Chapter 2

The Earth's Gravitational field

2.1 Global Gravity, Potentials, Figure of the Earth, Geoid

Introduction

Historically, gravity has played a central role in studies of dynamic processes in the Earth's interior and is also important in exploration geophysics. The concept of gravity is relatively simple, high-precision measurements of the gravity field are inexpensive and quick, and spatial variations in the gravitational acceleration give important information about the dynamical state of Earth. However, the study of the gravity of Earth is not easy since many corrections have to be made to isolate the small signal due to dynamic processes, and the underlying theory — although perhaps more elegant than, for instance, in seismology — is complex. With respect to determining the three-dimensional structure of the Earth's interior, an additional disadvantage of gravity, indeed, of any potential field, over seismic imaging is that there is larger ambiguity in locating the source of gravitational anomalies, in particular in the radial direction.

In general the gravity signal has a complex origin: the acceleration due to gravity, denoted by g, ($\mathbf{g}$ in vector notation) is influenced by topography, aspherical variation of density within the Earth, and the Earth's rotation. In geophysics, our task is to measure, characterize, and interpret the gravity signal, and the reduction of gravity data is a very important aspect of this scientific field. Gravity measurements are typically given with respect to a certain reference, which can but does not have to be an *equipotential surface*. An important example of an equipotential surface is the *geoid* (which itself represents deviations from a *reference spheroid*).

The Gravity Field

The law of gravitational attraction was formulated by Isaac Newton (1642-1727) and published in 1687, that is, about three generations after Galileo had determined the magnitude of the gravitational acceleration and Kepler had discovered his empirical "laws" describing the orbits of planets. In fact, a strong argument for the validity of Newton's laws of motion and gravity was that they could be used to derive Kepler's laws.

For our purposes, gravity can be defined as the force exerted on a mass m due to the *combination* of (1) the gravitational attraction of the Earth, with mass M or M_E and (2) the rotation of the Earth. The latter has two components: the centrifugal acceleration due to rotation with angular velocity ω and the existence of an equatorial bulge that results from the balance between self-gravitation and rotation.

The **gravitational force** between any two particles with (point) masses M at position $\mathbf{r}_0$ and m at position $\mathbf{r}$ separated by a distance r is an attraction along a line joining the particles (see Figure 2.1):

$$F = \|\mathbf{F}\| = G\frac{Mm}{r^2}, \tag{2.1}$$

or, in vector form:

$$\mathbf{F} = -G\frac{Mm}{\|\mathbf{r}-\mathbf{r_0}\|^3}(\mathbf{r}-\mathbf{r_0}) = -G\frac{Mm}{\|\mathbf{r}-\mathbf{r_0}\|^2}\hat{\mathbf{r}}'. \tag{2.2}$$

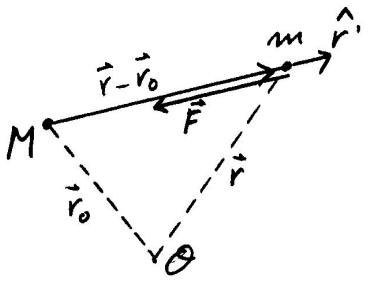

Figure 2.1: Vector diagram showing the geometry of the gravitational attraction.

where $\hat{\mathbf{r}}'$ is a unit vector in the direction of $(\mathbf{r}-\mathbf{r_0})$. The minus sign accounts for the fact that the force vector $\mathbf{F}$ points inward (i.e., towards M) whereas the unit vector $\hat{\mathbf{r}}'$ points outward (away from M). In the following we will place M at the origin of our coordinate system and take $\mathbf{r_0}$ at $\mathbf{O}$ to simplify the equations (e.g., $\mathbf{r}-\mathbf{r_0}=\mathbf{r}$ and the unit vector $\hat{\mathbf{r}}'$ becomes $\hat{\mathbf{r}}$) (see Figure 2.2).

G is the **universal gravitational constant**: $G = 6.673 \times 10^{-11}$ m^3 kg^{-1} s^{-2} (or N m^2 kg^{-2}), which has the same value for all pairs of particles. G must not be confused with $\mathbf{g}$, the **gravitational acceleration**, or force of a unit

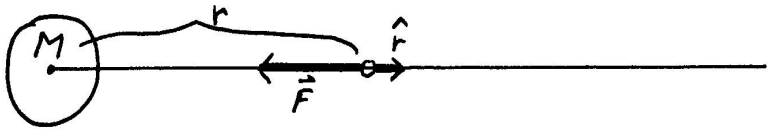

Figure 2.2: Simplified coordinate system.

mass due to gravity, for which an expression can be obtained by using Newton's law of motion. If M is the mass of Earth:

$$\mathbf{F} = m\mathbf{a} = m\mathbf{g} = -G\frac{Mm}{r^2}\hat{\mathbf{r}} \Rightarrow \mathbf{g} = \frac{\mathbf{F}}{m} = -G\frac{M}{r^2}\hat{\mathbf{r}} \tag{2.3}$$

and $$g = \|\mathbf{g}\| = G\frac{M}{r^2}. \tag{2.4}$$

The acceleration g is the length of a vector $\mathbf{g}$ and is by definition always positive: $g > 0$. We define the *vector* $\mathbf{g}$ as the **gravity field** and take, by convention, $\mathbf{g}$ positive towards the center of the Earth, i.e., in the $-\mathbf{r}$ direction.

The gravitational acceleration g was first determined by Galileo; the magnitude of $\mathbf{g}$ varies over the surface of Earth but a useful ball-park figure is g= 9.8 ms^{-2} (or just 10 ms^{-2}) (in S.I. — Système International d'Unités — units). In his honor, the unit often used in gravimetry is the *Gal*. 1 Gal = 1 cms^{-2} = 0.01 $\text{ms}^{-2} \approx 10^{-3}g$. Gravity anomalies are often expressed in *milliGal*, i.e., $10^{-6}g$ or *microGal*, i.e., $10^{-9}g$. This precision can be achieved by modern gravimeters. An alternative unit is the *gravity unit*, 1 gu = 0.1 mGal = $10^{-7}g$.

When G was determined by Cavendish in 1778 (with the Cavendish torsion balance) the mass of the Earth could be determined and it was found that the Earth's mean density, $\rho \sim 5,500\ \text{kgm}^{-3}$, is much larger than the density of rocks at the Earth's surface. This observations was one of the first strong indications that density must increase substantially towards the center of the Earth. In the decades following Cavendish' measurement, many measurements were done of g at different locations on Earth and the variation of g with latitude was soon established. In these early days of "geodesy" one focused on planet wide structure; in the mid to late 1800's scientists started to analyze deviations of the reference values, i.e. local and regional gravity anomalies.

Gravitational potential

By virtue of its position in the gravity field $\mathbf{g}$ due to mass M, any mass m has **gravitational potential energy**. This energy can be regarded as the work W done on a mass m by the gravitational force due to M in moving m from $\mathbf{r}_{\text{ref}}$ to $\mathbf{r}$ where one often takes $\mathbf{r}_{\text{ref}} = \infty$. The **gravitational potential** U is the potential energy in the field due to M per unit mass. In other words, it's the work done by the gravitational force $\mathbf{g}$ per unit mass. (One can define U as

either the *positive* or *negative* of the work done which translates in a change of sign. Beware!). The potential is a scalar field which is typically easier to handle than a vector field. And, as we will see below, from the scalar potential we can readily derive the vector field anyway.

(The gravity field is a **conservative field** so just *how* the mass m is moved from $\mathbf{r}_{\text{ref}}$ to $\mathbf{r}$ is not relevant: the work done only depends on the initial and final position.) Following the definition for potential as is common in physics, which considers Earth as a potential well — i.e. negative — we get for U:

$$U = \int_{\mathbf{r}_{\text{ref}}}^{\mathbf{r}} \mathbf{g} \cdot d\mathbf{r} = -\int_{\mathbf{r}_{\text{ref}}}^{\mathbf{r}} \frac{GM}{r^2}\hat{\mathbf{r}} \cdot d\mathbf{r} = GM \int_{\infty}^{r} \frac{1}{r^2} dr = -\frac{GM}{r} \tag{2.5}$$

Note that $\hat{\mathbf{r}} \cdot d\mathbf{r} = -dr$ because $\hat{\mathbf{r}}$ and $d\mathbf{r}$ point in opposite directions.

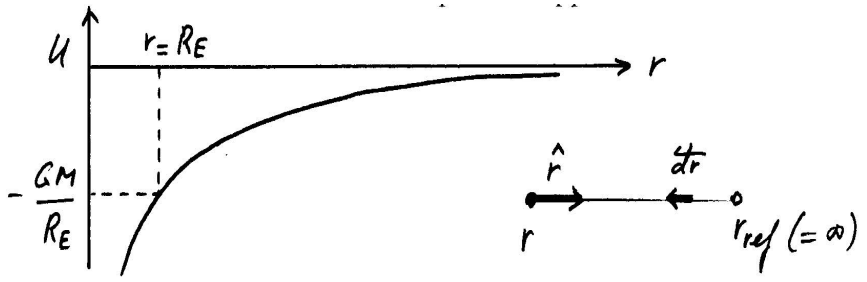

Figure 2.3: By definition, the potential is zero at infinity and decreases towards the mass.

U represents the gravitational potential at a distance r from mass M. Notice that it is assumed that $U(\infty) = 0$ (see Figure 2.3).

The potential is the integration over space (either a line, a surface or a volume) of the gravity field. Vice versa, the gravity field, the gravity force per unit mass, is the spatial derivative (gradient) of the potential.

$$\mathbf{g} = -\frac{GM}{r^2}\hat{\mathbf{r}} = \frac{\partial}{\partial \mathbf{r}}\left(\frac{GM}{r}\right) = -\frac{\partial}{\partial \mathbf{r}}U = -\text{grad}U \equiv -\boldsymbol{\nabla} U \tag{2.6}$$

Intermezzo 2.1 THE GRADIENT OF THE GRAVITATIONAL POTENTIAL

We may easily see this in a more general way by expressing $d\mathbf{r}$ (the incremental distance along the line joining two point masses) into some set of coordinates, using the properties of the dot product and the total derivative of U as follows (by our definition, moving in the same direction as $\mathbf{g}$ accumulates negative potential):

$$\begin{aligned} dU &= \mathbf{g} \cdot d\mathbf{r} \\ &= -g_x dx - g_y dy - g_z dz \end{aligned} \quad . \tag{2.7}$$

By definition, the total derivative of U is given by:

$$dU \equiv \frac{\partial U}{\partial x} dx + \frac{\partial U}{\partial y} dy + \frac{\partial U}{\partial z} dz \tag{2.8}$$

Therefore, the combination of Eq. 2.7 and Eq. 2.8 yields:

$$\mathbf{g} = -\left(\frac{\partial U}{\partial x}, \frac{\partial U}{\partial y}, \frac{\partial U}{\partial z} \right) = -\text{grad}\, U \equiv -\boldsymbol{\nabla} U \tag{2.9}$$

One can now see that the fact that the gravitational potential is defined to be negative means that when mass m approaches the Earth, its potential (energy) decreases while its acceleration due to attraction the Earth's center increases. The slope of the curve is the (positive) value of g, and the minus sign makes sure that the gradient U points in the direction of decreasing r, i.e. towards the center of mass. (The plus/minus convention is not unique. In the literature one often sees $U = GM/r$ and $\mathbf{g} = \boldsymbol{\nabla} U$.)

Some general properties:

- The gradient of a scalar field U is a vector that determines the rate and direction of change in U. Let an equipotential surface S be the surface of constant U and $\mathbf{r}_1$ and $\mathbf{r}_2$ be positions on that surface (i.e., with $U_1 = U_2 = U$). Then, the component of g along S is given by $(U_2 - U_1)/(\mathbf{r}_1 - \mathbf{r}_2) = 0$. Thus $\mathbf{g} = -\boldsymbol{\nabla} U$ has no components along S : the field is perpendicular to the equipotential surface. This is *always* the case, as derived in Intermezzo 2.2.

- Since fluids cannot sustain shear stress — the shear modulus $\mu = 0$, the forces acting on the fluid surface have to be perpendicular to this surface in steady state, since any component of a force along the surface of the fluid would result in flow until this component vanishes. The restoring forces are given by $F = -m\boldsymbol{\nabla} U$ as in Figure 2.4; a fluid surface assumes an equipotential surface.

- For a spherically symmetric Earth the equipotential would be a sphere and

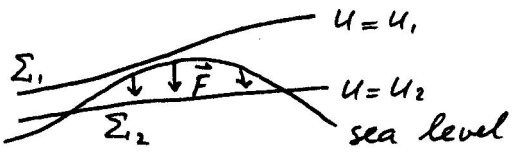

Figure 2.4: $F = -m\nabla U$ provides the restoring force that levels the sea surface along an equipotential surface.

g would point towards the center of the sphere. (Even in the presence of aspherical structure and rotation this is a very good approximation of **g**. However, if the equipotential is an ellipsoid, $\mathbf{g} = -\nabla U$ does not point to $r = 0$; this lies at the origin of the definition of geographic and geocentric latitudes.)

- Using gravity potentials, one can easily prove that the gravitational acceleration of a spherically symmetric mass distribution, at a point *outside* the mass, is the same as the acceleration obtained by concentrating all mass at the center of the sphere, i.e., a point mass.

  This seems trivial, but for the use of potential fields to study Earth's structure it has several important implications:

  1. Within a spherically symmetric body, the potential, and thus the gravitational acceleration **g** is determined only by the mass between the observation point at r and the center of mass. In spherical coordinates:

  $$g(r) = 4\pi \frac{G}{r^2} \int_0^r \rho(r') r'^2 \, dr' \tag{2.10}$$

  This is important in the understanding of the variation of the gravity field as a function of radius within the Earth;

  2. The gravitational potential by itself does not carry information about the radial distribution of the mass. We will encounter this later when we discuss more properties of potentials, the solutions of the Laplace and Poisson equations, and the problem of non-uniqueness in gravity interpretations.

  3. if there are lateral variations in gravitational acceleration on the surface of the sphere, i.e. if the equipotential is not a sphere there must be aspherical structure (departure from spherical geometry; can be in the shape of the body as well as internal distribution of density anomalies).

Intermezzo 2.2 GEOMETRIC INTERPRETATION OF THE GRADIENT

Let C be a curve with parametric representation $C(\tau)$, a vector function. Let U be a scalar function of multiple variables. The variation of U, *confined to the curve* C, is given by:

$$\frac{d}{dt}\left[U(C(t))\right] = \boldsymbol{\nabla} U(C(t)) \cdot \frac{dC(\tau)}{dt} \tag{2.11}$$

Therefore, if C is a curve of constant U, $\frac{d}{dt}\left[U(C(\tau))\right]$ will be zero.
Now let $C(\tau)$ be a straight line in space:

$$C(\tau) = \mathbf{p} + \mathbf{a}t \tag{2.12}$$

then, according to the chain rule (2.11), at $t_0 = 0$:

$$\left.\frac{d}{dt}\left[U(C(\tau))\right]\right|_{\mathbf{t=t_0}} = \boldsymbol{\nabla} U(\mathbf{p}) \cdot \mathbf{a} \tag{2.13}$$

It is useful to define the **directional derivative** of U in the direction of $\mathbf{a}$ at point $\mathbf{p}$ as:

$$D_A U(\mathbf{p}) = \boldsymbol{\nabla} U(\mathbf{p}) \cdot \frac{\mathbf{a}}{\|\mathbf{a}\|} \tag{2.14}$$

From this relation we infer that the gradient vector $\boldsymbol{\nabla} U(\mathbf{p})$ at $\mathbf{p}$ gives the direction in which the change of U is maximum. Now let S be an **equipotential surface**, i.e. the surface of constant U. Define a set of curves $C_i(\tau)$ on this surface S. Clearly,

$$\left.\frac{d}{dt}\left[U(C_i(\tau))\right]\right|_{\mathbf{t=t_0}} = \boldsymbol{\nabla} U(\mathbf{p}) \cdot \frac{dC_i}{dt}(t_0) = 0 \tag{2.15}$$

for each of those curves. Since the $C_i(\tau)$ lie completely on the surface S, the $\frac{dC_i}{dt}(t_0)$ will define a plane *tangent* to the surface S at point $\mathbf{p}$. Therefore, the gradient vector $\boldsymbol{\nabla} U$ is perpendicular to the surface S of constant U. Or: the field is perpendicular to the equipotential surface.

In global gravity one aims to determine and explain deviations from the equipotential surfaces, or more precisely the difference (height) between equipotential surfaces. This difference in height is related to the local $\mathbf{g}$. In practice one defines anomalies relative to reference surfaces. Important surfaces are:

Geoid the actual equipotential surface that coincides with the average sea level (ignoring tides and other dynamical effects in oceans)

(Reference) spheroid : empirical, longitude independent (i.e., zonal) shape of the sea level with a smooth variation in latitude that best fits the geoid (or the observed gravity data). This forms the basis of the international gravity formula that prescribes g as a function of latitude that forms the reference value for the reduction of gravity data.

Hydrostatic Figure of Shape of Earth : theoretical shape of the Earth if we know density ρ and rotation ω (ellipsoid of revolution).

We will now derive the shape of the reference spheroid; this concept is very important for geodesy since it underlies the definition of the International Gravity Formula. Also, it introduces (zonal, i.e. longitude independent) spherical harmonics in a natural way.

2.2 Gravitational potential due to nearly spherical body

How can we determine the shape of the reference spheroid? The flattening of the earth was already discovered and quantified by the end of the 18^{th} century. It was noticed that the distance between a degree of latitude as measured, for instance with a sextant, differs from that expected from a sphere: $R_E\,(\theta_1 - \theta_2) \neq R_E d\theta$, with R_E the radius of the Earth, θ_1 and θ_2 two different latitudes (see Figure 2.5).

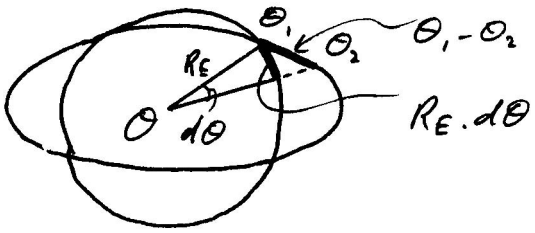

Figure 2.5: Ellipticity of the Earth measured by the distance between latitudes of the Earth and a sphere.

In 1743, Clairaut[1] showed that the reference spheroid can also be computed directly from the measured gravity field $\mathbf{g}$. The derivation is based on the computation of a potential $U(P)$ at point P due to a nearly spherical body, and it is only valid for points outside (or, in the limit, on the surface of) the body.

The contribution dU to the gravitational potential at P due to a mass element dM at distance q from P is given by

$$dU = -\frac{G}{q}\,dM \tag{2.16}$$

Typically, the potential is expanded in a series. This can be done in two ways, which lead to the same results. One can write $U(P)$ directly in terms of the known solutions of Laplace's equation ($\nabla^2 U = 0$), which are spherical harmonics. Alternatively, one can expand the term $1/q$ and integrate the resulting series term by term. Here, we will do the latter because it gives better

[1] In his book, *Théorie de la Figure de la Terre.*

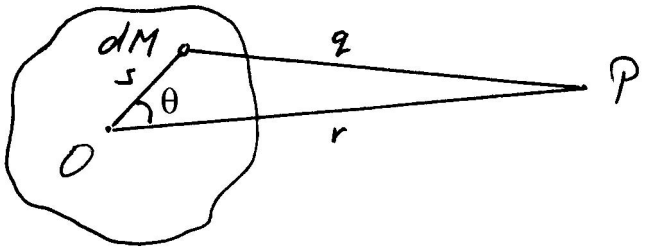

Figure 2.6: The potential U of the aspherical body is calculated at point P, which is external to the mass $M = \int dM$; $OP = r$, the distance from the observation point to the center of mass. Note that r is constant and that s , q, and θ are the variables. There is no rotation so $U(P)$ represents the gravitational potential.

understanding of the physical meaning of the terms, but we will show how these terms are, in fact, directly related to (zonal) spherical harmonics. A formal treatment of solutions of spherical harmonics as solutions of Laplace's equation follows later. The derivation discussed here leads to what is known as MacCullagh's formula[2] and shows how the gravity measurements themselves are used to define the reference spheroid. Using Figure 2.6 and the law of cosines we can write $q^2 = r^2 + s^2 - 2rs\cos\theta$ so that

$$dU = -\frac{G}{r\left[1 + \left(\frac{s}{r}\right)^2 - 2\left(\frac{s}{r}\right)\cos\Psi\right]^{\frac{1}{2}}}\, dM \tag{2.17}$$

We can use the Binomial Theorem to expand this expression into a power series of (s/r). So we can write:

$$\begin{aligned}\left[1 + \left(\frac{s}{r}\right)^2 - 2\left(\frac{s}{r}\right)\cos\theta\right]^{-\frac{1}{2}} &= 1 - \frac{1}{2}\left(\frac{s}{r}\right)^2 \\ &+ \left(\frac{s}{r}\right)\cos\theta + \frac{3}{2}\left(\frac{s}{r}\right)^2\left(\cos^2\theta\right) + \text{h.o.t.} \\ &= 1 + \left(\frac{s}{r}\right)\cos\theta + \frac{1}{2}\left(\frac{s}{r}\right)^2\left(3\cos^2\theta - 1\right) \\ &+ \text{h.o.t.} \end{aligned} \tag{2.18}$$

and for the potential:

$$U(P) = \int_V dU$$

[2]After James MacCullagh (1809–1847).

$$\begin{aligned} &= -\frac{G}{r}\int\left[1+\left(\frac{s}{r}\right)\cos\theta+\frac{1}{2}\left(\frac{s}{r}\right)^2\left(3\cos^2\theta-1\right)\right]\,dM \\ &= -\frac{G}{r}\int dM-\frac{G}{r^2}\int s\cos\theta\,dM \\ &- \frac{G}{2r^3}\int s^2\left(3\cos^2\theta-1\right)\,dM \end{aligned} \tag{2.19}$$

In Equation 2.19 we have ignored the higher order terms (h.o.t). Let us rewrite eq. (2.19) by using the identity $\cos^2\theta+\sin^2\theta=1$:

$$U(P)=-\frac{G}{r}\int dM-\frac{G}{r^2}\int s\cos\theta\,dM-\frac{G}{r^3}\int s^2\,dM+\frac{3}{2}\frac{G}{r^3}\int s^2\sin^2\theta\,dM \tag{2.20}$$

Intermezzo 2.3 BINOMIAL THEOREM

$$\begin{aligned} (a+b)^n &= a^n+na^{n-1}b+\frac{1}{2!}n(n-1)a^{n-2}b^2 \\ &+ \frac{1}{3!}n(n-1)(n-2)a^{n-3}b^3+\ldots \end{aligned} \tag{2.21}$$

for $|\frac{b}{a}|<1$. Here we take $b=\left[\left(\frac{s}{r}\right)^2-2\left(\frac{s}{r}\right)\cos\theta\right]$ and $a=1$.

Intermezzo 2.4 EQUIVALENCE WITH (ZONAL) SPHERICAL HARMONICS

Note that equation (2.19) is, in fact, a power series of (s/r), with the multiplicative factors functions of $cos(\theta)$:

$$\begin{aligned} U(P) &= -\frac{G}{r}\int\left[1\left(\frac{s}{r}\right)^0+\cos\theta\left(\frac{s}{r}\right)^1+\left(\frac{3}{2}\cos^2\theta-\frac{1}{2}\right)\left(\frac{s}{r}\right)^2\right]dM \\ &= -\frac{G}{r}\int\sum_{l=0}^{2}P_l(\cos\theta)\left(\frac{s}{r}\right)^l dM \end{aligned} \tag{2.22}$$

In spectral analysis there are special names for the factors P_l multiplying $(s/r)^l$ and these are known as **Legendre polynomials**, which define the *zonal surface spherical harmonics*[a].
We will discuss spherical harmonics in detail later but here it is useful to point out the similarity between the above expression of the potential $U(P)$ as a power series of (s/r) and $\cos\theta$ and the lower order spherical-harmonics. Legendre polynomials are defined as

$$P_l(\mu)=\frac{1}{2^l l!}\frac{d^l(\mu^2-1)^l}{d\mu^l} \tag{2.23}$$

with μ some function. In our case we take $\mu=\cos\theta$ so that the superposition of the Legendre polynomials describes the variation of the potential with latitude. At this stage we ignore variations with longitude. Surface spherical harmonics that depend on latitude only are known as *zonal* spherical harmonics. For $l=0,1,2$ we get for P_l

$$P_0(\cos\theta) = 1 \tag{2.24}$$
$$P_1(\cos\theta) = \cos\theta \tag{2.25}$$
$$P_2(\cos\theta) = \frac{3}{2}\cos^2\theta-\frac{1}{2} \tag{2.26}$$

which are the same as the terms derived by application of the binomial theorem. The equivalence between the potential expression in spherical harmonics and the one that we are deriving by expanding $1/q$ is no coincidence: the potential U satisfies Laplace's equation and in a spherical coordinate system spherical harmonics are the general solutions of Laplace's equation.

[a]Surface spherical harmonics are at the surface of a sphere what a Fourier series is to a time series; it can be thought of as a 2D Fourier series which can be used to represent any quantity at the surface of a sphere (geoid, temperature, seismic wave speed).

We can get insight in the physics if we look at each term of eq. (2.20) separately:

1. $-\frac{G}{r}\int dM=-\frac{GM}{r}$ is essentially the potential of a point mass M at O. This term will dominate for large r; at a large distance the potential due

to an aspherical density distribution is close to that of a spherical body (i.e., a point mass in O).

2. $\int s\cos\theta\, dM$represents a torque of mass $\times$ distance, which also underlies the definition of the center of mass $\mathbf{r_{cm}} = \int r\, dM / \int dM$. In our case, we have chosen O as the center of mass and $\mathbf{r_{cm}} = \mathbf{0}$ with respect to O. Another way to see that this integral must vanish is to realize that the integration over dM is essentially an integration over θ between 0 and 2π and that $\cos\theta = -\cos(\frac{\pi}{2} - \theta)$. Integration over θ takes $s\cos\theta$ back and forth over the line between O and P (within the body) with equal contributions from each side of O, since O is the center of mass.

3. $\int s^2\, dM$ represents the torque of a distance squared and a mass, which underlies the definition of the **moment of inertia** (recall that for a homogeneous sphere with radius R and mass M the moment of inertia is 0.4 MR^2). The moment of inertia is defined as $I = \int r^2\, dM$. When talking about moments of inertia one must identify the axis of rotation. We can understand the meaning of the third integral by introducing a coordinate system x, y, z so that $s = (x, y, z)$, $s^2 = x^2 + y^2 + z^2$ so that $\int s^2\, dM = \int(x^2+y^2+z^2)\, dM = 1/2[\int(y^2+z^2)\, dM + \int(x^2+z^2)\, dM + \int(x^2+y^2)\, dM]$ and by realizing that $\int(y^2 + z^2)\, dM$, $\int(x^2 + z^2)\, dM$ and $\int(x^2 + y^2)\, dM$ are the moments of inertia around the x-, y-, and z-axis respectively. See Intermezzo 2.5 for more on moments of inertia.

   With the moments of inertia defined as in the box we can rewrite the third term in the potential equation

$$-\frac{G}{r^3}\int s^2\, dM = -\frac{G}{2r^3}(A + B + C) \tag{2.27}$$

4. $\int s^2 \sin^2\theta\, dM$. Here, $s\sin\theta$ projects $\mathbf{s}$ on a plane perpendicular to OP and this integral thus represents the moment of inertia of the body around OP. This moment is often denoted by I.

Eq. (2.20) can then be rewritten as

$$U(P) = -\frac{GM}{r} - \frac{G}{2r^3}(\mathrm{A} + \mathrm{B} + \mathrm{C} - 3\mathrm{I}) \tag{2.28}$$

which is known as **MacCullagh's formula**.

At face value this seems to be the result of a straightforward and rather boring derivation, but it does reveal some interesting and important properties of the potential and the related field. Equation (2.20) basically shows that in absence of rotation the gravitational attraction of an irregular body has two contributions; the first is the attraction of a point mass located at the center of gravity, the second term depends on the moments of inertia around the principal axes, which in turn depend completely on the *shape* of the body, or, more precisely, on the deviations of the shape from a perfect sphere. This second

term decays as $1/r^3$ so that at large distances the potential approaches that of a point mass M and becomes less and less sensitive to aspherical variations in the shape of the body. This simply implies that if you're interested in small scale deviations from spherical symmetry you should not be to far away from the surface: i.e. it's better to use data from satellites with a relatively low orbit. This phenomenon is in fact an example of up (or down)ward continuation, which we will discuss more quantitatively formally when introducing spherical harmonics.

Intermezzo 2.5 MOMENTS AND PRODUCTS OF INERTIA

A moment of inertia of a rigid body is defined with respect to a certain axis of rotation.

For discrete masses: $I = m_1 r_1^2 + m_2 r_2^2 + m_3 r_3^2 + \ldots = \sum m_i r_i^2$
and for a continuum: $I = \int r^2 \, dM$

The moment of inertia is a tensor quantity

$$\mathrm{MI} = \int r^2 (\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}^{\mathrm{T}}) \, dM \tag{2.29}$$

Note: we revert to matrix notation and manipulation of tensors. $\mathbf{I}$ is a second-order tensor.
$(\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}^{\mathrm{T}})$ is a projection operator: for instance, $(\mathbf{I} - \hat{\mathbf{z}}\hat{\mathbf{z}}^{\mathrm{T}})\mathbf{a}$ projects the vector $\mathbf{a}$ on the (x,y) plane, i.e., perpendicular to $\hat{\mathbf{z}}$. This is very useful in the general expression for the moments of inertia around different axis.

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.30}$$

and

$$r^2 I = \begin{pmatrix} x^2+y^2+z^2 & 0 & 0 \\ 0 & x^2+y^2+z^2 & 0 \\ 0 & 0 & x^2+y^2+z^2 \end{pmatrix} \tag{2.31}$$

$$\begin{aligned} \text{and} \quad r^2 \hat{\mathbf{r}}\hat{\mathbf{r}}^{\mathrm{T}} &= r\hat{\mathbf{r}} \cdot r\hat{\mathbf{r}}^{\mathrm{T}} = \mathbf{r} \cdot \mathbf{r}^{\mathrm{T}} \\ &= \begin{pmatrix} x \\ y \\ z \end{pmatrix} \begin{pmatrix} x & y & z \end{pmatrix} = \begin{pmatrix} x^2 & xy & xz \\ yx & y^2 & yz \\ zx & zy & z^2 \end{pmatrix} \end{aligned} \tag{2.32}$$

So that:

$$r^2(\mathbf{I} - \hat{\mathbf{r}}\hat{\mathbf{r}}^{\mathrm{T}}) = \begin{pmatrix} y^2+z^2 & -xy & -xz \\ -yx & x^2+z^2 & -yz \\ -zx & -zy & x^2+y^2 \end{pmatrix} \tag{2.33}$$

The diagonal elements are the familiar **moments of inertia** around the x, y, and z axis. (The off-diagonal elements are known as the **products of inertia**, which vanish when we choose x, y, and z as the principal axes.)

Moment of Inertia around x-axis
around y-axis
around z-axis

$I_{xx} = \int (y^2 + z^2) \, dM = A$
$I_{yy} = \int (x^2 + z^2) \, dM = B$
$I_{zz} = \int (x^2 + y^2) \, dM = C$

We can pursue the development further by realizing that the moment of

inertia I around any general axis (here OP) can be expressed as a linear combination of the moments of inertia around the principal axes. Let l^2, m^2, and n^2 be the squares of the cosines of the angle of the line OP with the x-, y-, and z-axis, respectively. With $l^2 + m^2 + n^2 = 1$ we can write $\mathrm{I} = l^2\mathrm{A} + m^2\mathrm{B} + n^2\mathrm{C}$ (see Figure 2.7).

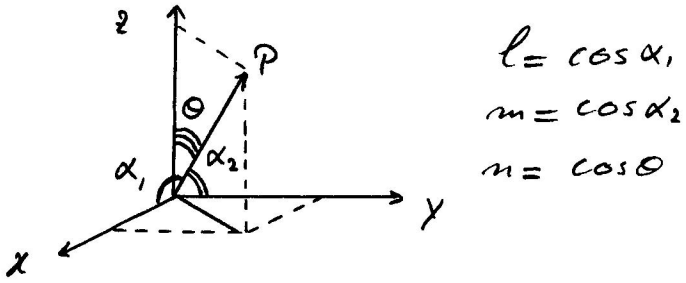

Figure 2.7: Definition of direction cosines.

So far we have not been specific about the shape of the body, but for the Earth it is relevant to consider rotational geometry so that $\mathrm{A} = \mathrm{B} \neq \mathrm{C}$. This leads to:

$$\mathrm{I} = \mathrm{A} + (\mathrm{C} - \mathrm{A})n^2 \tag{2.34}$$

Here, $n = \cos\theta$ with θ the angle between OP and the z-axis, that is θ is the **co-latitude**. ($\theta = 90 - \lambda$, where λ is the **latitude**).

$$\mathrm{I} = \mathrm{A} + (\mathrm{C} - \mathrm{A})\cos^2\theta \tag{2.35}$$

and

$$U(P) = -\frac{GM}{r} + \frac{G}{r^3}(\mathrm{C} - \mathrm{A})\left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right) \tag{2.36}$$

It is customary to write the difference in moments of inertia as a fraction J_2 of Ma^2, with a the Earth's radius at the equator.

$$\mathrm{C} - \mathrm{A} = J_2 M a^2 \tag{2.37}$$

so that

$$U(P) = -\frac{GM}{r} + \frac{GJ_2Ma^2}{r^3}\left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right) \tag{2.38}$$

J_2 is a measure of ellipticity; for a sphere $\mathrm{C} = \mathrm{A}$, $J_2 = 0$, and the potential $U(P)$ reduces to the expression of the gravitational potential of a body with spherical symmetry.

Intermezzo 2.6 ELLIPTICITY TERMS

Let's briefly return to the equivalence with the spherical harmonic expansion. If we take $\mu = \cos\theta$ (see box) we can write for $U(P)$

$$\begin{aligned} U(P) &= U(r,\theta) \\ &= -\frac{GM}{r}[J_0 P_0(\cos\theta) + J_1 \left(\frac{a}{r}\right) P_1(\cos\theta) \\ &+ J_2 \left(\frac{a}{r}\right)^2 P_2(\cos\theta)] \end{aligned} \tag{2.39}$$

The expressions (2.20), rewritten as (2.38), and (2.39) are identical if we define the scaling factors J_l as follows. Since $P_0(\cos\theta) = 1$, J_0 must be 1 because $-GM/r$ is the far field term; $J_1 = 0$ if the coordinate origin coincides with the center of mass (see above); and J_2 is as defined above. This term is of particular interest since it describes the oblate shape of the geoid. (The higher order terms (J_4, J_6 etc.) are smaller by a factor of order 1000 and are not carried through here, but they are incorporated in the calculation of the reference spheroid.)

The final step towards calculating the reference gravity field is to add a rotational potential.

Let $\boldsymbol{\omega} = \omega\hat{\mathbf{z}}$ be the angular velocity of rotation around the z-axis. The choice of reference frame is important to get the plus and minus signs right. A particle that moves with the rotating earth is influenced by a centripetal force $\mathbf{F}_{\text{cp}} = m\mathbf{a}$, which can formally be written in terms of the cross products between the angular velocity ω and the position vector as $m\boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{s})$. This shows that the centripetal acceleration points to the rotation axis. The magnitude of the force per unit mass is $s\omega^2 = r\omega^2 \cos\lambda$. The source of $\mathbf{F}_{\text{cp}}$ is, in fact, the gravitational attraction $\mathbf{g}$ ($\mathbf{g}_{\text{eff}} + \frac{\mathbf{F}_{\text{cp}}}{m} = \mathbf{g}$). The effective gravity $\mathbf{g}_{\text{eff}} = \mathbf{g} - \frac{\mathbf{F}_{\text{cp}}}{m}$ (see Figure 2.8). Since we are mainly interested in the radial component (the tangential component is very small) we can write $g_{\text{eff}} = g - r\omega^2 \cos^2\lambda$.

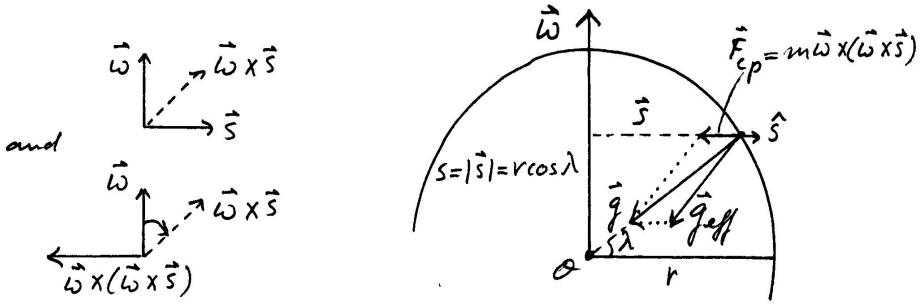

Figure 2.8: The gravitational attraction produces the centripetal force due to the rotation of the Earth.

In terms of potentials, the rotational potential has to be added to the grav-

itational potential $U_{\text{gravity}} = U_{\text{gravitation}} + U_{\text{rot}}$, with

$$U_{\text{rot}} = -\int r\omega^2 \cos^2\lambda \, dr = -\frac{1}{2}r^2\omega^2\cos^2\lambda = -\frac{1}{2}r^2\omega^2\sin^2\theta \tag{2.40}$$

(which is in fact exactly the rotational kinetic energy ($\mathrm{K} = \frac{1}{2}\mathrm{I}\omega^2 = \frac{1}{2}mr^2\omega^2$) per unit mass of a rigid body $-\frac{1}{2}\omega^2 r^2 = -\frac{1}{2}v^2$, even though we used an approximation by ignoring the component of $\mathbf{g}_{\text{eff}}$ in the direction of varying latitude $d\lambda$. Why? Hint: use the above diagram and consider the symmetry of the problem)

The geopotential can now be written as

$$U(r,\theta) = -\frac{GM}{r} + \frac{G}{r^3}J_2 M a^2\left(\frac{3}{2}\cos^2\theta - \frac{1}{2}\right) - \frac{1}{2}r^2\omega^2\sin^2\theta \tag{2.41}$$

which describes the contribution to the potential due to the central mass, the oblate shape of the Earth (i.e. flattening due to rotation), and the rotation itself.

We can also write the geopotential in terms of the latitude by substituting ($\sin\lambda = \cos\theta$):

$$U(r,\lambda) = -\frac{GM}{r} + \frac{G}{r^3}(\mathrm{C}-\mathrm{A})\left(\frac{3}{2}\sin^2\lambda - \frac{1}{2}\right) - \frac{1}{2}r^2\omega^2\cos^2\lambda \tag{2.42}$$

We now want to use this result to find an expression for the gravity potential and acceleration at the surface of the (reference) spheroid. The flattening is determined from the geopotential by defining the equipotential U_0, the surface of constant U.

Since U_0 is an equipotential, U must be the same (U_0) for a point at the pole and at the equator. We take c for the polar radius and a for the equatorial radius and write:

$$U_{0,\text{pole}} = U(c,90) = U_{0,\text{equator}} = U(a,0) \tag{2.43}$$

$$U_{\text{pole}} = -\frac{GM}{c} + \frac{G}{c^3}J_2 M a^2 \tag{2.44}$$

$$U_{\text{equator}} = -\frac{GM}{a} - \frac{G}{2a^3}J_2 M a^2 - \frac{1}{2}a^2\omega^2 \tag{2.45}$$

$$\tag{2.46}$$

and after some reordering to isolate a and c we get

$$f \equiv \frac{a-c}{a} \approx \frac{3}{2}\left(\frac{J_2 M a^2}{M a^2}\right) + \frac{1}{2}\frac{a\omega^2}{GM/a^2} = \frac{3}{2}J_2 + \frac{1}{2}m \tag{2.47}$$

Which basically shows that the geometrical flattening f as defined by the relative difference between the polar and equatorial radius is related to the ellipticity coefficient J_2 and the ratio m between the rotational ($a\omega^2$) to the

gravitational (GMa^{-2}) component of gravity at the equator. The value for the flattening f can be accurately determined from orbital data; in fact within a year after the launch of the first artificial satellite — by the soviets — this value could be determined with much more accuracy than by estimates given by many investigators in the preceding centuries. The geometrical flattening is small ($f = 1/298.257 \approx 1/300$) (but larger than expected from equilibrium flattening of a rotating body). The difference between the polar and equatorial radii is thus about $R_E f = 6371\text{km}/300 \approx 21$ km.

In order to get the shape of the reference geoid (or *spheroid*) one can use the assumption that the deviation from a sphere is small, and we can thus assume the vector from the Earth's center to a point at the reference geoid to be of the form

$$r_g \sim r_0 + dr = r_0(1+\epsilon) \quad \text{or, with} \quad r_0 = a \ , \ r_g \sim a(1+\epsilon) \tag{2.48}$$

It can be shown that ϵ can be written as a function of f and latitude as given by: $r_g \sim a(1 - f\sin^2\lambda)$ and (from binomial expansion) $r_g^{-2} \approx a^{-2}(1+2f\sin^2\lambda)$.

Geoid anomalies, i.e. the geoid "highs" and "lows" that people talk about are deviations from the reference geoid and they are typically of the order of several tens of meters (with a maximum (absolute) value of about 100 m near India), which is small (often less than 0.5%) compared to the latitude dependence of the radius (see above). So the reference geoid with $r = r_g$ according to (2.48) does a pretty good job in representing the average geoid.

Finally, we can determine the gravity field at the reference geoid with a shape as defined by (2.48) calculating the gradient of eqn. (2.42) and substituting the position r_g defined by (2.48).

In spherical coordinates:

$$\mathbf{g} = -\boldsymbol{\nabla} U = -\left(\frac{\partial U}{\partial r}, \frac{1}{r}\frac{\partial U}{\partial \lambda}\right) \tag{2.49}$$

$$g = |\mathbf{g}| = \left\{\left(\frac{\partial U}{\partial r}\right)^2 + \left(\frac{1}{r}\frac{\partial U}{\partial \lambda}\right)^2\right\}^{\frac{1}{2}} \sim \frac{\partial U}{\partial r} \tag{2.50}$$

because $\frac{1}{r}\frac{\partial U}{\partial \lambda}$ is small.

So we can approximate the magnitude of the gravity field by:

$$g = \frac{GM}{r^2} - 3\frac{GJ_2Ma^2}{r^4}\left(\frac{3}{2}\sin^2\lambda - \frac{1}{2}\right) - r\omega^2\cos^2\lambda \tag{2.51}$$

and, with $r = r_g = a\left(1 - f\sin^2\lambda\right)$

$$\begin{aligned} g &= \frac{GM}{a^2(1-f\sin^2\lambda)^2} - \frac{3GJ_2Ma^2}{a^4(1-f\sin^2\lambda)^4}\left(\frac{3}{2}\sin^2\lambda - \frac{1}{2}\right) \\ &- a\omega^2(1-f\sin^2\lambda)\cos^2\lambda \end{aligned} \tag{2.52}$$

or, with the approximation (binomial expansion) given below Eqn. (2.48)

$$\begin{aligned} g(\lambda) &= \frac{GM}{a^2}\left\{(1+2f\sin^2\lambda) - 3J_2\left(\frac{3}{2}\sin^2\lambda - \frac{1}{2}\right) - m(1-\sin^2\lambda)\right\} \\ &= \frac{GM}{a^2}\left\{(1+\frac{3}{2}J_2 - m) + \left(2f - \frac{9}{2}J_2 + m\right)\sin^2\lambda\right\} \\ &= \frac{GM}{a^2}(1+\frac{3}{2}J_2 - m)\left\{1 + \left(\frac{2f-(9/2)J_2+m}{1+(3/2)J_2-m}\right)\sin^2\lambda\right\} \\ &= \frac{GM}{a^2}(1+\frac{3}{2}J_2 - m)\left\{1 + f'\sin^2\lambda\right\} \end{aligned} \tag{2.53}$$

Eqn. (2.53) shows that the gravity field at the reference spheroid can be expressed as some latitude-dependent factor times the gravity acceleration at the equator:

$$g_{\rm eq}(\lambda = 0) = \frac{GM}{a^2}\left(1 + \frac{3}{2}J_2 - m\right) \tag{2.54}$$

Information about the flattening can be derived directly from the relative change in gravity from the pole to the equator.

$$g_{\rm pole} = g_{\rm eq}(1+f') \rightarrow f' = \frac{g_{\rm pole} - g_{\rm eq}}{g_{\rm eq}} \tag{2.55}$$

Eq. 2.55 is called **Clairaut's theorem**[3]. The above quadratic equation for the gravity as a function of latitude (2.53) forms the basis for the international gravity formula. However, this international reference for the reduction of gravity data is based on a derivation that includes some of the higher order terms. A typical form is

$$g = g_{\rm eq}(1 + \alpha\sin 2\lambda + \beta\sin^2 2\lambda) \tag{2.56}$$

with the factor of proportionality α and β depending on GM, ω, a, and f. The values of these parameters are being determined more and more accurate by the increasing amounts of satellite data and as a result the international gravity formula is updated regularly. The above expression (2.56) is also a truncated series. A closed form expression for the gravity as function of latitude is given by the **Somigliana Equation**[4]

$$g(\lambda) = g_{\rm eq}\left\{\frac{1 + k\sin^2\lambda}{\sqrt{1 - e^2\sin^2\lambda}}\right\}. \tag{2.57}$$

This expression has now been adopted by the Geodetic Reference System and forms the basis for the reduction of gravity data to the reference geoid (or reference spheroid). $g_{\rm eq} = 9.7803267714$ ms^{-2}; $k = 0.00193185138639$; $e = 0.00669437999013$.

[3] After Alexis Claude Clairaut (1713–1765).

[4] After C. Somigliana.

2.3 The Poisson and Laplace equations

The gravitational field of the Earth is caused by its density. The mass distribution of the planet is inherently three-dimensional, but we mortals will always only scratch at the surface. The most we can do is measure the gravitational acceleration at the Earth's surface. However, thanks to a fundamental relationship known as **Gauss's Theorem**[5], the link between a surface observable and the properties of the whole body in question can be found. Gauss's theorem is one of a class of theorems in vector analysis that relates integrals of different types (line, surface, volume integrals). Stokes's, Greens and Gauss's theorem are fundamental in the study of potential fields. The theorem due to Gauss relates the integral over the volume of some property (most generally, a tensor $\mathbf{T}$) to a surface integral. It is also called the divergence theorem. Let V be a volume bounded by the surface $S = \partial V$ (see Figure 2.9). A differential patch of surface $d\mathbf{S}$ can be represented by an outwardly pointing vector with a length corresponding to the area of the surface element. In terms of a unit normal vector, it is given by $\hat{\mathbf{n}}\|d\mathbf{S}\|$.

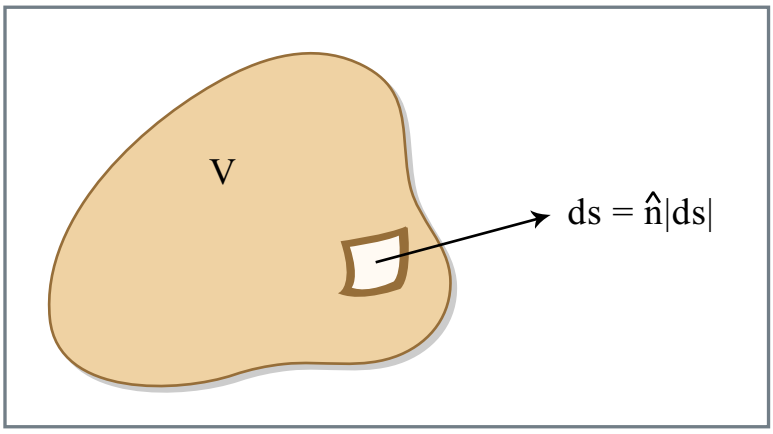

Figure by MIT OCW.

Figure 2.9: Surface enclosing a volume. Unit normal vector.

Gauss's theorem (for generic "stuff" $\mathbf{T}$) is as follows:

$$\boxed{\int_V \boldsymbol{\nabla} \cdot \mathbf{T}\, dV = \int_{\partial V} \hat{\mathbf{n}} \cdot \mathbf{T}\, dS.} \tag{2.58}$$

Let's see what we can infer about the gravitational potential within the Earth using only information obtained at the surface. Remember we had

$$g = \frac{GM}{r^2} \quad \text{and} \quad \mathbf{g} = -\boldsymbol{\nabla} U. \tag{2.59}$$

[5] After Carl-Friedrich Gauss (1777–1855).

Suppose we measure $\mathbf{g}$ everywhere at the surface, and sum the results. What we get is the **flux** of the gravity field

$$\int_{\partial V} \mathbf{g} \cdot \mathbf{dS}. \tag{2.60}$$

At this point, we can already predict that if S is the surface enclosing the Earth, the flux of the gravity field should be different from zero, and furthermore, that it should have something to do with the density distribution within the planet. Why? Because the gravitational field lines all point towards the center of mass. If the flux was zero, the field would be said to be **solenoidal** . Unlike the magnetic field the gravity field is essentially a monopole. For the magnetic field, field lines both leave and enter the spherical surface because the Earth has a positive and a negative pole. The gravitational field is only solenoidal in regions not occupied by mass.

Anyway, we'll start working with Eq. 2.60 and see what we come up with. On the one hand (we use Eq. 2.58 and Eq. 2.59)[6],

$$\int_{\partial V} \mathbf{g} \cdot \hat{\mathbf{n}}\, dS = \int_V \boldsymbol{\nabla} \cdot \mathbf{g}\, dV = -\int_V \boldsymbol{\nabla} \cdot \boldsymbol{\nabla} U\, dV = -\int_V \boldsymbol{\nabla}^2 U\, dV. \tag{2.61}$$

On the other hand (we use the definition of the dot product and Eq. 2.59, and define g_n as the component of $\mathbf{g}$ normal to dS):

$$\int_{\partial V} \mathbf{g} \cdot \hat{\mathbf{n}}\, dS = -\int_{\partial V} g_n\, dS = -4\pi r^2 \frac{GM}{r^2} = -4\pi G \int_V \rho\, dV. \tag{2.62}$$

We've assumed that S is a spherical surface, but the derivation will work for any surface. Equating Eq. 2.61 and 2.62, we can state that

$$\boxed{\nabla^2 U(\mathbf{r}) = 4\pi G \rho(\mathbf{r})} \quad \text{Poisson's Equation} \tag{2.63}$$

and in the homogeneous case

$$\boxed{\nabla^2 U(\mathbf{r}) = 0} \quad \text{Laplace's Equation} \tag{2.64}$$

The interpretation in terms of sources and sinks of the potential fields and its relation with the field lines is summarized in Figure 2.10:

Poisson's equation is a fundamental result. It implies

1. that the *total* mass of a body (say, Earth) can be determined from measurements of $\boldsymbol{\nabla} U = -\mathbf{g}$ at the surface (see Eq. 2.62), and

2. no information is required about how exactly the density is distributed within V

[6] Note that the identity $\boldsymbol{\nabla}^2 U = \boldsymbol{\nabla} \cdot \boldsymbol{\nabla} U$ is true for scalar fields, but for a vector field $\mathbf{V}$ we should have written $\nabla^2 \mathbf{V} = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \mathbf{V}) - \boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \mathbf{V})$.

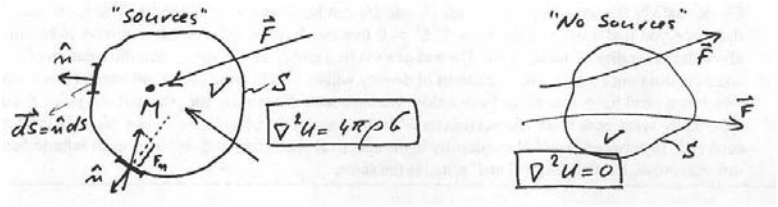

Figure 2.10: Poisson's and Laplace's equations.

If there is no potential source (or sink) enclosed by S Laplace's equation should be applied to find the potential at a point P *outside* the surface S that contains all attracting mass, for instance the potential at the location of a satellite. But in the limit, it is also valid at the Earth's surface. Similarly, we will see that we can use Laplace's equation to describe Earth's magnetic field as long as we are outside the region that contains the source for the magnetic potential (i.e., Earth's core).

We often have to find a solution for U of Laplace's equation when only the value of U, or its derivatives $|\boldsymbol{\nabla} U| = g$ are known at the surface of a sphere. For instance if one wants to determine the internal mass distribution of the Earth from gravity data. Laplace's equation is easier to solve than Poisson's equation. In practice one can usually (re)define the problem in such a way that one can use Laplace's equation by integrating over contributions from small volumes dV (containing the source of the potential dU, i.e., mass dM), see Figure 2.11 or by using Newton's Law of Gravity along with Laplace's equation in an iterative way.

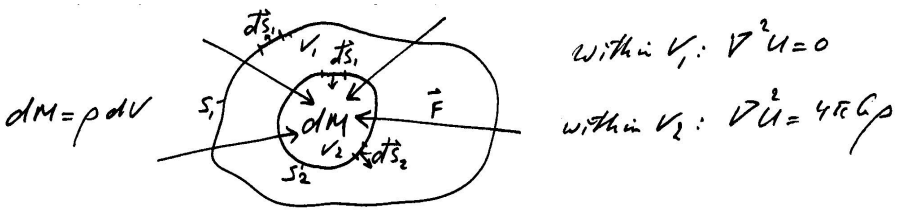

Figure 2.11: Applicability of Poisson's and Laplaces' equations.

See Intermezzo 2.7.

Intermezzo 2.7 NON-UNIQUENESS

One can prove that the solution of Laplace's equation can be uniquely determined if the boundary conditions are known (i.e. if data coverage at the surface is good); in other words, if there are two solutions U_1 and U_2 that satisfy the boundary conditions, U_1 and U_2 can be shown to be identical. The good news here is that once you find a solution for U of $\nabla^2 U = 0$ that satisfies the BC's you do not have to be concerned about the generality of the solution. The bad news is (see also point (2) above) that the solution of Laplace's equation does not constrain the variations of density within V. This leads to a fundamental non-uniqueness which is typical for potentials of force fields. We have seen this before: the potential at a point P outside a spherically symmetric body with total mass M is the same as the potential of a point mass M located in the center O. In between O and P the density in the spherical shells can be distributed in an infinite number of different ways, but the potential at P remains the same.

2.4 Cartesian and spherical coordinate systems

In Cartesian coordinates we write for ∇^2 (the Laplacian)

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{2.65}$$

For the Earth, it is advantageous to use spherical coordinates. These are defined as follows (see Figure 2.12):

$$\left\{ \begin{array}{lcl} x & = & r \sin\theta \cos\varphi \\ y & = & r \sin\theta \sin\varphi \\ z & = & r \cos\theta \end{array} \right. \tag{2.66}$$

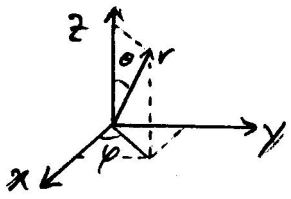

Figure 2.12: Definition of r, θ and φ in the spherical coordinate system.

where $\theta = 0 \rightarrow \pi$ = co-latitude, $\varphi = 0 \rightarrow 2\pi$ = longitude.

It is very important to realize that, whereas the Cartesian frame is described by the immobile unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$, the unit vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\varphi}}$ are dependent on the position of the point. They are local axes. At point P, $\hat{\mathbf{r}}$ points in the direction of increasing radius from the origin, $\hat{\boldsymbol{\theta}}$ in the direction of increasing colatitude θ and $\hat{\boldsymbol{\varphi}}$ in the direction of increasing longitude φ.

One can go between coordinate axes by the transformation

$$\begin{pmatrix} \hat{\mathbf{r}} \\ \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\varphi}} \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\varphi & \sin\theta\sin\varphi & \cos\theta \\ \cos\theta\cos\varphi & \cos\theta\sin\varphi & -\sin\theta \\ -\sin\varphi & \cos\varphi & 0 \end{pmatrix} \begin{pmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \\ \hat{\mathbf{z}} \end{pmatrix} \tag{2.67}$$

Furthermore, we need to remember that integration over a volume element $dx\,dy\,dz$ becomes, after changing of variables $r^2\sin\theta\,dr\,d\theta\,d\varphi$. This may be remembered by the fact that $r^2\sin\theta$ is the determinant of the Jacobian matrix, i.e. the matrix obtained by filling a 3×3 matrix with all partial derivatives of Eq. 2.66. After some algebra, we can write the spherical Laplacian:

$$\nabla^2 U = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial U}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial U}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\left(\frac{\partial^2 U}{\partial\varphi^2}\right) = 0. \tag{2.68}$$

2.5 Spherical harmonics

We now attempt to solve Laplace's Equation $\nabla^2 U = 0$, in spherical coordinates. Laplace's equation is obeyed by potential fields outside the sources of the field. Remember how sines and cosines (or in general, exponentials) are often solutions to differential equations, of the form $\sin kx$ or $\cos kx$, whereby k can take any integer value. The general solution is any combination of sines and cosines of all possible k's with weights that can be determined by satisfying the boundary conditions (BC's). The particular solution is constructed by finding a linear combination of these (basis) functions with weighting coefficients dictated by the BC's: it is a series solution. In the Cartesian case they are Fourier Series. In Fourier theory, a signal, say a time series $s(t)$, for instance a seismogram, can be represented by the superposition of cos and sin functions and weights can be found which approximate the signal to be analyzed in a least-squares sense.

Spherical harmonics are solutions of the spherical Laplace's Equation: they are basically an adaption of Fourier analysis to a spherical surface. Just like with Fourier series, the superposition of spherical harmonics can be used to represent and analyze physical phenomena distributed on the surface on (or within) the Earth. Still in analogy with Fourier theory, there exists a sampling theorem which requires that sufficient data are provided in order to make the solution possible. In geophysics, one often talks about (spatial) *data coverage*, which must be adequate.

We can find a solution for U of $\nabla^2 U = 0$ by the good old trick of separation of variables. We look for a solution with the following structure:

$$U(r,\theta,\varphi) = R(r)P(\theta)Q(\varphi) \tag{2.69}$$

Let's take each factor separately. In the following, an outline is given of how to find the solution of this elliptic equation, but working this out rigorously requires some more effort than you might be willing to spend. But let's not try to lose the physical meaning what we come up with.

Radial dependence: $R(r)$

It turns out that the functions satisfying Laplace's Equation belong to a special class of homogeneous[7] harmonic[8] functions. A first property of homogeneous functions that can be used to our advantage is that in general, a homogeneous function can be written in two different forms:

$$U_1(r,\theta,\varphi) = r^l Y_l(\theta,\varphi) \tag{2.70}$$

$$U_2(r,\theta,\varphi) = \left(\frac{1}{r}\right)^{(l+1)} Y_l(\theta,\varphi) \tag{2.71}$$

This, of course, gives the form of our radial function:

$$R(r) = \left\{ \begin{array}{c} r^l \\ \left(\frac{1}{r}\right)^{l+1} \end{array} \right\} \tag{2.72}$$

The two alternatives $R(r) = r^l$ and $R(r) = (1/r)^{l+1}$ describe the behavior of U for an external and internal field, respectively (in- and outside the mass distribution). Whether to use $R(r) = r^l$ and $R(r) = (1/r)^{l+1}$ depends on the problem you're working on and on the boundary conditions. If the problem requires a finite value for U at $r = 0$ than we need to use $R(r) = r^l$. However if we require $U \to 0$ for $r \to \infty$ then we have to use $R(r) = (1/r)^{l+1}$. The latter is appropriate for representing the potential outside the surface that encloses all sources of potential, such as the gravity potential $U = GMr^{-1}$. However, both are needed when we describe the magnetic potential at point r due to an internal and external field.

Longitudinal dependence: $Q(\varphi)$

Substitution of Eq. 2.69 into Laplace's equation with $R(r)$ given by Eq. 2.72, and dividing Eq. 2.69 out again yields an equation in which θ- and φ-derivatives occur on separate sides of the equation sign. For arbitrary θ and φ this must mean:

$$-\frac{\frac{d^2}{d\varphi^2}Q}{Q} = \text{constant}, \tag{2.73}$$

which is best solved by calling the constant m^2 and solving for Q as:

$$Q(\varphi) = A\cos m\varphi + B\sin m\varphi. \tag{2.74}$$

Indeed, all possible constants A and B give valid solutions, and m must be a positive integer.

[7] A homogenous function f of degree n satisfies $f(tx,ty,tz) = t^n f(x,y,z)$.

[8] By definition, a function which satisfies Laplace's equation is called harmonic.

Latitudinal dependence: $P(\theta)$

The condition is similar, except it involves both l and m. After some rearranging, one arrives at

$$\sin\theta\frac{d}{d\theta}\left(\sin\theta\frac{d}{d\theta}P(\theta)\right)+[l(l+1)\sin^2\theta-m^2]P(\theta)=0. \tag{2.75}$$

This equation is the **associated Legendre Equation**. It turns out that the space of the homogeneous functions has a dimension $2l+1$, hence $0\leq m\leq l$.

If we substitute $\cos\theta=z$, Eq. 2.75 becomes

$$(1-z^2)\frac{d^2}{dz^2}P(z)-2z\frac{d}{dz}P+\left[l(l+1)-\frac{m^2}{1-z^2}\right]P(z)=0. \tag{2.76}$$

Eq. 2.76 is in standard form and can be solved using a variety of techniques. Most commonly, the solutions are found as polynomials $P_l^m(\cos\theta)$. The associated Legendre Equation reduces to the Legendre Equation in case $m=0$. In the latter case, the longitudinal dependence is lost as also Eq. 2.74 reverts to a constant. The resulting functions $P_l(\cos\theta)$ have a rotational symmetry around the z-axis. They are called **zonal** functions.

It is possible to find expressions of the (associated) Legendre polynomials that summarize their behavior as follows:

$$P_l(z) = \frac{1}{2^l l!}\frac{d^l}{dz^l}(z^2-1)^l \tag{2.77}$$

$$P_l^m(z) = \frac{(1-z^2)^{\frac{m}{2}}}{l!2^l}\frac{d^{l+m}}{dz^{l+m}}(z^2-1)^l, \tag{2.78}$$

written in terms of the $(l+m)^{\text{th}}$ derivative and $z=\cos\theta$. It is easy to make a small table with these polynomials (note that in Table 2.1, we have used some trig rules to simplify the expressions.) — this should get you started in using Eqs. 2.77 or 2.78.

l	$P_l(z)$	$P_l(\theta)$
0	1	1
1	z	$\cos\theta$
2	$\frac{1}{2}(3z^2-1)$	$\frac{1}{4}(3\cos 2\theta+1)$
3	$\frac{1}{2}(5z^3-3z)$	$\frac{1}{8}(5\cos 3\theta+3\cos\theta)$

Table 2.1: Legendre polynomials.

Some Legendre functions are plotted in Figure 2.13.

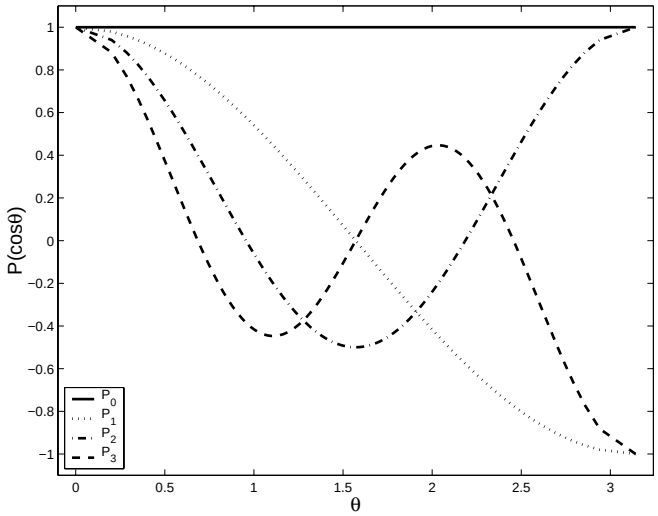

Figure 2.13: Legendre polynomials.

Spherical harmonics

The generic solution for U is thus found by combining the radial, longitudinal and latitudinal behaviors as follows:

$$U(r,\theta,\varphi) = \left\{ \begin{array}{c} r^l \\ \left(\frac{1}{r}\right)^{l+1} \end{array} \right\} [A_l^m \cos m\varphi + B_l^m \sin m\varphi] P_l^m(\cos\theta) \tag{2.79}$$

These are called the solid spherical harmonics of **degree** l and **order** m.

The spherical harmonics form a complete orthonormal basis. We implicitly assume that the full solution is given by a summation over all possible l and m indices, as in:

$$U(r,\theta,\varphi) = \sum_{l=0}^{\infty} \sum_{m=0}^{l} \left\{ \begin{array}{c} r^l \\ \left(\frac{1}{r}\right)^{l+1} \end{array} \right\} [A_l^m \cos m\varphi + B_l^m \sin m\varphi] P_l^m(\cos\theta) \tag{2.80}$$

The constants need to be determined from the boundary conditions. Because the spherical harmonics form a complete orthonormal basis, an arbitrary real function $f(\theta,\varphi)$ can be expanded in terms of spherical harmonics by

$$f(\theta,\varphi) = \sum_{l=0}^{\infty} \sum_{m=0}^{l} [A_l^m \cos m\varphi + B_l^m \sin m\varphi] P_l^m(\cos\theta). \tag{2.81}$$

The process of determining the coefficients A_l^m and B_l^m is analogous to that to determine the coefficients in a Fourier series, i.e. multiply both sides of Eq. 2.81 by $\cos m'\varphi P_{l'}^{m'}(\cos\theta)$ or $\sin m'\varphi P_{l'}^{m'}(\cos\theta)$, integrate, and use the orthogonality relationship — out comes A_l^m. For unequal data distributions, the coefficients may be found in a least-squares sense.

Visualization

It is important to visualize the behavior of spherical harmonics, as in Figure 2.14.

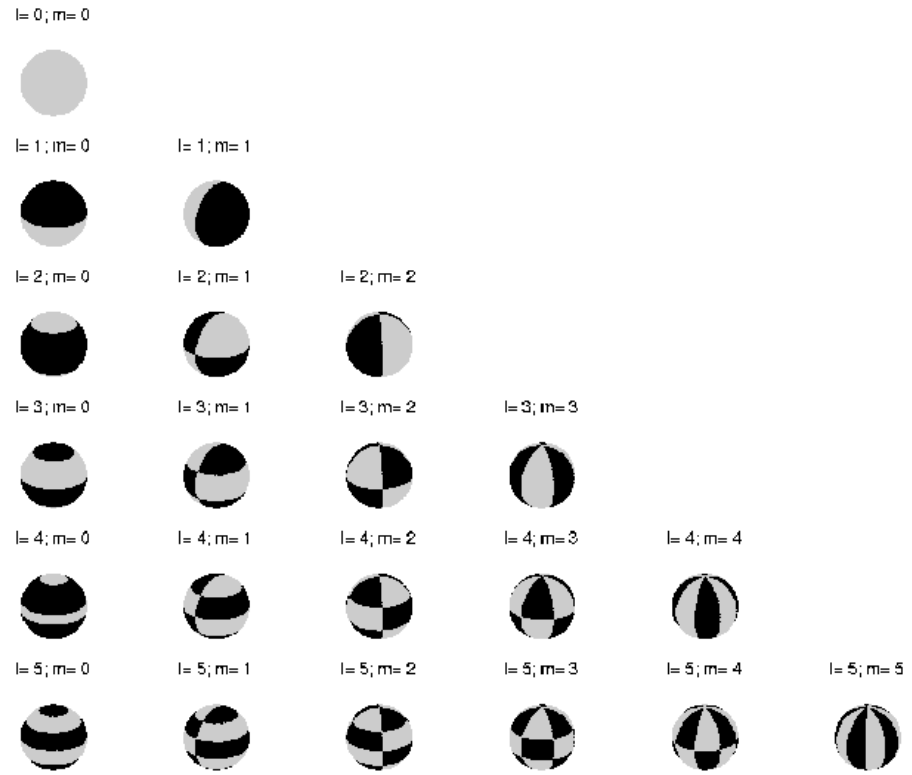

Figure 2.14: Some spherical harmonics.

Some terminology to remember is that on the basis of the values of l and m one identifies three types of harmonics.

- The **zonal** harmonics are defined to be those of the form $P_l^0(\cos\theta) = P_l(\cos\theta)$. The superposition of these Legendre polynomials describe variations with latitude; they do not depend on longitude. Zonal harmonics vanish at l small circles on the globe, dividing the spheres into latitudinal zones.

- The **sectorial** harmonics are of the form $\sin(m\varphi)P_m^m(\cos\theta)$ or $\cos(m\varphi)P_m^m(\cos\theta)$. As they vanish at $2m$ meridians (longitudinal lines, so m great circles), they divide the sphere into sectors.

- The **tesseral** harmonics are those of the form $\sin(m\varphi)P_l^m(\cos\theta)$ or $\cos(m\varphi)P_l^m(\cos\theta)$ for $l \neq m$. The amplitude of a surface spherical

harmonic of a certain degree l and order m vanishes at $2m$ meridians of longitude and on $(l-m)$ parallels of latitude.

Intermezzo 2.8 CARTESIAN VS SPHERICAL REPRESENTATION

If you work on a small scale with local gravity anomalies (for instance in exploration geophysics) it is not efficient to use (global) basis functions on a sphere because the number of coefficients that you'd need would simply be too large. For example to get resolution of length scales of 100 km (about 1°) you need to expand up to degree l=360 which with all the combinations $0 < m < l$ involves several hundreds of thousands of coefficients (how many exactly?). Instead you would use a Fourier Series. The concept is similar to spherical harmonics. A Fourier series is just a superposition of harmonic functions (sine and cosine functions) with different frequencies (or wave numbers $k = 2\pi/\lambda$, λ the wavelength):

$$g_z = \text{constant} \cdot \sin(k_x x) e^{-k_x z} = \text{constant} \cdot \sin\left(\frac{2\pi x}{\lambda_x}\right) e^{-\frac{2\pi z}{\lambda_x}} \tag{2.82}$$

(For a 2D field the expression includes y but is otherwise be very similar.) Or, in more general form

$$g_z = \sum_{n=0}^{\infty} \left[a_n \cos\left(\frac{n\pi x}{\lambda_x}\right) + b_n \sin\left(\frac{n\pi x}{\lambda_x}\right) \right] e^{-\frac{n\pi z}{\lambda_x}} \tag{2.83}$$

(compare to the expression of the spherical harmonics). In this expression the up- and downward continuation of the 1D or 2D harmonic field is controlled by an exponential form. The problem with downward continuation becomes immediately clear from the following example. Suppose in a marine gravity expedition to investigate density variation in the sediments beneath the sea floor, say, at 2 km depth, gravity measurements are taken at 10 m intervals on the sea surface (x_0 specifies the size of the grid at which the measurements are made). Upon downward continuation, the signal associated with the smallest wavelength allowed by such grid spacing would be amplified by a factor of $\exp(2000\pi/10) = exp(200\pi) \approx 10^{273}$. (The water does not contain any concentrations of mass that contribute to the gravity anomalies and integration over the surface enclosing the water mass would add only a constant value to the gravity potential but that is irrelevant when studying anomalies, and Laplace's equation can still be used.) So it is important to filter the data before the downward continuation so that information is maintained only on length scales that are not too much smaller than the distance over which the downward continuation has to take place.

In other words the degree l gives the total number of nodal lines and the order m controls how this number is distributed over nodal meridians and nodal parallels. The higher the degree and order the finer the detail that can be represented, but increasing l and m only makes sense if data coverage is sufficient to constrain the coefficients of the polynomials.

A different rendering is given in Figure 2.15.

An important property follows from the depth dependence of the solution:

From eqn. (2.80) we can see that (1) the amplitude of all terms will decrease with increasing distance from the origin (i.e., the internal source of the potential)

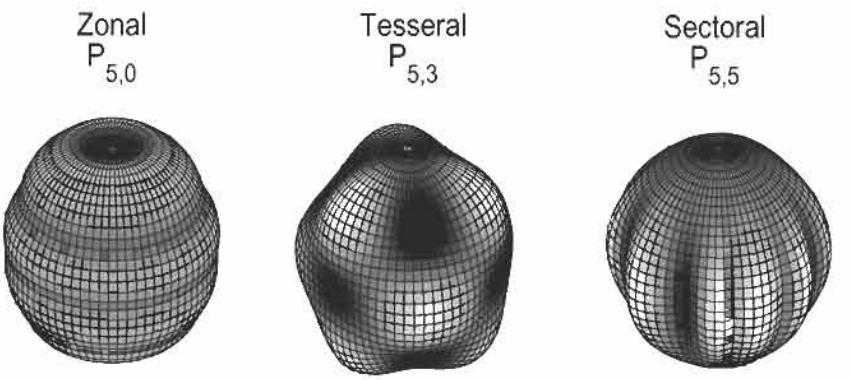

Figure 2.15: Zonal, tesseral and sectoral spherical harmonics.

and that (2) the rate of decay increases with increasing degree l. This has the following consequences:

1. with increasing distance the lower order harmonics become increasingly dominant since the signal from small-scale structure (large l's and m's) decays more rapidly. Recall that the perturbations of satellite orbits constrain the lower orders very accurately. The fine details at depth are difficult to discern at the surface of the Earth (or further out in space) owing to this spatial attenuation.

2. Conversely, this complicates the downward continuation of $U(r, \theta, \varphi)$ from the Earth's surface to a smaller radius, since this process introduces higher degree components in the solutions that are not constrained by data at the surface. This problem is important in the downward continuation of the magnetic field to the core-mantle-boundary.

2.6 Global gravity anomalies

Gravity in and outside a spherical mass sheet

The full solution to Laplaces equation was given in Equation 2.80. We've talked about the choice of radial functions. **Inside** the mass distribution, we use

$$U^{\text{in}}(r, \theta, \varphi) = \left\{ r^l \right\} [A_l^m \cos m\varphi + B_l^m \sin m\varphi] P_l^m(\cos\theta) \tag{2.84}$$

and outside, we use

$$U^{\text{out}}(r, \theta, \varphi) = \left\{ \frac{1}{r^{l+1}} \right\} [A_l^m \cos m\varphi + B_l^m \sin m\varphi] P_l^m(\cos\theta) \tag{2.85}$$

From now on, we'll add a normalization factor with a the radius at the equator:

$$U^{\text{out}}(r, \theta, \varphi) = -\frac{GM}{a} \sum_{l=0}^{\infty} \sum_{m=0}^{l} \left(\frac{a}{r} \right)^{l+1} [A_l^m \cos m\varphi + B_l^m \sin m\varphi] P_l^m(\cos\theta) \tag{2.86}$$

So in terms of a surface spherical harmonic potential $U(l)$ on the unit circle, we get the following equations for the field in- and outside the mass distribution:

$$\begin{aligned} U^{\text{in}}(r,l) &= \left(\frac{r}{a}\right)^{l} U(l) \\ U^{\text{out}}(r,l) &= \left(\frac{a}{r}\right)^{l+1} U(l) \end{aligned} \tag{2.87}$$

For gravity, this becomes:

$$\begin{aligned} \mathbf{g}^{\text{in}}(r,l) &= -\frac{l}{a^{l}} r^{l-1} U(l)\hat{\mathbf{r}} \\ \mathbf{g}^{\text{out}}(r,l) &= a^{l+1}(l+1)\frac{1}{r^{l+1}} U(l)\hat{\mathbf{r}} \end{aligned} \tag{2.88}$$

What is the gravity due to a thin sheet of mass of spherical harmonic degree l? Let's represent this as a sheet with vanishing thickness, and call $\sigma(l)$ the mass density per unit area. This way we can work at constant r and use the results for spherical symmetry. We know from Gauss's law that the flux through any surface enclosing a bit of mass is equal to the total enclosed mass (times $-4\pi G$). So constructing a box around a patch of surface $\mathbf{S}$ with area dS, enclosing a bit of mass dM, we can deduce that

$$g_{\text{out}} - g_{\text{in}} = 4\pi G\sigma(l) \tag{2.89}$$

On this shell — give it a radius a, we can use Eqs. 2.88 to find $g_{\text{out}} = U(l)(l+1)/a$ and $-g_{\text{in}} = -U(l)l/a$, and solve for $U(l)$ using Eq. 2.89 as $U(l) = 4\pi G\sigma(l)a/(2l+1)$. Plugging this into Eqs. 2.88 again we get for the gravity in- and outside this mass distribution

$$\begin{aligned} g^{\text{in}}(r,l) &= \frac{4\pi Gl}{2l+1}\sigma(l)\frac{r^{l-1}}{a^{l-1}} \\ g^{\text{out}}(r,l) &= \frac{4\pi Gl(l+1)}{2l+1}\sigma(l)\frac{a^{l+2}}{r^{l+2}} \end{aligned} \tag{2.90}$$

Length scales

Measurements of gravitational attraction are — as we have seen — useful in the determination of the shape and rotational properties of the Earth. This is important for geodesy. In addition, they also provide information about aspherical density variations in the lithosphere and mantle (important for understanding dynamical processes, interpretation of seismic images, or for finding mineral deposits). However, before gravity measurements can be used for interpretation several corrections will have to be made: the data reductions plays an important role in gravimetry since the signal pertinent to the structures we are interested in is very small.

Let's take a step back and get a feel for the different length scales and probable sources involved. If we use a spherical harmonic expansion of the field U we can see that it's the up- or downward continuation of the field and its dependence on r and degree l that controls the behavior of the solution at different depths (or radius) (remember Eq. 2.87).

With increasing r from the source the amplitude of the surface harmonics become smaller and smaller, and the decay in amplitude (spatial attenuation) is stronger for the higher degrees l (i.e., the small-scale structures).

Table 2.2 gives and idea about the relationship between length scales, the probable source regions, and where the measurements have to be taken.

wavelength λ	short wavelength ($\lambda < 1000$ km or $l > 36$)	long wavelength ($\lambda > 1000$ km or $l < 36$)
Source region	shallow: crust, lithosphere	probably deep (lower mantle) but shallower source cannot be excluded
Measurement: how, where?	close to source: surface, sea level, "low orbit" satellites, planes	Larger distance from origin of anomalies; perturbations of satellite orbits
Representation	values at grid points; 2D Fourier series	spherical harmonics
Coordinate system	cartesian	spherical

Table 2.2: Wavelength ranges of gravity anomalies

The free-air gravity anomaly

Let's assume that the geoid height N with respect to the spheroid is due to an anomalous mass dM. If dM represents excess mass, the equipotential is warped outwards and there will be a geoid high ($N > 0$); conversely, if dM represents a mass deficiency, $N < 0$ and there will be a geoid low.

We can represent the two potentials as follows: the actual geoid, $U(r, \theta, \varphi)$ is an equipotential surface *with the same potential* W_0 as the reference geoid U_0, only

$$U(r, \theta, \varphi) = U_0(r, \theta, \varphi) + \Delta U(r, \theta, \varphi) \tag{2.91}$$

We define the free-air gravity anomaly as the gravity $g(P)$ measured at point P minus the gravity at the projection Q of this point onto the reference geoid at r_0, $g_0(Q)$. Neglecting the small differences in direction, we can write for the magnitudes:

$$\Delta g = g(P) - g_0(Q) \tag{2.92}$$

In terms of potentials:

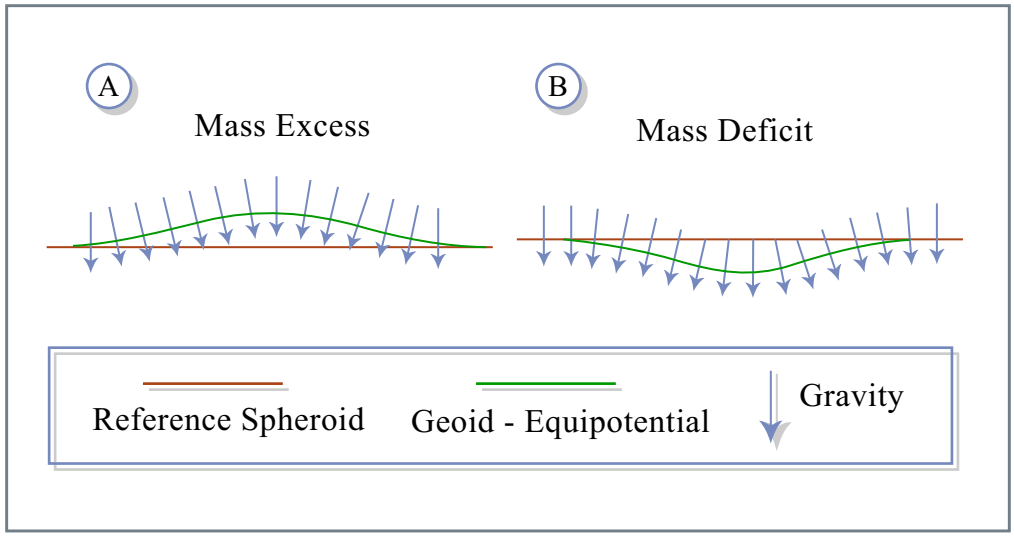

Figure by MIT OCW.

Figure 2.16: Mass deficit leads to geoid undulation.

$$\begin{aligned} U_0(P) &= U_0(Q) + \left(\frac{dU_0}{dr}\right)\Big|_{r_0} N \\ &= U_0(Q) - g_0 N \end{aligned} \tag{2.93}$$

(Remember that g_0 is the magnitude of the negative gradient of U and therefore appears with a positive sign.) We knew from Eq. 2.91 that

$$\begin{aligned} U(P) &= U_0(P) + \Delta U(P) \\ &= U_0(Q) - g_0 N + \Delta U(P) \end{aligned} \tag{2.94}$$

But also, since the potentials of U and U_0 were equal, $U(P) = U_0(Q)$ and we can write

$$g_0 N = -\Delta U(P) \tag{2.95}$$

This result is known as **Brun's formula**. Now for the gravity vectors $\mathbf{g}$ and $\mathbf{g}_0$, they are given by the familiar expressions

$$\begin{aligned} \mathbf{g} &= -\boldsymbol{\nabla} U \\ \mathbf{g_0} &= -\boldsymbol{\nabla} U_0 \end{aligned} \tag{2.96}$$

and the **gravity disturbance vector** $\delta\mathbf{g} = \mathbf{g} - \mathbf{g_0}$ can be defined as the difference between those two quantities:

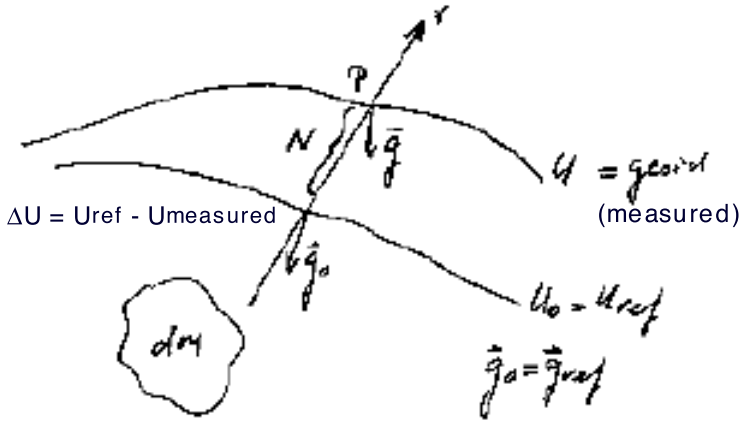

Figure 2.17: Derivation. Note that in this figure, the sign convention for the gravity is reversed; we have used and are using that the gravity is the negative gradient of the potential.

$$\delta \mathbf{g} = -\boldsymbol{\nabla} \Delta U$$
$$\delta g = g - g_0 = -\frac{\partial \Delta U}{\partial r} \tag{2.97}$$

On the other hand, from a first-order expansion, we learn that

$$g_0(P) = g_0(Q) + \left(\frac{dg_0}{dr}\right)\Big|_{r_0} N$$
$$g(P) = g_0(Q) + \left(\frac{dg_0}{dr}\right)\Big|_{r_0} N - \frac{d\Delta U}{dr} \tag{2.98}$$
$$\tag{2.99}$$

Now we define the **free-air gravity anomaly** as the difference of the gravitational accelartion measured on the actual geoid (if you're on a mountain you'll need to refer to sea level) minus the reference gravity:

$$\Delta g = g(P) - g_0(Q) \tag{2.100}$$

This translates into

$$\Delta g \quad = \quad \left(\frac{dg_0}{dr}\right)\Big|_{r_0} N - \frac{\partial \Delta U}{\partial r}$$

$$
\begin{aligned}
&= \frac{d}{dr}\left(\frac{GM}{r^2}\right)\Big|_{r_0} N - \frac{\partial \Delta U}{\partial r} \\
&= -\frac{2}{r_0}\left(\frac{GM}{r_0^2}\right) N - \frac{\partial \Delta U}{\partial r} \\
&= -\frac{2}{r_0} g_0 N - \frac{\partial \Delta U}{\partial r} \\
\Delta g &= \frac{2}{r_0}\Delta U - \frac{\partial \Delta U}{\partial r} \qquad (2.101)
\end{aligned}
$$

(2.102)

So at this arbitrary point P on the geoid, the gravity anomaly Δg due to the anomalous mass arises from two sources: the direct contribution dg_m due to the extra acceleration by the mass dM itself, and an additional contribution dg_h that arises from the fact that g is measured on height N above the reference spheroid. The latter term is essentially a **free air correction**, similar to the one one has to apply when referring the measurement (on a mountain, say) to the actual geoid (sea level).

Note that Eq. (2.95) contains the boundary conditions of $\nabla^2 U = 0$. The geoid height N at any point depends on the total effect of mass excesses and deficiencies over the Earth. N can be determined uniquely at any point (θ, φ) from measurements of gravity anomalies taken over the surface of the whole Earth — this was first done by Stokes (1849) — but it does not uniquely constrain the distribution of masses.

Gravity anomalies from geoidal heights

A convenient way to determine the geoid heights $N(\theta, \varphi)$ from either the potential field anomalies $\Delta U(\theta, \varphi)$ or the gravity anomalies $\Delta g(\theta, \varphi)$ is by means of spherical harmonic expansion of $N(\theta, \varphi)$ in terms of $\Delta U(\theta, \varphi)$ or $\Delta g(\theta, \varphi)$.

It's convenient to just give the coefficients of Eq. 2.86 since the basic expressions are the same. Let's see how that notation would work for eq. (2.86):

$$
\left\{ \begin{array}{c} U_A \\ U_B \end{array} \right\} = -\frac{GM}{a} \left\{ \begin{array}{c} A_l^m \\ B_l^m \end{array} \right\} \qquad (2.103)
$$

Note that the subscripts A and B are used to label the coefficients of the $\cos m\varphi$ and $\sin m\varphi$ parts, respectively. Note also that we have now taken the factor $-GMa^{-l}$ as the scaling factor of the coefficients.

We can also expand the potential U_0 on the reference spheroid:

$$
\left\{ \begin{array}{c} U_{0,A} \\ U_{0,B} \end{array} \right\} = -\frac{GM}{a} \left\{ \begin{array}{c} A_l'^m \\ 0 \end{array} \right\} \qquad (2.104)
$$

(Note that we did not drop the m , even though $m = 0$ for the zonal harmonics used for the reference spheroid. We just require the coefficient $A_l'^m$ to be zero for $m \neq 0$. By doing this we can keep the equations simple.)

The coefficients of the anomalous potential $\Delta U(\theta, \varphi)$ are then given by:

$$\left\{ \begin{array}{c} \Delta U_A \\ \Delta U_B \end{array} \right\} = -\frac{GM}{a} \left\{ \begin{array}{c} (A_l^m - A_l^{'m}) \\ B_l^m \end{array} \right\} \tag{2.105}$$

We can now expand $\Delta g(\theta, \varphi)$ in a similar series using eq. (2.101). For the ΔU, we can see by inspection that the radial derivative as prescribed has the following effect on the coefficients (note that the reference radius $r_0 = a$ from earlier definitions):

$$\frac{d\Delta U}{dr} \rightarrow -\frac{l+1}{r_0} \tag{2.106}$$

and the other term of Eq. 2.101 brings down

$$\frac{2}{r_0}\Delta U \rightarrow \frac{2}{a} \tag{2.107}$$

As a result, we get

$$\left\{ \begin{array}{c} \Delta g_A \\ \Delta g_B \end{array} \right\} = -\frac{GM}{a}\left(-\frac{l+1}{a}\right) \left\{ \begin{array}{c} (A_l^m - A_l^{'m}) \\ B_l^m \end{array} \right\} \tag{2.108}$$

$$= g_0(l-1) \left\{ \begin{array}{c} (A_l^m - A_l^{'m}) \\ B_l^m \end{array} \right\} \tag{2.109}$$

The proportionality with $(l-1)g_0$ means that the higher degree terms are magnified in the gravity field relative to those in the potential field. This leads to the important result that gravity maps typically contain much more detail than geoid maps because the spatial attenuation of the higher degree components is suppressed.

Using Eq. (2.95) we can express the coefficients of the expansion of $N(\theta, \varphi)$ in terms of either the coefficients of the expanded anomalous potential

$$g_0 \left\{ \begin{array}{c} N_A \\ N_B \end{array} \right\} = \frac{GM}{a} \left\{ \begin{array}{c} (A_l^m - A_l^{'m}) \\ B_l^m \end{array} \right\} \tag{2.110}$$

which, if we replace g by $\bar{g}$ and by assuming that $g \approx \bar{g}$ gets the following form

$$\left\{ \begin{array}{c} N_A \\ N_B \end{array} \right\} = a \left\{ \begin{array}{c} (A_l^m - A_l^{'m}) \\ B_l^m \end{array} \right\} \tag{2.111}$$

or in terms of the coefficients of the gravity anomalies (eqns. 2.109 and 2.111)

$$\left\{ \begin{array}{c} N_A \\ N_B \end{array} \right\} = \frac{a}{(l-1)g_0} \left\{ \begin{array}{c} \Delta g_A \\ \Delta g_B \end{array} \right\} = a \left\{ \begin{array}{c} (A_l^m - A_l^{'m}) \\ B_l^m \end{array} \right\} \tag{2.112}$$

The geoid heights can thus be synthesized from the expansions of either the gravity anomalies (2.112) or the anomalous potential (2.111). Geoid anomalies

have been constructed from both surface measurements of gravity (2.112) and from satellite observations (2.111). Equation (2.112) indicates that relative to the gravity anomalies, the coefficients of $N(\theta, \varphi)$ are suppressed by a factor of $1/(l-1)$. As a result, shorter wavelength features are much more prominent on gravity maps. In other words, geoid (and geoid height) maps essentially depict the low harmonics of the gravitational field. A final note that is relevant for the reduction of the gravity data. Gravity data are typically reduced to sea-level, which coincides with the geoid and not with the actual reference spheroid. Eq. 8 can then be used to make the additional correction to the reference spheroid, which effectively means that the long wavelength signal is removed. This results in very high resolution gravity maps.

2.7 Gravity anomalies and the reduction of gravity data

The combination of the reduced gravity field and the topography yields important information on the mechanical state of the crust and lithosphere. Both gravity and topography can be obtained by *remote sensing* and in many cases they form the basis of our knowledge of the dynamical state of planets, such as Mars, and natural satellites, such as Earth's Moon. Data reduction plays an important role in gravity studies since the signal caused by the aspherical variation in density that one wants to study are very small compared not only to the observed field but also other effects, such as the influence of the position at which the measurement is made. The following sum shows the various components to the observed gravity, with the name of the corresponding corrections that should be made shown in parenthesis:

Observed gravity = attraction of the reference spheroid, PLUS:

- *effects of elevation above sea level (***Free Air correction***), which should include the elevation (geoid anomaly) of the sea level above the reference spheroid*
- *effect of "normal" attracting mass between observation point and sea level (***Bouguer and terrain correction***)*
- *effect of masses that support topographic loads (***isostatic correction***)*
- *time-dependent changes in Earth's figure of shape (tidal correction)*
- *effect of changes in the rotation term due to motion of the observation point (e.g. when measurements are made from a moving ship. (Eötvös correction)*
- *effects of crust and mantle density anomalies ("geology" or "geodynamic processes").*

Only the **bold** corrections will be discussed here. The tidal correction is small, but must be accounted for when high precision data are required. The application of the different corrections is illustrated by a simple example of a small density anomaly located in a topography high that is isostatically compensated. See series of diagrams.

Free Air Anomaly

So far it has been assumed that measurements at sea level (i.e. the actual geoid) were available. This is often not the case. If, for instance, g is measured on the land surface at an altitude h one has to make the following correction :

$$dg_{\rm FA} = -2\frac{hg}{r} \tag{2.113}$$

For g at sea level this correction amounts to $dg_{\rm FA} = -0.3086h$ mgal or $0.3086h \times 10^{-5}$ ms^{-2} (h in meter). Note that this assumes no mass between the observer and sea level, hence the name "free-air" correction. The effect of ellipticity is often ignored, but one can use $r = R_{\rm eq}(1 - fsin^2\lambda)$. Note: per meter elevation this correction equals 3.1×10^{-6} ms^{-2} $\sim 3.1 \times 10^{-7}g$: this is on the limit of the precision that can be attained by field instruments, which shows that uncertainties in elevation are a limiting factor in the precision that can be achieved. (A realistic uncertainty is 1 mgal).

Make sure the correction is applied correctly, since there can be confusion about the sign of the correction, which depends on the definition of the potential. The objective of the correction is to compensate for the decrease in gravity attraction with increasing distance from the source (center of the Earth). Formally, given the minus sign in (2.114), the correction has to be subtracted, but it is not uncommon to take the correction as the positive number in which case it will have to be added. (Just bear in mind that you have to make the measured value larger by "adding" gravity so it compares directly to the reference value at the same height; alternatively, you can make the reference value smaller if you are above sea level; if you are in a submarine you will, of course, have to do the opposite).

The **Free Air anomaly** is then obtained by the correction for height above sea level and by subtraction of the reference gravity field

$$\Delta g_{\rm FA} = g_{\rm obs} - dg_{\rm FA} - g_0(\lambda) = (g_{\rm obs} + 0.3086h \times 10^{-3}) - g_0(\lambda) \tag{2.114}$$

(Note that there could be a component due to the fact that the sea level ($\equiv$ the geoid) does not coincide with the reference spheroid; an additional correction can then be made to take out the extra gravity anomaly. One can simply apply (2.114) and use $h' = h + N$ as the elevation, which is equivalent to adding a correction to $g_0(\lambda)$ so that it represents the reference value at the geoid. This correction is not important if the variation in geoid is small across the survey region because then the correction is the same for all data points.)

Bouguer anomaly

The free air correction does not correct for any attracting mass between observation point and sea level. However, on land, at a certain elevation there will be attracting mass (even though it is often compensated - isostasy (see below)). Instead of estimating the true shape of, say, a mountain on which the measurement is made, one often resorts to what is known as the "slab approximation" in which one simply assumes that the rocks are of infinite horizontal extent. The **Bouguer correction** is then given by

$$dg_B = 2\pi G\rho h \tag{2.115}$$

where G is the gravitational constant, ρ is the assumed mean density of crustal rock and h is the height above sea level. For $G = 6.67 \times 10^{-11}$ $\mathrm{m^3 kg^{-3} s^{-2}}$ and $\rho = 2,700\,\mathrm{kgm^{-3}}$ we obtain a correction of $1.1{\times}10^{-6}\,\mathrm{ms^{-2}}$ per meter of elevation (or 0.11 h mgal, h in meter). If the slab approximation is not satisfactory, for instance near the top of mountains, on has to apply an additional **terrain correction**. It is straightforward to apply the terrain correction if one has access to digital topography/bathymetry data.

The Bouguer anomaly has to be subtracted, since one wants to remove the effects of the extra attraction. The Bouguer correction is typically applied *after* the application of the Free Air correction. Ignoring the terrain correction, the Bouguer gravity anomaly is then given by

$$\Delta g_B = g_{\mathrm{obs}} - dg_{\mathrm{FA}} - g_0(\lambda) - dg_B = \Delta g_{\mathrm{FA}} - dg_{\mathrm{B}} \tag{2.116}$$

In principle, with the Bouguer anomaly we have accounted for the attraction of all rock between observation point and sea level, and Δg_B thus represents the gravitational attraction of the material below sea level. Bouguer Anomaly maps are typically used to study gravity on continents whereas the Free Air Anomaly is more commonly used in oceanic regions.

Isostasy and isostatic correction

If the mass between the observation point and sea level is all that contributes to the measured gravity one would expect that the Free Air anomaly is large, and positive over topography highs (since this mass is unaccounted for) and that the Bouguer anomaly decreases to zero. This relationship between the two gravity anomalies and topography is indeed what would be obtained in case the mass is completely supported by the strength of the plate (i.e. no isostatic compensation). In early gravity surveys, however, they found that the Bouguer gravity anomaly over mountain ranges was, somewhat surprisingly, large and negative. Apparently, a mass deficiency remained after the mass above sea level was compensated for. In other words, the Bouguer correction subtracted too much! This observation in the 19^{th} century lead Airy and Pratt to develop the concept of **isostasy**. In short, isostasy means that at depths larger than a certain compensation depth the observed variations in height above sea level

no longer contribute to lateral variations in pressure. In case of **Airy Isostasy** this is achieved by a compensation root, such that the depth to the interface between the loading mass (with constant density) and the rest of the mantle varies. This is, in fact *Archimedes' Law*, and a good example of this mechanism is the floating iceberg, of which we see only the top above the sea level. In the case of **Pratt Isostasy** the compensation depth does not vary and constant pressure is achieved by lateral variations in density. It is now known that both mechanisms play a role.

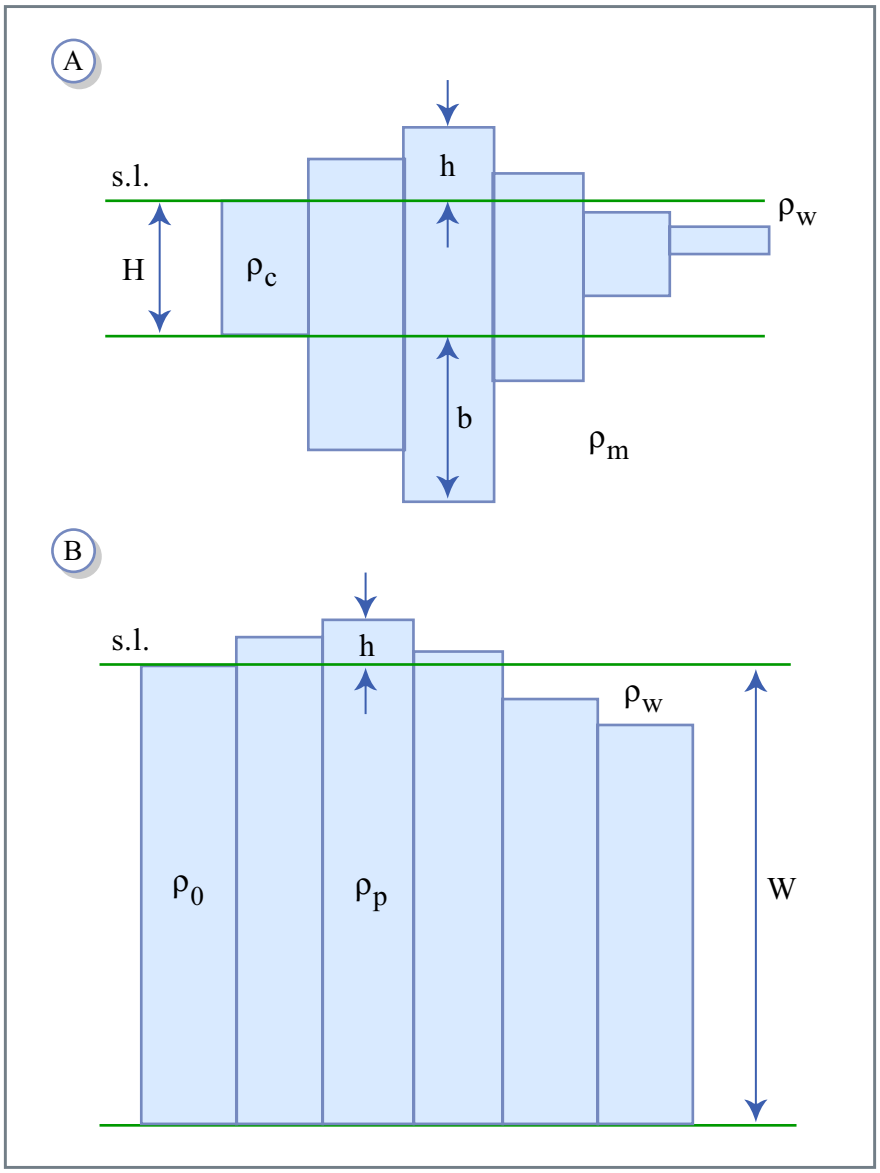

Figure by MIT OCW.

Figure 2.18: Airy (left) and Pratt (right) isostasy.

The basic equation that describes the relationship between the topographic height and the depth of the compensating body is (see Figure 2.19):

$$H = \frac{\rho_c h}{\rho_m - \rho_c} \tag{2.117}$$

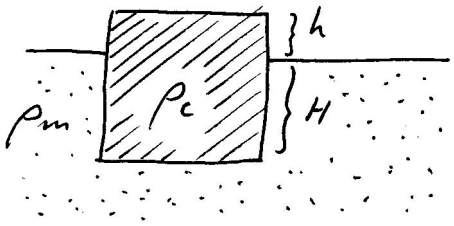

Figure 2.19: Airy isostasy.

Assuming Airy Isostasy and some constant density for crustal rock one can compute $H(x, y)$ from known (digital) topography $h(x, y)$ and thus correct for the mass deficiency. This results in the **Isostatic Anomaly**. If all is done correctly the isostatic anomaly isolates the small signal due to the density anomaly that is not compensated (local geology, or geodynamic processes).

2.8 Correlation between gravity anomalies and topography.

The correlation between Bouguer and Free Air anomalies on the one hand and topography on the other thus contains information as to what level the topography is isostatically compensated.

In the case of Airy Isostasy it is obvious that the compensating root causes the mass deficiency that results in a negative Bouguer anomaly. If the topography is compensated the mass excess above sea level is canceled by the mass deficiency below it, and as a consequence the Free Air Anomaly is small; usually, it is not zero since the attracting mass is closer to the observation point and is thus less attenuated than the compensating signal of the mass deficiency so that some correlation between the Free Air Anomaly and topography can remain.

Apart from this effect (which also plays a role near the edges of topographic features), the Free air anomaly is close to zero and the Bouguer anomaly large and negative when the topography is completely compensated isostatically (also referred to as "in isostatic equilibrium").

In case the topography is NOT compensated, the Free air anomaly is large and positive, and the Bouguer anomaly zero.

(This also depends on the length scale of the load and the strength of the supporting plate).

Whether or not a topographic load is or can be compensated depends largely on the strength (and the thickness) of the supporting plate and on the length scale of the loading structure. Intuitively it is obvious that small objects are not compensated because the lithospheric plate is strong enough to carry the load. This explains why impact craters can survive over very long periods of time! (Large craters may be isostatically compensated, but the narrow rims of the

crater will not disappear by flow!) In contrast, loading over large regions, i.e. much larger than the distance to the compensation depth, results in the development of a compensating root. It is also obvious why the strength (*viscosity*) of the plate enters the equation. If the viscosity is very small, isostatic equilibrium can occur even for very small bodies (consider, for example, the floating iceberg!). Will discuss the relationship between gravity anomalies and topography in more (theoretical) detail later. We will also see how viscosity adds a time dependence to the system. Also this is easy to understand intuitively; low strength means that isostatic equilibrium can occur almost instantly (iceberg!), but for higher viscosity the *relaxation time* is much longer. The flow rate of the material beneath the supporting plate determines how quickly this plate can assume isostatic equilibrium, and this flow rate is a function of viscosity. For large viscosity, loading or unloading results in a viscous delay; for instance the rebound after deglaciation.

2.9 Flexure and gravity.

The bending of the lithosphere combined with its large strength is, in fact, one of the compensation mechanisms for isostasy. When we discussed isostasy we have seen that the depth to the bottom of the root, which is less dense than surrounding rock at the same depth, can be calculated from **Archimedes' Principle**: if crustal material with density ρ_c replaces denser mantle material with density ρ_m a mountain range with height h has a compensating root with thickness H

$$H = \frac{h\rho_c}{\rho_m - \rho_c} \tag{2.118}$$

This type of compensation is also referred to as **Airy Isostasy**. It does not account for any strength of the plate. However, it is intuitively obvious that the depression H decreases if the strength (or the flexural rigidity) of the lithosphere increases. The consideration of lithospheric strength for calculates based on isostasy is important in particular for the loading on not too long a time scale.

An elegant and very useful way to quantify the effect of flexure is by considering the flexure due to a *periodic load*. Let's consider a periodic load due to topography h with maximum amplitude h_0 and wavelength λ: $h = h_0 \sin(2\pi x/\lambda)$. The corresponding load is then given by

$$V(x) = \rho_c g h_0 \sin\left(\frac{2\pi x}{\lambda}\right) \tag{2.119}$$

so that the flexure equation becomes

$$D\frac{d^4 w}{dx^4} + (\rho_m - \rho_c) g h = \rho_c g h_0 \sin\left(\frac{2\pi x}{\lambda}\right) \tag{2.120}$$

The solution can be shown to be

$$w(x) = \frac{h_0 \sin\left(\frac{2\pi x}{\lambda}\right)}{\left\{\frac{\rho_m}{\rho_c} - 1 + \frac{D}{\rho_c g}\left(\frac{2\pi}{\lambda}\right)^4\right\}} = w_0 \sin\left(\frac{2\pi x}{\lambda}\right) \tag{2.121}$$

From eq. (2.121) we can see that for very large flexural rigidity (or very large elastic thickness of the plate) the denominator will predominate the equation and the deflection will become small ($w \to 0$ for $D \to \infty$); in other words, the load has no effect on the depression. The same is true for short wavelengths, i.e. for $\lambda \ll 2\pi(D/\rho_c g)^{1/4}$. In contrast, for very long wave lengths ($\lambda \gg 2\pi(D/\rho_c g)^{1/4}$) or for a very weak (or thin) plate the maximum depression becomes

$$w_0 \approx \frac{\rho_c h_0}{\rho_m - \rho_c} \tag{2.122}$$

which is the same as for a completely compensated mass (see eq. 2.118). In other words, the plate "has no strength" for long wavelength loads.

The importance of this formulation is evident if you realize that any topography can be described by a (Fourier) series of periodic functions with different wavelengths. One can thus use Fourier Analysis to investigate the depression or compensation of any shape of load.

Eq. (2.121) can be used to find expressions for the influence of flexure on the Free Air and the Bouguer gravity anomaly. The gravity anomalies depend on the flexural rigidity in very much the same way as the deflection in (2.121).

Free-air gravity anomaly:

$$\Delta g_{\text{fa}} = 2\pi\rho_c G \left\{1 - \frac{e^{-2\pi b_m/\lambda}}{1 + \frac{D}{(\rho_m - \rho_c)g}\left(\frac{2\pi}{\lambda}\right)^4}\right\} h_0 \sin\left(\frac{2\pi x}{\lambda}\right) \tag{2.123}$$

Bouguer gravity anomaly:

$$\Delta g_{\text{B}} = \frac{-2\pi\rho_c G e^{-2\pi b_m/\lambda}}{1 + \frac{D}{(\rho_m - \rho_c)g}\left(\frac{2\pi}{\lambda}\right)^4} h_0 \sin\left(\frac{2\pi x}{\lambda}\right) \tag{2.124}$$

where b_m is the depth to the Moho (i.e. the depressed interface between ρ_c and ρ_m) and the exponential in the numerator accounts for the fact that this interface is at a certain depth (this factor controls, in fact, the downward continuation).

The important thing to remember is the linear relationship with the topography h and the proportionality with D^{-1}. One can follow a similar reasoning as above to show that for short wavelengths the free air anomaly is large (and positive) and that the Bouguer anomaly is almost zero. This can be explained by the fact that the flexure is then negligible so that the Bouguer correction successfully removes all anomalous structure. However, for long wavelength loads, the load is completely compensated so that after correction to zero elevation,

the Bouguer correction still 'feels' the anomalously low density root (which is not corrected for). The Bouguer anomaly is large and negative for a completely compensated load. Complete isostasy means also that there is no net mass difference so that the free air gravity anomaly is very small (practically zero). Gravity measurements thus contain information about the degree of isostatic compensation.

The correlation between the topography and the measured Bouguer anomalies can be modeled by means of eq. (2.124) and this gives information about the flexural rigidity, and thus the (*effective*!) thickness of the elastic plate. The diagram below gives the Bouguer anomaly as a function of wave length (i.e. topography was subjected to a Fourier transformation). It shows that topography with wavelengths less than about 100 km is not compensated (Bouguer anomaly is zero). The solid curves are the predictions according to eq. 2.124 for different values of the flexural parameter α. The parameters used for these theoretical curves are $\rho_m = 3400$ kgm^{-3}, $\rho_c = 2700$ kgm^{-3}, $b_m = 30$ km, $\alpha = [4D/(\rho_m - \rho_c)g]^{1/4} = 5$, 10, 20, and 50 km. There is considerable scatter but a value for α of about 20 seems to fit the observations quite well, which, with E= 60 GPa and σ= 0.25, gives an effective elastic thickness $h \sim 6$ km.

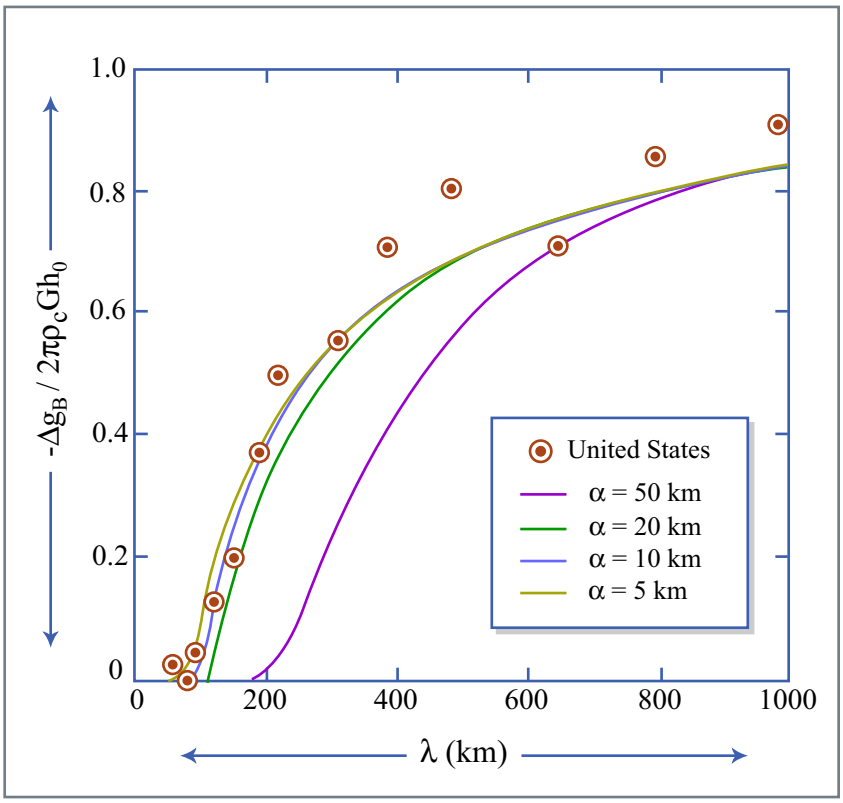

Figure by MIT OCW.

Figure 2.20: Bouguer anomalies and topography.

Post-glacial rebound and viscosity

So far we have looked at the bending or flexure of the elastic lithosphere to loading, for instance by sea mounts. To determine the deflection $w(x)$ we used the principle of isostasy. In order for isostasy to work the mantle beneath the lithosphere must be able to flow. Conversely, if we know the history of loading, or unloading, so if we know the deflection as a function of time $w(x,t)$, we can

investigate the flow beneath the lithosphere. The rate of flow is dependent on the **viscosity** of the mantle material. Viscosity plays a central role in understanding mantle dynamics. Dynamic viscosity can be defined as the ratio of the applied (deviatoric) stress and the resultant strain rate; here we mostly consider **Newtonian viscosity**, i.e., a linear relationship between stress and strain rate. The unit of viscosity is Pascal Second [Pa s].

A classical example of a situation where the history of (un)loading is sufficiently well known is that of **post-glacial rebound**. The concept is simple:

1. the lithosphere is depressed upon loading of an ice sheet (viscous mantle flow away from depression make this possible)
2. the ice sheet melts at the end of glaciation and the lithosphere starts rebound slowly to its original state (mantle flow towards the decreasing depression makes this possible). The uplift is well documented from elevated (and dated) shore lines. From the rate of return flow one can estimate the value for the viscosity.

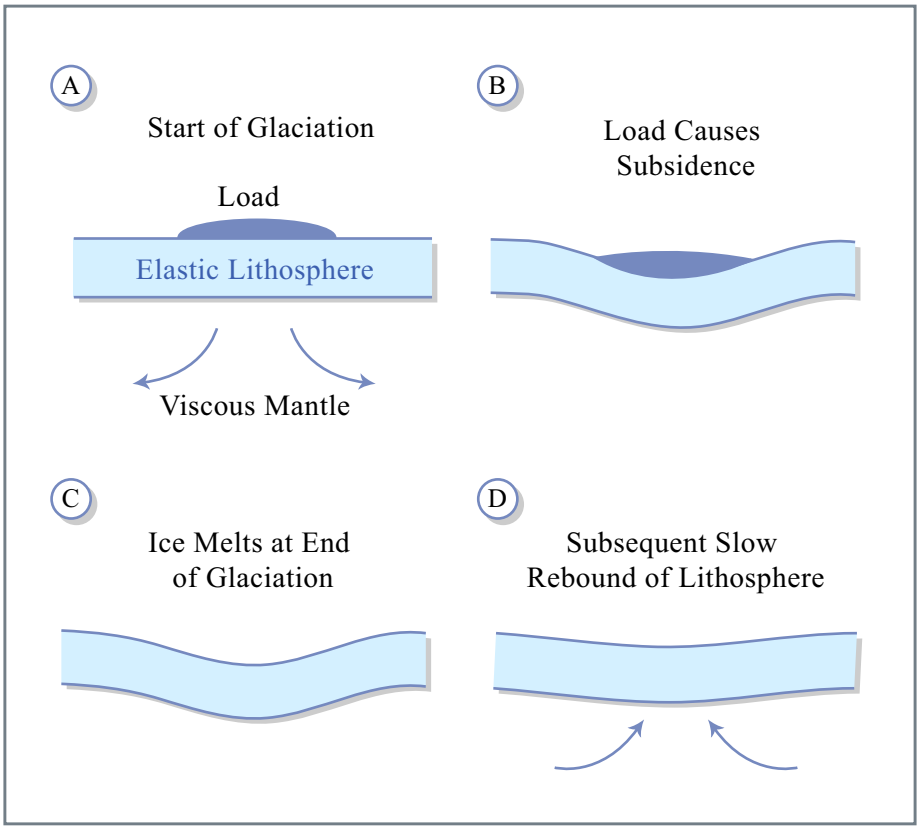

Figure by MIT OCW.

Two remarks:

1. the dimension of the load determines to some extend the depth over which the mantle is involved in the return flow → the comparison of rebound history for different initial load dimensions gives some constraints on the variation of viscosity with depth.
2. On long time scales the lithosphere has no "strength", but in sophisticated modeling of the post glacial rebound the flexural rigidity is still taken into

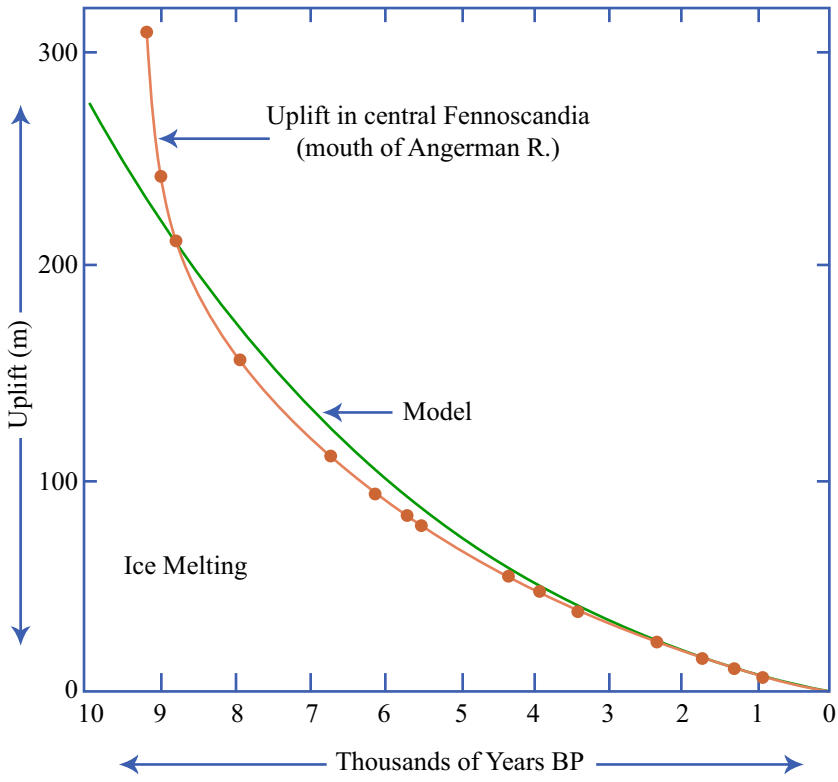

Figure 2.23. Uplift in central Fennoscandia calculated for a constant viscosity (10^{21} Pas) mantle (green line) and geological observations (circle) from the nothern Gulf of Bothnia.

Figure by MIT OCW.

account. Also taken into account in recent models is the history of the melting and the retreat of the ice cap itself (including the changes in shore line with time!). In older models one only investigated the response to instantaneous removal of the load.

Typical values for the dynamic viscosity in the Earth's mantle are 10^{19} Pa s for the upper mantle to 10^{21} Pa s for the lower mantle. The lithosphere is even "stiffer", with a typical viscosity of about 10^{24} (for comparison: water at room temperature has a viscosity of about 10^{-3} Pa s; this seems small but if you've ever dived of a 10 m board you know it's not negligible!)

Things to remember about these values:

1. very large viscosity in the entire mantle
2. lower mantle (probably) more viscous than upper mantle,
3. the difference is not very large compared to the large value of the viscosity itself.

An important property of viscosity is that it is *temperature dependent*; the viscosity decreases exponentially with increasing temperature as $\eta = \eta_0 e^{-30T/T_m}$, where T_m is the melting temperature and -30 is an empirical, material dependent value.

This temperature dependence of viscosity explains why one gets convection beneath the cooling lithosphere. As we have discussed before, with typical values for the geothermal gradient (e.g., 20 Kkm^{-1}) as deduced from surface heat flow using Fourier's Law the temperature would quickly rise to near the solidus, the temperature where the rock starts to melt. However, we know from several observations, for instance from the propagation of *S*- waves, that the temperature is below the solidus in most parts of the mantle (with the possible exception in the low velocity zone beneath oceanic and parts of the continental lithosphere). So there must be a mechanism that keeps the temperature down, or, in other words, that cools the mantle much more efficiently than conduction. That mechanism is *convection*. We saw above that the viscosity of the lithosphere is very high, and upper mantle viscosity is about 5 orders of magnitude lower. This is largely due to the temperature dependence of the viscosity (as mentioned above): when the temperature gets closer to the solidus (T_m) the viscosity drops and the material starts to flow.

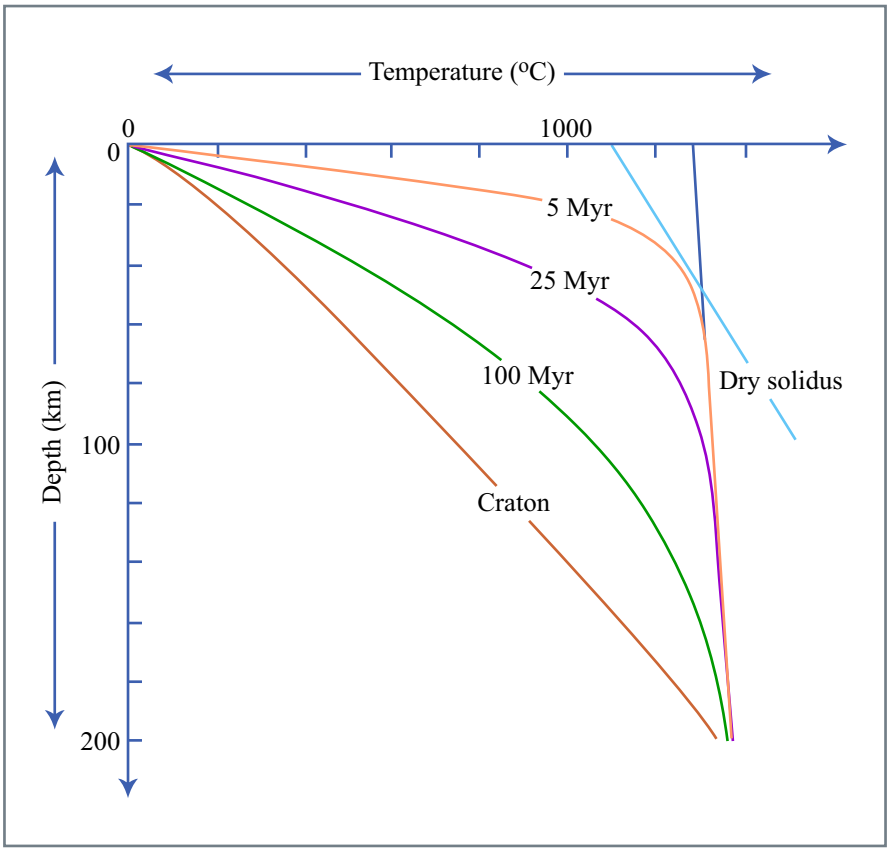

Figure by MIT OCW.