s)$ is called the **multipole strength**, and the term $\theta_{l,l}$ is called the **multipole phase.** Apparently, **frequency l and radial power k are coupled:** The lowest order in r that appears is l, and if the multipole strength is s-dependent, also the powers $l + 2$, $l + 4$, ... will appear.

For a multipole of order l, the potential has a total of $2l$ maxima and minima, and is so often called a **2l-pole.** Often Latin names are used for the $2l$ poles, and we have the following table:

l	**Leading Term in V**	**Name**
0	$M_{0,0}(s)\cos(\theta_{0,0})$	
1	$M_{1,1}(s)\cos(\phi + \theta_{1,1})\, r$	Dipole
2	$M_{2,2}(s)\cos(2\phi + \theta_{2,2})\, r^2$	Quadrupole
3	$M_{3,3}(s)\cos(3\phi + \theta_{3,3})\, r^3$	Sextupole
4	$M_{4,4}(s)\cos(4\phi + \theta_{4,4})\, r^4$	Octupole
5	$M_{5,5}(s)\cos(5\phi + \theta_{5,5})\, r^5$	Decapole

In many cases it is very important to study the Cartesian (and hence also particle optical) form of the fields of the elements. The case $k = 1$ with $V = M_{1,1}\cos(\phi + \theta_{1,1})\, r$ is quite trivial; for $\theta_{22} = 0$,we obtain $V = M_{1,1} \cdot x$, corresponding to a uniform field in x-direction, and for another important sub-case $\theta_{22} = -\pi/2$, we obtain $V = M_{1,1} \cdot y$, a uniform field in y-direction. In both of these cases, the reference orbit is indeed a straight line only in the limit of weak fields.

The case $k = 2$ has $V = M_{2,2}\cos(2\phi + \theta_{22})\, r^2$. Particularly important in practice will be the sub-cases $\theta_{22} = 0$ and $\theta_{22} = -\pi/2$. In the first case, we get

$$\begin{aligned} V &= M_{2,2}\cos(2\phi)\, r^2 \\ &= M_{2,2}\left(\cos^2\phi - \sin^2\phi\right) r^2 \\ &= M_{2,2}\left(x^2 - y^2\right), \end{aligned} \tag{77}$$

and in the second case we have

$$\begin{aligned} V &= M_{2,2}\cos(2\phi - \pi/2)\, r^2 \\ &= M_{2,2}\sin(2\phi)\, r^2 \qquad (78) \\ &= M_{2,2}(2\sin\phi\cos\phi)\, r^2 \\ &= M_{2,2}(2xy) \qquad (79) \end{aligned}$$

All other angles θ_{22} lead to formulas that are more complicated; they can be obtained from the ones here by subjecting the x, y coordinates to a suitable rotation. This again leads to terms of purely second order.

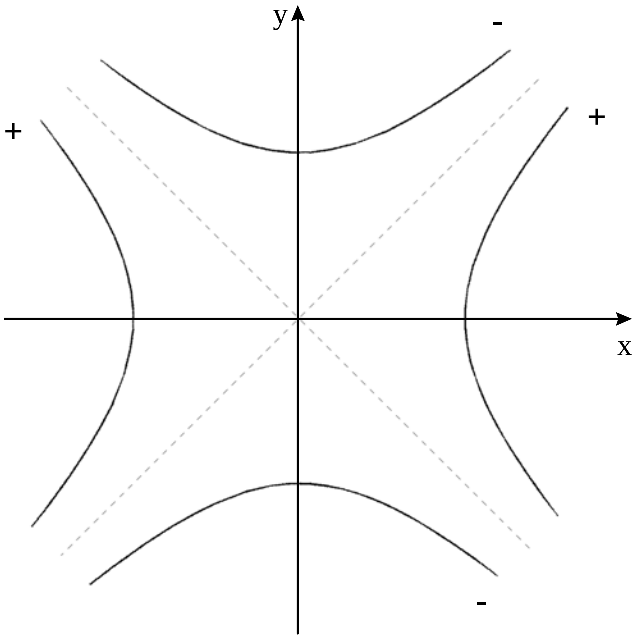
Because the potential is quadratic, the resulting fields $\vec{E}$ or $\vec{B}$ are **linear.** Indeed, the quadrupole is the **only s-independent element that**

leads to linear motion similar to that in glass optics, and thus has great importance. In the electric case, one usually chooses $\theta_{2,2} = 0$, resulting in the fields

$$
\begin{aligned}
E_x &= -2M_{2,2} \cdot x && (80)\\
E_y &= 2M_{2,2} \cdot y. && (81)
\end{aligned}
$$

So different from the case of glass optics, it turns out that the motion **cannot be rotationally symmetric** anymore: if there is focusing in the x-direction, there is defocusing in the y-direction, and vice versa! This effect, completely due to Maxwell's equations, turns out to be perhaps the biggest "nuisance" in beam physics: if one uses piecewise s-independent particle optical elements, the horizontal and vertical **planes are always different from each other.**

To make an electrostatic device that produces a quadrupole field, it is best to carve the electrodes along the equipotential surfaces, and utilize the fact that if "enough" boundary conditions are specified, the field is uniquely determined, and must hence be as specified by the formula used to determine the equipotential surfaces in the first place. So in practice, the electrodes of an electric quadrupole often look as shown in the picture.



In the magnetic case one indeed chooses $\theta_{2,2} = -\pi/2$, resulting in

$$
\begin{aligned}
B_x &= -2M_{22} \cdot y && (82)\\
B_y &= -2M_{22} \cdot x, && (83)
\end{aligned}
$$

and looking at the Lorentz forces that a particle moving mostly in s-direction experiences, we again see that if there is focusing in x-direction, there is defocusing in y-direction and vice versa.

To study higher orders in k, let us consider the case $k = 3$. For $\theta_{3,3} = 0$, we obtain

$$
\begin{aligned}
V &= M_{3,3} \cos(3\phi)\, r^3\\
&= M_{3,3} (\cos\phi \cos 2\phi - \sin\phi \sin 2\phi)\, r^3\\
&= M_{3,3} \left(xr^2 \cos 2\phi - yr^2 \sin 2\phi\right)\\
&= M_{3,3} \left\{x\left(x^2 - y^2\right) - 2xy^2\right\}\\
& \quad M_{3,3} \left(x^3 - 3xy^2\right). && (84)
\end{aligned}
$$

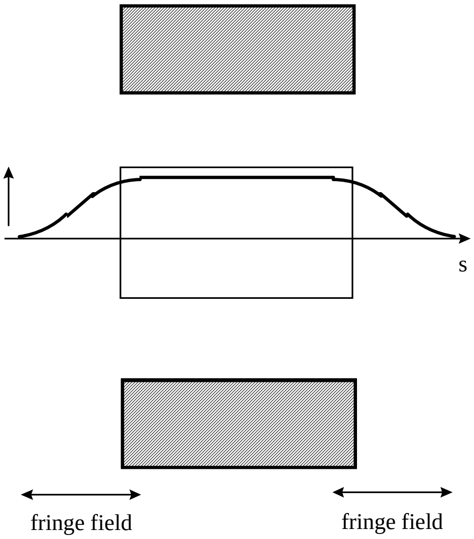
In this case, the resulting forces are quadratic, and are thus not suitable for the affecting the linear motion; but we shall see later that they are indeed very convenient for the correction of nonlinear motion, and they even have the nice feature of having **no influence** on the linear part of the motion! Another important case for $\theta_{3,3}$ is $\theta_{3,3} = -\pi/2$, in which case one can perform a similar argument and again obtain cubic dependencies on the position.

For all the **higher values of l**, corresponding to octupoles, decapoles, duodecapoles etc, the procedure is very similar. We begin with the addition theorem for $\cos(l\phi)$ or $\sin(l\phi)$, and by induction we see that each of which consists of terms that have a product of precisely l cosines and sines. Since each of these terms is multiplied with r^l, each cosine multiplied with one r translates into an x, and each sine multiplied with one r translates into a y; the end result is always a polynomial in x and y of exact order l.

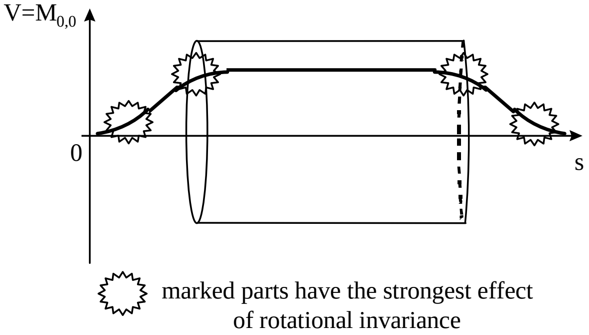
Because of their nonlinear field dependence, these elements will prove to have no effect on the motion up order $l-1$,and thus allow to selectively influence the higher orders of the motion without affecting the lower orders. And if it is the crux of particle optical motion that the horizontal and vertical linear motion cannot be affected simultaneously, it is its blessing that the **nonlinear effects can be corrected order-by-order**.

In the case there is no s-dependence, the potential terms that we have derived are the only ones; under the **presence of s-dependence**, as shown in eq. (76), to the given angular dependence there are higher order terms in r, the strengths of which are given by the s-derivatives of the multipole strength $M_{l,l}$. The computation of their Cartesian form is very easy once the Cartesian form of the leading term is known, because each additional term just differs by the previous one just by the factor of $r^2 = (x^2 + y^2)$.

In practice, of course, s-dependence is unavoidable: the field of any particle optical element has to begin and end somewhere, and it usually does this by rising and falling gently with s, entailing s-dependence. This actually entails another crux of particle optics: even the quadrupoles, **the "linear" elements, have nonlinear effects** at their edges, requiring higher order correction. The corrective elements in turn have higher order edge effects, possibly requiring even higher order correction, etc.



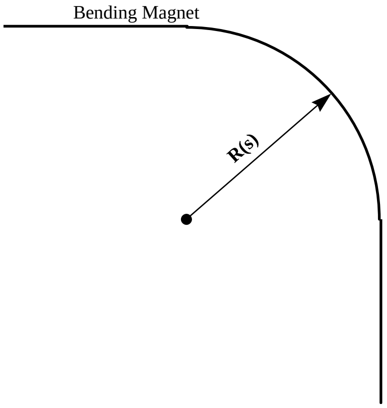
While without s-dependence, the case $l = 0$ corresponding to full rotational symmetry was not very interesting, if we consider s-dependence, it actually offers a remarkably useful effect. While there is no r-dependence in the leading term, the contributions through the derivatives of $M_{0,0}(s)$ entail terms with an r-dependence of the form $r^2, r^4, \ldots$ Of these, the r^2 terms will indeed produce **linear, rotationally symmetric fields,** similar to those in the glass lens. Unfortunately, in practice these fields are restricted to the entrance and exit fringe field regions and are comparatively weak; furthermore, there are usually quite large nonlinearities, and altogether these devices are usually used mostly for low-energy, small emittance beams, like those found in electron microscopes.


V=M$_{0,0}$
0
s
marked parts have the strongest effect of rotational invariance

B. Fields with Planar Reference Orbit

In the case of the straight reference orbit, we saw that Maxwell's equations entail a very clean connection between rotational symmetry and radial potential. As one may expect, in the case of a non-straight reference orbit, this is no longer the case; in this situation, Maxwell's equations have a rather different but not less interesting consequence as long as we restrict ourselves to the case in which the reference orbit stays in one plane.

As it turns out, in this case the arguments to express the Laplacian in the new coordinates is similar to that in cylindrical coordinates. Let us assume that the motion of the reference particle is in a plane, and that all orbits that are on this plane stay in it. Let $R(s)$ be the momentary radius of curvature as shown in the picture.



Then we have a situation very similar to cylindrical coordinates r, ϕ, z centered around the momentary origin of $R(s)$. In fact, setting $h(s) = 1/R(s)$, the particle optical coordinates s, x, y correspond to the cylindrical ones in the following way:

$$\begin{aligned} z &\leftrightarrow y \\ r &\leftrightarrow (1+hx)\cdot R(s) \\ \phi &\leftrightarrow \frac{s}{R(s)} \end{aligned}$$

As we recall from the previous section, in cylindrical coordinates the Laplacian had the form

$$\Delta V = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial V}{\partial r}\right) + \frac{1}{r}\frac{\partial}{\partial \phi}\left(\frac{1}{r}\frac{\partial V}{\partial \phi}\right) + \frac{\partial^2 V}{\partial z^2}$$

So we may expect that in particle optical coordinates, we in fact have

$$\begin{aligned} \Delta V &= \frac{1}{1+hx}\frac{\partial}{\partial x}\left((1+hx)\frac{\partial V}{\partial x}\right) \\ &+\frac{1}{1+hx}\frac{\partial}{\partial s}\left(\frac{1}{1+hx}\frac{\partial V}{\partial s}\right) \\ &+\frac{\partial^2 V}{\partial y^2}. \end{aligned} \quad (85)$$

A careful analysis based on the chain rule and determining the proper Jacobian reveals that this is indeed the case; the calculations are rather mechanical, but very involved and rather boring, and we skip them for the purposes of these lectures.

For the potential, we again make an expansion in transversal coordinates, and leave the longitudinal coordinates unexpanded. Since we are working now with x and y, both expansions are Taylor, and we have

$$V = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} a_{k,l}(s)\frac{x^k y^l}{k!l!}.$$

This expansion now has to be inserted into the Laplacian in particle optical coordinates. Besides the mere differentiation, we also have to Taylor expand $1/(1+hx) = 1-(hx)+(hx)^2-(hx)^3+\dots$ After gathering like terms and heavy arithmetic, and again using the convention that coefficients with negative indices are assumed to vanish, we obtain the recursion relation

$$\begin{aligned}
a_{k,l+2} &= -a''_{k,l} - kha''_{k-1,l} \\
&\quad +kh'a'_{k-1,l} - a_{k+2,l} \\
&\quad -(3k+1)\,ha_{k+1,l} \\
&\quad -3kha_{k-1,l+2} \\
&\quad -k\,(3k-1)\,h^2a_{k,l} \\
&\quad -3k\,(k-1)\,h^2a_{k-2,l+2} \\
&\quad -k\,(k-1)^2\,h^3a_{k-1,l} \\
&\quad -k\,(k-1)\,(k-2)\,h^3a_{k-3,l+2} \quad (86)
\end{aligned}$$

Although admittedly horrible and unpleasant, the formula apparently has the coefficient of highest total order $k+l+2$ on the left hand side, and thus recursively allows the calculation of coefficients. Indeed, the terms $a_{k,0}(s), a_{k,1}(s)$ can be chosen freely, and all others are uniquely determined through them.

To study the significance of the free terms, let us consider the electric and magnetic case separately. In the electric case, in order to assure that orbits that were in the plane stay there, there must not be any field components in the y-direction in the plane corresponding to $y=0$. Computing the gradient of the potential, we have

$$\begin{aligned}
E_x(x,y=0) &= -\sum_k a_{k,0}\frac{x^{k-1}}{(k-1)!} \\
E_y(x,y=0) &= -\sum_k a_{k,1}\frac{x^k}{k!} = 0
\end{aligned}$$

and looking at E_y,we conclude that $a_{k,1} = 0$ for all k. So the terms $a_{k,0}$ alone specify the field. Looking at E_x,we see that these are just the coefficients that specify the field within the plane, and so **the planar field determines the entire field.** Furthermore, looking at the details of the recursion relation, it becomes apparent that all second indices are either l or $l+2$. This entails that as long as $a_{k,1}$ terms do not appear, also $a_{k,3}$, $a_{k,5}$,... terms do not appear. Indeed, the resulting potential is fully symmetric around the plane, and the resulting field lines above and below the plane are mirror images.

In the magnetic field, the argument is rather similar: considering the fields in the plane, we have

$$\begin{aligned}
B_y(x,y=0) &= -\sum_k a_{k,1}\frac{x^k}{k!} \\
B_x(x,y=0) &= -\sum_k a_{k,0}\frac{x^{k-1}}{(k-1)!} = 0
\end{aligned}$$

In order for particles in the midplane to stay there, we must have that B_x vanishes in the midplane, which entails $a_{k,0} = 0$. So in the magnetic case, the coefficients $a_{k,1}$ specify everything. These coefficients, however, again describe the shape of the field in plane, and so again **the planar field determines the entire field.** In the magnetic case, the potential is fully antisymmetric around the plane, and again the resulting field lines are mirror images of each other.

VI. The Equations of Motion in Curvilinear Coordinates

There are a variety of methods to derive the equation of motion in curvilinear coordinates with the arclength s as the independent variable. Perhaps the most sophisticated and appealing of them is in the Lagrangian picture, in which one first expresses Cartesian variables by curvilinear coordinates and re-writes the Lagrangian. Then one proceeds to the Hamiltonian through a Legendre transformation. In the Hamiltonian picture, it is then possible to perform a change of independent variable from t to s while maintaining the Hamiltonian structure.

While very illuminating, the **Lagrangian-Hamiltonian** mechanism is **too involved for our purposes** and the limited amount of time we have, and we thus follow a more straightforward, classical way. For simplicity, we also restrict ourselves in that the reference orbit is allowed to bend in only one plane. As a function of the arclength s, we first define the momentary **curvature** of the reference orbit as $h(s)$. If the curvature is nonzero, the radius of curvature is then given by $R(s) = 1/h(s)$. We begin by studying the bend angle that the reference orbit experiences as we move from position s_0 to position s. We have

$$\alpha = \int_{\alpha_0} \alpha d\alpha = \int_{s_0}^{s} \frac{ds}{R(s)} = \int_{s_0}^{s} h(s) ds \qquad (87)$$

Let us remind ourselves that in Cartesian coordinates, the equations of motion have the Lorentz form

$$\frac{d}{dt}\vec{p} = \vec{F}, \text{ and } \frac{d}{dt}\vec{r} = \frac{\frac{\vec{p}}{m}}{\sqrt{1 + \frac{p^2}{m^2c^2}}} \qquad (88)$$

Note that in the second equation, we have expressed the velocity-dependent terms purely in terms of $\vec{p}$, as we want to maintain only one momentum or velocity component. For the purpose of our derivation, we re-write the equations as an integral equation:

$$\begin{aligned} \vec{p}(s) &= \vec{p}(s_0) + \int_{t(s_0)}^{t(s)} \vec{F}(t) dt \qquad (89) \\ &= \vec{p}(s_0) + \int_{s_0}^{s} \vec{F}(s) t^{'} ds, \end{aligned}$$

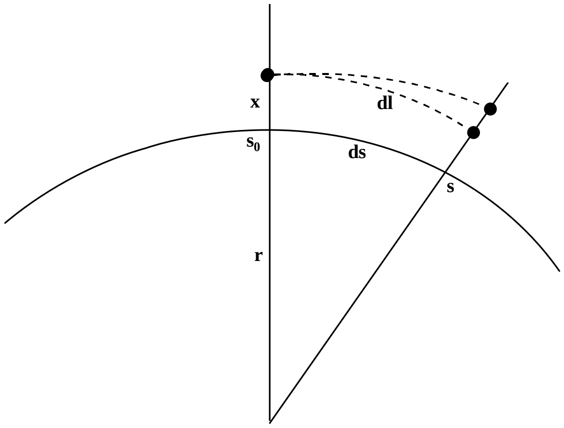
where we have used $t^{'} = dt/ds$, and it is worthwhile to remind ourselves that force $\vec{F}(s)$ still depends on both $\vec{x}$ and $\vec{p}$. As we have progressed from s_0 to s, the orientation of our locally attached particle optical coordinate system has changed, it was rotated by the angle α of eq. (87). So in the new local coordinates, we have

$$\begin{aligned} \vec{p_l}(s) &= \begin{pmatrix} \cos \int_{s_0}^{s} h & 0 & \sin \int_{s_0}^{s} h \\ 0 & 1 & 0 \\ -\sin \int_{s_0}^{s} h & 0 & \cos \int_{s_0}^{s} h \end{pmatrix} \qquad (90) \\ &\cdot \left(\vec{p}(s_0) + \int_{s_0}^{s} \vec{F}(\bar{s}) t^{'} d\bar{s} \right); \end{aligned}$$

the matrix will be denoted by $\hat{M}(s)$. Differentiating with respect to s , and evaluating at $s = s_0$ allows us to conveniently obtain the rate of change of the momentum $\vec{p_l}$, and we obtain

$$\begin{aligned} \vec{p_l}'(s) &= \left[\begin{array}{c} \widehat{M} \ (\overline{s}) \vec{F}(\overline{s}) t^{'} \\ + \widehat{M}^{'} \ (\overline{s}) \left(\vec{p}(s) + \int_{s}^{\overline{s}} \vec{F}(\widetilde{s}) t^{'} d\widetilde{s} \right) \end{array} \right]_{\overline{s}=s} \\ &= \vec{F}(s) t^{'} + \begin{pmatrix} 0 & 0 & h(s) \\ 0 & 0 & 0 \\ -h(s) & 0 & 0 \end{pmatrix} \vec{p}(s) (91) \end{aligned}$$

Note that the first resulting term depends on the actual forces and the factor t' accounts for the fact that we went to s as an independent variable. The second term is a pseudo force due to the fact that we are located in a rotating frame. Indeed, for $h = 0$, we obtain the conventional result. We also note in passing that if we were to allow out-of-plane motion of the reference orbit, then the matrix $\widehat{M}$ would depend on two curvatures. Unfortunately, in this case an additional complication arises from the fact that rotations around different axes don't generally commute.



Next we make an observation regarding the rate of change at which distances are covered at different positions x. Looking at the picture, we observe

$$\begin{aligned}\frac{dl}{ds} &= \frac{x+r}{r} = 1 + hx \\ \frac{dx}{ds} &= (1+hx)\frac{dx}{dl} = (1+hx)\frac{p_x}{p_s} \qquad (92)\end{aligned}$$

Similarly we obtain

$$\frac{dy}{ds} = (1+hx)\,\frac{p_y}{p_s}, \qquad (93)$$

and for the time of flight:

$$\begin{aligned}\frac{dt}{ds} &= \frac{1}{v}\sqrt{\left(\frac{dx}{ds}\right)^2 + \left(\frac{dy}{ds}\right)^2 + \left(\frac{dl}{ds}\right)^2} = \\ &= \frac{1}{v}\,(1+hx)\,\sqrt{1+\frac{p_x^2+p_y^2}{p_s^2}} = \\ &= \frac{1}{v}\,(1+hx)\,\frac{p}{p_s} \qquad (94)\end{aligned}$$

where $p = \sqrt{p_x^2+p_y^2+p_s^2}$ has been used.

Altogether, we have so far obtained the equations of motion in local coordinates with s as the independent variable. From there to the particle optical variables only a small step is left. We remind ourselves that the particle optical coordinates are

$$\begin{array}{cc} x & a = p_x/p_0 \\ y & b = p_y/p_0 \\ l = k(t-t_0) & \delta = (E-E_0)\,/E_0 \end{array} \qquad (95)$$

where p_0 is a fixed momentum and E_0 and t_0 are energy and time of flight of the reference particle, and E is the total (kinetic plus potential) energy of the particle under consideration. In order to study relativistic effects, it will turn out to be advantageous to introduce the relativistic measure

$$\eta = \frac{E - eV(x,y,s)}{mc^2}, \qquad (96)$$

the ratio of kinetic to rest mass energy. Obviously, we have

$$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} = 1+\eta \qquad (97)$$

and we also have

$$\begin{aligned}\frac{v}{c} &= \sqrt{1-\left(\frac{1}{1+\eta}\right)^2} \\ &= \sqrt{\frac{2\eta+\eta^2}{(1+\eta)^2}} = \frac{\sqrt{\eta\,(2+\eta)}}{1+\eta} \quad \text{and} \qquad (98) \\ \frac{p}{mc} &= \frac{mv\gamma}{mc} = \frac{v}{c}\gamma = \sqrt{\eta\,(2+\eta)} \qquad (99)\end{aligned}$$

As a first step, we express the rate of change of the particle optical variable l in terms of particle optical quantities. We obtain

$$\begin{aligned}l' &= \frac{dl}{ds} \\ &= k\left(t' - t_0'\right) = \frac{k}{v}\,(1+hx)\,\frac{p}{p_s} - kt_0' \\ &= (1+hx)\,\frac{k}{p_0}\frac{p}{v}\frac{p_0}{p_s} - \frac{k}{v_0} \\ &= (1+hx)\,(1+\eta)\,m\frac{k}{p_0}\frac{p_0}{p_s} - \frac{k}{v_0} \\ &= \left\{(1+hx)\,\frac{1+\eta}{1+\eta_0}\frac{p_0}{p_s} - 1\right\}\frac{k}{v_0} \qquad (100)\end{aligned}$$

where we have used the following relations:

$$
\begin{aligned}
t_0^{'} &= \left[\frac{1}{v}\left(1+hx\right)\frac{p_0}{p_s}\right]_0 = \frac{1}{v_0} \\
\frac{p}{v} &= \left(1+\eta\right)\cdot m \\
\frac{p_0}{v_0} &= \left(1+\eta_0\right)\cdot m_0
\end{aligned}
$$

Next we obtain for the positions

$$
\begin{aligned}
\frac{p_0}{p_s} &= \frac{p_0}{\sqrt{p^2-p_x^2-p_y^2}} \\
&= \left(\frac{p^2}{p_0^2}-a^2-b^2\right)^{-\frac{1}{2}} = \\
&= \left(\frac{\eta\left(2+\eta\right)}{\eta_0\left(2+\eta_0\right)}\cdot\frac{m^2}{m_0^2}-a^2-b^2\right)^{-\frac{1}{2}}.
\end{aligned}
$$

Because we always have $\vec{v} \parallel \vec{p}$, it follows that $\frac{\vec{v}}{v}=\frac{\vec{p}}{p}$,and altogether

$$
\begin{aligned}
&\frac{d}{ds}\left(\frac{p_x}{p_0},\frac{p_y}{p_0},\frac{p_s}{p_0}\right) \\
&= \frac{1}{p_0}\vec{F}\left(s\right)t^{'}+\begin{pmatrix} 0 & 0 & h \\ 0 & 0 & 0 \\ -h & 0 & 0 \end{pmatrix}\frac{\vec{p}}{p_0} = \\
&= ze\vec{E}\frac{t^{'}}{p_0}+ze\frac{\vec{v}}{v}\times\vec{B}\left(1+hx\right)\frac{p}{p_s} \\
&\quad +h\left(\frac{p_s}{p_0},0,-\frac{p_x}{p_0}\right) \\
&= \frac{\vec{E}}{\chi_{e0}}\left(1+\frac{l^{'}v_0}{k}\right)+\frac{\vec{p}}{p_0}\times\frac{\vec{B}}{\chi_{m0}}\left(1+hx\right)\frac{p_0}{p_s} \\
&\quad +h\left(\frac{p_s}{p_0},0,-\frac{p_x}{p_0}\right), \qquad (101)
\end{aligned}
$$

where the following abbreviations have been used:

$$
\begin{aligned}
\chi_{m0} &= \frac{p_0}{ze} \qquad (102) \\
\chi_{e0} &= \frac{p_0 v_0}{ze} \qquad (103)
\end{aligned}
$$

The quantities χ_m and χ_e are called the magnetic and electric rigidity, respectively; they describe directly to what extent the magnetic and electric fields influence the geometric motion of the particles. Continuing from eq. (101) we have

$$
\begin{aligned}
&= \frac{\vec{E}}{\chi_{e0}}\left(1+hx\right)\frac{1+\eta}{1+\eta_0}\frac{p_0}{p_z} \\
&\quad +\left(bB_z-\frac{p_z}{p_0}B_y,\frac{p_z}{p_0}B_x-aB_z,aB_y-bB_x\right) \\
&\quad \frac{p_0}{p_z}\frac{1}{\chi_{m0}}\left(1+hx\right)+h\left(\frac{p_z}{p_0},0,-\frac{p_x}{p_0}\right).
\end{aligned}
$$

Finally, we consider the change of the last variable in the particle optical coordinates, δ. Since it measures the deviations of kinetic and potential energies, as long as there is conservation of energy, we have $\delta'=0$. This is of course the case in all time-independent cases; in cavities, the situation is a little different, but we don't concern us here with this case.

Now we merely observe that $\vec{p}/p_0 = (a,b,p_s/p_0)$, and obtain the equations of motion in particle optical coordinates:

$$
\begin{aligned}
x^{'} &= a\left(1+hx\right)\frac{p_0}{p_s} \qquad (104) \\
y^{'} &= b\left(1+hx\right)\frac{p_0}{p_s} \qquad (105) \\
l^{'} &= \left\{\left(1+hx\right)\frac{1+\eta}{1+\eta_0}\frac{p_0}{p_s}-1\right\}\frac{k}{v_0} \qquad (106) \\
a^{'} &= \left[\frac{1+\eta}{1+\eta_0}\frac{p_0}{p_s}\frac{E_x}{\chi_{e0}}+b\frac{B_z}{\chi_{m0}}\frac{p_0}{p_s}-\frac{B_y}{\chi_{m0}}\right] (\cdot 107) \\
&\quad \cdot\left(1+hx\right)+h\frac{p_s}{p_0} \\
b^{'} &= \left[\frac{1+\eta}{1+\eta_0}\frac{p_0}{p_s}\frac{E_y}{\chi_{e0}}+\frac{B_x}{\chi_{m0}}-a\frac{B_z}{\chi_{m0}}\frac{p_0}{p_s}\right] (\cdot 108) \\
&\quad \cdot\left(1+hx\right) \qquad (109) \\
\delta^{'} &= 0 \qquad (110)
\end{aligned}
$$

where the abbreviations

$$
\eta = \frac{E-eV(x,y)}{mc^2} \text{ and } \qquad (111)
$$

$$\frac{p_0}{p_s} = \left(\frac{\eta\,(2+\eta)}{\eta_0\,(2+\eta_0)} \cdot \frac{m^2}{m_0^2} - a^2 - b^2\right)^{-\frac{1}{2}} \quad (112)$$

have been used.

A careful analysis of the equations of motion reveals that indeed if all the particle optical coordinates are small, so are their derivatives defined through the equations of motion; indeed, the system is weakly nonlinear. Furthermore, one can show that the system is even Hamiltonian; but for lack of time, we do not concern ourselves here with the development of the Hamiltonian function.

VII. The Linearization of the Equations of Motion

In order to develop a matrix theory of particle optics similar to the Gaussian theory in glass optics, we have to linearize the equations of motion. This is procedure is rather similar to other linearizations in Physics, in particular it is very similar to the study of so-called "small oscillations" in Mechanics. Since the solutions of linear systems depend linearly on the initial conditions, indeed the resulting transfer maps will be linear as needed.

We begin the actual process of linearization with the linearization of the fields, which corresponds to quadratic potentials. In the electric case, let us assume that there is no potential on axis, i.e. $a_{0,0} = 0$, and that in the midplane, we have

$$E_x = -E_{x0} \cdot (1 + n_e x)\,. \quad (113)$$

Because of the recursion relation for fields, we obtain an out-of-plane expansion of

$$E_y = E_{x0} \cdot (h + n_e) y \quad (114)$$

as well as a potential

$$V(x,y) = E_{x0} \cdot x + \frac{1}{2} E_{x0} (n_e x^2 - (h + n_e)\, y^2); \quad (115)$$

which is chosen in such a way as to vanish on the reference orbit. In the magnetic case, let the midplane field be given by

$$B_y = B_{y0} \cdot (1 + n_b x); \quad (116)$$

due to the recursion relation, we must then have

$$B_x = B_{y0} \cdot n_b y. \quad (117)$$

Before we even discuss linearization, let us consider the "zeroth order" of the motion: if the system is supposed to be origin preserving, then we must have from the equations of motion for a' that

$$\frac{E_x}{\chi_{e0}} - \frac{B_y}{\chi_{m0}} = h, \quad (118)$$

which in a natural and expected way couples the constant parts of the fields with the curvature of the reference orbit. Now we begin our process of linearization. It is easy to see that

$$\begin{aligned} x' &= a \quad (119)\\ y' &= b. \quad (120) \end{aligned}$$

We also obtain

$$\frac{\eta}{\eta_0} = \frac{m_0}{m} \cdot (1 + \delta - \mathcal{S} E_{x0} x)$$

and after more complicated expansions

$$\begin{aligned} \frac{1+\eta}{1+\eta_0} &= 1 + \frac{\eta_0}{1+\eta_0}\delta - \frac{ze}{(1+\eta_0)\,mc^2} E_{x0} x \\ \frac{2+\eta}{2+\eta_0} &= 1 + \frac{\eta_0}{2+\eta_0}\delta - \frac{ze}{(2+\eta_0)\,mc^2} E_{x0} x \end{aligned}$$

Similarly, we obtain

$$\begin{aligned} \frac{p_z}{p_0} &= 1 + \frac{1+\eta_0}{2+\eta_0}\delta \\ &- \frac{1}{2} x E_{x0} \left(\mathcal{S} + \frac{ze}{(2+\eta_0)\,mc^2}\right) \end{aligned}$$

where the abbreviation $\mathcal{S} = \frac{ze}{K_0} = \frac{ze}{mc^2\eta_0}$ has been used. After lengthy similar arguments, we also conclude

$$l^{'} =_1 \left\{ \begin{array}{c} hx - \frac{1}{(1+\eta_0)(2+\eta_0)}\delta \\ +E_{x0}x\mathcal{S}\frac{1}{(1+\eta_0)(2+\eta_0)} \end{array} \right\} \frac{k}{v_0} \tag{121}$$

as well as

$$\begin{aligned} a' &= \left\{ h\frac{1+\eta_0}{2+\eta_0} - \frac{E_{x0}}{\chi_{e0}}\frac{1}{(1+\eta_0)(2+\eta_0)} \right\} \delta - \\ &\quad -x \left[\begin{array}{c} h^2 - \frac{E_{x0}}{\chi_{e0}} n_e + \frac{B_{y0}}{\chi_{m0}} n_B \\ -\frac{E_{x0}}{\chi_{e0}} E_{x0}\mathcal{S}\frac{2+2\eta_0+\eta_0^2}{(1+\eta_0)(2+\eta_0)} \\ +\frac{B_{y0}}{\chi_{m0}} E_{x0}\mathcal{S}\frac{1+\eta_0}{2+\eta_0} \end{array} \right] \qquad (122) \\ b^{'} &= -\frac{E_{x0}}{\chi_{e0}}(h+n_e)y + \frac{B_{y0}}{\chi_{m0}} n_b y \qquad (123) \end{aligned}$$

Now that the equations of motion have been linearized, they have to be studied for a variety of different cases. We begin with the simplest case:

A. The Drift

In this case, the linearized equations of motion have the form

$$\begin{aligned} x^{'} &= a \\ y^{'} &= b \\ a^{'} &= 0 \\ b^{'} &= 0 \\ l^{'} &= -\frac{k}{v_0}\frac{1}{(1+\eta_0)(2+\eta_0)}\delta \\ \delta' &= 0 \qquad (124) \end{aligned}$$

where of course only the last equation is of any real interest. These equations are trivial to integrate, and we obtain

$$\begin{aligned} x_f &= x_i + a_i l \\ y_f &= y_i + b_i l \\ l_f &= -\frac{k}{v_0}\frac{l}{(1+\eta_0)(2+\eta_0)}\delta + l_i \\ a_f &= a_i \\ b_f &= b_i \\ \delta_f &= \delta_i \end{aligned}$$

which can be written in matrix form as

$$\begin{pmatrix} x_f \\ a_f \\ y_f \\ b_f \\ l_f \\ \delta_f \end{pmatrix} = \begin{pmatrix} 1 & l & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & l & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & D \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_i \\ a_i \\ y_i \\ b_i \\ l_i \\ \delta_i \end{pmatrix}$$

where $D = -\frac{k}{v_0}\frac{l}{(1+\eta_0)(2+\eta_0)}$ has been used.

B. The Electric Quadrupole without Fringe Field

More interesting is the case of the electric quadrupole. To remind ourselves, we have

$$\begin{aligned} V &= M_{2,2}\cos(2\phi)\,r^2 = M_{2,2}\left(x^2 - y^2\right) \\ \text{and } E_x &= -2M_{2,2}x \text{ , } E_y = 2M_{2,2}y \end{aligned}$$

The equations of motion have the form

$$\begin{aligned} x^{'} &= a \\ y^{'} &= b \\ a^{'} &= -x\frac{2M_{2,2}}{\chi_{e0}} \\ b^{'} &= y\frac{2M_{2,2}}{\chi_{e0}} \\ l^{'} &= -\frac{k}{v_0}\frac{1}{(1+\eta_0)(2+\eta_0)}\delta \\ \delta' &= 0 \qquad (125) \end{aligned}$$

Apparently we have sine-cosine solutions in the horizontal plane, and sinh-cosh solutions in the vertical plane. Calling $\omega = \sqrt{2M_{2,2}/\chi_{e0}}$, we obtain as the solution

$$\left\{\begin{array}{l} x_f = x_i \cos\omega L + a_i \frac{\sin\omega L}{\omega} \\ a_f = -\omega x_i \sin\omega L + a_i \cos\omega L \\ y_f = y_i \cosh\omega L + b_i \frac{\sinh\omega L}{\omega} \\ b_f = \omega y_i \sinh\omega L + b_i \cosh\omega L \\ l_f = -\frac{k}{v_0}\frac{1}{(1+\eta_0)(2+\eta_0)}\delta L + l_i \\ \delta_f = \delta_i \end{array}\right. .$$

Using the abbreviations $c_x = \cos(\omega L)$, $s_x = \sin(\omega L)/\omega$, $c_y = \cosh(\omega L)$, $s_y = \sinh(\omega L)/\omega$ as well as $D = -\frac{k}{v_0}\frac{l}{(1+\eta_0)(2+\eta_0)}$, this can be written in matrix form as

$$\begin{pmatrix} x_f \\ a_f \\ y_f \\ b_f \\ l_f \\ \delta_f \end{pmatrix} = \begin{pmatrix} c_x & s_x & 0 & 0 & 0 & 0 \\ c_x' & s_x' & 0 & 0 & 0 & 0 \\ 0 & 0 & c_y & s_y & 0 & 0 \\ 0 & 0 & c_y' & s_y' & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & D \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_i \\ a_i \\ y_i \\ b_i \\ l_i \\ \delta_i \end{pmatrix} \tag{126}$$

Observe that, as in the case of glass optics, the determinant is unity. Furthermore, note that if $M_{2,2} < 0$, ω is imaginary. In this case, the x- and y-planes exchange their roles, the quadrupole becomes focusing in the vertical direction and defocusing in the horizontal direction.

It is also worthwhile to briefly touch the case of fringe fields. In this case, $M_{2,2}$ changes as function of s. The resulting ODE is still linear, which entails that the result can be written in matrix form, but in most cases is impossible to solve it analytically.

C. The Magnetic Quadrupole without Fringe Field

In the case of the magnetic quadrupole, we have

$$\begin{aligned} V &= -2M_{2,2} x \cdot y \\ B_x &= 2M_{2,2} y \\ B_y &= 2M_{2,2} x \end{aligned}$$

which results in the linear equations

$$\begin{aligned} x' &= a \\ y' &= b \\ a' &= -\frac{2M_{2,2}}{\chi_{m0}} x \\ b' &= \frac{2M_{2,2}}{\chi_{m0}} y \\ l' &= -\frac{k}{v_0}\frac{1}{(1+\eta_0)(2+\eta_0)}\delta \\ \delta' &= 0 \end{aligned} \tag{127}$$

Similar to before, we introduce $\omega = \sqrt{2M_{2,2}/\chi_{m0}}$,and the resulting transfer matrix is the same as in the case of the electric quadrupole.

D. The Linear Magnetic Dipole

The next particle optical element we want to consider is the magnetic dipole, consisting of constant magnetic field in the y-direction. In terms of the quantities describing the linearized fields, we have

$$\begin{aligned} B_{y0} &= const,\, E_{x0} = 0 \\ n_e &= m_e = n_B = m_B = 0 \end{aligned}$$

Reminding ourselves about magnet design, such a field can be obtained very schematically with the following arrangement:

Let us now consider the equations of motion; we obtain

$$\begin{aligned} x' &= a \\ y' &= b \\ a' &= -xh^2 + \delta h \frac{1+\eta_0}{2+\eta_0} \\ b' &= 0 \\ l' &= \frac{k}{v_0}\left[hx - \frac{1}{(1+\eta_0)(2+\eta_0)}\delta\right] \\ \delta' &= 0 \end{aligned} \tag{128}$$

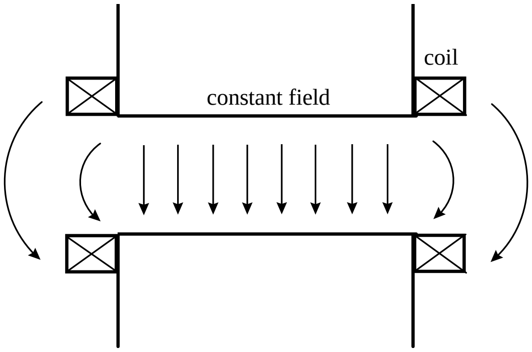


Figure 2:

First we observe that if we choose $h = 0$, we obtain $a^{'} = 0$, and the same situation as in the case of a drift. But even for the case of $h \neq 0$, the motion of the y-direction behaves simply like a drift, and we always have

$$\begin{aligned} b_f &= b_i \\ y_f &= y_i + b_i L \end{aligned}$$

where L is the arclength of the dipole.

Next we observe that as always, δ stays constant, and hence in the equation for a' plays the role of a parameter, making the differential equation inhomogeneous.. Finally we observe that since l does not couple into the horizontal or vertical motion, we can solve the equation for l after the horizontal motion is analyzed by a mere integration.

In order to solve the horizontal part of the motion, we first solve the homogenous part, which has the form

$$\begin{aligned} x' &= a \\ a' &= -x \cdot h^2 \end{aligned}$$

and setting $\omega = h$, we obtain as a solution

$$\begin{aligned} & x_f = x_0 \cos \omega L + \tfrac{1}{\omega} a_0 \sin \omega L \\ & = x_0 \cos \phi + R_0 a_0 \sin \phi \\ & a_f = -\omega x_0 \sin \omega L + a_0 \cos \omega L \\ & = -\tfrac{1}{R_0} x_0 \sin \phi + a_0 \cos \phi \end{aligned}$$

where we have used that $R_0 = 1/h$ and $L/R_0 = \phi$. Altogether, we have a behavior not much different from a focusing quadrupole. In order to treat the inhomogeneity, we perform a so-called ”Variation of the Parameters”, that is we make an Ansatz of the form

$$\left\{ \begin{array}{l} x(s) = \overline{x}_0(s) \cos \phi + R_0 \overline{a}_0(s) \sin \phi \\ a(s) = -\frac{1}{R_0} \overline{x}_0(s) \sin \phi + \overline{a}_0(s) \cos \phi \end{array} \right. ,$$

where now the original ”parameters” $\bar{x}_0$ etc. are viewed as functions of s. Inserting into the ODE, we obtain the following condition:

$$\begin{aligned} \overline{x}_0^{'}(s) \cos \omega s + \frac{1}{\omega} \overline{a}_0^{'}(s) \sin \omega s &= 0 \\ -\omega \overline{x}_0^{'} \sin \omega s + \overline{a}_0^{'} \cos \omega s &= \varkappa, \end{aligned}$$

where $\varkappa = \delta h \frac{1+\eta_0}{2+\eta_0}$. Rewriting in matrix form, this reads

$$\left(\begin{array}{cc} \cos \omega s & \frac{1}{\omega} \sin \omega s \\ -\omega \sin \omega s & \cos \omega s \end{array} \right) \left(\begin{array}{c} \overline{x}_0^{'} \\ \overline{a}_0^{'} \end{array} \right) = \left(\begin{array}{c} 0 \\ \varkappa \end{array} \right).$$

Multiplying with the inverse matrix and integrating, we obtain

$$\begin{aligned} \overline{x}_0^{'} &= \left(\int_0^s \left(-\frac{1}{\omega} \sin \omega s \right) \varkappa ds \right) + x_0 \\ &= \frac{1}{\omega^2} (\cos \omega s - 1) \varkappa + x_0 \\ \overline{a}_0^{'} &= \left(\int_0^s \cos \omega s \ \varkappa ds \right) + a_0 \\ &= \frac{1}{\omega} \sin \omega s \varkappa + a_0 \end{aligned}$$

So the complete solution of the inhomogeneous part has the form

$$\begin{aligned} x(s) &= \overline{x}_0(s) \cos \phi + R_0 \overline{a}_0(s) \sin \phi \\ &= \left[\frac{\varkappa}{\omega^2} (\cos \omega s - 1) + x_0 \right] \cos \omega s \\ & \quad + \frac{1}{\omega} \left[\frac{\varkappa}{\omega} \sin \omega s + a_0 \right] \sin \omega s \\ &= x_0 \cos \phi + \frac{1}{\omega} a_0 \sin \phi \\ & \quad + \frac{\varkappa}{\omega^2} (1 - \cos \phi) \end{aligned}$$

$$\begin{aligned} x_f &= x_0\cos\phi + R_0 a_0 \sin\phi \\ &\quad + R_0\left(1-\cos\phi\right)\frac{1+\eta_0}{2+\eta_0}\delta \\ a_f &= -\frac{1}{R_0}x_0\sin\phi + a_0\cos\phi \\ &\quad + \sin\phi\frac{1+\eta_0}{2+\eta_0}\delta \end{aligned}$$

Finally we have to study the case of the time of flight part, which as we said before can be obtained by mere integration. We have

$$\begin{aligned} l_f &= \frac{k}{v_0}\int_0^s \left\{ \begin{array}{c} hx \\ -\frac{1}{(1+\eta_0)(2+\eta_0)}\delta \end{array} \right\} ds + l_0 \\ &= \frac{k}{v_0}\int_0^s \left\{ \begin{array}{c} \frac{1}{R_0}x_0\cos\frac{s}{R_0} \\ +a_0\sin\frac{s}{R_0} \\ +\left(1-\cos\frac{s}{R_0}\right)\frac{1+\eta_0}{2+\eta_0}\delta \\ -\frac{1}{(1+\eta_0)(2+\eta_0)}\delta \end{array} \right\} ds + l_0 \\ &= \frac{k}{v_0}\left[\begin{array}{c} x_0\sin\frac{s}{R_0} \\ -R_0 a_0\cos\frac{s}{R_0} \\ -R_0\sin\frac{s}{R_0}\frac{1+\eta_0}{2+\eta_0}\delta \\ +\frac{1+\eta_0}{2+\eta_0}\delta s - \frac{1}{(1+\eta_0)(2+\eta_0)}\delta s \end{array} \right]_0^s + l_0 \\ &= \frac{k}{v_0}\left[\begin{array}{c} x_0\sin\frac{s}{R_0} \\ -R_0 a_0\left(\cos\frac{s}{R_0}-1\right) \\ -R_0\sin\frac{s}{R_0}\frac{1+\eta_0}{2+\eta_0}\delta \\ +\frac{1+\eta_0}{2+\eta_0}\delta s - \frac{1}{(1+\eta_0)(2+\eta_0)}\delta s \end{array} \right] + l_0 \\ &= \frac{k}{v_0}\left[\begin{array}{c} x_0\sin\phi \\ -R_0 a_0\left(\cos\phi-1\right) \\ -R_0\sin\phi\frac{1+\eta_0}{2+\eta_0}\delta \\ +\frac{1+\eta_0}{2+\eta_0}\delta s - \frac{1}{(1+\eta_0)(2+\eta_0)}\delta s \end{array} \right] + l_0 \end{aligned}$$

As a result, we see that all the final coordinates indeed depend on all initial coordinates in a linear fashion, and hence the relationship can be written in terms of a transfer matrix. The general shape of this matrix is now

$$\begin{pmatrix} x_f \\ a_f \\ y_f \\ b_f \\ l_f \\ \delta_f \end{pmatrix} = \begin{pmatrix} c_x & s_x & 0 & 0 & 0 & * \\ c'_x & s'_x & 0 & 0 & 0 & * \\ 0 & 0 & 1 & L & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ * & * & 0 & 0 & 1 & * \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_i \\ a_i \\ y_i \\ b_i \\ l_i \\ \delta_i \end{pmatrix} \tag{129}$$

where we have not explicitly filled in the ”*” terms because of space restrictions.

E. The Inhomogeneous Sector

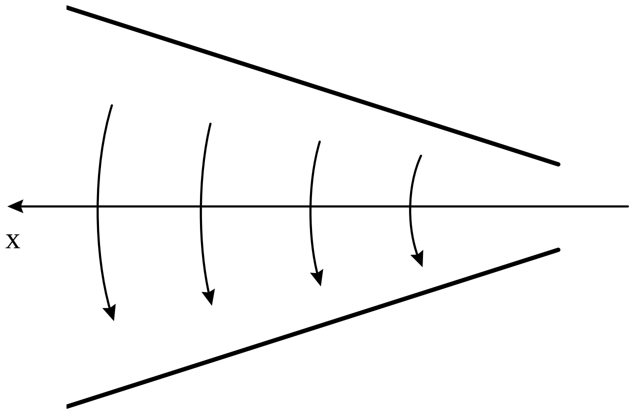
In the case of an inhomogeneous sector, there is a magnetic field that is constant in s-direction, but not constant in the x-direction; it has the shape

$$B_y = B_{y0}\left(1 + n\frac{x}{R_0}\right) \tag{130}$$

From the recursion relations for the fields, we infer that the corresponding horizontal field is

$$B_x = B_{y0} n \frac{y}{R_0} \tag{131}$$

In general terms, such a field is obtained by changing the distance between what generates the fields (coils or iron) as a function of x,similar to what is shown below for the case of $n < 0$.



weaker field

The linearized equation of motion has the form:

$$
\begin{aligned}
x^{'} &= a \\
a^{'} &= -x\left(h^2 + \frac{B_{y0}}{R_0\chi_m}n\right) + \delta h \frac{1+\eta_0}{2+\eta_0} \\
y^{'} &= b \\
b^{'} &= \frac{B_{y0}}{R_0\chi_m}ny \\
l^{'} &= \frac{k}{v_0}\left\{hx - \frac{1}{(1+\eta_0)(2+\eta_0)}\delta\right\} \\
\delta^{'} &= 0 \qquad (132)
\end{aligned}
$$

We observe that the horizontal motion is similar to the case of the homogeneous dipole, except that the strength of focusing now also depends on n, the field inhomogeneity. Different from the homogeneous dipole, there is now an effect in the vertical direction, which can be either focusing or defocusing, depending on the sign of n.

The solution of these equations of motion proceeds in the same way as before, first solve the homogeneous system, then address the inhomogeneity arising from δ via variation of parameters, and finally solve for l by a mere integration. In horizontal and vertical directions, the homogeneous solution corresponds to harmonic oscillators with frequencies

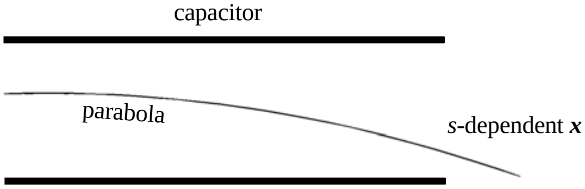
$$\omega_x = h\sqrt{1+n}; \omega_y = h\sqrt{-n}. \qquad (133)$$

An interesting case occurs for $n = -1/2$, in which case the magnet focuses x and y completely identically and represents a nice equivalent of the glass lens!

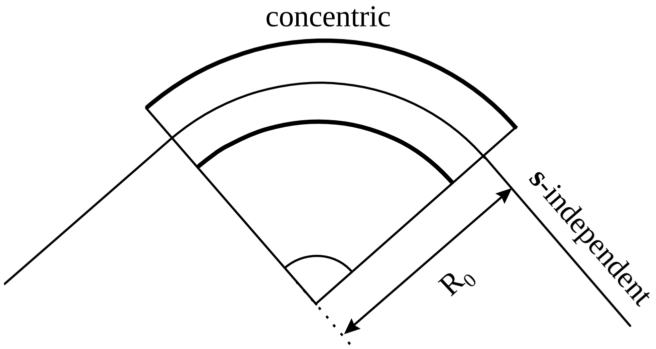
The remainder of the derivation is tedious algebra, and we will not list the details here.

F. The Inhomogeneous Electric Deflector

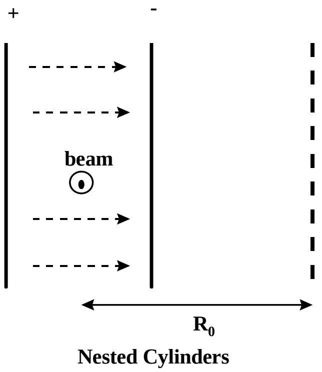
Rather commonly known is the motion of a particle in an electric capacitor; neglecting fringe fields, it follows a parabola as shown in the figure.

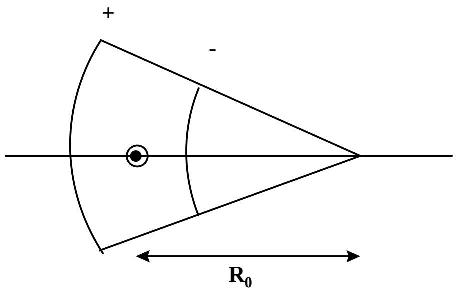


For particle optical purposes, such an arrangement is not particularly suitable for two reasons: first of all, the reference orbit has a curvature that depends on s, which makes the differential equations non-autonomous; and secondly, the potential along the reference orbit changes with s, which complicates the dynamics. Both of these problems do not appear if instead of a straight capacitor, one chooses a curved one in such a way that the reference orbit is concentric with the plates, as shown in the figure.



So far, no assumptions have been made about the vertical shapes of the electrodes, and in fact a variety of choices exist. Two common situations are the cylindrical plates and the spherical plates, as shown below.





In the case of the of the cylindrical field, we have from Gauss' law that $E \sim 1/R$, which implies the expansion

$$E_x = -E_{x0}\frac{R_0}{R_0+x} =_1 -E_{x0}\left(1-\frac{x}{R_0}\right) \qquad (134)$$

and hence corresponds to $n = -\dot{1}$ (or $n_e = -h$). In the spherical case, we have $E \sim 1/R^2$ and thus

$$E_{x_0} = -E_{x_0}\left(\frac{R_0}{R_0+x}\right)^2 =_1 -E_{x_0}\left(1-2\frac{x}{R_0}\right) \qquad (135)$$

and $n = -2$ (or $n_e = -2h$). The solution of the equations of motion is conceptually identical to the case of the inhomogeneous magnet, but practically even more involved, and we forgo it here for reasons of space.

VIII. Elementary Particle Optical Devices and Their Maps

A. The Map and its Aberrations

Recall that the transfer map of an optical system relates final coordinates to initial coordinates via

$$\vec{z}_f = \mathcal{M}(\vec{z}_i) \qquad (136)$$

where $\vec{z} = (x, a, y, b, l, \delta)$. In the previous sections, we were concerned mostly with the linearized part of the map, which describes the major part of the motion and which can be described by transfer matrices. The matrix elements were denoted as (x, a) etc.

In order to study the effects of the motion very precisely, it is necessary to also consider higher order or nonlinear effects. For this purpose we Taylor expand the map (in a rigorous sense the question whether the map can actually be Taylor expanded is rather nontrivial, but we ignore this here), and use names for the coefficients similar to what we had for the linear motion. We write

$$\begin{aligned}
x_f &= \sum \left(x|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}\right) x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}\\
a_f &= \sum \left(a|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}\right) x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}\\
y_f &= \sum \left(y|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}\right) x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}\\
b_f &= \sum \left(b|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}\right) x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}\\
l_f &= \sum \left(l|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}\right) x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}\\
\delta_f &= \sum \left(\delta|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}\right) x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta} \qquad (137)
\end{aligned}$$

where the sums go over all six-tuples $(i_x, i_a, i_y, i_b, i_l, i\delta)$; for convenience, they are usually sorted by total order. The Taylor coefficients belonging to terms of orders two or higher are usually called **aberrations,** as they describe corrections to the linear part of the map that are usually small if the phase space variables are small.

B. Symmetries of the Map

In most cases, the freedom of the aberration coefficients is severely restricted by the presence of a variety of symmetries. First, in many cases the motion of one of the variables does not depend on the values of some other variables. For example, if the motion is time independent, we have

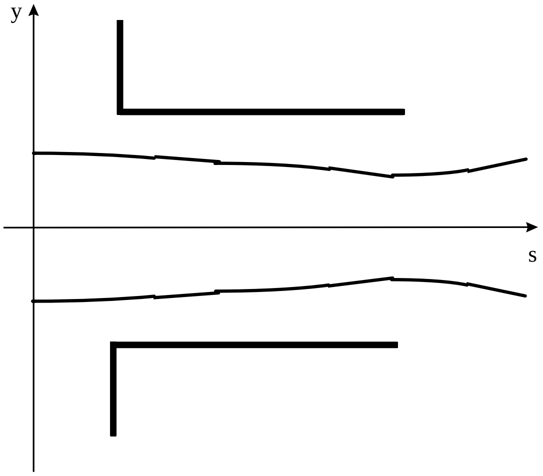
$$(*|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i\delta}) = 0 \text{ if } i_l \neq 0. \qquad (138)$$

Furthermore, in this case we know that the kinetic plus potential energy of the particle is conserved, and we have that

$$(\delta|...) = 0 \text{ except } (\delta|\delta) = 1. \qquad (139)$$

1. Horizontal Midplane Symmetry

This symmetry is perhaps the most important symmetry in Beam Physics, as it affects almost all devices: quadrupoles, higher multipoles, bending elements, cyclotrons, glass lenses, etc have it, and so do all their combinations. It says that motion is always symmetric around the midplane (the $x-z$ plane), and thus behaves like illustrated in the picture:



So for any orbit, going to the mirror image along the $x-z$ plane gives another valid orbit. Considering the phase space vector (x,a,y,b,l,δ), the mirror image is $(x,a,-y,-b,l,\delta)$. Thus flipping the signs of y_i, b_i simultaneously flips the signs of y_f, b_f , but leaves x_f, a_f, l_f, δ_f intact. Flipping sign of y_i, b_i simultaneously produces a sign of $(-1)^{i_y+i_b}$ in each monomial. So i_y+i_b must be odd for y_f, b_f and i_y+i_b must be even for all others. So we obtain

$$\begin{aligned}
\left(x|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i_\delta}\right) &= 0 \text{ for } i_y+i_b \text{ odd}\\
\left(a|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i_\delta}\right) &= 0 \text{ for } i_y+i_b \text{ odd}\\
\left(y|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i_\delta}\right) &= 0 \text{ for } i_y+i_b \text{ even}\\
\left(b|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i_\delta}\right) &= 0 \text{ for } i_y+i_b \text{ even}\\
\left(l|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i_\delta}\right) &= 0 \text{ for } i_y+i_b \text{ odd}\\
\left(\delta|x^{i_x}a^{i_a}y^{i_y}b^{i_b}l^{i_l}\delta^{i_\delta}\right) &= 0 \text{ for } i_y+i_b \text{ odd}
\end{aligned} \tag{140}$$

For the first order matrix, this requires the special form

$$M = \begin{pmatrix} * & * & 0 & 0 & * & * \\ * & * & 0 & 0 & * & * \\ 0 & 0 & * & * & 0 & 0 \\ 0 & 0 & * & * & 0 & 0 \\ * & * & 0 & 0 & * & * \\ * & * & 0 & 0 & * & * \end{pmatrix} \tag{141}$$

which is of compatible with what we saw before for the transfer matrices we encountered. Altogether, the symmetry entails that to any given order, roughly half of all aberrations are gone.

2. Double Midplane Symmetry

Several devices have a midplane symmetry not only around the horizontal plane, but also around a vertical plane. This is the case for all electric cylindrically symmetric devices, as well as quadrupoles, octupoles, and in general $4k$ poles. In this case, in addition to the requirements we just had, we obtain a second set in which the roles of x, a and y, b are interchanged. In this case we obtain

$$\begin{aligned}
(x|...) &= 0 \text{ for } i_y+i_b \text{ odd or } i_x+i_a \text{ even}\\
(a|...) &= 0 \text{ for } i_y+i_b \text{ odd or } i_x+i_a \text{ even}\\
(y|...) &= 0 \text{ for } i_y+i_b \text{ even or } i_x+i_a \text{ odd}\\
(b|...) &= 0 \text{ for } i_y+i_b \text{ even or } i_x+i_a \text{ odd}\\
(l|...) &= 0 \text{ for } i_y+i_b \text{ odd or } i_x+i_a \text{ even}\\
(\delta|...) &= 0 \text{ for } i_y+i_b \text{ odd or } i_x+i_a \text{ even}
\end{aligned} \tag{142}$$

and altogether, about 3/4 of all matrix elements vanish. To first order, the matrix must have the special form

$$M = \begin{pmatrix} * & * & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & * & 0 & 0 \\ 0 & 0 & * & * & 0 & 0 \\ 0 & 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & 0 & * & * \end{pmatrix} \tag{143}$$

which is what we observed in the case of the drift and the electric and magnetic quadrupoles.

3. Rotational Symmetry

One special case of the double midplane symmetry that we just discussed is the full rotational symmetry that round lenses satisfy. In this case there is a symmetry going beyond what double midplane symmetry requires, the map has to be invariant under a rotation in the $x-y$ plane. Let the rotation angle be ϕ and denote $\cos\phi$ by c and $\cos\phi$ by s. Then the matrix describing such a rotation of the particle optical variables is

$$\mathcal{R} = \begin{pmatrix} c & 0 & s & 0 & 0 & 0 \\ 0 & c & 0 & s & 0 & 0 \\ -s & 0 & c & 0 & 0 & 0 \\ 0 & -s & 0 & c & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{144}$$

and we must have that the transfer map satisfies

$$\mathcal{M} \circ \mathcal{R} = \mathcal{R} \circ \mathcal{M} \tag{145}$$

In the variables we are currently using, the study of the influence of the rotation on the map is somewhat cumbersome, and for this purpose it is actually better to choose complex coordinates

$$\begin{aligned} z &= x + iy \\ w &= a + ib \end{aligned} \tag{146}$$

as well as their complex conjugates

$$\begin{aligned} \bar{z} &= x - iy \\ \bar{w} &= a - ib. \end{aligned} \tag{147}$$

In these complex variables, the map $\mathcal{R}$ has the simple diagonal form

$$\mathcal{R} = \begin{pmatrix} e^{i\phi} & 0 & 0 & 0 & 0 & 0 \\ 0 & e^{i\phi} & 0 & 0 & 0 & 0 \\ 0 & 0 & e^{-i\phi} & 0 & 0 & 0 \\ 0 & 0 & 0 & e^{-i\phi} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{148}$$

and its effect in equation (145) is easy to study. It turns out that the in the map, only those terms that have the form

$$\begin{aligned} z_f &= z_i \cdot f_z(z\bar{z}, w\bar{w}) \\ w_f &= w_i \cdot f_w(z\bar{z}, w\bar{w}) \end{aligned} \tag{149}$$

are allowed to remain. For the expert reader, we note that this situation is remarkably similar to what happens in the theory of normal forms of repetitive motion.

4. Symplectic Symmetry

Another important symmetry of the motion is due to the fact that the motion is indeed obtained by solution of a Hamiltonian problem. In this case, one can show that the Jacobian $\tilde{M}$ of the transfer map $\mathcal{M}$, i.e. the matrix

$$\tilde{M} = \begin{pmatrix} \frac{\partial \mathcal{M}_1}{\partial z_1} & \frac{\partial \mathcal{M}_1}{\partial z_2} & \frac{\partial \mathcal{M}_1}{\partial z_3} & \frac{\partial \mathcal{M}_1}{\partial z_4} & \frac{\partial \mathcal{M}_1}{\partial z_5} & \frac{\partial \mathcal{M}_1}{\partial z_6} \\ \frac{\partial \mathcal{M}_2}{\partial z_1} & \frac{\partial \mathcal{M}_2}{\partial z_2} & \frac{\partial \mathcal{M}_2}{\partial z_3} & \frac{\partial \mathcal{M}_2}{\partial z_4} & \frac{\partial \mathcal{M}_2}{\partial z_5} & \frac{\partial \mathcal{M}_2}{\partial z_6} \\ \frac{\partial \mathcal{M}_3}{\partial z_1} & \frac{\partial \mathcal{M}_3}{\partial z_2} & \frac{\partial \mathcal{M}_3}{\partial z_3} & \frac{\partial \mathcal{M}_3}{\partial z_4} & \frac{\partial \mathcal{M}_3}{\partial z_5} & \frac{\partial \mathcal{M}_3}{\partial z_6} \\ \frac{\partial \mathcal{M}_4}{\partial z_1} & \frac{\partial \mathcal{M}_4}{\partial z_2} & \frac{\partial \mathcal{M}_4}{\partial z_3} & \frac{\partial \mathcal{M}_4}{\partial z_4} & \frac{\partial \mathcal{M}_4}{\partial z_5} & \frac{\partial \mathcal{M}_4}{\partial z_6} \\ \frac{\partial \mathcal{M}_5}{\partial z_1} & \frac{\partial \mathcal{M}_5}{\partial z_2} & \frac{\partial \mathcal{M}_5}{\partial z_3} & \frac{\partial \mathcal{M}_5}{\partial z_4} & \frac{\partial \mathcal{M}_5}{\partial z_5} & \frac{\partial \mathcal{M}_5}{\partial z_6} \\ \frac{\partial \mathcal{M}_6}{\partial z_1} & \frac{\partial \mathcal{M}_6}{\partial z_2} & \frac{\partial \mathcal{M}_6}{\partial z_3} & \frac{\partial \mathcal{M}_6}{\partial z_4} & \frac{\partial \mathcal{M}_6}{\partial z_5} & \frac{\partial \mathcal{M}_6}{\partial z_6} \end{pmatrix} \tag{150}$$

has to satisfy the condition

$$\tilde{M} \cdot \hat{J} \cdot \tilde{M}^t = \hat{J}, \tag{151}$$

where $\hat{J}$ is the totally antisymmetric matrix

$$\hat{J} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}. \qquad (152)$$

The proof of this so-called condition of symplecticity certainly goes beyond these lectures. But we can appreciate that the symplectic condition, which mixes in a very defined way the terms $\partial\mathcal{M}_i/\partial z_j$ that are themselves power series, entails a large variety of nonlinear restrictions between the aberrations.

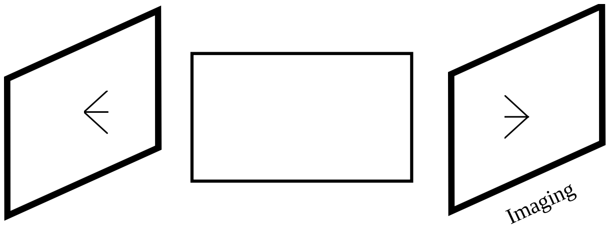
The detailed study of these is cumbersome and can be found in the literature (for example H. Wollnik and M. Berz, NIM238 (1985) p.127), and we want to restrict our attention to what happens in the linear case. Considering the constant part of the symplectic condition (151), we observe that what contributes via the Jacobian is just the transfer map. Assuming midplane symmetry and no acceleration, plugging into the symplectic condition yields the conditions

$$\begin{aligned} (x|x)\cdot(a|a) - (x|a)\cdot(a|x) &= 1 \quad (153) \\ (y|y)\cdot(b|b) - (b|y)\cdot(y|b) &= 1 \quad (154) \\ (l|x)\cdot 1 + (x|x)\cdot(a|\delta) - (a|x)\cdot(x|\delta) &= 0 \quad (155) \\ (l|a)\cdot 1 + (x|a)\cdot(a|\delta) - (a|a)\cdot(x|\delta) &= 0 \quad (156) \end{aligned}$$

The first two of these are old friends and describe the fact that the volume of phase space is preserved under the linear transformations generated by particle optical elements. The third and fourth conditions, however, represent an amazing connection between longitudinal and dispersive effects.

C. Imaging Devices

In the following sections, we want to discuss what specifically has to be done to a map of a system to make the system useful for a specific task. In many cases, this will require certain matrix elements to vanish, or sometimes to assume specific values. The most important device is probably the imaging device, in which final positions are not allowed to depend on initial angles, as shown schematically in the picture.



In the case of particle optical systems, this requires

$$\begin{aligned} (x|a) &= 0, \quad (157) \\ (y|b) &= 0. \quad (158) \end{aligned}$$

On the other hand, the final angles a_f and b_f are unimportant since it doesn't matter at what angle the rays strike at the image position; so all terms of the form $(a|...)$ or $(b|...)$ are insignificant. Additional requirements usually exist for the various sub-classes of imaging systems.

Note that one important application of beam physics that does indeed produce images of some sort, namely the X-ray machine, is actually not an imaging device: X-rays cannot be bent, and therefore it is impossible to achieve $(x|a) = (y|b) = 0$; in fact, if l is the length of the device we will have $(x|a) = (y|b) = l$. So X-ray systems should be short from object to "image", and they should have a source that produces beams with a very small a and b. Anything else increases the fuzziness that X-ray pictures usually exhibit.

1. The Television Tube

As far as practical use, impact on society, and revenues is concerned, sadly enough the TV tube (and more generally and not so sadly enough, the

CRT tube) is the most important application of particle optics. In this case, for each color an electron beam is deflected vertically and horizontally by two simple magnetic deflectors in order to sweep over the screen area, and its intensity is adjusted according to the color saturation at the respective point.

At any given point on the screen, the resulting spot should not be wider than about the distance between two pixels, so whatever size the beam had initially should not be amplified very much; so

$$(x|x) \text{ and } (y|y)$$

should not be large.

The requirements for aberrations are mostly somewhat benign as the beam phase space volume is small.

2. The Camera, the Electron Microscope

The purpose of a camera and an electron microscope is to create an image of an object through which the rays is moving. So the quantities

$$(x|x) \text{ and } (y|y) \text{ are magnifications,}$$

and in most cases it is desirable to have them equal. The electron microscope is just a special case in which both of these are made to be very large.

If a true image is desired, it is important that the relationship between final and initial coordinates be really linear, which requires that all higher order position dependent matrix elements vanish, and so

$$(x|xx) = 0, (y|yy) = 0, (x|xxx) = 0, ...$$

In reality, of course, it is sometimes difficult to do this to higher orders, and some distortions prevail. In the case of an electron microscope, this is often not detrimental as one can retroactively correct the effects by calculation. The effects that appear usually have the consequence that rectangles are distorted into either the shape of pincushions or into the shape of barrels; these effects are due to

$$(x|xyy) \text{ and } (y|yxx)$$

which entail that rays that simultaneously have x and y coordinates are either pushed out from the center (pincushion) or pulled in (barrel). Higher order terms in x and y produce similar effects.

There should no effect of energy on position, so

$$(x|\delta) = 0 \text{ and } (y|\delta) = 0$$

should be maintained. Similarly, all higher order aberrations involving δ should vanish; if this is not the case, some color-dependent blurring may occur, in particular for larger values of x and y, an effect that can be easily observed in the case of less expensive binoculars.

There should also be no effects of position on initial angles to higher order; so it is necessary that

$$(x|a^{i_a}b^{i_b}) = (y|a^{i_a}b^{i_b}) = 0, \tag{159}$$

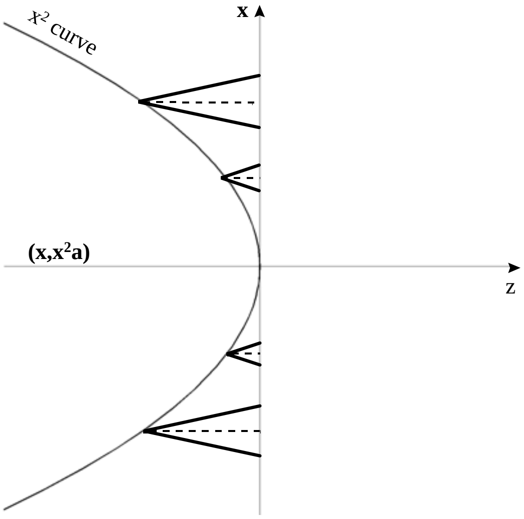
and since often the range of accepted angles corresponding to a and b is rather large, to correct these terms is often very important. If any of them prevail, they will entail a color-independent fuzziness; in case the order of the coordinates a and b is even, the fuzz will be oriented towards one side like the coma of a comet; if the powers are odd, it will lead to a uniformly distributed fuzziness.

Similarly, all aberrations involving positions and angles simultaneously should vanish, and hence it is necessary to have

$$(x|x^{i_x}y^{i_y}a^{i_a}b^{i_b}) = (y|x^{i_x}y^{i_y}a^{i_a}b^{i_b}) = 0; \tag{160}$$

if any of them prevail, they will entail a position-dependent fuzziness that becomes stronger with an increase of the positions x and y.

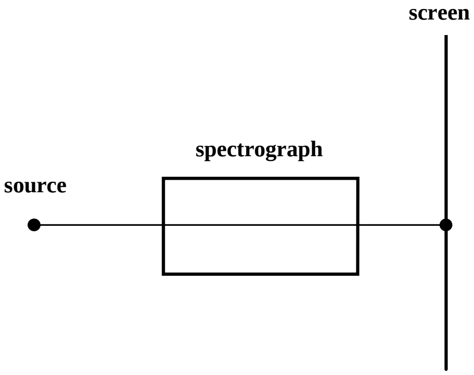
Interestingly enough, all higher order aberrations depending on a and b only linearly can be corrected by a re-shaping of the focal plane; in fact, $(x|xa)$ etc produce a tilt of the image, $(x|xxa)$ etc produce a curvature of the image. The image shows how the matrix element $(x|xxa)$ can be corrected by shaping the image position parabolically:



At any given position, due to the matrix element $(x|xxa)$, any ray with a given a is moved up or down in proportion to a, where the amount of deflection depends quadratically on x; so the rays arrive at the x plane as shown. However, tracing the rays backwards shows that they in fact all intersect before the plane, and the point where this happens depends quadratically on x. In similar ways, $(x|x^4a)$ etc can be corrected.

3. The Spectrograph

As we learned before, the purpose of the spectrograph is to translate energy information into position information, and in order to have high resolution, the position should not depend on anything else if possible. Rays originate from a source, travel through the spectrograph, and finally reach the screen, as shown in the picture.



It is possible to measure energies in terms of final positions by making

$$(x|\delta) \text{ large.} \tag{161}$$

In practice this requires the use of at least one bending element, because all other elements have vanishing $(x|\delta)$. The final position should not depend on anything else besides δ, and since it is important to be able to accept rays covering a wide range of angles, it is necessary to have

$$(x|a) = (y|b) = 0. \tag{162}$$

So the spot size is limited by $(x|x) = 1/(a|a)$, which is usually kept small, and the size of the object, the x-size of which is usually kept in the range of fractions of mm.

Any contribution to the final position should be due to energy, and so aberrations depending on initial angle should be avoided; so if possible, we want

$$\begin{aligned}(x|aa) &= (x|aaa) = 0\\ (x|bb) &= (x|abb) = 0\end{aligned}$$

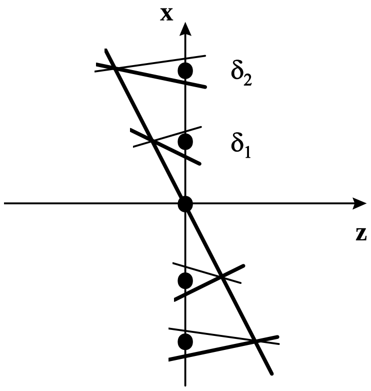
The aberrations involving also x-positions are less significant as positions are kept small. The ones involving also y positions are more important as y is not necessarily kept small; but if

$(y|y)$ is kept large enough, particles with significant initial y reach the focal plane with significant final y; the interplay of $(y|y)$ and $(x|yy)$ then leads to a parabolic shape of the resulting image, but the sharpness of the parabola, which determines the resolution, is unaffected by $(x|yy)$.

It is also important to consider aberrations involving energy. Of these, the terms depending only on energy like

$$(x|\delta^{i\delta}) \tag{163}$$

do not necessarily have to be corrected as long as they are known, as they just turn the relationship of final x and initial δ into a nonlinear one, which still allows an accurate measurement of δ. The most important aberrations are usually those that involve initial angles and energies simultaneously, as both of these can be large. Of these, the lowest order aberration $(x|a\delta)$ can be corrected by just tilting the focal plane: the final x of a particle, which depends mostly on δ, is moved up or down linearly depending on the value of a. As shown in the figure, similar to before, all these rays with different values of a go through a common point at a distance before or after the x plane, where the effect of $(x|a\delta)$ does not manifest itself.



In a similar way, spectrographs can also be used to measure masses of particles, and all previous arguments stay valid if the energy deviation δ is replaced by the mass deviation δ_m. If mass resolution is to be achieved to very high precision and the initial energy is not uniform, then in addition to the above requirements, it is also important that the final position does not depend on δ; this requires that

$$(x|\delta) = 0 \tag{164}$$

while of course at the same time trying to have

$$(x|\delta_m) \text{ large.} \tag{165}$$

The simultaneous satisfaction of these conditions is not possible using only magnetic devices; for low energies, it is usually achieved by combining magnetic and electric deflectors.

4. Telescopic Systems

Telescopic systems in particle optics are not as important as in the case of glass optics, as there is no direct equivalent to the human eye that directly uses angular information which may benefit from magnification.

D. Point-to-Parallel and Parallel-to-Point

There are several situations which require these systems; perhaps the most important is the transport of a beam over a long distance without blowing up its size very much.

1. The Beam Expander, the SDI Gun

If a beam travels over a long distance, its final size is governed by the angular divergence in the beam, and so it is necessary to make this angular divergence as small as possible. Since phase space volume is preserved, this seemingly paradoxically requires making the initial width of the beam large.

If the beam now originates from a source of small size, but under a variety of angles, than a point-to-parallel system will be able to turn this into a beam with rather small final angles, since in a telescopic system,

$$(a|a) = 0 \tag{166}$$

and $(a|x)$ is of reduced importance because of the initially small size. Depending on the emittance of the beam, aberrations are often also important, and we have to make sure that

$$(a|aa), (a|aaa), (a|abb) \text{ etc vanish}$$

or are at least small enough.

Perhaps the most famous point-to-parallel system (or depending on your viewpoint the most infamous) is the "Star Wars" or SDI device that was supposed to shoot down incoming missiles in space. In this case, the beam has to travel a very long distance (from the cannon to the missile) without too much increase in size, and this is best achieved by initially making it wide with a point-to-parallel system.

But also in regular applications, these systems are useful as they conveniently allow transport over larger distances.

2. The Final Focusing Section

Perhaps the opposite of the point-to-parallel system is the parallel-to-point system, in which an initially nearly parallel beam is brought down to a small point. This is indeed of prime importance in the case of the collider, where the number of collisions of the counter-rotating beams increases as the radii of the beams decrease.

But since the collision point is usually deep inside a detector, the necessary large angle requires a rather wide beam in the last focusing element, which because of its width at this point is nearly parallel there.

So the position at the final focus is not allowed to depend on the large position before the last focusing element, and so it is necessary to have

$$(x|x) = (y|y) = 0 \tag{167}$$

and hence a parallel-to-point system. Depending on the needed size of the focal point (which can be less than a micrometer), it is also very important to correct the position-dependent aberrations, in particular we must have that

$$(x, xx), (x, xxx), (x, xyy) \text{ etc vanish.}$$

E. The Periodic Transport

In the case of the periodic transport over long distances, the desire is not so much to give a special shape to the beam as the beam exits, but even much more simply, to just contain the beam. This is of key importance in all devices in which the beam repeatedly passes through the same (or a very similar) structure. We may wonder whether this again translates into the requirement that a certain matrix element vanish, but as we shall see, this is not quite the case.

Actually it is rather easy to formulate a necessary condition on the linear matrix: it is not allowed to have any eigenvalue of magnitude greater than unity. If the eigenvalue is real, the argument is simple: if this were the case, any particle that has its coordinates lined up with the corresponding real eigenvector will after one period end up on the same line, but all its coordinates would have increased by a factor equal to the eigenvalue.

If on the other hand the eigenvalue is complex, there is another eigenvalue that is conjugate and hence has the same magnitude. Similar to the eigenvalues, also the eigenvectors are conjugates of each other. Now simply consider the sum of the two eigenvectors, which is real; sending this sum through the matrix multiplies the first eigenvector by the first eigenvalue, and the second one by the conjugate, resulting in a sum that is again real and increased in size by the magnitudes of the eigenvalues.

In both cases, coordinates grow exponentially in time, and so eigenvalues that are even only a tiny amount above unity in magnitude are detrimental. Of course the nonlinear effects also influence the motion and break the purely exponential pattern, but all experience shows that it is not possible to correct linear instability with nonlinear means; in practice, things usually work quite to the contrary!

Because of emittance preservation, the fact that eigenvalues greater than unity are prohibited means that in fact, all eigenvalues have to have unit magnitude. Of these, the cases+1 and -1 are to be excluded too, since even the slightest imperfection in the machine may otherwise lead to instability. Altogether, in a periodic system, the eigenvalues must all be complex and of unit magnitude.

It is particularly interesting to study the special case of a matrix with midplane symmetry. In this case, the x and y motion decouple and can be described by individual matrices. We obtain for the eigenvalues for the two-by-two x submatrix, noting that the y submatrix is treated similarly:

$$\begin{aligned} 0 &= \left|\widehat{M}-\lambda\widehat{I}\right| \\ &= \begin{vmatrix} (x|x)-\lambda & (x|a) \\ (a|x) & (a|a)-\lambda \end{vmatrix} \\ &= \underbrace{(x|x)\,(a|a)-(x|a)\,(a|x)}_{1} \\ &\quad -\lambda\left[(x|x)+(a|a)\right]+\lambda^2 \end{aligned}$$

and so

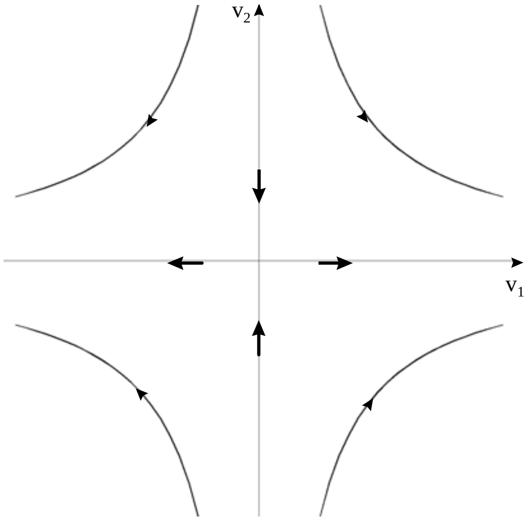
$$\begin{aligned} \lambda_{1,2} &= \frac{[(x|x)+(a|a)]\pm\sqrt{[(x|x)+(a|a)]^2-4}}{2} \\ &= \frac{tr\widehat{M}}{2}\pm\sqrt{\left(\frac{tr\widehat{M}}{2}\right)^2-1} \end{aligned} \qquad (168)$$

Hence to have complex eigenvalues requires the very simple condition

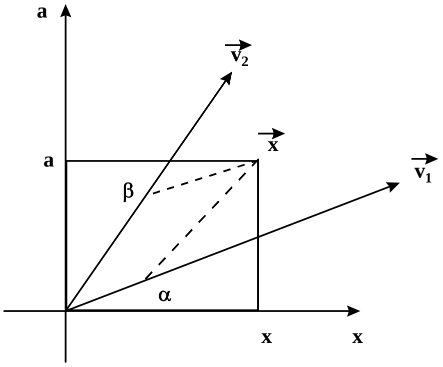
$$-2 < tr(M) < 2 \qquad (169)$$

A quick check of the four cases shows that this excludes the point-to-point case and the parallel-to-parallel case, as in both of these, the trace just equals two or exceeds two. The parallel-to-point or point-to-parallel case each have one element on the diagonal vanish, so they are permissible if the remaining diagonal matrix element is less than two in magnitude.

We also verify that for $tr\widehat{M}\neq 2$, the eigenvalues form a reciprocal pair, i.e. $\lambda_1\lambda_2=1$. Let us quickly re-visit the case $\left|tr\widehat{M}\right|>2$, for which the eigenvalues are real, and hence one of them is greater than unity, and as we had already concluded, the motion is unstable. Choosing a new basis, the so-called normal form basis, along the real eigenvectors $\vec{v}_1,\vec{v}_2$, we have that the repetitive motion asymptotically approaches the eigenvector $\vec{v}_1$ with eigenvalue greater than unity and becomes larger and larger.



A detailed analysis shows that the motion indeed follows hyperbolas; note that $\lambda_1\lambda_2=1$, and that $|\lambda_1|>1>|\lambda_2|$. Suppose we have a general vector expressed in the basis $(\vec{v}_1, \vec{v}_2)$ as shown in the figure



whose coordinates are now α and β, and thus

$$\vec{x} \equiv \begin{pmatrix} x \\ a \end{pmatrix} = \alpha\vec{v}_1 + \beta\vec{v}_2$$

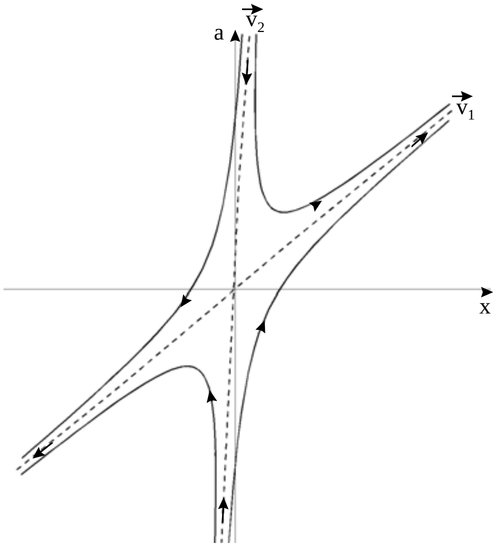
Applying the transfer matrix, we have

$$\begin{aligned} \widehat{M}\vec{x} &= \alpha\widehat{M}\vec{v}_1 + \beta\widehat{M}\vec{v}_2 \\ &= \alpha\lambda_1\vec{v}_1 + \beta\lambda_2\vec{v}_2 \end{aligned}$$

In normal form coordinates, the action of the transfer map is thus given by

$$\begin{pmatrix} \alpha\lambda_1 \\ \beta\lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

but since $\lambda_2 = 1/\lambda_1$, the product of the coordinates stays constant, characteristic of the motion along a hyperbola. In Cartesian coordinates, the motion looks more complicated as the hyperbolic structure is deformed:

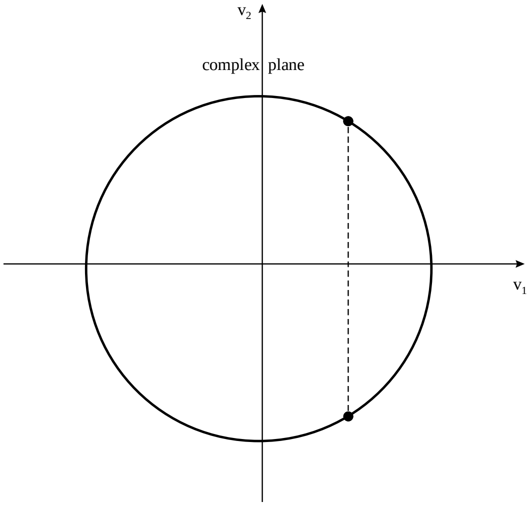


For practical purposes, this case is unstable and hence useless.

Let us now consider the case $\left|tr\widehat{M}\right| < 2$ in more detail. We have the complex eigenvalues that satisfy

$$\lambda_2 = \overline{\lambda}_1 \text{ and } \lambda_2 = \lambda_1^{-1}. \qquad (170)$$

So in the complex plane, λ_1 and λ_2 lie on a circle and form conjugate pairs, as shown in the figure:



The eigenvalues can hence be written as

$$\lambda_{1,2} = e^{\pm i\mu}, \qquad (171)$$

where μ is called the **tune** of the system. The eigenvectors $\vec{v}_{1,2}$ belonging to $\lambda_{1,2}$ also form conjugate pairs, since

$$\widehat{M}\overline{\vec{v}_2} = \overline{\widehat{M}\vec{v}_2} = \overline{\lambda_2\vec{v}_2} = \lambda_1\overline{\vec{v}_2}$$

Define now two new basis vectors $\vec{v}_+ = \mathrm{Re}\left(\vec{v}_1\right)$, $\vec{v}_- = \mathrm{Im}\left(\vec{v}_1\right)$ as the real and imaginary parts of the eigenvalues; they define what is called the normal form basis for stable motion. So we have

$$\begin{aligned}\vec{v}_1 &= \vec{v}_+ + i\vec{v}_- \\ \vec{v}_2 &= \vec{v}_+ - i\vec{v}_-. \end{aligned} \tag{172}$$

We now observe

$$\begin{aligned}\widehat{M}\vec{v}_1 &= \lambda_1\vec{v}_1 = e^{i\mu}\left(\vec{v}_+ + i\vec{v}_-\right) \\ &= \cos\mu\cdot\vec{v}_+ - \sin\mu\cdot\vec{v}_- \\ &\quad +i\left(\sin\mu\cdot\vec{v}_+ + \cos\mu\cdot\vec{v}_-\right)\end{aligned}$$

and similarly

$$\begin{aligned}\widehat{M}\vec{v}_2 &= \lambda_2\vec{v}_2 = e^{-i\mu}\left(\vec{v}_+ - i\vec{v}_-\right) \\ &= \cos\mu\cdot\vec{v}_+ - \sin\mu\cdot\vec{v}_- \\ &\quad -i\left(\sin\mu\cdot\vec{v}_+ + \cos\mu\cdot\vec{v}_-\right).\end{aligned}$$

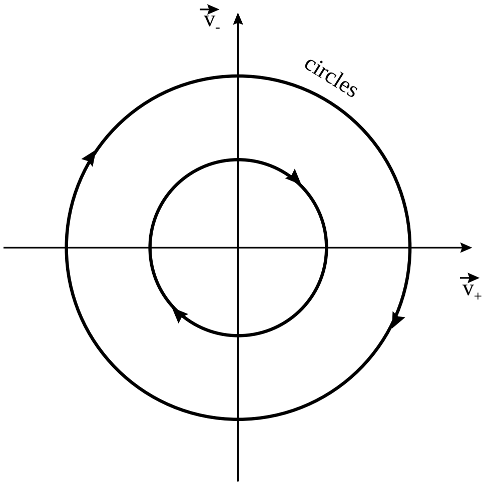
Now assume we have a general vector expressed in the basis vectors $\vec{v}_\pm$ with coefficients α and β, i.e. $\vec{x} = \alpha\vec{v}_+ + \beta\vec{v}_-$. Then we have

$$\begin{aligned}\widehat{M}\vec{x} &= \alpha\widehat{M}\vec{v}_+ + \beta\widehat{M}\vec{v}_- \\ &= \alpha\widehat{M}\frac{\vec{v}_1+\vec{v}_2}{2} + \beta\widehat{M}\frac{\vec{v}_1-\vec{v}_2}{2i} \\ &= \alpha\frac{\widehat{M}\vec{v}_1+\widehat{M}\vec{v}_2}{2} + \beta\frac{\widehat{M}\vec{v}_1-\widehat{M}\vec{v}_2}{2i} \\ &= \alpha\left(\cos\mu\cdot\vec{v}_+ - \sin\mu\cdot\vec{v}_-\right) \\ &\quad +\beta\left(\sin\mu\cdot\vec{v}_+ + \cos\mu\cdot\vec{v}_-\right) \\ &= \left(\alpha\cos\mu + \beta\sin\mu\right)\vec{v}_+ \\ &\quad +\left(-\alpha\sin\mu + \beta\cos\mu\right)\vec{v}_-\end{aligned}$$

So altogether, in normal form coordinates, we have

$$\begin{aligned}\hat{M}\begin{pmatrix}\alpha\\ \beta\end{pmatrix} &= \begin{pmatrix}\alpha\cos\mu + \beta\sin\mu \\ -\alpha\sin\mu + \beta\cos\mu\end{pmatrix} \\ &= \begin{pmatrix}\cos\mu & \sin\mu \\ -\sin\mu & \cos\mu\end{pmatrix}\begin{pmatrix}\alpha\\ \beta\end{pmatrix},\end{aligned}$$

and thus the transformation $\hat{M}$ simply performs a rotation!



The angle of the rotation in normal form coordinates is simply equal to the tune μ; and no wonder the motion is stable. To obtain the motion in the original Cartesian coordinates, we have to subject the circles to a linear transformation, which turns them into ellipses; so the motion looks as follows. The angle by which particles move in the original x, a coordinates is not necessarily μ anymore; but we can conclude that indeed if we look at the average angle advance over many turns, then this average converges to the tune μ,as at least the number of full revolutions that were experienced must agree in both coordinate systems.

It is also very illuminating to see what happens if the system is subjected to some small errors, which in reality of course always appear. If the eigenvalues were far enough from unity, even

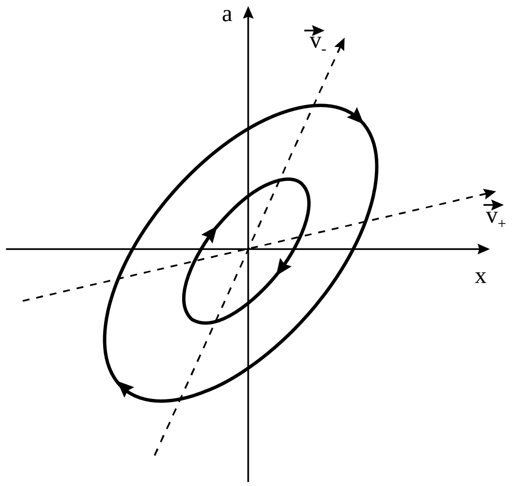
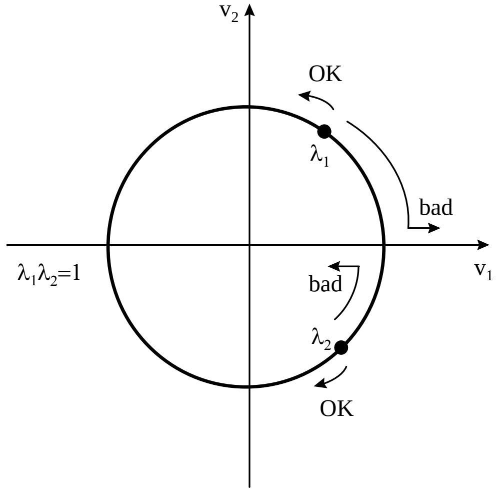


Figure 3:

under small errors we still have $\lambda_1 = \overline{\lambda}_2, \lambda_1 = \lambda_2^{-1}$, and while the tune μ may have changed a little, the qualitative behavior of stability is totally unaffected. So as long as we maintain that $\left|\left(tr\widehat{M}\right)/2\right| < 1$ is maintained, stability prevails. If on the other hand the perturbation is so large that this is violated, the perturbation leads to the loss of stability.



For the sake of completeness, let us also consider the case of $\left|tr\widehat{M}\right| = 2$. In this case, $\lambda_{1,2} = 1$,and

$$\widehat{M} = \pm\widehat{I}.$$

This is in principle stable forever; but under the slightest perturbation, there is danger of becoming unstable, and hence this case is practically useless.

1. The Invariant Ellipse

For many practical purposes it is particularly important to know in detail the parameters of the ellipse that is invariant under stable linear motion. For this purpose, let $\lambda_{1,2} = e^{\pm i\mu}$, and choose the sign of the tune μ such that $sign\,(\mu) = sign\,((x|a))$. We then define three parameters $\alpha_i, \beta_i, \gamma_i$ as

$$\begin{aligned} \alpha_i &= \frac{(x|x) - (a|a)}{2\sin\mu_i} \\ \beta_i &= \frac{(x|a)}{\sin\mu_i} \\ \gamma_i &= -\frac{(a|x)}{\sin\mu_i} \end{aligned} \qquad (173)$$

As we shall prove now, these three parameters describe the invariant ellipse via

$$\begin{pmatrix} x \\ a \end{pmatrix}^t \cdot \begin{pmatrix} \gamma_i & \alpha_i \\ \alpha_i & \beta_i \end{pmatrix} \cdot \begin{pmatrix} x \\ a \end{pmatrix} = 1, \qquad (174)$$

where the matrix describing the ellipse is called $\hat{T}$. To prove that $\hat{T}$ is actually invariant, we first express the transfer matrix in terms of the parameters. To this end, we observe that since

$$\lambda_{1,2} = \frac{tr\widehat{M}}{2} \pm \sqrt{\left(\frac{tr\widehat{M}}{2}\right)^2 - 1}$$

we have that

$$\begin{aligned} (x|x) + (a|a) &= tr\widehat{M} = \lambda_1 + \lambda_2 \\ &= e^{i\mu} + e^{-i\mu} = 2\cos\mu \end{aligned}$$

From the definition of α_i, we have $(x|x)-(a|a) = 2\sin\mu_i \cdot \alpha_i$, and hence

$$\begin{aligned}(x|x) &= \cos\mu_i + \alpha_i \sin\mu_i \\ (a|a) &= \cos\mu_i - \alpha_i \sin\mu_i\end{aligned}$$

On the other hand, from the definitions of β_i, γ_i , we have

$$\begin{aligned}(x|a) &= \beta_i \sin\mu_i, \\ (a|x) &= -\gamma_i \sin\mu_i,\end{aligned}$$

and so altogether

$$\widehat{M} = \begin{pmatrix} \cos\mu_i + \alpha_i \sin\mu_i & \beta_i \sin\mu_i \\ -\gamma_i \sin\mu_i & \cos\mu_i - \alpha_i \sin\mu_i \end{pmatrix} \tag{175}$$

Letting

$$\widehat{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \widehat{K} = \begin{pmatrix} \alpha_i & \beta_i \\ -\gamma_i & -\alpha_i \end{pmatrix}, \tag{176}$$

we have

$$\widehat{M} = \widehat{I}\cos\mu_i + \widehat{K}\sin\mu_i \tag{177}$$

Computing the inverse map of $\hat{M}$, we find

$$\widehat{M}^{-1} = \widehat{I}\cos\mu_i - \widehat{K}\sin\mu_i$$

where we used $\left|\hat{M}\right| = 1$, and as a consequence $\beta_i\gamma_i - \alpha_i^2 = 1$, which we infer as follows:

$$\begin{aligned} 1 &= \left|\widehat{M}\right| = (\cos\mu_i + \alpha_i \sin\mu_i)(\cos\mu_i - \alpha_i \sin\mu_i) \\ &\quad +\beta_i\gamma_i \sin^2\mu_i \\ &= \cos^2\mu_i + \left(\beta_i\gamma_i - \alpha_i^2\right)\sin^2\mu_i \\ &= 1 + \left(-1 + \beta_i\gamma_i - \alpha_i^2\right)\sin^2\mu_i; \end{aligned}$$

but since μ_i was not allowed to be zero or π because our requirement of stability, we must have $\beta_i\gamma_i - \alpha_i^2 = 1$.

We now are ready to study whether indeed the ellipse defined above is invariant under $\hat{M}$. This is the case if whenever a particle satisfies the ellipse equation

$$\begin{pmatrix} x \\ a \end{pmatrix}^t \cdot \begin{pmatrix} \gamma_i & \alpha_i \\ \alpha_i & \beta_i \end{pmatrix} \cdot \begin{pmatrix} x \\ a \end{pmatrix} = 1,$$

their image under $\hat{M}$, which is given by

$$\hat{M} \cdot \begin{pmatrix} x \\ a \end{pmatrix}, \tag{178}$$

also satisfies the ellipse equation. This means that also

$$\left\{\hat{M} \cdot \begin{pmatrix} x \\ a \end{pmatrix}\right\}^t \cdot \begin{pmatrix} \gamma_i & \alpha_i \\ \alpha_i & \beta_i \end{pmatrix} \cdot \left\{\hat{M} \cdot \begin{pmatrix} x \\ a \end{pmatrix}\right\} = 1.$$

This is the case if and only if

$$\hat{M}^t \cdot \hat{T} \cdot \hat{M} = \hat{T},$$

since every ellipse is described by a unique symmetric matrix and $\hat{M}^t \cdot \hat{T} \cdot \hat{M}$ is indeed symmetric. In order to execute the matrix multiplications necessary, we study various matrix products; let

$$\widehat{J} \overset{def}{=} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{179}$$

We then have:

$$\begin{aligned} \widehat{T}\widehat{K} &= \begin{pmatrix} \gamma_i & \alpha_i \\ \alpha_i & \beta_i \end{pmatrix} \begin{pmatrix} \alpha_i & \beta_i \\ -\gamma_i & -\alpha_i \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \widehat{J} \\ \widehat{K}^T\widehat{T} &= \widehat{K}^T\widehat{T}^T = \left(\widehat{T}\widehat{K}\right)^T \\ &= \widehat{J}^T = -\widehat{J} \\ \widehat{K}^T\widehat{J} &= \begin{pmatrix} \alpha_i & -\gamma_i \\ \beta_i & -\alpha_i \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \gamma_i & \alpha_i \\ \alpha_i & \beta_i \end{pmatrix} = \widehat{T} \end{aligned} \tag{180}$$

Now we are ready to compute the product $\left(\widehat{M}\right)^T \cdot \widehat{T} \cdot \widehat{M}$. We obtain

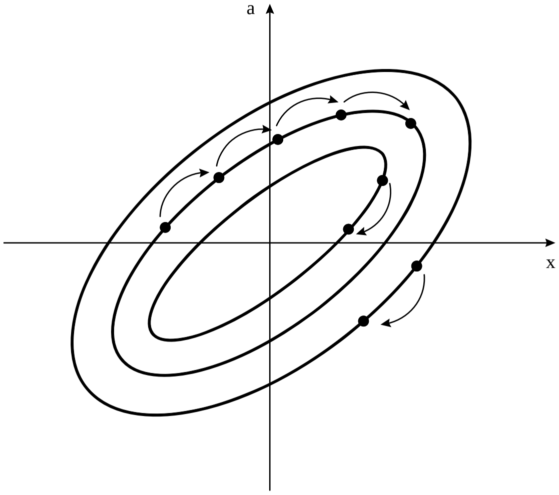


Figure 4:

$$
\begin{aligned}
& \left(\widehat{M}\right)^T \cdot \widehat{T} \cdot \widehat{M} \\
= \ & \left(\widehat{I}\cos\mu_i + \widehat{K}^T \sin\mu_i\right) \widehat{T} \left(\widehat{I}\cos\mu_i + \widehat{K}\sin\mu_i\right) \\
= \ & \left(\widehat{I}\cos\mu_i + \widehat{K}^T \sin\mu_i\right) \left(\widehat{T}\cos\mu_i + \widehat{J}\sin\mu_i\right) \\
= \ & \widehat{T}\cos^2\mu_i + \widehat{J}\sin\mu_i\cos\mu_i - \widehat{J}\sin\mu_i\cos\mu_i \\
& + \widehat{T}\sin^2\mu_i \\
= \ & \widehat{T}, \qquad (181)
\end{aligned}
$$

which is indeed what we needed to prove. To conclude we remark that there is not only one invariant ellipse, but even every ellipse that can be generated by stretching or shrinking from the original one is invariant. So altogether, we have a nested set of invariant ellipses, and particles will always stay contained on the invariant ellipse on which they are originally lying.

2. A Glimpse at Nonlinear Effects

Linear motion around a fixed point is completely classified by the two cases we discussed in the previous section, namely the stable or unstable case. This situation is **fundamentally different in the nonlinear case**, it is in fact much more complicated and interesting. Indeed there is even a whole modern research field dealing with just such questions, the field called **nonlinear dynamics**.

While this is not at all the place to try to develop a complete understanding of the nonlinear effects that may appear, let us spend some time to stake the territory and make some general observations. First we may expect that as long as the motion is "**close enough**" to the fixed point, it is **dominated by linear effects**, and depending on whether we have stability or not, we see either stable elliptic motion or unstable hyperbolic motion. While we may expect that linearly unstable motion will in most cases also stay unstable if we consider the nonlinear effects, **linear stable motion will not usually stay nonlinearly stable**. In fact, if the amplitudes of the motion become large, the effects of nonlinearity will become noticeable over-proportionally, and eventually they will become dominating, in most cases leading to instability for large amplitudes.

One can then try to heuristically separate the phase space into a region that appears stable for a reasonable number of turns, and a region that appears unstable. According to the previous arguments, in most cases the stable region will be near the fixed point, and the unstable region will be away from the fixed point. The region of transition between the apparently stable and apparently unstable parts is usually called the **dynamic aperture**, and it often looks like a deformed ellipse.

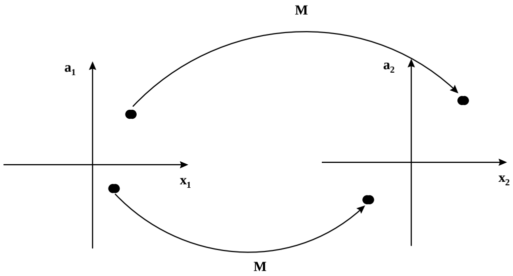
Let us now study a little what conditions seem to favor stable or unstable motion respectively. If we divide the phase space regions into parts in which the nonlinear effects have a tendency to pull particles away from the origin and those that tend to push the particles toward the origin, then we may expect that we want to avoid situations where the particles spend "too much time" in the "pull away" regions, and it is better if we sample the phase space uniformly, and thus averaging out the effects as much as possible.

A near uniform sampling of the phase space is favored if the linear **tune is not rational** multiple of 2π. On the other hand, if the tune is of the form $\mu_i = 2\pi\frac{p}{q}$, after q turns the particle will come back to where it was before and hence can see the same effect, a situation which we call **resonance**; so it is at least not a good idea to choose q too small, as repetition after large numbers of turns is not as critical.

We may also wonder to what extent it is possible to perform a transformation to normal form coordinates similar as in the linear case. As it turns out, "most" systems cannot be brought to a normal form in which the motion is exactly circular; the existence of such a transformation is tantamount to the system being **integrable,** i.e. having one integral of motion per phase space dimension. Truly integrable systems, however, are very rare. It turns out however, there is a beautiful order-by-order iterative procedure to turn a system into nonlinear normal form up to a given order. In practice, this procedure, together with the calculation of maps to higher order, is a prime application of the **differential algebraic methods** developed by us, and so far not possible in any other way; but this is a topic for PHY962!

IX. Linear Phase Space Motion

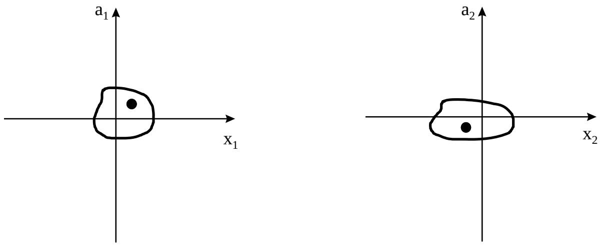
In this section, we want to study the action of transfer matrices on particles by looking in detail to what happens to **entire regions of phase space** as they are transported. This is important because the beam in an accelerator is just such a region, and of course we want to make sure that at any time, this region is within the beampipe!



Let us begin by collecting several observations about two dimensional transfer maps $\vec{\mathcal{M}}$.

1. $\vec{\mathcal{M}}$ preserves areas
2. Different initial points have different final points
3. Continuous curves stay continuous curves
4. Closed curves stay closed curves
5. A point inside a closed curve will stay inside of the closed curve

Let us remind ourselves that the first point led us to giving a name, namely "emittance," to the preserved area: the figure illustrates again how area is preserved.



The last two observations are particularly important, as they tells us that if we can enclose our beam within any closed boundary curve, then it is sufficient to study the dynamics of this boundary curve alone. It is interesting to note that while in the two-dimensional case, closed curves always stay closed curves, it is not generally true that in higher dimensions, closed surfaces stay closed surfaces. While this is true for linear higher-dimensional transformation, nonlinear maps can produce some "holes" in the surfaces through which particles that were initially "trapped" inside the surface may find a way to escape.

If in particular $\vec{\mathcal{M}}$ is linear, then we also have

1. Straight lines stay straight lines
2. Ellipses stay ellipses

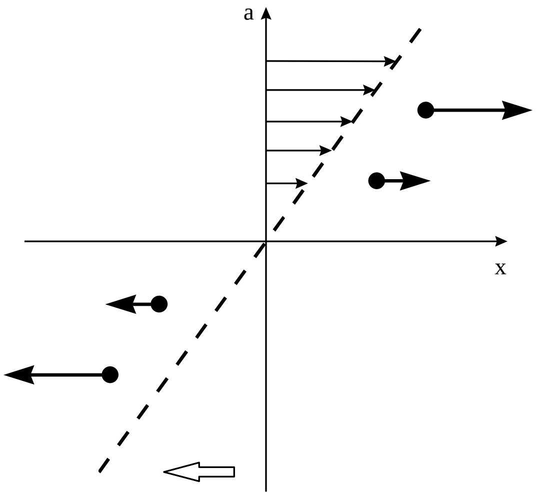
Since straight lines stay straight lines, we may "manufacture" such a boundary curve as a **polygon**; and to study its motion it is completely sufficient to move only the cornerpoints. Alternatively, we may try to enclose the beam by an **ellipse**. Before we follow these ideas, let us first study the action in phase space of some simple devices.

A. Phase Space Action of Drifts and Lenses

In the case of a drift, the matrix is given by

$$\hat{M} = \begin{pmatrix} 1 & l \\ 0 & 1 \end{pmatrix} \tag{182}$$

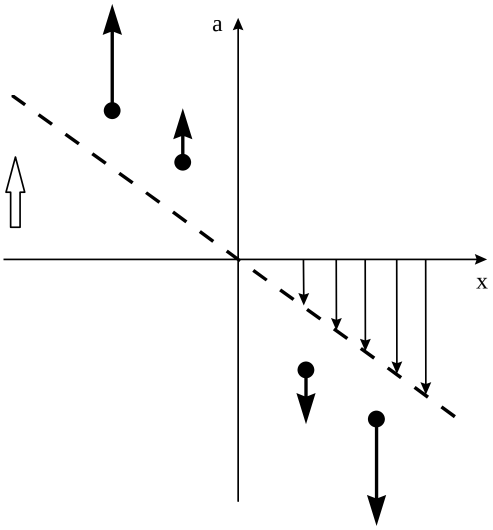
This matrix leaves a constant and moves x by an amount proportional to a; hence it perform a **horizontal shearing** in phase space.



Similarly, a lens has the drift matrix

$$\hat{M} = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix}; \tag{183}$$

it leaves x invariant and changes a by a value proportional to x; it performs a **vertical shearing.**

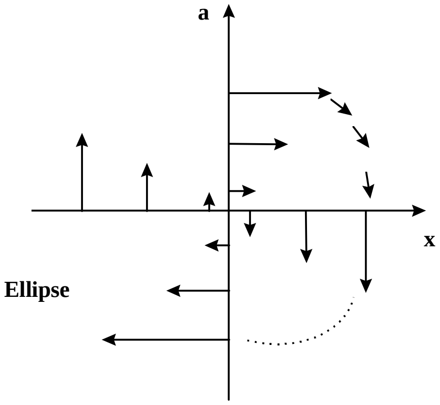


B. Phase Space Action of Quads and Dipoles

In the case of quads and dipoles, the matrices have the following form:

$$\hat{M} \propto \begin{pmatrix} \cos\phi & k\sin\phi \\ -\frac{1}{k}\sin\phi & \cos\phi \end{pmatrix}$$

This corresponds roughly to a rotation, except that the x and a coordinates are also stretched or compressed; the result is a motion on an ellipse.



In fact, computing the invariant ellipse of the motion following the procedure of the last section, we get from eq. (175) that

$$\alpha_i = 0 \quad (184)$$

$$\beta_i = k \quad (185)$$

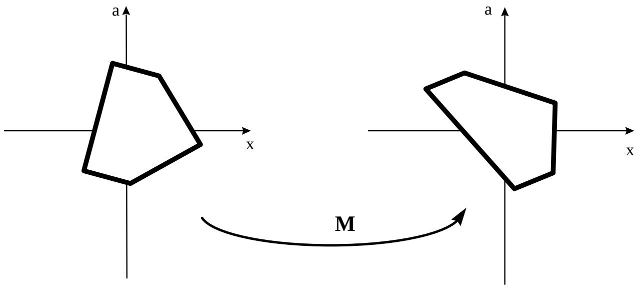
$$\gamma_i = \frac{1}{k}, \quad (186)$$

and as we expected, we see from $\alpha_i = 0$ that the ellipse is even upright.

In order to study the motion of ensembles of particles under linear transformations, it is useful by characterize them by certain simple geometric forms requiring few parameters in which the particles are contained. The two most useful such forms are the **polygon** and the **ellipse**.

C. Polygon-like Phase Space

A polygon in phase space is uniquely defined by its corner points; and since straight lines stay straight lines, it is sufficient to study just the motion of the corner points.

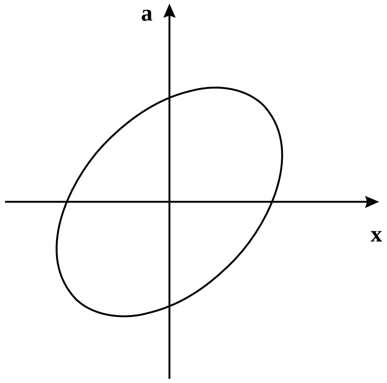


Frequently a polygon with just four points is chosen; if its lines are initially symmetrically arranged around the origin, they will stay symmetrically arranged. But a four-point polygon with symmetry around the origin is a parallelogram, and so parallelograms always stay parallelograms.

In many cases it is worth to study how the actual beam width changes as a function of s-position along the beamline. The beam width is apparently determined by the maximum of the horizontal positions of the corner points. In the special case in which we consider motion through a drift, each of the cornerpoints moves on a straight line. Furthermore, the cornerpoint that is furthest out will stay furthest out until it is possibly overtaken by another cornerpoint; during the time it determines the beamwidth, it entails that the beam width changes linearly with s. Since the outermost cornerpoint can change from time to time, the resulting beam width is piecewise linear with s.

D. Elliptic Phase Space

The other choice that is worth considering is that of an elliptic phase space.



In this case, the boundary of the phase space satisfies the ellipse condition

$$\gamma x^2 + 2\alpha x a + \beta a^2 = \varepsilon. \quad (187)$$

We first note that there is a redundancy in the description of the ellipse: obviously, doubling the values of α, β, γ as well as ε simultaneously leads to the same ellipse. In order to eliminate this redundancy, we demand that the determinant of the ellipse be unity, i.e.

$$\beta\gamma - \alpha^2 = 1. \quad (188)$$

With this choice of the matrix, the quantity ε is a unique measure of its area. We recall that the ellipse can be written in matrix form as

$$(x,a)\cdot\left(\begin{array}{cc}\gamma & \alpha\\ \alpha & \beta\end{array}\right)\cdot\left(\begin{array}{c}x\\ a\end{array}\right)=\varepsilon. \tag{189}$$

For future simplicity, we denote the matrix describing the ellipse by $\hat{\sigma}$.

Now we are ready to study the question how the phase space ellipse changes as we pass through a system. Let $\hat{M}$ be the transfer matrix of the system; then the coordinates x_1, a_1 are transformed to x_2, a_2 via

$$\left(\begin{array}{c}x_2\\ a_2\end{array}\right)=\widehat{M}\cdot\left(\begin{array}{c}x_1\\ a_1\end{array}\right);$$

and we also have

$$\left(\begin{array}{c}x_1\\ a_1\end{array}\right)=\widehat{M}^{-1}\cdot\left(\begin{array}{c}x_2\\ a_2\end{array}\right).$$

The new ellipse after the system characterized by $\hat{M}$ must obviously satisfy

$$(x_2,a_2)\cdot\widehat{\sigma}_2\cdot\left(\begin{array}{c}x_2\\ a_2\end{array}\right)=\varepsilon; \tag{190}$$

observe that if we demand that $\det(\hat{\sigma}_2)=1$, even the measure for the occupied area, ε, must be the same as before since we know the transfer map preserves area. We remind ourselves that the old coordinates satisfy

$$(x_1,a_1)\cdot\widehat{\sigma}_1\cdot\left(\begin{array}{c}x_1\\ a_1\end{array}\right)=\varepsilon$$

Expressing x_1, a_1 in terms of x_2, a_2, which is accomplished by the inverse matrix, we get

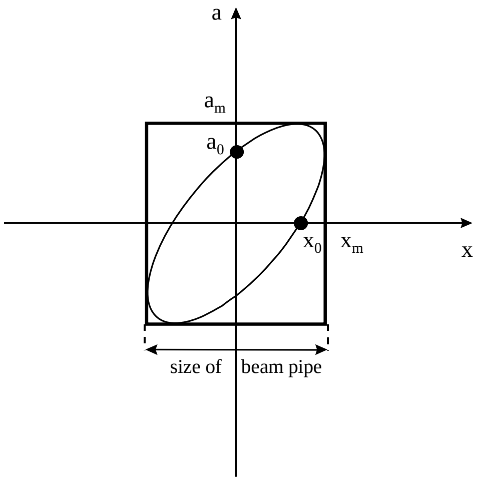
$$(x_2,a_2)\cdot\left(\left(\widehat{M}^{-1}\right)^T\cdot\widehat{\sigma}_1\cdot\widehat{M}^{-1}\right)\cdot\left(\begin{array}{c}x_2\\ a_2\end{array}\right)=\varepsilon; \tag{191}$$

we first conclude that the resulting object is again an ellipse; so ellipses are indeed preserved under linear transformation. But furthermore, since $\left(\widehat{M}^{-1}\right)^T\cdot\widehat{\sigma}_1\cdot\widehat{M}^{-1}$ is a symmetric matrix with unity determinant, such a representation of an ellipse by a symmetric matrix of unity determinant is unique, and because equations (190) and (191) hold at the same time, we must conclude that

$$\hat{\sigma}_2=\left(\widehat{M}^{-1}\right)^T\cdot\widehat{\sigma}_1\cdot\widehat{M}^{-1} \tag{192}$$

E. Practical Meaning of α, β, γ

As we propagate the beam through a system, the value of σ changes with s, and so do its three characteristic quantities α, β, γ. It is now important to study how the three quantities α, β and γ describe important characteristics of the beam like its width. Another important question relates to the shape and degree of deformation of the ellipse; together with the widths, this is characterized by the points at which the ellipse intersects the axes.



The question of intersection with the axes can be answered readily; in

$$\gamma x^2+2\alpha xa+\beta a^2=\varepsilon$$

we just set a and x to zero, and obtain

$$x_0=\sqrt{\frac{\varepsilon}{\gamma}}, a_0=\sqrt{\frac{\varepsilon}{\beta}}.$$

Now for the calculation of the maximal points x_m, and a_m, which characterize the width as well as the maximum angle in the ellipse. To this end, we view the elliptic shape as the contour line of a function, and remember that the gradient is always perpendicular to the contour lines. Hence the maximum position occurs where the angular component of the gradient vanishes, and the maximum angle occurs where the positional component of the gradient disappears. To

$$\begin{aligned} f(x,a) &= \gamma x^2 + 2\alpha x a + \beta a^2 \text{ we have} \\ \vec{\nabla} f &= (2\gamma x + 2\alpha a, 2\alpha x + 2\beta a), \end{aligned}$$

and we infer that for the maximum position, we must have $ax = -\beta a$, or $a = -\alpha/\beta \cdot x$. Inserting this into the ellipse yields

$$\begin{aligned} \gamma x^2 + 2\alpha x(-\frac{\alpha}{\beta}x) + \beta\left(-\frac{\alpha}{\beta}x\right)^2 &= \varepsilon \\ \beta\gamma x^2 - 2\alpha^2 x^2 + \alpha^2 x^2 &= \varepsilon\beta \\ \left(\beta\gamma - \alpha^2\right) x^2 &= \varepsilon\beta \end{aligned}$$

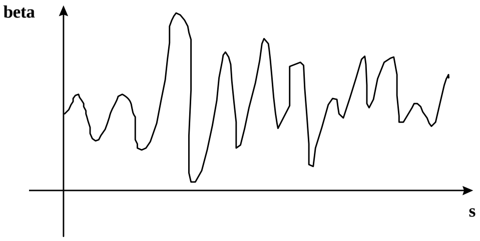
or

$$x_m = \sqrt{\varepsilon\beta}. \qquad (193)$$

Because of the symmetry of the equations with respect to interchange of x and a, we see that also

$$a_m = \sqrt{\varepsilon\gamma}. \qquad (194)$$

So the width in x direction is determined by the area of phase space ε as well as the function β. Thus, β plays an eminent role, as it immediately tells the width of a beam at a given point; and plots of its value for different positions around the accelerator are very common.



F. The Explicit Transformation of The Ellipse

For many practical purposes, it is useful to explicitly study the transformation of the ellipse (192) through the influence of the matrix $\hat{M}$. We first observe that if

$$\hat{M} = \begin{pmatrix} (x|x) & (x|a) \\ (a|x) & (a|a) \end{pmatrix},$$

then

$$\widehat{M}^{-1} = \begin{pmatrix} (a|a) & -(x|a) \\ -(a|x) & (x|x) \end{pmatrix},$$

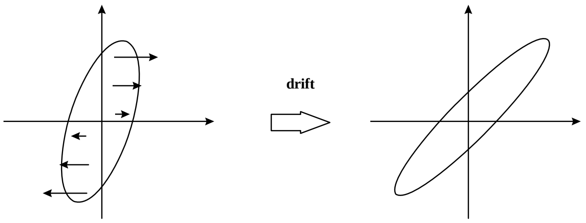
as simple arithmetic shows. So we have

$$\begin{aligned} \widehat{\sigma}_2 &= \left(\widehat{M}^{-1}\right)^T \cdot \begin{pmatrix} \gamma_1 & \alpha_1 \\ \alpha_1 & \beta_1 \end{pmatrix} \cdot \left(\widehat{M}^{-1}\right) \\ &= \begin{pmatrix} \gamma_2 & \alpha_2 \\ \alpha_2 & \beta_2 \end{pmatrix} \\ &= \begin{pmatrix} (a|a) & -(a|x) \\ -(x|a) & (x|x) \end{pmatrix} \cdot \\ & \quad \begin{pmatrix} \gamma_1 & \alpha_1 \\ \alpha_1 & \beta_1 \end{pmatrix} \cdot \begin{pmatrix} (a|a) & -(x|a) \\ -(a|x) & (x|x) \end{pmatrix} \end{aligned}$$

Performing the calculations, we see first of all that α_2, β_2, and γ_2 depend **linearly** on α_1, β_1, and γ_1, and hence the relationship can be written in matrix form (the third matrix, after $\hat{M}$ and $\hat{\sigma}$!, and this time three-by-three; aren't you getting your money's worth!) Explicity, we have

$$\begin{pmatrix} \beta_2 \\ \alpha_2 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} (x|x)^2 & -2(x|x)(x|a) & (x|a)^2 \\ -(x|x)(a|x) & (x|x)(a|a) + (x|a)(a|x) & -(x|a)(a|a) \\ (a|x)^2 & -2(a|x)(a|a) & (a|a)^2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \alpha_1 \\ \gamma_1 \end{pmatrix}.$$

One particularly interesting case is the one where we let an ellipse evolve under the action of a drift, as shown in the picture:



If we are interested in the way in which the width of the beam changes, we must look at the function $\beta(s)$. For the special case of the drift matrix with $(x|x) = (a|a) = 1$, $(a|x) = 0$ and $(x|a) = L$, we have

$$\begin{aligned}\beta(s) &= (x|x)^2\,\beta_1 - 2\,(x|x)\,(x|a)\,\alpha_1 + (x|a)^2\,\gamma_1\\ &= \beta_1 - 2L\alpha_1 + L^2\gamma_1\\ &= \gamma_1\left(L - \frac{\alpha_1}{\gamma_1}\right)^2 - \frac{\alpha_1^2}{\gamma_1} + \beta_1.\end{aligned}$$

So as a function of L, $\beta(s)$ changes quadratically! We also see readily that at the point where

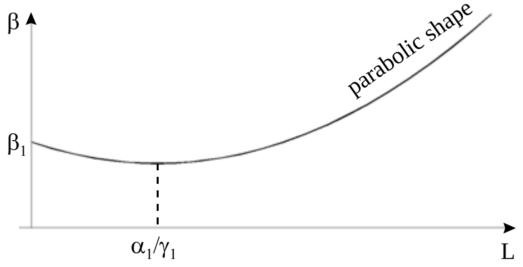
$$L = \frac{\alpha}{\gamma}, \tag{195}$$

the beam has minimum width, and we have what is called a "waist". Finally, we may ask ourselves about the rate of change of $\beta(s)$. Differentiating, we obtain

$$\frac{d\beta}{dL} = 2L\gamma_1 - 2\alpha_1,$$

and hence

$$\begin{aligned}\beta' &= -2\alpha & (196)\\ \beta'' &= 2\gamma & (197)\end{aligned}$$

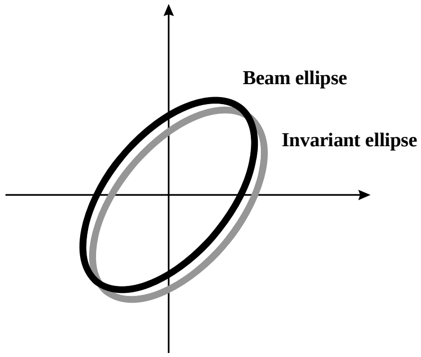


G. Invariant Ellipses Versus Beam Ellipses

The last important question remaining in this section is to put into perspective the parameters of the **beam** α, β, γ and the parameters α_i, β_i, γ_i describing the invariant ellipse of the cell of the **accelerator.** Are these greek letter equal, are they related, or do they have nothing to do with each other? This is actually a question that often throws off even die-hard accelerator physicists, and it is very much worthwhile to understand it in depth.

As far as the theory goes, these two sets of greek letters are actually **totally independent**. In fact, one describes some property of an accelerator, and the other describes a property of a beam; and of course we can feed any type of beam into a given accelerator.

However, if the goal is to fill the accelerator in the most efficient way, as it turns out this is accomplished if the beam's greek **agrees** with the accelerator's greek! In this case, after one revolution the phase space will occupy exactly the same area (although the individual particles in it are at different positions), as shown in the picture.



On the other hand, if one injects a beam with an ellipse that does not agree with the invariant ellipse of the accelerator, then the repetitive behavior of the beam ellipse shown solid in the next picture is determined by the dashed invariant ellipse it touches.

As we go around the repetitive system repeatedly, the beam ellipse stays within the dashed ellipse and touches it, but depending on the tune, will have a different orientation. In fact, if the tune isn't rational - something desireable for stability reasons - over time even all different orientations will occur. If we now want to operate the accelerator, we have to make sure we can handle everything inside the dashed ellipse, leading to considerable waste of area!

So it is best to operate a repetitive system in such a way that the beam ellipse is **matched** to the accelerator's invariant ellipse, and to avoid mismatching, the so-called **beating.**

X. Aberration Formulas

In this section, we want to derive a method to compute the Taylor transfer map of a general weakly nonlinear system that is origin preserving. So let us assume we are given the system

$$\frac{d}{ds}\vec{r} = \vec{f}\left(\vec{r}, s\right),$$

which satisfies $\vec{f}(\vec{0}, s) = \vec{0}$. We perform a Taylor expansion of the right hand side. Because the system is origin preserving, the first contribution is linear, and altogether we have

$$\frac{d}{ds}\vec{r} = \hat{M}\left(s\right) \cdot \vec{r} + \sum_{j=2}^{\infty} \vec{N}_j\left(\vec{r}, s\right),$$

where the $\vec{N}_j$ are polynomials of exact order j, the coefficients of which may depend on s.

The first step in obtaining a perturbative solution of the system is a **linearization;** we obtain

$$\frac{d}{ds}\vec{r} = \hat{M}\left(s\right) \cdot \vec{r}.$$

For this system, we determine a system of n independent solutions $\vec{l}_k\left(s\right)$, $k = 1, ..., n$, that satisfy the initial condition

$$\vec{l}_k\left(0\right) = (0, 0, \ldots, \underbrace{1}_{k-th}, \ldots 0, 0)^t.$$

We define the matrix

$$\hat{L}\left(s\right) = \left(\vec{l}_1\left(s\right), \vec{l}_2\left(s\right), \ldots, \vec{l}_n\left(s\right)\right),$$

and observe that the **general solution** of the linearized problem with initial condition $\vec{r}_i$ is then given by

$$\vec{r}\left(s\right) = \hat{L}\left(s\right) \cdot \vec{r}_i.$$

In practice, the determination of $\hat{L}$ may be possible in closed form, depending on the structure of $\hat{M}$, or may have to rely on numerical integration. For the special case that $\hat{M}$ is **piecewise constant**, then for every such piece, one can try the Ansatz $\vec{l}_k = \vec{v}_k \cdot \exp(\omega_k s)$, which leads to the condition

$$\omega_k \vec{v}_k \exp(\omega_k s) = \hat{M} \cdot \vec{v}_k \exp(\omega_k s),$$

an eigenvector problem. If $\hat{M}$ has n distinct eigenvalues, we are done, and depending on whether ω_k is real or complex, the solutions can also be expressed in terms of sin, cos or sinh, cosh. In case of multiple eigenvalues, often solutions of the form $s\cdot$ sin etc can be found.

The next step consists of an **expansion** of $\vec{r}\left(s\right)$ in a Taylor polynomial

$$\vec{r}\left(s\right) = \hat{L}\left(s\right) \cdot \vec{r}_i + \sum_{j=2}^{\infty} \vec{R}_j\left(s, \vec{r}_i\right),$$

where $\vec{R}_j$ denotes a polynomial of exact order j in the initial conditions, the coefficients of which may depend on s. We **insert** this expansion into the ODE and obtain

$$\frac{d}{ds}\hat{L}\left(s\right) \cdot \vec{r}_i + \sum_{j=2}^{\infty} \frac{d}{ds}\vec{R}_j\left(s, \vec{r}_i\right)$$

$$\begin{aligned} &= \hat{M}(s)\cdot\hat{L}(s)\cdot\vec{r}_i + \hat{M}(s)\cdot\sum_{j=2}^{\infty}\vec{R}_j(s,\vec{r}_i) \\ &\quad + \sum_{j=2}^{\infty}\vec{Q}_j(s,\widehat{L},\vec{R}_k), \end{aligned}$$

where the $\vec{Q}_j$ are polynomials of order $j \geq 2$, in $\vec{r}$, which result from inserting $\vec{r}$ into the $\vec{N}_j$. This insertion leaves no linear or constant parts, which is due to the fact that the ODE is origin preserving. This will prove crucial later on in the algorithm for the solution.

We now sort the result by order; the **linear** part has the form

$$\frac{d}{ds}\hat{L} = \hat{M}\cdot\hat{L}, \tag{198}$$

and the higher orders $j = 2, ...$ assume the form

$$\frac{d}{ds}\vec{R}_j(s,\vec{r}_i) = \hat{M}(s)\cdot\vec{R}_j(s,\vec{r}_i) + \vec{Q}_j(s,\hat{L},\vec{R}_k), \tag{199}$$

where the $\vec{Q}_j$ contains only $\vec{R}_k$ with $k < j$.

So for $j = 2, 3, ...$ we obtain a triangular system of ODEs. It can be solved iteratively in an **order-by-order** manner, and then each of the differential equations for $\vec{R}_j$ contains only lower order terms $\vec{R}_k$ that are already known. In this way, the ODEs **decouple** and become **inhomogeneous**. Because of the initial condition $\vec{r}(0) = \vec{r}_i$, we have that

$$\begin{aligned} \hat{L}(0) &= \hat{I}, \\ \vec{R}_j(0,\vec{r}) &= 0 \text{ for all } j = 2, 3, \ldots \end{aligned}$$

In order to solve the inhomogenous equation (199) of order j, we first determine the homogeneous solution, and then perform a so-called variation of parameters. The **homogenous** solution is just exactly the same as for the linearized case, and we have $\vec{R}_j(s) = \hat{L}(s)\cdot\vec{T}$, where $\vec{T} = \vec{R}_j(0)$. To obtain the **inhomogeneous** solution, we now make the Ansatz $\vec{R}_j(s) = \hat{L}(s)\cdot\vec{T}(s)$. Insertion in (199) yields

$$\begin{aligned} \frac{d}{ds}\vec{R}_j &= \left(\frac{d}{ds}\hat{L}(s)\right)\cdot\vec{T}(s) + \hat{L}(s)\cdot\frac{d}{ds}\vec{T}(s) \\ &= \hat{M}(s)\cdot\hat{L}(s)\cdot\hat{T}(s) + \vec{Q}_j(s,\hat{L},\vec{R}_k). \end{aligned}$$

Considering that $\hat{L}$ is a solution of the linear system, we obtain

$$\begin{aligned} \hat{L}(s)\cdot\frac{d}{ds}\vec{T}(s) &= \vec{Q}_j(s,\hat{L},\vec{R}_k) \text{ or} \\ \vec{T}(s) &= \int_0^s \hat{L}^{-1}(s^{'})\vec{Q}_j(s^{'},\widehat{L},\vec{R}_k)ds^{'}, \end{aligned}$$

where the choice of the lower integration boundary as 0 assures that $\vec{T}(0) = 0$, which is necessary to satisfies the initial conditions. Altogether we have

$$\vec{R}_j(s) = \hat{L}(s)\cdot\int_0^s \hat{L}^{-1}(s^{'})\vec{Q}_j(s^{'},\hat{L},\vec{R}_k)ds^{'}$$

The integral is often referred to as the **aberration integral**, and the argument of the integral as the **driving term**. The complete solution then is obtained as

$$\vec{r}(s) = \hat{L}(s)\cdot\vec{r}_i + \sum_{j=2}^{\infty}\vec{R}_j(s).$$

So apparently, once the linear solution is known, everything else just boils down to quadratures. If within a piece in which it is constant, $\hat{M}(s)$ is diagonalizable, the linear solutions can be written as combinations of sin, cos, sinh, cosh, and s. In other important cases where $\hat{M}(s)$ is singular, often a complete set of linear solutions that are polynomials in s can be obtained.

In both of these cases, the insertion into the polynomials $\vec{R}_j(s)$ leads to terms that are polynomials in sin, cos, sinh, cosh, and s. By expressing such functions in terms of exponentials times powers of s, one can show that the result of any integration can again be expressed as a polynomial of sin, cos, sinh, cosh, and s.

For practical cases, it is worthwhile to discuss the **complexity** of the procedure**.** With each new order, the expansion of the ODE becomes more complicated; then all previous orders have to be inserted, multiplied with the linear inverses, and integrated, resulting in substantially more terms than for the previous order. Altogether, the effort **increases extremely dramatically** with the order being considered, and for typical systems has proved practical only to orders around five.

Computer codes that use the above procedure usually contain a library of procedures that compute the aberrations for each particle optical element of interest. The aberrations of combined systems is then determined from those of the pieces with the help of a composition procedure. This approach was used first in the code **TRANSPORT** to second order, and then to third order in **TRIO** and the related **GIOS**. The subsequent code **MaryLie** achieved third order in a similar way using Poisson bracket methods. **COSY 5.0** contains libraries to fifth order, which were generated using a custom-made formula manipulator.

Modern codes, including **COSY INFINITY**, usually make use of the **Differential Algebraic** approach, which allows the computation of aberrations to any order in an elegant way without the need of explicit formulas for aberrations.

To illustrate the method of computation of aberrations with a simple **example**, let us consider the differential equation

$$\begin{aligned} x' &= a \\ a' &= -x + k \cdot x^2, \end{aligned}$$

which corresponds to the horizontal motion in a quadrupole with a superimposed sextupole. We first perform the **linearization** to obtain

$$\begin{aligned} \begin{pmatrix} x \\ a \end{pmatrix}' &= \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} x \\ a \end{pmatrix} \\ &= \hat{M}(s) \cdot \begin{pmatrix} x \\ a \end{pmatrix}. \end{aligned}$$

The linear solution then is

$$\begin{aligned} \begin{pmatrix} x_f \\ a_f \end{pmatrix} &= \begin{pmatrix} \cos(s) & \sin(s) \\ -\sin(s) & \cos(s) \end{pmatrix} \cdot \begin{pmatrix} x_i \\ a_i \end{pmatrix} \\ &= \hat{L}(s) \cdot \begin{pmatrix} x_i \\ a_i \end{pmatrix}. \end{aligned}$$

The next step is the **expansion** of the ODE, which is already done. We then **insert** the expansion of the transfer map

$$\begin{aligned} x(s) &= (x|x)\, x_i + (x|a)\, a_i \\ &\quad + (x|xx)\, x_i^2 + (x|xa)\, x_i a_i + (x|aa)\, a_i^2 \\ a(s) &= (a|x)\, x_i + (a|a)\, a_i \\ &\quad + (a|xx)\, x_i^2 + (a|xa)\, x_i a_i + (a|aa)\, a_i^2 \end{aligned}$$

into the ODE and obtain

$$\begin{aligned} & (x|x)'\, x_i + (x|a)'\, a_i + \\ & (x|xx)'\, x_i^2 + (x|xa)'\, x_i a_i + (x|aa)'\, a_i^2 \\ = \; & (a|x)\, x_i + (a|a)\, a_i + \\ & (a|xx)\, x_i^2 + (a|xa)\, x_i a_i + (a|aa)\, a_i^2 \end{aligned}$$

$$\begin{aligned} & (a|x)'\, x_i + (a|a)'\, a_i + \\ & (a|xx)'\, x_i^2 + (a|xa)'\, x_i a_i + (a|aa)'\, a_i^2 \\ = \; & -\left[\begin{array}{c} (x|x)\, x_i + (x|a)\, a_i \\ + (x|xx)\, x_i^2 + (x|xa)\, x_i a_i + (x|aa)\, a_i^2 \end{array} \right] \\ & +k \left[\begin{array}{c} (x|x)^2\, x_i^2 + 2\,(x|x)\,(x|a)\, x_i a_i \\ + (x|a)^2\, a_i^2 + \ldots \end{array} \right] \end{aligned}$$

Since we are interested only in order 2, we can ignore the higher order terms. The second order equations then read

$$\begin{aligned} & (x|xx)^{'} x_i^2 + (x|xa)^{'} x_i a_i + (x|aa)^{'} a_i^2 \\ = \quad & (a|xx)\, x_i^2 + (a|xa)\, x_i a_i + (a|aa)\, a_i^2 \end{aligned}$$

and

$$\begin{aligned} & (a|xx)^{'} x_i^2 + (a|xa)^{'} x_i a_i + (a|aa)^{'} a_i^2 \\ = \quad & -(x|xx)\, x_i^2 - (x|xa)\, x_i a_i - (x|aa)\, a_i^2 \\ & +k\,(x|x)^2\, x_i^2 + 2k\,(x|x)\,(x|a)\, x_i a_i + k\,(x|a)^2\, a_i^2 \end{aligned}$$

where the last line contains the inhomogenous part. Following the algorithm, we make the ansatz

$$\begin{aligned} & \begin{pmatrix} (x|xx)\, x_i^2 + (x|xa)\, x_i a_i + (x|aa)\, a_i^2 \\ (a|xx)\, x_i^2 + (a|xa)\, x_i a_i + (a|aa)\, a_i^2 \end{pmatrix} \\ = \quad & \begin{pmatrix} \cos s & \sin s \\ -\sin s & \cos s \end{pmatrix} \cdot \vec{T}(s)\,, \end{aligned}$$

which then leads to

$$\begin{aligned} & \vec{T}(s) \\ = \quad & \int_0^s \begin{pmatrix} \cos s^{'} & -\sin s^{'} \\ \sin s^{'} & \cos s^{'} \end{pmatrix} \cdot \\ & \begin{pmatrix} 0 \\ \{k \cos^2 s^{'} \cdot x_i^2 \\ +2k \cos s^{'} \sin s^{'} \cdot x_i a_i \\ +k \sin^2 s^{'} \cdot a_i^2\} \end{pmatrix} ds^{'} \\ = \quad & k \cdot \begin{pmatrix} \int_0^s \{-\cos^2 s^{'} \cdot \sin s^{'} \cdot x_i^2 \\ -2 \cos s^{'} \cdot \sin^2 s^{'} \cdot x_i a_i \\ -\sin^3 s^{'} \cdot a_i^2 \,\}\, ds' \,, \\ \int_0^s \{\cos^3 s^{'} \cdot x_i^2 \\ +2 \cos^2 s^{'} \cdot \sin s^{'} \cdot x_i a_i \\ +\cos s^{'} \sin^2 s^{'} \cdot a_i^2 \,\}\, ds' \end{pmatrix} \\ = \quad & k \cdot \begin{pmatrix} \{1/3(\cos^3 s - 1) \cdot x_i^2 - 2/3 \sin^3 s \cdot x_i a_i \\ +(\cos s - 1/3 \cos^3 s - 2/3) \cdot a_i^2\} \,, \\ \{(\sin s - 1/3 \sin^3 s) \cdot x_i^2 \\ -2/3 \cos^3 s \cdot x_i a_i + 1/3 \sin^3 s \cdot a_i^2\} \end{pmatrix} \end{aligned}$$

Then we obtain $\vec{R}_2(s) = \hat{L}(s) \cdot \vec{T}(s)$, which yields the matrix elements

$$\begin{aligned} (x,xx) &= k \cdot \left[\frac{1}{3}\sin^2 s - \frac{1}{3}\cos s + \frac{1}{3}\right] \\ (x,xa) &= k \cdot \left[-\frac{2}{3}\sin s \cos s + \frac{2}{3}\sin s\right] \\ (x,aa) &= k \cdot \left[\frac{1}{3}\cos^2 s - \frac{2}{3}\cos s + \frac{1}{3}\right] \\ (a,xx) &= k \cdot \left[\frac{2}{3}\sin s \cos s + \frac{1}{3}\sin s\right] \\ (a,xa) &= k \cdot \left[+\frac{2}{3}\sin^2 s - \frac{2}{3}\cos^2 s + \frac{2}{3}\cos s\right] \\ (a,aa) &= k \cdot \left[-\frac{2}{3}\sin s \cos s + \frac{2}{3}\sin s\right] \end{aligned}$$

In a similar fashion, but with much more effort, one could now compute the third and higher order terms.

	Examples	typical E
Van de Graff	Many fixed stage	30 MeV
Linac	SLAC,	50 GeV (e^{-})
Betatron	Kerst	300 MeV e, 50 Mev p
Microtron	(CEBAF)	
Cyclotron	LBL	20 Mev (p)
Synchrocyclotron	Dubna	500 Mev (p)
Isochro Cyclotron	NSCL	200 MeV/A
Synchrotron ($p^{\pm}$)	Tevatron ; LHC	1TeV; 8 TeV
Synchrotron ($e^{\pm}$)	LEP	200 GeV
Wake Field	-	1000 TeV (??)