Geodesy and Gravity

Class Notes

John Wahr

Department of Physics
University of Colorado
Boulder, CO 80309

Samizdat
Press

Published by the Samizdat Press

Center for Wave Phenomena

Department of Geophysics

Colorado School of Mines

Golden, Colorado 80401

and

New England Research

76 Olcott Drive

White River Junction, Vermont 05001

Release 2.0, January 1999

Contents

Chapter 1

Introduction

1.1 Introduction

There are three required geophysics courses in the University of Colorado geophysics program:

1. Seismology: seismic waves, earthquakes, earth structure.
2. Geodesy and gravity. (This course.)
3. Heat flow, mantle convection, fluid dynamics, the earth's magnetic field.

Plate tectonics is the unifying theory for most of modern-day geophysics and, to a large extent, geology. According to this theory, the earth's surface is composed of about twenty disjoint plates which move with respect to each other. This motion is responsible, directly or indirectly, for most surface features (e.g. ocean basins, mountains, etc.) and for earthquakes and volcanos.

The driving force for plate tectonics is mantle convection of some sort: the plates are thermal boundary layers of convective cells in the mantle (over long time periods, the mantle behaves as a viscous fluid). But the details are not yet well understood. People have deduced the rate of motion averaged over millions of years by looking at magnetic anomalies in material on the sea floor. The motion of the plates on a year-to-year time scale is now beginning to be observed. One of the goals of modern geodesy is to detect the year-to-year motion. Is it the same as the long-term mean? Do the plates move rigidly? Etc.

1.2 Geodesy

Geodesy has a reasonable claim to being the oldest branch of geophysics. Originally it was solely concerned with global surveying. Its primary goal was, and probably still is, to tie local survey nets together by doing careful surveying over long distances. Geodesists tell local surveyors where their lines are with respect to the rest of the world. That includes telling them their elevation above sea level. This is still the major function of most geodesists, most of whom are not geophysicists.

To measure long baselines and to determine global positions you need:

1. more accurate observing instruments than in surveying — although frequently surveyors and geodesists use the same instruments.

2. complicated mathematical techniques to take into account things like the earth's curvature and, especially, the gravity field.

3. measurements of the gravity field.

The effects of the gravity field are especially difficult to deal with. Why should gravity enter in? It's because many geodetic instruments use gravity as a reference. For example, when geodesists or surveyors say a surface is horizontal, what they really mean is that it is a surface of constant gravitational potential (think of the way a carpenter's level works, for example). So, geodesists have always had to measure gravity — in addition to relative positions, which is why gravity has historically come under the heading of geodesy.

Out of these gravity observations came the first modern geophysical interpretations of any sort. This was the development of the idea of *isostasy* around 1840. We'll get into this, later.

Nowadays, what can gravity and point positioning observations do for geophysicists? Briefly, static gravity observations (that is, observations of the time-independent field) give information on strength within the earth (the earth's surface bends under loads, with resulting effects on gravity), on composition near the earth's surface (mineral prospectors use gravity), and on long term dynamical processes within the earth (eg. density contrasts associated with mantle convection/plate tectonics, and postglacial rebound).

Static (time-independent values) of positions have not given much useful information.

Recently, new geodetic techniques have begun to give useful observations of time-dependent gravity variations and positions of surface points. This turns out to be really exciting to geophysicists, because it allows people to be able to see the plates move and deform in real time.

1.3 Course Organization

Chapter 2 Observational techniques

Chapter 3 Potential theory (mathematical theory of gravity)

Chapter 4 Physical geodesy problems (for example, how do you determine the earth's shape)

Chapter 5 Stress/strain laws. Viscosity (for use in later interpretation)

Chapter 6 Interpretation of observed gravity anomalies

Chapter 7 Postglacial rebound

Chapter 8 Earth tides

Chapter 9 Earth rotation

Chapter 2

Observational Techniques

2.1 Instruments

2.1.1 Gravity meters

These measure g, the acceleration due to gravity. At the earth's surface,

$$g \approx 980 \text{ cm/sec}^2 \equiv 980 \text{ gal}.$$

(So, $g \approx 10^3$ gal.)

Surveying and static gravity observations require accuracies of approximately 1 mgal (10^{-6} accuracy). To measure changes in gravity you'd like 1 μgal ($\sim 10^{-9}$) accuracies. For example, if you move radially outward by 3 mm, gravity decreases (because you are moving further from the center of the earth) by about 1 μgal. So, 1 μgal accuracy corresponds to about 3 mm of vertical displacement. Most secular vertical crustal motions, when they occur, are on the order of up to a few mm per year. So a few years of gravity measurements at this accuracy level would allow those motions to be estimated.

There are two types of gravimeters:

Absolute meters Where g can be directly determined by measuring a length and/or a time.

Relative meters Where g depends on things like spring constants, which cannot be so readily determined. Relative instruments can only tell you the relative difference in g between two points or between two times.

2.1.2 Pendulums

These are the oldest gravimeters. They can be either absolute or relative instruments.

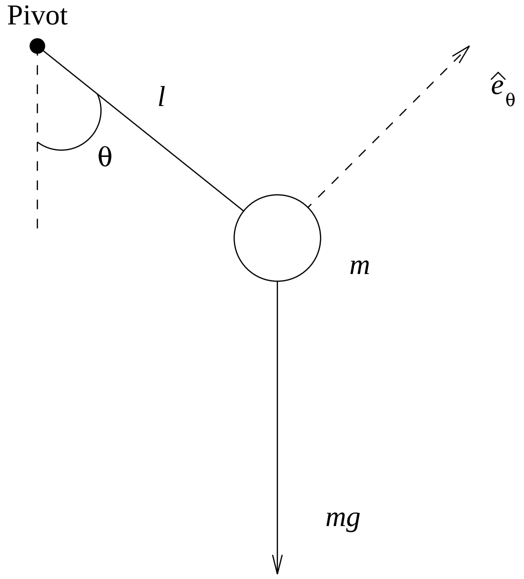

Figure 2.1:

First, consider a point mass, m, on a massless string, confined to move in a plane (i.e. a simple pendulum). See Figure 2.1. The acceleration in the $\hat{e}_\theta$ direction is $l\ddot{\theta}$. The gravitational force in the $\hat{e}_\theta$ direction is $-mg\sin\theta$. So

$$l\ddot{\theta} = -g\sin\theta.$$

For small motion $\sin\theta \approx \theta$, and so $\ddot{\theta} \cong -(g/l)\theta$.

The solution is $\theta \cong A\cos(\omega t + \phi)$ where

$$\omega = \sqrt{\frac{g}{l}} \tag{2.1.1}$$

is the angular frequency of the motion. Or:

$$g = l\omega^2 = l\,\frac{(2\pi)^2}{T^2}$$

where T = period. Note that by measuring l and T, you get g. So this is an *absolute* instrument.

Probably the biggest problem, here, is that there is no *massless* string. The result of Equation 2.1.1 is, thus, only approximate at best. (The approximation $\sin\theta \approx \theta$ can be removed by using elliptic integrals.)

Instead, a real pendulum is a solid object — preferably a rod of some sort — swung about a fixed point in the rod (a physical pendulum), and constrained to move in a plane. See Figure 2.2. Let l = distance between pivot and the center of mass, CM. If N =

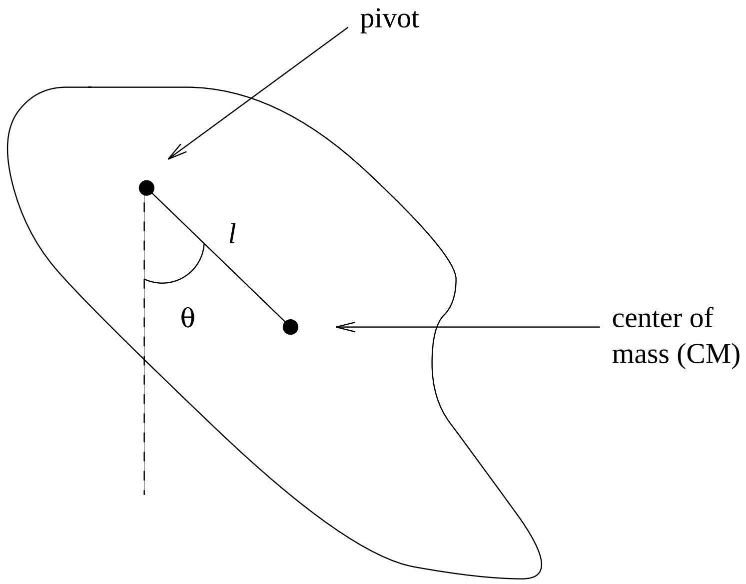

Figure 2.2:

torque about the pivot due to gravity, and I = moment of inertia about the axis passing through the pivot (and perpendicular to the plane of motion), then

$$N = I\ddot{\theta}$$

$$N = -(mg)(l)\sin\theta$$

where mg is the force, and $l\sin\theta$ is the moment arm. So,

$$\ddot{\theta} = -\left(\frac{mgl}{I}\right)\sin\theta.$$

For small motion (again, this assumption can be removed) $\sin\theta \approx \theta$ and

$$\theta = A\cos(\omega t + \phi)$$

where

$$\omega^2 = \left(\frac{mgl}{I}\right)$$

or

$$g = \frac{I\omega^2}{ml} = \frac{I}{ml}\frac{(2\pi)^2}{T^2}. \tag{2.1.2}$$

Thus, g depends on I — which is not directly measurable.

So in this mode a physical pendulum (that is, any *real* pendulum) is not an absolute instrument. It is, though, a relative instrument. I, m and l will presumably not change if you move the instrument to a new place, so that any change in g causes a corresponding change in T. So, if g_1 and g_2 are values of g at two points, and T_1 and T_2 are the observed values of T at those points, then:

$$\frac{g_1}{g_2} = \left(\frac{T_2}{T_1}\right)^2.$$

Thus you can determine g_1/g_2 by simply measuring times. In this way, relative gravity can be determined with accuracies of about 0.5 mgal — and pendulums were used until WWII in this way.

It is also possible to use a trick to turn a physical pendulum into an absolute meter. You can, in general, swing the pendulum about two different pivot points on opposite sides of the center of mass and get the same frequency. To see this, note that by the parallel-axis theorem, $I = I_{\text{CM}} + ml^2$ where I_{CM} = moment about the center of mass. So, $\omega^2 = mgl/\left(I_{\text{CM}} + ml^2\right)$, and thus the two pivot points must have the same $l/(I_{\text{CM}} + ml^2)$. If l_1 and l_2 denote the l's for the two points, then:

$$\frac{l_1}{I_{\text{CM}} + ml_1^2} = \frac{l_2}{I_{\text{CM}} + ml_2^2}$$

or

$$l_2^2(ml_1) - l_2\left[I_{\text{CM}} + ml_1^2\right] + [I_{\text{CM}}l_1] = 0.$$

Solving for l_2 using the quadratic formula, gives:

$$l_2 = \begin{cases} l_1 \\ \dfrac{I_{\mathrm{CM}}}{ml_1} \end{cases}$$

The non-trivial result is $l_2 = (I_{\mathrm{CM}}/ml_1)$. Solving for I_{CM} in terms of l_2, and using the result in Equation 2.1.2 gives:

$$g = \left(\frac{I_{\mathrm{CM}} + ml_1^2}{ml_1}\right)\frac{(2\pi)^2}{T^2} = \left(\frac{l_2 ml_1 + ml_1^2}{ml_1}\right)\frac{(2\pi)^2}{T^2}$$

or

$$g = \frac{L(2\pi)^2}{T^2} \tag{2.1.3}$$

where $L = l_1 + l_2 =$ distance between the two pivots (since the two points are on opposite sides of the center of mass).

Run in this manner, the pendulum is an absolute meter. First you choose a pivot and measure T. Then you find another pivot point that gives the same T, making sure that this second point is on the opposite side of the center of mass from the first point and is along the line that includes the first point and the center of mass, You measure the distance between pivots, L, and use Equation 2.1.3. Note that you find g by measuring only a distance and a time. Typically accuracies in this absolute mode are about 0.5 mgal, the same as for a pendulum used as a relative instrument.

Obviously, an absolute pendulum can be a lot of work to run — since you must search for the second pivot point. Still, until very recently pendulums were used as absolute instruments, long after they were abandoned as relative meters. People would maintain absolute pendulums at a few sites around the world (or else periodically revisit those sites with an absolute pendulum), and then people using relative meters out in the field would occasionally visit those sites to recalibrate their relative instruments.

The pendulum results can be corrected for air density (that is, for friction), for temperature, and for the fact that $\sin\theta \neq \theta$.

The principal problems for these instruments are:

1. the fact that the pivot is not a knife edge — the pivot point moves around a little in an unknown way;

2. the base of the pendulum moves around as the pendulum oscillates.

2.1.3 Springs

These replaced pendulums as relative instruments after WWII. They cannot be modified to give absolute g.

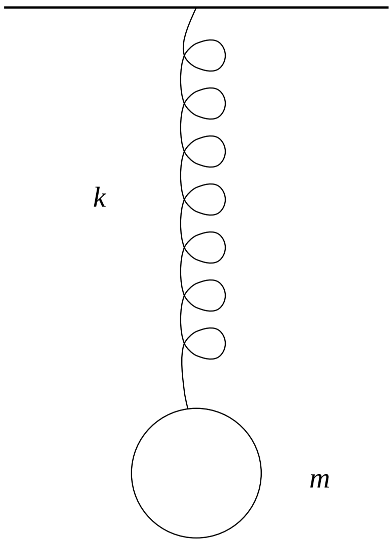

Figure 2.3: Simple Spring.

Consider a mass on a massless spring; see Figure 2.3. The period of oscillation of the spring is independent of g, so it can't be used for anything $\left(T = 2\pi\sqrt{m/k}\right)$. But at equilibrium the amount the spring has stretched is $l = mg/k$. So, $g = lk/m$. You cannot measure k/m, so a spring is not an absolute meter. But if you change g you change l. Note:

$$\frac{g_1}{g_2} = \frac{l_1}{l_2}.$$

So if you change g_2 to $g_1 = g_2 + \Delta g$, then you change l_2 to $l_2 + \Delta l$, where

$$\frac{\Delta l}{l_2} = \frac{\Delta g}{g_2}.$$

This lets you measure *changes* in g, by measuring changes in l.

But this simple mass on a spring is an impractical gravity meter. The departure from equilibrium, l_2, might be 1 m at most. So to measure changes in g at the 0.5 mgal level (the accuracy of a pendulum), you would have to measure changes in length to better than

$$\Delta l = l_2 \frac{\Delta g}{g_2} \approx 10^2 \frac{5 \times 10^{-4}}{10^3} \text{ cm} = 5 \times 10^{-5} \text{ cm}$$

which is awfully small. So a simple mass on a spring is not too useful.

The best way to make a spring work is to modify it somehow so that it has a very long natural period. Note that for our mass on a spring

$$g = \frac{lk}{m} = \frac{l(2\pi)^2}{T^2}.$$

So, another way to relate the change in gravity to the change in the distance the spring stretches would be:

$$\Delta g = \Delta l \frac{(2\pi)^2}{T^2}.$$

If you can increase the natural period, T, then a small change in g might produce a reasonably large, and so measurable, change in l. Increasing T would be equivalent to making a spring which had a larger l. But there are better ways to make a spring-type instrument with a much longer period than to simply lengthen the spring.

2.1.4 LaCoste-Romberg Meter

The relative meter now used almost exclusively is the LaCoste-Romberg meter. This meter has a design which gives almost an infinite period.

Here's the setup. (See Figure 2.4.) The spring in Figure 2.4 is called a "zero length spring." It is designed so that $F = kL$, where $L =$ *total* length of the spring. There are several ways to make such a spring. The wire can, for example, be twisted as it is wound. The beam of length b — with the mass on it — is free to pivot about the lower left hand point. That beam is assumed here to be massless.

This system has an infinite period: if the beam is at equilibrium for *one* value of θ, then it is at equilibrium for *every* value of θ. What happens qualitatively is that if you decrease θ (move m upward), the counter-clockwise spring torque lessens because L decreases; but the clockwise torque from gravity also lessens, because the angle between

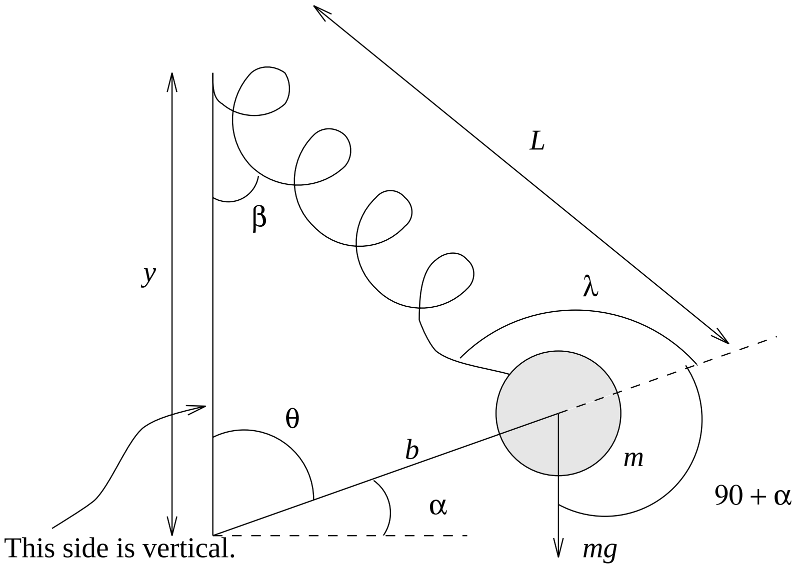

Figure 2.4: Set up for the LaCoste-Romberg Meter.

$m\hat{g}$ and the moment arm $\hat{b}$ decreases. These two effects cancel, and there is still no net torque.

To see this more quantitatively, the counterclockwise torque on m from the spring is $(kL)(b)\sin\lambda$. The angle λ is: $\lambda = \beta + \theta$. The clockwise torque from gravity is $(mg)b\sin(90+\alpha)$. Note that the angle $90+\alpha = 90+(90-\theta) = 180-\theta$.

Equating these two torques gives the equilibrium condition:

$$kLb\sin(\beta+\theta) = mgb\sin\theta. \tag{2.1.4}$$

To remove L, note that from the law of sines:

$$\frac{L}{\sin\theta} = \frac{y}{\sin(180-\theta-\beta)} = \frac{y}{\sin(\theta+\beta)}.$$

So $L\sin(\theta+\beta) = y\sin\theta$, and Equation 2.1.4 becomes

$$ky = mg.$$

So the equilibrium condition is independent of θ: any θ will do, so long as

$$y = \frac{mg}{k}. \tag{2.1.5}$$

If $y = mg/k$ does not hold, then the beam b swings all the way clockwise or counterclockwise until it hits the stops.

So the idea is to set the thing up, and then adjust y until the meter is at equilibrium. Then you know that

$$g = \left(\frac{k}{m}\right) y.$$

If you measure a change in y, you can then infer a change in g. Note that this is a relative meter because of (k/m) in the expression for g. Actually, if the meter were built as described above, it would be *too* sensitive. You could never find a y which agreed exactly with Equation 2.1.5, so you could never find the equilibrium point. Instead, what's done is to tip the side y, slightly, from vertical. This gives a large but finite, natural period, and it results in a usable but still highly sensitive meter.

The accuracies of this relative meter are on the order of 5 μgal (0.005 mgal), which is much better than pendulums. Fancy LaCoste-Romberg tide meters can do even better if kept running continuously at the same place.

There are two sorts of errors. One is in measuring y. y is changed by turning a screw, but the relation between the number of turns and the change in y is not always precisely known. These errors can be as large as 50 μgal for large changes in y. The second is spring hysteresis. As the spring is stretched, or even as it ages while doing nothing, the spring constant k can change, and $F = kL$ becomes $F = kL + d$, where d is some small constant. This causes drifts in measured gravity that can be as large as several hundred μgal per month. This means that, in practice, the meter must be frequently brought back to a base station and recalibrated.

2.1.5 Superconducting Gravimeter

This is a relative instrument developed by people at the University of California at San Diego. They are now produced commercially.

The idea is to levitate a superconducting ball in a magnetic field. If gravity changes, an electrostatic force is applied to the ball to keep it level. The instrument output is voltage. (So it's a relative meter.) Essentially it's an electro-magnetic spring.

The instrument is designed to stay fixed at one spot and measure changes in gravity with time. It is definitely not portable. Furthermore, interruptions (such as refilling liquid helium) can reset the calibration factor (the relation between voltage and g). One method of determining the calibration is to roll a heavy ball of known mass underneath the meter, and to measure the resulting signal. Other calibration methods are also used, including one in which the meter is subject to a known acceleration.

This instrument is very accurate, with sub μgal accuracies (as good or slightly better than the best LaCoste-Romberg tide meters). Making the instrument superconducting keeps currents in the ball stable. In fact, the meter is exceptionally stable, which is its main advantage over good LaCoste-Romberg meters. The big disadvantage is that it is not portable.

2.2 Relative Gravity at Sea and on Planes

The instruments described above are primarily land instruments. Trying to measure g from ships or planes leads to all sorts of additional problems. But, you certainly need to know g over the oceans. And sometimes measurements from planes or helicopters are desirable to get to inaccessible regions; or to speed up a survey.

The extra problems are associated with the fact that ships and planes are moving objects. The errors from the motion can be severe, and can easily swamp instrumental noise. Thus, it can be more important to use meters that minimize the motion errors, than to use meters designed simply to measure g extremely precisely. For example, pendulums were used in ships until recently, since the additional accuracy of a LaCoste-Romberg is not always needed.

One problem is knowing where you are making your measurements from. Nowadays, you can figure that out very well using the space positioning measurements that we'll discuss later. In the past, though, people had to use astronomy (e.g. observing the stars and the sun) to determine their latitude and longitude, and then had to somehow find the height of the meter above mean sea level.

A second, more serious problem is knowing what your velocity is. The problem is that since the earth is rotating, any moving object experiences Coriolis forces which can be confused as gravitational forces.

To understand the biggest effect, suppose you're in a ship moving east or west — the same direction as the motion due to rotation. Let's think of the resulting force as an increased (or decreased) centrifugal force, rather than as a Coriolis force. Suppose you attach a coordinate system to the earth and ship so that it rotates about the North Pole with the angular velocity of the ship: $\Omega = \Omega_0 + v/(a\sin\theta)$ where Ω_0 = rotation rate of the earth, v = eastwards velocity of the ship, a = earth's radius, and θ = colatitude of the ship. There is no Coriolis force, because the ship doesn't move with respect to the coordinate system. The centrifugal acceleration on the ship and every object in it is

$$\overline{\Omega} \times \left(\overline{\Omega} \times \overline{r}\right) = \overline{\Omega}\Omega \cdot r - \overline{r}\Omega^2. \tag{2.2.1}$$

The radial component of the right hand side will look exactly like gravity to the meter:

$$\begin{aligned} \hat{r} \cdot \left(\overline{\Omega} \times \left(\overline{\Omega} \times \overline{r}\right)\right) &= \left[\frac{(r\cdot\Omega)^2 - r^2\Omega^2}{r}\right] \\ &= -a\Omega^2 \sin^2\theta \end{aligned}$$

where $r = a$ at the earth's surface, and $r \cdot \Omega = a\Omega\cos\theta$. So, the "gravitational" acceleration shown on the meter will really be $g - a\Omega^2\sin^2\theta$, instead of g. People on land routinely subtract off the $a\Omega^2\sin^2\theta$ term, assuming $\Omega = \Omega_0$. But, from a ship (or a plane), Ω also depends on v. To lowest order in v, the measured acceleration is approximately

$$g - a\Omega_0^2 \sin^2\theta - \sin^2\theta a \left[2\Omega_0 \frac{v}{a\sin\theta}\right]. \tag{2.2.2}$$

So, the lowest order correction due to east–west velocity is:

$$\Delta g = -2\Omega_0 v \sin\theta.$$

This is the "Eotvos correction." It works out to be $(7.5\sin\theta)$ mgal/knot. (1 knot = 1.15 mile/hour = 50 cm/sec.)

For ships, the Eotvos correction is typically 50 mgal. For planes, which move faster, it is typically 1000 mgal. So it is important to make this correction — and thus to know your velocity accurately. Nowadays, the space positioning measurements described below can be used to determine your velocity well enough.

There is also a contribution from north–south motion, but it's only about 1 mgal in planes and is negligible in ships. It's small because the Coriolis force it induces is mostly horizontal, not radial, and so it is not confused with g as easily.

But, the biggest problems come from unknown accelerations of the ship or plane. For example, if the meter support accelerates, the meter can confuse that acceleration with gravity.

Horizontal accelerations and tilting of the ship or plane are less of a problem than vertical accelerations. There are two ways to remove the effects of tilting and horizontal accelerations, both of them instrumental (see Figure 2.5):

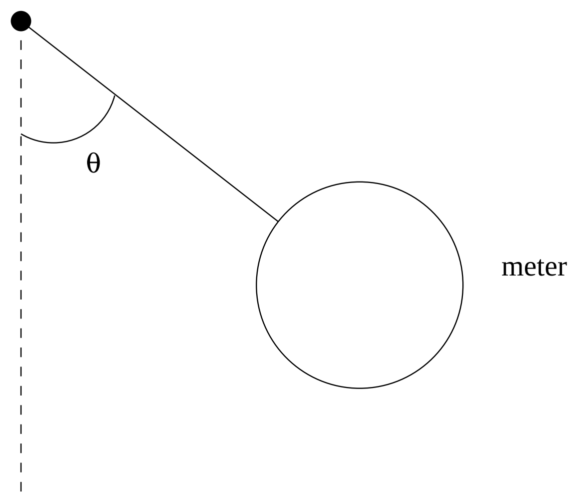

Figure 2.5:

1. Swing the entire meter about a pivot point. A horizontal acceleration starts it swinging. You measure θ and can thereby remove the swinging motion from the meter reading.

2. Put the meter on a stabilized platform with gyroscopes to keep it level and horizontal accelerometers to determine the horizontal acceleration.

Vertical accelerations are *the* biggest problem. A vertical acceleration is indistinguishable from the gravitational acceleration. Typically, vertical accelerations in ships are equivalent to 10^4–10^5 mgal in gravity. So, they can be huge.

The only way to get rid of the effects of vertical accelerations is to average over time. Some instruments do the averaging, partially, for you — by including damping to filter out high frequencies. For example, sea-going LaCoste-Romberg meters do it that way. But even in that case, at least some additional averaging of the data is needed.

For example, the instrument measures $g + \ddot{x}$, where $\ddot{x}$ = vertical acceleration. The averaged output over time T is:

$$g_{\text{avg}} \equiv \frac{1}{T}\int_0^T (g + \ddot{x})\,dt = g + \frac{\dot{x}(T) - \dot{x}(0)}{T}.$$

So, the longer the averaging time (T), the smaller the error. But, long term accelerations can't be removed.

Planes are smoother than ships, but you can't average for as long in planes because they don't stay in one place very long.

Typically, errors in g from ships are approximately 1 mgal. Errors in g from helicopters are approximately 2 mgal. Errors in g from planes are worse.

These numbers are large enough that sometimes other meters besides LaCoste-Rombergs are used in ships and planes. Maybe the most common of these is a "vibrating string" meter, which can achieve accuracies of only 1 mgal on land. This meter has a mass on the end of a thin metal ribbon. The ribbon is shaken and the mass and ribbon vibrate at a characteristic frequency, which depends on g. The frequency is measured to find g. The frequency is not particularly dependent on accelerations.

2.3 Free Fall Meters

These are absolute instruments. They replaced pendulums maybe 20 years or so ago. The idea is to watch an object fall and thereby determine its acceleration. If v_0 is the initial velocity, and x_1 is the vertical distance fallen after time t_1, then: $x_1 = v_0 t_1 + g t_1^2/2$.

If x_2 is the distance fallen after time t_2, then $x_2 = v_0 t_2 + g t_2^2/2$. Eliminating v_0 gives:

$$g = \frac{2}{t_1 - t_2}\left(\frac{x_1}{t_1} - \frac{x_2}{t_2}\right).$$

So, if you measure (x_1, t_1) and (x_2, t_2) you can find g. Since x is a distance and t is a time, this is an absolute meter. No calibration is necessary (except of your measuring tape and your clock). In practice, you probably want to measure lots of x's and t's, and then determine g by least squares fitting it to $x = v_0 t + g t^2/2$.

The earliest free fall meter was simply a meter bar which was dropped alongside a fixed pointer. A strobe light was flashed at a known frequency and a number was read off the meter stick at each flash as it fell. See Figure 2.6.

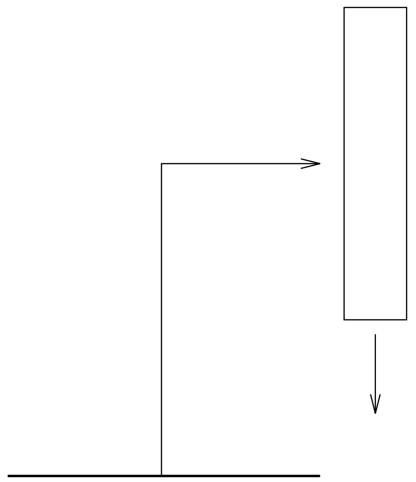

Figure 2.6:

Nowadays, free fall meters are much fancier. They involve laser interferometers.

2.3.1 Launch Type

You throw a mass, m, upwards — and measure two time *intervals* (see Figure 2.7):

1. t^A = time between upward and downward crossing of A;

2. t^B = time between upward and downward crossing of B.

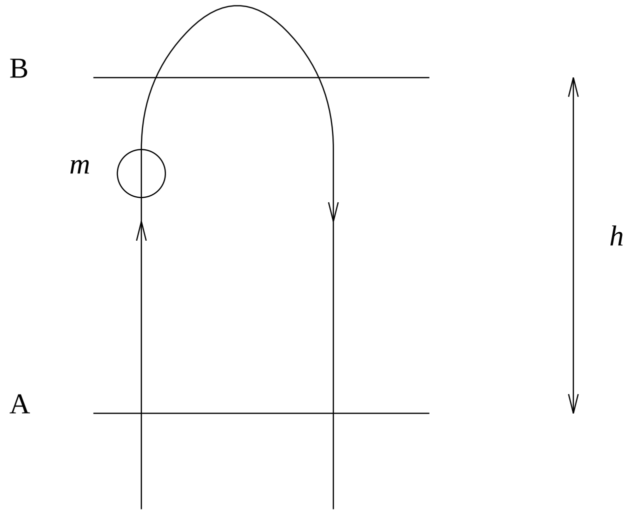

Figure 2.7:

Then, $t^A/2$ = time to get from A to the top of the trajectory. Since at the top the velocity is 0, then:

$$0 = v_A - g\frac{t^A}{2}$$

where v_A = upward velocity at A. So,

$$v_A = g\,\frac{t^A}{2} \tag{2.3.1}$$

Also, $(t^A - t^B)/2$ = time to go from A to B. So, if h is the height of B above A, then:

$$h = v_A\left(\frac{t^A - t^B}{2}\right) - \frac{g}{2}\left(\frac{t^A - t^B}{2}\right)^2 .$$

Using Equation 2.3.1 for v_A gives h in terms of t^A, t^B and g. Or inverting to find g:

$$g = \frac{8h}{(t^A)^2 - (t^B)^2}$$

So, measuring a distance (h) and two time intervals gives g.

The best launch instruments are very good. They have been developed by people in Japan.

2.3.1.1 Advantages over free-fall-only instruments

1. Air resistance cancels to 1st order. That's because the drag force is downwards on the way up (so in same the direction of gravity) , and upwards on the way down

(so opposing gravity). The instrument is run in a vacuum, anyway. But, the effects of the remaining air in the vacuum are small.

2. Timing biases are not as important as in a free-fall-only meter. (A timing bias refers to the time it takes, after the passage of the mass, for the instrument to register the time of passage.) That's because, here, it's time *intervals* that are measured.

2.3.1.2 Disadvantages

Launching the mass deforms and shakes the entire instrument in an unmodelable way, and that produces errors in g.

2.4 Free-Fall-only Meters

These involve masses which are dropped instead of launched. Jim Faller's group at the Joint Institute for Laboratory Astrophysics (JILA) has played an important role in developing this type of meter. Only a very few places in the world make them. I'll describe the JILA meter in some detail.

See Figure 2.8. The laser emits light with wavelength λ. The path length from the beam splitter to the fixed cube and back is $2L_1$. The path length from the beam splitter to the falling cube and back is $2L_2(t)$. $L_2(t)$ depends on time as the cube falls. L_1 is independent of time.

The detector sees a mixture of light from each corner cube. The phase difference of the two beams is the wave number, $k = 2\pi/\lambda$, multiplied by the difference in the distances the two beams have traveled $(2L_2(t) - 2L_1)$. So, $\Delta\phi$ = phase difference = $4\pi\left(L_2(t) - L_1\right)/\lambda$. Thus, if $2(L_2(t) - L_1)$ is an integer times λ, then the two beams are in phase and you get a large signal. If $2(L_2(t) - L_1)$ is a half integer times λ, the two beams are $180°$ out of phase and you get a small signal.

So, you drop the cube and look at the detected signal as a function of time. You use a zero-crossing discriminator to give a pulse at every zero crossing (that is, when the two beams are out of phase by $180°$). Every 3000 pulses, or so, you record the

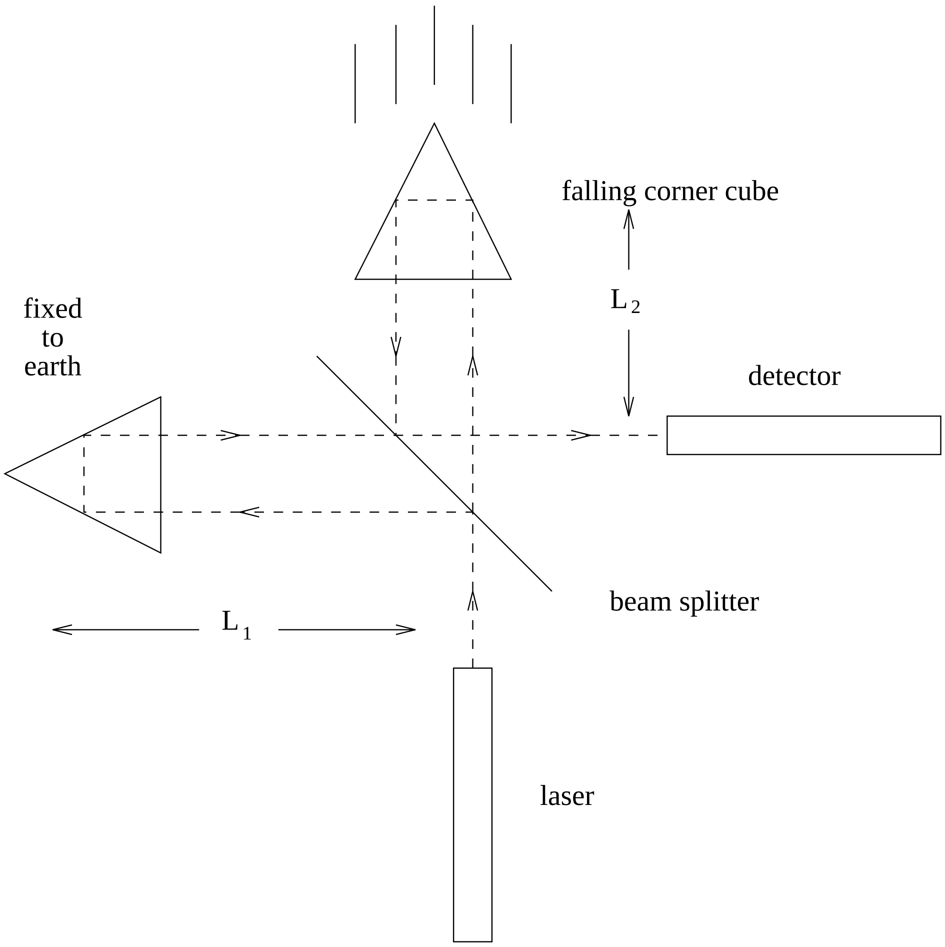

Figure 2.8:

time. So, you know the times corresponding to distance intervals of $(3000) \times \lambda/2 \approx (3 \times 10^3) \times (3 \times 10^{-5})$ cm ≈ 1 mm (the laser wavelength is about $\lambda = 6 \times 10^{-5}$cm). For a drop of approximately 20 cm you get 200 values of x and t which you fit to $x = v_0 t + gt^2/2$ to get g. It takes about 0.2 seconds to make a drop, and the whole process of dropping, computing g, and raising back to the top takes about 10 seconds. So, there are about 6 drops/minute.

There are four sources of error:

1. uncertainty in λ

2. electronic counting and timing errors

3. ground acceleration

4. non-gravitational forces: for example, electro-magnetic forces (EM) and air resistance

The objective is to obtain accuracies of about 1 μgal (1 part in 10^{-9}), which corresponds to 3 mm of vertical motion.

2.4.1 Uncertainty in λ

This is no problem nowadays. There are lasers with frequencies that are stable to 10^{-11}. You either use one of those directly, or you periodically calibrate your laser against one.

2.4.2 Electronic counting and timing errors

This is ok, nowadays. You need to know the timing to

$$\underbrace{(1 \times 10^{-9})}_{\substack{\text{accuracy}\\ \text{in}\\ \text{gravity}}} \times \underbrace{(0.2)}_{\substack{\text{time of}\\ \text{the}\\ \text{drop}}} \text{ sec} \approx 0.2 \text{ nsec.}$$

You can live with errors larger than this if they are constant during the drop. And, any errors larger than this are probably due to the time delay between fringe crossings and timer response, which *would* be nearly constant.

2.4.3 Ground Acceleration

This *can* be a problem. The typical ground acceleration (called "microseisms") due to the interaction of the oceans with the sea floor is on the order of $10^{-6}g$ and has energy peaked at periods of around 6 seconds. This acceleration affects the fixed corner cube and can map into errors in g. Its effects on the laser and detector don't matter, because both beams travel through the laser and detector in the same way — and it's the *difference* in the beam paths that you measure. Because 6 sec > 0.2 sec (the time duration of a drop), the acceleration is not averaged out during a drop.

One way to get rid of the problem is to average over lots of drops. But to obtain 1×10^{-9} accuracy with 10^{-6} random errors, you'd need to make $(1/(1 \times 10^{-3}))^2 \approx 10^6$ drops. This would take approximately 1000 hours.

To reduce this averaging time, they use a very long spring. They suspend the fixed cube from the spring. The spring has about a 60 sec period. Since 60 sec $\gg$ 6 sec, the cube at the end of the spring doesn't move much due to the microseisms. To get a spring with a 60 sec period, note that $kl = mg$ and $T = 2\pi\sqrt{m/k} \Rightarrow l = gT^2/(2\pi)^2 \approx 10^3 \cdot (60)^2/6^2$ cm ≈ 1 km, where $l =$ amount the spring is stretched at equilibrium. So, the spring, itself, would have to be substantially longer than 1 km.

Instead, they use what they call a "super spring" — a device that electronically mimics a 1 km spring. The idea is that in a real spring, points on the spring near the bottom move almost — but not quite — like points *at* the bottom. They use a sensor to detect differences between the bottom and the top of a *short* (20 cm) spring, and they move the top as though it and the bottom were, instead, the very bottom part of a 1 km spring.

The super spring reduces number of drops needed by a factor of about 400.

2.4.4 Non-gravitational forces

Electro-magnetic forces used to be a problem because the falling cube was dropped by shutting off a magnetic field. That caused eddy currents in the cube which interacted with other electro-magnetic fields and perturbed the fall. Now the cube is dropped mechanically, instead.

Air resistance is a potential problem. To get it down to an acceptable level (1 μgal errors in g), some instruments use a very high vacuum (about 10^{-7} mm of mercury, which is approximately 10^{-10} of atmospheric pressure). This low pressure causes mechanical problems (more friction, more equipment). Instead, the JILA instrument pumps down only to 10^{-5}–10^{-6} mm of mercury, and then uses a falling chamber that moves with the cube during its drop.

As the cube is dropped, it is tracked by the laser interferometer through a hole in the bottom of the chamber. The position of the cube relative to the chamber is also optically sensed as the cube falls. The chamber is then driven downward mechanically at the right speed so that the position of the cube doesn't change relative to the chamber. Thus, the chamber pushes the remaining air out of the way.

This also provides the mechanical dropping mechanism: Initially the cube rests on the chamber. The motion is started by suddenly driving the chamber downward, which causes the cube to come free and to start to fall.

2.4.5 Accuracy

The JILA instruments and their derivatives are believed to be accurate to the 1 μgal level, for averaging times of a few days to a couple weeks.

2.4.6 Satellites

Another way to measure gravity, and certainly the best way to obtain gravity at global scales, is to use satellites. We'll talk about this later, when we discuss space geodesy.

2.5 Positioning

The other broad goal of observational geodesy is to determine the locations of points on the earth's surface. Geophysicists are interested in this type of observation, because it allows them to look for displacements of the earth's crust. What level of accuracy do you need to begin to detect crustal motion? To measure continental drift you'd like to be able to determine the relative positions of two points to roughly 1 cm over baseline lengths (the "baseline" is the vector between the two points) of at least a few thousand km, which is the typical width of a plate. The relative motion between plates is generally on the order of a few cm/yr, so that this level of accuracy would let you determine plate motions with relatively small errors after a very few years. Note that 1 cm accuracy over a baseline of a few thousand km requires an accuracy of 1 cm/a few $\times\ 10^8$ cm = a few $\times\ 10^{-9}$.

The detection of local and regional tectonic deformation near plate boundaries does not require such high accuracies, because the deformation occurs over shorter baselines. For example, strain rates (the strain = the change in baseline length divided by the baseline length) over baselines of tens to hundreds of km are typically 10^{-7} per year in California, a tectonically active region. And a conventional geodesist — who only wants to tie local survey nets together — can settle for much worse accuracies.

2.5.1 Ground Based Techniques

To find positions, people use combinations of four sorts of measurements:

1. measurements of horizontal angles between points
2. measurements of elevation differences between points
3. measurements of distances between points
4. measurements of a point's coordinates with respect to the stars or to some other special astronomical coordinate system. (Or, what is equivalent, measurements of a point's latitude and longitude.)

You need to measure latitude and longitude because measurements of horizontal angles, elevation differences, and distances, just tell you the position of one point relative to another. Latitudes and longitudes tell you something about one point relative to inertial space.

How do you measure these things using ground-based techniques?

(Note: space techniques, that we'll talk about later, determine all coordinates of a point or a baseline at once. There is no natural separation into four categories.)

2.5.1.1 Horizontal Angles

Horizontal angles can be used to help find where points are on the earth's surface relative to one another. (Vertical angles, incidentally, tell you about the shape of that surface.)

You can't determine relative locations, though, without also measuring distances between points.

A typical instrument that measures horizontal angles consists of two telescopes that pivot about a common, vertical axis, and some sort of level used to make sure that axis is vertical. You want to measure the horizontal angle θ between lines $\overline{AB}$ and $\overline{AC}$. See

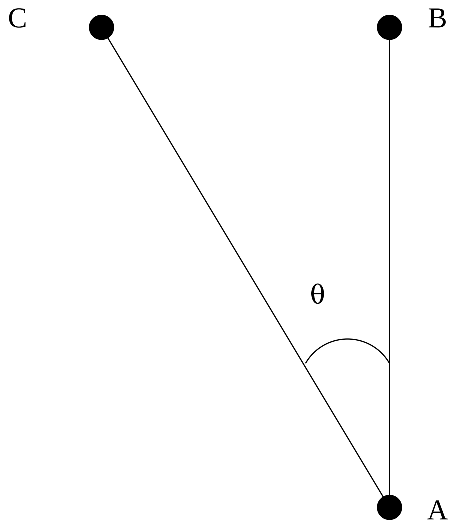

Figure 2.9:

Figure 2.9. You put the instrument at A, sight one telescope towards B and one towards C. You level the two telescopes so that they are both horizontal, and then measure the angle, θ, between the telescopes.

Suppose your goal is to find out where C is with respect to A. Suppose you have determined the absolute direction of the line $\overline{AB}$. Maybe you've done this with astronomical techniques. Or maybe you've arbitrarily defined your coordinate system so that $\overline{AB}$ has a specific direction (that is, you've tied your system to that line somehow). Then to find the position of C, you measure θ — and you measure the distance between A and C. Until the 1960's, distance measuring was tedious: you'd just roll a tape along the ground. So, you'd rather measure angles than distances. In that case, what you'd do was to measure θ. Then, you'd also measure the angle between $\overline{BA}$ and $\overline{BC}$. And, you'd measure the distance $\overline{BA}$ with the tape. That would give C. The advantage of this was that if you then wanted to find the position of some other point, D, you would measure the angle between $\overline{AD}$ and $\overline{AB}$, and the angle between $\overline{BA}$ and $\overline{BD}$, and then use the distance $\overline{AB}$ that you've already measured. So, you could do lots of points by measuring no distance except $\overline{AB}$.

This technique is called "triangulation." It is not used much anymore for geodetic/geophysical applications. Its accuracy is approximately 10^{-5}, which is not good enough for most geophysical studies. Its advantage was that it minimized tiresome distance measurements.

Nowadays, geophysical geodesists don't often measure horizontal angles at all. Instead, they use only distance measurements, which are now easy to make and are very accurate. The techniques involve laser interferometers and I'll wait a little to describe them.

Meantime, how does this new method work? Suppose you know the locations of two points, A and B. See Figure 2.10. You want to know the location of C. You measure $\overline{AC}$ and $\overline{BC}$, which then tells you that C lies on the intersection of two known circles.

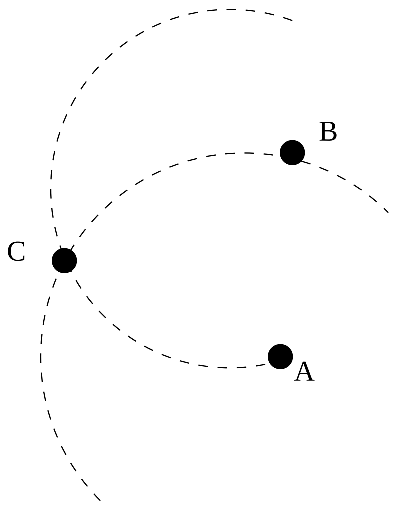

Figure 2.10:

That determines C uniquely.

There is, sometimes (not often used), a slight twist to this. You can measure the *difference* between $\overline{AC}$ and $\overline{AB}$, and then the *difference* between $\overline{BC}$ and $\overline{BA}$. By comparing those two distance differences, C can be found. There can be advantages to measuring distance *differences*, rather than absolute distances.

Whichever method you use is called *trilateration*. It is much more accurate than triangulation. I'll describe the accuracies when I describe distance measuring instruments, below.

2.5.1.2 Elevation

The basic methods of measuring elevations haven't changed, as far as I know, for hundreds of years. You want to measure the elevation difference between A and B. See Figure 2.11.

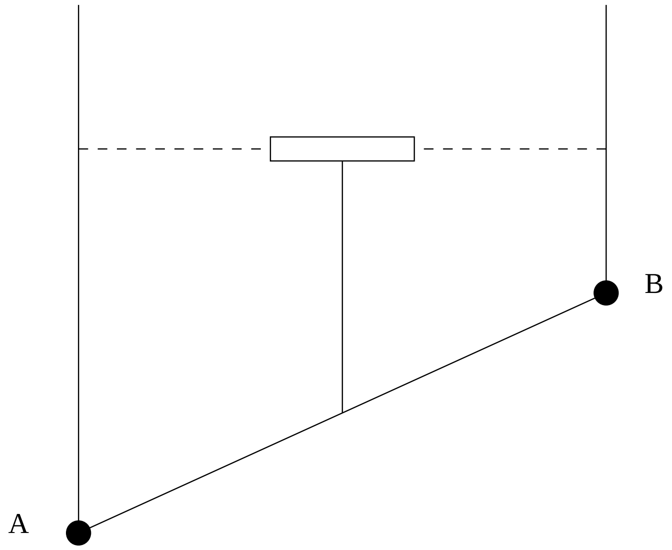

Figure 2.11:

You put two vertical rods with meter (and cm, mm, etc.) markings on them, at A and B. You put the leveling instrument — consisting of a telescope and a level — in between A and B. You use the level to keep the telescope horizontal, and sight first on the A rod, and then on the B rod. You read off the numbers you sight on from the two rods, and the difference between those numbers is the difference in elevation. Here you can see the importance of gravity for determining elevation: the orientation of a horizontal surface depends on the direction of gravity.

One obvious limitation of this technique is the finite rod height. The standard height is 2.5 m, so you can't measure over a baseline with elevation changes of more than this.

Surveying this way in rough topography is tedious. An alternative is to measure vertical angles, by training a telescope at the end of the baseline, and determining the angle between the telescope and the horizontal plane.

The errors in leveling are, traditionally, assumed to be 10^{-5}; that is, 1 cm over a 1 km difference in elevation. But, leveling hasn't always been done that well in the past. There are two sources of systematic error that can cause topographic-related errors larger than this, if not properly accounted for:

2.5.1.2.1 Leveling Errors

2.5.1.2.1.1 Rod calibration errors If meter marks on the rods aren't in quite the right places, you'll get topographic-related errors. If you switch rods from one survey to the next, these errors will make you think the ground has moved. You get a similar effect if you change the lengths of the individual baselines you sight over. In that case you will be sighting on different sets of meter marks which will have different errors.

2.5.1.2.1.2 Refraction This is a hard problem. The light going from the rod to the telescope is bent due to temperature gradients along the ground. So the number you read off the rod, and the telescope, are not quite on the same horizontal surface. People do know how to correct for this effect given the temperature on the ground. But temperature measurements were not always made in past surveys. The corrections are dependent on the variation in topography along the line. If you do not make these corrections, or if they are not made properly, there can be topographic-related errors as large as 10^{-4}; This effect can be particularly severe if you change the baseline lengths you level over, since the size of the correction varies with baseline length in a non-linear fashion.

2.5.1.2.2 Fluid Tiltmeters One other method is occasionally used to measure elevation differences between points. This method involves fluid tiltmeters. There are two versions:

1. You measure the height of the fluid column in each of the two vertical segments of the container shown in Figure 2.12. The difference in heights = the elevation difference.

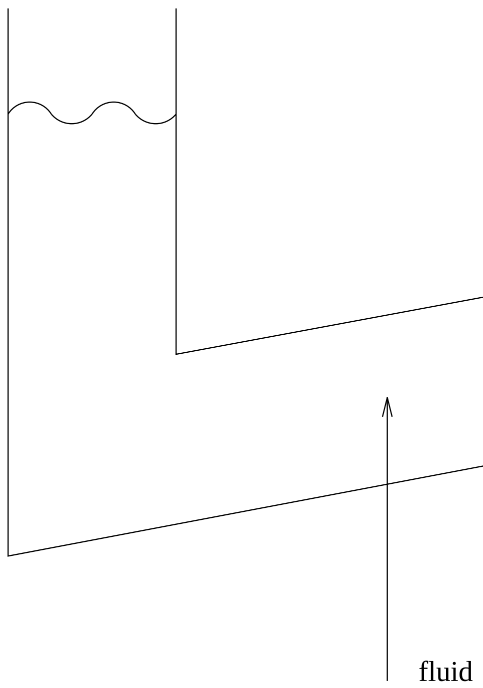

Figure 2.12:

2. You measure the fluid pressure at each end of the pipe in Figure 2.13.

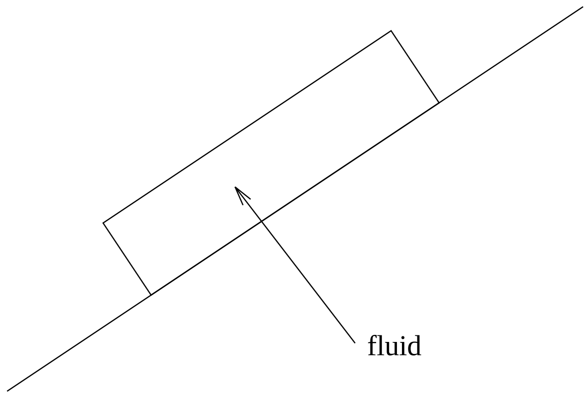

Figure 2.13:

These instruments can measure 10 m elevation differences — and so can be useful when there is a lot of topography. They are usually used, though, as fixed instruments to look for time dependent tilts at one place. The results can be affected by temperature, so the instruments must be thermally insulated. Instruments have even been built which have two fluids with different thermal expansion coefficients, so the thermal effect can be removed.

There are related, short baseline instruments called "tiltmeters." These are designed to measure time dependent changes in the normal to the surface at a fixed point. Tiltmeters can be fluid tubes, as described above, but here they are only a few cm long. Or, they can be horizontal pendulums.

2.5.1.3 Latitude and Longitude

Latitude and longitude measurements used to be made at certain points along a survey to fix those points with respect to the stars, and to reduce cumulative errors which can build up over a survey. There were also fixed stations around the globe which measured their latitude and longitude on a routine basis. Results from the fixed stations were used to fix the terrestrial coordinate system to the stars, and thereby to look for variations in the earth's rotation. These stations were all closed down about a decade or so ago. But, they provided information about the earth's rotation going back into the 19th century that is still being used today to study long period rotation fluctuations.

How did people measure their latitude and longitude relative to the stars?

2.5.1.3.1 Latitude You align a rod along the local vertical. You watch a star with a telescope as the star moves from east to west across the sky. The star is at its highest point as it passes the north–south meridian running through the observer. At that point you measure the angle z (the "zenith distance" of the star) between the telescope and the vertical rod. Then θ = co-latitude $= \delta - z$, where $\delta =$ *known* stellar co-declination = angle between the star and the earth's mean (ie. time-averaged) rotation axis, $\overline{\Omega}$. See Figure 2.14.

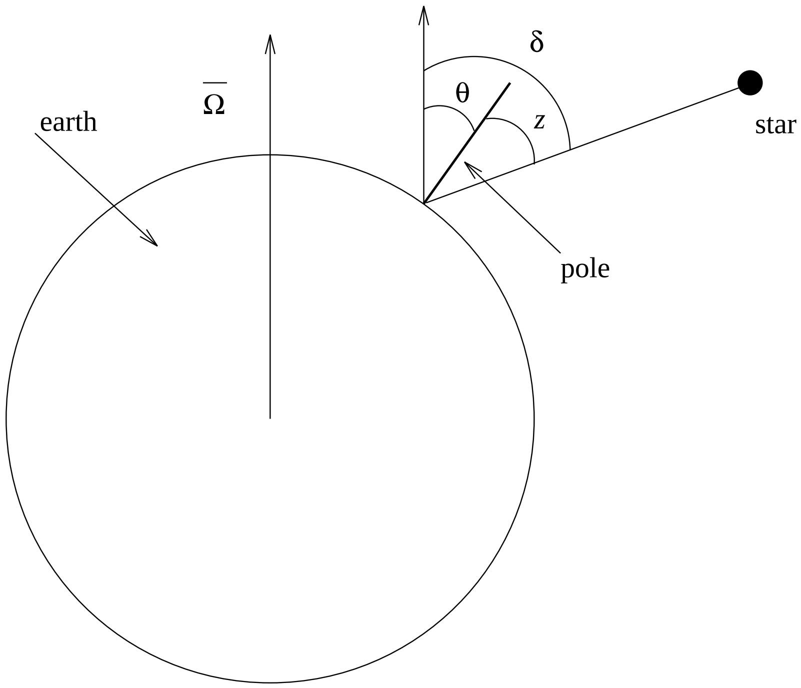

Figure 2.14:

Refraction of the starlight as it passes through the atmosphere causes the light to bend, and results in errors in θ. These can be partially modeled, but remaining errors are approximately 1 second of arc, which corresponds to a position error of the point on the earth's surface of about 30 m.

To reduce this error, people often looked at a pair of stars on two sides of the local vertical. See Figure 2.15. You measure z_1 and z_2. Then $\theta = \delta_1 + z_1$ and $\theta = \delta_2 - z_2$. So,

$$\theta = \frac{1}{2}\left[\delta_1 + \delta_2 + z_1 - z_2\right].$$

So, you can determine θ by measuring $z_1 - z_2$: the *difference* in angles. For z_1 and z_2 about equal and very small, the refraction errors mostly cancel in $z_1 - z_2$. Accuracies are approximately 0.1 arc seconds, which maps into roughly 3 m of surface position error. This is still pretty large, but if lots of stars are observed, you can reduce the error further.

For example, the big international services which routinely measured latitude at fixed stations, could determine θ to approximately 0.01 arc seconds when averaged over a month or so (about 30 cm at the surface), although they still got errors of 1–3 meters at

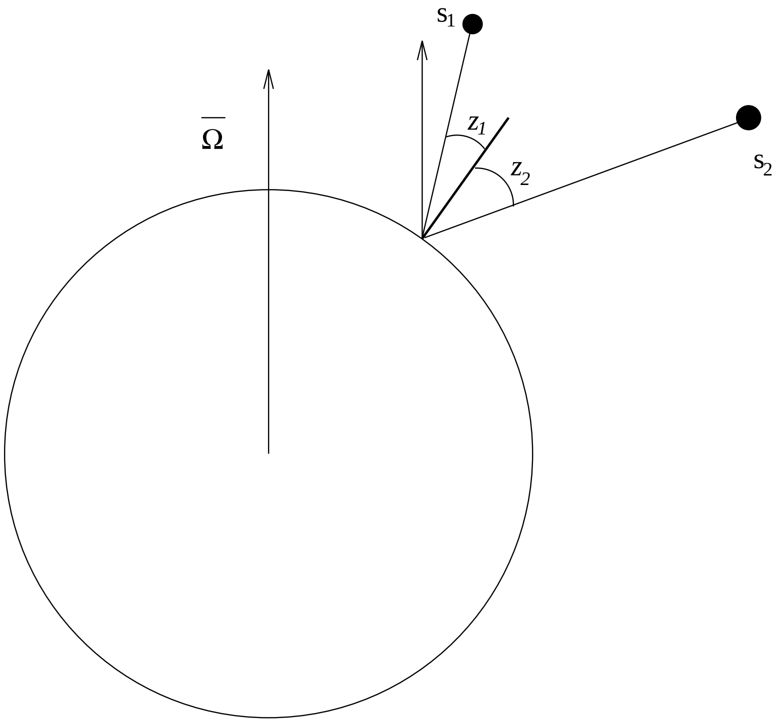

Figure 2.15:

the surface at the annual frequency, due to annual variability in the atmosphere which can't be averaged out.

2.5.1.3.2 Longitude The method was similar to that used for latitude observations. You watch a star with a telescope until it passes the local meridian. Then you record the time using an accurate clock. You know from a table the time the star passed the Greenwich meridian. Let this time be t_G. Then ϕ, the angle west of Greenwich, is

$$\phi = (t - t_G)\Omega$$

where Ω = rotation rate of earth. Individual determinations of east–west position through the mid-1960's were accurate to 10–30 m, due to refraction and, especially, to clock errors. You can average the errors down assuming random clock errors.

Again, gravity affects both the latitude and longitude measurements: stellar positions are recorded relative to the local vertical, and the vertical depends on local gravity.

2.5.1.4 Distances

There have been rapid developments in distance-measuring instruments over the last few decades. Tape measures have been replaced by instruments which send and receive electro-magnetic radiation — usually light. The method is to shine a continuous beam of mono-chromatic radiation on a distant target. The beam is usually then reflected back to the source, where the phase difference between the emitted and received waves is measured. This phase difference tells you something about the distance between the two points. For example, suppose the two phases differ by δ. Then, you could deduce that the round trip distance between the two points was an integer number of wavelengths *plus* $\left(\frac{\delta}{2\pi}\right)$ wavelengths. If D is the source-target distance, so that $2D$ is the round-trip distance, and if λ = wavelength, then $2D = l\lambda + \lambda\left(\frac{\delta}{2\pi}\right)$, where l is an integer. You wouldn't, of course, know l. This problem (finding l) is called the "2π ambiguity." To find l you must either:

1. already know the distance to within one wavelength through some other means; or

2. change λ slightly (typically by on the order of 1%), do the experiment over again, and combine the results from both experiments. This allows you to reduce the 2π ambiguity considerably, as we will see.

In either case, using un-modulated light doesn't work well. The wavelength of light is approximately 5×10^{-7} m, which is simply too small. For this wavelength, the 2π ambiguity would require you to already know distances to $5{\times}10^{-7}$ m. By using method 2, above, you could reduce the 2π problem, but this would be hard to implement effectively in such an extreme case.

Instead, you use light as a carrier, but you modulate it at low frequencies and compare phases of the modulation. The modulation can be in amplitude, frequency, phase or polarization. For amplitude modulation, the signal looks like Figure 2.16. For typical modulations — usually radio frequencies — λ can be tens of cm's to a few m's. So the

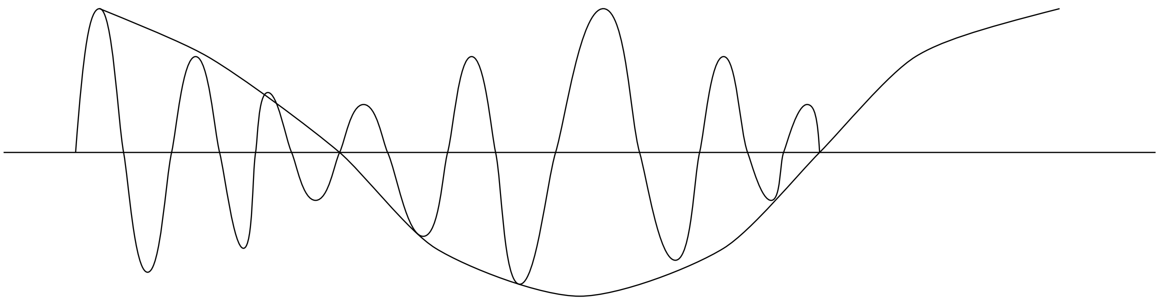

Figure 2.16:

2π problem is manageable, in the sense that you can readily overcome it by changing λ slightly and re-doing the experiment.

Why not use radio waves directly (no carrier)? The answer is that light waves are easier to use, particularly if you can use lasers.

Why not use a modulation wavelength of much larger than tens of cm's to a few m's, so as to further minimize the 2π problem? The answer is that the use of longer wavelengths tends to reduce accuracies. For example, if you are able to measure phase differences to an accuracy of Δ radians, then you will have an accuracy in your distance measurements of $\lambda\Delta/2\pi$, which is proportional to λ. So, the smaller the wavelength, λ, the smaller the error.

These distance measuring devices are called *Geodimeters*. Here's a short history of their development.

2.5.1.4.1 White light These were the first geodimeters. They predated lasers. They worked sort of like this:

A light signal is modulated by shining it through a revolving wheel with a slit. When the slit points, say, directly downward, the slit is in front of a light source. In that case light gets through and travels toward a mirror several km away. Thus, an observer at the mirror sees what looks to be a strobe light blinking on and off. The light from the source has had its amplitude modulated by the wheel. The modulation is a series of square waves: equal to 1 when the slit is in front of the source, and equal to zero when the slit is not in front of the source.

The light hits the mirror and returns to the wheel. If the returning light hits the wheel when the slit is pointing directly downward again, the light passes through and is recorded by a detector located next to the original light source. If the slit is not pointed directly downward, the returning light doesn't get through. (This is basically the way Michelson and Morley measured the speed of light.)

Let τ = the time it takes the wheel to revolve once = the period of the modulation. Let $2D$ be the round trip distance traveled by the light (between the wheel and the mirror and back again) so that the round trip travel time of the light is $2D/c$, where c is the velocity of light. Thus, the light gets through to the detector if $2D/c = \tau l$ where l is an integer. So, light is detected if

$$2D = lc\tau. \tag{2.5.1}$$

The product $c\tau$ is the wavelength of the modulation. To determine the distance D, you adjust the rotation period, τ, until light does get back through to the detector. In that case you know that D satisfies Equation 2.5.1. The unknown integer l represents the 2π ambiguity.

Why not slow the wheel down — increase τ — to reduce the 2π problem? For example, if you know the distance is around, say, 10 km, why not make τ so large that $c\tau/2 \approx 10$ km, and then vary τ around this value so that light gets back through? Then, you'd know $l = 1$. The problem with this is that your estimate of the time that the light takes to arrive back at the receiver, has an uncertainty that depends on the width of the slit and on how long the slit is in front of the detector. In other words, the uncertainty in D is

$$\frac{\text{width of slit}}{\text{circumference of wheel}} \times \frac{c\tau}{2}$$

which increases with τ. Another way to say this is that the phase difference between the transmitted and returning signals can be determined only to within

$$\frac{\text{width of slit}}{\text{circumference of wheel}} \times 2\pi \text{ radians.} \tag{2.5.2}$$

The corresponding error in distance is one-half the wavelength over 2π (that is, $\frac{c\tau}{2} \cdot \frac{1}{2\pi}$) times Equation 2.5.2. So, you don't want to pick τ large.

So how do you deal with the 2π problem? Suppose you pick, say, $c\tau = 10$ cm. To find l you would have to already know D to 5 cm, which is unlikely. However, suppose you change τ a little and make the measurement again. You find τ_1 and τ_2 that both work. So:

$$D = l_1 \left(\frac{c\tau_1}{2}\right)$$

and

$$D = l_2 \left(\frac{c\tau_2}{2}\right)$$

where τ_1, τ_2 are known, and l_1 and l_2 are unknown integers. So,

$$\frac{l_1}{l_2} = \frac{\tau_2}{\tau_1}$$

Suppose τ_2 and τ_1 are very close: maybe, say

$$\frac{\tau_2}{\tau_1} = 1.01.$$

Then, since l_1 and l_2 are integers, you know that $l_1/l_2 = 101/100$, or $202/200$, or $j(101/100)$ where $j =$ integer. So, now you know that $l_2 = 100$, or 200, or 300, etc., and so

$$D = j\,100 \left(\frac{c\tau_2}{2}\right)$$

where $j = 1, 2, \ldots$. So, you've reduced the 2π uncertainty down to 100×10 cm $=$ 10m, and you'd only have to know the distance beforehand to 10 m. This is the general approach used by all geodimeters to reduce the 2π problem.

The conventional white light instrument described above became obsolete when lasers became available. Its modulation was somewhat different than I have described it, but the idea was similar. The accuracies were a little worse than 10^{-6} over 5–50 km baselines. So, it could do horizontal positions about one order of magnitude better than could triangulation. But, it only worked at night.

2.5.1.4.2 Radio signals To get something that worked during the day, people invented the *Tellurometer*. It used un-modulated radio waves with no carrier. Its accuracy was about 3×10^{-6} over baselines of up to 70 km.

2.5.1.4.3 Lasers All geodimeters now use lasers for the carrier signal. Lasers can be aimed better than white light or radio waves, and they work in the daytime. They are modulated electronically, not mechanically. Typical modulation frequencies are approximately 50 MHz, which gives a modulation wavelength of around 5 m. So, the 2π ambiguity is around 2–3 meters. You reduce the ambiguity by changing λ slightly and repeating the phase measurement, as described above.

There are various ways to detect the phase shift of the returning signal. One way is to mix it with the transmitted signal. Then you adjust the modulation frequency slightly until the combined modulation signals are in phase. You achieve this by maximizing the combined modulated signal. Then you know that at that frequency the round trip distance is exactly an integral number of modulation wavelengths.

The accuracies with the laser instruments are about the same as with white light: about 10^{-6}. The biggest error source comes from the atmosphere the wave propagates through. To find the distance traveled by the wave, you need to know the wavelength of the modulation. Instead, what you *do* know is the angular frequency, ω, of the modulation. To deduce λ from ω you need to know the velocity of light, and that depends on the index of refraction n:

$$\lambda = \frac{c2\pi}{n\omega}.$$

And, n depends on the density of the air the wave travels through, and on the water vapor content of that air. The dependence on water vapor is relatively weak at optical frequencies (it is the index of refraction at the frequency of the *carrier* wave that is pertinent), though it is non-negligible.

So, to improve the accuracies, you must determine the densities of the air and of water vapor integrated along the path of the signal. The usual way to do that is to fly an airplane along the path while you make the measurement, and to monitor pressure, temperature, and humidity from the plane. The densities can then be estimated from these measurements. The accuracies people obtain using this method are approximately 3×10^{-7} over baselines of 30 km or less.

Airplanes are expensive and hard to schedule. An alternative is to use a multi-color instrument. Larry Slater (formerly of CIRES) has designed and built two-color instruments, which he has used in California and in New Mexico. Judah Levine in JILA has built a three-color instrument. I will briefly describe both of these.

2.5.1.4.3.1 Two-color meters The idea is to use the known dispersive properties of light to variations in atmospheric density: make measurements at two carrier frequencies — red and blue, say — and then compare to deduce the index of refraction at either frequency.

Specifically, $n-1$ ($=0$ for a vacuum) depends on the carrier frequency ω. You can write $n-1=\alpha\rho_{\text{air}}+\beta\rho_{\text{vapor}}$ where ρ_{air} and ρ_{vapor} are the densities of dry air and water vapor, and α and β depend on ω in a known way. The $\beta\rho_{\text{vapor}}$ term is much smaller than the $\alpha\rho_{\text{air}}$ term at optical frequencies. The two-color meter just tries to deduce the $\alpha\rho_{\text{air}}$ term. The $\beta\rho_{\text{vapor}}$ term is inferred by making meteorological measurements at both ends of the baseline.

So you know α but not ρ_{air}. If $n(\omega_1^c)$ and $n(\omega_2^c)$ are the values of n at the two carrier frequencies, then you know

$$\frac{1-n\left(\omega_1^c\right)}{1-n\left(\omega_2^c\right)}\equiv\Delta \tag{2.5.3}$$

without making any measurements (assuming you ignore the $\beta\rho_{\text{vapor}}$ term).

You then measure the distance D at the two frequencies. Suppose you find that $D=A\lambda_1$ and $D=B\lambda_2$ where A and B are your measured results, and λ_1 and λ_2 are the wavelengths of the modulation. (For both carrier frequencies, you have presumably gone through the usual trick of adjusting the modulation frequency slightly to resolve the 2π problem when determining A and B.) But, what are λ_1 and λ_2? If, in each case, the modulated frequency is ω, then

$$\begin{aligned}\lambda_1 &= \frac{c2\pi}{n(\omega_1^c)\omega} &(2.5.4)\\ \lambda_2 &= \frac{c2\pi}{n(\omega_2^c)\omega}.\end{aligned}$$

So, equating the two results for D gives:

$$A\frac{c2\pi}{n(\omega_1^c)\omega} = B\frac{c2\pi}{n(\omega_2^c)\omega}.$$

Or

$$n(\omega_2^c) = \frac{B}{A}n(\omega_1^c). \tag{2.5.5}$$

Then, using Equation 2.5.5 in Equation 2.5.3 gives

$$n(\omega_1^c)\left[\frac{B}{A}\Delta - 1\right] = \Delta - 1.$$

So, your measurements of B and A give $n(\omega_1^c)$, which then goes in Equation 2.5.4 to find λ_1 and then in $D = A\lambda_1$ to give D.

In his two-color meter, Larry Slater modulates his carrier waves with a rotating polarizing crystal. He de-modulates by having the wave pass back through the crystal when it returns. (So, it's like the white light passing through the rotating cog wheel.) The returning signal will be maximum if the crystal has exactly the same orientation when the wave returns as when it was sent. He adjusts the modulation frequency slightly until that maximum is obtained. (Actually, he tries to minimize the signal rather then to maximize it — so the round-trip distance works out to be an odd number of half wavelengths.) His modulation works out to $\lambda \approx 10$ cm. So, the 2π ambiguity in one way distance is approximately 5 cm.

His accuracy is approximately 1×10^{-7}, which is slightly better than can be obtained using a one color geodimeter with a plane. But his meter is much cheaper to run and easier to operate. It has a somewhat shorter range, about 10–15 km. And, it's not very portable. The accuracy is limited by the uncertainty in the water vapor along the path.

2.5.1.4.3.2 Three-color meter All of these problems are reduced in Judah Levine's three-color meter. This instrument uses red and blue to get ρ_{air}, and an un-modulated microwave signal ($\lambda \approx 4$ cm) to infer ρ_{vapor}. You've got to use a microwave signal, instead of a third optical carrier, because the effects of water vapor are not notably dispersive at optical frequencies. That is, β is about the same at all optical frequencies.

The range is increased, too, to 30–50 km. The reason is that the optical portion of Judah's meter is one way instead of round trip, and less of the laser beam is lost if it travels a shorter total distance. (Though the microwave signal used to infer the water vapor is two way.) The meter is also portable. It fits into a four-wheel-drive vehicle or a helicopter.

In this instrument, red and blue, modulated in polarization, are sent from one end to the other, where they are received and the phase difference is measured. This gives $\rho_{\rm air}$. (In fact, in this configuration the instrument could be used with conventional one-color geodimeters to replace the plane.) Then, only red is sent back to the starting end. When this second red signal is sent, the phase of its modulation is locked to the modulation phase of the first red beam, so it's just like a reflection. The red is received at the starting end, the modulations of the returning signal and the outgoing signal are compared to determine the phase shift, and the distance is inferred.

The 2π problem is resolved by changing the modulation frequency by either 1% or 0.1%; there is a switch for each. A change in the frequency of 1% reduces the 2π ambiguity by a factor of 100, from about 10 cm to 10 m. And, a change of 0.1% reduces the ambiguity by another factor of 10: to about 100 m. For long-term monitoring the meter is usually run at a single modulation frequency for a few days. The frequency is then switched over to the slightly altered value for an hour or so, until the 2π problem has been resolved. The modulation frequency is then switched back to its original value.

The overall accuracies for this instrument are about 2–3 $\times 10^{-8}$, which are substantially better than those of any other geodimeter. One error source is that red and blue are bent by the air along slightly different paths, and so sample different n's.

2.5.1.4.4 Strainmeters These are another sort of distance-measuring instruments, but they have a different purpose. They are designed to sit at a fixed location and measure the change in distance with time between two closely separated points. They are not survey instruments.

They are very accurate. The most accurate of them can measure a change in a fixed baseline to one part in 10^{10}. The baselines, though, are less than 1 km: usually tens of meters.

The earliest strainmeters were two piers, fixed in the ground, with a horizontal quartz rod attached to one pier and almost touching the other. See Figure 2.17.

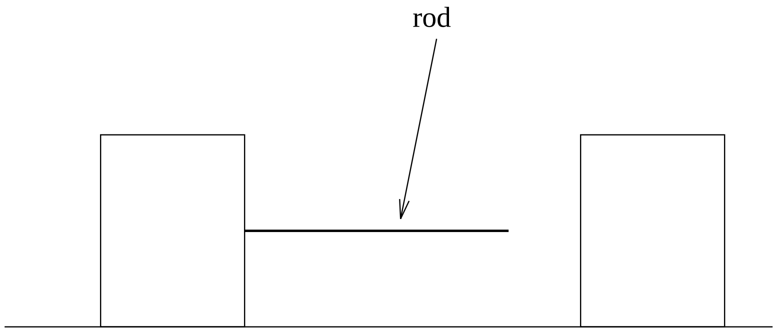

Figure 2.17:

The distance between the rod end and the second pier was continually monitored. The typical baseline, here, would be less than 100 m. Many of these are still used.

The newer and more accurate strainmeters are laser interferometers. There aren't many of these. Judah Levine built a strainmeter with a 30 m baseline that he operated for a while in the Poor Man mine, west of Boulder, Colorado. It has been shut off now for a number of years.

Judah's instrument used two lasers. One was kept running at a stable frequency (it had to be continually calibrated). The other sent light from one point in the tunnel, to a point 30 m away, and then back again. The path was evacuated of air to eliminate refraction. The frequency of this second laser was continually tuned so that the returning wave was in phase with the transmitted wave; thus, the round trip travel time was an integral number of wavelengths. The wavelength was then determined by beating the output from that laser against the output from the stable laser and looking at the beat frequency. A change in the beat frequency meant a change in the path length. Note that the strainmeter does not modulate the light; it works with the phases at the optical frequencies. This does not work well for geodimeters, because the 2π problem would be too severe. But for strainmeters, the 2π problem is not an issue. For strainmeters you are not trying to determine the absolute distance between the points. Instead, you are interested in time-dependent changes in the distance. And, since you monitor the baseline continuously, changes in the baseline can be determined without knowing the integral number of whole wavelengths traveled by the wave.

2.5.2 Space Techniques

There are two types of space positioning techniques:

1. those that involve measuring distances to objects in space, usually artificial satellites (using either lasers or radio waves), though laser ranging measurements to the moon have also provided useful geophysical information.

2. a method that compares radio signals from quasars received at two widely separated radio antennas (called very-long-baseline-interferometry — VLBI).

The satellite techniques can also provide information about the earth's gravity field. The lunar ranging technique tells you about the moon. And the VLBI technique tells you about the structure of quasars.

2.5.2.1 Satellites

These have been used for geodetic purposes for 30 years or more. The first satellites were "passive," and were used to tie local or regional survey lines together. This was done by observing the satellite and determining its coordinates from each line, simultaneously. This was particularly useful for tying surveys together across large bodies of water. Before satellites people used flares for this purpose, although those weren't good over as long a distance. For this application, you don't care about the satellite's precise position, so long as you can see it.

Nowadays, all satellites which give scientifically useful results are "active." You track the satellite from the ground, to determine its distance away from you. If you know where the satellite is in some terrestrial or inertial space coordinate system, you can then figure out where you are in that same coordinate system. To know where the satellite is, you must have information about the earth's gravitational field, since that field determines the satellite's orbit. Conversely, you can use the observed orbit to better constrain the gravity field.

Actually, the first useful information from these satellites was better information about the gravity field. That's because, initially, satellite orbits could be modelled much less accurately than could the positions of the ground stations. As the gravity field improved, the orbit errors became small enough that people could begin to use the satellite results to determine variations in the earth's rotation better than could be obtained using traditional astrometric methods. And now, and for the last 5–10 years or so, people have been able to see the tectonic motion of individual ground points. During all this time, of course, the gravity field results have continued to improve.

To maximize the gravitational and positioning applications from a geodetic satellite, the orbit should have a high inclination (the angle between the orbital plane and the equator) so that the satellite ground track covers as much of the earth as possible. To solve effectively for the orbit, and to use the orbital results to learn more about the earth's gravity field, a geodesist must deal with certain non-gravitational orbital effects. These include:

Atmospheric drag

You can reduce this by going to a high orbit where there's not much atmosphere. This also tends to reduce the effects of uncertainties in the gravity field, since those die away rapidly with altitude. On the other hand, if your goal is partly to learn about the gravity field, then you want to be closer to the surface.

Solar and terrestrial radiation pressure

This is hard to deal with. You either model it empirically, or try to average or filter it out.

Differential gravity acting across a satellite

A a satellite is not a point mass, and gravity is not uniform across it. This can cause motion of the satellite which might affect the geophysical observations, but which tells you nothing about the earth. The best solution is to make the satellite spherically symmetric, though this cannot always be done.

The earliest active satellites had flashing lights. You tracked them against the stellar background using cameras, and so learned something about the orbit. These methods are no longer used.

Now there are two methods used to range to a satellite. You can use radio waves, which originate from the satellite and are recorded on the ground. Or, you can use lasers, which sit on the ground and reflect a light beam off of the satellite.

The laser satellites have all been designed primarily to do geophysics. They are used to find the positions of the lasers, while also providing information about the gravity field. There have been several of these: GEOS1, GEOS2, Starlette, GEOS3, PAGEOS, LAGEOS I and II, and more recent satellites launched by Japan and by the Soviet Union. So far, the LAGEOS satellites have proven to be the most useful of these.

Most of the radio-ranging satellites are designed primarily for navigation, and are funded by the US Department of Defense. Geodesists have usually been able to use these satellites for scientific purposes. I will describe the radio-ranging satellites first.

2.5.2.1.1 Radio Ranging Satellites The navigation system used today by the Department of Defense is the Global Positioning System (GPS). This system measures phase shifts/time delays of radio signals transmitted from a constellation of GPS satellites. Prior to GPS, Department of Defense navigation was done using Doppler measurements of satellite-based radio transmissions. Both these techniques will be discussed here.

2.5.2.1.1.1 Doppler Satellites When this navigation system was operable, there were six Doppler satellites orbiting the globe at any one time. In this technique, each satellite transmits radio waves of known frequency, f_S. The satellite oscillators are stable, so that f_S is known very well. The signal is recorded at a receiver on the ground, and the frequency, f_R, is measured. $f_R \neq f_S$, because the satellite is moving with respect to the ground. Specifically:

$$f_R = f_S \left[1 + \frac{(\overline{v}_S - \overline{v}_R) \cdot \hat{r}}{c}\right]$$

where $\overline{v}_S$, $\overline{v}_R$ = source and receiver velocities, and $\hat{r}$ = unit vector from source to receiver. If ρ = the range between the source and receiver, then

$$(\overline{v}_S - \overline{v}_R) \cdot \hat{r} = -\partial_t \rho = - \text{ range rate.}$$

So,

$$\frac{f_R - f_S}{f_S} = -\frac{\partial_t \rho}{c}. \tag{2.5.6}$$

So you can find $\partial_t \rho$ by measuring f_R. Because you know the satellite orbit (supplied by the Department of Defense, as determined from about a dozen fixed tracking stations), you can use the results for $\partial_t \rho$ to determine your position. The results from the fixed tracking stations could also be averaged together to give earth's rotation results that were considerably better than those obtained by observing stars with telescopes. The positions of the fixed stations could be determined to a few tens of cm — close to 1 m. So it was not good enough to measure continental drift effectively, for which you'd like accuracies closer to 1 cm.

The accuracy limitations for this technique were mostly instrumental. But there were external systematic effects active here which affect all radio transmission-type satellites, and so I'll describe them here.

The problem is, as for all space-geodetic experiments, you must know the velocity of the radio wave. That means, you must know the index of refraction, n. For example, the c in Equation 2.5.6 should really be c/n.

We saw, earlier, that at optical frequencies n depends on the dry air density and on the density of water vapor along the path. The same is true for radio waves, except that in the radio band the atmosphere is not dispersive. That is, n is approximately frequency independent and so, unlike with the two-color geodimeters, it does *not* help to use two radio frequencies. Another difficulty that is not an issue at optical frequencies, is that at radio frequencies n depends on the number of charged particles (ions) along the path. So,

n is perturbed relatively strongly in the ionosphere (the region from 100 km to several hundred km altitudes above the earth).

Of these effects on the index of refraction, the ionospheric effect is the largest at radio frequencies. It can cause tens of meters of position error if uncorrected. Luckily, the ionospheric effect is dispersive at radio frequencies. So the way to get around this problem, for any radio-ranging technique, is to range at two radio frequencies. n has the form $n = 1 + \gamma\rho_{\rm ion}$, where γ depends on frequency and is known, but where $\rho_{\rm ion}$, the ion density along the path, is unknown. By combining the ranging results for the two frequencies you can infer $\rho_{\rm ion}$, which allows you to deduce n. The result for n can then be used with either one of the range measurements, to compute the receiver-satellite distance (or to give the range rate, in the case of the Doppler satellites).

The second largest effect is from the dry air. If uncorrected, it can perturb the inferred receiver-satellite distance by approximately 2 meters. People usually model this effect by making pressure and temperature observations at the receiver and using them in some semi-empirical atmospheric model. This seems to work well, as long as the satellite isn't too low in the sky. (It is harder to model the *horizontal* variations in pressure and temperature, than it is to model the vertical variations.) In general, the dry air error can be reduced to the 1–2 mm level.

The third largest effect is the water vapor density. It's not important enough to affect the Doppler data, given the relatively large error bars of the Doppler data. But, it is a limiting error source for GPS and VLBI (see below). The effect of water vapor on n gives position errors that can vary widely, from 5 to 30 cm. Because vapor density is highly variable spatially, it is hard to model. One procedure is to make surface atmospheric observations and to use them in an empirical formula. A related, and very effective alternative, is to solve for both the dry air and water vapor effects using the ranging data, right along with station positions and clock errors. This is possible because as the satellite moves across the sky, the total amount of atmosphere along the range path varies as the cosecant of the satellite's elevation angle. The amplitude of the cosecant term is proportional to the average density along the path, and can be included in the set of solution parameters.

People also occasionally use water vapor radiometers, which are instruments that point at the sky and detect microwaves emitted by water vapor. The instruments measure microwave amplitudes at two frequencies. The results are fit to an empirically-determined microwave emission curve for water vapor, and this allows the total vapor content along the path to be determined.

The goal of all these methods, at least for GPS and VLBI, is to model the effects of water vapor to the sub-cm level, even during moderately stormy weather.

One of the most important advantages that space geodetic techniques have over geodimeters, is that people don't have to worry as much about the dry air using the space techniques as they do when reducing geodimeter data; at least not over long baselines. For example, if you measure a 1000 km baseline with geodimeters, you must range through 2x1000 = 2000 km of atmosphere — and all near the ground where the atmosphere is densest. But with satellites (either radio or laser ranging) you're looking upwards through only a few tens of kms of atmosphere at each point. And much of that atmosphere is thin. So the total effects of atmospheric refraction are greatly reduced using the space techniques. On the other hand, the atmosphere can make satellite techniques less useful at very short baselines, since those measurements always involve ranging through a few tens of km of atmosphere, no matter how short the baseline is. Still, there are techniques designed to measure the two ends of a baseline simultaneously (such as GPS and VLBI, described below). In those cases atmospheric errors at the two endpoints tend to cancel, thereby reducing their effects.

2.5.2.1.1.2 GPS (Global Positioning System) The Department of Defense's primary navigation system consists of 24 radio-transmitting GPS satellites. The first satellite was launched in 1978. Although the system was designed solely for navigation purposes, geodesists and geophysicists have found that it can also be used for high-precision geodetic positioning. It is probably fair to characterize the scientific GPS effort as one of the most visible programs of any sort in geophysics.

The GPS satellites are at about 20,000 km altitude, and have about 12-hour orbits. The orbits are chosen so that between six and twleve satellites are visible at any time from any point on the earth's surface. Because the altitude is so high, the satellites experience little atmospheric drag. On the other hand, this high orbit means that the GPS satellites can not provide any information about the earth's gravity field that is not given much better from the lower-orbiting laser ranging satellites described below.

Here is a description of how the system operates for navigational applications: Each satellite emits a radio signal — the carrier. The signal is modulated by inserting abrupt 180° phase shifts. The phase shifts are inserted at irregular times, according to a secret code. If you record the signal and have a code book you can figure out the times at which certain bits of the signal left the satellite. You have an accurate clock at your receiver. By comparing the transmission and arrival times, and correcting for ionospheric effects on n, you can deduce the distance to the satellite. The transmitted signal also includes information about the satellite orbit (ie. about the Keplerian elements: semi-major axis, eccentricity, etc.), so that you know the location of the source of the transmission. From simultaneous results to four satellites, you can then solve for the three components of your position and for the offset of your clock with respect to the satellite clocks (the satellite clocks are controlled from a central command point on earth, and are kept synchronous). In this way you can instantaneously determine your position to a few meters. There's a non-secret "civilian" code that allows you to determine positions to about 15 m, though the Department of Defense now routinely degrades the satellite transmissions (using what they call "selective availability") so that civilian accuracies are usually closer to 100 m.

Because GPS is a radio technique, it is sensitive to the sort of atmospheric problems described above for the Doppler satellites. The ionospheric effects are removed by transmitting at two frequencies from each satellite: 1.2 and 1.6 GHz, which correspond to wavelengths of 25 cm and 18.75 cm. The dry air and water vapor effects are not important at the level of accuracy needed for navigation.

There are two general modifications that geodesists make to this procedure in order to use the GPS satellites to determine positions at the 1 cm level. First, they make phase measurements at two or more points simultaneously, and difference the results. In navigational GPS, the measurements are made at a single point. Second, geodesists spend considerable time processing the data after the measurements are taken. Unlike DOD users, geodesists do not need the answer in real time.

As it happens, the code is not a useful part of the signal to geophysicists. The frequencies of the code are near 10 MHz, which corresponds to modulation wavelengths near 30 m. To do 1 cm geodesy, you'd have to determine the phase accurately to

$$\frac{1 \text{ cm}}{3 \times 10^3 \text{ cm}} \approx 3 \times 10^{-4},$$

which would not be easy.

Instead, what geodesists want to do is to remove the code and just work with the carrier. Since the carrier wavelength is only about 20 cm, it provides more accurate position measurements. There are several commercial receivers now being used to do geodesy. They are all portable, ranging from about 20–100 lbs. Different receivers use different techniques for removing the code. One method, for example, involves squaring the signal. The 180° phase shifts, which just reverse the sign of the carrier, then disappear. Alternatively, some instruments require the code which is removed directly from the recorded signal.

The general strategy is to receive a satellite's signal at each end of a baseline. The code is removed from both signals, and the phase at some specified time is recorded at each end (the receiver must include an accurate clock). The two phase results are brought to some central processing site and compared. The phase difference, together with information about n in the atmosphere, gives the difference in distances between the satellite and the two endpoints. See Figure 2.18.

There is still the 2π ambiguity problem. For example, if the observed phase shift is δ, then the distance D is $D = l\lambda + \delta\frac{\lambda}{2\pi}$, where l is an unknown integer. To find l, you observe over an hour or two. The satellite moves around during this time in a known way (you know the orbit), and with enough data you can remove the 2π problem and find D. This is equivalent to changing the wavelength slightly in geodimeters. Here, instead, you keep the wavelength the same but change the distance in a known way. With four

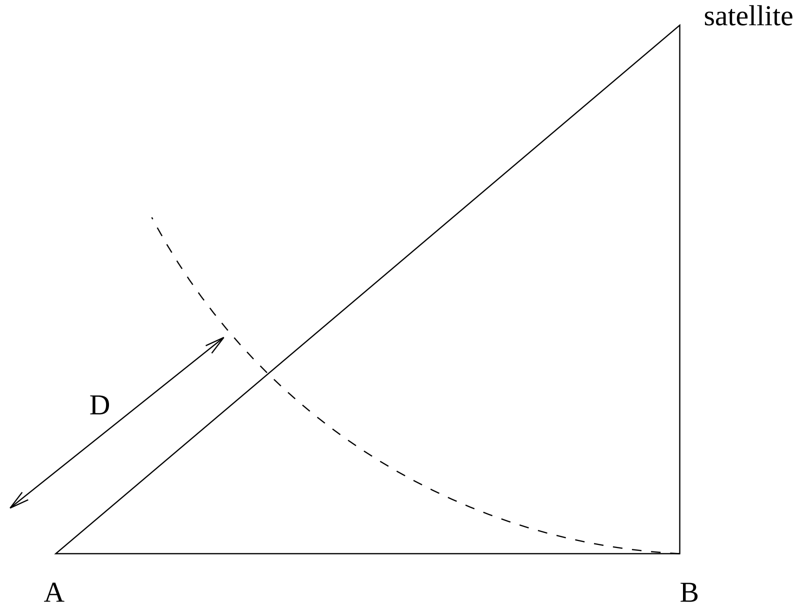

Figure 2.18:

satellites (though you don't need this many if you wait a sufficiently long time) you can find all three components of the baseline, including its length.

Orbit errors are a potential problem. The Department of Defense determines orbits only to about 20 m accuracy, using a set of fixed ground stations. This is not accurate enough for geophysical applications. So geodetic users have pooled resources and established centers for determining and distributing more accurate satellite orbits. These centers now routinely determine the orbits to 5-10 cm accuracy, using tracking from about 200 permanent receivers scattered around the globe. One of the main reasons that these orbits are so much better than the Department of Defense's orbits, is that the DOD orbits are predicted orbits: they must be predicted beforehand so that they can be transmitted down to the user. On the other hand, the geodetic orbits use tracking data from stations around the globe to fit the orbital parameters. Those orbits are made available to the geodetic community well after the signals have been recorded.

A 10 cm orbit error causes a baseline error that is smaller than 10 cm. Specifically, the errors in the baseline components corresponding to a δ meter orbit error are approximately

$$\text{baseline error} \approx \frac{\delta \text{ meter}}{2 \times 10^7 \text{ meter}} \times L, \tag{2.5.7}$$

where 2×10^7 meter is the satellite altitude, L is the baseline length, and δ is the orbit error in meters.

To see an example of this, suppose the satellite is directly above the midpoint of the baseline. See Figure 2.19. Then you would measure no phase difference between the

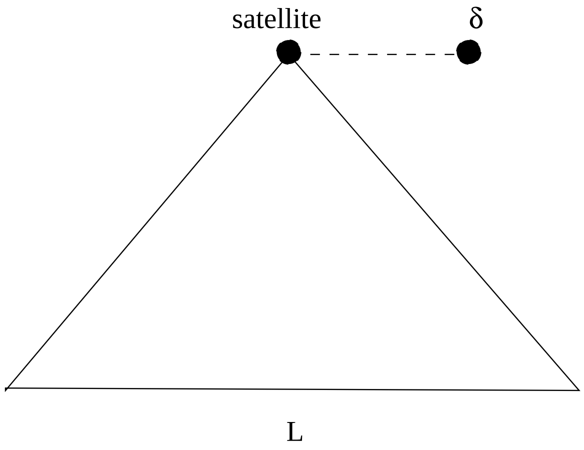

Figure 2.19:

signals received at the two endpoints. Instead, though, suppose you *think* the satellite is at a position a horizontal distance δ from where it really is (so δ is the orbit error). Since you measure no phase difference, you conclude that the baseline is rotated through the angle θ from its actual orientation. See Figure 2.20. Roughly, $\theta \approx \delta/H$, where H is the satellite altitude. And so the error in the vertical component is $L\theta \approx L\delta/H$, which is the same as Equation 2.5.7.

So a 10 cm error in the orbit would give baseline accuracies of 5×10^{-9}, which is not bad for short and medium length baselines. It turns out that you can sometimes do even better than this by using your GPS data to estimate your own orbital corrections to the geodetic orbits, at the same time that you estimate station positions.

Another important error source is the water vapor density along the path. This affects mostly the vertical components of baselines. The effects are reduced by either

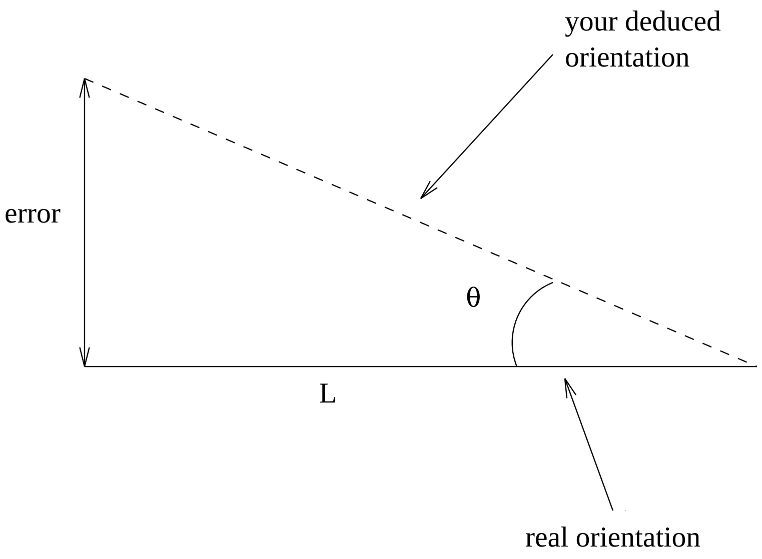

Figure 2.20:

(1) using ground-based meteorological measurements; (2) using radiometers; (3) fitting the atmospheric effects using the GPS data themselves.

And there is an additional problem, called "multi-pathing", that is caused by reflections of the signal from nearby electrical conductors. If they exist, these reflections cause multiple arrivals at the receiver which can complicate the problem of determining the phase and resolving the 2π ambiguity.

Present accuracies seem to depend on baseline length, though not necessarily in a linear way. Baselines of less than about 1000 km have accuracies of on the order 1 mm in the horizontal and maybe 3 mm in the vertical, for averaging time of a week. For baselines greater than about 1000 km the orbit error dominates, and weekly accuracies of a few mm in the horizontal and maybe 10 mm in the vertical are typical for sites with good monumentation.

GPS has had a large impact on conventional surveying as well — even where they don't need these high accuracies. One problem with conventional methods (leveling, geodimeters) is that they are line-of-sight. To range from A to B, you've got to see from A to B. Often, trees and hills, etc., get in the way. Then the geodesist must build towers that stick up above the projections, and range between the towers. With GPS that's not necessary since you've only got to see the sky. The initial equipment investment is high for GPS (tens of thousands of dollars for one receiver), but the operating costs are only about 5% of conventional operating costs.

Another application of GPS is to put the receivers in other satellites to monitor the positions of those satellites. In many cases this may give accuracies significantly better than can be obtained with other methods. Probably its main advantage is that it allows the satellite to be tracked continuously, so that the orbit determination process requires far less dynamic modeling than is required for other forms of tracking. For other tracking techniques (for example, with laser tracking) the satellite can be out of view much of the time, and so the resulting orbit solutions are much more dependent on the dynamic force models.

There are also applications that involve using the GPS refraction estimates. For example, the ionospheric corrections to GPS measurements are of interest, since if there are enough of them they can be used to construct maps of electron density in the ionosphere. This information can be useful for helping to monitor the effects of solar activity on the ionosphere. And, the GPS maps and have been used on occasion to remove ionospheric effects from radio tracking observations of other satellites.

In a similar vein, there was a recent proposed GPS program named GPS/MET that would involve lots of small satellites, each with an on-board GPS receiver. Each satellite would receive continuous GPS signals from the transmitting GPS satellites. By monitoring the phase arrival times as the satellites move (especially valuable would be phase measurements taken when the transmitting GPS satellite is near occultation), and correcting for the effects of orbital motion and of the ionosphere, it would be possible to produce global maps of atmospheric refraction: a linear combination of the dry air and water vapor densities. Thus, these measurements would provide some of the information currently provided by radiosondes (balloon-borne instrument packages), but with vastly improved global coverage and at a much lower cost. A trial satellite was launched in 1995, and performed well. GPS/MET is now defunct as a formal program, and has been replaced by a joint US-Taiwanese project called COSMIC. COSMIC will consist of

8 satellites, and has been approved for a 2001 launch. Two years of funding are now in-place. The satellites are being designed to have much longer lifetimes than that, with the expectation that more years of funding will eventually be secured.

2.5.2.1.1.3 DORIS This is a satellite tracking system developed and used by the French, that relies on Doppler radio measurements. DORIS stands for Doppler orbitography and radiopositioning integrated by satellite. DORIS was developed as a means of providing 10-cm-level radial orbit determination for low-earth satellites. At the moment, DORIS is being used to track the SPOT-series of satellites (French remote-sensing satellites), and the TOPEX/Poseidon radar altimeter satellite. The TOPEX/Poseidon satellite, which will be discussed further below, is a joint US/French mission which also employs both laser and GPS satellite tracking methods.

There are about 50 ground stations in the DORIS network, spread uniformly enough over the earth's surface that a satellite is almost always being tracked by at least one station no matter where that satellite is in its orbit. By contrast, with satellite laser ranging (described below) the station network is sparse enough that a satellite spends much of its time out of range of a ground-based laser. (Though laser ranging makes up for this disadvantage by delivering range measurements of higher precision than those obtainable with DORIS.)

Each DORIS station broadcasts a radio signal at a stable frequency. The signal is received and recorded by the satellite, and the receiving frequency is measured by comparing against the output of an on-board ultrastable radio oscillator. The transmitted and received frequencies differ because of the Doppler shift, and the observed difference is used to solve for the rate of change of the satellite-to-station distance. The orbit solution process must include parameters to represent the possible drift of the satellite oscillator frequency relative to the oscillator frequencies on the ground. The Doppler measurements are actually made at 2 frequencies (the ground stations broadcast at both 2036 MHz and 401 MHz) so that the effects of the ionosphere can be removed.

The radio beacons on the ground broadcast uniformly in all directions, and so do not need to be aimed at the satellite. All that is required to add another satellite to the DORIS system, is to include a DORIS receiver on-board the new satellite. This is an advantage over laser ranging systems, which must be pointed at the satellite being tracked. The addition of a new laser ranging satellite can result in decreased tracking of other satellites, depending on how fully-utilized the laser tracking network is.

2.5.2.1.2 Laser ranging satellites The most useful laser ranging satellites have been the NASA satellite LAGEOS, launched in 1976, and a nearly identical satellite, LAGEOS II, launched in 1993 as a joint project between NASA and Italy. Other NASA LAGEOS-type satellites are being proposed. And, there have been similar satellites launched over the past few years by other countries, including Japan and the Soviet Union.

2.5.2.1.2.1 LAGEOS (LAser GEOdetic Satellite) LAGEOS is a small (60 cm diameter), heavy (about 400 kg) sphere covered with reflecting corner cubes. A laser on the ground sends a pulse of light up to the satellite. The light is reflected and returns to the laser, where it is detected. The round trip travel time is measured, which gives the range between the satellite and the laser. (Using a continuous laser and measuring phase differences would require too much power.) The orbit altitude is about 6000 km. That's high enough to get the satellite out of most of the atmosphere, but low enough to keep it within reach of moderately powerful lasers. LAGEOS is more sensitive to lateral variations in the earth's gravity field than are the GPS satellites (20,000 km elevation), and that's both a disadvantage and an advantage. It's a disadvantage because it means that uncertainties in the gravity field can make it more difficult to model the orbit accurately. But it's an advantage in that it means that LAGEOS is able to provide useful information about the earth's gravity field, in addition to its role in point positioning.

LAGEOS is currently tracked by about 40 fixed stations around the globe. Results from these stations are used to give the orbit, which then provides information about the earth's gravity field. The results are also routinely combined to determine variations in the earth's rotation, and to give the tectonic motion at the stations.

At one time there were also about half a dozen mobile layers. The smallest of these were housed in small trucks and fit easily into a plane. They could be taken to a point to determine the position, and then brought back again later to see if the point has moved. It sometimes took several weeks at a spot to fix its position, and ranging had to be done in clear weather. Note that this technique gives the position of a single point: it does not require simultaneous observations from two ends of a baseline.

The laser systems vary in accuracy. The fixed stations have more powerful lasers, and so are more accurate. The best lasers are able to range with precisions of better than 1 cm for a single pulse, and of about 1–3 mm for a normal point (an average of the individual pulses over a few minutes). Not all the fixed lasers are this good. Some of them are probably at least an order-of-magnitude worse.

There are two sources of systematic errors that make the final positioning accuracies worse than might be inferred from the laser precisions alone. One is atmospheric refraction. Refraction problems are less severe at optical frequencies than at radio frequencies. The ionosphere has virtually no effect, and the effects of the dry and wet air components are smaller. Still, uncertainties due to the dry air density can often be important at almost the 1 cm level. One way of reducing that error in the future might be to lase at two optical frequencies, and then to use the known dispersive properties of dry air at optical frequencies to solve for the index of refraction. This, of course, is the approach used in multi-color geodimeters.

The other important systematic error comes from uncertainties in modeling the satellite orbit: uncertainties both in the gravity field and in the non-gravitational forcing. At present, radial orbit errors are probably 1–2 cm, which probably translates to about a 1 cm error in position measurements. Note the difference from GPS, where a 1–2 cm orbit error would likely translate into a much smaller position error. That's because GPS observes from both ends of a baseline simultaneously, whereas LAGEOS does not. Thus, the orbit errors tend to cancel in GPS.

Another laser mission in NASA's future plans, involves putting a laser on a satellite and using it to range to the earth's surface. In its originally proposed form, this mission was called GLRS (Geoscience Laser Ranging System), was intended to measure distances to retro-reflectors scattered over the earth's surface. This mission has now been de-scoped to something called GLAS (Geoscience Laser Altimeter System), and is scheduled for a 2001 launch with a lifetime of 3-5 years. GLAS will mainly use the on-board laser to range to polar ice, with the goal of monitoring temporal changes in ice sheet elevations. GLAS, at an altitude of 600 km, should be able to determine the change in ice sheet elevation to an accuracy of about 13 cm over a 70 meter spot size. By averaging together the results at many spots in a region, accuracies of 15 mm over areas of 100 km x 100 km should be obtainable. GLAS will also be able to determine the heights of cloud tops.

2.5.2.1.3 Lunar Laser Ranging (LLR) This technique is similar to satellite laser ranging. You measure the round trip travel time of laser pulses reflected off of corner cubes left on the moon's surface by Apollo astronauts and by unmanned Soviet spacecraft. LLR was initiated in the 1960's, so it's an older experiment than LAGEOS. There are three U.S. and two Soviet reflectors on the moon. Most of the existing lunar ranging data come from two laser stations, one in Texas and one in France. The Texas station has been operative, in one form or another, since the inception of the experiment some 25-30 years ago.

You need much more powerful lasers to range to the moon than to range to LAGEOS. Depending on the pulse duration, between 10^{14} and 10^{20} photons leave the laser and head towards the moon. But, on average, less than 1 photon is in the returning beam. At present, the earth–moon distance can be determined to 1 cm when averaged over a few minutes.

These are the objectives of LLR, as originally stated in the 1960's:

1. determine the lunar orbit

2. determine the positions of the retroflectors — to give control for lunar geodesy and mapping

3. determine the "librations" of the moon (rigid body rotations of the moon about its

principal axes)

4. determine the positions of the lasers

5. determine the rotation of the earth

6. test general relativity

The orbital motion has been determined extremely well — maybe three orders of magnitude improvement over earlier results. This has helped people learn about tidal energy dissipation on earth (the relevance of one to the other will be discussed later in the course). Also, the orbit has been used to look for the Nordvedt effect. This is a difference between the gravitational and inertial masses of the moon, due to gravitational self-energy of the moon. It is predicted by some gravitational theories, but not by relativity. LLR has shown the effect is 0 to within an observational error.

The observed lunar librations give the lunar moments of inertia and some other gravity harmonics. People have noticed large phase shifts between the librational motion and external torques. These results have been tentatively interpreted as possible evidence of energy dissipation within a fluid lunar core.

LLR gives information on variations in the earth's rotation, particularly on the rate of rotation, and has proven useful for constraining the longest period terms of the earth's nutational motion. In general, though, rotational motion can now be obtained more accurately using other techniques. LLR has not given much useful tectonic information, because there have been so few stations lasing to the moon.

2.5.2.1.4 VLBI (very-long-baseline-interferometry) VLBI is similar in spirit to GPS, but for VLBI the radio signal comes from a quasar instead of from a satellite.

VLBI was developed to do radio astronomy. You have two radio antennas a long distance apart (up to several thousand km). You observe the same radio source from each end, simultaneously, and compare results. This sort of emulates one enormous antenna, and so is useful for studying the very long wavelength radio waves emitted by quasars. The results are used to investigate the positions and structures of quasars (believed to be extremely distant galaxies in formation).

People realized early on that you could also use VLBI to do geodesy. If you measure the time delay between the signals, you can solve for a component ('d' in Figure 2.21) of the baseline between the two points.

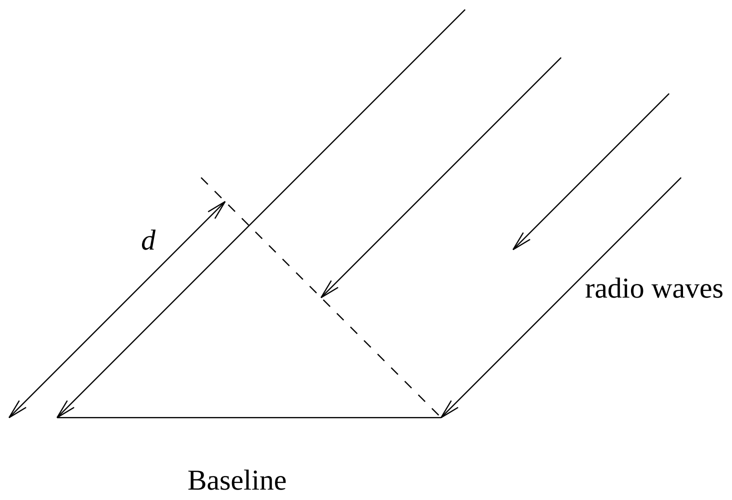

Figure 2.21:

If you measure time delays for four different quasars, you can solve for all three components of the baseline and the clock difference between the two sites. Usually, people look at about twelve sources per day.

Here is how VLBI works as a geophysical experiment: You receive the radio signal, which has content across a wide band of frequencies. You filter the signal to get a reasonably pure frequency, and record the signal on tape. You then take the tape to a central site and compare it with a similar tape from the other antenna. Both tapes have accurate time marks on them. You use the time marks to deduce the phase shift between the two signals, by cross-correlating the two tapes.

The wavelength of the signal is about 3 cm. So the 2π ambiguity is about 3 cm. With geodimeters, you reduce the problem by changing the wavelength slightly and ranging again. You do the same sort of thing with VLBI. That is, you keep reasonably broad sidebands when you filter. So on your tape you've really got a band of frequencies, and you cross-correlate the whole band. This reduces the 2π problem to about 30 m. Often you already know the baseline to 30 m. If not, looking at twelve sources instead of just four can resolve the problem.

Incidentally, why not choose an even a wider band of frequencies, to reduce the 2π problem further? One reason is that this would require too much data storage and handling. Also, without extra effort, it's harder to get accurate phase delays when cross-correlating signals with lots of frequency content.

2.5.2.1.4.1 Problems

ionospheric effects on n

As with GPS and Doppler, you remove the ionospheric effects by recording data at two radio frequencies, and using the known dispersion to pull out the ion content.

dry air

As in GPS and Doppler: radio waves are non-dispersive to the dry air effect on n. Typically, the effect on baselines is about 2 m. It can be corrected to about 2 mm (0.1%) by using surface data and by estimating the effect directly from the data as the quasars move across the sky.

water vapor

This is probably the limiting error source for VLBI. Again, as in GPS, the uncorrected effect is typically 10 cm, but can be as large as 30 cm. The signal is non-dispersive to the effects of water vapor. You can use surface measurements or radiometers *or* estimate the effect directly from the VLBI results.

2.5.2.1.4.2 Accuracy For the best determined VLBI baselines, the horizontal baseline components can probably be determined to 1–3 mm and the vertical components to maybe 7–12 mm, or so. Some (most?) baselines can not do this well. The vertical is worse than the horizontal, because the vertical is affected more by tropospheric (i.e. water vapor) errors.

2.5.2.2 Advantages of one space positioning technique over another

The three space positioning techniques are VLBI, SLR (satellite laser ranging), and GPS. (LLR doesn't provide positions of enough points to make it competitive.)

1. VLBI results are not affected by satellite orbit errors, but SLR and GPS are. (Conversely, VLBI can't give the gravity field, whereas SLR can.)

2. VLBI and GPS are all-weather. SLR needs clear skies.

3. GPS fits in a jeep easily; in fact, it can be carried around by hand. Mobile SLR instruments fit in a small truck. Mobile VLBI requires a tractor trailer.

4. VLBI and GPS can do short baselines better than SLR. (VLBI and GPS may not be as good as the very best geodimeters over very short baselines.) That's because VLBI and GPS are differencing techniques, so that the errors tend to cancel over short baselines. SLR is usually used as a single-point technique, so that the errors don't cancel as effectively.

5. VLBI and SLR do long baselines better than GPS, at least at the moment. Orbit problems limit GPS at long baselines, though the orbits are continually improving.

6. SLR can probably determine vertical positions better than GPS and VLBI, because of its relative insensitivity to tropospheric errors.

2.5.2.3 Some Results

It is hard to determine the accuracies of these various techniques. Inter-comparisons between different techniques are valuable, but are not often done. (When inter-comparisons have been done in the past, people have often found discrepancies between techniques of 2–3 cm, which is larger than would be inferred from the usually accepted accuracy estimates for each technique individually.) Instead, the usual approach is to look at repeatability: you measure the same baseline over and over again, and compare the results. The resulting scatter gives an estimate for the accuracy. This estimate, though, is only a lower bound for the total error, since there is usually no guarantee that you don't have systematic errors that are constant over the entire observing period.

People are now obtaining results from all of these techniques (GPS, SLR, VLBI) that agree well with geological models of plate motion on a global scale. The rates of motion obtained by fitting linear trends to baseline measurements over several years, are usually in line with the geologic rates predicted by paleomagnetism. It suggests that plate motion is probably reasonably steady in time: that the plates move about the same over a few years as they do over millions of years.

Not surprisingly, geodetic rates do not always agree with geological rates near plate boundaries. Plate boundaries are often areas with substantial local and regional deformation, due to the stresses caused by the plates moving past one another. California, where the Pacific and North American plates are sliding past one another, is perhaps the most intensely studied of these boundary regions. California is crossed by numerous leveling and geodimeter lines, and is extensively monitored with GPS, VLBI, and SLR instruments. To get an idea of the amount of regional deformation, note that the NUVEL-1 geological plate motion model predicts about 46 mm/yr relative motion between the Pacific and North American plates. Of this amount, about 12 mm/yr occurs across the San Andreas fault. There are numerous faults associated with the plate boundary, however, which are often collectively referred to as the San Andreas Fault System. The total slip across this entire system is closer to 35 mm/yr, which is still less than the total plate motion rate. Most of the difference comes from the spreading of the Basin and Range region (i.e. Nevada and portions of surrounding states), which has been detected, using space geodetic techniques, to be between 5 and 7 mm/yr. Most of the remaining discrepancy is apparently due to motion on faults outside of the San Andreas Fault System, as it is usually defined. When the geodetic rates over this entire region of western North America are added up, the disagreement with the NUVEL-1 plate motion model is only at the 1–2 mm level, which is probably insignificant.

2.5.2.4 Altimeters

This is another sort of satellite ranging technique. It's not really a point positioning technique, but it has geophysical, geodetic, and, most of all, oceanographic applications.

The goal is to map the shape of the sea surface as a function of time. It works like this: A radar pulse is emitted by the satellite and reflected off the ocean surface, from which it returns to the satellite. The round trip travel time of the pulse is measured, and so the distance between the satellite and the ocean surface can be determined. The satellite is also tracked from the ground, using either lasers or radio signals (or both), to obtain the satellite orbit in a geocentric coordinate system. By combining these results you can determine the geocentric locations of the sub-satellite points on the ocean surface (the points from which the radar pulses are reflected), and so determine the shape of the sea surface. The area of the sea surface sampled by the radar pulse is about 20 km^2. So, the effects of waves and other local disturbances tend to average out. The cancellation is not perfect, and some effects of sea surface roughness do affect the data. But even these can be largely removed, because the roughness also affects the shape of the returning pulse and that change in shape can be measured. In fact, one of the indirect goals of altimeters is to relate the observed roughness to local wind speed.

Altimeters can also range to land and ice, in principle. But land and ice topography can vary considerably over 20 km^2, and so the topographic effects tend not to average out effectively in the returning pulse. In fact, altimetry is not a generally useful technique for determining the shape of the land surface. It has, however, proven useful in mapping large ice sheets.

Why do you want to know the shape of the sea surface? One reason is that the sea surface should be nearly a surface of constant gravitational potential. Thus, the shape of the surface provides information about the gravity field over the oceans. The short wavelength features (tens to a few hundreds of km) in the gravity field are due, mostly, to the mass of the underlying sea floor topography. So, the data let you deduce accurate maps of ocean bathymetry. The longer wavelength features can tell you about the earth's interior, though the longest wavelength features (half-wavelengths longer than about 300 km) can also be determined using other means (i.e. from SLR).

By determining the sea surface, you can also learn about ocean currents. In fact, oceanographic applications are the primary motivation for modern satellite altimeter missions. The sea surface will not exactly coincide with a surface of constant potential (that surface is called the "geoid"), because of vertical displacements associated with horizontal currents. What happens is that the Coriolis force due to the currents is balanced by pressure forces caused by the departure of the surface from the geoid. The excess height above or below the geoid can be as large as a meter or so, though it is usually much less than that. So if you know the height above the equipotential surface, you can then learn about the currents. To learn about the time-averaged mean currents, you must know the geoid by some independent means. To learn about time dependent variations in currents, however, you don't need the gravity field. You only need to assume that the geoid doesn't change with time.

Outside of early testing, the useful altimeter satellites have been SEASAT, GEOSAT, ERS-1 and ERS-2, and TOPEX/Poseidon. All of these satellites have had to deal with atmospheric propagation errors, just like any radio technique. The ionosphere is removed by looking at two radar pulses which have different frequencies. The dry air is removed by using surface meteorological observations (usually in the form of gridded, global models generated by meteorological centers). The wet air is sometimes removed by using those same meteorological data. Better, though, is to use downward-pointing water vapor radiometers.

SEASAT:

SEASAT was launched in 1978 by NASA. It was tracked from the ground using lasers. It gave excellent results for the geoid, and then short-circuited after about three months. The precision of the on-board altimeter was about 5 cm. The main limitation on the accuracy was uncertainties in the radial component of the satellite orbit. The orbit accuracies achieved at the time were on the order of 1 m for global-scale wavelengths, which maps directly into a 1 m error in sea surface. Accuracies were much better (about 10 cm) for time-dependent, short wavelength variability.

The earth's gravity field is known much better today than it was during the time of SEASAT. These new gravity field models have recently been used to re-fit the SEASAT orbits, to obtain global-scale accuracies of about 20 cm.

GEOSAT:

GEOSAT was launched in 1985 by the U.S. Navy. Data from the first 18 months were classified up until 1995, when they were released to the general public. During that initial phase of the mission, the satellite had very dense ground tracks (spacing on the order of 10 km), so that bathymetry could be estimated in considerable detail. After that, the satellite orbit was changed, so that the ground tracks were less dense (several tens to about 100 km spacing), but so the satellite would repeat the same track every 17 days. The data from this part of the mission were released to the public as they were obtained. GEOSAT was turned off in early 1990. Orbit accuracies at the time were on the order of 50 cm at long wavelengths, which translates to about the same accuracies in the sea surface. The accuracies were much better at shorter wavelengths. Recent re-analysis of the data using improved gravity field models has resulted in orbits that are accurate to 10 cm at even the longest wavelengths.

TOPEX/Poseidon:

TOPEX/Poseidon was launched in 1992. It is a joint NASA/French mission. It is still functioning normally, and is expected to continue to return data for at least several more years. Its orbit repeats every 10 days, with spacing between the ground tracks of about 300 km at the equator. TOPEX/Poseidon is tracked with lasers and with DORIS; and there is also an on-board GPS receiver. Before launch, the sea surface height measurements were expected to be accurate to about 14 cm, with the limitation coming primarily from uncertainties in the orbit. But because of impressive and unanticipated improvements in knowledge of the gravity field, people have found that the orbital errors are probably closer to 3–4 cm, and that the total TOPEX/Poseidon sea surface error budget from all sources and for all wavelengths, is probably about 5 cm.

ERS:

ERS-1 and ERS-2 were launched in 1991 and 1995 by the European Space Agency. These are almost identical platforms carrying various remote sensing instruments, including an altimeter. The primary method of determining platform position is a two-way microwave ranging system called PRARE, which is supposed to provide orbits accurate to the 10 cm level. There have been some hardware problems with PRARE, and so laser ranging, which was supposed to be the secondary tracking method, has largely been providing the orbits. I don't know how accurate the sea surface height measurements are from these satellites, though I do know that the orbit solutions are less accurate than those for TOPEX/Poseidon.

2.5.2.5 GRACE: A Dedicated Gravity Mission

The best models of the earth's global gravity field come from satellite laser ranging. Tracking of LAGEOS, in particular, has provided most of the high-quality, long wavelength data used in recent gravity solutions. In the most complete models, the SLR data are supplemented with satellite altimetry data to provide short wavelength information over the oceans, and with surface gravity data to improve the short wavelengths over the continents.

The best gravity field models based solely on satellite orbit solutions are spherical harmonic expansions complete through degree and order of 70 (corresponding to half-wavelengths of about 300 km at the earth's surface). The highest resolution models intended for surface gravity studies are complete to degree and order 360, which corresponds to a half-wavelength of about 50 km. These latter models combine satellite gravity measurements with radar altimeter and ground-based gravimeter measurements.

The accuracies of these models are somewhat uncertain. Formal error estimates suggest the satellite solutions for the geoid height, up through degree and order 70, are probably accurate to about 20 cm. The shorter-scale errors in the high-resolution models are likely to be larger, particularly over the land where complications caused by incorporating hundreds of different ground-based gravity profiles into the solution can be considerable.

There has long been interest within the geophysical community in improving the gravity field down to very short wavelengths through the use of a dedicated satellite gravity mission. Results from such a mission would be useful to oceanographers as well. To determine the time-averaged mean currents from satellite altimetry data, an oceanographer needs to know the geoid. Given the estimated accuracies of existing geoid models, it is not possible to obtain reliable results for the time averaged mean currents at half-wavelengths below about 1300 km.

Dedicated gravity missions have been proposed several times over the past few decades. A requirement for any such mission is that the satellite be placed in a low altitude orbit, typically a few hundred km above the earth's surface. LAGEOS, and other existing laser ranging satellites, can't give good results at shorter wavelengths because their altitudes are too high: short wavelength anomalies die out quickly with altitude.

There have been two competing ideas over the past few years for the design of a dedicated gravity mission. One is to use a single satellite with an on-board gravity gradiometer to measure spatial derivatives of gravity. In the most recent version of this idea, the gradiometer would consist of 6 closely-spaced proof masses. The motion of those masses relative to one another would be monitored. An acceleration of one mass relative to another would be a reflection of a non-zero gravity gradient along the direction between the two masses. The entire instrument would be only a few tens of cm on a side, and would be superconducting to reduce Brownian motion of the proof masses which might otherwise be mis-interpreted as gravity signals.

In the other generic design, the mission would consist of two satellites that range to each other and that are tracked from the earth. As the satellites fly over a gravity high, for example, the lead satellite is accelerated ahead thereby increasing the range. As the trailing satellite passes over the high, it is accelerated so as to catch up with the lead satellite, thus decreasing the range back to where it was originally. Thus, the range measurements reflect the gradient of gravity in the satellite-to-satellite direction. The mission can be thought of as a one-component gradiometer, where the "proof masses" are the satellites and are on the order of hundreds of km apart rather than a few cm.

A mission using this latter dual-satellite approach has now been approved by NASA for a 2001 launch. The mission, named GRACE (Gravity Recovery and Climate Experiment), is being administered by the University of Texas and JPL and is planned to have a nominal 5-year lifetime. The two satellites will be launched by the German Space Agency, using a single Russian COSMOS rocket. The satellites will be in identical, or nearly-identical orbits, probably at 400-450 km altitude and 100-400 km apart. The satellite-to-satellite distance will be monitored by determining the travel times of microwaves beamed between the two satellites. Those travel times will be determined by measuring the phase difference between the transmitted and received signals. Each satellite will include an on-board GPS receiver to monitor its position relative to the earth and to provide clock synchronization.

Each satellite will also include an accelerometer to monitor the non-gravitational accelerations of the satellite so that their effects can be removed from the range measurements. The accelerometer measures its acceleration relative to the spacecraft. A gravitational acceleration affects both the spacecraft and the accelerometer, and so is not detected in the accelerometer output. Atmospheric drag, however, affects only the spacecraft and so its effects on the spacecraft can be detected using the accelerometer output. The accelerometer is vital to the success of the mission, since the spacecraft will be flying so low.

The GRACE results will lead to gravity field improvement of several orders of magnitude at all wavelengths, and would determine gravity accurately down to spatial scales as short as 100 km. Furthermore, GRACE will deliver a map of the gravity field at this resolution every 2-3 weeks. This will allow GRACE to determine time-variations in the gravity field at these temporal and spatial scales, and so will permit estimates of mass redistribution both within the earth and on or above its surface. There are a large number of dynamic processes that cut across almost all the geophysical sciences, that involve a large enough mass redistribution to be seen with GRACE. Time-variable GRACE measurements should have implications for hydrology, oceanography, meteorology, glaciology, and the physics of the solid earth.

2.5.2.6 SAR Interferometry

As a geophysical technique, SAR (Synthetic Aperture Radar) is mostly used as a remote sensing tool. A radar image of the earth's surface provides a map of surface scattering at radar wavelengths, which yields information about a number of physical surface properties, including surface roughness, surface topography, and near-surface water content. But obtaining fine-scale spatial resolution with radar is difficult. The resolution of an image is an inversely proportional to the size of the antenna that transmits and receives the radar beam. In SAR, the radar antenna is placed either in a plane or a satellite, and the motion of the antenna is used to simulate the effects of a very large antenna.

Suppose, for example, the antenna is in a satellite that is traveling from north to south. A radar pulse is beamed, say, in the eastward direction at an angle that could be typically in the range of 20-50 degrees. Resolution perpendicular to the direction of motion (ie. in the E-W direction) is obtained by noting that back-scattered waves from regions closest to the satellite track will arrive back at the antenna prior to those waves scattered from regions further away. Along-track (ie. N-W) resolution is obtained by noting that the frequency of a back-scattered wave will be Doppler shifted by an amount that depends on how far north or south the scattering region lies relative to the satellite. For example, a region south of the satellite is moving away from the antenna, so it produces a back-scattered wave that is shifted downward in frequency as observed at the antenna.

SAR images of the same region on the surface, taken from different satellite locations and at different times, t_1 and t_2, can be combined to map any surface displacements that occurred during the time interval $[t_1, t_2]$. The technique is commonly called SAR interferometry. It works as follows: In a SAR scene from a satellite, the scattering region is divided into a very large network of pixels, usually a few meters to a few tens of meters on a side. The SAR data consist of an amplitude and a phase (ie. the phase shift between the transmitted and received radar waves) for the scattered wave for every one of those pixels. In SAR interferometry, the phase measurements taken from one satellite location are differenced from those taken from another location. It requires considerable processing to line up the pixel positions in the two images so that they correspond to one another. The two satellite locations must be within 100 m or less for this to work. The two measurements are taken on different orbital passes, separated typically by days or months or longer. The time separation is determined by the time scale of the geophysical process being observed.

Differences in phase can be due to: (1) changes in the position of the earth's surface along the direction to the satellite; (2) the fact that back-scattering from surface topography will occur off of slightly different portions of the topographic feature (ie. that the topography is seen in stereo from the two satellite locations); (3) changes in the location of the satellite; and (4) changes in the dielectric constant of the surface (due to such things as changes in snow cover, vegetation, etc.). Effect (3) is reduced as far as possible using knowledge of the satellite position, and is then reduced further by adjusting the satellite's orbital parameters so that the phase differences are minimized at the corners of the imaged region. Effect (2) must be reduced either by taking an image from a third nearby satellite location, or else using a second pair of closely-spaced images. Effect (4) can only be overcome by using two images that are spaced closely enough in time, or in the time of year, that the surface properties are about the same at those two times. Effect (1) is what the SAR analysis is trying to extract, and is obtained once effects (2)-(4) are minimized.

The first satellite SAR was on SEASAT in 1978, followed by a number of prototype instruments flown on the shuttle. The European Space Agency included a SAR instrument on both ERS-1 (launched in 1991) and ERS-2 (launched in 1995). And Canada launched the RADARSAT-1 SAR in 1995. Satellite SAR interferometry has been used to detect various sorts of signals, including co-seismic displacement fields, ground deformation caused by volcanic activity, and the motion of ice streams. The ultimate precision obtainable with this technique is believed to be at the mm level.

Chapter 3

Potential Theory

This chapter describes the mathematical theory of gravity. It pretty much reduces to a study of Poisson's equation. Much of the material may look familiar, because it is similar to what you see in a course in electrostatics. Later in this course, we will apply these results to learn about the earth.

3.1 Introductory remarks

The problem Newton posed was: given a density distribution, can you find the gravitational field? He solved that problem by formulating Newton's Law of Gravity. Another type of problem that often comes up when dealing with the earth's gravity field, is: given a volume $\mathcal{V}$ bounded by a surface $\mathcal{S}$, and given some information about gravity on $\mathcal{S}$, can you find gravity inside $\mathcal{V}$? In this case, $\mathcal{V}$ may or may not contain a mass distribution.

Both these types of problems are best solved using the gravitational potential, rather than $\overline{g}$. The potential is a scalar, and scalars are easier to work with than vectors. ($\overline{g}$ is the gradient of the potential.) The boundary value problem is then: given information about the potential on $\mathcal{S}$, what is the potential in $\mathcal{V}$? The solution to this problem exists and is unique, as long as the right sort of boundary conditions are specified on $\mathcal{S}$. The 'right' boundary conditions are that the potential is specified everywhere, *or* that the normal gradient of the potential is specified everywhere, *or* that one is specified over part of $\mathcal{S}$ and the other over the rest of $\mathcal{S}$. But they can't both be specified at the same places — that's too much information and, in general, a solution won't exist. (Though for a real situation, of course, the potential on $\mathcal{S}$ and the normal gradient of the potential on $\mathcal{S}$ *will* both be consistent with a solution in $\mathcal{V}$.)

Consider the earth, where $\mathcal{S}$ is the earth's surface and $\mathcal{V}$ could be the volume either outside or inside the earth. What sort of gravity information is available on $\mathcal{S}$? Well, people measure the gravitational acceleration on $\mathcal{S}$. And there are also leveling data, which give the direction of the gravity vector with respect to the surface normal. Thus you know the magnitude and direction of $\overline{g}$, which means you know all three components of $\overline{g}$ on the surface. In terms of the potential, this means you know both the potential and the normal component of the gradient of the potential on the surface $\mathcal{S}$. Only, you don't know $\mathcal{S}$.

Finding $\mathcal{S}$ from these boundary conditions is one of the fundamental mathematical problems in physical geodesy. It's called "Molodensky's problem." What you're after is the shape of the earth. There is no known closed-form mathematical solution to this problem. It is usually solved iteratively, as we will discuss later. In practice, the problem is further complicated by observational errors and incomplete data. We will not give a rigorous treatment of Molodensky's problem here. Instead, we will concentrate on the more familiar and more tractable problems:

1. given the density distribution, find the potential;

2. given information about the potential over a known surface, find the potential in the volume bounded by that surface.

First, let's deal with 1.

3.2 Finding the gravitational field from knowledge of the density

For a point mass M at $\overline{x}'$, the gravitational acceleration at $\overline{x}$ is:

$$\overline{g}(\overline{x}) = -\frac{GM(\overline{x} - \overline{x}')}{|\overline{x} - \overline{x}'|^3} \tag{3.2.1}$$

where $G = 6.672 \times 10^{-8}$ cm^3/gm s^2. If a mass m is placed at $\overline{x}$, then the gravitational force on m is $\overline{F} = m\overline{g}(\overline{x})$. For a continuous mass density $\rho(\overline{x}')$, you replace M in Equation 3.2.1 with $\rho(\overline{x}')\, d^3\overline{x}'$, and sum (i.e. integrate) over $\overline{x}'$ to get:

$$\overline{g}(\overline{x}) = \int_{\substack{\text{all}\\ \text{space}}} \frac{G\rho(\overline{x}')(\overline{x}' - \overline{x})}{|\overline{x} - \overline{x}'|^3}\, d^3\overline{x}'.$$

You can define the gravitational potential scalar $V(\overline{x})$ as

$$\overline{\nabla} V(\overline{x}) = \overline{g}(\overline{x}). \tag{3.2.2}$$

V is the negative potential energy per unit mass. Note the word *negative* — that's the usual geophysical convention for V, and it means that there are sign differences between the results here and many of the corresponding results you may have seen in electrostatics. You can show, without much trouble, that

$$V(\overline{x}) = \int \frac{G\rho(\overline{x}')}{|\overline{x} - \overline{x}'|}\, d^3\overline{x}'. \tag{3.2.3}$$

Note that because of Equation 3.2.2, you can add a spatial constant to Equation 3.2.3 without affecting the physical significance of $V(\overline{x})$.

Thus, given $\rho(\overline{x}')$ everywhere, you can find $V(\overline{x})$ and so you can infer $\overline{g}(\overline{x})$. So if you are given an anomalous density distribution in the earth, you can estimate its effects on gravity. As you might expect, problems like this arise often in geophysics.

3.2.1 $V(\overline{x})$ for a uniform sphere

Let's find $V(\overline{x})$ for certain simple density distributions. First, consider a uniform sphere. In this case $\rho = \text{constant} = \rho_0$ for $r' < R$ ($R =$ the radius of the sphere), where r' is the radial spherical coordinate using the center of the sphere as the origin. So:

$$V(\overline{x}) = G\rho_0 \int_0^R (r')^2\, dr' \int_0^\pi \sin\theta'\, d\theta' \int_0^{2\pi} \frac{d\phi'}{|\overline{x} - \overline{x}'|}. \tag{3.2.4}$$

We define our coordinate system so that $\overline{x}$ is on the $\hat{z}$ axis, a distance d from the origin. See Figure 3.1. Then,

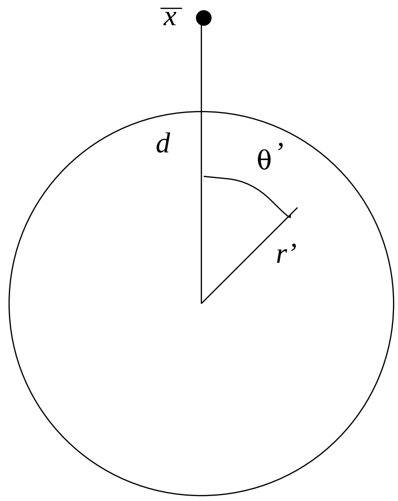

Figure 3.1:

$$|\overline{x} - \overline{x}'| = \left[d^2 + (r')^2 - 2r'd\cos\theta'\right]^{1/2}$$

which is independent of ϕ'. So the $d\phi'$ integral in Equation 3.2.4 gives 2π, and thus

$$V(\overline{x}) = 2\pi G\rho_0 \int_0^R (r')^2\, dr' \int_0^\pi \frac{\sin\theta' d\theta'}{[d^2 + (r')^2 - 2r'd\cos\theta']^{1/2}}.$$

The $d\theta'$ integral is relatively easy because $\sin\theta' d\theta' = -d(\cos\theta')$. That integral works out to be $2/d$. So:

$$V(\overline{x}) = \frac{4\pi G\rho_0}{d} \int_0^R (r')^2\, dr' = \frac{4}{3}\frac{R^3\pi\rho_0 G}{d} = \frac{MG}{d}, \tag{3.2.5}$$

where M is the total mass of the sphere. So, the gravitational field depends on how much mass is in the sphere and on where the sphere is, but it is independent of the size of the sphere. The field is the same whether the sphere has a large radius and a small density, or has a small radius and a large density. In fact, the sphere could have a radius of zero with an infinite density (i.e. a point mass). This illustrates a fundamental limitation for using observed gravity to learn about the earth's interior: different density distributions can give the same gravitational field. So, knowledge of the field can constrain the density, but can not determine it uniquely.

The result in Equation 3.2.5 can be extended to any spherically symmetric density distribution. If $\rho(\overline{x}') = \rho(r')$ where $\rho(r')$ is not necessarily constant, then the potential at a distance d from the origin is

$$V(\overline{x}) = \frac{MG}{d} \tag{3.2.6}$$

where M is the total mass inside the sphere of radius d. Equation 3.2.6 is, in fact, the lowest order approximation to the earth's gravitational field, since the earth is close to being spherically symmetric.

3.2.2 $V(\overline{x})$ for a thin disc

A second useful example is the field due to a thin disc of uniform density, where the field point is on the axis of the disc. See Figure 3.2. In this case, using cylindrical coordinates with the origin at the center of the disc:

$$V(\overline{x}) = G \int_0^R r'\, dr' \int_0^{2\pi} d\phi' \int_0^h \frac{\rho\, dz'}{|\overline{x} - \overline{x}'|}$$

where R and h are the disc radius and thickness.

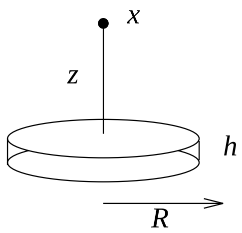

Figure 3.2:

For a thin disc, $1/|\overline{x} - \overline{x}'|$ is approximately independent of z'. Let $\sigma = \int_0^h \rho dz' =$ (for constant ρ) $\rho h =$ constant. σ can be interpreted as a surface mass density. Then

$$V(\overline{x}) \quad = \quad G\sigma \int_0^R r'\, dr' \int_0^{2\pi} d\phi' \frac{1}{|\overline{x} - \overline{x}'|}$$

$$\begin{aligned} &= G\sigma \int_0^R r'\,dr' \int_0^{2\pi} d\phi' \frac{1}{\sqrt{z^2+(r')^2}} \\ &= 2\pi G\sigma \left[\left(z^2+R^2\right)^{1/2} - z \right]. \end{aligned} \tag{3.2.7}$$

To find the gravitational potential due to an infinite plane sheet of mass, it seems logical to use Equation 3.2.7 to find the limit of V as $R \to \infty$. But that limit is ∞. The problem is that V is determined only to within a spatial constant. And we need to choose that constant more carefully before taking the limit. Specifically, we add a constant, $-2\pi G\sigma R$, to Equation 3.2.7 so that $V(z=0)=0$. Then, V becomes:

$$V(\overline{x}) = 2\pi G\sigma \left[(z^2+R^2)^{1/2} - (z+R) \right]. \tag{3.2.8}$$

For an infinite plane, we take the limit of Equation 3.2.8 as $R \to \infty$, keeping leading terms in z. For example, for $z \ll R$: $(z^2+R^2)^{1/2} \approx R(1+\frac{1}{2}\frac{z^2}{R^2})$, so that

$$V \approx \lim_{R\to\infty} 2\pi G\sigma \left[R + \frac{1}{2}\frac{z^2}{R} - z - R \right] = -2\pi G\sigma z.$$

So, for an infinite plane, $\overline{g} = \overline{\nabla} V = -2\pi G\sigma \hat{z}$, which is independent of z. The explanation for this is that for an infinite plane there is no scale length. You look down at the plane, and no matter how far away you are, it looks the same to you.

3.2.3 $V(\overline{x})$ for a line mass

Finally, let's find the potential due to a thin line mass, where the field point is above the midpoint of the line. See Figure 3.3. Then:

$$V(\overline{x}) = \int_{-R/2}^{R/2} dl' \int_{\substack{\text{cross-}\\\text{sectional}\\\text{area}}} \frac{G\rho\,dA}{|\overline{x}-\overline{x}'|}$$

where dA is a differential element of the line's cross-sectional area. For a line with negligible cross-sectional area, $1/|\overline{x}-\overline{x}'|$ is essentially independent of the angular coordinates that describe the area.

Define the mass/length as

$$\lambda = \int_{\substack{\text{cross-}\\\text{sectional}\\\text{area}}} \rho\,dA.$$

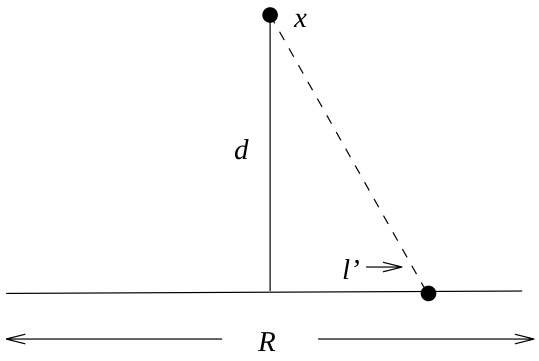

Figure 3.3:

Then

$$\begin{aligned} V(\overline{x}) &= \int_{-R/2}^{R/2} \frac{G\lambda\, dl'}{\sqrt{d^2 + (l')^2}} \\ &= 2G\lambda \ln\left[\frac{R + \sqrt{R^2 + 4d^2}}{2d}\right]. \end{aligned} \tag{3.2.9}$$

For an infinite line, let $R \to \infty$. Then, $V(\overline{x})$ from Equation 3.2.9 $\to \infty$. Again, the problem here is that we need to add a constant to Equation 3.2.9 before we take the limit. We choose the constant to be $-2G\lambda \ln R$, so that Equation 3.2.9 becomes:

$$\begin{aligned} V &= 2G\lambda \ln\left[\frac{R + \sqrt{R^2 + 4d^2}}{2d}\right] - 2G\lambda \ln R \\ &= 2G\lambda \ln\left[\frac{R + \sqrt{R^2 + 4d^2}}{2dR}\right]. \end{aligned}$$

Then, for an infinite line:

$$V_{R\to\infty} = 2G\lambda \ln\left(\frac{2R}{2dR}\right) = -2G\lambda \ln(d).$$

so that

$$\overline{g} = \overline{\nabla} V = -\frac{2G\lambda}{d}.$$

3.2.4 A numerical method for arbitrary mass anomalies

Prospectors and, sometimes, geophysicists have a standard method for solving the forward gravity problem (i.e. for finding $\overline{g}$ from an assumed density distribution, or from an hypothesized mass inclusion of arbitrary shape). They divide the underlying mass into many thin discs. They then approximate the contribution of each disc to $|\overline{g}|$ in the following way. See Figure 3.4. The contribution from the disc to the vertical component

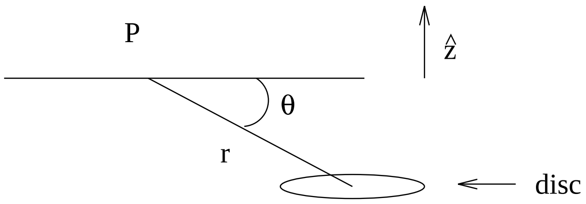

Figure 3.4:

of $\overline{g}$ at P (only the vertical component will affect the amplitude of $\overline{g}$ to first order) is

$$\Delta g = -G \int_{\text{disc}} \frac{\rho \sin\theta}{r^2}\, d\mathcal{V}$$

where $\rho =$ disc density, and $\sin\theta$ is included to give the vertical component of $\overline{g}$. For a thin enough disc, θ and r^2 do not depend on the z coordinate of the point within the disc (that is, the entire disc has about the same z coordinate). If $\int \rho\, dz \equiv \sigma$ is the mass/area of the disc, then:

$$\Delta g \approx -G\sigma \int_{\text{disc}} \frac{\sin\theta}{r^2}\, dA$$

where $dA =$ element of surface area. Note that $dA \sin\theta / r^2 =$ solid angle subtended at P by the infinitesimal area dA. (($dA \sin\theta$) is the area, dA, projected onto a sphere of radius r centered about P; and that projected area is $r^2\, d\Omega$, where $d\Omega$ is the infinitesimal solid angle.) So:

$$\Delta g = -G\sigma \int_{\text{disc}} d\Omega = -G\sigma\Omega$$

where $\Omega =$ total solid angle subtended by the disc. The prospector computes Ω for each disc, and then adds up all the discs. Discs are used because it's easy to find Ω for them. To include all the mass that is present in the anomaly, you've got to let the discs overlap somehow. All this is done on a computer. The only thing you've got to worry about is: are the discs thin enough? Programs that do this sort of thing are available commercially.

Incidentally, to see that this gives the right answer for an infinite plane, note that in that case $\Omega = 2\pi$ (= half of the spherical result $\Omega = 4\pi$). So $\Delta g = -G\sigma 2\pi$, which is the infinite plane result we obtained earlier.

3.3 Poisson's equation

3.3.1 Derivation

There is another way to write Newton's Law of Gravity. You can represent it as a differential equation (Poisson's equation) instead of as an integral. Here, we will derive the differential equation from the integral.

Consider the Laplacian operator, ∇^2, defined as

$$\nabla^2 V(\overline{x}) \equiv {\partial_x}^2 V + {\partial_y}^2 V + {\partial_z}^2 V.$$

From Newton's integral for V, (Equation 3.2.3):

$$\nabla^2 V = G \int \rho(\overline{x}') \nabla^2 \left(\frac{1}{|\overline{x} - \overline{x}'|} \right) d^3\overline{x}'$$

where the Laplacian is with respect to $\overline{x}$. But

$$\nabla_x^2 \frac{1}{|\overline{x} - \overline{x}'|} = \nabla_{x'}^2 \frac{1}{|\overline{x} - \overline{x}'|}$$

where $\nabla_{x'}^2$ is with respect to $\overline{x}'$. So:

$$\nabla^2 V = G \int \rho(\overline{x}') \nabla_{x'}^2 \left(\frac{1}{|x - x'|} \right) d^3\overline{x}'.$$

Divide the integration volume (all space) into $D + \epsilon$, where $\epsilon =$ a sphere of radius R centered about $\overline{x}$, and D is the rest of space. We'll let $R \to 0$, later. For simplicity, we temporarily define our coordinate system so that $\overline{x} = 0$. Then (dropping the prime's in the integrand)

$$\nabla^2 V(\overline{x} = 0) = G \int_D \rho(\overline{x}) \nabla^2 \left(\frac{1}{|\overline{x}|} \right) d^3\overline{x} + G \int_\epsilon \rho(\overline{x}) \nabla^2 \left(\frac{1}{|\overline{x}|} \right) d^3\overline{x}. \qquad (3.3.1)$$

We will use spherical coordinates to do these integrals, so that

$$\nabla^2 \frac{1}{|\overline{x}|} = \nabla^2 \left(\frac{1}{r} \right).$$

In spherical coordinates, and if $r \neq 0$, the Laplacian operator has the form:

$$\nabla^2 = \frac{1}{r^2}\partial_r(r^2\partial_r) + \frac{1}{r^2\sin^2\theta}\partial_\theta\left(\sin\theta\partial_\theta\right) + \frac{1}{r^2\sin^2\theta}\partial_\phi^2. \tag{3.3.2}$$

Thus,

$$\nabla^2\left(\frac{1}{r}\right) = \frac{1}{r^2}\partial_r(r^2\partial_r\frac{1}{r}) = 0$$

and so the integral over D in Equation 3.3.1 is 0. The integral over ϵ is *not* zero, because Equation 3.3.2 is not valid at $r = 0$ (the center of ϵ).

To reduce the ϵ integral, note that

$$\overline{\nabla}\cdot\left(\rho\overline{\nabla}\frac{1}{r}\right) = \overline{\nabla}\rho\cdot\overline{\nabla}\frac{1}{r} + \rho\nabla^2\frac{1}{r}.$$

So,

$$\nabla^2 V(0) = G\int_\epsilon \overline{\nabla}\cdot\left(\rho\overline{\nabla}\frac{1}{r}\right)\,d^3\overline{x} - G\int_\epsilon \overline{\nabla}\rho\cdot\left(\overline{\nabla}\frac{1}{r}\right)\,d^3\overline{x}. \tag{3.3.3}$$

Consider the second integral on the right hand side of Equation 3.3.3. Note that

$$\overline{\nabla}\left(\frac{1}{r}\right) = -\hat{e}_r\frac{1}{r^2}$$

where $\hat{e}_r$ is the unit vector in the r direction. And, $d^3\overline{x}$ is proportional to r^2. So, if ρ is continuous at r=0 so that $\overline{\nabla}\rho$ is bounded inside ϵ for ϵ small enough, then the integrand is bounded. In that case, we know the integral must be less than the bound of the integrand, multiplied by the volume of ϵ. So, if we take the limit as $R \to 0$ (that is we let $\epsilon \to 0$), this second integral vanishes. Note, though, that this argument does not work if ρ is discontinuous at r=0. This means that Poisson's equation may not be valid at discontinuities. We will have to come back later and deal with this, by deriving continuity equations on V across discontinuities.

To reduce the first integral on the right hand side of Equation 3.3.3, we use the divergence theorem:

$$\int_\epsilon \overline{\nabla}\cdot\overline{T} = \int_\mathcal{S} \hat{n}\cdot\overline{T}$$

where $\mathcal{S}$ is the surface bounding ϵ, $\hat{n}$ is the outward normal to $\mathcal{S}$, and $\overline{T}$ is any arbitrary vector. In our integral ϵ is a sphere, so that $\hat{n} = \hat{r}$. Thus:

$$\nabla^2 V(0) \quad = \quad \lim_{R\to 0}\left[G\int_\mathcal{S} \hat{e}_r\cdot\left(\rho\overline{\nabla}\frac{1}{r}\right)\,d^2\overline{x}\right]$$

$$
\begin{aligned}
&= \lim_{R\to 0}\left[G\int_{\mathcal{S}}\rho\partial_r\left(\frac{1}{r}\right)\Big|_{r=R} R^2\sin\theta\, d\theta\, d\phi\right] \\
&= -G4\pi\lim_{R\to 0}\left[\rho(R)\right] \\
&= -G4\pi\rho(0).
\end{aligned}
$$

If we now change back to our original coordinate system, so that the field point is at $\overline{x}$ instead of at 0, we have the final form of Poisson's equation:

$$\nabla^2 V(\overline{x}) = -4\pi G\rho(\overline{x}).$$

We have shown that Newton's Law of Gravity implies Poisson's equation. You can also go backwards and show that Poisson's equation implies Newton's Law of Gravity, but we won't do that here. That proof involves the use of Green's functions ($GM/|\overline{x}-\overline{x}'|$ is the Green's function for Poisson's equation).

So, we have another way to find the gravity field: solve Poisson's equation. If you are given $\rho(\overline{x})$ everywhere in space, it's probably easiest to find V using Newton's Law of Gravity. But if you don't know ρ everywhere, but you do have information about V or about $\overline{g}$ over some bounding surface, then Poisson's equation is apt to be more useful. Differential equations are better suited for the use of boundary values. There is no obvious place to put them into Newton's Law of Gravity.

3.3.1.1 Gauss' Law

Poisson's equation has an integral form called *Gauss' Law*. Consider any volume $\mathcal{V}$ with surface $\mathcal{S}$ and normal $\hat{n}$. Integrating Poisson's equation over $\mathcal{V}$, and using the divergence theorem, gives:

$$
\begin{aligned}
\int_{\mathcal{V}}\nabla^2 V d\mathcal{V} &= \int_{\mathcal{S}}\hat{n}\cdot\overline{\nabla}V\, d\mathcal{S} \\
&= \int_{\mathcal{S}}\hat{n}\cdot\overline{g}s\mathcal{S} \ \ .
\end{aligned}
$$

We also know that

$$\int_{\mathcal{V}}\nabla^2 V d\mathcal{V} = -4\pi G\int_{\mathcal{V}}\rho\, d\mathcal{V} \tag{3.3.4}$$

from Poisson's equation. So, Gauss' Law is

$$\int_{\mathcal{S}} \hat{n} \cdot \overline{g} d\mathcal{S} = -4\pi G \int_{\mathcal{V}} \rho d\mathcal{V} \tag{3.3.5}$$

In words: the integral over a surface of the normal component of $\overline{g}$ is proportional to the total mass inside the surface.

3.3.1.2 Continuity conditions

The derivation given above for Poisson's equation is not valid at points where the density is discontinuous. So, what do you do if you have discontinuities. For example, suppose you have two regions, labeled as $\mathcal{V}_1$ and $\mathcal{V}_2$ in Figure 3.5, separated by the surface $\mathcal{S}$.

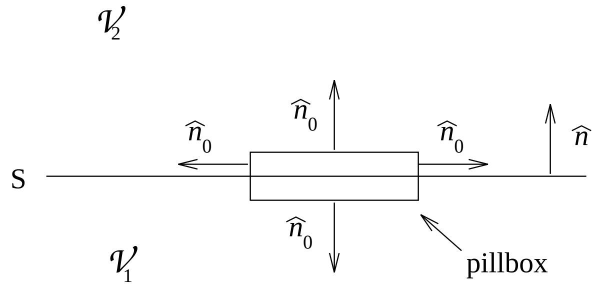

Figure 3.5:

And, suppose the density is discontinuous across $\mathcal{S}$. In fact, let's make it worse than that. Let's suppose there might even be a surface mass density, σ, on $\mathcal{S}$. Surface mass densities do not exist in the real world, but they are often used to approximate thin mass layers to simplify the mathematics. You can think of a surface mass density this way: you start with a mass anomaly distributed across a thin, but non-zero layer. You compress the layer down to zero-thickness, keeping the total mass in the layer the same. You end up with a layer of zero thickness but with non-zero mass, so that the volumetric density, ρ, is infinite in the layer. σ is the density multiplied by the layer thickness (so it has units of mass/area), and so it is a finite quantity. So, in our example, we are assuming not only that ρ is discontinuous across $\mathcal{S}$, but that it might even be infinite on $\mathcal{S}$.

The surface $\mathcal{S}$ causes no particular difficulties if you are using Newton's integral, Equation 3.2.3, to find V. But, how do you incorporate $\mathcal{S}$ into Poisson's equation?

What you do is to solve Poisson's equation in volumes $\mathcal{V}_1$ and $\mathcal{V}_2$ separately, and then match the solutions across $\mathcal{S}$ using continuity conditions. Here we derive those continuity conditions. There are two of them.

The first condition is that $V_1 = V_2$ on $\mathcal{S}$, where V_1 and V_2 are the solutions just inside $\mathcal{V}_1$ and $\mathcal{V}_2$, respectively. This can be seen directly from Newton's integral (Equation 3.2.3), noting that $\overline{x}$ in the integrand is continuous across $\mathcal{S}$.

The second continuity condition is that

$$\hat{n} \cdot \overline{\nabla} V_2 - \hat{n} \cdot \overline{\nabla} V_1 = -4\pi G\sigma, \tag{3.3.6}$$

where $\hat{n}$ is the unit normal to $\mathcal{S}$ pointing from $\mathcal{V}_1$ into $\mathcal{V}_2$. Let's derive Equation 3.3.6 using Gauss' Law applied to the pillbox shown in Figure 3.5. The pillbox is infinitesimally small.

Gauss' Law says that

$$\int_{\substack{\text{pillbox}\\ \text{surface}}} \hat{n}_0 \cdot \overline{\nabla} V \, d\mathcal{S} = -4\pi G \int_{\substack{\text{pillbox}\\ \text{volume}}} \rho d\mathcal{V}$$

where $\hat{n}_0$ is the outward normal to the pillbox. For a very small pillbox, $\overline{\nabla} V$ on the right hand surface of the pillbox is approximately equal to $\overline{\nabla} V$ on the left hand surface. But $\hat{n}_0$ has opposite signs on the two sides. Thus, the contributions to $\int_{\substack{\text{pillbox}\\ \text{surface}}} \hat{n}_0 \cdot \overline{\nabla} V$ from the vertical sides vanish. So, Gauss' Law reduces to

$$\int_{\text{top}} \hat{n}_0 \cdot \overline{\nabla} V \, d\mathcal{S} + \int_{\text{bottom}} \hat{n}_0 \cdot \overline{\nabla} V \, d\mathcal{S} = -4\pi G \int_{\substack{\text{pillbox}\\ \text{volume}}} \rho \, d\mathcal{V}.$$

Or, noting that $\hat{n}_0 = \hat{n}$ in $\mathcal{V}_2$, and $\hat{n}_0 = -\hat{n}$ in $\mathcal{V}_1$, then

$$\int_{\text{top}} \hat{n} \cdot \overline{\nabla} V_2 \, d\mathcal{S} - \int_{\text{bottom}} \hat{n} \cdot \overline{\nabla} V_1 \, d\mathcal{S} = -4\pi G \int_{\substack{\text{pillbox}\\ \text{volume}}} \rho \, d\mathcal{V}.$$

If A is the area of $\mathcal{S}$ inside the pillbox, then A is also the area of the top and of the bottom of the pillbox, and Gauss' Law reduces to

$$A \left[\hat{n} \cdot \overline{\nabla} V_2 - \hat{n} \cdot \overline{\nabla} V_1 \right] = -4\pi G\sigma A.$$

Dividing by A gives the continuity condition, (Equation 3.3.6). By using this continuity condition and the condition that V is continuous across $\mathcal{S}$, you can use Poisson's equation to uniquely determine the gravitational potential in the presence of a discontinuity and/or a surface density.

It is possible to also show that the tangential derivative of V is continuous across any discontinuity, even one with a surface mass. This additional continuity condition follows directly from the continuity of V across $\mathcal{S}$, and it provides no new information about the solution. Nevertheless, I will derive it here.

Consider the two volumes $\mathcal{V}_1$ and $\mathcal{V}_2$, separated by the surface $\mathcal{S}$, as shown below. $\mathcal{S}$ is a surface of discontinuity in density, and $\mathcal{S}$ could even possess a surface density, σ.

$$\begin{array}{llcr} \mathcal{V}_2 & \mathbf{1}\,\cdot & & \cdot\,\mathbf{2} \\ \hline \mathcal{V}_1 & \mathbf{3}\,\cdot & & \cdot\,\mathbf{4} \end{array}\ \mathcal{S}$$

Let the points $\mathbf{1}$ and $\mathbf{2}$ approach one another, and let $\mathbf{3}$ and $\mathbf{4}$ do the same. The tangential derivative in $\mathcal{V}_2$ is then

$$\lim_{\mathbf{1}\to\mathbf{2}} \left[\frac{V(\mathbf{1}) - V(\mathbf{2})}{\mathbf{1} - \mathbf{2}}\right], \tag{3.3.7}$$

and that in $\mathcal{V}_1$ is

$$\lim_{\mathbf{3}\to\mathbf{4}} \left[\frac{V(\mathbf{3}) - V(\mathbf{4})}{\mathbf{3} - \mathbf{4}}\right]. \tag{3.3.8}$$

But before we take the limits above, suppose we let the points $\mathbf{1}$ and $\mathbf{3}$ approach each other, as well as the points $\mathbf{2}$ and $\mathbf{4}$. As $\mathbf{1} \to \mathbf{3}$ and $\mathbf{2} \to \mathbf{4}$, the denominators in Equations 3.3.7 and 3.3.8 converge to the same value. And the numerators do too, because V is continuous across $\mathcal{S}$. So, the tangential derivatives are continuous.

3.3.2 Laplace's equation

If you are in a region of space where there is no mass density, then Poisson's equation reduces to $\nabla^2 V(\overline{x}) = 0$, which is called Laplace's equation. Not surprisingly, Laplace's equation is easier to solve than Poisson's equation. It turns out that you never actually need to solve Poisson's equation. Instead, you can always get by with a combination of Laplace's equation and Newton's Law of Gravity (Equation 3.2.3).

For example, the most general problem you might come up against is: given the volume $\mathcal{V}$ with boundary values specified on the bounding surface $\mathcal{S}$, and given some density distribution inside $\mathcal{V}$, what is the potential inside $\mathcal{V}$? In general, you might start off by trying to solve Poisson's equation. Instead, you can do the problem in two steps, as follows.

First, you ignore the boundary, take ρ as given inside $\mathcal{V}$ and 0 outside $\mathcal{V}$, and use Newton's Law to find V everywhere in space. Let this solution be $V = V_1$. V_1 will probably not satisfy the boundary conditions on $\mathcal{S}$. So, you write the actual solution in $\mathcal{V}$ as $V = V_1 + V_2$, and try to find V_2. Note that $\nabla^2 V = -4\pi G\rho$. And, you have constructed V_1 so that $\nabla^2 V_1 = -4\pi G\rho$ inside $\mathcal{V}$. So, $\nabla^2 V_2 = 0$ inside $\mathcal{V}$, which is Laplace's equation for V_2. You solve this for V_2 using the boundary conditions. Only, the boundary conditions for V_2 are different than for V. Suppose, for example, the boundary conditions are: $V = V_0$ (= some known function) on $\mathcal{S}$. Then, the conditions on V_2 are: $V_2 = V_0 - V_1$ on $\mathcal{S}$.

For the remainder of this chapter, we will only consider Laplace's equation. Suppose the volume $\mathcal{V}$ is bounded by the surface $\mathcal{S}$. Consider the following problem: $\nabla^2 V = 0$, with either V or $\partial_n V$ specified on $\mathcal{S}$. A solution to this problem exists and is unique (to within, possibly, a constant). I'll just ask you to believe the statement that a solution does exist. But, I will show the uniqueness of the solution. To do this, we need:

3.3.2.1 Green's Theorems

These are useful in many geodetic and seismic applications. They are general relations between functions, and do not require that the functions satisfy Laplace's equation.

Start with the divergence theorem:

$$\int_{\mathcal{V}} (\overline{\nabla} \cdot \overline{F})\, d\mathcal{V} = \int_{\mathcal{S}} \hat{n} \cdot \overline{F}\, d\mathcal{S} \tag{3.3.9}$$

where $\hat{n}$ is the outward normal to $\mathcal{S}$. We choose $\overline{F}$ to have the form $\overline{F} = U\overline{\nabla}T$, where U and T are scalars. Then:

$$\overline{\nabla} \cdot \overline{F} = \overline{\nabla}U \cdot \overline{\nabla}T + U\nabla^2 T.$$

And:

$$\hat{n} \cdot \overline{F} = U\hat{n} \cdot \overline{\nabla}T.$$

So, Equation 3.3.9 is:

$$\int_{\mathcal{V}} \overline{\nabla} U \cdot \overline{\nabla} T + \int_{\mathcal{V}} U \nabla^2 T = \int_{\mathcal{S}} U \hat{n} \cdot \overline{\nabla} T. \tag{3.3.10}$$

This is Green's first identity. Now, interchange U and T in Equation 3.3.10 and subtract the results from (Equation 3.3.10). You get:

$$\int_{\mathcal{V}} \left[T\nabla^2 U - U\nabla^2 T\right] = \int_{\mathcal{S}} \left[T\hat{n} \cdot \overline{\nabla} U - U\hat{n} \cdot \overline{\nabla} T\right]$$

This is Green's theorem.

We can use the first identity to prove the uniqueness of solutions to Laplace's equation in the case where either V or $\hat{n} \cdot \nabla V$ is given on $\mathcal{S}$. To do that, we show that if both V_1 and V_2 are solutions, then they can differ by at most a constant.

Suppose V_1 and V_2 satisfy Laplace's equation and satisfy the same boundary conditions: either V or $\hat{n} \cdot \overline{\nabla} V \equiv \partial_n V$ specified on $\mathcal{S}$. Construct $V \equiv V_1 - V_2$. I will show that V is constant everywhere in $\mathcal{V}$, so that $V_1 = V_2 +$ constant.

First, $\nabla^2 V = 0$ (since $\nabla^2 V_1 = \nabla^2 V_2 = 0$), and either $V = 0$ or $\partial_n V = 0$ on $\mathcal{S}$. Then, letting $U = T = V$ in Equation 3.3.10 gives

$$\int_{\mathcal{V}} |\overline{\nabla} V|^2 + \int_{\mathcal{V}} \underbrace{V\nabla^2 V}_{=0} = \int_{\mathcal{S}} \underbrace{V\partial_n V}_{=0}.$$

Or

$$\int_{\mathcal{V}} |\overline{\nabla} V|^2 = 0.$$

Since $|\overline{\nabla} V|^2 \geq 0$ everywhere in $\mathcal{V}$, then $\overline{\nabla} V = 0$ everywhere (otherwise $\int |\overline{\nabla} V|^2 > 0$). So, $V =$ constant in $\mathcal{V}$. Note that if the boundary condition is that V is specified on $\mathcal{S}$, then the constant would have to be 0.

As a corollary, if V is a solution to Laplace's equation and is constant on $\mathcal{S}$, then V is constant throughout $\mathcal{V}$. This is because $V =$ constant is a solution to the problem, and we now know that the solution is unique.

Another property of harmonic functions is that if two harmonic functions, which are defined in all of space, agree everywhere in some volume, then they agree everywhere. The trick to proving this is to consider some closed surface, $\mathcal{S}$, in that volume. The two functions agree on $\mathcal{S}$ and in the volume inside of $\mathcal{S}$, by assumption. Since $\mathcal{S}$ is the only boundary, they must also agree in the volume outside of $\mathcal{S}$ due to the uniqueness theorem. So they agree everywhere.

Incidentally, ∞ is considered a boundary point for volumes that extend to infinity. The usual boundary condition applied at ∞ is that V must go to 0 with increasing r at least as fast as $1/r$, *except* for a possible additive constant. This boundary condition is necessary to ensure uniqueness for external volumes. It is not hard to show that this boundary condition must hold if the mass distribution that causes V is finite. If the field did not go to 0 this quickly, then the left hand side of Equation 3.3.5 (Gauss' Law) would go to ∞ as the surface $\mathcal{S}$ goes to infinity. But, the right hand side of Equation 3.3.5 must remain bounded for finite total mass. So this is impossible.

Another interesting property of solutions to Laplace's equation is that those solutions cannot have a local maximum or minimum within any volume $\mathcal{V}$, except at a boundary point of $\mathcal{V}$. As a consequence, the gravitational potential can't have either a minimum or a maximum in free space.

A crude way to see this is: Suppose $\mathcal{P}$ is a point in $\mathcal{V}$ away from the boundary. Suppose V is a solution to Laplace's equation and has a local maximum at $\mathcal{P}$. The $\mathcal{P}$ is surrounded by points with smaller V. That is, you can enclose $\mathcal{P}$ with a surface, $\mathcal{S}$, where $\hat{n} \cdot \overline{\nabla} V < 0$ everywhere on $\mathcal{S}$, where $\hat{n}$ points away from $\mathcal{P}$. ($\hat{n} \cdot \overline{\nabla} V < 0$ implies that V is decreasing as you move away from $\mathcal{P}$.) But, from Laplace's equation and from the divergence theorem

$$0 = \int_{\mathcal{V}\text{ inside } S} \nabla^2 V = \int_{\mathcal{S}} \hat{n} \cdot \nabla V < 0,$$

since the last integrand is < 0. This is a contradiction. Thus, $\mathcal{P}$ can not be a local maximum (similarly, it can't be a local minimum).

3.3.3 Solutions to Laplace's equation.

How do you solve $\nabla^2 V = 0$, if you are given information about V or about $\partial_n V$ on the boundary? A general method is to find a set of solutions which satisfy $\nabla^2 V = 0$, but which (probably) do not satisfy the boundary conditions. You then add these solutions together to satisfy the boundary conditions. The sum of solutions should still satisfy $\nabla^2 V = 0$, because ∇^2 is linear. The trick is to find a set of solutions which can easily be summed to satisfy the boundary conditions. The solutions you choose will depend on the shape of the boundary. You can only usefully apply this technique to simple boundary shapes. For the earth, the useful boundaries are planes and spheres. For planes, the appropriate solutions are usually trigonometric functions. For spheres, they are spherical harmonics.

3.3.3.1 Planar boundaries

If you are interested in gravity variations that have wavelengths $\ll$ earth's radius, then you can pretend the earth's surface is a plane. In that case, you are often faced with the following sort of problem:

V on the plane $z = 0$ is specified to be $V(x, y, 0) = V_0(x, y)$. V satisfies Laplace's equation in the half-space $z > 0$. Find V for all $z > 0$ (V must also be finite at $z \to \infty$).

The method, as described above, is to find a complete set of solutions to $\nabla^2 V = 0$. Try solutions of the form $V(x, y, z) = X(x)Y(y)Z(z)$. These are called separable solutions in x, y, z. Using $V(x, y, z)$ in $\nabla^2 V = 0$ gives

$$\left(\partial_x^2 X\right) YZ + \left(\partial_y^2 Y\right) XZ + \left(\partial_z^2 Z\right) XY = 0.$$

Or, dividing by XYZ:

$$\frac{\partial_x^2 X}{X} + \frac{\partial_y^2 Y}{Y} + \frac{\partial_z^2 Z}{Z} = 0.$$

$\partial_x^2 X/X$ is a function of x, $\partial_y^2 Y/Y$ of y, and $\partial_z^2 Z/Z$ of z. The only way these three terms can sum to 0 is if all three are constants. That is, if

$$\begin{aligned} \partial_x^2 X &= aX \\ \partial_y^2 Y &= bY \\ \partial_z^2 Z &= cZ, \end{aligned}$$

where a, b, and c are arbitrary constants and $a + b + c = 0$. But, we are not necessarily trying to find all possible solutions to Laplace's equations, or even all separable solutions.

We are just trying to find enough solutions so that we can satisfy the boundary values. It turns out that to do this we only need to consider solutions where a and b are negative real numbers, and where c is a positive real number. So, we choose $a = -k_1^2$, $b = -k_2^2$, and $c = k_3^2$, where k_1, k_2, and k_3 are, for the moment, arbitrary real numbers. Then, the differential equations for X, Y, and Z have the form:

$$\begin{aligned} \partial_x^2 X &= -k_1^2 X \\ \partial_y^2 Y &= -k_2^2 Y \\ \partial_z^2 Z &= k_3^2 Z \end{aligned}$$

where $k_1^2 + k_2^2 - k_3^2 = 0$. We will see later that these are all the solutions we need.

The solutions to these equations are

$$\begin{aligned} X &= Be^{ik_1 x} \\ Y &= Ce^{ik_2 y} \\ Z &= De^{k_3 z} \end{aligned}$$

where B, C, and D are arbitrary constants. (Alternatively, you could write the solutions as cosines and sines, if you prefer those to complex exponential functions.) Note that $X = e^{-ik_1 x}$ is another solution to the differential equation for X, but that it has not been included in the list above. The reason is that $X = e^{-ik_1 x}$ is equivalent to $X = e^{ik_1 x}$, since k_1 can be any real number, either positive or negative. The situation is similar for Y and Z.

The total separable solution is then:

$$V(x, y, z) = Ae^{ik_1 x} e^{ik_2 y} e^{k_3 z} \tag{3.3.11}$$

where A is an arbitrary complex constant, and where $k_3^2 = k_1^2 + k_2^2$. To satisfy the condition that V be finite at $z = \infty$, we require that k_3 be negative. Because of this, a slightly more convenient way to represent the separable solutions is to switch the sign on

k_3, and to require that k_3 be positive. So, our separable solutions have the form

$$V(x, y, z) = Ae^{ik_1x}e^{ik_2y}e^{-k_3z} \tag{3.3.12}$$

where $k_3 = \sqrt{k_1^2 + k_2^2}$, and where k_1 and k_2 can be any real numbers.

A sum of these solutions for different k_1 and k_2 must also be a solution. Since k_1 and k_2 can be any real numbers, then a sum over k_1 and k_2 becomes an integral. So,

$$V(x, y, z) = \int_{-\infty}^{\infty} dk_1 \int_{-\infty}^{\infty} dk_2\, A(k_1, k_2)e^{ik_1x}e^{ik_2y}e^{-k_3z} \tag{3.3.13}$$

satisfies $\nabla^2 V = 0$ for any complex function $A(k_1, k_2)$ for which the integral exists. To verify this, simply take the Laplacian of Equation 3.3.13, and note that the derivatives move through the integral sign and act directly on the product of exponentials. (Note: there is no integral over k_3 in Equation 3.3.13, because k_3 is a function of k_1 and k_2.)

We now have our general solution to Laplace's equation. We need to see if it is general enough. That is, by choosing $A(k_1, k_2)$ appropriately, can we satisfy the boundary values?

The answer is: yes. Letting $z = 0$ in Equation 3.3.13 for V, and setting the result equal to V_0, gives:

$$V_0(x, y) = V(x, y, z = 0) = \int_{-\infty}^{\infty} dk_1 \int_{-\infty}^{\infty} dk_2\, A(k_1, k_2)e^{ik_1x}e^{ik_2y}. \tag{3.3.14}$$

For any well-behaved V_0, we can always find a function $A(k_1, k_2)$ that solves Equation 3.3.14. A is the two-dimensional Fourier transform of V_0, and is given by:

$$A(k_1, k_2) = \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, V_0(x, y)e^{-ik_1x}e^{-ik_2y}. \tag{3.3.15}$$

If A satisfies Equation 3.3.15, then V satisfies Equation 3.3.14. So, V is a solution. In this way you can solve Laplace's equation for any V_0 on $z = 0$. (Again, if you don't like complex exponentials you can use sines and cosines instead.) You do the integrals in Equation 3.3.15 to find $A(k_1, k_2)$. And then you use the result in Equation 3.3.13, and integrate to find V for $z > 0$. This method can be easily modified to solve problems where it is $\partial_z V$ that is given at $z = 0$, rather than V.

A common situation we'll run into is one where $V_0 = V_0(x)$ is independent of y. That occurs when the surface values are dominated by long, linear features. In that case, the results become a little simpler.

One approach to this special case, is to note that if V_0 is independent of y, then the result in Equation 3.3.15 for $A(k_1, k_2)$ gives a Dirac delta function for k_2 when integrated over y. That's the right answer, but I don't want to work with Dirac delta functions in this course.

Instead, to find $V(x, y, z)$ when $V_0 = V_0(x)$, let's go back to finding separable solutions, under the assumption that V doesn't depend on y either. In that case, the separable solutions will have the form

$$V(x, z) = Ae^{ik_1 x}e^{-k_3 z}$$

where $k_3 = |k_1|$, and where k_1 can be any real number. We add these solutions together to get the general solution:

$$V(x, z) = \int_{-\infty}^{\infty} A(k_1)e^{ik_1 x}e^{-|k_1|z}\, dk_1. \tag{3.3.16}$$

This $V(x, z)$ satisfies Laplace's equation. To find $A(k_1)$, V must satisfy

$$V_0(x) = V(x, 0) = \int_{-\infty}^{\infty} A(k_1)e^{ik_1 x}\, dk_1.$$

This can be inverted, for any V_0 (provided V_0 is integrable, etc.), to give

$$A(k_1) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dx\, V_0(x)e^{-ik_1 x}.$$

Note that $k_1 = 2\pi/\lambda$, where λ is the horizontal wavelength. In terms of λ, a separable solution has the form $e^{i2\pi x/\lambda}e^{-z/(\lambda/2\pi)}$. This solution decays with elevation (ie. with z) with a scale length of $\lambda/2\pi$. Thus, terms in the potential with short horizontal wavelength die away more rapidly with elevation than do terms with longer wavelengths.

3.3.3.2 Cylindrical coordinates

Another way to solve plane boundary problems is to use cylindrical coordinates. These are useful if the boundary values on the plane $z = 0$ are circularly symmetric about the origin. The separable solutions in that case include Bessel functions, instead of $e^{ik_1 x}$ or $e^{ik_2 y}$. I won't go into this. In this course, rectangular coordinates will work just fine for all planar problems we consider.

3.3.3.3 Spherical boundaries

In global problems you're interested in finding V outside the earth, given information about V on what is approximately a spherical surface. In this case, a spherical coordinate system (r, θ, ϕ) is appropriate. In spherical coordinates:

$$\nabla^2 V = \frac{1}{r^2}\partial_r(r^2\partial_r V) + \frac{1}{r^2 \sin\theta}\partial_\theta(\sin\theta\,\partial_\theta V) + \frac{1}{r^2\sin^2\theta}\partial_\phi^2 V.$$

We first try to find a complete set of solutions to $\nabla^2 V = 0$ in spherical coordinates, without trying to satisfy the boundary conditions. Let's look for separable solutions of the form:

$$V(r, \theta, \phi) = T(\theta)L(\phi)R(r).$$

Then, T, L and R must satisfy (using $V(r, \theta, \phi)$ in $\nabla^2 V = 0$, and dividing by V):

$$\left[\frac{1}{R}\partial_r\left(r^2\partial_r R\right)\right] = -\left[\frac{1}{T\sin\theta}\partial_\theta(\sin\theta\,\partial_\theta T) + \frac{1}{L\sin^2\theta}\partial_\phi^2 L\right].$$

The $1/\sin^2\theta$ in the $\partial_\phi^2 L$ term prevents a complete separation of these terms into r, θ, and ϕ at this stage, but at least r is separated from θ and ϕ. That is, the left hand side depends only on r, while the right hand side depends only on θ and ϕ. This equation can be satisfied only if the two sides are equal to a constant. Let's write the constant as $l(l+1)$. (l will turn out to be an integer, but so far we don't know that. At the moment, l could be any complex number.) Then:

$$\begin{aligned} \partial_r(r^2\partial_r R) &= Rl(l+1) && (3.3.17)\\ \frac{1}{T\sin\theta}\partial_\theta\left(\sin\theta\,\partial_\theta T\right) + \frac{1}{L\sin^2\theta}\partial_\phi^2 L &= -l(l+1). && (3.3.18)\end{aligned}$$

Equation 3.3.17 is:

$$\partial_r^2 R + \frac{2}{r}\partial_r R - \frac{R}{r^2}l(l+1) = 0.$$

There are two solutions:

$$R = \begin{cases} r^l \\ r^{-(l+1)}. \end{cases}$$

To find T and L, we multiply Equation 3.3.18 by $\sin^2\theta$ to give:

$$\frac{1}{T}\sin\theta\,\partial_\theta\left(\sin\theta\,\partial_\theta T\right) + l(l+1)\sin^2\theta = -\frac{1}{L}\partial_\phi^2 L.$$

The left hand side depends only on θ, and the right hand side depends only on ϕ. So both sides equal a constant. Call the constant m^2. (m will turn out to be an integer, but we don't know that yet, either.) So:

$$\begin{aligned}\partial_\phi^2 L &= -m^2 L & (3.3.19)\\ \sin\theta\,\partial_\theta\left(\sin\theta\,\partial_\theta T\right) &= \left(m^2 - l(l+1)\sin^2\theta\right)T. & (3.3.20)\end{aligned}$$

The solutions to Equation 3.3.19 are

$$L = \begin{cases} e^{im\phi} \\ e^{-im\phi}. \end{cases}$$

Continuity of L requires that $L(\phi = 2\pi)$ must equal $L(\phi = 0)$, and this will occur only if m is an integer. So the solution for L is

$$L = e^{im\phi}$$

where m can be any integer: positive, negative, or 0.

The hard part is now to solve Equation 3.3.20 for T. Define $x \equiv \cos\theta$. Then,

$$\frac{dT}{d\theta} = \frac{dx}{d\theta}\frac{dT}{dx} = -\sin\theta\frac{dT}{dx},$$

and

$$\sin\theta = \sqrt{1-x^2}.$$

Using these results in Equation 3.3.20, gives a differential equation for T in terms of x:

$$\partial_x\left[\left(1-x^2\right)\partial_x T\right] + \left(l\,(l+1) - \frac{m^2}{1-x^2}\right)T = 0. \qquad (3.3.21)$$

This equation can be solved by expanding T using a power series in x:

$$T = \left(1-x^2\right)^{m/2}\sum_{n=0}^{\infty} a_n x^n.$$

We are interested in solutions for T which are finite on $[-1, 1]$, which is why we include no negative exponents in the sum over n. To find the a_n's, we use this power series in Equation 3.3.21 to find recursion relations for the a_n's. We find that the series diverges at $x = \pm 1$ unless l is a non-negative integer, and unless $l \geq |m|$. In fact, what we find is that in that case, the power series truncates to a finite-order polynomial in x: powers of x^n with $n > l - |m|$ have coefficients $a_n = 0$. From this point on we require that l and m be integers, and that $l \geq |m|$.

When $m = 0$, these polynomial solutions are written as $T(x) = P_l(x)$, and are called "Legendre polynomials." These polynomials reduce to the form

$$P_l(x) = \frac{1}{2^l}\frac{1}{l!}\frac{d^l}{dx^l}(x^2 - 1)^l.$$

Note that Equation 3.3.21 determines T only to within a multiplicative constant: the function P_l could be multiplied by any additional constant, and it would still be a solution. The leading factor of $\left(\frac{1}{2^l}\frac{1}{l!}\right)$ in the definition of P_l is simply the conventional normalization factor.

For $m \neq 0$, the solutions are written as $T(x) = P_l^m(x)$, and are called "associated Legendre functions." They have the form

$$P_l^m(x) = (-1)^m (1 - x^2)^{m/2} \frac{d^m}{dx^m} P_l(x) \tag{3.3.22}$$

for $m \geq 0$. What about for $m < 0$? Note that the differential equation for T has an m^2 dependence. So, solutions for negative m should be the same as for positive m. It is not usual to define P_l^m for $m < 0$. But, you could if you wanted — by changing all the m's on the right hand side of Equation 3.3.22 into $|m|$'s. Note, also, that $P_l^0 = P_l$.

The condition that $|m| \leq l$, which is required in order for T to be finite on $[-1, 1]$, is automatically guaranteed by the expressions for P_l^m and P_l. The highest power of x in $P_l(x)$ is x^l. That's because $(x^2 - 1)^l$ has highest power x^{2l}, and after you have taken l derivatives, you are left with x^l. Then, to find P_l^m you take $|m|$ more derivatives. If $|m| > l$, that will give 0.

So, we've found separable solutions. They are of the form:

$$V = \left\{ \begin{array}{c} r^l \\ \text{or} \\ r^{-(l+1)} \end{array} \right\} \times P_l^m(\cos\theta) e^{im\phi}$$

(using $x = \cos\theta$ as the argument of P_l^m). It is common to lump the θ and ϕ dependence together into a single set of functions called spherical harmonics, and defined as:

$$Y_l^m(\theta,\phi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}\,P_l^m(\cos\theta)e^{im\phi} \tag{3.3.23}$$

for $m \geq 0$. For negative m, the spherical harmonics are defined using $Y_l^m \equiv (-1)^m Y_l^{(-m)*}$, where Y_l^{m*} indicates complex conjugation. With the use of these functions, the separable solutions to $\nabla^2 V = 0$ have the form:

$$V = \left\{ \begin{array}{c} r^l \\ \\ r^{-(l+1)} \end{array} \right\} \times Y_l^m(\theta,\phi) \tag{3.3.24}$$

where l (≥ 0) and m ($|m| \leq l$) are integers.

The leading square root factor in the definition of Y_l^m is used so that we have the convenient normalization:

$$\int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta\, d\theta\, [Y_l^m(\theta,\phi)Y_l^{m*}(\theta,\phi)] = 1$$

Note that it doesn't matter how the Y_l^m's are normalized, in the sense that if the Y_l^m's are multiplied by any additional constant, Equation 3.3.24 will still satisfy Laplace's equation. *Sometimes in geophysics other normalizations are used — so be careful.*

3.3.4 Properties of the Y_l^m's and P_l^m's

I'll give, without proof, some of the more useful properties. There are lots of good references for these.

3.3.4.1 Parity

$$P_l^m\left(\cos(\pi-\theta)\right) = P_l^m(-\cos\theta) = (-1)^{l+m}P_l^m(\cos\theta)$$

$$e^{im(\phi+\pi)} = (-1)^m e^{im\phi}.$$

So:

$$Y_l^m(\pi-\theta,\phi+\pi) = (-1)^l Y_l^m(\theta,\phi).$$

This tells you what happens to Y_l^m due to an inversion of the coordinate system through the origin.

3.3.4.2 Recursion Relations

$$\begin{aligned}(2l+1)xP_l^m(x) &= (l+1-m)P_{l+1}^m(x) + (l+m)P_{l-1}^m(x)\\ (1-x^2)\partial_x P_l^m &= -lxP_l^m + (l+m)P_{l-1}^m.\end{aligned}$$

There are lots of other recursion relations, but these two are probably the most useful.

3.3.4.3 Specific Results

I'll only give two, both of which will be useful later in the course.

$$\begin{aligned}Y_0^0 &= \frac{1}{\sqrt{4\pi}}\\ Y_2^0 &= \sqrt{\frac{5}{16\pi}}\,(3\cos^2\theta - 1).\end{aligned}$$

3.3.4.4 Orthogonality

$$\int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta\, d\theta\, Y_l^m(\theta,\phi) Y_{l'}^{m'*}(\theta,\phi) = \delta_{ll'}\,\delta_{mm'}$$

where δ is the Kronecker delta (i.e. $\delta_{ll'} = 1$ if $l = l'$, and $= 0$ otherwise).

3.3.4.5 Completeness

If $V_0(\theta,\phi)$ is any sufficiently smooth, complex function over the unit sphere (that is, on $0 \le \theta \le \pi$ and $0 \le \phi \le 2\pi$), with $V_0(\phi = 0) = V_0(\phi = 2\pi)$, then you can always find complex numbers A_l^m such that

$$V_0(\theta,\phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_l^m Y_l^m(\theta,\phi). \tag{3.3.25}$$

To find the A_l^m, multiply Equation 3.3.25 by $Y_{l'}^{m'*}$ and integrate over the sphere. Using the orthogonality property gives:

$$A_l^m = \int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta\, d\theta\, V_0(\theta,\phi) Y_l^{m*}(\theta,\phi).$$

3.3.4.6 The wavelengths of the Y_l^m's

The values of l and m tell you something about the spatial wavelengths of the Y_l^m. Since $P_l^0 = P_l$ is a polynomial of degree l in $\cos\theta$, then P_l^m is $(1-x^2)^{m/2} = \sin^m(\theta)$ times a polynomial of degree $(l-m)$. (The "$l-m$" is from the $\partial_x^m P_l$ in Equation 3.3.22.) So, $P_l^m(x)$ has $l-m$ zeros between the North and South Poles, plus additional zeros *at* the poles (for $m \neq 0$) due to the $\sin^m(\theta)$ factor.

The ϕ dependence of Y_l^m is $e^{im\phi}$. So, the real and imaginary parts of Y_l^m each have $2m$ zeros for $\phi \in [0, 2\pi)$.

Since a function has two zeros per wavelength, then the north–south wavelength over the earth's surface is roughly

$$\frac{\pi\ 6371}{\frac{1}{2}(l-m)}\ \text{km} \approx \frac{40000}{l-m}\ \text{km}$$

(where 6371 km = earth's radius). And the east–west wavelength at the equator is

$$\frac{2\pi\ 6371}{m}\ \text{km} \approx \frac{40000}{m}\ \text{km}.$$

Another way to think of this is that Y_l^m has a wavelength, independent of direction, that is roughly $40{,}000/l$; and that the value of m tells you about the orientation of the spatial pattern. For example, for $m = 0$ the wavelengths are $\frac{40{,}000}{l}$ km and are oriented north–south, while for $m = l$ they are oriented east–west. (Actually, there are north–south variations even when $l = m$, because of the $sin^m(\theta)$ factor.)

It may be more accurate to describe the spatial scale of a function as equivalent to its half-wavelength. The half-wavelength of a spherical harmonic is $\frac{20{,}000}{l}$ km.

3.3.4.7 The Addition Theorem

Let $\overline{x}(= (r, \theta, \phi))$ and $\overline{x}'(= (r', \theta', \phi'))$ be two vectors separated by the angle γ. Then

$$P_l(\cos\gamma) = \frac{4\pi}{2l+1} \sum_{m=-l}^{l} Y_l^{m*}(\phi', \phi') Y_l^m(\theta, \phi)$$

(where $\cos\gamma = \cos(\theta)\cos(\theta') + \sin(\theta)\sin(\theta')\cos(\phi - \phi')$). This is the addition theorem, which has been given here without proof.

Here's one way to interpret the addition theorem. Consider θ' and ϕ' as fixed, so that the $Y_l^{m*}(\theta', \phi')$ are just numbers. Then, the theorem tells us how to use those numbers to construct a linear combination of the $Y_l^m(\theta, \phi)$ which, for all θ, ϕ, reduces to an $m = 0$ spherical harmonic. Think of the $Y_l^{m*}(\theta', \phi')$'s as rotation coefficients, due to rotating the coordinate system from a coordinate system where the z-axis points towards $\overline{x}'$, to a new system where the z-axis is along $\hat{e}_z$. This rotation changes the ϕ-independent $P_l(\cos\gamma)$ to a sum of ϕ-dependent Y_l^m. But note that it doesn't change the value of l. This theorem further supports the suggestion above, that l characterizes the shape of a Y_l^m while m describes its orientation.

3.3.5 Applications

The most common application is something like the following. We are given V on a spherical surface ($r = R$) about the origin: $V(r = R, \theta, \phi) = V_0(\theta, \phi)$. We assume that $V(r = \infty) = 0$. Find the solution to $\nabla^2 V = 0$ outside the sphere ($r > R$).

To do this, we add together the separable solutions. Keep the $r^{-(l+1)}$ term in Equation 3.3.24, but discard the r^l term since $r^l \to \infty$ as $r \to \infty$. So, try:

$$V(r, \theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_l^m r^{-(l+1)} Y_l^m(\theta, \phi). \tag{3.3.26}$$

To find the complex coefficients A_l^m, use $r = R$ in Equation 3.3.26, and set the result equal to V_0. Then, from the completeness property of the Y_l^m's,

$$A_l^m = R^{l+1} \int \sin\theta\, d\theta\, d\phi\, V_0(\theta, \phi) Y_l^{m*}(\theta, \phi).$$

(If the problem were, instead, to find V *inside* $r = R$, then we'd keep the r^l term in Equation 3.3.24 instead of the $r^{-(l+1)}$ term.)

This approach can be changed slightly to solve problems where you are given $\partial_r V$, instead of V, on $r = R$.

3.3.5.1 Example 1.

Suppose $V_0 =$ constant on the sphere $r = R$. Then,

$$A_l^m = R^{l+1} V_0 \int \sin\theta \, Y_l^{m*} \, d\theta \, d\phi \tag{3.3.27}$$

Note that $1 = \sqrt{4\pi}\, Y_0^0$. So Equation 3.3.27 can be written as:

$$\begin{aligned} A_l^m &= R^{l+1} V_0 \sqrt{4\pi} \int \sin\theta \, (Y_0^0 \, Y_l^{m*}) \, d\theta \, d\phi \\ &= R^{l+1} V_0 \sqrt{4\pi} \, \delta_{l0} \, \delta_{m0}. \end{aligned}$$

where the last equality follows from the orthogonality of the Y_l^m's. So: $A_0^0 = R V_0 \sqrt{4\pi}$, and all other $A_l^m = 0$. Putting these results for A_l^m back into $V(r, \theta, \phi)$ gives:

$$V(r, \theta, \phi) = \frac{A_0^0}{r} \, Y_0^0 = \frac{R V_0 \sqrt{4\pi}}{r} \frac{1}{\sqrt{4\pi}} = \frac{R V_0}{r}$$

which is the result we obtained earlier for a sphere using direct integration.

3.3.5.2 Example 2.

For the real earth, V is not quite constant over the surface, and the earth's surface is not quite a sphere. A better approximation for V on $r = R$, than that described by Example 1, is:

$$V(r = R, \theta, \phi) = V_0 \left[1 - J_2 P_2(\cos\theta)\right]$$

where J_2 and V_0 are constants. (Note that we are still assuming that the surface is a sphere.) What is V in this case for $r > R$? (We'll see, later, why the biggest perturbation to V_0 is a P_2 term. It has to do with the earth's rotation.) From the results above, we conclude that

$$A_l^m = V_0 R^{l+1} \int \sin\theta \, d\theta \, d\phi Y_l^{m*} \left[1 - J_2 P_2(\cos\theta)\right].$$

To write the integrand as products of Y_l^m's, we use:

$$1 = \sqrt{4\pi} Y_0^0 \qquad\qquad P_2 = P_2^0 = \sqrt{\frac{4\pi}{5}} \, Y_2^0.$$

So:

$$
\begin{aligned}
A_l^m &= V_0 R^{l+1} \int \sin\theta \, d\theta \, d\phi \, \sqrt{4\pi} \left[Y_l^{m*} Y_0^0 - J_2 Y_l^{m*} Y_2^0 \frac{1}{\sqrt{5}} \right] \\
&= V_0 \sqrt{4\pi} R^{l+1} \left[\delta_{l0} \, \delta_{m0} - \frac{J_2}{\sqrt{5}} \, \delta_{l2} \, \delta_{m0} \right]
\end{aligned}
$$

where the last equality follows from the orthogonality of the Y_l^m's. So:

$$
\begin{aligned}
A_0^0 &= V_0 R \sqrt{4\pi} \\
A_2^0 &= -V_0 R^3 J_2 \sqrt{\frac{4\pi}{5}}
\end{aligned}
$$

which gives

$$
\begin{aligned}
V(r, \theta, \phi) &= \frac{V_0 R \sqrt{4\pi}}{r} \, Y_0^0 - \frac{V_0 R^3 J_2 \sqrt{\frac{4\pi}{5}}}{r^3} \, Y_2^0 \\
&= \frac{V_0 R}{r} - \frac{V_0 R^3 J_2}{r^3} P_2(\cos\theta) \\
&= V_0 \left[\left(\frac{R}{r} \right) - J_2 \left(\frac{R}{r} \right)^3 P_2(\cos\theta) \right].
\end{aligned}
$$

3.3.5.3 Radial dependence as a function of l

One general comment: Note that if you're outside the earth, an individual Y_l^m term in V will decrease with increasing radius as $r^{-(l+1)}$ (see Equation 3.3.24). So, components with larger l decrease more rapidly. Since large l corresponds to short horizontal wavelengths, this implies that the shorter wavelength signals die away more rapidly as you proceed away from the earth. This is why a satellite must be in a relatively low orbit if it is to be useful for determining short wavelength terms in the earth's gravity field.

3.3.6 The use of Y_l^m's in direct integration

Spherical harmonics can also be useful in direct integrations problems: where you are trying to find V for a given density distribution inside the earth. Note that this is an application which doesn't involve solving a boundary value problem.

Suppose you know $\rho(\overline{x}')$. You can always expand ρ into spherical harmonics, to obtain:

$$\rho(\overline{x}') = \sum_{l',m'} \rho_{l'}^{m'}(r') Y_{l'}^{m'}(\theta', \phi'). \tag{3.3.28}$$

You can do this, because, as stated above, the spherical harmonics are a complete set of functions on every sphere $r' =$ constant. In fact,

$$\rho_{l'}^{m'}(r') = \int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta \, d\theta \, \rho(r', \theta, \phi) Y_{l'}^{m'*}(\theta, \phi). \tag{3.3.29}$$

It is not obvious, but it turns out that you can also expand $1/|\overline{x} - \overline{x}'|$ into spherical harmonics as:

$$\frac{1}{|\overline{x} - \overline{x}'|} = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \frac{1}{2l+1} \frac{r_<^l}{r_>^{l+1}} Y_l^{m*}(\theta', \phi') \, Y_l^m(\theta, \phi)$$

where

$$\left.\begin{array}{rcl} r_< & = & r' \\ r_> & = & r \end{array}\right\} \text{ if } \; r' < r$$

$$\left.\begin{array}{rcl} r_< & = & r \\ r_> & = & r' \end{array}\right\} \text{ if } \; r' > r.$$

3.3.6.1 Finding V outside a surface

These expansions of $\rho(\overline{x}')$ and of $1/\left|\overline{x} - \overline{x}'\right|$, can be used in Equation 3.2.3 to find V outside the earth. In that case $r > r'$ for all r' inside the earth, and so Equation 3.2.3 becomes

$$\begin{aligned} V(\overline{x}) &= G \int \frac{\rho(\overline{x}')}{|\overline{x} - \overline{x}'|} d^3x' \\ &= 4\pi G \sum_{l,m} \sum_{l',m'} \left[\frac{1}{2l+1} \times \right. \\ & \left. \int \frac{(r')^l}{r^{l+1}} \rho_{l'}^{m'}(r') \, Y_{l'}^{m'}(\theta', \phi') \, Y_l^{m*}(\theta', \phi') \, Y_l^m(\theta, \phi) (r')^2 \, dr' \sin\theta' \, d\theta' \, d\phi' \right]. \end{aligned}$$

The integral over $d\theta' d\phi'$ is easy, due to the orthogonality of the Y_l^m's. We get 0 from the integral unless $l = l'$ and $m = m'$. So, the sum over l', m' goes away, and we can replace

$\rho_{l'}^{m'}$ with ρ_l^m, to obtain:

$$V(\overline{x}) = 4\pi G \sum_{l,m} \frac{Y_l^m(\theta,\phi)}{2l+1} \frac{1}{r^{l+1}} \left[\int (r')^{l+2} \rho_l^m(r')\, dr' \right]. \tag{3.3.30}$$

So, to find V, you find the $\rho_l^m(r')$ using Equation 3.3.29, and then you integrate over r' and sum over l and m in Equation 3.3.30.

We will use this result later in these notes. We will also occasionally want to compute V when there is a surface density, σ, concentrated on the spherical surface $r = R$. In that case, the radial integral in Equation 3.3.30, reduces to $R^{l+2}\sigma_l^m$, and so:

$$V = 4\pi G R \sum_{l,m} \frac{\sigma_l^m}{2l+1} Y_l^m(\theta,\phi) \left(\frac{R}{r}\right)^{l+1}. \tag{3.3.31}$$

Note: suppose you know V, and you want to learn about $\rho(r)$. The best you can do is to use Equation 3.3.30. This equation doesn't give you ρ as a function of r; but at least it does give you a weighted average of $\rho_l^m(r)$ for each l and m. Note that the weighting factor, r^{l+2}, is largest near the surface, especially for large l (short horizontal wavelengths).

As an application of Equation 3.3.30, suppose we have a spherically symmetric earth, so that $\rho(\overline{x'}) = \rho(r')$. Then, from Equation 3.3.29,

$$\begin{aligned}
\rho_{l'}^{m'}(r') &= \int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta\, d\theta\, \rho(r') Y_{l'}^{m'*}(\theta,\phi) \\
&= \int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta\, d\theta\, \rho(r') \sqrt{4\pi}\, Y_0^0 Y_{l'}^{m'*}(\theta,\phi) \\
&= \begin{cases} \sqrt{4\pi}\rho(r') & \text{if } l' = m' = 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

So $\rho_0^0(r') = \sqrt{4\pi}\rho(r')$, and all other $\rho_{l'}^{m'}(r') = 0$. Using this result in Equation 3.3.30, gives:

$$V(\overline{x}) = 4\pi G \frac{Y_0^0(\theta,\phi)}{r} \int (r')^2 \rho_0^0(r')\, dr' = \frac{GM}{r}$$

where M is the total mass of the earth, given by

$$M = \int \rho(\overline{x}')\,(r')^2 \sin\theta'\, dr'\, d\theta'\, d\phi' = 4\pi \int (r')^2 \rho(r')\, dr' = \sqrt{4\pi} \int (r')^2 \rho_0^0(r')\, dr'.$$

This result for V agrees with the result that we obtained in Section 3.2.1 without using spherical harmonics.

3.3.6.2 Finding V inside a surface

Here's another result we'll need later. Consider a point, $\overline{x}$, *inside* a hollow object which has a mass density of

$$\rho(r', \theta', \phi') = \sum_{l,m} \rho_l^m(r')\, Y_l^m(\theta', \phi').$$

What is V at $\overline{x}$?

The answer is obtained as in the case where you are outside the sphere, except that here $r_< = r$ and $r_> = r'$. (Assume $r < r'$, for all r' in the massive object.) The result is:

$$V(\overline{x}) = 4\pi G \sum_{l,m} \frac{Y_l^m(\theta, \phi)}{2l+1}\, r^l \int \frac{\rho_l^m(r')}{(r')^{l-1}}\, dr'.$$

As a special case, suppose the hollow object is a sphere, and that all of the mass is described with a surface mass density, σ, on the spherical surface $r = R$. Then

$$V(\overline{x}) = r4\pi G \sum_{l,m} \frac{Y_l^m(\theta, \phi)}{2l+1} \left(\frac{r}{R}\right)^{l-1} \sigma_l^m, \tag{3.3.32}$$

where

$$\sigma = \sum_{l,m} \sigma_l^m Y_l^m(\theta, \phi).$$

Note that if $\overline{x}$ is inside an object that is *not* hollow, so that there is some mass inside of $\overline{x}$ and some outside, then V would be obtained by adding together the solutions for external and internal mass distributions described above.

3.3.6.3 A slightly non-spherical surface

Finally, here is another example that will prove useful later. Suppose we model the earth as having a homogeneous mass density ($\rho = \rho_0 =$ constant), and an outer surface that is almost, but not quite, spherical. Suppose the outer surface consists of the points (r', θ', ϕ') that satisfy:

$$r' = r_0 \left[1 + \sum_{l',m'} \epsilon_{l'}^{m'}\, Y_{l'}^{m'}(\theta', \phi')\right]$$

where

$$\epsilon_{l'}^{m'} \ll 1$$

and where r_0 (= constant) is the mean spherical radius of the surface. What is V inside and outside this earth?

We could do this exactly by integrating over $dr', d\theta', d\phi'$ as described above, after expanding $\rho(\overline{x'})$ in terms of Y_l^m. But there is an easier way to obtain an approximate result, that uses the assumption that $\epsilon_{l'}^{m'} \ll 1$.

If all of the $\epsilon_{l'}^{m'}$'s were zero, our earth would be a sphere and the solution would be easy (described in several places above). Because the $\epsilon_{l'}^{m'}$'s are small, you can think of their effects on V as, to lowest order, the effects of a surface mass. You think of this earth as the sum of a sphere of radius r_0, and a thin shell at $r = r_0$, with thickness $\sum_{l',m'} \epsilon_{l'}^{m'} Y_{l'}^{m'}(\theta', \phi')$.

This thickness can be negative for some (θ', ϕ'), but that's OK. It just means that at those (θ', ϕ') we must remove mass from the r_0 sphere to obtain the shape of the object. (See Figure 3.6.)

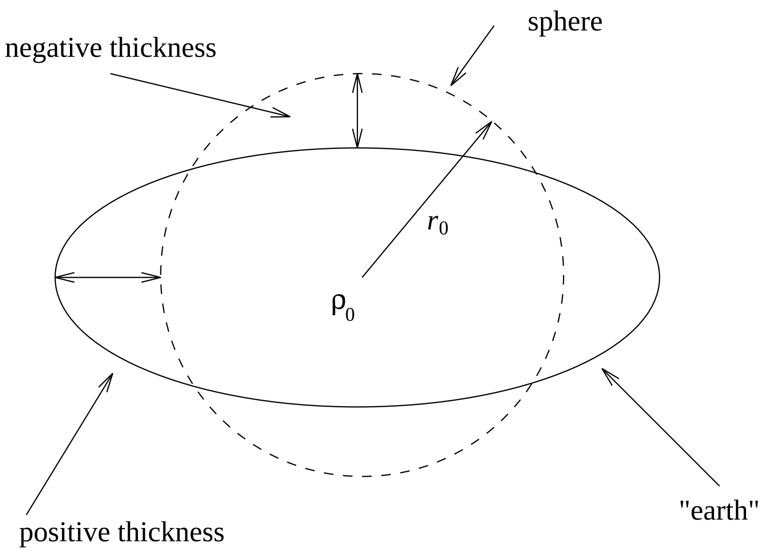

Figure 3.6:

We can easily find V due to the sphere. The difficult part is to find V due to the thin shell. To lowest order, we assume that the mass distribution of the shell can be described with a surface density at $r = r_0$, given by

$$\sigma(\theta', \phi') = r_0 \rho_0 \sum_{l',m'} \epsilon_{l'}^{m'} Y_{l'}^{m'}(\theta', \phi').$$

The expansion of σ into spherical harmonics is:

$$\sigma(\theta', \phi') = \sum_{l',m'} \sigma_{l'}^{m'} Y_{l'}^{m'}(\theta', \phi')$$

where

$$\sigma_{l'}^{m'} = r_0 \rho_0 \epsilon_{l'}^{m'}. \tag{3.3.33}$$

Outside. Suppose $\overline{x}$ is outside our earth, so that $r \geq r'$, for all r' in the earth. Then, V due to the $r = r_0$ sphere is

$$\frac{GM}{r} = \frac{4}{3}\pi r_0^3 \rho_0 \frac{G}{r}.$$

And we can use Equation 3.3.31 to find the effects of the surface mass at a point outside the sphere. Using Equation 3.3.33 for σ_l^m in Equation 3.3.31, and adding the spherical component, gives a total V of

$$V(\overline{x}) = \frac{4}{3}\pi \frac{r_0^3 \rho_0 G}{r} + 4\pi G r_0^2 \rho_0 \sum_{l,m} \left(\frac{\epsilon_l^m}{2l+1} \right) Y_l^m(\theta, \phi) \left(\frac{r_0}{r} \right)^{l+1}.$$

Inside. Suppose $r < r'$, for all r' on the surface of the object.

The potential at r due to the sphere $(r = r_0)$ is

$$V = \frac{2}{3}\pi \rho_0 G \left(3r_0^2 - r^2 \right).$$

I haven't derived this in these notes, but it's easy to do that on your own. The $3r_0^2$ term is a constant, added to V so that V is continuous at $r = r_0$. To find the effects of the thin shell, we use the Equation 3.3.33 for σ_l^m in the result for V from Equation 3.3.32. The solution for the total contribution to V is:

$$V = \frac{2}{3}\pi \rho_0 G (3r_0^2 - r^2) + 4\pi G \rho_0 r^2 \sum_{l,m} \frac{\epsilon_l^m}{2l+1} Y_l^m(\theta, \phi) \left(\frac{r}{r_0} \right)^{l-2}.$$

Chapter 4

Physical Geodesy Problems

4.1 The Figure of the Earth: The Geoid

The geoid is the surface of constant potential that coincides with mean sea level over the oceans. ("Potential ," here, refers to the gravitational plus centrifugal potential.) This is the standard definition of the geoid, but it's a sloppy definition. For one thing, mean sea level is not quite a surface of constant potential, due to dynamic processes within the ocean. You can imagine, though, turning off the dynamic processes so that sea level does become a constant potential surface. (We'll show later that the surface of an equilibrium fluid *is* a constant potential surface.) For another thing, wherever there are continents, the geoid lies beneath the earth's surface. As a result, the actual equal potential surface under continents is warped by the gravitational attraction of the overlying mass. But geodesists define the geoid as though that mass were underneath the geoid instead of above it. In other words, their geoid is not truly an equipotential surface. We'll worry about a more precise definition, later.

Probably the main function of the geoid in physical geodesy is to serve as a reference surface for leveling. To see the connection, suppose the earth's surface was covered with a layer of water. The geoid would then coincide with the surface of that water layer. If you leveled over this surface, your results would show that the entire surface was at the same elevation: the fluid level in the instrument would always indicate "horizontal" as being parallel to the water surface.

So what you really measure when you level, are the elevations above (or below) the geoid. Thus, to find the actual shape of the earth you need to determine the shape of the geoid. And, for that, you need gravity observations.

Let's consider the shape of the geoid and its relation to measured gravity for some simple, but progressively more complicated, earth models.

4.1.1 Spherically Symmetric, Non-rotating Earth

In this case, the surface of the earth is $r = a$, the potential outside the earth is $V = GM/r$ where $M =$ earth's mass, and gravity at the earth's surface is $g = GM/a^2$. The potential V, is constant for constant r. If, for example, we want the geoid to have a mean radius equal to the radius of the earth, then we choose that constant value of r to equal "a," so that $r = a$ is the geoid. So, in this case, the geoid is the outer surface. Note that $g =$ constant on the geoid. This will not be true, in general, for a more realistic earth.

4.1.2 Spherically Symmetric, Rotating Earth

If you take a spherical earth and start it rotating, it will not remain spherical. Centrifugal forces deform the earth into an ellipse. The ellipticity has as big an effect on the geoid as does the rotation. But, let's look at the two effects separately. In this section we assume the earth is rotating but that it remains spherical.

In this case the outer surface is $r = a$. The *gravitational* potential outside is $V = GM/r$. The gravitational acceleration outside is GM/r^2, and is directed radially inwards.

The geoid is a surface of constant gravitational plus centrifugal potential. The centrifugal force per unit mass is

$$
\begin{aligned}
-\overline{\Omega} \times (\overline{\Omega} \times \overline{r}) &= -\overline{\Omega}\,\Omega \cdot r + \overline{r}\Omega^2 && (4.1.1) \\
&= \overline{\nabla}\underbrace{\left(\frac{1}{2}\left[r^2\Omega^2 - \left(\overline{\Omega}\cdot\overline{r}\right)^2\right]\right)}_{\text{centrifugal potential}} && (4.1.2)
\end{aligned}
$$

where $\overline{\Omega}$ = the earth's rotation vector. So, the sum of the gravitational and centrifugal potentials is

$$V_T = \frac{GM}{r} + \frac{1}{2}\left[r^2\Omega^2 - \left(\overline{\Omega}\cdot\overline{r}\right)^2\right].$$

If we define the coordinate system so that $\overline{\Omega}$ is along $\hat{e}_z$, then

$$\overline{\Omega}\cdot\overline{r} = \Omega z = \Omega r\cos\theta.$$

So:

$$V_T = \frac{GM}{r} + \frac{1}{2}\Omega^2 r^2 \sin^2\theta.$$

Using Legendre polynomials:

$$\begin{aligned} \sin^2\theta &= \frac{2}{3} - \frac{2}{3}\underbrace{\left[\frac{1}{2}\left(3\cos^2\theta - 1\right)\right]}_{P_2} \\ &= \frac{2}{3} - \frac{2}{3}P_2(\cos\theta). \end{aligned}$$

So

$$V_T = \left(\frac{GM}{r} + \frac{1}{3}\Omega^2 r^2\right) - \frac{1}{3}\Omega^2 r^2 P_2(\cos\theta). \tag{4.1.3}$$

So rotation modifies the θ-independent term in V_T, and introduces a P_2 term. (We are using P_2, here, instead of Y_2^0 simply to follow convention.)

The rotational terms are small. For example:

$$\frac{\frac{1}{3}\Omega^2 r^2}{GM/r} \approx \frac{1}{870} \text{ (at earth's surface).}$$

$P_2(\cos\theta)$ varies from 1 to $-\frac{1}{2}$ over the surface, so the P_2 term is approximately $(1 - (-\frac{1}{2}))\left(\frac{1}{870}\right) = \frac{1}{580}$ of the gravitational term.

Rotation also affects the observed acceleration. A gravimeter records the sum of the gravitational acceleration and the centrifugal acceleration. That sum is best found by taking the gradient of V_T:

$$\overline{g}_T = \overline{\nabla}V_T \tag{4.1.4}$$

$$\begin{aligned} &= \hat{e}_r\partial_r V_T + \frac{1}{r}\hat{e}_\theta\partial_\theta V_T \\ &= \hat{e}_r\left[-\frac{GM}{r^2} + \frac{2}{3}\Omega^2 r - \frac{2}{3}\Omega^2 r P_2(\cos\theta)\right] + \hat{e}_\theta\left[-\frac{1}{3}\Omega^2 r\partial_\theta P_2(\cos\theta)\right]. \end{aligned} \tag{4.1.5}$$

So, $\overline{\Omega}$ modifies the radial component of $\overline{g}_T$, and introduces a θ-component.

Note that neither the amplitude nor the direction of $\overline{g}_T$ is constant over the surface $r = a$.

What is the shape of the geoid for this example? In other words, V_T in Equation 4.1.3 is a function of r and θ: $V_T = V_T(r, \theta)$. Find r as a function of θ ($r = r(\theta)$), so that $V_T(r(\theta), \theta)$ is independent of θ. The surface described by $r(\theta)$ is then a surface of constant V_T, and so is a candidate for the geoid.

To find $r(\theta)$, we assume that $r(\theta) \approx a$, so that the effects of rotation on the geoid are small. This is probably a pretty good approximation, since the effects of rotation on V_T are less than $1/500$. We write $r(\theta) \approx a + \delta r(\theta)$, where $\delta r / a \ll 1$, and we try to find a first-order approximation for δr.

V_T in Equation 4.1.3 has a GM/r term and two $r^2\Omega^2$ terms. The $r^2\Omega^2$ terms are first order already, so we can use the approximation $r = a$ in $r^2\Omega^2$. The GM/r term is zero order, so we use $r = a + \delta r$ in that term, along with the approximation $(a + \delta r)^{-1} \approx \frac{1}{a}[1 - \frac{\delta r}{a}]$. Then:

$$V_T(r = a + \delta r, \theta) \approx \frac{GM}{a}\left[1 - \frac{\delta r(\theta)}{a}\right] + \frac{1}{3}\Omega^2 a^2 - \frac{1}{3}\Omega^2 a^2 P_2(\cos\theta). \tag{4.1.6}$$

For Equation 4.1.6 to be independent of θ, $\delta r(\theta)$ must satisfy:

$$\left(\frac{GM}{a^2}\right)\delta r(\theta) = -\frac{1}{3}\Omega^2 a^2 P_2(\cos\theta) + \text{constant}.$$

where the constant can be any number, so long as it is independent of θ. Or, since $GM/a^2 \cong g_T$ at $r = a$ ($g_T \equiv |\overline{g}_T|$) to zero order in Ω^2, then:

$$\delta r(\theta) = -\frac{1}{3}\frac{\Omega^2 a^2}{g_T} P_2(\cos\theta) + \text{constant}.$$

The constant in these expressions does nothing interesting. It is not uniquely determined by the requirement that $V_T =$ constant on this surface. It just tells you which of the V_T=constant surfaces you are on. If you want the geoid to be the V_T=constant surface that has its mean radius $= a$, then the constant in the above equations is zero (the average of $P_2(\cos\theta)$ is 0), and the geoid is described by:

$$r = a - \frac{1}{3}\frac{\Omega^2 a^2}{g_T} P_2(\cos\theta). \tag{4.1.7}$$

Equation 4.1.7, then, would be the shape of the mean sea surface. The second term on the right-hand-side of Equation 4.1.7 varies with θ (as θ varies between 0 and 2π) by approximately $(1/580) \times a \approx 11$ km. So, sea level has large departures from a spherical surface. Note that if you leveled over the earth's surface, assumed here to be $r = a$, you would conclude that there were 11 km differences in elevation over that surface.

Incidentally, suppose we write the geoid in the form: $r = a + \delta r(\theta, \phi)$, where δr is the departure of the geoid from a sphere. Write V_T at (r, θ, ϕ) (Equation 4.1.3) as $V_T = V_T^0 + \delta V(r, \theta, \phi)$ where V_T^0 is independent of both θ and ϕ ($V_T^0 = \frac{GM}{r} + \frac{1}{3}\Omega^2 r^2$) and δV ($= -\frac{1}{3}\Omega^2 r^2 P_2(\cos\theta)$ in this example) represents the angular dependence of V_T. Then, note that:

$$\delta r \equiv \frac{\delta V(r = a)}{g_T}.$$

This result will also hold in all later examples, as we increase the complexity of the model earth.

Finally, suppose we could measure $|\overline{g}_T|$ on the geoid. How would the result vary with position on the earth's surface (i.e. with θ)? That is, construct the vector field $\overline{g}^g = \overline{g}_T(r = a + \delta r, \theta, \phi)$ to represent $\overline{g}$ on the geoid. What is g^g ($\equiv |\overline{g}^g|$) as a function of θ?

$$\begin{aligned} g^g &= |g_r^g \hat{e}_r + g_\theta^g \hat{e}_\theta| \\ &\approx |g_r^g| \qquad \text{(to 1st order in small quantities)} \\ &\approx |g_T(r = a) + \delta r \partial_r g_T(r = a)|. \end{aligned}$$

Then from Equation 4.1.5 above, ignoring terms in $\partial_r g_T$ of the order of Ω^2, since those terms are second order after being multiplied by δr, we obtain:

$$\begin{aligned} g^g &\approx \left| -\underbrace{\frac{GM}{a^2}}_{\substack{\text{dominant}\\ \text{term}}} + \frac{2}{3}\Omega^2 a \underbrace{-\frac{2}{3}\Omega^2 a P_2(\cos\theta)}_{\partial_r \delta V} - \underbrace{\left[\frac{1}{3}\frac{\Omega^2 a^2}{(\frac{GM}{a^2})}\right] P_2(\cos\theta)}_{\delta r = \delta V / g_T} \frac{2GM}{a^3} \right| \\ &= \frac{GM}{a^2} - \frac{2}{3}\Omega^2 a + \frac{4}{3}\Omega^2 a P_2(\cos\theta) \end{aligned}$$

$$= \frac{GM}{a^2} - \frac{2}{3}\Omega^2 a + \delta g, \tag{4.1.8}$$

where all the angular dependence in this final result is included in:

$$\delta g \equiv \frac{4}{3}\Omega^2 a P_2(\cos\theta) = -\partial_r \delta V(r=a) - \frac{2\delta V(r=a)}{a} \tag{4.1.9}$$

This result, Equation 4.1.9, relating the angular dependent component of g^g to the angular dependent component of V_T, is another result that will hold for the more complicated earth models described below.

Given that ($\delta g = \frac{-2\delta V}{a} - \partial_r \delta V$) and ($\delta r = \delta V/g_T$) will hold for an arbitrarily complicated earth model (we have yet to demonstrate that), we can anticipate now how geodesists determine the earth's shape.

First, they level the earth. This tells them how the earth's surface differs from the geoid. To determine the geoid, they measure gravity over the surface. They then use those results to deduce gravity on the geoid. They do that using the leveling data. Those data give them the radial distance between the surface and the geoid. Call that distance H. Note that $H/a \ll 1$. If $g_s = g$ as measured on the surface, then g^g ($=$ the value of g on the geoid) is approximately:

$$\begin{aligned}
g^g &\approx g_s - H\partial_r g_s \\
&\approx \text{(ignoring terms that are 2nd order in the departure from spherical symmetry)} \\
&\quad g_s - H\partial_r \left(\frac{GM}{r^2}\right)_{r=a} \\
&\approx g_s \left[1 + \frac{2}{a}H\right].
\end{aligned}$$

So, now they have g on the geoid. They then find δg (the angular-dependent component of g^g), and solve $\partial_r \delta V + \frac{2}{a}\delta V = -\delta g$ to find δV. (We'll talk about how you solve for δV later.) And $\delta r = \delta V/g$ gives the geoid shape.

4.1.3 Rotating, Elliptically Symmetric Earth

The earth is not a sphere. The biggest contribution to V_T from the non-spherical density components is the addition of a $P_2(\cos\theta)$ term to the $P_2(\cos\theta)$ term that is already present due to the rotation. This new P_2 term is caused by a P_2 deformation of each constant density surface inside the earth which, in turn, is caused by the centrifugal force associated with the earth's rotation. We'll model this P_2 deformation, later. Meanwhile, just take my word that the P_2 contribution to V_T is by far the largest non-spherical contribution.

Let J_2 be a dimensionless parameter which describes the size of the P_2 contribution to V_T. Define J_2 so that outside the earth:

$$V_T = \frac{GM}{r} + \frac{1}{3}\Omega^2 r^2 - \frac{1}{3}\Omega^2 r^2 P_2(\cos\theta) - \underbrace{\frac{MG}{r^3} a^2 J_2 P_2(\cos\theta)}_{\text{new term}}. \tag{4.1.10}$$

The radial dependence of a spherical harmonic of order l is $r^{-(l+1)}$, which is why the term proportional to P_2, above, has a radial dependence of r^{-3} (P_2 is proportional to Y_2^0). The amplitude of the total acceleration (gravitational + centrifugal) outside the earth is then:

$$\begin{aligned}
g_T &\equiv |\overline{g}_T| \\
&\equiv |\overline{\nabla} V_T| \\
&\approx |\partial_r V_T| \qquad \text{(to 1st order in small quantities)} \\
&= \left| -\frac{GM}{r^2} + \frac{2}{3}\Omega^2 r - \frac{2}{3}\Omega^2 r P_2 + \frac{3MG}{r^4} a^2 J_2 P_2 \right| \\
&= \frac{GM}{r^2} - \frac{2}{3}\Omega^2 r + \left[\frac{2}{3}\Omega^2 r - \frac{3MG}{r^4} a^2 J_2 \right] P_2.
\end{aligned}$$

To find the shape of the geoid, we write the geoid as $r = a + \delta r(\theta, \phi)$, and try to find the function δr for which $V_T(r = a + \delta r, \theta, \phi)$ is independent of θ and ϕ. As we saw above, the $\Omega^2 r$ terms are small at $r = a$. And it turns out that $J_2 \ll 1$, as well. (After all, the P_2 density perturbation is due to the $\Omega^2 r$ terms in the potential — as we will see later — and those terms are small.) So we can assume $\delta r / a \ll 1$. Then, putting $r = a + \delta r$ into Equation 4.1.10, and keeping terms 1st order in small quantities, gives the following approximation for V_T on the geoid:

$$V_T \text{ on geoid } = \frac{GM}{a}\left[1 - \frac{\delta r}{a}\right] + \frac{1}{3}\Omega^2 a^2 - \frac{1}{3}\Omega^2 a^2 P_2 - \frac{MG}{a} J_2 P_2.$$

This expression is independent of θ (it is obviously independent of ϕ), if:

$$-\frac{GM}{a^2}\delta r - \left[\frac{1}{3}\Omega^2 a^2 + \frac{MG}{a}J_2\right] P_2 = \text{constant.}$$

Or, if:

$$\delta r \cong -\left[\frac{1}{3}\frac{\Omega^2 a^2}{g_T} + aJ_2\right] P_2 + \text{constant.} \tag{4.1.11}$$

with $g_T \approx GM/a^2$. As in Section 4.1.2, the constant determines *which* constant potential surface we are on. If we want 'a' to equal the mean radius, then the constant in Equation 4.1.11 must equal 0. As in Section 4.1.2, note that

$$\delta r \cong \frac{\delta V(a)}{g_T} \tag{4.1.12}$$

where δV is the angular-dependent part of V_T:

$$\delta V(r) = -\left[\frac{1}{3}\Omega^2 r^2 + \frac{MGa^2}{r^3}J_2\right] P_2. \tag{4.1.13}$$

Also, we can estimate g_T on the geoid, as:

$$\begin{aligned} g^g &\equiv g_T(r = a + \delta r, \theta, \phi) \\ &\approx g_T(a, \theta, \phi) + \delta r \partial_r g_T(r = a) \\ &\approx \frac{GM}{a^2} - \frac{2}{3}\Omega^2 a + \left[\frac{2}{3}\Omega^2 a - \frac{3GM}{a^2}J_2\right] P_2 - 2\frac{GM}{a^3}\delta r. \end{aligned} \tag{4.1.14}$$

where the last equality follows using Equation 4.1.11 for g_T, and keeping only the lowest order term (in g_T) where it multiplies the first-order quantity: δr. If δg is the angular-dependent part of g^g, then using the result for δr gives:

$$\begin{aligned} \delta g &\cong \underbrace{\left[\frac{2}{3}\Omega^2 a - 3g_T J_2\right] P_2}_{-\partial_r \delta V} + 2\frac{g_T}{a}\underbrace{\left[\frac{1}{3}\frac{\Omega^2 a^2}{g_T} + aJ_2\right] P_2}_{-\delta r = -\delta V/g_T} \\ &= \left[\frac{4}{3}\Omega^2 a - g_T J_2\right] P_2 \\ &= -\left[\partial_r \delta V + \frac{2}{a}\delta V\right]_{r=a}. \end{aligned} \tag{4.1.15}$$

Equation 4.1.15 was also valid in Section 4.1.2.

So, to find the shape of the geoid, you solve Equation 4.1.15 for δV, given observations of δg (actually, as described in Section 4.1.2 above, you obtain δg by extending surface gravity data to the geoid using observed leveling elevations). Then, you use δV in Equation 4.1.12 to get δr. Note that once you have δV, you can determine J_2 from Equation 4.1.13. And, a value for J_2 is potentially interesting since it tells you something about the earth's density distribution.

The notation J_2 comes from satellite geodesy. Other subdisciplines use different parameters to describe the P_2 distribution.

For example, the "flattening" is defined as

$$f = \frac{3}{2}J_2 + \frac{a\Omega^2}{2g_T}.$$

The significance of f is that for the simple elliptical earth model discussed here, the geoid height is

$$\delta r = -af\frac{2}{3}P_2 = af\left[\frac{1}{3} - \cos^2\theta\right].$$

So, the *difference* in geoid height between the equator and the poles is

$$af\left[\left(\frac{1}{3} - \cos^2 90^\circ\right) - \left(\frac{1}{3} - \cos^2 0^\circ\right)\right] = af.$$

Another parameter, sometimes written as B_2, is defined as

$$B_2 = \frac{5}{2}\frac{a\Omega^2}{g_T} - f.$$

The significance of B_2 is that the gravitational acceleration observed on the geoid is (for our simple elliptical model)

$$g^g \cong \frac{GM}{a^2} - \frac{2}{3}\Omega^2 a + g_T\frac{2}{3}B_2P_2(\cos\theta).$$

So, the difference between equatorial and polar gravity on the geoid is

$$g_TB_2.$$

4.1.4 The ellipsoid

The real earth is, of course, more complicated than a simple ellipse. There are contributions to V_T from *all* the Y_l^m, not just from P_2. But, the P_2 term is by far the largest.

It is convenient (and usual) to remove the P_2 term from all observations. We will then define "gravity anomalies" and "geoid anomalies" as the differences from the constant plus P_2 terms.

The P_2 term is well determined from satellite observations. The accepted values for J_2, B_2, and f are:

$$\begin{aligned} J_2 &= 1.0826 \times 10^{-3} \\ f &= \frac{1}{298.26} \\ B_2 &= 5.28 \times 10^{-3}. \end{aligned} \tag{4.1.16}$$

These numbers (look, particularly at f) predict a difference in the geoid radius between the poles and the equator of about 21 km. This is much bigger than typical elevation differences associated with surface topography, and is about twice as large as the contribution from rotation alone.

The mathematical surface described by $r = a + \delta r$, where δr is a $P_2(\cos\theta)$ term chosen to be consistent with the numerical values listed in Equation 4.1.16 (in other words, the geoid for the simplified earth in Section 4.1.3) is called the *ellipsoid*. The ellipsoid is a pretty good approximation to the geoid for the real earth — better than a sphere, for example. Later, when we talk about geoid heights for the real earth, we will mean the difference between the geoid and the ellipsoid.

4.2 Clairaut's Differential Equation

Before we go ahead and consider a more realistic earth, with more Y_l^m contributions to V_T, let's see if we can't understand the P_2 term a little better. Why does rotation cause such a term, and why are the results for J_2, B_2, and f equal to the numbers shown above? What do those numerical values tell us about the earth?

It turns out that the numerical results are consistent with the assumption that the earth responds to the centrifugal force of rotation as though it were a fluid. To understand this result, let's suppose we have a spherically symmetric fluid sphere (density = constant on spherical surfaces). We spin the fluid about the $\hat{e}_z$ axis with an angular velocity of rotation $= \Omega$. After the fluid has come to equilibrium, what does the internal density distribution look like? We will assume that the centrifugal force is $\ll$ the gravitational force, so that surfaces of constant density are *almost* spheres — as they were before spin up.

In other words, we assume that a surface of constant density which was $r = r_0$ before spin up, is now

$$r = r_0 \left[1 - \frac{2}{3} \sum_{l=0}^{\infty} \epsilon_l(r_0) P_l(\cos\theta) \right]$$

where $\epsilon_l(r_0) \ll 1$. Why just include Y_l^0, instead of Y_l^m (P_l is proportional to Y_l^0)? Because the centrifugal force, which causes the aspherical shape, is symmetric about the $\hat{e}_z$ axis, so that we expect surfaces of constant density to be independent of ϕ. The use of P_l instead of Y_l^0, and the inclusion of the factor $-2/3$, are consistent with convention. Note that with this convention, $\epsilon_2(a) = f$ (the flattening defined above). Our objective here is to find the $\epsilon_l(r_0)$ for all l and r_0.

Let $V_T(\overline{x})$ = total potential (gravitational plus centrifugal). I claim that for a fluid at equilibrium, surfaces of constant density are also surfaces of constant $V_T(\overline{x})$. To see this, note that at equilibrium there is no motion, and that the only forces are those due to pressure, gravity, and the centrifugal force. If $\mathbf{P}$ = pressure, then the equilibrium condition is

$$\overline{\nabla}\mathbf{P} = \rho \overline{\nabla} V_T$$

where ρ = density. Take $\overline{\nabla}\times$ each side of this equation. Then, since the curl of a gradient is zero:

$$\begin{aligned} 0 &= \overline{\nabla} \times \left(\rho \overline{\nabla} V_T \right) \Rightarrow \\ 0 &= \overline{\nabla}\rho \times \overline{\nabla} V_T. \end{aligned}$$

So, the normals to constant ρ and constant V_T surfaces are everywhere parallel ($\overline{\nabla}\rho$ is along the normal to the constant ρ surface, for example). So, the two surfaces coincide.

It is possible to derive integral/differential equations for the ϵ_l. We will do that here, but only for ϵ_2. You can show from the integral/differential equations for arbitrary l, that $\epsilon_l = 0$ for $l \neq 2$. We'll skip that here, but physically it is because the centrifugal potential includes only $l = 2$ aspherical terms.

So, we assume the constant density/V_T surfaces have a shape described by:

$$r = r_0\left[1 - \frac{2}{3}\epsilon(r_0)P_2(\cos\theta)\right]$$

where $\epsilon \ll 1$. We need to find an equation for $\epsilon(r_0)$.

To do this, we need to find V_T at an arbitrary point (r, θ, ϕ) inside the earth. The centrifugal potential is $\frac{1}{3}\Omega^2 r^2 - \frac{1}{3}\Omega^2 r^2 P_2(\cos\theta)$, as we saw in Equation 4.1.2. The gravitational potential is harder.

To find the gravitational potential, note that we can describe the density of the deformed earth by using a radially-dependent function: $\rho(r_0)$. The real density is dependent on θ, as well, and it can be determined from $\rho(r_0)$. Here's how we do that. We choose a point (r, θ, ϕ) in the earth. This point is on the constant density surface described by r_0, where r_0 satisfies:

$$r = r_0\left[1 - \frac{2}{3}\epsilon(r_0)P_2(\cos\theta)\right].$$

(To first order in ϵ, $r_0\epsilon(r_0) = r\epsilon(r)$. So r_0 is given, to first order, by $r_0 \cong r[1 + \frac{2}{3}\epsilon(r)P_2(\cos\theta)]$.) Define $\rho(r_0)$ as the density on the surface described by the parameter r_0. So the density at (r, θ, ϕ) is $\rho(r_0 = r[1 + \frac{2}{3}\epsilon(r)P_2(\cos\theta)])$, to first order in ϵ.

Now, break the earth up into thin, constant-density shells. Suppose the lower boundary of a shell is the surface described by r_0. The upper boundary is the surface described by $r_0 + dr_0$ where dr_0 is infinitesimal. So, the upper boundary of the shell is

$$r = (r_0 + dr_0)\left[1 - \frac{2}{3}\epsilon(r_0 + dr_0)P_2(\cos\theta)\right].$$

The lower boundary is

$$r = r_0\left[1 - \frac{2}{3}\epsilon(r_0)P_2\right].$$

Let's find V due to the shell, both above the shell (i.e. outside the shell's outer surface) and below the shell (i.e. inside the shell's inner surface). First, we consider a point outside the outer surface. That is, we find V at (r, θ, ϕ), where r is larger than any r in the shell. To do this, we write down the expression for V caused by a uniform object that has density $\rho_0 = \rho(r_0)$, and an outer surface equal to the $(r_0 + dr_0)$ surface (the outer surface of the shell); and we subtract from it the expression for V caused by a uniform object with $\rho_0 = \rho(r_0)$ and an outer surface equal to the r_0 surface (the inner surface of the shell).

From Chapter 3 we know that for small ϵ the potential due to a homogeneous object with outer surface $r = r_0[1 + \epsilon Y_l^m]$ (where $\epsilon \ll 1$) is

$$\underbrace{V(r,\theta,\phi)}_{\text{outside the object}} = \frac{4}{3}\pi\frac{(r_0)^3\rho_0 G}{r} + 4\pi G(r_0)^2\rho_0\frac{\epsilon}{2l+1}\left(\frac{r_0}{r}\right)^{l+1} Y_l^m(\theta,\phi).$$

In our case, $l = 2$ and $m = 0$. And we can replace Y_2^0 with $-\frac{2}{3}P_2$, since that's just a matter of re-defining ϵ in these Chapter 3 results.

So, for the $r_0 + dr_0$ surface:

$$\begin{aligned} V_{r_0+dr_0} &= \frac{4}{3}\pi\frac{(r_0+dr_0)^3\rho_0 G}{r} \\ &\qquad +4\pi G(r_0+dr_0)^2\rho_0\left(-\frac{2}{3}\epsilon(r_0+dr_0)\right)\frac{1}{5}\left(\frac{r_0+dr_0}{r}\right)^3 P_2 \\ &\approx \text{(to first order in } dr_0\text{)} \\ &\qquad \frac{4}{3}\pi\frac{\rho_0 G}{r}r_0^3\left[1+3\frac{dr_0}{r_0}\right] \\ &\qquad -\frac{8}{15}\pi G\frac{P_2(\cos\theta)}{r^3}r_0^5\left[\epsilon(r_0)+dr_0\left(5\frac{\epsilon(r_0)}{r_0}+\partial_{r_0}\epsilon(r_0)\right)\right]\rho_0. \end{aligned}$$

For the r_0 surface:

$$V_{r_0} = \frac{4}{3}\pi\frac{r_0^3\rho_0 G}{r} - \frac{8}{15}\pi G\rho_0 r_0^5\frac{\epsilon(r_0)}{r^3}P_2(\cos\theta).$$

So, the contribution to V from the shell alone is the difference of these V's.

$$\begin{aligned} \underbrace{V_{\text{shell}}}_{\text{outside}} &= V_{r_0+dr_0} - V_{r_0} \\ &= dr_0\, 4\pi\rho_0 G\left[\frac{r_0^2}{r} - \frac{2}{15}\frac{r_0^4}{r^3}\left[5\epsilon(r_0)+r_0\partial_{r_0}\epsilon(r_0)\right]P_2(\cos\theta)\right] \end{aligned}$$

where $\rho_0 \equiv \rho(r_0)$.

Now, let's find V_{shell} *inside* the inner surface of the shell. From the results of Chapter 3, if you are inside a homogeneous body with density ρ_0 and outer surface $r = r_0[1 + \epsilon Y_l^m]$, then

$$\underbrace{V}_{\text{inside}} = 2\pi\rho_0 G\left(r_0^2 - \frac{r^2}{3}\right) + 4\pi G\rho_0 r^l \frac{\epsilon}{2l+1} r_0^{(2-l)} Y_l^m.$$

In our case $l = 2$, $m = 0$, and we again replace Y_l^m with $-\frac{2}{3}P_2$. The effect on V from an object with surface described by $r_0 + dr_0$ is

$$\begin{aligned} V_{r_0+dr_0} &= 2\pi\rho_0 G\left((r_0+dr_0)^2 - \frac{r^2}{3}\right) - \frac{8\pi G\rho_0}{15} r^2 \epsilon(r_0+dr_0) \\ &\approx \text{(to 1st order in } dr_0) \\ &\qquad 2\pi\rho_0 G\left[r_0^2 + 2r_0 dr_0 - \frac{r^2}{3}\right] - \frac{8\pi G\rho_0}{15} r^2 \left[\epsilon(r_0) + dr_0 \partial_{r_0}\epsilon\right] P_2. \end{aligned}$$

The contribution from a homogeneous object with an r_0 surface is

$$V_{r_0} = 2\pi\rho_0 G\left(r_0^2 - \frac{r^2}{3}\right) - \frac{8\pi G\rho_0}{15} r^2 P_2.$$

So, the contribution from the shell is the difference:

$$\begin{aligned} \underbrace{V_{dr_0}}_{\text{inside}} &= V_{r_0+dr_0} - V_{r_0} \\ &= \left[4\pi\rho_0 G r_0 - \frac{8\pi G\rho_0}{15} r^2 \partial_{r_0}\epsilon(r_0) P_2(\cos\theta)\right] dr_0. \end{aligned}$$

Incidentally, you might wonder: why not estimate the effects of the shell by finding the mass in a radial column of the shell (that mass would equal the density times the thickness) and then pretend the mass is a surface density at $r = r_0$. After all, that's what we did to find the potential from a homogeneous object with a slightly non-spherical surface. But, that doesn't work here. That will give an answer correct to 1st order in the shell thickness. So, for the homogeneous body that means correct to order ϵ. But our shell differs from a spherical surface mass by terms of order dr_0 *and* ϵ. If we were to pretend it was a spherical surface mass we would be throwing away $(dr_0)^2$ and ϵ^2 terms — which is ok — but also the $(dr_0\epsilon)$ terms — which are important. So, that method doesn't work.

So, we now have V from a shell. To find $V_T(r,\theta,\phi)$ from the entire object, we sum over shells. That means we integrate V_{dr_0} over r_0. We use the outer solution for $0 \le r_0 \le r$.

And the inner solution for $r \le r_0 \le a$, where $a =$ earth's radius. And, we include the centrifugal potential. The result is:

$$\begin{aligned} V_T(r,\theta,\phi) = &\int_0^r 4\pi G\rho(r_0)\left[\frac{r_0^2}{r} - \frac{2}{15}\frac{r_0^4}{r^3}\left[5\epsilon(r_0) + r_0\partial_{r_0}\epsilon(r_0)\right]P_2(\cos\theta)\right] dr_0 \\ &+ \int_r^a 4\pi\rho(r_0)G\left[r_0 - \frac{2}{15}r^2\partial_{r_0}\epsilon(r_0)P_2(\cos\theta)\right] dr_0 \\ &+ \frac{1}{3}\Omega^2 r^2 - \frac{1}{3}\Omega^2 r^2 P_2(\cos\theta) \end{aligned} \tag{4.2.1}$$

Incidentally, there's another thing you might worry about. Our result for V inside a shell was really only valid at points (r,θ,ϕ), where r was less than every r in the shell. And, our result for V outside was only good when r was larger than every r in the shell. So, the integrals in Equation 4.2.1 should really have the limits $\int_0^{r-\Delta r_1}$ and $\int_{r+\Delta r_2}^a$ where Δr_1 and Δr_2 are of order ϵ. And then there should be a 3rd integral between $r - \Delta r_1$ and $r + \Delta r_2$, where the integrand is some more complicated expression for V_{dr_0} which is valid when r is "in" the shell. In fact, you can see that something is amiss by looking at the shell potentials V_{dr_0} inside and outside the shell. The two are not continuous when $r = r_0$, but V should be continuous. What happens is that V_{dr_0} should be modified slightly at r, within ϵ or so of r_0.

But, this all gives a 2nd order (in ϵ) effect on $V_T(r)$. Comparing V_{dr_0} inside and outside at $r = r_0$, shows that the discontinuity is of order ϵ. So the true V_{dr_0} near $r = r_0$ differs from the V_{dr_0} used here, by terms of order ϵ. And, the size of the region you integrate over is order ϵ. So, the result is a 2nd order effect in ϵ, and thus the result above for V_T is ok to 1st order in ϵ.

We still must find an equation for ϵ. By assumption, a constant density surface is described by

$$r = r_1\left[1 - \frac{2}{3}\epsilon(r_1)P_2(\cos\theta)\right].$$

(Use r_1 instead of r_0 to distinguish between the field (r_1) point and the mass (r_0) point.) For this to also be a constant potential surface, $V_T(r = r_1\left[1 - \frac{2}{3}\epsilon P_2\right], \theta, \phi)$ must be independent of θ and ϕ. To 1st order in ϵ,

$$V_T\left(r = r_1\left[1 - \frac{2}{3}\epsilon P_2\right], \theta, \phi\right) = V_T(r_1,\theta,\phi) - \frac{2}{3}r_1\epsilon(r_1)P_2\partial_r V_T(r_1,\theta,\phi). \tag{4.2.2}$$

If we only require Equation 4.2.2 to be accurate to first order in ϵ, then we can approximate $\partial_r V_T$ by setting $\epsilon = 0$ in V_T before differentiating (since $\partial_r V_T$ is already multiplied by ϵ in Equation 4.2.2). We get (ignoring the Ω^2 terms in $\partial_r V_T$ as well, since those terms are the same size as the terms that are first-order in ϵ):

$$\begin{aligned}
\partial_r V_T &\cong 4\pi G \partial_r \left[\int_0^r \rho(r_0) \frac{r_0^2}{r} \, dr_0 + \int_r^a \rho(r_0) r_0 \, dr_0 \right] && (4.2.3)\\
&= 4\pi G \left[\rho(r) \frac{r^2}{r} - \int_0^r \rho(r_0) \frac{r_0^2}{r^2} \, dr_0 - \rho(r) r \right] && (4.2.4)\\
&= -\frac{4\pi G}{r^2} \int_0^r \rho(r_0) r_0^2 \, dr_0. && (4.2.5)
\end{aligned}$$

So V_T on this constant-density surface, from Equation 4.2.2, and using Equations 4.2.1 and 4.2.5, is:

$$\begin{aligned}
V_S &\equiv V_T \left(r = r_1 \left[1 - \frac{2}{3} \epsilon P_2 \right], \theta, \phi \right) \\
&= 4\pi G \int_0^{r_1} \rho(r_0) \left[\frac{r_0^2}{r_1} - \frac{2}{15} \frac{r_0^4}{r_1^3} \left[5\epsilon(r_0) + r_0 \partial_{r_0} \epsilon(r_0) \right] P_2(\cos\theta) \right] dr_0 \\
&\quad + 4\pi G \int_{r_1}^a \rho(r_0) \left[r_0 - \frac{2}{15} r_1^2 \partial_{r_0} \epsilon(r_0) P_2(\cos\theta) \right] dr_0 \\
&\quad + \frac{1}{3} \Omega^2 r_1^2 - \frac{1}{3} \Omega^2 r_1^2 P_2(\cos\theta) \\
&\quad + \frac{2}{3} r_1 \epsilon(r_1) P_2 \frac{4\pi G}{r_1^2} \int_0^{r_1} \rho(r_0) r_0^2 \, dr_0.
\end{aligned}$$

We want V_S to be independent of θ. So, we set the sum of the $P_2(\cos\theta)$ coefficients equal to zero, which results in the condition:

$$\begin{aligned}
4\pi G \int_0^{r_1} \rho(r_0) \left(-\frac{2}{15} \right) \frac{r_0^4}{r_1^3} \left[5\epsilon(r_0) + r_0 \partial_{r_0} \epsilon(r_0) \right] dr_0 \\
+ 4\pi G \int_{r_1}^a \rho(r_0) \left(-\frac{2}{15} \right) r_1^2 \partial_{r_0} \epsilon(r_0) \, dr_0 \\
- \frac{1}{3} \Omega^2 r_1^2 + \frac{8}{3} \frac{\epsilon(r_1) \pi G}{r_1} \int_0^{r_1} \rho(r_0) r_0^2 \, dr_0 = 0.
\end{aligned}$$

A slight re-organization gives:

$$
\begin{aligned}
-\frac{8}{5}\pi G \left[\frac{1}{r_1^3} \int_0^{r_1} \right. & \rho(r_0) r_0^4 \left[5\epsilon(r_0) + r_0 \partial_{r_0} \epsilon(r_0) \right] dr_0 \\
& + r_1^2 \int_{r_1}^{a} \rho(r_0) \partial_{r_0} \epsilon(r_0)\, dr_0 \\
& \left. - 5\frac{\epsilon(r_1)}{r_1} \int_0^{r_1} \rho(r_0) r_0^2\, dr_0 \right] = \Omega^2 r_1^2. \qquad (4.2.6)
\end{aligned}
$$

Equation 4.2.6 is an integral equation for ϵ. We can transform it into a differential equation by:

1. multiplying by r_1^3 and taking ∂_{r_1} of the result; and then
2. multiplying by r_1^{-4} and taking ∂_{r_1} again.

The result is Clairaut's differential equation for $\epsilon(r)$, which, after changing r_1 to r, has the form:

$$
\overline{\rho}(r) \left(\partial_r^2 \epsilon(r) - \frac{6}{r^2}\epsilon(r) \right) + \frac{6\rho(r)}{r} \left(\partial_r \epsilon(r) + \frac{\epsilon(r)}{r} \right) = 0 \qquad (4.2.7)
$$

where

$$
\overline{\rho}(r) = \frac{3}{r^3} \int_0^r \rho(r_0) r_0^2\, dr_0.
$$

Note that all Ω^2 terms are missing in Equation 4.2.7. That means that Clairaut's equation doesn't have enough information to uniquely determine $\epsilon(r)$. Instead, to find ϵ you find a general solution to Clairaut's equation. (There will be two independent solutions since the equation is 2nd order.) Then you find the coefficients in the general solution, by requiring that the result satisfies the integral equation for ϵ, Equation 4.2.6 above.

4.2.1 Example: A uniform (homogeneous) earth

Assume $\rho = \rho_0 =$ constant. Then:

$$
\overline{\rho}(r) = \frac{3}{r^3}\rho_0 \int_0^r r_0^2\, dr_0 = \rho_0.
$$

In this case, Clairaut's equation reduces to:

$$\partial_r^2 \epsilon + \frac{6}{r}\partial_r \epsilon = 0.$$

We try a solution of the form $\epsilon = r^n$, and find that

$$n(n-1) + 6n = 0.$$

Or

$$n(n+5) = 0 \qquad \Rightarrow \qquad n = \begin{cases} 0 \\ -5. \end{cases}$$

So, the general solution is:

$$\epsilon = b + \frac{c}{r^5}.$$

Now, we find b and c. We note right away that $c = 0$. Otherwise, $\epsilon \to \infty$ near the earth's center. This would mean that the radii of constant density/potential surfaces $\to \infty$ near the center, which we know cannot be true.

So $c = 0$, and the general solution is: $\epsilon = b =$ constant. To find b in terms of Ω^2, we put $\epsilon = b$ into the integral equation Equation 4.2.6. Then $\partial_{r_0}\epsilon = 0$, and Equation 4.2.6 reduces to:

$$-\frac{8}{15}\pi G \left[\frac{1}{r_1^3} 5b\rho_0 \int_0^{r_1} r_0^4 \, dr_0 - 5\frac{b}{r_1}\rho_0 \int_0^{r_1} r_0^2 \, dr_0\right] = \frac{1}{3}\Omega^2 r_1^2$$

or:

$$-\frac{8}{15}\pi G \left[\frac{5b\rho_0}{r_1^3}\frac{r_1^5}{5} - 5\frac{b}{r_1}\rho_0 \frac{r_1^3}{3}\right] = \frac{1}{3}\Omega^2 r_1^2$$

or:

$$b = \frac{15}{16}\left(\frac{\Omega^2}{\pi\rho_0 G}\right). \tag{4.2.8}$$

So, ellipticity = constant = Equation 4.2.8.

The real density inside the earth is not constant, but increases with depth. It turns out that then Clairaut's equation implies that ϵ increases with radius. So the outer portions of our fluid earth are more aspherical than the inner portions.

ϵ also always turns out to be positive, which implies that the surfaces are squashed down at the poles. To see this, note that (remembering that $P_2 = \frac{1}{2}(3\cos^2\theta - 1)$):

$$r(\theta = 90^\circ) - r(\theta = 0^\circ) \;=\; -\frac{2}{3}\epsilon r_0 \left[P_2(90^\circ) - P_2(0^\circ)\right]$$

$$
\begin{aligned}
&= -\frac{2}{3}\epsilon r_0 \left[-\frac{1}{2} - 1\right] \\
&= \epsilon r_0 .
\end{aligned}
$$

So, $\epsilon > 0 \Rightarrow r(90^\circ) > r(0^\circ)$.

If you use the best seismic results for the density function $\rho(r_0)$ and solve Clairaut's equation, you find:

$$\epsilon(a) = f(= \text{ flattening }) = \frac{1}{299.7}.$$

The observed f, using satellite data for J_2, is

$$f_{\text{obs}} = \frac{1}{298.257}.$$

The difference is $1/2\%$ (i.e. a relative accuracy of 5×10^{-3}).

This agreement is good, but the observational errors for J_2 are better than 10^{-6}. So, why is there this discrepancy?

One possibility is that the 1st order theory is not good enough, so that you should go to 2nd order. People have constructed a 2nd order theory, but the 2nd order terms affect the solution at only the 5×10^{-4} level (.05%), which is not large enough.

Another possibility is that the seismic density model is not quite right. You can modify the density slightly so that it is still reasonably consistent with seismic data and predicts the correct f. But, you still have trouble. That's because of the following:

We'll see, later, that the earth precesses due to the gravitational attraction of the sun and moon, in the same way a top nutates due to the gravitational attraction of the earth. The period of the precession is well known from observations — it is roughly 26,000 years. The period depends on the "dynamical ellipticity" of the earth:

$$H = \frac{C - A}{C}$$

where C and A are principal moments of inertia of the earth. We'll see all this, later.

It turns out (you can get this by appropriate integration of the internal density) that H is proportional to the coefficient of the $P_2(\cos\theta)$ term in the density, integrated through the earth. In other words, H is proportional to the radial integral (through the entire earth) of $\epsilon(r)$. f (or J_2), on the other hand, is proportional to $\epsilon(r)$ at the outer surface $r = a$.

To first order in ϵ you can relate H to f from Clairaut's equation. It turns out that:

$$H = \frac{f - \frac{1}{2}m}{1 - \frac{2}{5}\sqrt{\frac{5}{2}\frac{m}{f} - 1}}$$

where $m = \Omega^2 a/g_T$.

So, once f is known — H follows directly. And, note that the relation between H and f is independent of the density.

Well, the observed H and f don't satisfy this relation (m is known). So, no matter what the density is, you can't satisfy the observations of H and f from Clairaut's equation. Using the best seismic density results in Clairaut's equation gives a predicted H of 1/308.0. The observed H is 1/304.437. The discrepancy is approximately 1%. If you adjust the density so that f is right, then H is still wrong by 0.5%. If you adjust the density so H is right, then f is wrong by approximately 0.7%. Or, you can adjust the density so that both are right to approximately 0.4%.

The only conclusion possible is that the earth is not exactly in hydrostatic equilibrium with the centrifugal potential. Note that a 0.5% error in f is approximately 0.005×21 km ≈ 100 m error in the pole-equator elevation difference in the ellipsoid, and that's about the same order as the geoid anomalies for other Y_l^m terms. (Actually, this term is still somewhat larger than the other Y_l^m terms.)

So, what can cause the ellipticity to be different from the expected value? One unlikely explanation is that the earth has some finite strength. That is, it can support some shear stress over long time periods, and so is not a perfect fluid. Most geophysicists don't believe that. It is true that the earth's crust and lithosphere can support shear stresses for long times. After all, there is topography on the surface, and if the crust behaved as a fluid that topography would presumably 'flow' laterally until it had disappeared. But, in order to have a significant non-fluid response to the centrifugal force, there'd have to be finite strength over a much thicker region than the thin crust/lithosphere. And, most geophysicists are reluctant to believe that.

A related possibility is that the mantle is a very viscous fluid. That is, the viscosity is large and so it takes a reasonably long time for the mantle fluid to reach equilibrium. The relevance of this is that the earth's rotation is decreasing approximately linearly with time (we'll get into this later). And so the centrifugal force is decreasing. If the viscosity is high enough that the mantle response time is longer than the time it takes for the centrifugal force to decrease significantly, then the present shape of the earth might be reflecting the older, faster rotation. This would lead to ellipticities larger than those predicted from Clairaut's equation. Most geophysicists don't believe this explanation, either. There are estimates of the mantle viscosity from postglacial rebound (we'll talk about that, later), and they suggest the viscosity is not large enough.

The most likely explanation is that the earth *is* approximately fluid over long time periods, but that there are other factors besides rotation that cause lateral variations in density. That is certainly the case: thermal anomalies must exist within the earth if there is to be on-going mantle convection. Those thermal anomalies must be associated with density anomalies (it is buoyancy forces caused by the density anomalies that directly drive the convection). So, presumably the discrepancy from Clairaut's equation is due to the Y_2^0 component of those thermal/density anomalies. The non-Y_2^0 part of the geoid is also presumably due to this effect.

4.3 A More Realistic Earth

In Section 4.1, we described the geoid for an elliptical earth. Here, we extend that description to include non-elliptical contributions.

First, let's discuss the definition of the geoid a little more. The definition we have used so far is: the surface of constant potential which coincides with mean sea level over the oceans. If you cut thin canals across the continents joining the oceans, the water in the canals would be on that constant potential surface.

But, this is not really the geoid. The real geoid is a surface which is defined in some sense by the leveling process: it is the reference surface for leveling. For example, you level and find that Boulder, Colorado is approximately 5000 ft above some reference surface — that surface is the geoid.

Suppose you make leveling measurements everywhere over the earth's surface. Over the oceans, the leveling results would indicate no change in elevation (except for the effects on sea level from dynamical oceanographic processes — like currents). So, over the oceans the geoid does coincide with a constant potential surface. What about under land?

There certainly is a constant potential surface under the land connecting the ocean surfaces. But, it is not exactly coincident with the geoid (although it's pretty close). The horizontal surface that the leveling instrument uses as a reference is, indeed, an equipotential surface. But it is the equipotential surface that runs through the position of the instrument, which at Boulder is about 5000 ft above the reference surface you are trying to define. And the equipotential surface down at that depth is not necessarily parallel to the one that runs through Boulder.

For example, consider the possible situation pictured in Figure 4.1. The leveling

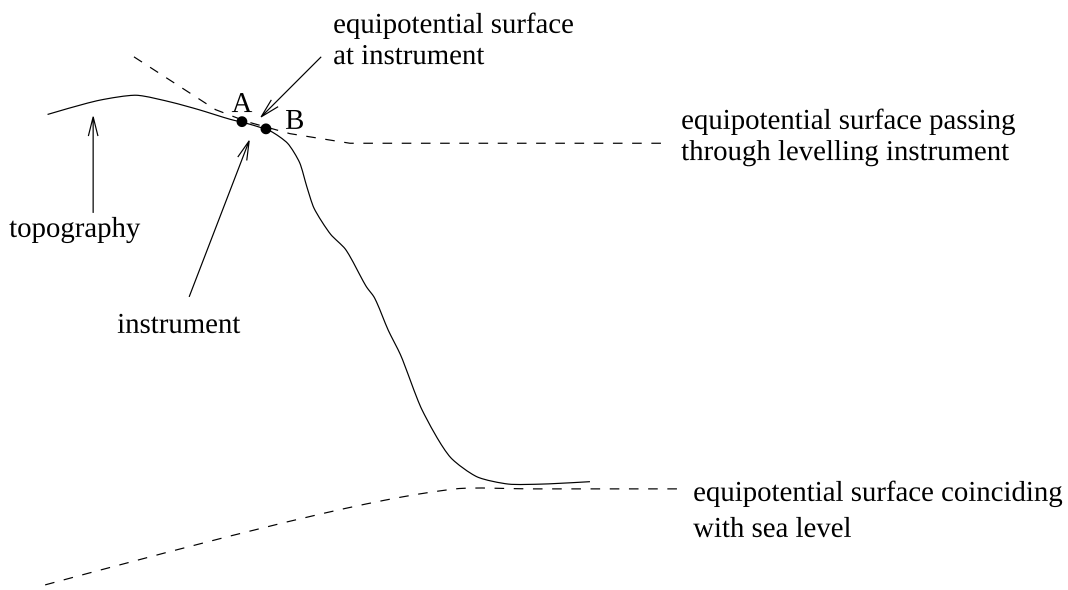

Figure 4.1:

instrument would show that the two endpoints, A and B, of line $\overline{AB}$, were at the same elevation because they lie on the same equipotential surface. So, we would want to define the geoid so that A and B were the same distance above it. But in this example, the two endpoints are *not* equal distances above the lower equipotential surface, because the equipotential surfaces above and below the topographic feature are not parallel to one another. (The figure shows the equipotential surface as bending downward beneath the continent; this is only artistic license: it would be highly unlikely.)

Incidentally, there is another type of problem that sometimes comes up when trying to use leveling observations to determine elevations. It turns out that for particularly unfavorable geometries, the apparent elevation difference between two locations can depend on the leveling route you use. For example, consider the mountain shown in Figure 4.2, where there is more mass on one side than the other. You want to level from A to B.

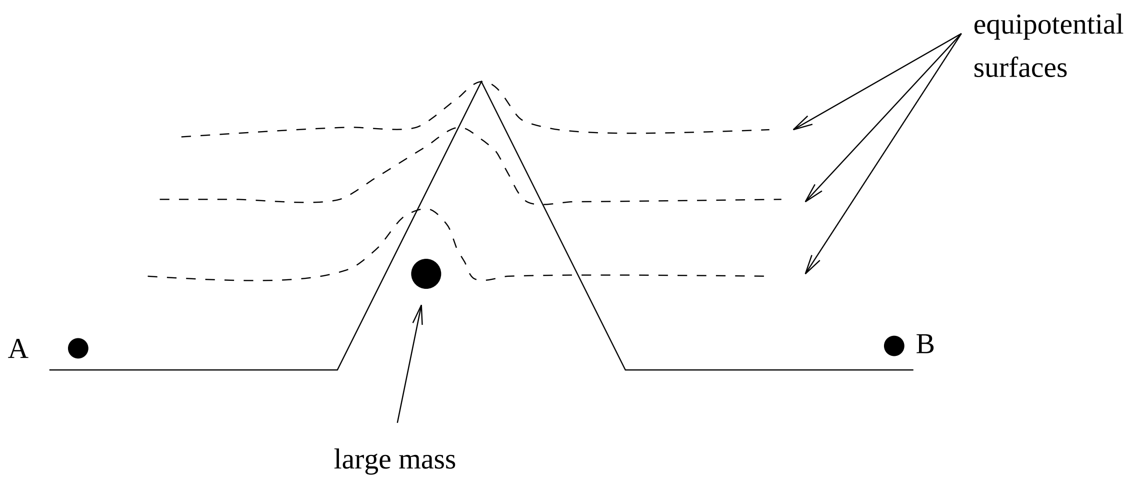

Figure 4.2:

One leveling route you can take is around the mountain: out of the page and around to B, where the equipotential surfaces are presumably all flat and parallel to the surface. Using this route, you would conclude that A and B were at the same elevation. Another route you could take would be over the mountain. As you go up the left hand side, the equipotential surfaces are tipped so that they are more parallel to the mountain surface than are the equipotential surfaces on the right hand side. So, leveling would not register as much elevation gain going up the left hand side as the measured elevation loss going down the right hand side. You would then conclude that B was lower than A.

These sorts of errors are usually negligible. They are only important when you've got large elevation changes along the leveling route. Geodesists have developed methods to reduce them (usually involving gravity measurements) although you can't eliminate them entirely. What this problem shows, though, is that the idea of a well-defined surface (the geoid) acting as a reference surface for leveling is not really valid. With the geoid, we are simply trying to find a surface which comes as close to a useful reference surface as possible.

So, how do we find such a surface? We want to try to relate it to gravity somehow, so that we can determine it from gravity observations. Over the ocean the geoid *will* be an equipotential surface. Under continents, we want the geoid to be parallel to the equipotential surface at the leveling instrument overhead. External equipotential surfaces are approximately parallel to one another, so long as they are not separated by too great a distance. But, an equipotential surface inside the earth is apt to be distinctly tilted with respect to an external equipotential surface.

For example, consider the topography in Figure 4.3. The land mass warps the internal

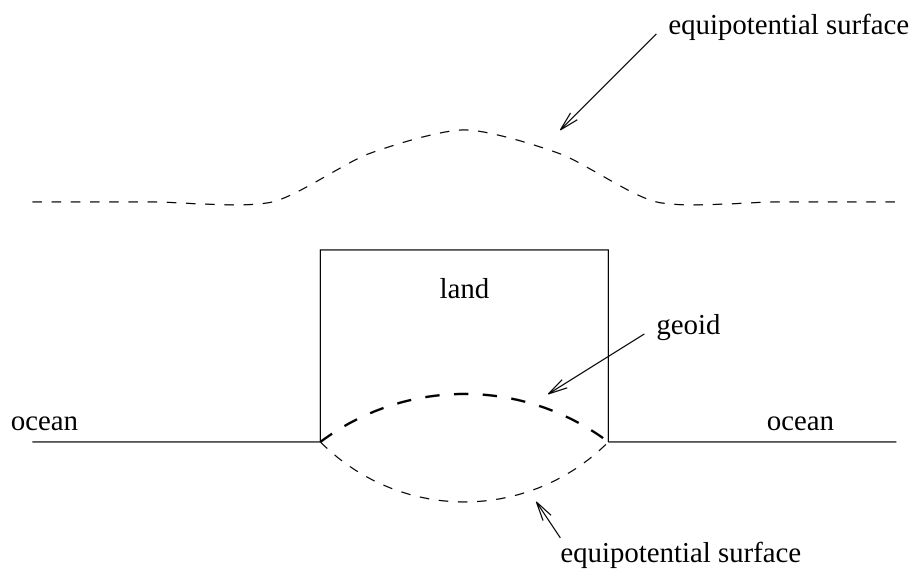

Figure 4.3:

and external equipotential surfaces so that they are not parallel. (Again, it is highly unlikely that an equipotential surface would be warped downward, as illustrated in the figure.) The geoid should be defined so that it is parallel to the external surface. Thus the geoid in this example will *not* coincide with an equipotential surface.

So, what do we do? We use the fact that external equipotential surfaces are reasonably parallel. For example, one way to define a reference surface for leveling would be to choose the equipotential surface that lies just above Mt. Everest. That surface, since it is everywhere external to the earth, is nearly parallel to the local equipotential surfaces. And, it is a physically meaningful surface (i.e. it is an equipotential surface). But, it obviously does not coincide with sea level over the oceans. In fact, if we used this as a reference surface, we would end up assigning all elevations to be negative, determined by the number of feet below Mt. Everest.

In a sense, this *is* what people do. Only they (sort of) add the height of Mt. Everest to every such elevation. More precisely, they expand the potential just outside Mt. Everest as a sum of spherical harmonics. They then continue this expansion down to mean sea level by assuming that the radial dependence of each Y_l^m term is $r^{-(l+1)}$, which is the radial dependence for an external field. This results in a sort of quasi-potential at sea level. They then define the geoid as the surface over which this quasi-potential is constant. Note that this does not give the true potential down at sea level, because the actual radial dependence is much more complicated when you continue down through continental mass. Instead, it is nearly equivalent to compressing all the continental mass down into a thin layer lying just below sea level, and then finding the equipotential surface for this re-organized earth. The resulting surface (the geoid) will thus not really be an equipotential surface under continents, but it will be nearly parallel to the equipotential surfaces up where the leveling observations are made. It is a more useful reference surface than the Mt. Everest surface, because most of the earth's surface is at or near sea level.

Incidentally, you need to know the geoid not just for conventional geodetic applications (that is, for global surveying), but also to interpret long wavelength surface gravity observations. Suppose you measure gravity on the surface. You get some anomalous results. You want to use the results to learn about the earth's density. But, are the results anomalous because of anomalous density inside the earth, or because there is topography on the earth's surface. Specifically, if you observe a gravity high, is it due to excess underlying density, or to a depression in the surface at that point — causing the gravimeter to be closer to the earth's center where the gravitational acceleration is larger? The only way to know is to determine the shape of the earth — and that requires knowledge of the geoid.

4.3.1 The Geoid

Let's find a relation between the geoid and the earth's external potential, for an earth with an arbitrary aspherical shape. The external potential for a realistic earth (including rotation) has the form

$$\begin{aligned} V_T \;=\; & \underbrace{\left[\frac{GM}{r}+\frac{1}{3}\Omega^2 r^2\right]}_{Y_0^0} - \underbrace{\left[\frac{1}{3}\Omega^2 r^2 + \frac{GM}{r^3}a^2 J_2\right]}_{Y_2^0} P_2(\cos\theta) \\ & - \frac{GM}{r}\sum_{\substack{l=2 \\ (l,m)\neq(2,0)}}^{\infty}\sum_{m=-l}^{l} V_l^m Y_l^m(\theta,\phi)\left(\frac{a}{r}\right)^l \end{aligned} \tag{4.3.1}$$

where the V_l^m are constants. Here we have separated the Y_0^0 and Y_2^0 terms from the sum over l and m, only because we have already considered their effects on the geoid.

There is no $l = 1$ term in the sum over l because we are assuming the origin of the coordinate system is the earth's center of mass. What's the connection? Back in Chapter 3, I showed that if you have a density distribution $\rho(\overline{x}') = \sum_{l,m} \rho_l^m(r')Y_l^m(\theta',\phi')$, then the gravitational potential outside the object would be

$$V(\overline{x}) = 4\pi G\sum_{l,m}\frac{Y_l^m(\theta,\phi)}{2l+1}\frac{1}{r^{l+1}}\left[\int (r')^{l+2}\rho_l^m(r')\,dr'\right].$$

So, the Y_1^m coefficient in V is proportional to

$$\int (r')^3 \rho_1^m(r')\,dr'. \tag{4.3.2}$$

where

$$\rho_1^m(r) = \int_0^{2\pi} d\phi \int_0^{\pi} \sin\theta\,d\theta\,\rho(r,\theta,\phi)Y_1^{m*}(\theta,\phi).$$

So Equation 4.3.2 becomes (dropping the primes on r):

$$\int_{\text{earth}} \rho(\overline{x}) r Y_1^{m*}(\theta, \phi)\, d^3\overline{x} \tag{4.3.3}$$

where $d^3\overline{x} = r^2\, dr\, \sin\theta\, d\theta\, d\phi$. In terms of rectangular coordinates, the Y_1^m are:

$$\begin{aligned} rY_1^{0*} &= \sqrt{\frac{3}{4\pi}}\, z \\ rY_1^{1*} &= -\sqrt{\frac{3}{8\pi}}\, (x - iy) \\ rY_1^{-1*} &= \sqrt{\frac{3}{8\pi}}\, (x + iy). \end{aligned}$$

So, the integral in Equation 4.3.3 is proportional to $[\int_{\text{earth}} \rho z\, d^3\overline{x}]$, $[\int \rho(x - iy)\, d^3\overline{x}]$, $[\int \rho(x + iy)\, d^3\overline{x}]$ depending on the value of m. But, if the origin is chosen to be the center of mass, then $\int \rho z = \int \rho x = \int \rho y = 0$, so that the integrals for $l = 1$ vanish for all m. Thus, there are no $l = 1$ terms in the expansion of V_T.

By collecting all the angular-dependent terms of V_T — including the P_2 term — into $\delta V(r, \theta, \phi)$, we can write:

$$V_T = \left[\frac{GM}{r} + \frac{1}{3}\Omega^2 r^2\right] + \delta V(r, \theta, \phi) \tag{4.3.4}$$

Note that δV is much smaller then GM/r.

Equation 4.3.4 is only valid *outside* the earth. But let's now just consider it as a mathematical expression and extend the expression, as written, down into the earth far enough that it coincides with mean sea level over the oceans. We then define the geoid as the surface

$$r = a + \delta r(\theta, \phi),$$

where the mean of δr is zero, and where $V_T(r = a + \delta r, \theta, \phi)$ is independent of θ and ϕ. So, V_T is constant on the geoid. Thus, the geoid won't really be an equipotential surface — since our expression for V_T does not really equal the potential inside the earth — but it will be nearly parallel to the equipotential surface at the earth's outer surface. So, by assuming that $\delta r/a \ll 1$, and that $\delta V/(GM/r) \ll 1$ and $\frac{1}{3}\Omega^2 r^2/(GM/r) \ll 1$ near $r = a$, we obtain:

$$V_T(r = a + \delta r) \approx \frac{GM}{a + \delta r} + \frac{1}{3}\Omega^2 a^2 + \delta V(a, \theta, \phi).$$

In other words, to lowest order we only need to include δr in the leading GM/r term. Expanding $GM/(a+\delta r)$ to 1st order gives the geoid condition

$$\frac{GM}{a}\left[1-\frac{\delta r}{a}\right]+\frac{1}{3}\Omega^2 a^2+\delta V(a,\theta,\phi)= \text{ independent of } \theta,\phi.$$

So, since GM/a and $\frac{1}{3}\Omega^2 a^2$ are already independent of θ and ϕ, we get:

$$\delta r\left(\frac{GM}{a^2}\right)=\delta V(a,\theta,\phi)+ \text{ constant.}$$

The constant $=0$ if the average of δr is to be 0 (since the average of each Y_l^m, and so of δV, is zero). To lowest order: $GM/a^2 \cong g_T(r=a) =$ the acceleration at the surface. So:

$$\delta r \cong \left.\left(\frac{\delta V}{g_T}\right)\right|_{r=a}.$$

That's the result we derived for the simpler earth models above. Evidently it is always true.

4.3.2 Stoke's Formula

You can find δV either from surface gravity observations or from satellite ranging data. Stoke's formula is a method for finding the geoid from surface gravity. Here's how it works.

The total acceleration vector (gravitational + centrifugal) that affects a gravimeter is $\overline{g}_T=\overline{\nabla}V_T$. The amplitude of $\overline{g}_T$ (the quantity actually observed) is, to first order in δV, equal to the negative of the radial component of $\overline{g}_T$ (it's the *negative* of the radial component because the acceleration is downward, and the positive radial direction is upward). So

$$g_T=-\partial_r V_T=\left[\frac{GM}{r^2}-\frac{2}{3}\Omega^2 r\right]-\partial_r\delta V. \tag{4.3.5}$$

In practice, you measure g_T on the earth's surface. That surface is

$$r=a+\delta r(\theta,\phi)+H(\theta,\phi)$$

where $H(\theta,\phi)$ is the surface elevation as determined by leveling. (Leveling gives the elevation above the geoid.) So, from gravity and leveling, you can determine g on the geoid: g^g. What you measure with gravimeters is $g_T(r = a + \delta r + H)$. And $g^g \equiv g_T(r = a + \delta r)$. To 1st order in H/a:

$$\begin{aligned} g^g &= g_T(r = (a + \delta r + H) - H) \\ &= g_T(r = a + \delta r + H) - H\ \partial_r g_T|_{r=a+\delta r+H}\,. \end{aligned}$$

And, to *zero* order in H/a and $\delta r/a$,

$$\partial_r g_T|_{r=a+\delta r+H} \cong -\frac{2GM}{r^3}\bigg|_{r=a+\delta r+H} \approx -\frac{2GM}{a^3} \cong -\frac{2}{a}g_T\bigg|_a .$$

So:

$$g^g \approx \underbrace{g_T(r = a + \delta r + H)}_{g_{\rm obs}} + \frac{2}{a}g_T\bigg|_{r=a} H.$$

You measure g_{obs} with gravimeters, and H by leveling. Thus, you can deduce g^g from your observations.

Now, how does g^g depend on δV?

$$g^g = g_T(r = a + \delta r) \cong g_T(r = a) + \delta r \partial_r g_T(r = a).$$

To zero order in small quantities:

$$\partial_r g_T|_{r=a} = -\frac{2GM}{a^3} = -\frac{2}{a}g_T\bigg|_a .$$

So, to first order, and using Equation 4.3.5:

$$g^g \cong \left(\frac{GM}{a^2} - \frac{2}{3}\Omega^2 a\right) - \partial_r \delta V|_{r=a} - \frac{2}{a}g_T\bigg|_a \delta r.$$

But, $\delta r \approx (\delta V/g_T)\,|_{r=a}$. So:

$$g^g \approx \left(\frac{GM}{a^2} - \frac{2}{3}\Omega^2 a\right) - \left[\partial_r \delta V + \frac{2}{a}\delta V\right]\bigg|_{r=a} .$$

So:

$$g^g - \left[\frac{GM}{a^2} - \frac{2}{3}\Omega^2 a\right] = -\left[\partial_r \delta V + \frac{2}{a}\delta V\right]\bigg|_{r=a} . \qquad (4.3.6)$$

This (Equation 4.3.6) is another result that I had claimed, earlier in this section, was always true. (For the simpler earth models considered earlier, I had defined the left hand side of Equation 4.3.6 as δg.)

Both $g^g - [\frac{GM}{a^2} - \frac{2}{3}\Omega^2 a]$ and δr are dominated by the P_2 terms. It is usual to remove those terms from g^g and δr. δr *minus* the P_2 terms is usually written as N. The interpretation of N is that it is the height of the geoid above the *ellipsoid* (δr = height of geoid above the surface $r = a$). So:

$$N = \left.\frac{\Delta V}{g_T}\right|_{r=a}$$

where $\Delta V = \delta V$ after subtracting all P_2 terms.

The relation between g^g and ΔV is

$$g^g - \left[\frac{GM}{a^2} - \frac{2}{3}\Omega^2 a\right] + \left.\left[\partial_r \delta V_{l=2}^{m=0} + \frac{2}{a}\delta V_{l=2}^{m=0}\right]\right|_{r=a} = -\left.\left[\partial_r \Delta V + \frac{2}{a}\Delta V\right]\right|_{r=a}.$$

The left hand side is:

$$\underbrace{\left[g_{\mathrm{obs}} + \frac{2}{a} H \left. g_T\right|_a\right]}_{g^g} - \underbrace{\left[\left(\frac{GM}{a^2} - \frac{2}{3}\Omega^2 a\right) - \left.\left[\partial_r \delta V_{l=2}^{m=0} + \frac{2}{a}\delta V_{l=2}^{m=0}\right]\right|_{r=a}\right]}_{\approx \gamma_0}.$$

There is an internationally-sanctioned equation that approximates the large second bracketed term in this equation, and that is used to define the official "Geodetic Reference System"

$$\gamma_0 = 978.03188\left[1 + 0.0052788\cos^2\theta + 0.0000236\cos^4\theta\right]. \qquad (4.3.7)$$

These numbers are based on global observations. I've talked as though we only want to remove the constant and P_2 terms from g^g — but note the $\cos^4\theta$ term in γ_0 (Equation 4.3.7). That term represents *second* order effects of J_2 and Ω^2 on g^g. It's a small term, and it is often not necessary to include it when reducing data.

So, defining $\Delta g = g_{\mathrm{obs}} - \gamma_0 + \frac{2}{a} H g_T|_a$, we obtain

$$\Delta g = -\left.\left[\partial_r \Delta V + \frac{2}{a}\Delta V\right]\right|_{r=a}. \qquad (4.3.8)$$

You know the left hand side from gravity and leveling observations. You then solve Equation 4.3.8 for $\Delta V|_{r=a}$. You use the result in $N = (\Delta V/g_T)|_a$ to get the geoid anomaly, N.

So, given δg, how do you solve Equation 4.3.8 for ΔV? You start by noting that Δg is a function of θ and ϕ, and so can be expanded as:

$$\Delta g = \sum_{l,m} g_l^m Y_l^m(\theta, \phi) \tag{4.3.9}$$

where the g_l^m are constants. ΔV is a function of r, θ, and ϕ, and so has the expansion (see Equation 4.3.1):

$$\Delta V = -\frac{GM}{r} \sum_{l,m} V_l^m \left(\frac{a}{r}\right)^l Y_l^m(\theta, \phi). \tag{4.3.10}$$

In both expansions (Equations 4.3.9 and 4.3.10), the sum over l starts at $l = 2$, and the $l = 2$, $m = 0$ terms are missing. Then, using these expansions in Equation 4.3.8 gives:

$$\begin{aligned} \sum g_l^m Y_l^m &= \frac{GM}{a} \sum_{l,m} V_l^m Y_l^m \left[-\frac{l+1}{a} + \frac{2}{a} \right] \\ &= \frac{GM}{a^2} \sum_{l,m} V_l^m (1-l) Y_l^m. \end{aligned}$$

So, equating coefficients gives:

$$g_l^m = \underbrace{\frac{GM}{a^2}}_{g_T|_a}(1-l) V_l^m \tag{4.3.11}$$

Or:

$$V_l^m = \frac{g_l^m}{(1-l) g_T|_a}. \tag{4.3.12}$$

We can relate g_l^m directly to the geoid, by expanding

$$N = \sum_{l,m} N_l^m Y_l^m(\theta, \phi)$$

with N_l^m = constants. Then

$$N = \left.\frac{\Delta V}{g_T}\right|_{r=a} \quad \Rightarrow \quad N_l^m = \frac{\frac{-GM}{a} V_l^m}{g_T} = -a V_l^m. \tag{4.3.13}$$

So:

$$N_l^m = -\left[\frac{a}{g_T|_{r=a}}\right] \left[\frac{g_l^m}{1-l}\right].$$

So if we had the g_l^m coefficients from surface observations, we could easily find the N_l^m, and so find N.

Note, incidentally, that short wavelength features are less prominent in the geoid than they are in gravity. You can deduce this from the $(1-l)$ in the denominator of N_l^m. For short horizontal wavelengths, l is large, and so N_l^m is proportionally smaller than g_l^m.

One consequence of this result (i.e. that the geoid is "smoother" than gravity) is that surface gravity observations are not as useful for finding the geoid as are satellite observations. Surface gravity gives good short wavelength gravity results, but poor long wavelength components. And, the geoid is more sensitive to the long wavelength (small l) terms. Satellites, on the other hand, are *best* at providing the longer wavelength terms in V_T.

Nevertheless, let's continue this discussion of how you obtain N from surface gravity observations. That's all people *could* do before satellites. And, in some situations, people still do it that way. Only, they don't do it by finding the g_l^m coefficients. Instead, they use Stoke's formula. I'll just outline the derivation.

We have seen that

$$N = -\frac{a}{g_T|_a} \sum_{l,m} \frac{g_l^m}{1-l} Y_l^m(\theta, \phi). \tag{4.3.14}$$

The g_l^m, defined in Equation 4.3.9, are given by:

$$g_l^m = \int \Delta g(\theta', \phi') Y_l^{m*}(\theta', \phi') \sin\theta' \, d\theta' \, d\phi'.$$

We substitute this integral into Equation 4.3.14, to obtain:

$$N = -\frac{a}{g_T|_a} \int \Delta g(\theta', \phi') \left[\sum_l \left(\frac{1}{1-l} \right) \sum_m \left[Y_l^m(\theta, \phi) Y_l^{m*}(\theta', \phi') \right] \right] \sin\theta' \, d\theta' \, d\phi'. \tag{4.3.15}$$

The addition theorem says that the sum over m in Equation 4.3.15 is

$$\sum_m \left[Y_l^m(\theta, \phi) Y_l^{m*}(\theta', \phi') \right] = \frac{2l+1}{4\pi} P_l(\cos\gamma)$$

where $\gamma =$ angle between (θ, ϕ) and (θ', ϕ'). So:

$$N = -\frac{a}{4\pi g_T|_a} \int \Delta g(\theta', \phi') \left[\sum_l \left(\frac{2l+1}{1-l} \right) P_l(\cos\gamma) \right] \sin\theta' \, d\theta' \, d\phi'.$$

You can do the sum over l. Define this sum as $f(\gamma)$, so that:

$$f(\gamma) = \frac{1}{2} \sum_l \left(\frac{2l+1}{1-l} \right) P_l(\cos\gamma).$$

It turns out [see Garland, "Introduction to Geophysics," W.B. Saunders Co., 1979, p161]

$$f(\gamma) = \left[\frac{1}{2}\csc\frac{\gamma}{2} - 1 - \cos\gamma\right] + 3\left[1 - \cos\gamma - 2\sin\frac{\gamma}{2} - \cos\gamma\ln\left[\sin\frac{\gamma}{2} + \sin^2\frac{\gamma}{2}\right]\right].$$

So:

$$N = -\frac{a}{2\pi g_T|_a}\int \Delta g(\theta', \phi') f(\gamma)\sin\theta'\, d\theta'\, d\phi'. \qquad (4.3.16)$$

Equation 4.3.16 is Stoke's formula. You determine δg from observations. You then integrate Equation 4.3.16 to get N. This doesn't really get around the problem of not knowing the long wavelength g_l^m very well from surface observations. The reason you don't know them well is because there are many areas of the globe where you don't have good surface gravity observations: over the oceans, for example. It turns out that $f(\gamma)$ doesn't go to 0 very quickly as $\gamma \to \pi$. That is, there will be significant contributions to the integral in Equation 4.3.16 from δg all over the globe. So, areas of sparse data are still a problem when using Stoke's formula.

4.4 Satellite Geoids

It is much easier nowadays to determine the global geoid from satellite observations than from gravity measurements on the ground. From Equation 4.3.13 we know that the geoid is

$$N = -a\sum_{l,m} V_l^m Y_l^m, \qquad (4.4.1)$$

where the V_l^m are the potential coefficients defined in Equation 4.3.10 (and in Equation 4.3.1). Satellite ranging gives the V_l^m, by fitting to the satellite's orbital motion. And the V_l^m give N directly, using Equation 4.4.1. (Actually, satellite geodesists usually use a different normalization for their spherical harmonics, so that their coefficients, written as C_{lm} and S_{lm}, are proportional to our V_l^m's.)

We'll discuss the observed geoid and its geophysical interpretation, later. But as a preview: people have now demonstrated pretty convincingly that the long wavelength geoid (l from 2 to about 10) is caused by density anomalies associated with convection, *and* by convection-driven perturbations of surfaces that have density discontinuities across them (e.g. the outer surface; the core-mantle-boundary).

Chapter 5

Stress/Strain Relations

Observational results for surface gravity and for the shape of the geoid tell geophysicists about the internal density of the earth. Sometimes that density is what the geophysicist is after. But often the geophysicist wants to learn about the mechanism causing the density anomaly. For that sort of application you need to know something about how the earth responds to external and internal forcing. That means you have to know something about stress and strain inside a solid and how to incorporate stress and strain into the equations of motion for the earth. That's what this chapter is about.

First, we'll talk about stress and strain inside a solid and how they are related. Here is a very crude qualitative description:

Suppose you have a block of solid material. You deform it somehow. There are internal forces in the block which resist the deformation, so that when you let go of the block it tries to spring back to its original shape (unless you've deformed it too much). Think of the deformation as the strain. Think of the restoring forces as the stress. Somehow we must relate stress to strain. That is, can we infer the restoring forces if we know the deformation? (Alternatively, if we know the "restoring" forces, can we model the deformation?) For the earth (and for most materials), as long as the deformation is small, the stress will be nearly linearly proportional to the strain. That's a generalization of Hooke's Law for a one-dimensional spring: $F = kx$. Only the stress/strain relation looks a lot more complicated than this. That's because:

1. You can't describe deformation or internal forces with simple scalars like F and x. Both stress and strain turn out to be tensors.

2. A solid cannot be easily modeled as a set of discrete objects, like springs. It is better modeled as a continuum. That means we must deal with *continuum* mechanics — rather than with the sort of discrete mechanics you learn about in most physics mechanics courses. To develop the equations of continuum mechanics you break the medium up into small, discrete blocks and look at the forces on each block, but then you let the block size go to zero.

Now, let's get into the details. First, I'll define the stress tensor, and show how a solid deforms under certain stresses. Then, I'll define the strain tensor, and write down the general relationship between stress and strain.

After that, I'll describe: incorporating stress and strain into the equations of motion; waves in a solid; anelasticity.

5.1 Stress tensor

A stress tensor describes internal forces in an object. Consider a small volume from the interior of a solid body, as in Figure 5.1. What external forces act on this block? There

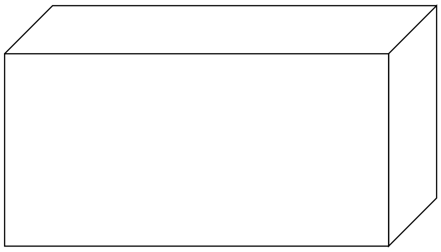

Figure 5.1:

are "body forces" — that is, forces that originate from an object outside the block and which act on every molecule within the block. For the earth, the most pertinent of these would be gravity. But, electro-magnetic (EM) forces also fall into this category.

There are also "surface forces," which are the forces that act on the surfaces of the block from the surrounding molecules — through atomic and molecular bonding. (In a fluid, the surface force is pressure.) Those forces are primarily between neighbors and next-nearest neighbors. The surface force per unit area of the surface is called the stress. Stresses are really internal atomic and molecular forces (EM bonding) — but it is easiest to model them as acting between and across surfaces, even though there are probably not well defined surfaces inside the solid.

Pick, say, the $y =$ constant face of the block. See Figure 5.2. The net force acting

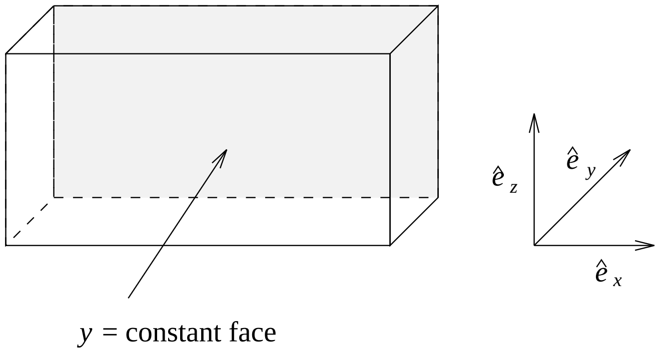

Figure 5.2:

on that face from the atoms just outside the block is a vector, and so has $\hat{e}_x$, $\hat{e}_y$, $\hat{e}_z$ components: F_x, F_y, F_z. The stress tensor, $\overline{\tau}$, is defined so that its elements are the net force components divided by the area of the face. At least, they are defined that way when the area of the face is infinitesimal. So:

$$\begin{aligned}
\tau_{yx} &= \frac{\hat{e}_x \text{ component of force on } y = \text{ constant face}}{\text{area of } y = \text{ constant face}} \\
\tau_{yy} &= \frac{\hat{e}_y \text{ component of force on } y = \text{ constant face}}{\text{area of } y = \text{ constant face}} \\
\tau_{yz} &= \frac{\hat{e}_z \text{ component of force on } y = \text{ constant face}}{\text{area of } y = \text{ constant face}}
\end{aligned}$$

Similarly, you can define τ_{xx}, τ_{xy}, τ_{xz}, τ_{zx}, τ_{zy}, and τ_{zz}. The first index describes the orientation of the face, and the second describes the component of the force. Since the block is infinitesimal, the nine numbers really represent surface forces at a *point*, $\mathcal{P}$, in the medium.

You organize these nine numbers into a tensor:

$$\overleftrightarrow{\tau} = \begin{pmatrix} \tau_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \tau_{yy} & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \tau_{zz} \end{pmatrix}$$

Don't worry about what a tensor is. Just think of it as a matrix. It has the nice property that if we pick any infinitesimal surface passing through the point $\mathcal{P}$, and if $\hat{n}$ is the normal to that surface, then the (force/area) acting across the surface is: $(\overline{F}/\text{area}) = \hat{n}\cdot \overleftrightarrow{\tau}$, where this dot product represents vector/matrix multiplication:

$$\hat{n}\cdot \overleftrightarrow{\tau} = \hat{e}_i \sum_{j=1}^{3} n_j \tau_{ji}$$

where $\hat{e}_1 = \hat{e}_x$, $\hat{e}_2 = \hat{e}_y$, and $\hat{e}_3 = \hat{e}_z$. (We need to be more careful about the signs. If one side of the surface is **1** and the other side **2**, and if $\hat{n}$ goes from **1** into **2**, then $\hat{n}\cdot \overleftrightarrow{\tau}$ is the force/area on **1** from **2**.) The nice thing is that this result for the force doesn't require the surface to be parallel to the xy, xz, or yz planes.

It turns out that the stress tensor is symmetric. That is: $\tau_{ij} = \tau_{ji}$. To show this, consider the very small internal block in Figure 5.3, where dx, dy, and dz are infinitesimal.

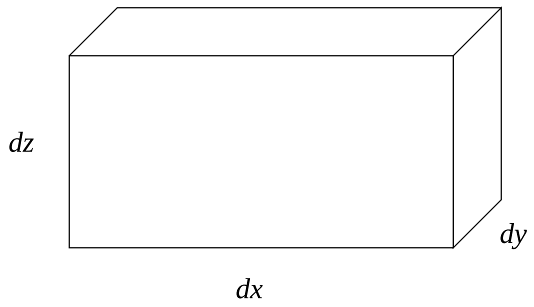

Figure 5.3:

Newton's Second Law for the block as a whole is:

$$m\overline{a} = \overline{F}_{\text{surface}} + \overline{F}_{\text{body}} \tag{5.1.1}$$

where m and $\overline{a}$ are the mass and acceleration of the block, $\overline{F}_{\text{surface}}$ is the sum of all surface forces acting on the block, and $\overline{F}_{\text{body}}$ is the sum of all body forces acting on the cube.

$$\overline{F}_{\text{body}} \quad = \quad (\text{body force density}) \times \text{volume}$$

$$
\begin{aligned}
m &= (\text{mass density}) \times \text{volume} \\
\overline{F}_{\text{surface}} &= (\text{surface stress}) \times \text{area}
\end{aligned} \tag{5.1.2}
$$

The volume is proportional to $(dx\,dy\,dz)$. The area is proportional to $(dx\,dy)$ or $(dx\,dz)$ or $(dy\,dz)$, depending on which face we are considering. So, for infinitesimal dx, dy, and dz, the volume $\ll$ area. So, for a small block, Equations 5.1.1 and 5.1.2 imply that $\overline{F}_{\text{surface}} = 0$ to lowest order in dx, dy and dz. You can go through the same reasoning to conclude that the total torque on the infinitesimal cube due to surface forces (stresses) is zero, to lowest order.

These results imply that $\overleftrightarrow{\tau}$ is symmetric (actually you only need the zero torque result). Consider, for simplicity, a two-dimensional stress field, where $\tau_{yi} = \tau_{iy} = 0$ for all i. See Figure 5.4.

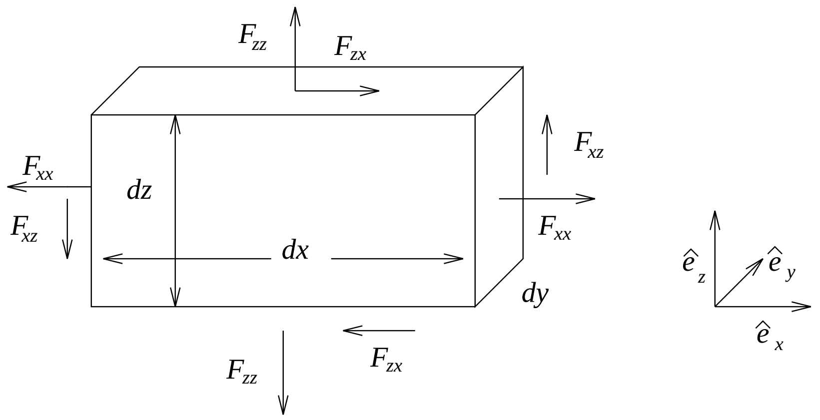

Figure 5.4:

The F_{ij} represent the force on a face. For example:

$$
\begin{aligned}
F_{xx} &= \tau_{xx}\,dy\,dz \qquad & F_{zz} &= \tau_{zz}\,dx\,dy \\
F_{xz} &= \tau_{xz}\,dy\,dz \qquad & F_{zx} &= \tau_{zx}\,dx\,dy
\end{aligned} \tag{5.1.3}
$$

Note that I have labeled these forces so that the forces on opposite faces are equal. In other words, F_{xx} on the right hand face $=$ F_{xx} on the left hand face, and similarly for F_{xz}, F_{zz}, and F_{zx}. Why can I do that? Well, really F_{xx} on the left hand face should be $\tau_{xx}(x)\,dy\,dz$ and F_{xx} on the right hand face should be $\tau_{xx}(x+dx)\,dy\,dz$. That is, the forces are different because τ_{xx} is a function of x. But: $\tau_{xx}(x+dx) \approx \tau_{xx}(x) + dx\,\partial_x \tau_{xx}$. And, to zeroth order in dx (remember — "no surface torques" is only valid to zeroth order), $dx\,\partial_x \tau_{xx}$ is negligible. So $\tau_{xx}(x+dx) \approx \tau_{xx}(x)$, to zeroth order. Thus, F_{xx} is the same on both faces. Similarly for the other F_{ij}'s. This means that the requirement that there be no net force on the cube is automatically satisfied. The condition of no net torque means:

$$\frac{dz}{2}\,2F_{zx}\ (=\text{clockwise torque}) = \frac{dx}{2}\,2F_{xz}\ (=\text{counterclockwise torque}).$$

Thus, $dz\,F_{zx} = F_{xz}\,dx$. Combining this result with Equation 5.1.3 gives $\tau_{zx} = \tau_{xz}$. You can go through a similar argument in the full three-dimensional case to conclude that $\tau_{ij} = \tau_{ji}$ for all i, j.

So, there are really only six independent stress components. I'll label them as $\tau_{11}, \tau_{12}, \ldots$, instead of $\tau_{xx}, \tau_{xy}, \ldots$, so that:

$$\overset{\leftrightarrow}{\tau} = \begin{pmatrix} \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{12} & \tau_{22} & \tau_{23} \\ \tau_{13} & \tau_{23} & \tau_{33} \end{pmatrix}$$

5.2 Stress-Induced Deformation

So, now we know something about stresses and how they relate to forces. How do stresses deform material? For the moment we'll only worry about the deformation induced by the diagonal terms in τ: $\tau_{11}, \tau_{22}, \tau_{33}$.

Assume, first, that τ_{11} is the only component not equal to zero. Start with the infinitesimal block shown in Figure 5.5, which has sides of length dx, dy, dz. We apply the forces F_{11}, as shown, where $F_{11} = \tau_{11}\,dy\,dz$. The block deforms. One thing that happens is that the width of the block lengthens by Δx. It is like stretching a spring. Hooke's Law says that the restoring force on a spring is linearly proportional to the amount the spring has stretched. It works well, so long as the spring is elastic and is not stretched too much.

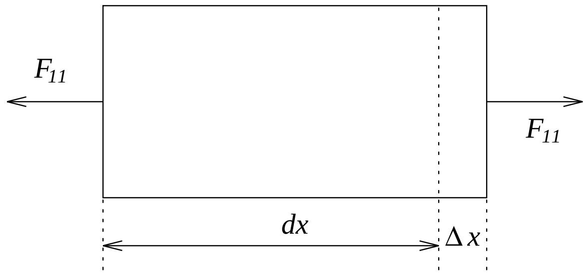

Figure 5.5:

Let's assume the solid in Figure 5.5 is elastic and is not deformed too much. This assumption means that the block returns to its initial state immediately after the forces are removed. The elastic assumption is pretty good for the earth at short time periods (for example, at seismic periods of minutes or less). It is not as good at very long time periods (thousands to millions of years). We'll extend our results to include anelasticity later.

For an elastic, slightly deformed solid, we might expect to find a generalization of Hooke's Law:

$$F_{11} = k\Delta x. \tag{5.2.1}$$

It turns out that this generalization works pretty well for small Δx. The constant k depends on the material *and* on the size and shape of the block. Let's try to find the dependence on the size and shape.

First, how does k depend on the unstretched length, dx? Imagine cutting the block in half, as in Figure 5.6. Consider, say, the left half. The right half exerts the force F_{11}

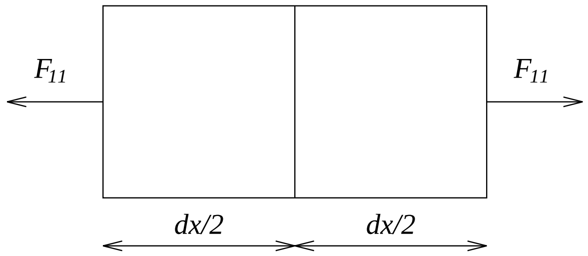

Figure 5.6:

on the left half (see Figure 5.7). We know this must be true, because the net force on the infinitesimal half-block vanishes (to lowest order in the size of the block). By symmetry,

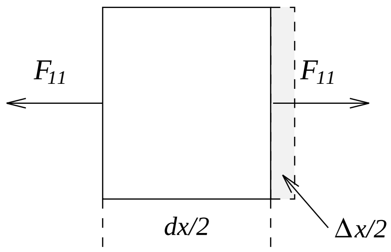

Figure 5.7:

the left half must absorb half the increase in length. So, the left half stretches by the amount $\Delta x/2$. So, for this half block:

$$F_{11} = k_{\frac{1}{2}} \left(\frac{\Delta x}{2}\right)$$

where, $k_{1/2}$ represents the value of k for the half-block. By equating the F_{11} equations for the full block and the half block, we note that $k_{1/2} = 2k$. So, when we divide the block by 2 we double k. We can obviously extend this result by dividing the original block into smaller blocks of any size. We conclude that k is inversely proportional to the length. That is, k is proportional to $1/dx$. Define k' so that $k = k'/dx$. Using this in Equation 5.2.1 gives

$$F_{11} = k' \frac{\Delta x}{dx} \tag{5.2.2}$$

Next, how does k' depend on the cross-sectional area: $dy\,dz$ in this case? Go back to the original block and cut it in half lengthwise. See Figure 5.8. The cross-sectional area

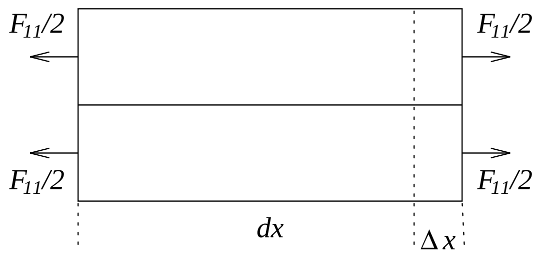

Figure 5.8:

on one face of the half-block is now $(1/2)\,dy\,dz$, instead of $dy\,dz$. By symmetry, the force on a half-block face should be $F_{11}/2$, as shown.

The stretching of each-block is still Δx, and the unstretched length is dx. Hooke's Law (Equation 5.2.2) applied to a half-block is now

$$\frac{1}{2}F_{11} = k'_{\frac{1}{2}}\frac{\Delta x}{dx} \tag{5.2.3}$$

where $k'_{1/2} = k'$ for the half-block. Comparing Equations 5.2.2 and 5.2.3 gives: $k'_{1/2} = k'/2$. So, reducing the area by 1/2, reduces k' by 1/2, implying that k' must be proportional to $dy\,dz$. Define $E \equiv$ Young's modulus, so that $k' = E\,dy\,dz$. Then Equation 5.2.2 becomes:

$$F_{11} = E\,dy\,dz\frac{\Delta x}{dx}.$$

Or:

$$\tau_{11} = \frac{F_{11}}{dy\,dz} = E\frac{\Delta x}{dx}.$$

Define $\epsilon_{11} \equiv \Delta x/dx =$ (change in length)/(length). The reason for the two subscripts on ϵ will be clear later. Then:

$$\tau_{11} = E\epsilon_{11}. \tag{5.2.4}$$

Equation 5.2.4 is a generalized Hooke's Law for solids. E depends on the material the block is made out of, but not on its size or shape. $E \geq 0$ for reasonable materials. Otherwise, when you pull on an object it would contract.

The forces F_{11} will change the other dimensions, as well. For example, (see Figure 5.9)

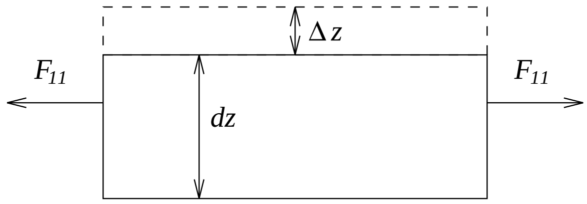

Figure 5.9:

assume the height before deformation is dz, and that when you apply F_{11} on each side, the height increases by Δz (Δz will probably be negative). You can go through an argument similar to the one above for Δx — using, again, a generalization of Hooke's Law — to get a relation similar to Equation 5.2.4:

$$\tau_{11} = -\left(\frac{E}{\nu}\right)\frac{\Delta z}{dz} = -\left(\frac{E}{\nu}\right)\epsilon_{33}$$

where $\epsilon_{33} \equiv \Delta z/dz =$ (change in height)/(height). We write the proportionality constant here as $(-E/\nu)$, instead of using a single variable, to conform with convention. ν is called Poisson's ratio and is dimensionless. (E, here, is the same Young's modulus that appears in Equation 5.2.4.) The minus sign is present because for most materials $\nu \geq 0$. (If you pull on the block, its height decreases.) Note that this result for ϵ_{33} can be combined with Equation 5.2.4 to give:

$$\epsilon_{33} = -\nu\epsilon_{11}$$

You can do something similar for $\Delta y/dy \equiv \epsilon_{22}$ to find

$$\tau_{11} = -(E/\nu)\epsilon_{22} \qquad \text{and} \qquad \epsilon_{22} = -\nu\epsilon_{11}.$$

For isotropic materials — and we will always assume isotropic materials here (isotropic means that the material properties are independent of the orientation of the applied stresses) — the ν's in the ϵ_{22} and ϵ_{33} equations are the same.

It turns out that for "reasonable" materials, $\nu \leq 1/2$. To see this, suppose we apply stresses $\tau_{11} = \tau_{22} = \tau_{33} = -\mathbf{P}$, and that $\tau_{ij} = 0$ for $i \neq j$. The symbol $\mathbf{P}$ is appropriate, because if the τ_{ii}'s are equal, and $\tau_{ij} = 0$ for $i \neq j$, then the τ_{ii}'s are the negative of the pressure (a positive τ_{ii} means an *outward* force on an object — but pressure is *inward*). Each of the τ_{ii} perturb the three sides dx, dy, dz — increasing them by Δx, Δy, Δz. From the results above we conclude that for an isotropic material:

$$\begin{aligned}
\Delta x &= \tau_{11}\frac{1}{E}\,dx - \tau_{22}\frac{\nu}{E}\,dx - \tau_{33}\frac{\nu}{E}\,dx \\
\Delta y &= -\tau_{11}\frac{\nu}{E}\,dy + \tau_{22}\frac{1}{E}\,dy - \tau_{33}\frac{\nu}{E}\,dy \\
\Delta z &= -\tau_{11}\frac{\nu}{E}\,dz - \tau_{22}\frac{\nu}{E}\,dz + \tau_{33}\frac{1}{E}\,dz.
\end{aligned}$$

(Without the isotropic assumption, the ν's and E's above might all be different.) So:

$$\begin{aligned}
\Delta x &= (\tau_{11} - \nu\tau_{22} - \nu\tau_{33})\,\frac{dx}{E} = -\mathbf{P}\,(1-2\nu)\,\frac{dx}{E} \\
\Delta y &= (\tau_{22} - \nu\tau_{11} - \nu\tau_{33})\,\frac{dy}{E} = -\mathbf{P}\,(1-2\nu)\,\frac{dy}{E} \\
\Delta z &= (\tau_{33} - \nu\tau_{11} - \nu\tau_{22})\,\frac{dz}{E} = -\mathbf{P}\,(1-2\nu)\,\frac{dz}{E}.
\end{aligned}$$

So, the new volume of the solid is:

$$(dx + \Delta x)(dy + \Delta y)(dz + \Delta z) = dx\,dy\,dz \left[\left(1 + \frac{\Delta x}{dx}\right) \left(1 + \frac{\Delta y}{dy}\right) \left(1 + \frac{\Delta z}{dz}\right) \right].$$

To lowest order in the deformation ($\Delta x/dx$, $\Delta y/dy$, $\Delta z/dz$ are small), this new volume is:

$$dx\,dy\,dz \left[1 + \frac{\Delta x}{dx} + \frac{\Delta y}{dy} + \frac{\Delta z}{dz}\right] = dx\,dy\,dz \left[1 - 3(1-2\nu)\frac{\mathbf{P}}{E}\right] \tag{5.2.5}$$

The original volume was $dx\,dy\,dz$. So, the *change* in volume is Equation 5.2.5 minus $dx\,dy\,dz$. Thus, the change in volume ($\equiv \Delta V$), divided by the original volume ($\equiv dV$) is:

$$\frac{\Delta V}{dV} = -\left(\frac{3(1-2\nu)}{E}\right)\mathbf{P}.$$

If $\mathcal{K} \equiv$ bulk modulus $\equiv E/3(1-2\nu)$ then

$$\mathbf{P} = -\mathcal{K}\frac{\Delta V}{dV} \tag{5.2.6}$$

Since $E > 0$, then $\nu > 1/2$ would imply that $\mathcal{K} < 0$. That would mean that if you squeezed the block ($\mathbf{P} > 0$), the volume would increase. That's unphysical, and so $0 < \nu \leq 1/2$.

5.3 Response to τ_{ij} for $i \neq j$

The section above shows that we can describe the response of a solid, isotropic block to τ_{11}, τ_{22}, and τ_{33}, using two independent constants: either (E and ν) or ($\mathcal{K}$ and ν) (or (E and $\mathcal{K}$), but that combination is rarely used).

I claim that we can describe the response to the other τ_{ij} ($i \neq j$) with the same two elastic constants. The reason for this is that for any stress field you can come up with, there is a coordinate system in which $\tau_{ij} = 0$ for $i \neq j$. Another way to put that is that because $\overleftrightarrow{\tau}$ is symmetric, we can always find a coordinate system where $\overleftrightarrow{\tau}$ is diagonal: i.e. where only τ_{11}, τ_{22}, and $\tau_{33} \neq 0$. And, the response to the τ_{ii}'s can be described with just those two constants.

Let's try to see, physically, why any stress field looks like a normal stress field in some coordinate system. Suppose, for example, that we choose an infinitesimal cube with sides along the $\hat{e}_x$, $\hat{e}_y$, and $\hat{e}_z$ axes, as shown in Figure 5.10, and we find that the diagonal elements of the stress tensor are zero (i.e. $\tau_{ii} = 0$ for every i). In fact, suppose,

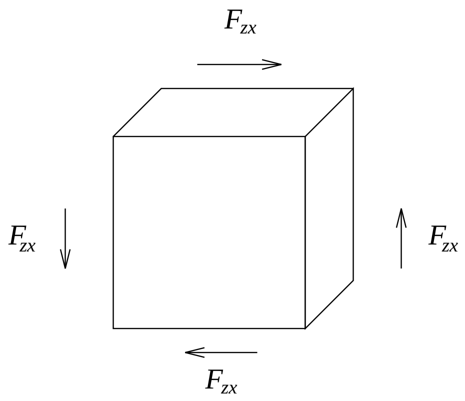

Figure 5.10:

for simplicity, that all stress components with a y-index are 0, so that the only surface forces acting on the cube are those shown in Figure 5.10. Each of the four force arrows in Figure 5.10 has the same length (F_{zx}), because the stress tensor is symmetric and because each face is assumed to have the same area. We bisect this cube along the upper left-lower right diagonal, so that the volume above the diagonal is as shown in Figure 5.11.

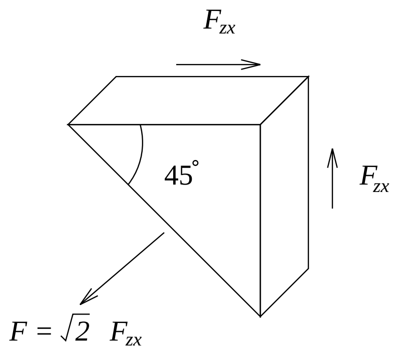

Figure 5.11:

The arrow pointing away from the diagonal in Figure 5.11 is the force on the diagonal face of the half-cube, and it is normal to this face (with length $\sqrt{2}F_{zx}$). That's because the sum of all the surface forces acting on an infinitesimal object must vanish, to lowest order in the dimensions of the object (we saw that, earlier). Similarly, if we bisect the Figure 5.10 cube between the upper right and lower left corners, we find that the force on *that* diagonal face is normal to the face. See Figure 5.12.

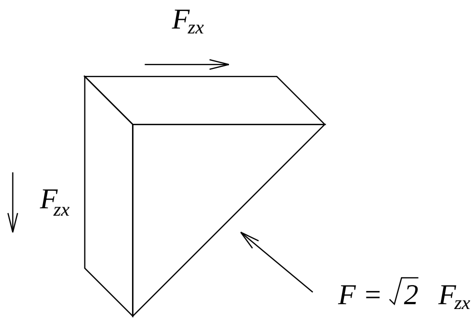

Figure 5.12:

So, we have found two mutually perpendicular planes where the surface forces are normal to the surfaces. That means, we didn't cut our infinitesimal cube in a clever way. We should have cut out a cube oriented as in Figure 5.13. Then, the surface forces would

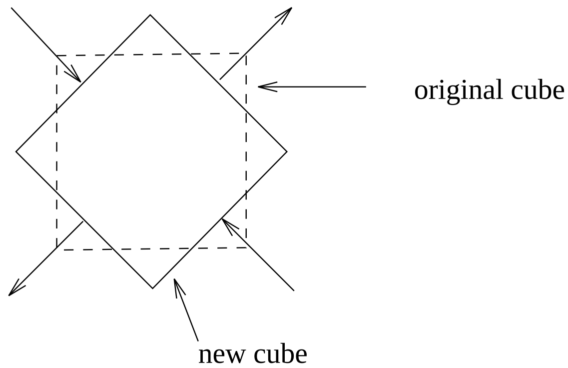

Figure 5.13:

be normal to the new cube. And, if our local $\hat{e}_x$, $\hat{e}_y$, $\hat{e}_z$ axes are re-defined to be parallel to the faces of the new cube, $\overleftrightarrow{\tau}$ would be diagonal in the new system.

The above argument can be extended to a more general three-dimensional case, where the τ_{ii} are not zero, only it's harder. It is easier to simply trust in the result from linear algebra, that says that any symmetric matrix can be diagonalized.

In the diagonal system:

$$\overleftrightarrow{\tau} = \begin{pmatrix} \tau_{11} & 0 & 0 \\ 0 & \tau_{22} & 0 \\ 0 & 0 & \tau_{33} \end{pmatrix}.$$

The τ_{ii} may not be equal. If they were equal, then the internal stress field would be equivalent to a pure pressure field: i.e. τ would have the form $\overleftrightarrow{\tau} = -\mathbf{P}\overleftrightarrow{I}$ where $\tau_{11} = \tau_{22} = \tau_{33} = -\mathbf{P}$, and $\overleftrightarrow{I}$ is the identity matrix (tensor). But, the identity tensor is unaffected by rotations into new coordinate systems. So, $\overleftrightarrow{\tau} = -\mathbf{P}\overleftrightarrow{I}$ in *every* system.

If the τ_{ii} are *not* equal in the diagonal system, then there will be shear stresses ($\tau_{ij} \neq 0$ for $i \neq j$) in other systems. But, no matter what the system, $[\tau_{11} + \tau_{22} + \tau_{33}]$ is always the same — this is a property of matrices (the sum of the diagonal components is the same in every system). So, for any stress tensor we define the *pressure* as

$$\mathbf{P} = -\frac{1}{3}\left[\tau_{11} + \tau_{22} + \tau_{33}\right].$$

Then, we think of $\overleftrightarrow{\tau}$ as consisting of pressure ($-\mathbf{P}\overleftrightarrow{I}$) plus a remainder. The remainder is called the deviatoric stress, and is written as $\overleftrightarrow{\delta\tau}$:

$$\overleftrightarrow{\delta\tau} = \overleftrightarrow{\tau} + \mathbf{P}\overleftrightarrow{I}.$$

All shear stresses are described by $\overleftrightarrow{\delta\tau}$. For a fluid, $\overleftrightarrow{\delta\tau} = 0$.

To find the deformation produced by an arbitrary stress tensor $\overleftrightarrow{\tau}$ we can proceed in three steps:

The Procedure:

1. Transform the stress tensor to the diagonal ("principal-axis") system.

2. Find the deformation (Δx, Δy, Δz) in that diagonal system, using (E, ν) or $(\mathcal{K}, \nu)$, and the results of the preceding section.

3. Transform the deformation field back into the original system.

Of course, we don't want to have to go through this procedure every time we are given a new stress tensor. Instead, we want to work the method through just *once* for an arbitrary stress tensor, to obtain a general relation between stress and deformation that can be easily applied to any situation. The algebra is messy. It involves finding the diagonalizing rotation matrix for an arbitrary stress tensor, taking its inverse, multiplying the matrices together, etc. I'm not going to go through it all here, but will simply write down the answer. But, I can't do that quite yet, because I haven't defined enough quantities to describe deformation in an arbitrary system. All I've done, so far, is to describe the quantities $\Delta x/dx$, $\Delta y/dy$, and $\Delta z/dz$, which describe what happens to line elements when they are stretched outward or inward, along their lengths.

To describe deformation in the more general case, I need to define a matrix, called the strain tensor. The three numbers $\Delta x/dx$, $\Delta y/dy$, and $\Delta z/dz$, turn out to be the diagonal elements of that tensor.

5.3.1 Strain Tensor

It turns out that you need nine numbers to describe an arbitrary displacement of the material in a small block. Suppose you have two points in a solid, initially both on the $\hat{e}_x$ axis a small distance dx apart, as shown in Figure 5.14. The solid is then deformed,

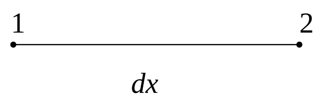

Figure 5.14:

so that the right hand point moves up along the $\hat{e}_z$ axis and out along the $\hat{e}_x$ axis, as shown in Figure 5.15. We can describe this deformation with two numbers: Δz and Δx.

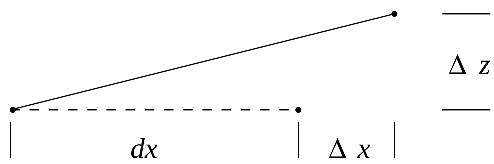

Figure 5.15:

If there is also a displacement in the $\hat{e}_y$ direction, then we need a third number: Δy. In

Section 5.2, this dx line element only stretched in the $\hat{e}_z$ direction, so we only needed to introduce the quantity Δx to describe it. Here, we need three quantities. Only, let's use ratios with respect to dx for our three numbers:

$$\frac{\Delta x}{dx}, \qquad \frac{\Delta y}{dx}, \qquad \frac{\Delta z}{dx}.$$

Similarly, if two points separated along the $\hat{e}_z$ direction are displaced by Δx, Δy, Δz, we need the numbers:

$$\frac{\Delta x}{dz}, \qquad \frac{\Delta y}{dz}, \qquad \frac{\Delta z}{dz}$$

to describe the deformation of the line between those two points. ($dz =$ unperturbed line length.) And:

$$\frac{\Delta x}{dy}, \qquad \frac{\Delta y}{dy}, \qquad \frac{\Delta z}{dy}$$

are needed to describe the deformation of a line along the $\hat{e}_y$ axis. (Note: Δx, for example, is different for each of the three lines.)

So, it looks like we need nine numbers to describe deformation: we must specify what happens to the original $\hat{e}_x$, $\hat{e}_y$, and $\hat{e}_z$ axes at any point — and it takes three numbers to describe the deformation of each of these three axes. This suggests that a reasonable definition of the strain tensor ought to be the matrix with elements $\Delta x/dz$, etc. For example, the (ij)th element would be $\Delta r_i/dr_j$ (where $dr_1 = dx$, $dr_2 = dy$, etc.).

Well, we could certainly construct a 3×3 matrix this way, but it turns out it wouldn't be the "right" matrix. What happens is that when we go through "The Procedure" described above, we find that only six independent, linear combinations of the nine numbers appear in the final result. Three independent, linear combinations do not appear. The three combinations that do not appear represent rotations of infinitesimal blocks.

Here's how these linear combinations are defined: The quantities $\Delta x/dz$, etc., can be affected either by deformation of the infinitesimal block, or by rotation. By "deformation" I mean displacements which can be caused by stress on the block. We have seen that you can always find a coordinate frame where the stress tensor is diagonal. And, in that frame, the displacements represent pure stretching. That is, $\Delta x/dx$, $\Delta y/dy$, and $\Delta z/dz$ are, in general, non-zero, but all other $\Delta r_i/dr_j = 0$. So, by "deformation," I mean values of $\Delta x/dx$, $\Delta x/dy$, etc., which when transformed to some coordinate system, represent pure stretching. I claim that means that deformational displacements must satisfy $\Delta z/dx = \Delta x/dz$, $\Delta z/dy = \Delta y/dz$, and $\Delta y/dx = \Delta x/dy$.

There are two ways to see this. One is that for deformation, the matrix of $\Delta x/dx$, $\Delta y/dx$, etc., can be diagonalized by a rotation (a unitary transformation). That means the matrix must be symmetric.

The other way is to define two matrices ϵ_{ij} and ω_{ij} so that

$$\epsilon_{13} = \frac{1}{2}\left[\frac{\Delta x}{dz} + \frac{\Delta z}{dx}\right], \qquad \omega_{13} = \frac{1}{2}\left[\frac{\Delta x}{dz} - \frac{\Delta z}{dx}\right]$$

etc., for the other ϵ_{ij}, ω_{ij} (note: $\omega_{ii} = 0$). Then, think of the elements $\Delta x/dx$, $\Delta x/dy$, etc., as linear combinations of the ϵ's and ω's. For example:

$$\frac{\Delta x}{dz} = \epsilon_{13} + \omega_{13} \qquad \frac{\Delta z}{dx} = \epsilon_{13} - \omega_{13}.$$

I claim that the ω_{ij} describe rotations — *not* deformation. For example, suppose $\omega_{13} \neq 0$, but — for simplicity — suppose $\epsilon_{13} = 0$. Then

$$\frac{\Delta x}{dz} = \omega_{13} = -\frac{\Delta z}{dx}. \tag{5.3.1}$$

Consider the central point, O, of a cube, as shown in Figure 5.16. Consider the four line segments connected to O with endpoints at dx, $-dx$, dz, and $-dz$. The arrows in

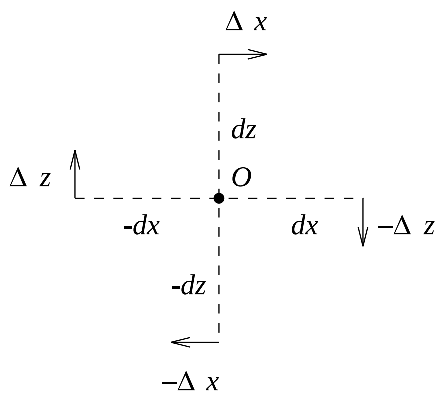

Figure 5.16:

Figure 5.16 represent the displacements corresponding to Equation 5.3.1. They describe a *rotation*, not a deformation. They look like a rotation in *every* frame — never a "stretching."

So, for deformation $\omega_{13} = 0 \Rightarrow \Delta x/dz = \Delta z/dx$. Similarly for $\Delta x/dy$, etc.

The result of all this is that the ϵ_{ij} describe the deformational contributions to $\Delta x/dx$, $\Delta x/dy$, etc. So, we define the strain tensor, $\overleftrightarrow{\epsilon}$, so that its elements are

$$\epsilon_{ij} = \frac{1}{2}\left[\Delta r_i/dr_j + \Delta r_j/dr_i\right]. \tag{5.3.2}$$

It turns out that this is a perfectly good tensor. In fact, it is symmetric, and the diagonal elements are $\epsilon_{11} = \Delta x/dx$, $\epsilon_{22} = \Delta y/dy$, $\epsilon_{33} = \Delta z/dz$.

5.3.2 Relation between stress and strain

We are now ready to write down the results of applying "The Procedure" described above, to find the deformation caused by an arbitrary, non-diagonal stress tensor. We find that in the original system the stress and strain are related to each other linearly, according to:

$$\tau_{ij} = \sum_{k,l} \Lambda_{ijkl}\, \epsilon_{kl} \tag{5.3.3}$$

where Λ is a 4th rank tensor. Think of Λ as the product of rotation matrices, with the principal-axis stress-strain relation stuck in the middle.

For an isotropic material, such as we're considering here, it just takes two parameters to characterize Λ. Those parameters could be (ν, E) or $(\nu, \mathcal{K})$. Instead, at this point, it's more convenient to define two new independent parameters — called *Lamé parameters*:

$$\begin{aligned} \mu &= \frac{E}{2(1+\nu)} \\ \lambda &= 2\mu\frac{\nu}{1-2\nu}. \end{aligned}$$

In terms of these parameters, it turns out that

$$\Lambda_{ijkl} = 2\mu\, \delta_{ik}\, \delta_{jl} + \lambda\, \delta_{ij}\, \delta_{kl} \tag{5.3.4}$$

where the δ's are Kronecker delta's. Putting Equation 5.3.4 into Equation 5.3.3 gives:

$$\tau_{ij} = 2\mu\epsilon_{ij} + \lambda\delta_{ij}\sum_{k=1}^{3}\epsilon_{kk}. \tag{5.3.5}$$

5.4 Relation between stress/strain and the displacement field

The problem with the strain tensor as defined in Equation 5.3.2, is that it involves displacements of line elements, which don't have a lot of physical significance. When solving a geophysical problem, you are usually interested in knowing the displacements of individual points inside the earth. Here we define the displacement field, and show how to represent the strain tensor in terms of the displacement components.

Suppose we have a solid object. We deform the object so that the displacement of any point originally at $\overline{x}$ is $\overline{s}(\overline{x})$. So, the point is now at $\overline{x} + \overline{s}(\overline{x})$. What is the strain tensor at that point, in terms of $\overline{s}$? Our result above for strain is in terms of displacements of infinitesimal line elements. So, how do those displacements relate to $\overline{s}$? Suppose, for example, that all displacements are in the $\hat{e}_x$ direction. So, *before* the deformation we have Figure 5.17. After the deformation (if $\overline{s}$ has only a $\hat{e}_x$ component) we have

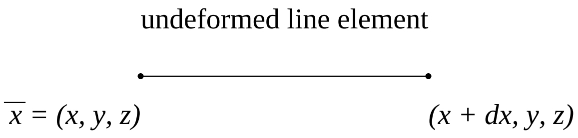

Figure 5.17:

Figure 5.18. So:

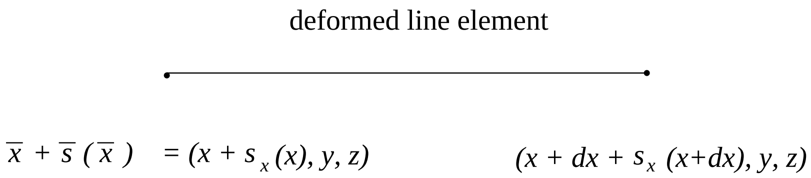

Figure 5.18:

$$\begin{aligned}
\Delta x &= [[x + dx + s_x(x + dx)] - (x + s_x(x))] - dx \\
&= [\text{deformed length}] - \text{undeformed length} \\
&= s_x(x + dx) - s_x(x).
\end{aligned}$$

So in the limit as $dx \to 0$:

$$\frac{\Delta x}{dx} \to \partial_x s_x.$$

Similarly: $\Delta z/dx \to \partial_x s_z$, $\Delta y/dx \to \partial_x s_y$, $\Delta x/dz \to \partial_z s_x$, etc. So:

$$\epsilon_{ij} = \frac{1}{2}\left[\partial_i s_j + \partial_j s_i\right]. \qquad (5.4.1)$$

Equation 5.4.1 is the strain tensor in terms of the displacement field, $\overline{s}(\overline{x})$.

Now, suppose we take this result (Equation 5.4.1) for ϵ_{ij}, and use it in the stress/strain equation (Equation 5.3.2). We obtain:

$$\overset{\leftrightarrow}{\tau} = \mu\left[\overline{\nabla}\overline{s} + \left(\overline{\nabla}\overline{s}\right)^T\right] + \lambda\overline{\nabla}\cdot\overline{s}\ \overset{\leftrightarrow}{I} \qquad (5.4.2)$$

where $\overset{\leftrightarrow}{I}$ = identity tensor (matrix), and $\overline{\nabla}\overline{s}$, $\left(\overline{\nabla}\overline{s}\right)^T$ are tensors:

$$\begin{aligned}\left(\overline{\nabla}\overline{s}\right)_{ij} &= \partial_i s_j\\ \left(\overline{\nabla}\overline{s}\right)^T_{ij} &= \partial_j s_i\end{aligned}$$

where T means transpose.

5.5 Remarks

1. μ is called the rigidity. If $\mu = 0$, then $\tau_{ij} = 0$ if $i \neq j$, and τ_{ii} is the same for all i. In other words, $\overset{\leftrightarrow}{\tau}$ is diagonal with equal diagonal elements. So, $\overset{\leftrightarrow}{\tau}$ represents a pure pressure — with no shear stresses. In this case, the material is a fluid: it cannot support shear.

2. Suppose $\mu \neq 0$. Then, there are shear stresses, in general. We still can define the pressure as:

$$\begin{aligned}\mathbf{P} &= -\frac{1}{3}\sum_i \tau_{ii}\\ &= -\frac{1}{3}\sum_i\left[2\mu\epsilon_{ii} + \lambda\underbrace{\delta_{ii}}_{=1}\sum_k \epsilon_{kk}\right]\\ &= -\frac{2\mu+3\lambda}{3}\sum_k \epsilon_{kk}.\end{aligned}$$

For an infinitesimal block:

$$\sum_k \epsilon_{kk} = \frac{\Delta x}{dx} + \frac{\Delta y}{dy} + \frac{\Delta z}{dz} = \frac{\Delta V}{dV}$$

where dV = the unperturbed volume of the block, and ΔV = the change in volume. So:

$$\mathbf{P} = -\frac{2\mu + 3\lambda}{3}\frac{\Delta V}{dV}.$$

Thus, by comparing with Equation 5.2.6, we would expect that $(2\mu + 3\lambda)/3 = \mathcal{K}$ = bulk modulus, and this is indeed what we find.

3. For an *incompressible* material, we can apply any pressure we want and ΔV should be 0. That's the definition of incompressible: you can't change the volume of any small interior cube. That means, $\mathcal{K} = \infty$ characterizes "incompressible." So, $2\mu + 3\lambda = \infty$ for an incompressible solid. But, μ is finite (at least, for a non-rigid material). So, $\lambda = \infty$ for an incompressible material. But, you don't want ∞'s in your equations.

   So, what happens to Equation 5.4.2 for an incompressible material? If $\lambda = \infty$ and $\overleftrightarrow{\tau}$= finite in that equation, then

$$\overline{\nabla} \cdot \overline{s} = 0 \tag{5.5.1}$$

   must hold. (Another way to obtain this result is to note that $\overline{\nabla} \cdot \overline{s} = \sum \epsilon_{kk} = \Delta V/dV = 0$ for an incompressible material.) Equation 5.5.1 is the condition for incompressibility. But, what do we do about the term

$$\lambda(\overline{\nabla} \cdot \overline{s}) = \infty \cdot 0 = \text{indeterminate} \tag{5.5.2}$$

   in Equation 5.4.2? Well, it turns out that this term equals $-\mathbf{P}$. That's because $\mathbf{P} = -\frac{2\mu}{3}\sum \epsilon_{kk}(=0) - \lambda \sum \epsilon_{kk} = -\lambda \sum \epsilon_{kk} = -\lambda(\overline{\nabla} \cdot \overline{s})$. So, for an incompressible material Equation 5.4.2 reduces to:

$$\begin{aligned} \overleftrightarrow{\tau} &= \mu\left[\overline{\nabla}\overline{s} + (\overline{\nabla}\overline{s})^T\right] - \mathbf{P}\,\overleftrightarrow{I} \\ \overline{\nabla} \cdot \overline{s} &= 0. \end{aligned} \tag{5.5.3}$$

Note that we have had to introduce the extra unknown, **P**, and that we have an extra equation ($\overline{\nabla} \cdot \overline{s} = 0$). Equation 5.5.3.

People often use the incompressible assumption for the earth because: (1) it's not too bad an assumption for many applications; and (2) it usually simplifies the mathematics, despite the fact that it requires an extra equation and an extra unknown.

5.6 Putting $\overleftrightarrow{\tau}$ into $\overline{F} = m\overline{a}$

We now know how to find $\overleftrightarrow{\tau}$ from $\overline{s}$. The equation we use for this, Equation 5.4.2, is the continuum mechanics analogue of Hooke's Law, $F = kx$, for a spring. The stresses, in turn, act across internal surfaces to cause the displacement field $\overline{s}$. To describe the latter process, we need to derive the continuum mechanics analogue of $F = ma$.

To obtain this equation we divide the material into blocks, and write down $F = ma$ for each block. The internal stresses enter into the $F = ma$ equation for a block, because they represent surface forces acting across the faces of the block. But the separation into individual blocks is purely artificial: it is unlikely that there are well-defined surfaces inside the material. Ultimately, we want to somehow come up with a body force representation for the stresses, that we can then use in our continuum mechanics version of $F = ma$.

Suppose we start with a block of material from inside the object, as in Figure 5.19. The block need not be infinitesimal. Let the block have volume $\mathcal{V}$ and surface $\mathcal{S}$. Let ρ

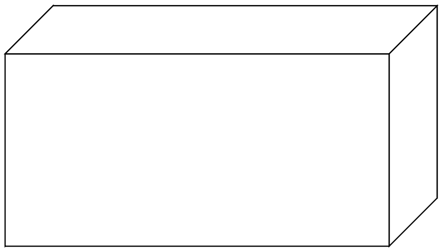

Figure 5.19:

be the density inside the block. We apply forces to the block: $\overline{F}_{\text{body}}$ and $\overline{F}_{\text{surface}}$ where $\overline{F}_{\text{body}}$ is the net body force (gravity, etc.) and $\overline{F}_{\text{surface}}$ is the surface force (stress $\times$ area). So:

$$\overline{F}_{\text{body}} = \int_{\mathcal{V}} \overline{f}\, d\mathcal{V}$$

where $\overline{f}$ = body force per unit volume (for gravity, $\overline{f} = \rho\overline{g}$), and

$$\overline{F}_{\text{surface}} = \int_{\mathcal{S}} \hat{n}\cdot \overleftrightarrow{\tau}\; d\mathcal{S}$$

where $\hat{n}$ = normal to $\mathcal{S}$, pointing outward from the block. These forces cause the block to accelerate. The acceleration is

$$\begin{aligned} \overline{a} &= \partial_t^2 \overline{x}_{\text{CM}} \\ &= \frac{\partial_t^2 \int_{\mathcal{V}} \rho\overline{s}\, d\mathcal{V}}{m} \\ &= \frac{\int_{\mathcal{V}} \rho\ddot{\overline{s}}\, d\mathcal{V}}{m} \end{aligned}$$

where $\overline{x}_{\text{CM}}$ is the position of the block's center of mass, $\overline{s}$ is the displacement field within the block, and m is the mass of the block. Then, $F = ma$ for the block becomes:

$$\int_{\mathcal{V}} \rho\ddot{\overline{s}}\, d\mathcal{V} = \int_{\mathcal{V}} \overline{f}\, d\mathcal{V} + \int_{\mathcal{S}} \hat{n}\cdot \overleftrightarrow{\tau}\; d\mathcal{S}.$$

By the divergence theorem:

$$\int_{\mathcal{S}} \hat{n}\cdot \overleftrightarrow{\tau}\; d\mathcal{S} = \int_{\mathcal{V}} \nabla\cdot \overleftrightarrow{\tau}\; d\mathcal{V} \tag{5.6.1}$$

where $\nabla\cdot \overleftrightarrow{\tau}$ is a vector with ith component:

$$\left(\nabla\cdot \overleftrightarrow{\tau}\right)_i = \sum_{j=1}^{3} \partial_j\, \tau_{ji}.$$

(You may not be familiar with the *tensor* divergence theorem. But you can obtain this result by separating $\overleftrightarrow{\tau}$ into column vectors, and using the *vector* divergence theorem on each of those vectors.) So, $\overline{F} = m\overline{a}$ for the block reduces to:

$$\int_{\mathcal{V}} \left[\rho\ddot{\overline{s}} - \overline{f} - \nabla\cdot \overleftrightarrow{\tau}\right] d\mathcal{V} = 0 \tag{5.6.2}$$

Since Equation 5.6.2 is true for arbitrary $\mathcal{V}$, the integrand must vanish everywhere:

$$\rho\ddot{\overline{s}} = \overline{f} + \overline{\nabla}\cdot \overleftrightarrow{\tau} \tag{5.6.3}$$

We identify $\overline{\nabla}\cdot \overleftrightarrow{\tau}$ as the body force due to stress.

Equation 5.6.3 plus the relation Equation 5.4.2 between $\overleftrightarrow{\tau}$ and $\overline{\nabla}\overline{s}$, etc., is a complete set of differential equations for $\overline{s}$. These equations tell us how the earth responds to internal and external forcing.

5.6.1 Example: Waves in a homogeneous, isotropic, ∞ medium

Suppose there are no body forces. Then $\rho\ddot{\overline{s}} = \overline{\nabla}\cdot \overleftrightarrow{\tau}$. Or, in component form:

$$\rho\ddot{s}_i = \sum_j \partial_j\, \tau_{ji} \qquad \text{for } i = 1, 2, 3. \tag{5.6.4}$$

And for an isotropic material:

$$\tau_{ij} = \mu\left[\partial_i\, s_j + \partial_j\, s_i\right] + \lambda\delta_{ij}\sum_k \partial_k\, s_k. \tag{5.6.5}$$

Put Equation 5.6.5 for τ_{ij} into Equation 5.6.4, and we have an equation for $\overline{s}$. Most of seismology involves solving this equation for specified functions of position $\mu(\overline{x})$ and $\lambda(\overline{x})$, and for specified external boundaries of the material. (We haven't talked about what happens at an external boundary of the object. If there are no applied external surface forces on the object, then there are homogeneous boundary conditions on $\overleftrightarrow{\tau}$. Specifically: $\hat{n}\cdot \overleftrightarrow{\tau}$ = surface force = 0 on the boundary. If there is an applied, external surface force, $\overline{f}_s$, then the boundary condition is $\hat{n}\cdot \overleftrightarrow{\tau} = \overline{f}_s$.)

For a homogeneous material μ and λ are constants, and so:

$$\begin{aligned}
\sum_j \partial_j\, \tau_{ji} &= \sum_j \left[\mu\left(\partial_i\,\partial_j\, s_j + \partial_j^2 s_i\right)\right] + \lambda\sum_j \partial_j\left(\delta_{ij}\sum_k \partial_k\, s_k\right) \\
&= \mu\partial_i\left(\overline{\nabla}\cdot\overline{s}\right) + \mu\nabla^2 s_i + \lambda\partial_i\left(\overline{\nabla}\cdot\overline{s}\right).
\end{aligned}$$

Or:

$$\overline{\nabla}\cdot \overleftrightarrow{\tau} = \mu\nabla^2\overline{s} + (\mu+\lambda)\,\overline{\nabla}\left(\overline{\nabla}\cdot\overline{s}\right).$$

Using this in Equation 5.6.4 gives:

$$\rho\ddot{\overline{s}} = \mu\nabla^2\overline{s} + (\mu + \lambda)\,\overline{\nabla}\left(\overline{\nabla}\cdot\overline{s}\right) \tag{5.6.6}$$

This is the partial differential equation for $\overline{s}$ in a homogeneous solid.

There are two types of solutions to this equation:

5.6.2 p-waves: longitudinal or sound waves

Take $\overline{\nabla}\cdot$ Equation 5.6.6. Then, since $\overline{\nabla}$ commutes with ∇^2 and with ∂_t^2:

$$\rho\partial_t^2(\overline{\nabla}\cdot\overline{s}) = \mu\nabla^2\overline{\nabla}\cdot\overline{s} + (\mu+\lambda)\nabla^2(\overline{\nabla}\cdot\overline{s}).$$

Or, defining $\Delta \equiv \overline{\nabla}\cdot\overline{s} \equiv$ *dilatation*:

$$\rho\ddot{\Delta} = (2\mu+\lambda)\nabla^2\Delta. \tag{5.6.7}$$

This is a wave equation for the scalar Δ. Solutions have the form:

$$\Delta = e^{i(\overline{k}\cdot\overline{x}-\omega t)} \tag{5.6.8}$$

where $\overline{k}$ = wave vector ($\overline{k}/|\overline{k}|$ is the direction of wave propagation, and $k = |\overline{k}| = 2\pi/\text{wavelength}$), and ω = angular frequency. Putting Equation 5.6.8 into Equation 5.6.7 gives: $\rho\omega^2 = (2\mu+\lambda)k^2$. So:

$$\frac{\omega}{k} = \sqrt{\frac{2\mu+\lambda}{\rho}}.$$

And, ω/k = wave velocity = v_p. The subscript 'p' stands for *primus*. This letter was originally assigned to these waves, because they are the first waves to arrive at a seismometer following an earthquake. Typically, within the earth:

$$v_p \approx \begin{cases} 6\ \frac{\text{km}}{\text{sec}} & \text{near the surface} \\ 13\ \frac{\text{km}}{\text{sec}} & \text{in the lower mantle.} \end{cases}$$

These "p-waves" are compressive, or sound, waves. They involve a change in volume of the material the wave passes through. That's because

$$\Delta = \overline{\nabla}\cdot\overline{s} = \sum_k \epsilon_{kk} = \frac{\Delta V}{dV}$$

for a small block. (Note: for an incompressible material which *cannot* undergo volume changes, $\lambda = \infty$, so that $v_p = \infty$.)

These waves also involve displacements along the direction of motion. To see this, suppose

$$\overline{s} = \overline{A} e^{i(\overline{k}\cdot\overline{x} - \omega t)}. \tag{5.6.9}$$

Then:

$$\Delta = \overline{\nabla} \cdot \overline{s} = i(\overline{k} \cdot \overline{A}) e^{i(\overline{k}\cdot\overline{x} - \omega t)}.$$

If $\Delta \neq 0$ (as it must if there is a p-wave), then $\overline{k} \cdot \overline{A} \neq 0$, so the direction of motion ($\overline{k}/k$) and the direction of displacement ($\overline{A}/A$) are not perpendicular. This is why p-waves are sometimes called longitudinal waves.

5.6.3 s-waves: transverse waves

Take $\overline{\nabla}\times$ Equation 5.6.6. Then, since $\overline{\nabla} \times (\overline{\nabla} f) = 0$ for any scalar f, we get:

$$\rho \partial_t^2 (\overline{\nabla} \times \overline{s}) = \mu \nabla^2 (\overline{\nabla} \times \overline{s}). \tag{5.6.10}$$

Equation 5.6.10 is a wave equation for the vector $\overline{\nabla} \times \overline{s}$. Solutions have the form:

$$\overline{\nabla} \times \overline{s} = \overline{A} e^{i(\overline{k}\cdot\overline{x} - \omega t)}. \tag{5.6.11}$$

Equation 5.6.11 is a traveling wave with wave vector $\overline{k}$ and frequency ω. It satisfies the differential equation if $\omega/k =$ wave velocity $= \sqrt{\mu/\rho}$. This wave velocity is usually written as v_s. The 's' stands for *secondus*. These waves arrive at a seismometer after p-waves (they are called s-waves). Note that $v_s < v_p$ (because $\mu < 2\mu + \lambda$). Typically, within the earth:

$$v_s \approx \begin{cases} 3.5 \ \frac{\mathrm{km}}{\mathrm{sec}} & \text{near the surface} \\ 7 \ \frac{\mathrm{km}}{\mathrm{sec}} & \text{in the lower mantle.} \end{cases}$$

There are no s-waves in the fluid core, since $\mu = 0$ in a fluid. The absence of s-waves traveling through the core is why we know there is a fluid core.

S-waves involve *no* displacements along the direction of motion, which is why they are called transverse waves. That's hard to show using this method for deriving s-waves, so I'll skip it.

5.6.4 The general case

After an earthquake, the real earth has all sorts of waves traveling through it. The earthquake generates p- and s-waves, which travel through the earth until they hit a discontinuity — a place where the material properties change abruptly. There, they generate transmitted and reflected waves. A p-wave hitting a boundary can cause reflected and transmitted p- *and* s-waves. The situation is similar for an incident s-wave. And, there are lots of boundaries.

As a further complication, the earth's free outer surface gives rise to another sort of wave: a "surface wave." (The p- and s-waves are called "body waves.") A surface wave is a solution to Equation 5.6.6 which varies periodically with distance along the surface, but decays exponentially with depth into the earth. These waves can't exist in an infinite medium, because they would grow exponentially with radius, and so be unbounded.

5.7 Departure From Perfect Elasticity

Real materials do not deform elastically. If you apply a stress to a material it does *not* immediately deform to its final state. There will probably be an instantaneous response followed by some sort of slow deformation.

People have come up with all sorts of mathematical models to describe anelasticity. In all cases, the stress is modeled as dependent on strain in some way. Except that for an anelastic material, the dependence is not simple multiplication by a constant.

The most mathematically complicated models are those where the stress is not a linear function of strain. Laboratory materials often behave in a non-linear manner when subjected to large stress. A familiar model is one where the strain rate (the time derivative of strain) is proportional to stress raised to the nth power, where n is greater than 1 (typically, $n \approx 3$ or so). For small stresses, though, people often find they can get away with a linear stress-strain relation. That's certainly true for seismic applications. And it is usually valid for studies of the response of the earth to surface loading, which we will consider later. Linear models are convenient because they are relatively easy to implement.

It should not be surprising that linear models tend to work reasonably well for small stress perturbations. You obtain a linear model by expanding a non-linear model to first-order in the stress perturbation. For example, if τ_0 is the background stress level, and $\Delta\tau$ is the perturbation in stress caused by whatever geophysical process you are considering, then the first-order Taylor series expansion of $\tau^n = (\tau_0 + \Delta\tau)^n$ is $\tau_0^n + n\tau_0^{n-1}\Delta\tau$, which is linear in $\Delta\tau$.

The consequence is that geophysicists usually assume the rheology for an anelastic material is linear (consistent with the standard scientific procedure of using the simplest model which can explain the observations). Except that here the "linear relation" is not simply multiplication by a constant (if it were, the material would be elastic). It involves, also, taking time derivatives or integrating over time.

A simple example of anelastic behavior in the real world is air resistance. In that case, the anelastic force is proportional to the velocity — which is the time derivative of the displacement ($F = b\dot{x}$). This is analogous to the stress being proportional to the strain rate. Because the time derivative is a linear operator, the stress-strain relation in this case is linear.

5.7.1 The general linear model

The most general, linear relation for an anelastic material is of the form

$$\sum_{k,l} W_{ijkl}\,\tau_{kl} = \sum_{k,l} X_{ijkl}\,\epsilon_{kl} \qquad \text{for all } i,j \tag{5.7.1}$$

where W and X may involve multiplication by constants, time derivatives of any order, and time integrals of any order. (For example: a second order integration term in X would look like $\int_0^t (\int_0^{t'} \epsilon_{kl}(t'')\,dt'')\,dt'$.)

People have used all sorts of different W's and X's to model geophysical processes. They find they need different W's and X's for different processes. For example, different models are required at the short periods (seconds to minutes) and small stresses appropriate to seismic waves, than at the long periods (thousands to millions of years) and relatively large stresses appropriate for tectonics and much of geodesy. It is not yet clear what happens at intermediate periods. Earth tides and earth rotation can maybe help with that in the future.

Evidently, different mechanisms are dominant in these different regimes. For seismic waves, the anelastic mechanism may have something to do with the opening and closing of pre-existing cracks in the medium, as the wave passes through. Or, it may involve the two sides of a crack sliding past one another (equivalent to simple friction).

In the tectonic regime the important mechanisms probably involve either *diffusion* (at high temperatures, molecular bonds are weak and so molecules diffuse over time) or *dislocation creep* (a dislocation, such as in Figure 5.20, that moves through a crystal over time).

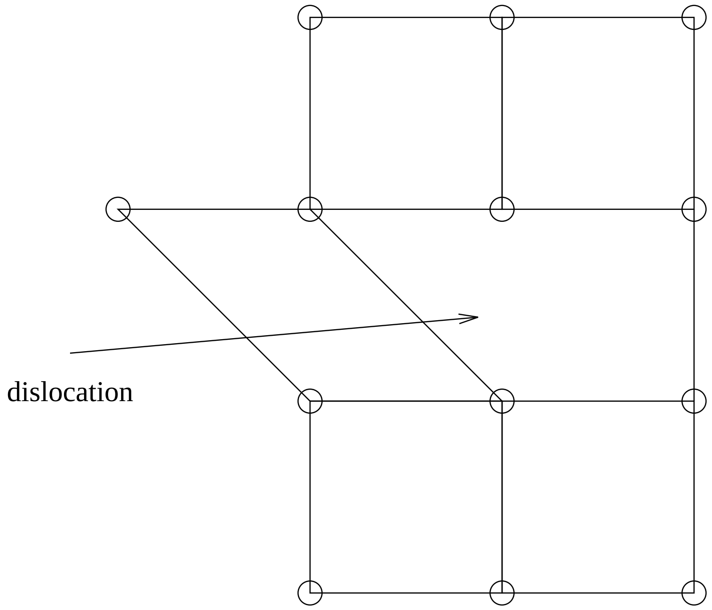

Figure 5.20:

Since people don't really understand the mechanisms very well, the usual mathematical approach used for modeling in-situ geophysical observations is to try to find the simplest W and X which explain those observations, and then to see what those mathematical results might tell you, if anything, about the mechanisms.

5.7.2 The Frequency domain

People usually work in the frequency regime. That is, they Fourier transform Equation 5.7.1 from the time domain to the frequency domain. They also transform the equation for conservation of momentum, Equation 5.6.2, into the frequency regime (for example, replacing $\ddot{\overline{s}}$ with $-\omega^2\overline{s}$). They then solve all these equations for the Fourier transform of $\overline{s}$, and take the inverse Fourier transform to get $\overline{s}$ in the time domain.

Why go to the frequency domain? Because then $\partial_t \to i\omega$ and $\int dt \to 1/i\omega$. So, in the frequency domain, W and X get replaced by simple multiplication operators — only where the multiplicative factors are functions of frequency (ω). You can, of course, solve *elastic* problems in the frequency domain, too. Only there's not much point in it, since in that case W and X are *already* simple multiplication operators in the time domain ($W_{ijkl} = \delta_{ik}\,\delta_{jl}$ and $X_{ijkl} = \Lambda_{ijkl}$), and so are unaffected by the Fourier transform.

So in the Fourier transform domain for a linear, anelastic material, where $\widetilde{W}$ and $\widetilde{X}$ are multiplication tensors (the $\tilde{}$ denotes Fourier transform), you can invert $\widetilde{W}$ to obtain a relation of the form:

$$\tilde{\tau}_{ij} = \sum_{k,l} \widetilde{\Lambda}_{ijkl}\,\tilde{\epsilon}_{kl}, \tag{5.7.2}$$

where $\tilde{\tau}_{ij}$ and $\tilde{\epsilon}_{kl}$ are the Fourier transforms of stress and strain, and $\widetilde{\Lambda} = \widetilde{W}^{-1}\widetilde{X}$ depends on ω and is, in general, complex.

Equation 5.7.2 looks like the elastic stress/strain relation. In fact, for isotropic materials, you can write:

$$\widetilde{\Lambda}_{ijkl} = 2\tilde{\mu}\delta_{ik}\,\delta_{jl} + \tilde{\lambda}\delta_{ij}\,\delta_{kl}$$

just as in the elastic case, but where $\tilde{\mu}$ and $\tilde{\lambda}$ are now complex functions of ω.

The result of all this is that the formalism used to solve elastic problems can also be applied to anelastic problems. We just replace μ and λ by functions of ω, and invert the solution from the frequency domain where the equations are solved, to the time domain.

We can make some general remarks about the functions $\tilde{\mu}(\omega)$ and $\tilde{\lambda}(\omega)$. For example, one issue is whether there is dissipation of shear energy or of compressional energy. In the first case (shear dissipation), we want the relation between $\mathbf{P}$ and $\Delta V/dV$ to be unaffected by anelasticity. In other words, no dissipation of compressional energy implies that the volume responds instantaneously to any applied pressure. That means, we want

$$\widetilde{\mathcal{K}} = \text{bulk modulus} = \frac{3\widetilde{\lambda} + 2\widetilde{\mu}}{3} \tag{5.7.3}$$

to be independent of ω. This must hold, even though both $\widetilde{\mu}$ and $\widetilde{\lambda}$ are likely to be complex functions of ω.

In the second case (compressive dissipation), $\widetilde{\mathcal{K}}$ *is* frequency dependent, but $\widetilde{\mu}$ is not. That's because $\widetilde{\mu}$ describes shear — after all, $\widetilde{\mu}$ is responsible for the off-diagonal terms in the stress tensor. So, $\widetilde{\lambda}$ depends on frequency while $\widetilde{\mu}$ does not.

For the earth, compressional dissipation is believed to be much less important than shear dissipation — although there is some evidence from free oscillation observations that suggests there may be a small amount of compressional dissipation. Here, we will only consider shear dissipation.

So, what do people use for $\widetilde{\mu}$ and $\widetilde{\lambda}$ in the shear dissipation case? It depends on whether they are interested in short-period or long-period behavior. We'll be more interested in long-period phenomena, but first we briefly consider:

5.7.3 Short-period behavior

This is relevant to seismic waves. In the seismic regime, the effects of anelasticity are relatively weak. So, $\widetilde{\mu}$ and $\widetilde{\lambda}$ depend only slightly on frequency. In this case, the frequency-dependent parts of $\widetilde{\mu}$ and $\widetilde{\lambda}$ are described with a parameter Q. Q is defined in seismology much as it is defined in electrical engineering. If you have a shear wave (ie. an s-wave) propagating through the medium, then

$$\frac{2\pi}{Q} \equiv \frac{\text{energy lost per cycle of the wave}}{\text{peak energy in the wave}}. \tag{5.7.4}$$

Q may depend on the frequency of the wave. For an elastic material, $Q = \infty$ (a "high quality" material).

Seismologists start by trying to find Q from observations, without worrying about the relationship between $\widetilde{\mu}$ and Q, or between $\widetilde{\lambda}$ and Q. They do this by observing the attenuation of body or surface waves as those waves travel through the earth. They can also find Q by observing free oscillations, since the same parameter, Q, can also be related to the free oscillation decay times. (A free oscillation — or normal mode of the earth — is a standing wave, and as such can be envisioned as the sum of traveling waves, each of which loses energy according to Equation 5.7.4.)

By using body wave attenuation measurements or free oscillation decay times to find $Q(\omega)$, seismologists are learning about the imaginary parts of $\widetilde{\mu}$ and $\widetilde{\lambda}$, because it is the imaginary part of a modulus that is responsible for energy dissipation. Once they know the imaginary part, they can infer information about the real part using what's called the "Kramers-Kronig relation," which follows from the simple and obvious requirement that a seismic wave can't arrive at a point until after the wave has been generated. The general form of the real part depends on what the functional form for $Q(\omega)$ is. For example, most seismic data are reasonably consistent with a frequency-independent Q. In that case, $\widetilde{\mu}$ turns out to be:

$$\widetilde{\mu}(\omega) = \mu_0 \left[1 + \left[\frac{2}{\pi} \ln\left(\frac{\omega}{\omega_m} \right) + i \right] \frac{1}{Q} \right] \tag{5.7.5}$$

where $\widetilde{\mu}_0$, ω_m and Q are all constants. (Equation 5.7.5 is actually only an approximation valid for large Q. There are an infinite number of possible functions $\widetilde{\mu}(\omega)$ which give a frequency independent Q. But they all reduce to Equation 5.7.5 for large Q.)

There are, however, reasons for believing that there might be a slight variation of Q with frequency, usually parameterized as

$$Q(\omega) = Q_0 \left(\frac{\omega}{\omega_m} \right)^{\alpha} \tag{5.7.6}$$

where Q_0, ω_m, and α are constants. In that case (again, an approximation for large Q):

$$\widetilde{\mu}(\omega) = \mu_0 \left(1 + \left\{ \cot\left(\frac{\alpha\pi}{2} \right) \left[1 - \left(\frac{\omega_m}{\omega} \right)^{\alpha} \right] + i \left(\frac{\omega_m}{\omega} \right)^{\alpha} \right\} \frac{1}{Q_0} \right). \tag{5.7.7}$$

More complicated functional forms for $Q(\omega)$ lead to more complicated relations between $\widetilde{\mu}(\omega)$ and Q. But people rarely use anything more complicated than Equation 5.7.6 for linear anelasticity. It has even been used successfully in explaining variations in the earth's rotation at periods longer than one year (specifically, the 14 month Chandler wobble, that we'll talk about later).

Laboratory experiments suggest that $\alpha \approx 0.3$ over a wide range of frequencies, including over the short-period portion of the seismic band. (To my knowledge, laboratory measurements of Q have not been performed at the lower frequencies characteristic of free oscillations.) The commonly accepted explanation for this frequency dependence is that the physical mechanisms responsible for dissipation in earth materials give rise to an absorption band: a frequency interval in which signals are highly damped. Disturbances at frequencies longer than those in the absorption band, should have $Q(\omega)$ values that vary with ω according to Equation 5.7.6. Presumably the absorption band frequencies are at shorter periods than the laboratory experiments, which would mean they are also at shorter periods than seismic observations. The seismic results, on the other hand, suggest that $\alpha = 0$, to within an error that is substantially smaller than 0.3. It is not yet clear why the seismic and laboratory results disagree.

Note that $\tilde{\mu}(\omega)$ has both real and imaginary parts in Equations (5.7.5) and (5.7.7). The imaginary part represents energy dissipation (it causes the stress to be out-of-phase with the strain). The ω-dependent real part represents dispersion. That is, waves with different frequencies travel at different speeds. This causes a seismic pulse to spread out as it travels.

5.7.4 Long-Period behavior

At periods of hundreds of years and longer, people have been reasonably successful at explaining observations using a "Maxwell solid" model.

5.7.4.1 Maxwell solid

The best way to think of a Maxwell solid is in terms of a spring and a dashpot in series as in Figure 5.21. The spring is elastic with spring constant 2μ (the "2" is used here so that this μ will end up corresponding to the Lamé parameter, μ), and the dashpot has viscosity 2η.

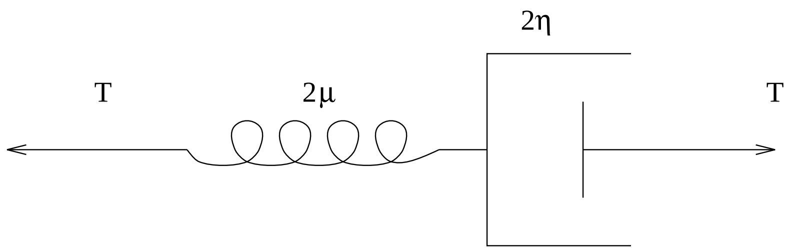

Figure 5.21:

Suppose that at time $t = 0$ you apply the outward force T to each side, as shown. The configuration immediately responds elastically — through stretching of the spring. You continue to hold the two ends apart so that the total stretched length doesn't change. Gradually the dashpot pulls apart and the spring closes, thus relieving the stress. In the limit $t \to \infty$ there is no stress (you no longer have to pull to keep the thing stretched) and the spring is closed.

That's a Maxwell solid. The idea, when generalized to a three-dimensional solid, is that the solid initially responds "elastically" to the sudden application of an external force. Shear stresses are initially set up in the material. But, gradually, the material inside the solid flows until the shear stresses are relieved. At $t = \infty$, the only stress remaining is pressure (this assumes we are talking about *shear* anelasticity, rather than *compressional* anelasticity).

Let's construct a mathematical model for this behavior. We start with the spring and dashpot shown in Figure 5.21. Define:

$$\begin{aligned} E_S &= \text{amount the spring has stretched} \\ E_D &= \text{amount the dashpot has stretched} \\ E = E_S + E_D &= \text{amount the total configuration has stretched.} \end{aligned}$$

Let T be the force applied to each end, as shown. In general, T and the E's are functions of time. How are they related? Across the spring:

$$T = (2\mu)E_S. \qquad (2\mu = \text{ spring constant})$$

Taking d/dt of this gives:

$$\dot{T} = 2\mu \dot{E}_S. \tag{5.7.8}$$

Across the dashpot: (where the force = viscosity $\times$ d/dt (stretching)):

$$T = 2\eta \dot{E}_D. \tag{5.7.9}$$

Add Equations 5.7.8 and 5.7.9 to get:

$$\frac{\dot{T}}{2\mu} + \frac{T}{2\eta} = \dot{E}_S + \dot{E}_D = \dot{E}.$$

Or:

$$2\mu \dot{E} = \dot{T} + \frac{1}{\tau_0} T \tag{5.7.10}$$

where $\tau_0 = \eta/\mu$ is called the "relaxation time."

Note that if there is no dashpot, then η = viscosity = ∞ (infinite viscosity $\Leftrightarrow$ elastic material), so $1/\tau_0 = 0$ and $2\mu\dot{E} = \dot{T}$. Or, integrating: $2\mu E = T$, which is the spring equation. So, Equation 5.7.10 has the right elastic limit.

Why is τ_0 called the relaxation time? Suppose we suddenly stretch the configuration by an amount E_0 at time $t = 0$, and we continue to hold the configuration apart so that it remains stretched by E_0 at all later times. A plot of E versus time is a step function, as shown in Figure 5.22. Let's solve Equation 5.7.10 to find the force T as a function of

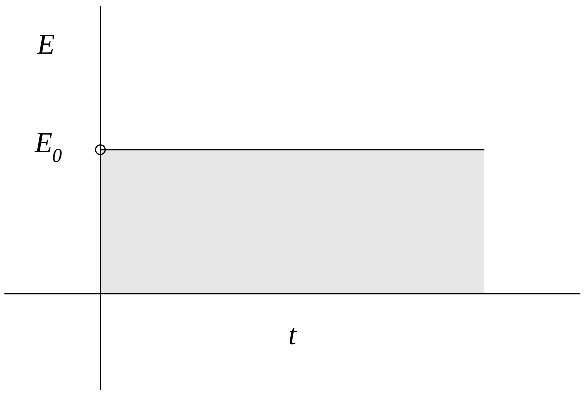

Figure 5.22:

time.

We know that T must be 0 at times $t < 0$. What is T right at $t = 0$, immediately after we apply the displacement E_0? To find the answer, we note that Equation 5.7.10 at $t = 0$ implies that

$$2\mu \lim_{\mathbf{dt}\to\mathbf{0}} \frac{E(t=0) - E(t=-dt)}{dt} = \lim_{\mathbf{dt}\to\mathbf{0}} \frac{T(t=0) - T(t=-dt)}{dt} + \frac{1}{\tau_0} T(t=0).$$

Or, since $T(t=-dt) = E(t=-dt) = 0$, and $E(t=0) = E_0$:

$$2\mu \lim_{\mathbf{dt}\to\mathbf{0}} \frac{E_0}{dt} = \lim_{\mathbf{dt}\to\mathbf{0}} \frac{T(t=0)}{dt} + \frac{1}{\tau_0} T(t=0)$$

T must always be finite. Therefore,

$$\lim_{\mathbf{dt}\to\mathbf{0}} \frac{2\mu E_0 - T(t=0)}{dt} = \frac{1}{\tau_0} T(t=0)$$

must be finite. That means that

$$T(t=0) = 2\mu E_0.$$

Physically, what happens is that at infinitely short times you see the spring but not the dashpot. And, for the spring alone, $T(0) = 2\mu E_0$.

So, for our example:

$$\begin{aligned} T &= 2\mu E_0 && \text{for } t = 0 \\ \dot{T} + \frac{1}{\tau_0} T &= 0 && \text{for } t > 0 \end{aligned}$$

(since $\dot{E} = 0$ for $t > 0$). The solution is

$$T = (2\mu E_0) e^{-t/\tau_0}.$$

So, the force decays exponentially to 0 as $t \to \infty$, with time constant τ_0.

A Maxwell solid is a special case of a *visco-elastic solid*. Visco-elastic solids can all be represented by springs and dashpots in series and/or in parallel. Because of the presence of dashpots, visco-elastic solids are described by a viscosity parameter (possibly more than one, if there are two or more dashpots with different viscosities). Two examples besides a Maxwell solid which are sometimes used in geophysics:

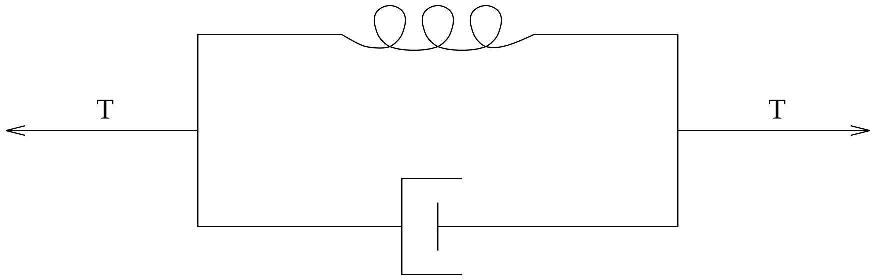

Figure 5.23:

5.7.4.2 Kelvin-Voigt solid

A Kelvin-Voigt solid is a spring and dashpot in parallel, as shown in Figure 5.23. If you turn on the force T at $t = 0$, and leave it on, the response is

1. initially, no stretching;

2. gradually the dashpot opens and the force gets transferred to the spring.

At $t \to \infty$, the behavior is controlled only by the spring, and the relation between T and the amount of stretching is given by the elastic relation for the spring.

5.7.4.3 Standard Linear solid

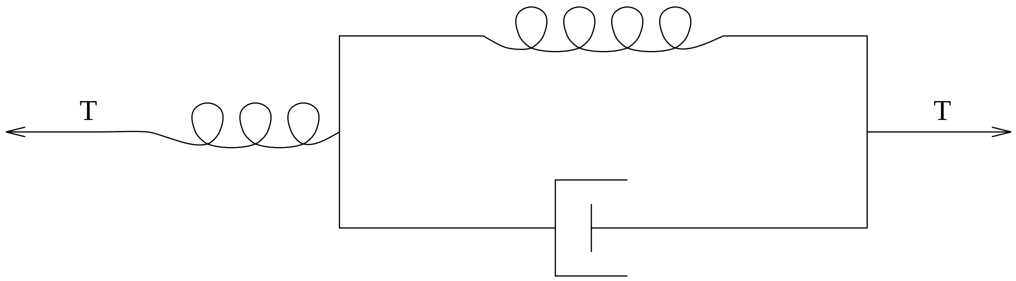

Figure 5.24:

A standard linear solid is a spring in series with a Kelvin-Voigt element, as shown in Figure 5.24. This is a combination of a Maxwell solid and a Kelvin-Voigt solid. The initial response is elastic, and is controlled by the spring on the left in Figure 5.24. The configuration relaxes because of the dashpot. The $t = \infty$ response is controlled by both springs.

5.7.4.4 Three-dimensional Maxwell solid rheology

Consider, again, a Maxwell solid. How do we extend the result Equation 5.7.10, that we derived for a spring and dashpot, to a three-dimensional solid? We assume there is no dissipation of compressional energy, so that the elastic result

$$-\frac{1}{3}\sum_k \tau_{kk} = -\left(\frac{2\mu+3\lambda}{3}\right)\sum_k \epsilon_{kk} \tag{5.7.11}$$

holds, where μ and λ are the elastic Lamé parameters. (Note that we are working in the time domain at the moment, not the frequency domain.) We assume that the Maxwell rheology applies only to the dissipation of shear energy. The shear stress is described by the deviatoric stress, which is $\overset{\leftrightarrow}{\tau}$ minus the pressure terms:

$$\delta\tau_{ij} = \tau_{ij} - \frac{1}{3}\sum_k \tau_{kk}\delta_{ij}. \tag{5.7.12}$$

The shear strain is similarly defined as $\overset{\leftrightarrow}{\epsilon}$ minus the volume terms:

$$\delta\epsilon_{ij} = \epsilon_{ij} - \frac{1}{3}\sum_k \epsilon_{kk}\delta_{ij}. \tag{5.7.13}$$

To describe a Maxwell solid rheology, we replace T and E in Equation 5.7.10 with $\delta\tau_{ij}$ and $\delta\epsilon_{ij}$, respectively, to obtain:

$$2\mu\left[\dot\epsilon_{ij} - \frac{1}{3}\sum_k \dot\epsilon_{kk}\delta_{ij}\right] = \dot\tau_{ij} - \frac{1}{3}\sum_k \dot\tau_{kk}\delta_{ij} + \frac{\mu}{\eta}\left[\tau_{ij} - \frac{1}{3}\sum_k \tau_{kk}\delta_{ij}\right]. \tag{5.7.14}$$

I claim that μ in Equation 5.7.14 is the elastic Lamé parameter, which is why I called it 'μ' originally. To see that, suppose we have an elastic material, where $\eta = \infty$. Then, Equation 5.7.14 relates $\dot\epsilon$ to $\dot\tau$ (no τ terms). Integrating over time gives:

$$2\mu\left[\epsilon_{ij} - \frac{1}{3}\sum_k \epsilon_{kk}\delta_{ij}\right] = \tau_{ij} - \frac{1}{3}\sum_k \tau_{kk}\delta_{ij}$$

which is the result for an elastic solid so long as $\mu =$ elastic Lamé parameter.

Now, go back to Equation 5.7.14, for $\eta \neq \infty$. Use Equation 5.7.11 to remove the τ_{kk} term in Equation 5.7.14, and use the derivative with respect to time (d/dt) of Equation 5.7.11 to remove the $\dot\tau_{kk}$ term. We find:

$$\dot\tau_{ij} + \frac{\mu}{\eta}\tau_{ij} = 2\mu\dot\epsilon_{ij} + \lambda\sum_k \dot\epsilon_{kk}\delta_{ij} - \frac{\mu}{\eta}\left(\frac{2\mu+3\lambda}{3}\right)\sum_k \epsilon_{kk}\delta_{ij}. \tag{5.7.15}$$

Except for the last term on each side, this equation is the time derivative of the elastic stress/strain relation.

Equation 5.7.15 is the anelastic stress/strain relation in the time domain. It is of the form

$$\sum_{k,l} W_{ijkl}\,\tau_{kl} = \sum_{kl} X_{ijkl}\epsilon_{kl}$$

where W and X include time derivatives. What happens in the frequency domain? Suppose we replace d/dt with $i\omega$. The left hand side is then $(i\omega + \mu/\eta)\widetilde{\tau}_{ij}$. We divide each side by $(i\omega + \mu/\eta)$, to obtain:

$$\widetilde{\tau}_{ij} = 2\widetilde{\mu}\widetilde{\epsilon}_{ij} + \widetilde{\lambda}\sum_k \widetilde{\epsilon}_{kk}\delta_{ij} \tag{5.7.16}$$

where

$$\begin{aligned}\widetilde{\mu} &= \mu\left(\frac{i\omega}{i\omega + \frac{\mu}{\eta}}\right)\\ \widetilde{\lambda} &= \lambda\left[\frac{i\omega + \frac{\mu}{\eta}\left(\frac{2\mu}{3\lambda}+1\right)}{i\omega + \frac{\mu}{\eta}}\right].\end{aligned}$$

So, in the frequency domain the stress/strain relation looks like the elastic result, except that $\widetilde{\mu}$ and $\widetilde{\lambda}$ are functions of ω, and are complex. Note that in the elastic limit (where $\eta \to \infty$), $\widetilde{\mu} \to \mu$ and $\widetilde{\lambda} \to \lambda$. Also: at high frequencies $(\omega \gg (\mu/\eta) = (1/\tau_0))$, $\widetilde{\mu} \to \mu$ and $\widetilde{\lambda} \to \lambda$, which is the elastic limit. And at low frequencies $(\omega \ll (\mu/\eta) = (1/\tau_0)$, $\widetilde{\mu} \to 0$ and $\widetilde{\lambda} \to (2\mu + 3\lambda)/3 = \mathcal{K}$. So, at long periods the material behaves as a fluid.

The Maxwell rheology has been especially useful for modeling postglacial rebound, as we shall see. To model the response of a Maxwell earth to any externally applied force, you work in the frequency domain (so you must transform the applied force to the frequency domain). You solve the equations, which are:

1. the stress/strain relation Equation 5.7.16 ; and

2. the conservation of momentum equation $(-\rho\omega^2\overline{s} = \nabla\cdot \overleftrightarrow{\tau} + \overline{f})$ and the equation

relating $\overleftrightarrow{\epsilon}$ to derivatives of $\overline{s}$, Equation 5.4.1, both of which are unaffected by anelasticity.

Once you obtain a solution for $\overline{s}$, you transform it to the time domain.

It turns out that for many applications, including for postglacial rebound, you work in the Laplace transform domain rather than the frequency domain. That means in all the above equations, you replace $(i\omega)$ with s, the Laplace transform variable. But, the frequency domain is always ok. It's just that sometimes the Laplace transform simplifies the algebra.

Chapter 6

Interpretation of Observed Gravity Anomalies

6.1 Surface Gravity Anomalies

As discussed in Chapter 4, a plot of the geoid provides a good representation of the long wavelength features in the earth's gravity field. But, it's usually *not* the best way to represent short wavelength features, since short wavelengths are less prominent in the geoid. Instead, short wavelength features are more clearly evident in maps of gravitational acceleration. Although the gravitational acceleration, as a function of position, can be inferred from satellite ranging data (see Equation 4.3.11 and the discussion in Section 4.4), those data do not resolve short wavelengths well. The short wavelength terms are best determined from surface gravity observations.

But, there's a complication. The reason geophysicists are interested in gravity, is because it allows them to learn about the earth's internal density distribution. But observed gravity also depends on the radial coordinate of the instrument. For example, if the gravimeter is further from the center of the earth, such as on top of a mountain, g is smaller.

So, to use observed gravity to learn about the earth's interior, you must first remove the effects of the earth's non-spherical surface. In principle, this means you should correct observed gravity for the topography (the elevation above the geoid) and for the geoid shape (the difference between the geoid and the mean spherical surface). But, at short wavelengths — say less than 1000 km — you don't need to correct for the geoid. That's because the geoid is smooth, with little power at those wavelengths. So if you only care about interpreting gravity over regions of a thousand km and smaller (and for regions much larger than that you would probably be representing gravity using geoid anomalies instead of surface gravity), then geoid anomalies are roughly the same over the entire region, and so by ignoring the geoid corrections you are not introducing any important *relative* errors across the region. And, it's only the relative errors which are apt to affect your interpretation. *Except* that people usually *do* correct for the P_2 component of the geoid: the ellipsoidal component. It's true that this component has a very long wavelength. But, it also has an enormous amplitude. So, if you don't remove it, your surface gravity observations could well show a linear decrease from North to South. You remove this component by subtracting the International Gravity Formula (see Equation 4.3.7) from your data. This is equivalent to removing the effects of the centrifugal force and of the P_2 internal density distribution.

The data should be corrected for topography, however, because topography can have significant power at short wavelengths. Let $h(\theta, \phi)$ be the elevation of the surface at (θ, ϕ). To remove the effects of h you construct what are called *Free Air anomalies*:

$$\begin{aligned} g_{\mathrm{FA}} &= (g_{\mathrm{obs}} - \gamma_0) - h\partial_r g \\ &= (g_{\mathrm{obs}} - \gamma_0) + h\left(\frac{2}{a} g_{\mathrm{obs}}\right) \\ &\approx (g_{\mathrm{obs}} - \gamma_0) + h[\text{in m}] \times [0.3086 \text{ mgal/m}] \end{aligned} \tag{6.1.1}$$

where γ_0 is the ellipsoidal contribution used to define the official "Geodetic Reference System". Equation 6.1.1 reduces the observed surface gravity to a common spherical surface (except, as described above, for the effects of the height of the geoid above the ellipsoid). This g_{FA} I denoted as Δg back in Section 4.3.2.

The free air anomaly, g_{FA}, can be interpreted in terms of the earth's density distribution. Think of g_{FA} as the acceleration you would observe at a surface of constant elevation outside the earth: the surface in Figure 6.1, for example. That's not really quite what it

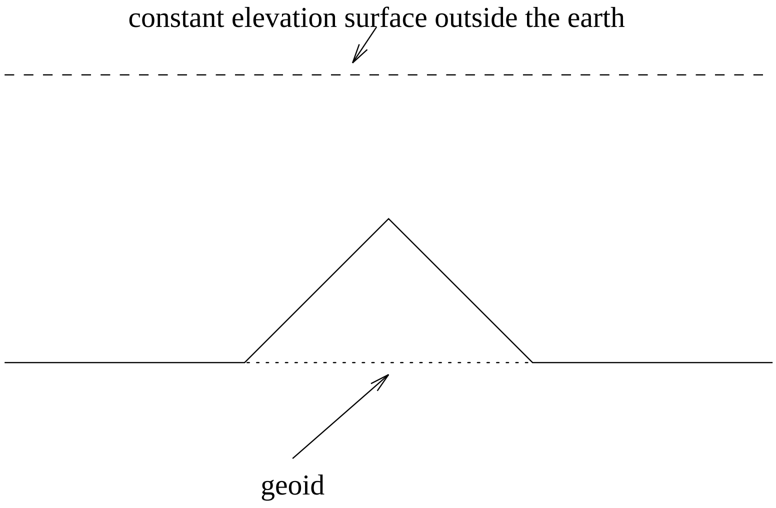

Figure 6.1:

is. It's really the acceleration you would observe on the *geoid* (except for the effects of subtracting γ_0 in Equation 6.1.1); *except* that, because the geoid is likely to be below the earth's surface, when we extrapolate from the surface down to the geoid we are ignoring the mass that we are extrapolating through (i.e. we are assuming that $\partial_r g = -2g/a$). This is the same sort of conceptual problem we had to deal with when we defined the geoid. In fact, g_{FA} *does* turn out to equal g on an external level surface, plus a constant. So, think of it this way, if you prefer.

g_{FA} is affected by the direct attraction of topographic mass. For example, g_{FA} will be large where there is a mountain, because the mountain has mass. You would have to classify this as an effect due to the earth's internal density (as opposed to the free air correction, $(2g/a)h$, which is an effect due to the position of the instrument), but it's not an interesting effect. There are better ways to learn about the topography than by using gravity: leveling, for example. Gravity is most useful because it can provide information about density *inside* the earth.

So, to learn about the interior of the earth, it is useful to subtract off the direct attraction of the underlying topography. There's a crude way of doing this that almost all geophysicists and most geodesists use. Assume the topography is smooth, in some sense.

See Figure 6.2. Then, if you are at elevation h when you make your measurement, you approximate all the mass above the geoid as an infinite slab of thickness h. You compute

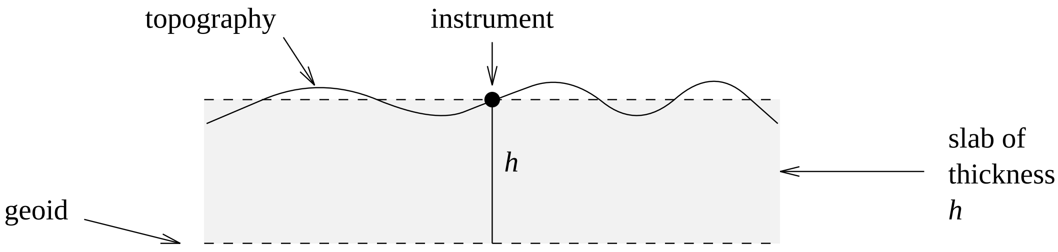

Figure 6.2:

g due to this imaginary slab. Earlier, in Section 3.2.2, we found that the downward gravitational acceleration caused by a homogeneous, infinitesimally thin, infinite plane with surface density σ, is $2\pi G\sigma$. This result does not depend on how far you are from the plane. So, g from our slab is:

$$\int_0^h 2\pi G(\rho\, dz) = 2\pi G\rho h$$

where ρ is the density of the slab. (Think of the slab as composed of lots of thin planes, each with $\sigma = \rho\, dz$, and sum over the planes.) This is the extra downward attraction due to the topographic mass, assuming the infinite slab approximation is ok. To learn about the earth's interior, you want to subtract this slab contribution from $g_{\rm FA}$. In this way you construct *Bouguer anomalies*:

$$\begin{aligned} g_{\rm B} &= g_{\rm FA} - 2\pi G\rho h \\ &= (g_{\rm obs} - \gamma_0) + \left(\frac{2}{a} g_{\rm obs} - 2\pi G\rho\right) h. \end{aligned}$$

For typical crustal rocks, ρ is often taken to be $\rho = 2.67$ gm/cm^3. Then, $2\pi G\rho =$ 0.1118 mgal/m, which is about 1/3 of the free air $(2g_{\rm obs}/a)$ correction.

So, by using $g_{\rm obs}$, together with h from leveling, we have a result, $g_{\rm B}$, that can be used to learn about the earth's interior. Sometimes geodesists use a better method of subtracting the effects of topographic density than simply using an infinite slab. They

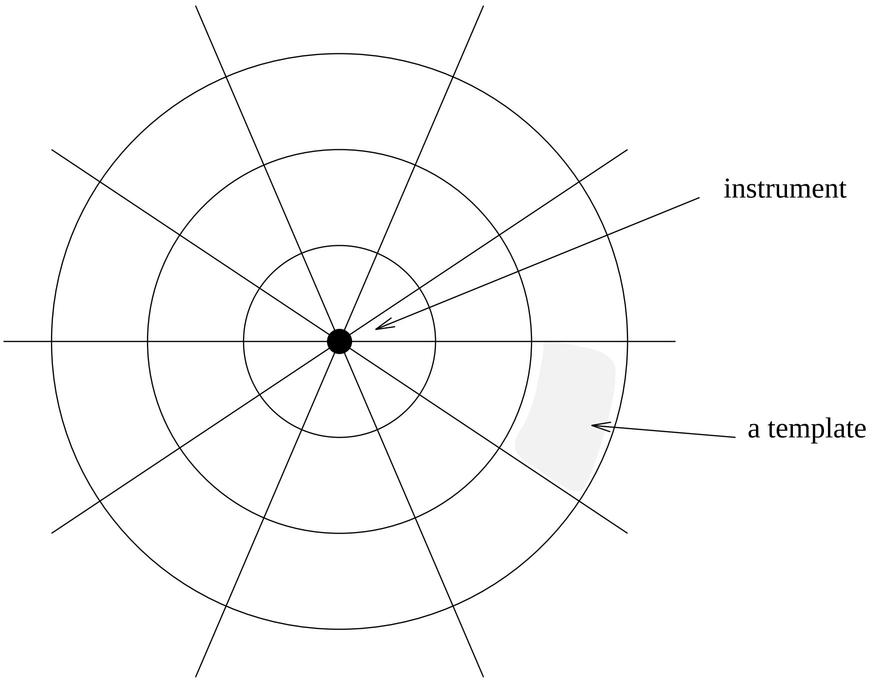

Figure 6.3:

break up the surrounding region into "templates." See Figure 6.3. They find the mean elevation (and density) of each template, compute the gravitational effect from the mass in each template, and sum over all templates. There are standard tables they use to look up the gravitational effect from a template — given the density and elevation. Geophysicists, though, almost always use the cruder Bouguer anomalies.

I've described what to do about topography when you're measuring g over the continents. The Bouguer anomaly essentially shaves off all mass down to the geoid. But, what do you do when you're observing g over the oceans? You must remove the effects of sea floor topography, *and* the effects of having low density water beneath you instead of higher density rocks. If these latter effects are not removed, then gravity anomalies across an ocean-continent boundary have a discontinuity at the the shoreline, as illustrated in Figure 6.4.

To construct Bouguer anomalies over the oceans, you imagine replacing the water with crustal rock. Suppose you observe g at sea level, where the ocean has depth H. You approximate the underlying ocean as an infinite water slab of thickness H. The

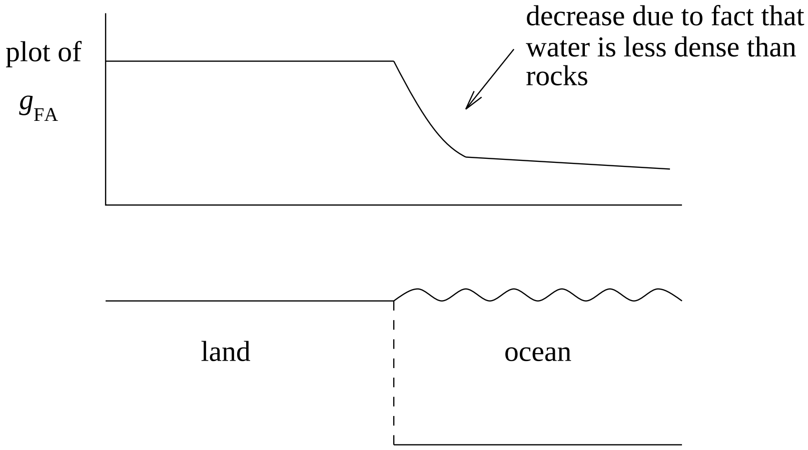

Figure 6.4:

gravitational acceleration toward the slab is $2\pi G\rho_w H$ where $\rho_w =$ density of sea water. If you replace the slab with crustal rocks of density ρ, the new slab produces a gravitational acceleration of $2\pi G\rho H$. So, to construct Bouguer anomalies over the ocean, you add $2\pi G(\rho - \rho_w)H$ to g_{FA} to obtain:

$$g_{\mathrm{B}} = g_{\mathrm{FA}} + 2\pi G\left(\rho - \rho_w\right) H.$$

Note that over the oceans: $g_{\mathrm{FA}} = g_{\mathrm{obs}} - \gamma_0$. No free air correction is required, because, except for dynamic topography (which is small), the oceans have zero elevation.

6.2 Isostasy

Suppose you measure g over the earth's surface, and construct Bouguer anomalies. You are now ready to learn about the earth's interior. But, you find a strange thing. At short horizontal wavelengths — say a few 10's of km or shorter — everything looks like you'd expect it. That is, there is little correlation between your g_{B} results and the topography. The Bouguer anomalies do a pretty good job of removing the topography. But, at wavelengths of 100 km and longer, the Bouguer anomalies look like the inverse of the topography. In fact, at these wavelengths the free air anomalies show little correlation with the topography.

This result was first discovered around 1850 by geodesists surveying in the Himalayas. Only they first noticed it in observations of the *direction* of gravity, rather than the amplitude. The direction is measured by finding the angle between a plumb bob and vectors to stars. The geodesists expected the plumb bob to be attracted by the Himalayas. But, they found the observed angle didn't change as they moved closer to the mountains.

6.2.1 Airy compensation

These results don't, of course, mean that the topographic mass doesn't affect gravity Instead, they demonstrate that there is material with anomalously low density beneath topography, with gravitational effects that tend to cancel the effects of the topography at long wavelengths. This is the theory of *isostasy.* It was proposed soon after the Himalayan results were found.

One version of this idea is that mountains have roots. See Figure 6.5. The crust is

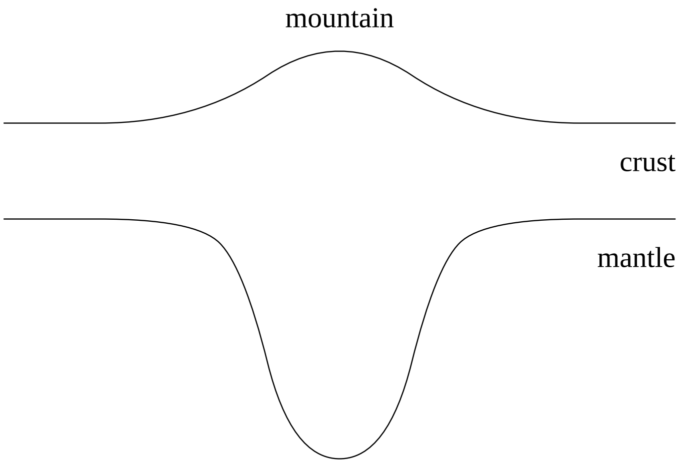

Figure 6.5:

lighter than the mantle. A mountain has excess mass above the geoid. This is balanced by the light crust extending down further into the mantle, so that the total mass in a vertical column is the same, whether the column is beneath the mountain or not. The result is that the gravitational effect from the root will nearly cancel the effect from the mountain in g_{FA}. For example, suppose we approximate the root as an infinite slab with *negative* density $= (\rho_{\text{crust}} - \rho_{\text{mantle}})$. Because of the compensation, the mass/area in the root is the negative of the mass/area in the infinite slab used to approximate the topography. So, since g from an infinite slab is independent of the distance to the slab, the effects of the two slabs cancel.

To find the size of a root, note from Figure 6.6 that equal masses in vertical columns implies that:

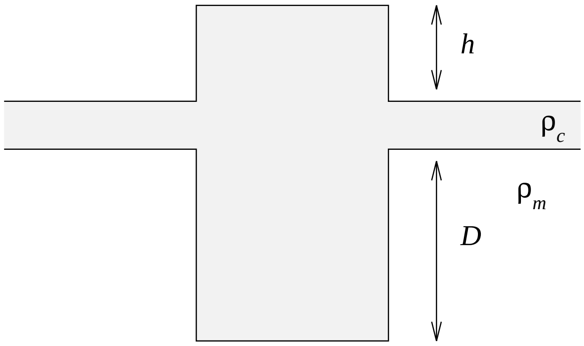

Figure 6.6:

$$h\rho_c = D(\rho_m - \rho_c).$$

Or:

$$D = h\left(\frac{\rho_c}{\rho_m - \rho_c}\right).$$

(Note: crustal thicknesses are usually much larger than both h and D, contrary to what is shown in Figure 6.6.) Typical crustal densities are 2–3 gm/cm^3. The density contrast between crust and mantle is typically: $\rho_m - \rho_c \cong 0.4$–$0.7$ gm/cm^3. Using these numbers, we can infer that D is 3–8 times h. So, the roots are large.

This model of isostasy is called *Airy compensation*, after a mathematician named Airy, who proposed it. The idea, physically, is that

1. the mantle is fluid over long time periods

2. the crust breaks up under long wavelength topographic loading, and the resulting blocks float on the mantle like icebergs float on water.

As we'll see, number 2 is an over-simplified idea.

Even if this idea of isostasy is valid (that is, if there is equal mass in all vertical columns), there should still be a small correlation between $g_{\rm FA}$ and topography. The root and the mountain have opposite masses, but they are not really infinite slabs, and so their gravitational effects depend on how far away they are. Thus, the contributions will not exactly cancel. By doing the calculations more exactly, and comparing the results with observations, people have been able to estimate the thickness of the crust.

6.2.2 Pratt compensation

Another type of isostasy, *Pratt compensation*, was proposed by Pratt — the person who had noticed the discrepancy in the Himalayan data — a couple years after Airy came up with his explanation. Pratt agreed that there was equal mass in equal columns. But, he argued that this was achieved by changing the density in a column, rather than by extending or contracting the bottom of the column. In this theory, mountains would be supported by a lower density crust, rather than by a thickened crust. See Figure 6.7. If h

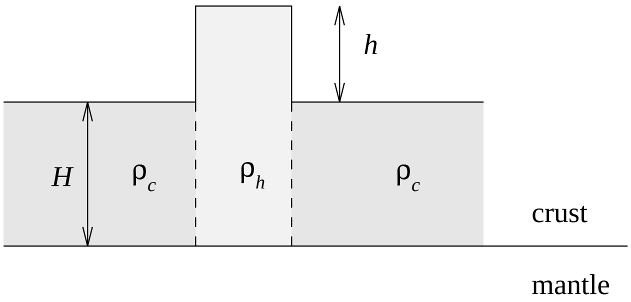

Figure 6.7:

= topography, H = crustal thickness in the absence of topography, ρ_c = normal crustal density, and ρ_h = crustal density beneath topography, then equal mass in equal columns gives: $H\rho_c = (H+h)\rho_h$. So

$$\rho_h = \left(\frac{H}{H+h}\right)\rho_c.$$

Again, this model predicts little correlation between $g_{\rm FA}$ and topography, since to lowest order you approximate the underlying column as an infinite slab with mass/area equal to the total mass/area of the column. And each column has the same mass/area no matter what h is, so that there is no spatial variation of g_{FA}. Of course the underlying column is *not* an infinite slab, and so there will be some correlation between h and g_{FA}. Again, this can be used to find the depth to the crust/mantle boundary.

A more general version of Pratt compensation would not require ρ to be constant in a vertical column. All you need specify is the total mass in a vertical column. How you distribute the mass vertically is up to you. Then ρ_h and ρ_c in $\rho_h = (\frac{H}{H+h})\rho_c$ are the *average* densities beneath the mountain and the non-mountainous surface.

6.2.3 Airy versus Pratt

Which of these models (Airy or Pratt) is correct? Pratt compensation might be valid if the topography were the result of thermal expansion or contraction in the crust or upper mantle. The Airy model is probably valid if the topography can be viewed as an applied load on the earth.

People have estimated the thickness of the crust for both Pratt and Airy cases (by correlating the observed g with topography). They find crustal thicknesses of about 50 km under land and of about 10 km under oceans. These results are consistent with seismic observations of the crustal thickness. The crust/mantle boundary is a chemical boundary (different materials on different sides of the boundary) and so represents a discontinuity in material properties. Seismic waves are reflected from the boundary. (The boundary is known to seismologists as the *Moho*, short for the Mohorovičić discontinuity.) By comparing the arrival times of waves reflected from the boundary with the time of the earthquake, seismologists can infer the boundary depth. They obtain depths of 30–60 km under the continents, and 5–15 km under the oceans. Furthermore, the Moho depths generally show that the crust *is* thicker under mountainous regions, which tends to support Airy over Pratt in those regions.

Why, then, don't you see the effects of isostasy at short wavelengths (10's of km and shorter)? At those wavelengths the topographic signal is present in g_{FA}, but not in g_{B}. Is it because isostasy is not operative at short wavelengths? Perhaps short wavelength loads don't break up the crust? The answer is: "yes." In fact, we'll see later that even long wavelength loads don't actually break up the crust. But, there's more to the story than this. Even if the crust *did* break up under short wavelength loads, there would still be topographic effects evident in g_{FA}. To understand why, let's try to do a better job of finding the effects of topography and its compensation on g_{FA} than simply using infinite slabs.

6.3 A Better Model for Airy Compensation Effects on Gravity

Consider, first, Airy compensation. Suppose, for simplicity, that the earth is flat and symmetric in the $\hat{e}_y$-direction, so that the topography is a function only of x: $h = h(x)$. Suppose $h(x) = h_0 e^{ikx}$ where h_0 and k are constants ($k = 2\pi$/wavelength). (Topography is not a complex-valued function, of course. But, imagine we have expanded $h(x)$ into a complex Fourier series, and that here we are considering just one term in that series.) Assume $k \geq 0$. For perfect compensation, $h(x)$ is supported by a root of thickness

$$D(x) = \left(\frac{\rho_c}{\rho_m - \rho_c}\right) h(x) = \left(\frac{\rho_c}{\rho_m - \rho_c}\right) h_0 e^{ikx}.$$

Let H = crustal thickness. Assume that $|h| \ll H$ and $|D| \ll H$. See Figure 6.8. Then,

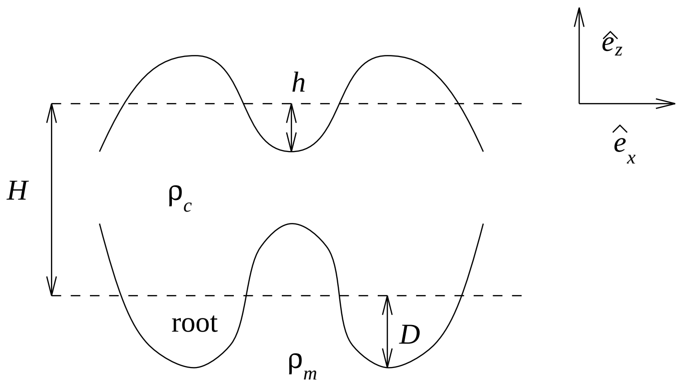

Figure 6.8:

we approximate the topography with a surface mass $\sigma = \rho_c h(x)$ at $z = 0$. And, we approximate the root as a surface mass $\sigma = (\rho_c - \rho_m)D(x) = -\rho_c h(x)$ at $z = -H$.

Let's find a better approximation to g from these surface masses, than the infinite slab approximation.

Suppose we are trying to find g at the point P shown in Figure 6.9, due to the surface mass $\sigma(x')$. Suppose P is located at $(x,\, z = a)$, and that the surface mass is on the plane $z = 0$ (so the surface mass is a distance a below P). Break the surface into thin strips of

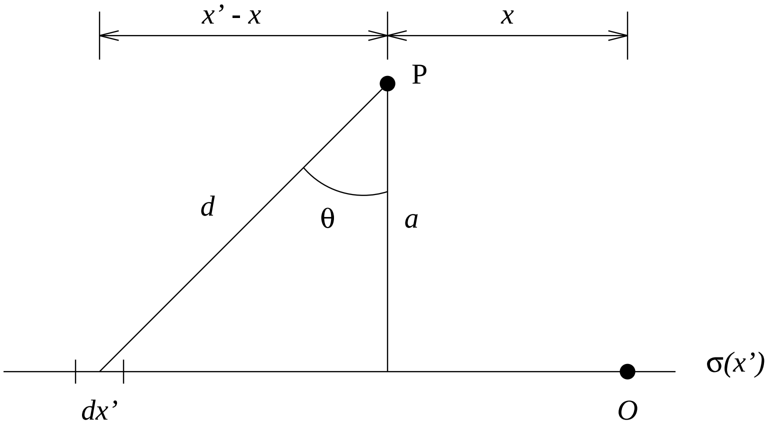

Figure 6.9:

width dx', as shown. Each strip looks like an infinite line extending in the $\hat{e}_y$ direction, with constant mass/length $= \lambda = \sigma(x')\,dx'$. The distance between P and the line is $d = \sqrt{a^2 + (x' - x)^2}$. From Section 3.2.3, we know that g due to the line has magnitude

$$g_{\rm line} = \frac{2G\lambda}{d} = \frac{2G\sigma(x')\,dx'}{\sqrt{a^2 + (x' - x)^2}},$$

and is directed from P towards dx'. We only want the component of acceleration in the $-\hat{e}_z$ direction (this direction is downwards) since, to first order, only this component affects $|\overline{g}|$. So, the $-\hat{e}_z$ contribution to $\overline{g}$ from the line is

$$\frac{2G\sigma(x')\,dx'}{\sqrt{a^2 + (x' - x)^2}}\cos\theta = \frac{2G\sigma(x')\,dx'\,a}{a^2 + (x' - x)^2}$$

where $\cos\theta = a/d$.

We add up (i.e. integrate) over all lines to get the total contribution from the plane:

$$g(x, a) = \int_{-\infty}^{\infty} \frac{2Ga\sigma(x')\,dx'}{a^2 + (x' - x)^2} \qquad (6.3.1)$$

where $g(x, a)$ is the $-\hat{e}_z$ component of $\overline{g}$ at the point P. Suppose $\sigma(x') = \sigma_0 e^{ikx'}$. Then we can do the integral in Equation 6.3.1 to get

$$g(x, a) = 2G\pi\sigma_0 e^{-ka} e^{ikx} \tag{6.3.2}$$

(for $k < 0$ we get the same result, except that $-ka$ is replaced with $-|k|a$).

Return, now, to our original problem of finding g from the topography and its compensation. Suppose we want to find g at any elevation, z, above the earth's outer surface ($z = 0$). (Gravity observations are made only at $z = 0$, but we want to keep $z \geq 0$ arbitrary at this time, because it will be useful when we compute the geoid, later.) We use Equation 6.3.2 for each of the two planes. One plane has $\sigma_0 = \rho_c h_0$ and $a = z$. The other has $\sigma_0 = -\rho_c h_0$ and $a = H + z$. Using Equation 6.3.2 for each plane, and adding to get the total g, gives:

$$g(x, z) = 2G\pi\rho_c h_0 e^{ikx} \left[1 - e^{-kH}\right] e^{-kz}. \tag{6.3.3}$$

$g(x, z = 0)$ from Equation 6.3.3 is our prediction for g_{FA}.

Suppose the horizontal wavelength of the topography is large, so that k is small. Specifically, suppose that $kH \ll 1$, so that the wavelength $\gg$ crustal thickness. Then $e^{-kH} \approx 1$, and so $g(x, z = 0) = 0$. But, suppose $kH \gg 1$, so that the wavelength is $\ll$ crustal thickness. Then $e^{-kH} \to 0$, so that $g(x, z = 0) \to 2\pi G\rho_c h(x)$, which is the infinite slab result for the topography alone.

So for long wavelengths there is little correlation between g_{FA} and topography. And for short wavelengths there is little correlation between the *Bouguer* anomalies and topography. "Short" and "long" mean compared with the crustal thickness. In fact, the way geophysicists learn about the crustal thickness is by comparing Equation 6.3.3 for $g_z(x, z = 0)$ (or its two-dimensional equivalent $g_z(x, y, z = 0)$) with observations of g_{FA}.

What's happening here is that short wavelength gravity variations on a plane die away rapidly with distance from the plane. The decay distance is about equal to a horizontal wavelength. So, if the crustal thickness is larger than a horizontal wavelength, the gravitational signal from the root will not be seen at the surface. But, if the crustal thickness is much less than a wavelength, the signal from the root will not have decayed much, and so will cancel the signal from the topography.

6.4 A Better Model for Pratt Compensation Effects on Gravity

Let's go through a similar derivation for Pratt compensation to see how it differs from Airy compensation.

Suppose the surface topography is $h = h(x) = h_0 e^{ikx}$, where $k \geq 0$. Then, g at (x, z) due to the surface topography is

$$2\pi G \rho_c h_0 e^{ikx} e^{-kz} \tag{6.4.1}$$

as in the Airy case.

The Pratt model proposes that the topography is supported by density variations in the crust. Suppose the crust fills the region $-H < z' < 0$ (see Figure 6.10). Let

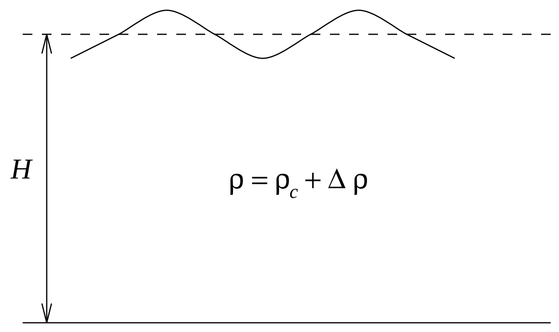

Figure 6.10:

$\Delta\rho(x', z')$ be the crustal density anomaly at the location (x', z'), so that the total density inside the crust is $\rho(x', z') = \rho_c + \Delta\rho(x', z')$. (We used ρ_c in Equation 6.4.1, instead of ρ, because $h_0 \Delta\rho$ is 2nd order in small quantities.) We assume that there is no density anomaly below the crust ($z' < -H$).

When we introduced Pratt compensation above, we assumed that $\Delta\rho(x', z')$ was independent of z'. That is, we assumed that the entire crustal column beneath a topographic feature has a uniform density anomaly. We defined $\rho_c + \Delta\rho(x', z') \equiv \rho_h$, and then concluded that

$$\begin{aligned} \rho_h = \left(\frac{H}{H+h}\right)\rho_c \;\; &\approx \;\; \rho_c - \frac{h}{H}\rho_c \qquad \Rightarrow \\ \Delta\rho \;\; &= \;\; -\frac{h_0}{H}\rho_c. \end{aligned} \tag{6.4.2}$$

I mentioned that you could generalize Pratt compensation to cases where the density anomaly depends on z', with the result that

$$\rho_h \approx \rho_c - \frac{h}{H}\rho_c \tag{6.4.3}$$

is valid in an average sense. I want to consider that more general case here. We don't really need to do that to address the topic of this section: finding something better than an infinite slab to approximate the gravity anomaly from Pratt compensation. But, the results of the more general case will be useful, later, in applications.

To have equal masses in all columns, $\Delta\rho$ must satisfy:

$$\underbrace{\int_{-H}^{h(x)} \left(\rho_c + \Delta\rho(x, z')\right)\, dz'}_{\text{mass in a column}} = \underbrace{H\rho_c}_{\substack{\text{mass in a}\\ \text{column with}\\ \text{no topography}}}$$

for all x (the value of x defines which column is being considered: x is the horizontal coordinate of the column). Or:

$$\int_{-H}^{0} \Delta\rho(x, z')\, dz' = -\rho_c h_0 e^{ikx} \tag{6.4.4}$$

for all x. (For the upper limit of the integral we have used 0 instead of h, because $\Delta\rho$ and h are both 1st order, implying that the difference between using 0 and using h as the limit, is 2nd order.) Since Equation 6.4.4 must hold for all x, we expect $\Delta\rho(x, z')$ to be proportional to e^{ikx}, so that:

$$\Delta\rho(x, z') = \delta\rho(z') e^{ikx}.$$

With this definition of $\delta\rho$, Equation 6.4.4 implies that:

$$\int_{-H}^{0} \delta\rho(z')\, dz' = -\rho_c h_0. \tag{6.4.5}$$

Note that for $\delta\rho(z') = \delta\rho = \text{constant}$, Equation 6.4.5 reduces to $\delta\rho = -\frac{h_0}{H}\rho_c$, as in Equation 6.4.2.

So, Equation 6.4.5 is a more general version of Pratt compensation, valid for depth-dependent density anomalies. What effect does the anomalous density $\Delta\rho(x, z') = \delta\rho(z')e^{ikx}$ have on g? Think of $\delta\rho(z')e^{ikx}$ as spread over a lot of thin horizontal planes of thickness dz', each with surface density $\sigma(x', z') = \delta\rho(z')e^{ikx}\,dz'$. The plane at z' is a distance $z - z'$ from the field point (note that $z' < 0$). From Equation 6.3.2, we know that $g(x, z)$ caused by the surface density on that plane is:

$$2G\pi e^{kz'}e^{ikx}\delta\rho(z')\,dz'\,e^{-kz}.$$

We add up the contributions from all planes, from $z' = -H$ to $z' = 0$. We also add the effect, Equation 6.4.1, from the surface topography. We find that the total g is:

$$g(x, z) = 2\pi G\left[\rho_c h_0 + \int_{-H}^{0}\delta\rho(z')e^{kz'}\,dz'\right]e^{ikx}e^{-kz}.$$

This is the result for generalized Pratt compensation.

6.4.1 Remarks

- Suppose the wavelength $\gg$ crustal thickness. Then $kH \ll 1$, so that $|kz'| \ll 1$ for z' between $-H$ and 0. Then, crudely, $e^{kz'} \to 1$. So

  $$g \to 2\pi G\left(\rho_c h_0 + \int_{-H}^{0}\delta\rho(z')\,dz'\right)e^{ikx}e^{-kz} = 0$$

  since the integral here is equal to $-\rho_c h_0$ for perfect compensation (see Equation 6.4.5). So, there is no correlation between g_{FA} and topography in this long wavelength limit.

- Suppose the wavelength is $\ll$ crustal thickness. Then $kH \gg 1$, so that $e^{kz'} \approx 0$ except for z' close to 0. So, as long as $\delta\rho(z')$ is not all concentrated right near the surface ($z' = 0$), the integral $\int_{-H}^{0}\delta\rho(z')e^{kz'}\,dz'$ is small. So, $g(z = 0) \to 2\pi G\rho_c h_0$, which is the infinite slab result = the Bouguer correction. So, short wavelength topography is not correlated with g_{B}.

- Consider the usual Pratt compensation, where $\delta\rho = \text{constant} = -\frac{h_0}{H}\rho_c$. Then

$$\int_{-H}^{0} \delta\rho(z') e^{kz'}\, dz' = -\frac{h_0}{H}\rho_c \left(\frac{1-e^{-kH}}{k}\right).$$

So:

$$g(x,z) = 2\pi G\rho_c h_0 \left[1 - \frac{1-e^{-kH}}{kH}\right] e^{ikx} e^{-kz}. \qquad (6.4.6)$$

At long wavelengths ($kH \ll 1$), let's expand Equation 6.4.6 to 1st order in kH, rather than to 0th order (it vanishes to 0th order, as we've seen).

$$\frac{1-e^{-kH}}{kH} \cong \frac{1 - \left[1 - kH + \frac{(kH)^2}{2}\right]}{kH} = 1 - \frac{kH}{2}.$$

Thus, we obtain the first order Pratt result:

$$g(z=0) \approx 2\pi G\rho_c h \left(\frac{kH}{2}\right). \qquad (6.4.7)$$

How does the first order Pratt result (Equation 6.4.7) differ from the 1st order Airy result?

First Order Airy:

$$\begin{aligned} g(z=0) &= 2\pi G\rho_c h \left[1 - e^{-kH}\right] \\ &\approx 2\pi G\rho_c h (kH). \end{aligned}$$

So, to first order at long wavelengths, the Pratt g is half the Airy g. That does make some sense. It sort of implies that the compensation is in the middle of the crust for the Pratt case, rather than at the bottom. And, after all, the Pratt compensation is distributed uniformly throughout the entire crust, so that the average compensation *is* in the middle.

6.5 Geoid anomalies

Isostasy should show up somehow in the geoid, too. After all, the geoid is just another way to represent the gravity field. If we knew how the different isostasy models affected the geoid, we could maybe compare the geoid with topography and so learn about crustal thicknesses, etc. But, why bother? Useful information on isostasy will only appear at wavelengths of less than a few thousand km. At longer wavelengths you start to see effects of density anomalies deep inside the earth, that tend to mask any near-surface isostatic signal. And, I've been telling you that shorter wavelength features are more clearly evident in gravitational acceleration maps than in the geoid.

Nevertheless, people do sometimes look at isostatic effects on the geoid, particularly over the oceans. There aren't many surface gravity observations over the oceans. But, satellite altimeter data do give good results for the geoid, even at very short wavelengths. Of course, you could always transform the geoid results to obtain gravitational acceleration maps, and thus amplify the shorter wavelength terms. But there can be reasons to work, instead, with the geoid results, given that they are the direct output of the altimeter analysis.

What are the effects of Airy and Pratt compensation on the geoid? The reason I kept the e^{-kz} terms in the results above for g (rather than setting $z = 0$) is so that we can answer this question. Remember that

$$\text{geoid anomaly } = \left.\frac{\Delta V}{g}\right|_{z\,=\,0}$$

where ΔV is the perturbation in the potential due to the underlying topography and its compensation. And:

$$\Delta V(z) = -\int^{z} g(z')\, dz' \qquad (6.5.1)$$

where g is the effect of the topography plus compensation on the vertical acceleration. (The minus sign in Equation 6.5.1 is because we have defined g to be positive downwards: in the negative $\hat{e}_z$-direction.) (Back in Chapter 4 we showed that $(\partial_r \Delta V + \frac{2}{a}\Delta V)|_{\text{surface}} = \Delta g$ (Equation 4.3.8). But, the Δg in this equation is g on the geoid; whereas the g in Equation 6.5.1 is g in space, as a function of z.)

We just finished finding $g(z)$ for different types of compensation. In all cases the z

dependence was e^{-kz}. So, the integral over z is $-\frac{1}{k}e^{-kz}$. So:

Airy

$$N(x) = \frac{2\pi G\rho_c h(x)H}{g}\left[\frac{1-e^{-kH}}{kH}\right] \tag{6.5.2}$$

Generalized Pratt

$$N(x) = \frac{2\pi G}{gk}\left[\rho_c h(x) + \left(\int_{-H}^{0}\delta\rho(z')e^{kz'}\,dz'\right)e^{ikx}\right] \tag{6.5.3}$$

Constant density Pratt

$$N(x) = \frac{2\pi G\rho_c h(x)H}{gk^2H^2}\left[-1+kH+e^{-kH}\right] \tag{6.5.4}$$

6.5.1 One Final Remark

Suppose we are considering long wavelengths, where $kH \ll 1$. Let's expand all three results, above, to zero order in kH. So, use $e^{-kH} \approx 1-kH$ in Equation 6.5.2, $e^{kz'} \approx 1+kz'$ in Equation 6.5.3, and $-1+kH+e^{-kH} \approx -1+kH+1-kH+\frac{(kH)^2}{2}$ in Equation 6.5.4. Then,

Airy

$$N(x) \approx \frac{2\pi G\rho_c Hh(x)}{g}$$

Generalized Pratt

$$\begin{aligned} N(x) &\approx \frac{2\pi G}{kg}\left[\underbrace{\rho_c h(x) + \int_{-H}^{0}\delta\rho(z')\,dz'\,e^{ikx}}_{\text{These terms cancel}} + k\int_{-H}^{0}\delta\rho(z')z'\,dz'\,e^{ikx}\right] \\ &= \frac{2\pi G}{g}\left(\int_{-H}^{0}\delta\rho(z')z'\,dz'\right)e^{ikx} \end{aligned}$$

Constant density Pratt

$$N(x) \approx \frac{\pi G\rho_c Hh(x)}{g}$$

All three of these results for $N(x)$ have the same x-dependence as $h(x)$: e^{ikx}. And, all three are independent of k (except for the k in e^{ikx}). That has an interesting implication.

It means that if you have a complicated topography — not a simple e^{ikx} — and if you transform the topography to a sum or integral of e^{ikx} terms, the contribution to $N(x)$ from each term is independent of k. Or, to be more accurate, it depends on k in exactly the same way as h_0 does. So, when you add up all the contributions to N from the different e^{ikx} terms, you find that the resulting sum has exactly the same x-dependence as the total topography. In other words, $N =$ constant times topography, no matter how complicated the topography is (so long as we only include long wavelength topography).

This result does not hold for gravity at these long wavelengths. The gravity contributions from e^{ikx} depend linearly on k (for $kH \ll 1$), so adding up the e^{ikx} contributions to g does not give (constant) $\times$ (topography).

6.6 Lithospheric Bending

There *are* cases where topography is caused by thermal anomalies in the crust and upper mantle. In these cases, some sort of Pratt isostasy is probably operable.

But, most topography is due to other causes. For example: ongoing collisions of plates, or relics of past collisions. Or volcanic activity. Or erosion. In these cases, you might expect Airy compensation to be at work. But, Airy compensation is too simple a picture of what really happens. It assumes that when you put a load on the crust, the crust breaks up. Actually, the crust bends. And, it's not only the crust that bends. It's the entire lithosphere. It turns out that for loads with long wavelengths, the lithosphere bends easily, and you get what look like roots "supporting" the topography. At short wavelengths the lithosphere does not bend much, and so you get smaller roots. Thus, at short wavelengths, the effects of topography on g are not exactly cancelled by the effects of the root. And it's not just because the roots are far below the surface compared with the wavelength. It is also because the roots are small.

6.6.1 Lithosphere

Before modeling this bending, I need to define and describe the lithosphere.

There are at least two ways you can subdivide the upper few hundred km or so of the earth. You can separate it into the crust and mantle. As I've noted, the crust/mantle boundary is a chemical boundary, with significant density contrasts across it. The crust is less dense than the mantle.

You can also divide the region into a lithosphere overlying an asthenosphere. See Figure 6.11. The idea is that the lithosphere is the top layer of convection in the mantle. It has cooled by radiating away heat into the atmosphere, and has become brittle. It behaves like a solid even at extremely long time periods. The region beneath the lithosphere is the asthenosphere. It is still pretty hot, and so behaves like a fluid over long time periods. The crust is perhaps the top one-quarter to one-half of the lithosphere. Typical lithospheric

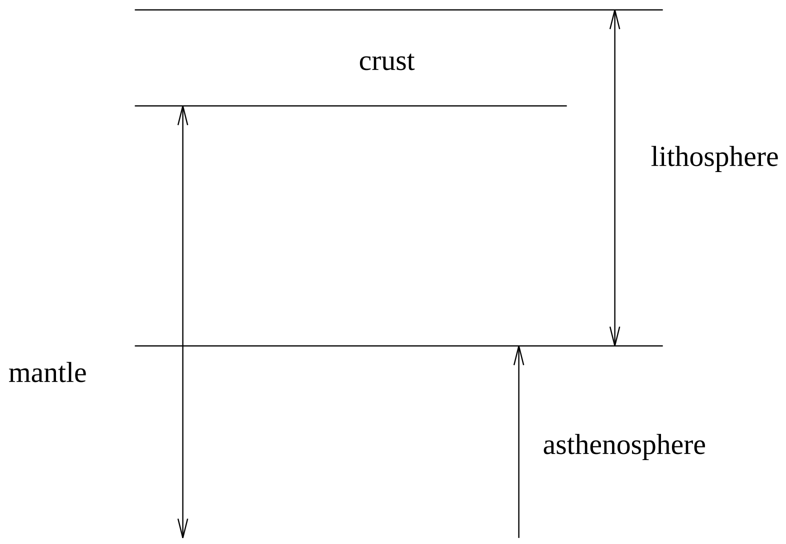

Figure 6.11:

thicknesses are probably 50 km or larger (maybe much larger) under the continents, and possibly somewhat less than 50 km under the oceans. The density contrast between the asthenosphere and the overlying lithosphere is not nearly as pronounced as between the crust and mantle. But, neither is it entirely negligible. The lithosphere is denser than the asthenosphere (the lithosphere is cooler). How can you have a denser material on top and still have a stable configuration? You can't. In fact, one of the most important forces driving continental drift (maybe *the* most important) is the downward gravitational pull on the lithosphere at subduction zones.

What evidence is there of the lithosphere/asthenosphere boundary and its depth? We expect to find a boundary (or, at least a region of transition between solid and fluid) somewhere down there. We know the earth's outer surface is brittle — otherwise all topography would flow away. And, we know the interior must be fluid at long time periods. Otherwise there would be no convection. So, there must be a boundary — or at least a transition region. Why not at the crust/mantle boundary, as assumed by Airy compensation?

Part of the reason is that seismic travel time observations show a layer of material with low shear velocities and low Q (ie. high attenuation) between about 100 and 200 km depth. This suggests that region may be hot and, quite possibly, a region of partial melt. Conceivably, this depth (100–200 km) is a dividing line between brittle and less brittle material. This depth is below the Moho. In fact, the shear velocity actually *increases* below the Moho. But, seismic observations sample the earth at periods from seconds to minutes. Over thousands to millions of years, material above this dividing line might flow. So, the lithosphere/asthenosphere boundary region is apt to be above the 100–200 km depth.

In fact, one of the best ways to determine the depth to the lithosphere/asthenosphere boundary — or at least the depth appropriate for the time scales of mantle convection — is with gravity and leveling observations. That's what this section is all about.

The idea is to assume there is a brittle lithosphere over a fluid asthenosphere. Assume the crust/mantle boundary, with its significant density contrast, is inside the lithosphere somewhere. Maybe it *is* coincident with the lithosphere/asthenosphere boundary. That's one of the things you'd like to find out. Suppose you put a mass load on the crust. The lithosphere bends, supported below by buoyancy forces from the asthenosphere. You can see the bending in two ways. First, the surface around the load might show the effects. So, by mapping the surface you might learn something about the bending. And, when the lithosphere bends, so does the crust/mantle boundary embedded in it. Since there is a density contrast at that boundary (the smaller density contrast at the lithosphere/asthenosphere boundary can usually be ignored), the bending shows up in gravity (remember, it's this mechanism which is replacing Airy compensation). So, by looking at topography and gravity, you can learn about the lithospheric plate and its thickness.

All of this is really only useful — with a few exceptions — for finding the thickness beneath the oceans. For the continents, erosion quickly smooths out the slight topography caused by the bending. And there can easily be significant lateral variations in density in continental areas (due to chemical differences) which can obliterate the small part of the gravity signal needed to find the thickness.

6.6.2 Theory of Flexure

To model the bending process described above, we need to find equations describing the bending of thin plates under a surface load.

Suppose we have an initially horizontal plate of thickness H. The plate's surface is perpendicular to the $\hat{e}_z$ axis, and the plate extends to infinity in the $\hat{e}_x$ and $\hat{e}_y$ directions. The undeformed upper surface is at $z = H/2$, and the lower surface is at $z = -H/2$. The plate is homogeneous (uniform thickness and material properties). See Figure 6.12.

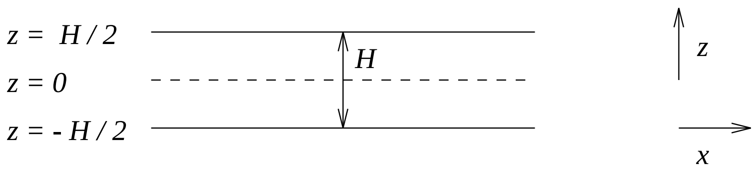

Figure 6.12:

Suppose we apply pressure loads on the upper and lower surfaces of the plate: $Q_1(x)$ above and $Q_2(x)$ below, see Figure 6.13.

We assume Q_1 and Q_2 are independent of y, so that $Q_1(x)$ for a fixed x really represents an infinite line load in the $\hat{e}_y$-direction. In our later applications, Q_1 will be due to the weight of some topographic feature, and Q_2 will represent upward buoyancy forces from the asthenosphere. How does the plate deform? Specifically, what is the shape of the deformed surface?

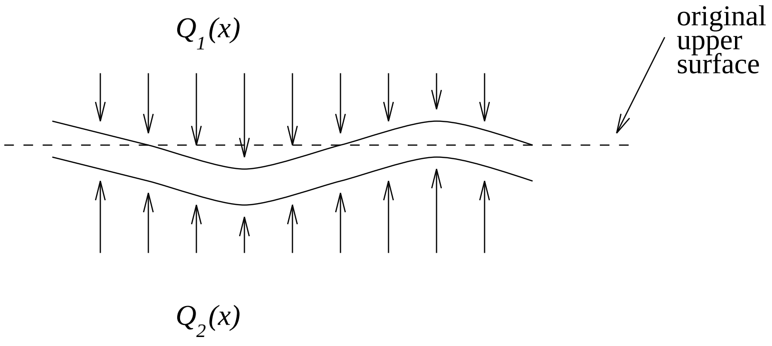

Figure 6.13:

We will make two assumptions:

1. The plate does not deform much. Specifically, the displacements are $\ll H$.

2. The horizontal wavelengths of Q_1 and Q_2 are much larger than the plate thickness, H. In other words, the plate is "thin."

Let $\omega(x)$ = distance between the deformed upper surface and the original upper surface, see Figure 6.14. $w(x)$ is positive if the surface at x has been displaced downward

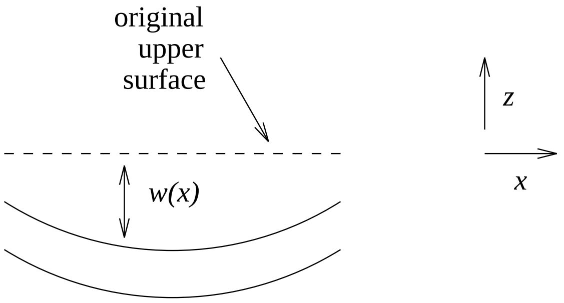

Figure 6.14:

from its undeformed position. I want to find a differential equation for $\omega(x)$ in terms of $Q_1(x)$ and $Q_2(x)$.

First, by symmetry, there can be no displacements in the $\hat{e}_y$ direction, and no variable can depend on y. Among the consequences are that the derivative of any variable with respect to y must vanish; and that the strain components $\epsilon_{xy} = \epsilon_{yy} = \epsilon_{zy} = 0$ (this last conclusion follows from the relation, Equation 5.4.1, between the strain tensor and the displacement field). And, using these results for ϵ in Equation 5.3.5, shows that $\tau_{xy} = \tau_{zy} = 0$,

Second, we assume the plate is at equilibrium, so that there is no motion. And, we ignore all body forces on the plate, including gravity. Then $\nabla \cdot \overleftrightarrow{\tau} = 0$ throughout the plate. This implies that:

$$\partial_x \tau_{xx} + \partial_z \tau_{xz} = 0 \tag{6.6.1}$$

$$\partial_x \tau_{xz} + \partial_z \tau_{zz} = 0. \tag{6.6.2}$$

Suppose we integrate Equation 6.6.2 vertically through the plate. Then:

$$\int_{-H/2}^{H/2} \partial_x \tau_{xz}\, dz + \tau_{zz}\,(z = H/2) - \tau_{zz}\,(z = -H/2) = 0.$$

(Here we can use upper and lower limits on z of $\pm H/2$, instead of using the z-coordinates of the deformed surface. The difference is only second order in the deformation, because τ is already first order.)

But

$$\begin{aligned} \tau_{zz}(z = \frac{H}{2}) &= -Q_1(x) \\ \tau_{zz}(z = -\frac{H}{2}) &= -Q_2(x). \end{aligned}$$

So:

$$\int_{-H/2}^{H/2} \partial_x \tau_{xz}\, dz = Q_1(x) - Q_2(x). \tag{6.6.3}$$

Next, suppose we multiply Equation 6.6.1 by z and integrate vertically. Then:

$$\partial_x \left[\int_{-H/2}^{H/2} z \tau_{xx}\, dz \right] + \int_{-H/2}^{H/2} z \partial_z \tau_{xz}\, dz = 0. \tag{6.6.4}$$

Integrating the right-hand integral in Equation 6.6.4 by parts gives:

$$\begin{aligned} \int_{-H/2}^{H/2} z \partial_z \tau_{xz}\, dz &= \int_{-H/2}^{H/2} \partial_z (z \tau_{xz})\, dz - \int_{-H/2}^{H/2} \tau_{xz}\, dz \\ &= (z\tau_{xz})\,|_{z=H/2} - (z\tau_{xz})\,|_{z=-H/2} - \int_{-H/2}^{H/2} \tau_{xz}\, dz. \end{aligned}$$

But, $\tau_{xz}|_{z=H/2} = \tau_{xz}|_{z=-H/2} = 0$, since there are no applied shear tractions at the outer surfaces. So, Equation 6.6.4 reduces to

$$\partial_x \left(\int_{-H/2}^{H/2} z\tau_{xx}\, dz \right) = \left(\int_{-H/2}^{H/2} \tau_{xz}\, dz \right). \tag{6.6.5}$$

(You can derive these another way. Take an infinitesimally thin (dx) cross-section of the plate, see Figure 6.15. Then, Equation 6.6.3 is the condition that the net vertical force

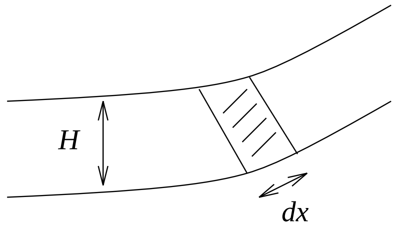

Figure 6.15:

on the cross-section vanishes, and Equation 6.6.5 is the condition that the net torque vanishes.)

Now, take ∂_x of Equation 6.6.5, and add to Equation 6.6.3. You get:

$$\partial_x^2 \left(\int_{-H/2}^{H/2} z\tau_{xx}\, dz \right) = Q_1 - Q_2. \tag{6.6.6}$$

The integral in Equation 6.6.6 is sometimes called bending moment, and written as M.

We now try to relate τ_{xx} to $\omega(x)$. First, we relate τ_{xx} to the strain ϵ_{xx}. Since $\epsilon_{yy} = 0$, we know that

$$\begin{aligned} \tau_{xx} &= (2\mu + \lambda)\epsilon_{xx} + \lambda\epsilon_{zz} \\ \tau_{zz} &= (2\mu + \lambda)\epsilon_{zz} + \lambda\epsilon_{xx}. \end{aligned}$$

We can use these equations to find ϵ_{xx} in terms of τ_{xx} and τ_{zz}:

$$\epsilon_{xx} = \frac{1}{4\mu(\mu + \lambda)} \left[(2\mu + \lambda)\tau_{xx} - \lambda\tau_{zz} \right]. \tag{6.6.7}$$

Now, I claim that for a thin plate, $\tau_{zz} \ll \tau_{xx}$. The reason is that for a thin plate, ∂_x(of anything) is $\ll \partial_z$(of anything). That is, things vary much more slowly in the horizontal direction than they do in the vertical direction, because the plate is thin. So, since $\partial_x \tau_{xz} = -\partial_z \tau_{zz}$ (from the z-component of $\nabla \cdot \overleftrightarrow{\tau} = 0$) then $\tau_{zz} \ll \tau_{xz}$. And, since $\partial_x \tau_{xx} = -\partial_z \tau_{xz}$ (from the x-component of $\nabla \cdot \overleftrightarrow{\tau} = 0$) then $\tau_{xz} \ll \tau_{xx}$. So: $\tau_{zz} \ll \tau_{xx}$. So, we ignore the τ_{zz} term in Equation 6.6.7, to obtain

$$\epsilon_{xx} = \left(\frac{2\mu + \lambda}{4\mu(\mu + \lambda)}\right) \tau_{xx} = \left(\frac{E}{1 - \nu^2}\right)^{-1} \tau_{xx}$$

if you prefer to work with E and ν. So:

$$\tau_{xx} = \frac{E}{1 - \nu^2} \epsilon_{xx}.$$

And:

$$\left(\frac{E}{1 - \nu^2}\right) \partial_x^2 \left[\int_{-H/2}^{H/2} z \epsilon_{xx} \, dz\right] = Q_1 - Q_2. \tag{6.6.8}$$

The last step is to relate ϵ_{xx} to ω (note that $\omega = -s_z$). By definition of ϵ_{ij}: $\epsilon_{xx} = \partial_x s_x$. For a thin plate, we can expand $s_x(x, z)$ in a Taylor series in z, and just keep the constant and linear terms:

$$s_x \approx a(x) + b(x) z \quad .$$

So

$$\int_{-H/2}^{H/2} z \epsilon_{xx} \, dz = \int_{-H/2}^{H/2} z(\partial_x a(x) + z \partial_x b(x)) \, dz = \frac{H^3}{12} \partial_x b(x). \tag{6.6.9}$$

Now, $b(x) \cong \partial_z s_x$. And I claim that $\partial_z s_x = \partial_x \omega$, for small deformation. To see this, consider the undeformed plate in Figure 6.16. After deformation, the upper-right-hand

Figure 6.16:

corner of the shaded region looks like Figure 6.17. Looking at the angles in Figure 6.17, gives $-\partial_x \omega = -\partial_z s_x$. So:

$$\int_{-H/2}^{H/2} z \epsilon_{xx} \, dz = \frac{H^3}{12} \partial_x^2 \omega \tag{6.6.10}$$

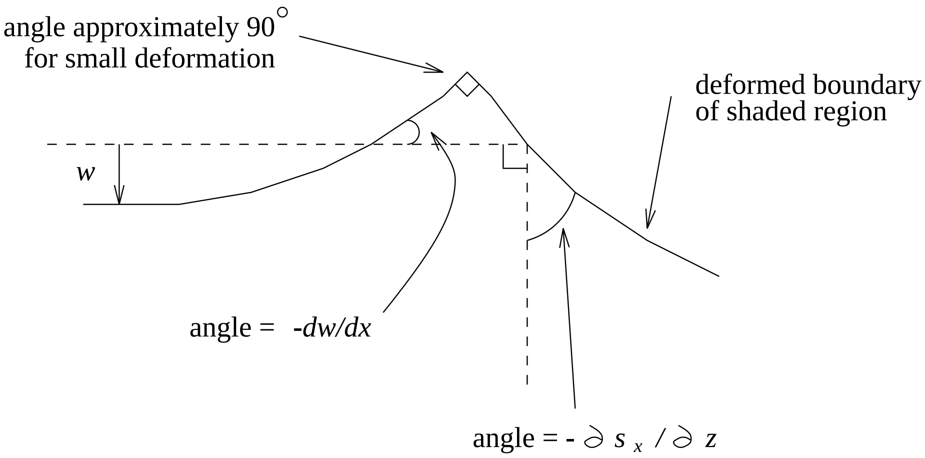

Figure 6.17:

and Equation 6.6.8 becomes:

$$Q_1 - Q_2 = \frac{EH^3}{12(1-\nu^2)} \partial_x^4 \omega.$$

Or, defining

$$D \equiv \frac{EH^3}{12(1-\nu^2)} \equiv \text{“flexural rigidity,”}$$

$$D\partial_x^4 \omega = Q_1 - Q_2. \tag{6.6.11}$$

Given loads Q_1 and Q_2 on a thin plate, we can solve Equation 6.6.11 to find the "deflection," ω. This is the result I was after.

6.6.3 Application

For us, the principal application of all this is: suppose the lithosphere is a thin plate, floating on a fluid asthenosphere. Suppose the crust/mantle boundary is inside the lithosphere, a distance d below the top. See Figure 6.18. Assume the asthenosphere and lower lithosphere have the same density, ρ_m (the actual density contrast is non-zero, but is small). The crustal density is ρ_c.

We load the lithosphere by placing crustal material (density ρ_c) with height $h'(x)$ on the surface, as shown in Figure 6.19. The outer surface and crust/mantle interface get displaced downward a distance ω. The depression at the outer surface might show up in

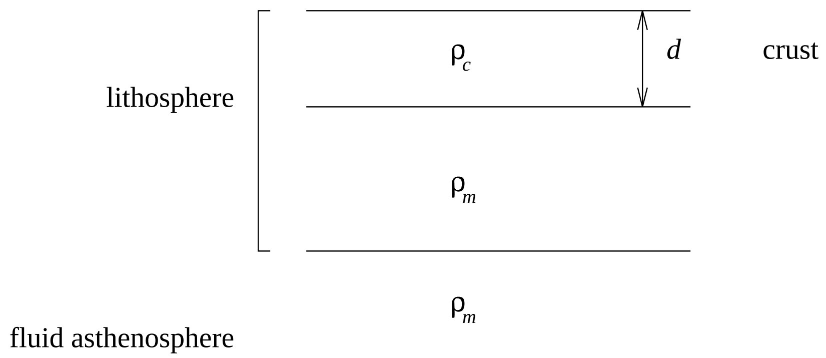

Figure 6.18:

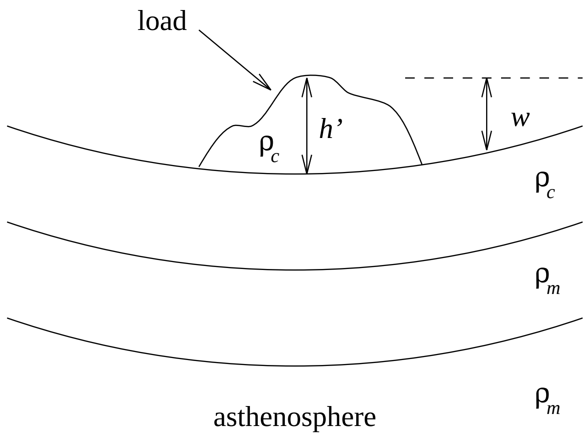

Figure 6.19:

the surface topography (especially if the surface is the sea floor, so there's no erosion). And the depression at the crust/mantle boundary might affect surface gravity.

To find the differential equation for ω, we need to relate Q_1 and Q_2 to h' and ω. The pressure load on the upper surface is $Q_1 = \rho_c g h'(x)$. It is harder to find Q_2 (= pressure on lower surface of lithosphere). This pressure is due to buoyancy: the fluid asthenosphere holds up the lithosphere.

Since the asthenosphere is fluid, the only possible horizontal internal force is the horizontal gradient of the pressure (gravity, the other force in the system, is vertical). If there is no motion in the fluid, then all forces inside the fluid must cancel, and so there can be no horizontal pressure gradients. Thus, the pressure is constant over horizontal surfaces.

A fluid asthenosphere also implies that the vertical derivative of the pressure inside the asthenosphere is the negative of the local gravitational force: $\partial_z P = -\rho_m g$. In other words, the pressure difference between two horizontal surfaces in the fluid separated by a distance z, see Figure 6.20, is

$$P_1 = P_2 + \rho_m g z.$$

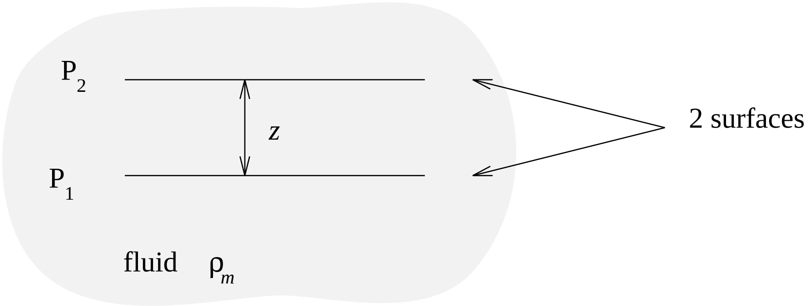

Figure 6.20:

Now, consider the deformed lithosphere sticking a distance ω down into the fluid, as in Figure 6.21. The pressure on the base of the lithosphere is $Q_2(x) = P_0 + \rho_m g \omega(x)$,

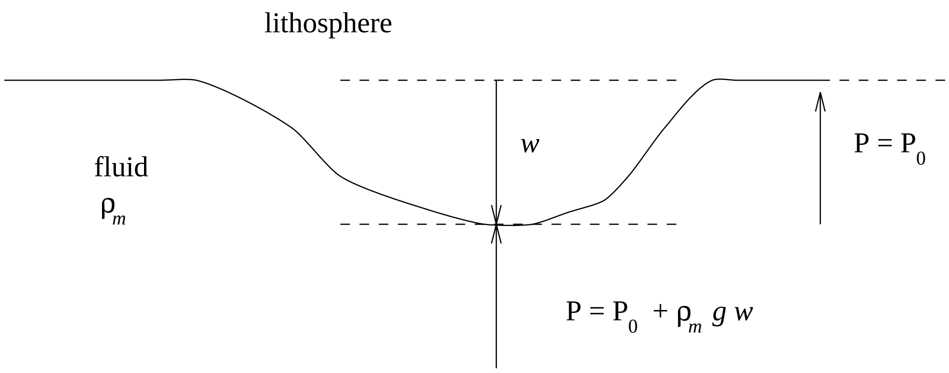

Figure 6.21:

where P_0 = fluid pressure on the base of the lithosphere at places where $\omega = 0$.

Suppose the mass load placed on the upper surface is localized in space and is finite (as all mass loads are). Then we can take $P_0 = 0$. The reason is that the total externally-applied force on the slab must vanish. Otherwise the entire slab would accelerate in the direction of the net force. The total downward force is

$$\int_{\text{lithosphere}} (Q_1(x) - Q_2(x))\,dx = \int_{\text{lithosphere}} (\rho_c g h'(x) - P_0 - \rho_m g \omega(x))\,dx = 0 \quad (6.6.12)$$

If the mass load is finite then the integral of $h'(x)$ is finite. And, if the mass load is localized then so is $\omega(x)$, and so the integral of $\omega(x)$ is also finite. Thus, the integral

$$\int_{\text{lithosphere}} P_0\,dx = P_0 \int_{\text{lithosphere}} dx \quad (6.6.13)$$

is finite. Since the domain of integration is infinite, then $\int_{\text{lithosphere}} dx = \infty$, so that the only way (6.6.13) can be finite is if $P_0 = 0$.

So, we will now assume $P_0 = 0$. In that case

$$Q_2(x) = \rho_m g \omega(x).$$

Putting Q_1 and Q_2 into Equation 6.6.11, gives:

$$D\partial_x^4 \omega + g\rho_m \omega = g\rho_c h'(x). \quad (6.6.14)$$

The best way to solve Equation 6.6.14 for arbitrary $h'(x)$, is to expand h' into a Fourier series or Fourier integral, solve Equation 6.6.14 for ω for each term in the series, and then add the ω terms together.

What does the solution look like for one of these terms? Suppose, for example, that

$$h' = h'_0 \cos(kx). \quad (6.6.15)$$

What is ω? This example is useful, also, because it tells us what happens at short and long wavelengths. In Equation 6.6.15, the only restriction we need make on k is that $2\pi/k \gg$ lithospheric thickness ($2\pi/k =$ horizontal wavelength of the load), since we assumed the horizontal wavelength $\gg$ lithospheric thickness to get our flexure equation, Equation 6.6.11.

We try a solution of the form $\omega = \omega_0 \cos kx$. This solution will work if:

$$(Dk^4 + \rho_m g)\omega_0 = g\rho_c h'_0.$$

Or:

$$\omega_0 = \left[\frac{g\rho_c}{Dk^4 + \rho_m g}\right] h'_0. \tag{6.6.16}$$

In practice, you might not know h'_0 or h', particularly over continents. $h'(x)$ describes the thickness of the load — or how much the load sticks up above the lithosphere. But, you might not know where the top of the lithosphere is.

Instead, what you might know are surface elevations, as determined from leveling. In that case, you know the distance above the geoid. You assume that before deformation the lithospheric surface was parallel to the geoid. So, if $h(x)$ = measured elevation, then $h'(x) = h(x) + \omega(x)$ (that is, the lithosphere is now $\omega(x)$ below the constant potential surface, and the load is $h(x)$ above that surface).

In that case, the differential equation for ω is

$$D\partial_x^4\omega + g\rho_m\omega = g\rho_c(h + \omega)$$

or:

$$D\partial_x^4\omega + g(\rho_m - \rho_c)\omega = g\rho_c h.$$

If $h = h_0 \cos kx$, and if we assume that $\omega = \omega_0 \cos kx$, then

$$\omega_0 = \left[\frac{g\rho_c}{Dk^4 + (\rho_m - \rho_c)g}\right] h_0. \tag{6.6.17}$$

6.6.3.1 Limiting cases

Let's see what happens to ω_0 at short and long wavelengths. Consider Equation 6.6.16, that relates ω_0 to h'_0.

Short wavelengths

Suppose k is large enough that $k^4 D \gg \rho_m g > \rho_c g$. (Though we must also assume that the wavelengths are $\gg$ lithospheric thickness, so that the derivation of Equation 6.6.11 remains valid.) Then

$$\frac{g\rho_c}{Dk^4 + \rho_m g} \ll 1,$$

so that:

$$\omega_0 \ll h_0'.$$

In other words, there is very little deflection. This is part of the reason why topography does not show up in Bouguer anomalies at short wavelengths: the compensation is small. It's *not* only that the compensation is deep.

Long wavelengths

Suppose $k^4 D \ll \rho_m g$. Then, Equation 6.6.16 is

$$\omega_0 \approx \left(\frac{\rho_c}{\rho_m}\right) h_0' = \frac{\rho_c}{\rho_m}\left(h_0 + \omega_0\right).$$

Or:

$$\omega_0 = \left(\frac{\rho_c}{\rho_m - \rho_c}\right) h_0$$

which is the familiar Airy compensation result. (ω_0 = root thickness.) This is why Airy compensation works so well at long wavelengths. The lithosphere is easy to bend — it mirrors the topography — at long wavelengths. Shear stresses within the lithosphere are not strong enough to support the load at long wavelengths. Instead, the load must be fully supported by buoyancy forces in the fluid asthenosphere — which is the Airy compensation hypothesis.

We need to be more quantitative for what is meant by "long" and "short" wavelengths. The wavelength is $\lambda = 2\pi/k$. The separation between short and long corresponds to $k^4 D = \rho_m g$. Or: $\lambda = 2\pi(D/\rho_m g)^{1/4}$. For typical oceanic values of E, ν, and H ($H \approx$ 30 km): $D \approx 2 \times 10^{30}$ dyne-cm. For typical ρ_m and g: $\lambda = 2\pi(D/\rho_m g)^{1/4} \approx 300$ km is the dividing line between short and long wavelengths. So, for the ocean

$$\begin{aligned} \lambda \ll 300 \text{ km} \quad &\Longrightarrow \quad \text{compensation is small} \\ \lambda \gg 300 \text{ km} \quad &\Longrightarrow \quad \text{Airy compensation.} \end{aligned}$$

The values are not much different than this for continents. Note that in order for all these equations to be valid, λ must also be $\gg H \approx 30$ km for the ocean. More general results, without the the thin plate assumption, show the lithosphere doesn't bend at short λ, even if $\lambda \gg H$ doesn't hold.

Go back to Equations 6.6.16 and 6.6.17. These results relating ω_0 to h'_0 and to h_0 can be inverted from the Fourier transform (k) domain back to the x-domain, to give a relation between $\omega(x)$ and $h'(x)$ or $h(x)$. In some oceanic areas, the deflection $\omega(x)$ can be measured. Then, you can fit the theory to the observations and solve for D. Using estimates for E and ν, your results for D can be used to infer H: the thickness of the thermal boundary layer for mantle convection beneath the oceans.

6.6.3.2 Example: Flexure from an island chain

We model an island chain as an infinite line load concentrated at $x = 0$, with mass/length $= M$. (If you are familiar with Dirac delta functions: $\rho_c h'(x) = M\delta(x)$.) What is $\omega(x)$ in this case? (We assumed the load is independent of y to derive Equations 6.6.16 and 6.6.17.)

We can use Equation 6.6.16 if we can expand $h'(x)$ in terms of cosines. We are assuming the island chain is infinitely-thin, so that $h'(x) = 0$ for all x except $x = 0$ — where $h'(x = 0)$ is infinite. To find the expansion coefficients, we first approximate this infinitely-thin load as a block of width 'a' (so that the height of the block is $M/\rho_c a$) centered at $x = 0$, as shown in Figure 6.22. When this block is expanded as a Fourier

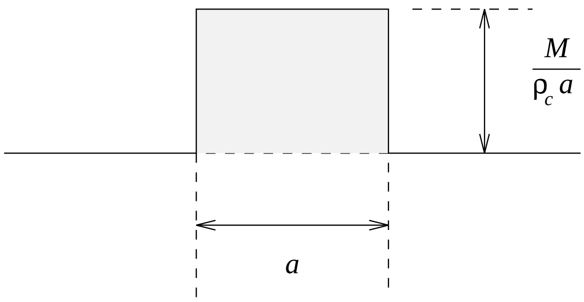

Figure 6.22:

integral of sines and cosines (I won't derive it here) we obtain:

$$h'(x) = \frac{M}{\rho_c 2\pi} \int_{-\infty}^{\infty} \frac{\sin(ak/2)}{(ak/2)} \cos(kx)\, dk.$$

If we now take the limit as $a \to 0$ (so the height $\to \infty$) we obtain (noting that $\lim_{y \to 0}(\sin y/y) = 1$):

$$h'(x) = \frac{M}{\rho_c 2\pi} \int_{-\infty}^{\infty} \cos(kx)\, dk. \tag{6.6.18}$$

We take the integral in Equation 6.6.18 as our infinitely-thin load. Note that this integral, though, is meaningless: the integral is not well-defined. We should, really, find $\omega(x)$ caused by the block of non-zero width a, and *then* set $a \to 0$ in $\omega(x)$. But, using Equation 6.6.18 as the load, without worrying about its convergence, is easier and gives the same answer.

Equation 6.6.18 implies that if we separate $h'(x)$ into cosine terms, the amplitude, $h'_0(k)$, of each cosine is $M/(2\pi\rho_c)$ — independent of k. The response, ω_0, to each cosine term is then: (from Equation 6.6.16)

$$\omega_0 = \left[\frac{g\rho_c}{Dk^4 + \rho_m g}\right] \frac{M}{2\pi\rho_c}.$$

(We use Equation 6.6.16 instead of Equation 6.6.17 because in this case the island mass you observe sticking up above the sea floor is $h'(x)$. The lithosphere does not bend at short wavelengths, and so ω directly under the load is approximately ω just outside the edges of the load — so you *do* know where the surface of the lithosphere under the load is — it is at the depth of the surrounding sea floor.)

To find $\omega(x)$, invert ω_0 back to the x-domain:

$$\omega(x) = \int_{-\infty}^{\infty} \omega_0(k)\cos(kx)\, dk = \frac{gM}{2\pi} \int_{-\infty}^{\infty} \left(\frac{\cos kx}{Dk^4 + \rho_m g}\right) dk.$$

We can find this integral in integral tables. Define $\alpha \equiv (4D/\rho_m g)^{1/4} \equiv$ "Flexural parameter." Then, we find that

$$\omega(x) = \left(\frac{gM}{8D}\right) \alpha^3 e^{-|x|/\alpha} \left[\cos\left(\frac{|x|}{\alpha}\right) + \sin\left(\frac{|x|}{\alpha}\right)\right]. \tag{6.6.19}$$

This result is symmetric about $x = 0$, as it should be. A plot of $-\omega(x)$ for $x \geq 0$ looks something like Figure 6.23. Note that the ocean floor depresses under and near to the island, but that there is a region of uplift away from the island. Let x_b be the distance between the island chain and the point of maximum uplift. Then:

$$\left.\frac{dw}{dx}\right|_{x=x_b} = 0.$$

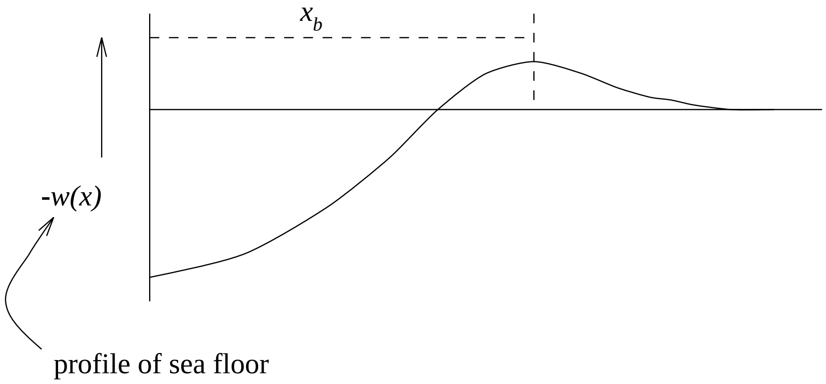

Figure 6.23:

This derivative is easy to find, and the resulting equation is $\sin(x_b/\alpha)e^{-x_b/\alpha} = 0$. Or (for $x_b \neq \infty$): $x_b = n\pi\alpha$. The point of maximum uplift closest to the load corresponds to $n = 1$, where

$$x_b = \pi\alpha.$$

Here's how these results can be used. Probably the most successful application has been to the Hawaiian-Emperor seamount chain.

1. Determine the mass/length of the seamount chain, using bathymetry data.

2. Pretend the mass is all located at $x = 0$ — so that the chain is approximated as an infinitely-thin line mass. This, of course, is not correct. The Hawaiian Islands are about 150 km wide. This means that our infinitely-thin approximation causes errors at wavelengths of less than 150 km. But, at those wavelengths the lithosphere doesn't bend much, anyway. So there should only be a small short wavelength error in $\omega(x)$. (Incidentally, this suggests a way out of another problem. Equation 6.6.16 was derived assuming the horizontal wavelengths were $\gg H$. That's not true for a line load. But, by using Equation 6.6.16 even at short wavelengths we are including no significant short wavelength terms in $\omega(x)$. And, as mentioned above, the exact solution without the thin plate assumption predicts no significant short wavelength terms in $\omega(x)$. So, our answer is pretty good.)

3. Use the result for $\omega(x)$, Equation 6.6.19, to compare with the observed bathymetry.

Specifically, find the maximum uplift point using the observations, and determine $\alpha = x_b/\pi$. Once you have α, you infer $H =$ lithosphere thickness. (Note that you don't need the mass/length, M, to do this.) One omission in the results above, is that I neglected to include the weight of the ocean water. The water fills in depressed regions and makes $\omega(x)$ larger. I won't go through the re-derivation, but will just give the result: by including the weight of the water, all results for $\omega(x)$ are the same, except that

$$\alpha \to \left[\frac{4D}{(\rho_m - \rho_w)g}\right]^{\frac{1}{4}}$$

where $\rho_w =$ water density.

For Hawaii, x_b is found to be approximately 250 km, and the solution for H is $H \approx$ 30 km.

6.6.4 Gravity

The deflection $\omega(x)$ also shows up in gravity. Surface gravity is affected by:

A. The mass load $\sigma(x) = \rho_c h'(x)$ at the upper surface.

B. The (negative) mass load $\sigma(x) = -\rho_c \omega(x)$ at the upper surface.

C. The (negative) mass load $\sigma(x) = (\rho_c - \rho_m)\omega(x)$ at the crust/mantle boundary. (The effective density here is $\rho_c - \rho_m$ because we are replacing mantle with crust, when $\omega(x) > 0$.)

If you observe $h(x) = h'(x) - \omega(x)$ instead of $h'(x)$, then you replace A. and B. with "AB: The mass load $\sigma(x) = \rho_c h(x)$ at the upper surface."

Let's assume that you measure h' instead of h (it just makes the algebra a little simpler). Then, you approximate the $\rho_c(h'(x) - \omega(x))$ upper surface load as a surface mass at $z = 0$. And, you approximate the $(\rho_c - \rho_m)\omega(x)$ load as a surface mass at $z = -d$ ($d =$ crust/mantle depth).

Suppose $h'(x) = h'_0 \cos kx$, so that Equation 6.6.16 is valid. Back in Section 6.3 we found that g_z at $z = 0$ due to a surface mass $\sigma_0 \cos(kx)$ a distance 'a' beneath $z = 0$, is $2\pi G \sigma_0 e^{-ka} \cos kx$. So, for our two surface masses:

$$\begin{aligned} g_z(x, z=0) &= 2\pi G \left[\rho_c(h'_0 - \omega_0) + (\rho_c - \rho_m)\omega_0 e^{-kd}\right] \cos kx \\ &= 2\pi G \rho_c h'_0 \cos kx \left[1 - \frac{g\rho_c}{Dk^4 + \rho_m g} + \frac{g(\rho_c - \rho_m)}{Dk^4 + \rho_m g} e^{-kd}\right] \quad (6.6.20) \end{aligned}$$

where the last equality follows from Equation 6.6.16.

For short wavelengths, where $Dk^4 \gg \rho_m g$, $g_z \to 2\pi G \rho_c h'(x) \approx 2\pi G \rho_c h(x)$ ($h' \approx h$ since $\omega \approx 0$) which is the Bouguer correction. And, for long wavelengths (where $Dk^4 \ll \rho_m g$, *and* $kd \ll 1$):

$$g_z \to 2\pi G \rho_c h'(x) \left[1 - \frac{\rho_c}{\rho_m} + \frac{\rho_c - \rho_m}{\rho_m}\right] = 0 \quad (6.6.21)$$

which is the Airy result.

More generally, given observations of $h'(x)$ (or of $h(x)$) you expand $h'(x)$ (or $h(x)$) in terms of cosines and sines, find g_z for each term in the expansion, and then add the g_z's for the different cosines and sines together. You compare with observations to find D (which gives H) *and* d (= the crustal thickness). For long wavelength loads, you'll end up with about the same estimate for d as you would if you had started with Airy compensation. That's because Equation 6.6.21 equals the Airy result.

6.6.5 Remarks

These gravity equations, which can be used to estimate H from observations of g_z, are best applied in oceanic areas. They have, for example, been used to confirm the 30 km estimate of H beneath Hawaii, as determined from the bathymetry. They are not as useful for finding H in continental areas, because there are often substantial horizontal chemical- and structural-related variations in g_z over continents, which mask the effects of deflection on the crust/mantle boundary. There are, though some things which have been successfully tried beneath continents. For example, people have looked at the gravity signal related to loading from sedimentary basins. They use seismic techniques to find the basement rock beneath the sediments. They then know the sediment thickness, which allows them to treat the sediments as a known load, estimate g_z, and compare with data.

Another thing that's been done is to collect lots of gravity data over a single continent, compare with the measured elevation h, and try to fit H to the data. When this is done for the United States, for example, the Bouguer anomalies are found to $\rightarrow 0$ as the wavelength $\rightarrow 0$. That implies that there is little gravity compensation at short wavelengths. At the other extreme, the free air anomalies $\rightarrow 0$ at large wavelengths, implying that long wavelengths are compensated.

So, the picture is qualitatively what one would expect. But, if you fit the result Equation 6.6.20 to the U.S. g_z data, and solve for H, you find that $H \approx 5$ km. That's way too small. This is largely because the most useful Bouguer anomaly results come from places with high topography. And, most of the high U.S. topography is in the west. It turns out that the western U.S. is sitting on anomalously hot material in the upper mantle. The heat has melted the lower lithosphere, and so the lithosphere *is* thinner there than you would normally expect. But, more fundamentally, thermal expansion has uplifted the entire region — so that interpreting long wavelength gravity from the west in terms of lithospheric loading may not make sense. Pratt isostasy is likely to be more appropriate.

6.6.6 Examples of Pratt compensation

6.6.6.1 Basin and Range

Let's consider the western U.S. in more detail. Specifically, we consider the Basin and Range geological province, which includes all of Nevada, plus some of California, Arizona, Idaho, and Utah. Broad-scale free air anomalies from this region are close to zero, so that Bouguer anomalies are the mirror image of the topography. This implies the region is compensated at long wavelengths. The region has high overall elevation. So, if it is Airy compensation that is at work here, you'd expect the crust to be thick underneath the region. The hypothesis that regions of high topography are underlain by anomalously thick crust, is found to hold in most places around the world. This is confirmed not only by the sort of gravity-flexure theory studies described in the previous sections. But it's also supported by seismic results for the depth to the Moho (the crust/mantle discontinuity). Those observed depths are obviously independent of any assumptions about the compensation mechanism, and they are generally large for high elevations.

But, not for the Basin and Range. There, the Moho depth is about the same as elsewhere — in low elevation regions — beneath the U.S. So, Airy compensation is not supporting the Basin and Range topography. (Which means, also, that flexure is not operative, since the effects of flexure would cause a deeper crust/mantle boundary, just as in Airy compensation.) Instead, the conclusion is that there must be some sort of Pratt compensation at work here. In other words, there is material of anomalously low density beneath the Basin and Range. This is further supported by high heat flow observed at the surface: when material is heated, its density decreases.

Is this hot material in the crust or in the upper mantle? Seismic observations show nothing particularly anomalous about seismic p-wave speeds in the Basin and Range crust. But, they do show that p-wave speeds in the upper mantle beneath the Basin and Range are slower than they are at similar depths in other regions of the globe. There is laboratory evidence of an approximately linear relationship between density and v_p, when the temperature of the material is changed: a decrease in v_p implies a decrease in ρ. This relation might appear to be the opposite of that implied by the density dependence of v_p: $v_p = \sqrt{(2\mu + \lambda)/\rho}$, which suggests that a decrease in $v_p \Rightarrow$ *increase* in ρ. But, it turns out that although an increase in temperature reduces ρ, it reduces $2\mu + \lambda$ even more. So, the observed p-wave velocities suggest that the anomalously light material beneath the Basin and Range is in the upper mantle.

What is going on in this region? The Basin and Range is so-named because it is marked with long mountain ranges trending from the northeast to the southwest, separated by wide basins. The ranges are spaced roughly 20–30 km apart. Geologists have found that at the edge of each range there is a normal fault, see Figure 6.24. Normal

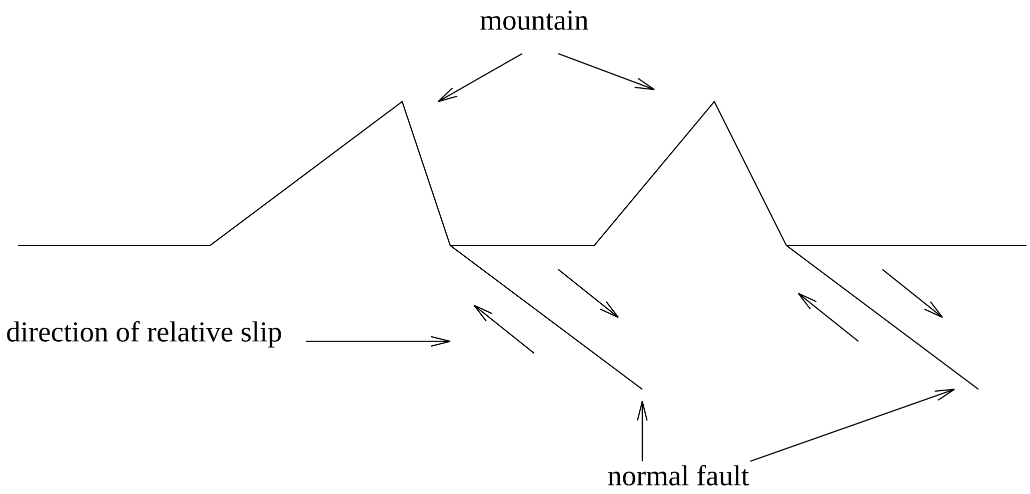

Figure 6.24:

faults occur when material is being pulled apart — or extended. In fact, that is apparently what's happening to the Basin and Range. As the region extends, normal faulting occurs. The crust and, probably, the lithosphere break up into blocks separated by the normal faults, and the blocks tip — causing mountain ranges. See Figure 6.25. Erosion

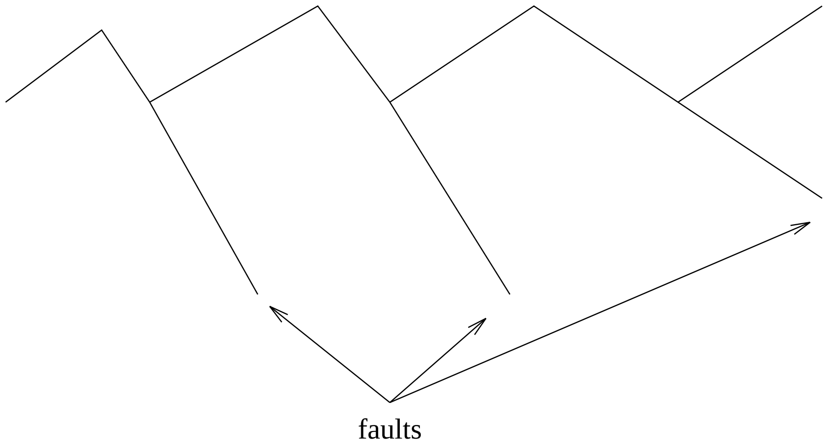

Figure 6.25:

fills in the depressions, producing sedimentary basins. In this sort of geological structure, the high points (the mountains) are called *horsts*, and the low points (basins) are called *grabens*.

Further evidence of Basin and Range extension comes from geodetic (e.g. VLBI) measurements of the on-going extension. Those measurements suggest an extension rate of about 5–7 mm/yr, which is consistent with the geologically-inferred rate over millions of years.

So, what's causing the Basin and Range to extend? Presumably it is the same mechanism that is causing the high temperatures in the upper mantle. Nobody knows for sure, but there are a number of theories. One of the more generally accepted hypotheses (though by no means the only hypothesis) runs approximately as follows. Tens of millions of years ago there was an oceanic plate between the North American and Pacific plates, called the Farallon Plate. The Farallon and Pacific plates joined at a spreading center — where Pacific sea floor was created. The eastern edge of the Farallon plate subducted beneath the western edge of the North American plate. See Figure 6.26.

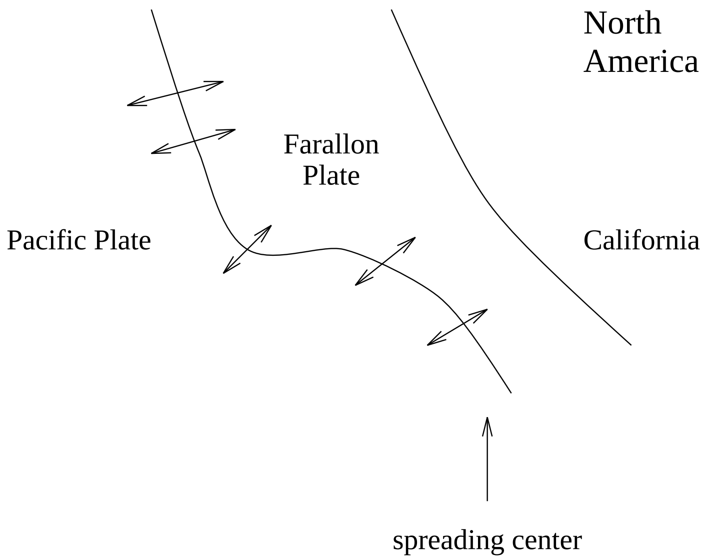

Figure 6.26:

See Figure 6.27 for a cross-section view.

The material in the North American plate just above the subducting Farallon plate could well be hot, due to frictional heating as the Farallon plate descends into the mantle. Frictional heating could be at least partly responsible for the 'back-arc' volcanism that often occurs above subducting slabs.

The North American plate was moving westward faster than the Farallon plate was

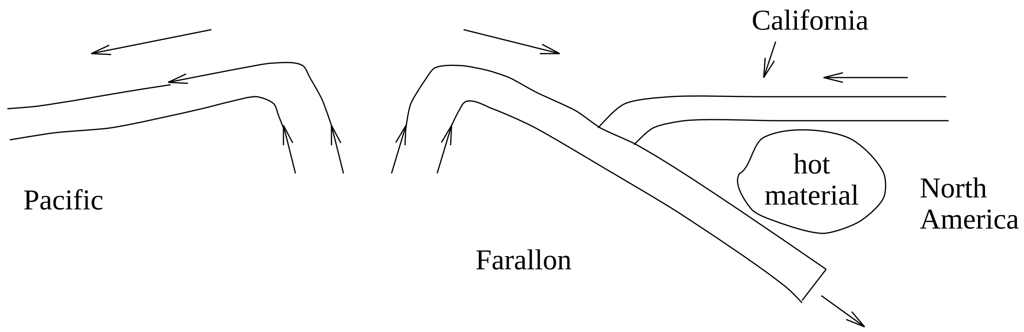

Figure 6.27:

moving eastward. So, about 30 million years ago, the spreading center began to be subducted beneath the North American plate. All that remains of the Farallon plate at the present time is the Cocos plate just off Latin America, and the Juan de Fuca plate just west of Washington. Eventually these will be subducted, too.

It is not clear, but it seems probable that the upwelling at a mid-ocean ridge turns off when the ridge is subducted. This conclusion might be hard to accept if you believe in a picture of upper mantle convection where the plates mirror the shape of the underlying upper mantle convection cells, such as illustrated in Figure 6.28.

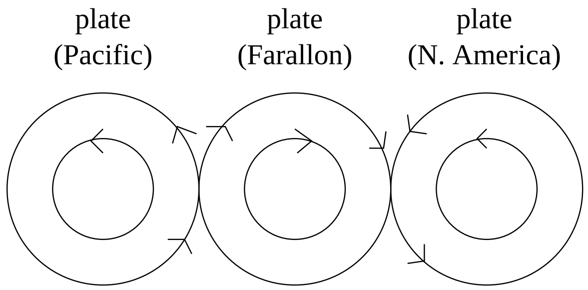

Figure 6.28:

But it is easier to understand if, instead, plate boundaries simply correspond to places where the original lithosphere was weak. For example, convection causes stresses on the lithosphere. The lithosphere could well split apart where it is weakest — not necessarily at the convection cell boundaries. Mantle material would come to the surface at the rupture to replace the separating lithospheric material, and the result is a spreading center. So, if the spreading center eventually gets subducted, the original lithospheric rupture is essentially healed, and the spreading stops.

In any case, the subduction of the Pacific/Farallon boundary is likely to have been the event that initiated Basin and Range extension. There are several ways this could have happened. As one hypothesis, when the the spreading center was subducted and the upwelling of mantle material ceased, the Farallon plate no longer pushed eastward on the North American plate. As a result, the region, which is hotter and higher in elevation than its surroundings, is better able to expand, leading to extension of the surface.

6.6.6.2 Hot Spots

Hot spots are another example of Pratt compensation.

I've mentioned the application of flexure theory to lithospheric loading by island chains. You get subsidence beneath the chain, and a region of slight uplift a couple hundred km away. People also find longer wavelength swells in the bathymetry around some islands such as Hawaii and Bermuda. These swells are maybe 1000–2000 km in horizontal extent. All the shorter wavelength wiggles due to the flexure are super-imposed on the swell. There is a corresponding swell across the island axes in free-air gravity. The free-air anomalies are small — maybe 20 mgal or so. The swell in the bathymetry can be over 1 km in elevation. So, the Bouguer correction would be approximately $(0.1118\ \mathrm{mgal/m}) \times 10^3\ \mathrm{m} \cong 100\ \mathrm{mgal}$. So, the swell must be reasonably well compensated. The crust isn't thicker under swells than elsewhere. So, the compensation is likely to be Pratt compensation of some sort. The observed free-air anomaly can be used to learn about the underlying density anomaly, by assuming a compensation depth (the depth level above which all columns have the same mass); *or* to learn about the compensation depth, by assuming a known density anomaly.

The free-air anomalies are far enough from zero over the swell, to imply that the underlying density anomaly is well below the crust mantle boundary — further down into the upper mantle. If, instead, the compensation were at shallower depths, there would likely be more cancellation between the gravitational effects of the compensation and those of the topography.

Presumably both the low density and the swell itself are caused by thermal expansion. This suggestion is supported by the observed high heat flow in these regions. In fact, in the most notable example of one of these regions — Hawaii — there is even volcanic activity.

These sorts of island arcs are believed to be caused by hot plumes of material rising up out of the lower mantle. The surface expressions of the plumes are called hot spots. It's not clear what causes the plumes or where they come from; they could well originate from near the core/mantle boundary.

Several dozen tentative hot spots have been identified around the globe. They occur on continents as well as on oceans. They appear to remain fixed with respect to each other over tens to hundreds of millions of years or longer. The plates move over the hot spots, and so a hot spot leaves a track behind it. For example, the Hawaiian islands are the end of a long chain of islands and seamounts. The seamounts — at the other end of the Hawaiian island chain — used to be islands when they lay over the hot spot. When they moved off the hot spot, the entire region subsided again and the islands ended up as under-water seamounts. Because hot spots do not appear to undergo relative motion, they have been useful in determining an absolute reference frame to describe plate motion.

6.6.6.3 Mid-Ocean Ridges

Here's another example of Pratt compensation. Plates are pulling apart at mid-ocean ridges. The bathymetry at these ridges shows under-water mountains and other short wavelength features caused by the upwelling mantle material. But, all this is superimposed on a very long wavelength swell, several thousand km wide — essentially across the entire ocean basin, or at least over much of it.

The free-air gravity anomalies show that the swell is mostly compensated — although the compensation is not perfect. The departure from perfect compensation can be used to learn about the density, the compensation depth, etc. The anomalies are large-scale enough that they must first be corrected for the horizontal variation of the geoid. Remember: the free air correction only corrects for the variation of gravity due to the elevation of the gravimeter above the geoid. It does not correct for the variations in the radius of the geoid. The geoid variations usually have much longer wavelengths than the interesting gravity signal. But not in this case. In fact this swell is so broad that people often work with it by looking at its effect on the *geoid*, rather than on surface gravity.

People have been able to explain the swell in the bathymetry, and the observed geoid anomalies, as the result of thermal subsidence as the lithosphere moves away from the ridge where it was created, and then cools and thickens. See Figure 6.29. It thickens,

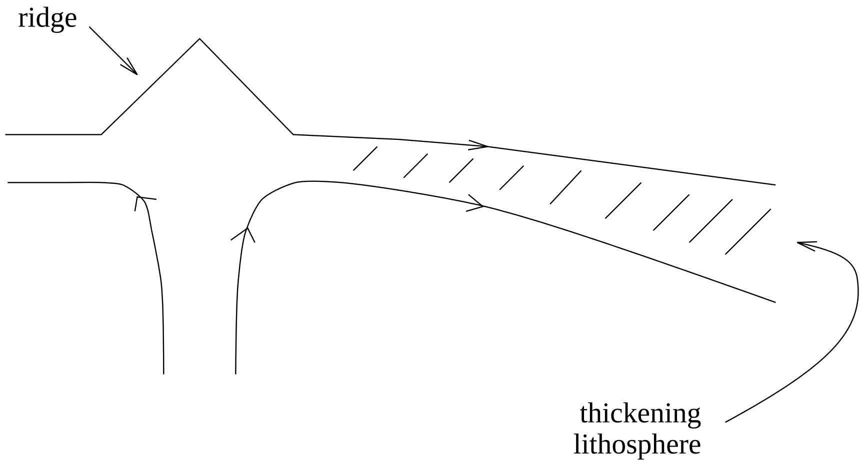

Figure 6.29:

because material from the underlying asthenosphere cools and becomes solid — and so becomes part of what we call the lithosphere. It subsides because as it cools its density increases — it takes up less volume. The amount of subsidence is, presumably, consistent with the isostatic assumption of equal mass in all vertical columns.

People observe that the ocean depth (which reflects the subsidence) varies as the square root of the distance from the ridge. For constant spreading rates, the distance from the ridge is proportional to the age of the lithosphere. So, the ocean depth varies as $\sqrt{\text{age}}$.

People observed this $\sqrt{\text{age}}$ dependence back before the theory of continental drift. Nowadays, people can explain it using simple heat flow models. This is one of the major triumphs of the theory of continental drift, and I will derive it here. The phenomenon is referred to as *Thermal Isostasy*.

6.6.6.3.1 Thermal Isostasy The model is simple. We take a sample two-dimensional cross-section, and follow it as it moves away from the ridge axis. See Figure 6.30. We find the temperature profile as a function of time. We do this by solving the equations

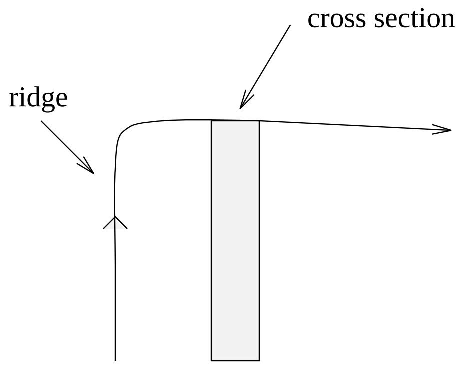

Figure 6.30:

for heat conduction inside the cross-section, assuming all heat conduction is vertical, so that the column does not interact thermally with its neighbors. This assumption is probably pretty good since the temperature gradient in the lithosphere will be predominately vertical: the outer surface is much cooler than the interior of the mantle.

So, the problem reduces to a one-dimensional conduction problem. See Figure 6.31. Let z be positive downwards. Let $z = 0$ be the top of the column (the surface of the

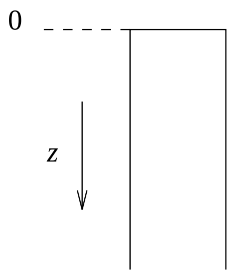

Figure 6.31:

earth). Assume the column extends to $z = \infty$. The one-dimensional equation for heat conduction in the $\hat{e}_z$ direction is:

$$\partial_t T = \kappa \partial_z^2 T \tag{6.6.22}$$

where T = temperature and κ = thermal diffusivity. Equation 6.6.22 assumes there are no internal heat sources or sinks. We must solve this equation for given initial conditions and boundary conditions:

Initial condition:

$T = T_m =$ constant at $t = 0$. (Except we assume that $T(z=0, t=0) = T_s$ — see Equation 6.6.23 below.) Here, $T_m =$ temperature of the mantle, where the column originated. We assume the vertical gradient of the initial temperature is zero (i.e. $T =$ constant) because convection in the earth is much faster than conduction: the column rises to the surface much faster than it can lose heat through the outer ($z = 0$) surface.

Boundary conditions:

For all $t > 0$:

$$\begin{aligned} T(z=0) &= T_s = \text{ surface temperature} \qquad (6.6.23) \\ T(z=\infty) &= T_m. \end{aligned}$$

I claim the solution to the partial differential equation with these initial and boundary conditions is unique, and is given by:

$$T(z,t) = T_m + (T_s - T_m)\left[1 - \text{erf}\left(\frac{z}{2\sqrt{\kappa t}}\right)\right] \qquad (6.6.24)$$

where

$$\text{erf}(\eta)(\equiv \text{ error function}) = \frac{2}{\sqrt{\pi}}\int_0^{\eta} e^{-\eta_0^2}\, d\eta_0.$$

To verify that Equation 6.6.24 solves the initial and boundary conditions, note that $\text{erf}(\infty) = 1$ (since $\int_0^\infty e^{-\eta_0^2}\, d\eta_0 = \sqrt{\pi}/2$) and $\text{erf}(0) = 0$. So,

$$\begin{aligned} T(z,0) &= T_m + (T_s - T_m)\underbrace{[1 - \text{erf}(\infty)]}_{=0} &= T_m \\ T(0,t) &= T_m + (T_s - T_m)\underbrace{[1 - \text{erf}(0)]}_{=1} &= T_s \\ T(\infty,t) &= T_m + (T_s - T_m)\underbrace{[1 - \text{erf}(\infty)]}_{=0} &= T_m. \end{aligned}$$

To verify that Equation 6.6.24 solves the differential equation, we use the results:

$$\begin{aligned}
\partial_t T &= -(T_s - T_m)\left[\partial_\eta \mathrm{erf}(\eta)|_{\eta=z/2\sqrt{\kappa t}}\right]\frac{z}{2\sqrt{\kappa}}\left(-\frac{1}{2t^{3/2}}\right) \\
\partial_z T &= -(T_s - T_m)\left[\partial_\eta \mathrm{erf}(\eta)|_{\eta=z/2\sqrt{\kappa t}}\right]\frac{1}{2\sqrt{\kappa t}} \\
\partial_z^2 T &= -(T_s - T_m)\left[\partial_\eta^2 \mathrm{erf}(\eta)\Big|_{\eta=z/2\sqrt{\kappa t}}\right]\frac{1}{4\kappa t}.
\end{aligned}$$

And:

$$\partial_\eta \mathrm{erf}(\eta) = \frac{2}{\sqrt{\pi}}e^{-\eta^2}.$$

So:

$$\partial_\eta^2 \mathrm{erf}(\eta) = -\frac{4}{\sqrt{\pi}}\eta e^{-\eta^2} = -\frac{2}{\sqrt{\pi}}\frac{z}{\sqrt{\kappa t}}e^{-\eta^2}.$$

So:

$$\begin{aligned}
&-\partial_t T + \kappa\partial_z^2 T = \\
&\qquad -(T_s - T_m)\frac{2}{\sqrt{\pi}}e^{-\eta^2}\Big|_{\eta=z/2\sqrt{\kappa t}}\frac{z}{4\sqrt{\kappa t}\,t} \\
&\qquad + \kappa(T_s - T_m)\frac{2}{\sqrt{\pi}}\frac{z}{\sqrt{\kappa t}}\,e^{-\eta^2}\Big|_{\eta=z/2\sqrt{\kappa t}}\frac{1}{4\kappa t} \\
&\qquad = 0
\end{aligned}$$

So, Equation 6.6.24 does, indeed, solve the differential equation.

We now have an expression for the temperature $T(z,t)$. What is the density? For small temperature variations, $(T - T_m)$:

$$\rho \approx \rho_m\left[1 + \alpha(T_m - T)\right]$$

where ρ_m = density at the mantle temperature T_m, and α = volumetric coefficient of thermal expansion. So, if $\Delta\rho \equiv [\rho(z,t) - \rho_m]$ is the increase in density, then:

$$\Delta\rho = \rho_m\alpha(T_m - T_s)\left[1 - \mathrm{erf}\left(\frac{z}{2\sqrt{\kappa t}}\right)\right]. \qquad (6.6.25)$$

Note that $T_m > T_s$, $\alpha > 0$ (a decrease in T causes an increase in ρ), and $\mathrm{erf}(\eta) \leq 1$ (with $\mathrm{erf}(\eta) = 1$ only if $\eta = \infty$). So, $\Delta\rho \geq 0$, implying an increase in density throughout the column. In fact, since $\text{erf}(\eta)$ decreases with decreasing η, we can conclude that $\Delta\rho$ increases with time. That suggests that the column should subside with time, due to isostasy. To estimate that subsidence, we impose the isostatic condition of equal masses in vertical columns.

In Figure 6.32, the column on the right has subsided by h. Since all this is under

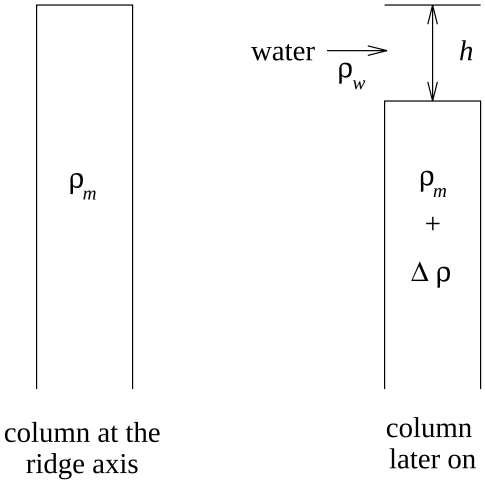

Figure 6.32:

water, the height h is filled up with water of density ρ_w. There should be equal masses in these two columns. That's nonsense, because each column has infinite mass. But that's just an artifact of our infinite column model.

Instead, just consider the *difference* in mass between the two columns. The left hand solid column has a uniform density of ρ_m between $z = 0$ and $z = \infty$. The right hand *solid* column has density $\rho_m + \Delta\rho$ between $z = 0$ and $z = \infty$ (the reference level $z = 0$ goes up and down with the top of the column). So, the difference in density is $\Delta\rho$ between $z = 0$ and $z = \infty$. *Except*, the right hand *solid* column is a length h shorter than the left hand column. That is, it is missing the mass $\rho_m h$. It has replaced the solid mass with water, so now has added the mass $\rho_w h$.

So, the extra mass in the right hand column is

$$\int_0^\infty \Delta\rho(z,t)\,dz - \rho_m h + \rho_w h = 0$$

(the extra mass is set to zero, here, because of our isostatic assumption). Or:

$$(\rho_m - \rho_w)h = \int_0^\infty \rho_m \alpha (T_m - T_s) \left[1 - \operatorname{erf}\left(\frac{z}{2\sqrt{\kappa t}} \right) \right] dz.$$

It turns out that

$$\int_0^\infty [1 - \operatorname{erf}(x)]\, dx = \frac{1}{\sqrt{\pi}} \tag{6.6.26}$$

Using Equation 6.6.26 gives $h(t)$ = depth of ocean at time t:

$$h(t) = \alpha (T_m - T_s) 2 \sqrt{\frac{\kappa t}{\pi}} \left(\frac{\rho_m}{\rho_m - \rho_w} \right). \tag{6.6.27}$$

Typically, $h(t) \approx$ 3–4 km depth a long way from the ridge. The important result is that h is proportional to $\sqrt{t}$, as is observed. (Alternatively, because the distance from the ridge = (plate velocity) $\times\, t$, then h is proportional to the square root of that distance.)

People find that the $\sqrt{t}$ dependence of h tends to disappear a few thousand km from the ridge. There, the ocean floor flattens out. Probably the background heat flow through the entire ocean floor counteracts the cooling described here, and keeps the ocean floor elevated to some minimal level.

So, we now have a model of the bathymetry. What about the geoid? Specifically, what effect does $\delta\rho$ have on the geoid height N?

The situation here is an example of what was referred to as "Generalized Pratt compensation" in Section 6.3. There, we considered the case of topography with an x-dependence of e^{ikx}. If $\delta\rho(z')e^{ikx}$ is the density anomaly that supports the topography, and if the compensation is perfect (as it is here), then the "Generalized Pratt" result in Section 6.5.1 is:

$$N(x) = \frac{2\pi G}{g} \left(\int_{-H}^{0} \delta\rho(z') z' \, dz' \right) e^{ikx} \tag{6.6.28}$$

where H is the compensation depth (that is, where $\delta\rho(z')e^{ikx} = 0$ below $z' = -H$). Equation 6.6.28 is valid so long as the horizontal wavelength is much larger than the compensation depth: $kH \ll 1$.

Since the amplitude of $N(x)$ is independent of k, we can generalize Equation 6.6.28 to arbitrary x dependence in $\delta\rho$ (so long as it is long wavelength) to obtain:

$$N(x) = \frac{2\pi G}{g} \int_{-H}^{0} \delta\rho(x, z') z' \, dz'. \tag{6.6.29}$$

How do we modify this result so that it is appropriate for our thermal isostasy problem? First, the $\delta\rho$ in Equation 6.6.29 is different from our $\Delta\rho$: $\delta\rho$ *also* includes the mass deficiency $(\rho_w - \rho_m)$ in the depth $h(x)$, due to replacing rock with water. Second, our depth of compensation is at ∞, and our z-axis is positive downwards. So:

$$N(x) = \frac{2\pi G}{g}\left[\int_{h(x)}^{0}(\rho_w - \rho_m)z\,dz + \int_{\infty}^{h(x)}\Delta\rho z\,dz\right]. \tag{6.6.30}$$

For the second integral in Equation 6.6.30 we can replace the upper limit $h(x)$ with 0, since h and $\Delta\rho$ are both small.

The second integral in Equation 6.6.30 needs some more justification. We needed to assume $|kH| \ll 1$ to derive Equation 6.6.30, but for $|H| = \infty$ this is obviously wrong. On the other hand, for the thermal isostasy case, $\Delta\rho$ goes to 0 pretty rapidly as z gets large. Most of the density anomaly is near the surface. So, all we really need to assume to get a pretty reliable answer, is that the horizontal wavelength is $\gg$ the thickness of this near surface layer. And, that's a pretty good approximation in our case since the wavelength of the swell is several thousand km. So, Equation 6.6.30 is probably pretty good.

Using our result (Equation 6.6.25) for $\Delta\rho$, we get:

$$N = \frac{2\pi G}{g}\left[-\frac{(\rho_w - \rho_m)h^2}{2} + \rho_m\alpha(T_m - T_s)\int_{\infty}^{0}\left[1 - \mathrm{erf}\left(\frac{z}{2\sqrt{\kappa t}}\right) z\,dz\right]\right]. \tag{6.6.31}$$

From integral tables, we find that: $\int_0^\infty \eta[1 - \mathrm{erf}(\eta)]\,d\eta = \frac{1}{4}$. So, simplifying the integral in Equation 6.6.31 and using Equation 6.6.27, for h in terms of t, gives N as a function of age:

$$N(t) = -\left[\frac{2\pi G\rho_m\alpha(T_m - T_s)\kappa}{g}\right]\left[1 + \frac{2\rho_m\alpha(T_m - T_s)}{\pi(\rho_m - \rho_w)}\right]t. \tag{6.6.32}$$

Note that $N(t)$ is linearly proportional to the age, t (or, equivalently, to the distance from the ridge). This linear dependence agrees well with what is observed. Note, also, that $N(t)$ is < 0, so N decreases away from the ridge axis.

The effects of the $\sqrt{t}$ bathymetry have been noticed in another way, near transform faults along mid-ocean ridges. Mid-ocean ridges are not strictly linear — but consist of linear segments separated by transform faults, as in Figure 6.33. Along a transform fault,

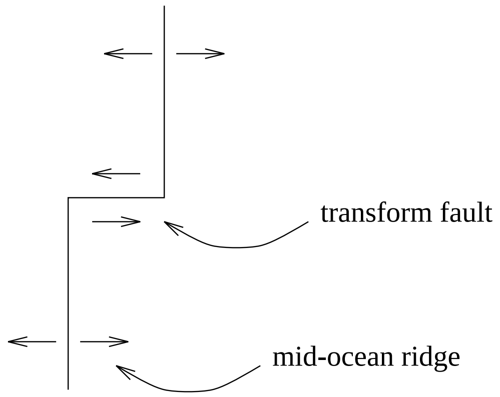

Figure 6.33:

the direction of motion is parallel to the fault. Along the continuation of a transform fault, material on different sides will be of different ages. See Figure 6.34. Because of the time-dependent subsidence, the younger material will have greater elevation. So, if we

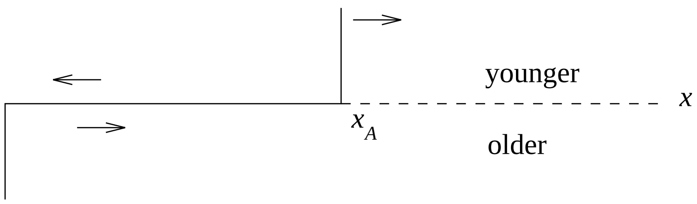

Figure 6.34:

look at Figure 6.34 in cross-section from along the x-axis, we'd see Figure 6.35. At least,

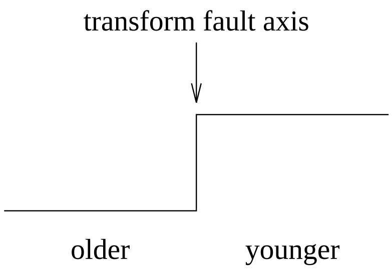

Figure 6.35:

this is what we'd see for a cross-section right next to the upper arm of the spreading ridge (at x_A in Figure 6.34). At this point the younger lithosphere has just been created and has just come in contact with the older lithosphere on the other side of the fault axis. The elevation difference is determined by the age of the older lithosphere at this time.

As the cross-section moves away from the ridge, both sides cool and subside. But, the subsidence is not linear in time — it varies as $\sqrt{t}$. This implies that the elevation difference between the two sides should decrease with time.

People first noticed and explained this sort of feature using bathymetry results inferred from satellite altimeter data. Offsets in elevation across the fault axis can be as large as 2 km. The results confirm the $\sqrt{t}$ age dependence of the subsidence. But, the results also show something else. As a cross-section moves away from the ridge, the elevation difference between material on the two sides does decrease, consistent with the $\sqrt{t}$ subsidence. But *not* right across the transform fault axis. There, the elevation difference remains fixed — determined by the initial elevation difference when the two sides first came in contact. The pattern across the fault looks like that shown in Figure 6.36. The

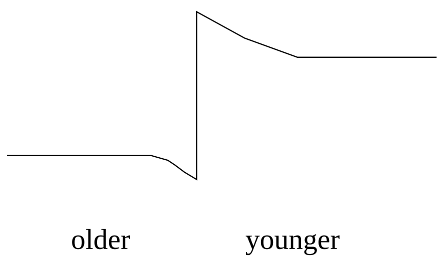

Figure 6.36:

size of the discontinuity between the older and younger rocks shown in Figure 6.36 is the initial offset.

This shows that the lithosphere can support shear stress over many tens of millions of years (you see this feature in cross-sections many tens of millions of years old). The slopes in Figure 6.36 in the bathymetry near the fault are determined by how thick the lithosphere is. For example, thin lithosphere is easier to bend and so the slopes are larger than for thick lithosphere. In fact, people have used thin plate flexure theory to try and reproduce the observed slopes by varying the flexural rigidity parameter D. (In this application, the plate surface is required to have the observed discontinuity in elevation right at the fault, and also to have the observed elevation difference between $x = \infty$ and $x = -\infty$.) From this they infer the thickness of the lithosphere.

One more observation of the $\sqrt{t}$ result comes from observations of island arc loading from the Hawaii-Emperor Seamount chain, and other island arcs. I described how you can infer the lithospheric thickness from observations of bathymetry and gravity around island arcs. For example, you measure the distance between the arc axis and the point of maximum uplift that usually occurs a couple hundred km away. When people do this for different islands in the Hawaii-Emperor chain, they infer different lithospheric thicknesses. The apparent thickness is correlated with the age of the sea floor at the time the island was formed (as determined by geological dating). In fact, the observed thickness appears to be proportional to $\sqrt{\text{that age}}$.

This says two things:

1. The results are consistent with the thermal isostasy model, which predicts lithosphere thicknesses which depend on age as $\sqrt{t}$. The idea is that the lithosphere/asthenosphere boundary is presumably determined by temperature. It is an isotherm, so that its depth z satisfies: $T(z) = \text{constant} \Rightarrow z/2\sqrt{\kappa t} = \text{constant}$.

2. The lithospheric thickness which determines the flexural response is evidently the thickness at the time of loading, *not* the thickness at the present time (since the present thicknesses are larger than the original thicknesses).

6.7 Convergence Zones

We've described the interpretation of gravity anomalies and bathymetry using flexure models (essentially Airy compensation) and Pratt models (such as thermal isostasy). Now, as a final example, we consider a situation that involves both sorts of models — and includes some uncompensated loading.

People long ago observed large (typically 150–200 mgal, and sometimes even larger) negative free-air anomalies over oceanic trenches, with somewhat smaller but still large (20–80 mgal) positive anomalies on either side of the trench. These signals are large enough that they show up clearly in maps of the oceanic geoid, even though they are relatively short-wavelength features: typically 50-100 km in width. The negative geoid anomaly over a trench, for example, is typically 10-20 m. The maximum negative gravity and geoid anomalies are often centered a few km away from the point of maximum sea floor depth, rather than directly over it.

The negative gravity anomaly over a trench is very large. It implies that the trench is nowhere close to being compensated. Since the trench is a depression in the earth's surface, an uncompensated trench *would* cause a negative free air anomaly. But the observed negative anomaly has an amplitude that in most cases is too large to be explained even by an uncompensated trench. In other words, the Bouguer anomaly is usually still negative so that there must be a mass deficiency somewhere beneath the trench. Before the theory of plate tectonics was developed, this characteristic gravity signature of trenches was a mystery. Nowadays, the ideas of plate tectonics and the role of subducting slabs in the convection process, can be used to explain the observations quite nicely.

At a trench, one plate is diving down beneath the other. See Figure 6.37.

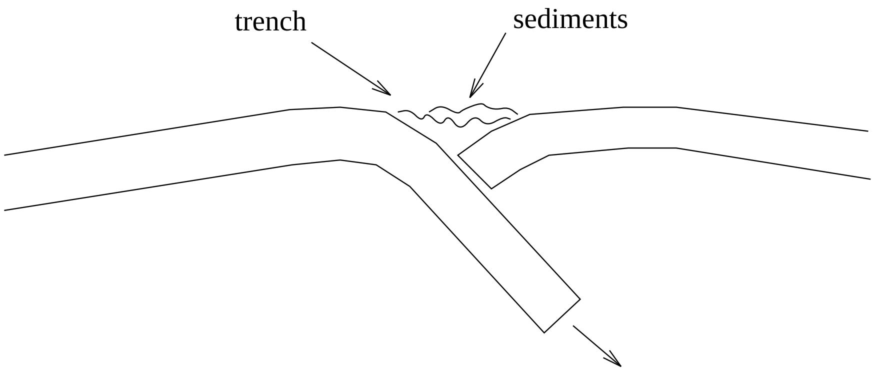

Figure 6.37:

The compressive forces between the plates cause them to bend, thereby creating the trench. Since the trench, which is typically a few km deep, is largely supported by stresses within the slabs, mass compensation is not required.

The right-hand plate scrapes sediments off the left-hand plate as it descends. Depending on the trench system, the volume of these sediments can be negligible or, more typically, can fill much of the trench. Sediments have low density, and so the sediment fill causes a negative gravity anomaly that is in addition to the negative anomaly caused by the depression in the sea floor. Moreover, the sediment-filled trench is usually centered further toward the subducting side of the cross-section then would be inferred from simply looking at the sea floor topography, and this leads to the observed offset between the gravity low and the axis of maximum sea floor depression. The amplitude of the free-air anomaly over a trench tells you how deep the sediment wedge is (typically, tens of km).

The positive anomaly on the right hand side is usually pretty large: 50–75 mgal. It is presumably due to the descending slab. That slab is cool and so is denser than the asthenosphere it is passing through (that density difference is pulling the slab downwards). The slab is not compensated by lower elevations at the surface, because the region is under compression, keeping the topography high. So, gravity above the slab is anomalously large.

The positive anomaly on the left is somewhat smaller: maybe 20–50 mgal. It is caused by buckling of the left hand plate, in response to the resistant pressure force at the trench from the other plate. The buckling increases the topography there, and it is uncompensated: it is *not* supported from below. So, there is excess mass in the vertical column beneath that point. So, gravity is large. This buckling has been modeled with thin plate flexure theory, to learn about the thickness of the lithosphere *and* the resistant pressure force at the trench.

6.8 The Long-Wavelength Geoid

Most of the discussion so far in this chapter has involved the interpretation of gravity at wavelengths $\ll$ radius of earth. What about at longer wavelengths (many thousands to tens of thousands of km)? Those are best studied by using the geoid, rather than gravity anomalies, because longer wavelengths are more prominent in the geoid.

At these very long wavelengths, the relation between topography and the geoid can be quite complicated. Some of the global-scale topography is undoubtedly supported by near-surface compensation. For example, the mean ocean-continent elevation difference is largely compensated by a corresponding difference in ocean-continent crustal thicknesses. If long wavelength topography is compensated at depths of a few tens of km, the gravitational signal from the compensation should almost exactly cancel the signal from the root: a signal with a wavelength of many thousands of km will decrease only slightly over a vertical distance of tens of km. In that case, there would be virtually no correlation between topography and the geoid.

But, some of the longest-wavelength topography is supported by density anomalies at greater depths in the mantle — such as those caused by the thermal anomalies associated with mantle convection. The geoid is sensitive to the earth's deep structure at global-scale wavelengths, because a long-wavelength gravitational signal decreases only slowly with distance. But the relation between a deep internal density anomaly and the topography it causes is not straightforward.

Beginning in the mid-1980's, when three-dimensional, global seismic velocity results started to appear, people began to successfully model and interpret the observed global-scale geoid. The idea goes like this:

There are density anomalies throughout the mantle caused by thermal anomalies (e.g., colder material is denser). These density anomalies drive mantle convection, through the buoyancy forces that act on them. The density anomalies affect the geoid to an extent that depends on how deep they are.

At the same time, the density anomalies tend to be compensated. For example, a positive density anomaly will push down the material underneath it, and pull down the material above it. So it will cause depressions at the outer surface, at the core/mantle boundary, and at every other internal boundary. These depressions act like negative mass anomalies, and their effects on the geoid tend to offset the direct effects of the density anomaly. The situation is thus similar to that for the near-surface compensation that we have been considering throughout this chapter. The degree to which the two effects cancel in the geoid depends on which boundaries are perturbed and by how much. The answer to that problem depends on where the density anomaly is, and on how effectively stresses can be supported by the material that lies between the anomaly and the boundary. That, in turn, depends on the earth's viscosity as a function of depth.

What people do, is to infer the internal density anomalies from the three-dimensional seismic velocity maps. They convert seismic velocity anomalies to density anomalies by, usually, multiplying the velocity anomalies by a constant. The constant is sometimes chosen to be depth-dependent. They then estimate the vertical displacement of the earth's surface and of all internal boundaries, by solving differential equations that describe the earth's viscous response to internal loads. The boundary displacement depends on the viscosity profile. The viscosity profile is adjusted until the geoid prediction (coming from the combined effects of density anomalies and boundary displacements) best matches the observed geoid.

People have obtained good agreement with the observed long-wavelength geoid. They have been able to explain about 70 to 80% or more of the variance of the observed geoid, for spherical harmonics of degree and order of, say, 8 or less (corresponding to half-wavelengths of 2500 km and longer). And most studies find that in order to obtain these good fits, a large increase in viscosity is needed between the upper and lower mantles: lower mantle viscosity approximately 1 to 2 orders of magnitude larger than upper mantle viscosity, with upper/lower mantle dividing line at approximately 660 km depth.

Chapter 7

Postglacial Rebound

During the last ice age, enormous volumes of ice accumulated over Canada and Scandinavia, with thicknesses as large as 3–4 km. These loads depressed the underlying earth. The ice began to melt about 22,000 years ago, and had disappeared about 13,000 years later (i.e. 9000 years ago).

As the ice melted, the earth initially rebounded elastically, and then continued to uplift due to the viscous relaxation of the shear stresses inside the earth. (Before melting, the ice had presumably been in place long enough that the shear stresses below the elastic lithosphere had almost entirely disappeared due to viscous flow. So, the removal of the load *caused* shear stresses.) In fact, the land is still uplifting. For example, Hudson Bay — a depression left by the ice — still exists. By constructing models of the rebound, and comparing the model predictions against observations of such things as the uplift rates, people have been able to infer the viscosity of the earth, and to place constraints on models of the ice load (i.e. on the volume and horizontal extent of the ice, and on the time history of the melting). This has resulted in what are probably the best estimates for the earth's viscosity.

7.1 Theory

Suppose the earth has a Maxwell solid rheology. Suppose we add or remove a load from the earth's surface. How do we compute the uplift of the surface as a function of time?

To address this question, we will consider a simple earth model. Our earth is a homogeneous, incompressible, half space. Furthermore, we assume the surface load depends only on one horizontal coordinate, so that we have a two-dimensional problem (with horizontal and vertical coordinates x and z, respectively). We are assuming that the earth is homogeneous, so that the entire earth is a Maxwell solid. Consequently, there is no elastic lithosphere.

This is the simple model people used when they first considered the problem about 30 years ago. It provides a pretty good qualitative picture of visco-elastic rebound. And, it even gives an analytical relation between the relaxation time and the viscosity, that is reasonably consistent with the numerical results obtained from more recent, complicated models. Later, I'll describe how the newer models have extended the simple model used here.

We will work in the Fourier-transformed frequency domain, where the independent variable is the frequency, ω, rather than the time, t. The advantage of doing this is that then we can use the correspondence principle. That is, we can solve the problem for an elastic earth, and then let $\mu \rightarrow \overline{\mu}(\omega)$ and $\lambda \rightarrow \overline{\lambda}(\omega)$ and transform back to the time domain.

7.2 Elastic Problem

Suppose a pressure $\mathbf{P}_0(x)$ is applied to the earth's surface ($z = 0$). See Figure 7.1. Assume the earth is a homogeneous, incompressible, *elastic* half space. We try to find the surface displacement, $\overline{s}(x, z = 0)$. Later, we will use the correspondence principle to obtain the visco-elastic solution.

Note that we are assuming that $\mathbf{P}_0(x)$ is independent of y. By symmetry, this implies that $s_y(x, z) = 0$, and that no variable can depend on y (so that $\partial_y(\text{anything}) = 0$).

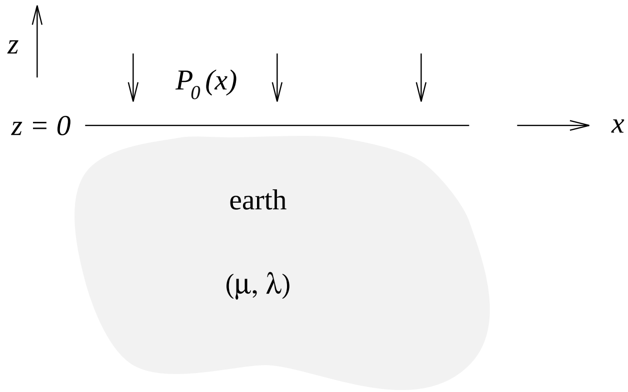

Figure 7.1:

The stress-strain relation is

$$\tau_{ij} = -\mathbf{P}\delta_{ij} + \mu\left[\partial_i s_j + \partial_j s_i\right] \tag{7.2.1}$$

where $\mathbf{P}$ = pressure. Normally, $\mathbf{P} = -\lambda\overline{\nabla}\cdot\overline{s}$. But for an incompressible solid, $\lambda = \infty$ and

$$\nabla\cdot\overline{s} = 0, \tag{7.2.2}$$

and so $\mathbf{P}$ cannot be determined from $\overline{s}$. So, we have an extra equation (Equation 7.2.2) and an extra unknown ($\mathbf{P}$).

The momentum equation ($\overline{F} = m\overline{a}$) in the time domain is $\rho\partial_t^2\overline{s} = \overline{\nabla}\cdot\overline{\tau}$. We are in the frequency domain, where $\partial_t^2 \to -\omega^2$. We are only interested in the long period response (long compared to seismic wave periods of seconds to minutes). In that case ω is small, so that there is negligible acceleration, and the $\rho\partial_t^2\overline{s}$ term can be ignored. So:

$$\overline{\nabla}\cdot\overset{\leftrightarrow}{\tau}\approx 0. \tag{7.2.3}$$

By using Equation 7.2.1 in Equation 7.2.3 to eliminate the tensor $\overset{\leftrightarrow}{\tau}$ from the equations, and noting that μ = constant from the homogeneous earth assumption, we obtain:

$$-\overline{\nabla}\mathbf{P} + \mu\left[\nabla^2\overline{s} + \overline{\nabla}\,\overline{\nabla}\cdot\overline{s}\right] = 0.$$

Or, using Equation 7.2.2 to eliminate $\overline{\nabla}\cdot\overline{s}$

$$\overline{\nabla}\mathbf{P} = \mu\nabla^2\overline{s}. \tag{7.2.4}$$

Equations 7.2.4 and 7.2.2 are our differential equations, and $\mathbf{P}$ and $\overline{s}$ are our unknown variables.

We also have boundary conditions at $z = 0$: $\tau_{zz}(z = 0) = -\mathbf{P}_0$. Or, using Equation 7.2.1:

$$\mathbf{P}(x, z = 0) - 2\mu\partial_z s_z(x, z = 0) = \mathbf{P}_0(x). \tag{7.2.5}$$

Also, $\tau_{13}(x, z = 0) = 0$, since there is no shear traction at the upper surface. This implies, using Equation 7.2.1 for τ_{13} that:

$$\partial_x s_z(z = 0) + \partial_z s_x(z = 0) = 0. \tag{7.2.6}$$

Also, all variables must remain finite as $z \to -\infty$.

The differential equations are easier to solve if we take their derivatives. First, we take $\overline{\nabla}\cdot$ (Equation 7.2.4) and use Equation 7.2.2, to obtain:

$$\nabla^2\mathbf{P} = 0. \tag{7.2.7}$$

Second, we take $\nabla\times$ (Equation 7.2.4), to get:

$$\nabla^2\left(\overline{\nabla}\times\overline{s}\right) = 0. \tag{7.2.8}$$

Since $\overline{\nabla}\cdot\overline{s} = 0$, and since $\partial_y(\text{anything}) = 0$, then there is some scalar, $f = f(x, z)$, for which

$$\begin{aligned} s_x &= \partial_z f & (7.2.9)\\ s_z &= -\partial_x f. & (7.2.10) \end{aligned}$$

(Equations 7.2.9 and 7.2.10 follow because $\nabla\cdot s = 0 \Rightarrow s = \overline{\nabla}\times\overline{\phi}$ for some vector $\overline{\phi}$. Then, we use the fact that $\partial_y\phi_i = 0$ for all i, and define $f = -\phi_y$.)

Equation 7.2.8 is a vector equation. But $\overline{\nabla}\times\overline{s}$ has $\hat{e}_x$ and $\hat{e}_z$ components that vanish, because $s_y = 0$ and $\partial_y = 0$. So, the only non-zero component of Equation 7.2.8 is the $\hat{e}_y$-component, which reduces to:

$$\left(\partial_z^4 + 2\partial_x^2\partial_z^2 + \partial_x^4\right) f = 0. \tag{7.2.11}$$

Equations 7.2.11 and 7.2.7 are now our differential equations, and $\mathbf{P}$ and f are our unknowns. The boundary conditions, Equations 7.2.5 and 7.2.6, become:

$$\mathbf{P}(x, z=0) + 2\mu\partial^2_{xz} f(x, z=0) = \mathbf{P}_0 \tag{7.2.12}$$

$$\partial^2_z f(x, z=0) = \partial^2_x f(x, z=0). \tag{7.2.13}$$

We've lost some of the original information in deriving these equations for $\mathbf{P}$ and f, because we obtained them by taking derivatives of the original equations. We will find, as a consequence, that the general solution for $\overline{s}$ will include an arbitrary multiplicative constant. To find the constant we will have to substitute our solution into one of the original equations: Equation 7.2.4.

Let's see how this works. First, we solve the partial differential equations, Equations 7.2.11 and 7.2.7, making sure we satisfy the boundary conditions Equations 7.2.12 and 7.2.13. We do this in the wave number domain. That is, we assume

$$\mathbf{P}_0(x) = \overline{\mathbf{P}}_0 e^{ikx} \tag{7.2.14}$$

where $\overline{\mathbf{P}}_0$ is a constant. The rationale for this is that any $\mathbf{P}_0(x)$ can be expanded as a Fourier integral of e^{ikx} terms. So if we can solve our equations for a single e^{ikx} term, we can add together (i.e. integrate) the results to obtain the solution for an arbitrary $\mathbf{P}_0(x)$. To be definite, and to avoid having to use absolute values in what follows, we will assume that $k \geq 0$.

We look for solutions of the form

$$\begin{aligned} \mathbf{P}(x, z) &= \overline{\mathbf{P}}(z) e^{ikx} \\ f(x, z) &= \overline{f}(z) e^{ikx} \end{aligned}$$

Then $\partial_x \to ik$, and Equations 7.2.11 and 7.2.7 become:

$$\begin{aligned} \partial^2_z \overline{\mathbf{P}} &= k^2 \overline{\mathbf{P}} \\ (\partial^4_z - 2k^2\partial^2_z + k^4)\overline{f} &= 0. \end{aligned}$$

The boundary conditions are

$$\begin{aligned} \overline{\mathbf{P}}(z=0) + 2\mu i k \partial_z f(z=0) &= \overline{\mathbf{P}}_0 \\ \partial_z^2 \overline{f}(z=0) &= -k^2 \overline{f}(z=0). \end{aligned}$$

The most general solution is

$$\begin{aligned} \overline{\mathbf{P}}(z) &= \overline{\mathbf{P}}_0 e^{kz} \\ \overline{f}(z) &= A(1-kz)e^{kz} \end{aligned}$$

where we have not included the e^{-kz} solutions to the differential equations, since they $\to \infty$ as $z \to -\infty$; and where A is an arbitrary constant, at this stage. So

$$\begin{aligned} \mathbf{P}(x,z) &= \overline{\mathbf{P}}_0 e^{kz} e^{ikx} \\ s_x(x,z) &= -Ak^2 z e^{kz} e^{ikx} \\ s_z(x,z) &= -Aik(1-kz)e^{kz}e^{ikx}. \end{aligned}$$

To find A, we use these results in $\overline{\nabla}\mathbf{P} = \mu \nabla^2 \overline{s}$ (Equation 7.2.4), and obtain:

$$A = \frac{-i}{2\mu k^2}\overline{\mathbf{P}}_0.$$

Thus, the vertical displacement at the surface ($z = 0$) — which is the quantity that can be observed — is:

$$s_z(x, z=0) = \frac{-\overline{\mathbf{P}}_0}{2\mu k} e^{ikx} = \frac{-\mathbf{P}_0(x)}{2\mu k}. \tag{7.2.15}$$

If we were to take this result (Equation 7.2.15) directly over to the Maxwell solid case by replacing μ by $\overline{\mu}(\omega)$, we'd find a distressing thing: If you put a mass on the surface and wait, the mass sinks out of sight as $t \to \infty$. You can see this from Equation 7.2.15, by noting that for a Maxwell solid at long periods, $\overline{\mu} \to 0$ and so $s_z(z=0) \to \infty$.

The problem is, we have ignored the effects of gravity acting on the deformed earth. Instead of the equation $\nabla \cdot \tau = 0$, we should have used $\nabla \cdot \tau + \rho \overline{g} = 0$. Luckily, we can include these effects as sort of an afterthought, as follows.

Suppose we load the earth with a surface mass per unit area $= M(x)$. The earth's surface rises by the amount $s_z(x, z=0)$ ($s_z(x, z=0)$ will be negative for a positive surface mass load — note the negative sign on the right-hand-side of Equation 7.2.15). The total pressure at $z = 0$ is the pressure due to $M(x)$, *plus* the pressure due to gravity pulling down on the extra mass above $z = 0$. This extra mass per area is $\rho s_z(x, z = 0)$, where ρ = density of the earth. So, Equation 7.2.15 for $s_z(x, z = 0)$ is ok, as long as we use

$$\mathbf{P}_0(x) = gM(x) + \rho g s_z(x, z = 0).$$

Using this result for $\mathbf{P}_0$ in Equation 7.2.15 gives:

$$s_z(x, z = 0) = \frac{-gM(x)}{2\mu k}\left[\frac{1}{1 + \dfrac{\rho g}{2\mu k}}\right] = \frac{-gM(x)}{2\mu k + \rho g}. \qquad (7.2.16)$$

Because the additional pressure from the deformed surface has the opposite sign of the pressure from the applied load, the effects tend to offset one another. What happens is that the earth depresses under an applied load until perfect compensation is obtained. At that point, the pressures from the deformed surface and from the applied load exactly cancel, and the displacement stops.

We now transform our earth into a Maxwell solid. We are in the frequency domain, so we can do this by modifying Equation 7.2.16 so that:

$$\mu \to \overline{\mu}(\omega) = \mu\left(\frac{i\omega}{i\omega + \dfrac{1}{\tau_0}}\right) \qquad (7.2.17)$$

where $\tau_0 = \eta/\mu$ and η is the viscosity. By using Equation 7.2.17 in Equation 7.2.16, we obtain:

$$\left[2\mu k \frac{i\omega}{i\omega + \dfrac{1}{\tau_0}} + \rho g\right] s_z(x, z = 0) = -gM(x).$$

Or:

$$\left[(2\mu k + \rho g)i\omega + \frac{\rho g}{\tau_0}\right] s_z(x, z = 0) = -g\left[i\omega + \frac{1}{\tau_0}\right] M(x) \qquad (7.2.18)$$

Equation 7.2.18 is valid in the frequency (ω) and wave number (k) domains. We can transform back to the time domain by replacing $i\omega$ with ∂_t:

$$\left[(2\mu k + g\rho)\partial_t + \frac{\rho g}{\tau_0}\right] s_z(x, z = 0, t) = -g\left[\partial_t + \frac{1}{\tau_0}\right] M(x, t). \tag{7.2.19}$$

Although we could also transform back to the x-domain, we choose, instead, to leave Equation 7.2.19 as it is.

Let's solve this differential equation, Equation 7.2.19, in the time domain for the case where the load $M_0 e^{ikx}$ is removed instantaneously at $t = 0$:

$$M(x, t) = \begin{cases} M_0 e^{ikx} & t < 0 \\ 0 & t \geq 0. \end{cases}$$

What is $s_z(x, z = 0, t)$?

First, for $t < 0$ we'll assume the load has been in place long enough that all motion has stopped. Then: $\partial_t s_z = 0$. So, from Equation 7.2.19

$$s_z(x, z = 0, t) = \left[\frac{-\dfrac{g}{\tau_0}}{\dfrac{\rho g}{\tau_0}}\right] M_0 e^{ikx} = -\frac{M_0 e^{ikx}}{\rho} \qquad t < 0. \tag{7.2.20}$$

That's the result we'd get for a mass floating on a fluid.

Next, at $t = 0$ all $\partial_t \to \infty$ ($\partial_t M$ = Dirac delta function). So, since $\rho g/\tau_0$ and $1/\tau_0$ are finite, the equation for s_z at $t = 0$ reads:

$$(2\mu k + \rho g)\partial_t s_z = -g\partial_t M.$$

Writing $\partial_t s_z = \Delta s_z/\Delta t$, $\partial_t M = \Delta M/\Delta t$, and cancelling Δt's, gives:

$$\Delta s_z = \frac{-g\Delta M}{2\mu k + \rho g}.$$

Here, ΔM is the change in $M(t)$ in going from $t < 0$ to $t > 0$, and similarly for Δs_z. Since $\Delta M = -M_0 e^{ikx}$, then:

$$\Delta s_z = \frac{g M_0 e^{ikx}}{2\mu k + \rho g} \tag{7.2.21}$$

By adding Equation 7.2.21 to Equation 7.2.20, we obtain s_z immediately after the removal of the load:

$$s_z = \left[\frac{g}{2k\mu + \rho g} - \frac{1}{\rho}\right] M_0 e^{ikx} \qquad t = 0. \tag{7.2.22}$$

Note that there is, as yet, no dependence on τ_0. So, the initial response doesn't depend on the viscosity.

Now, let's consider $t > 0$. For $t > 0$, $M = 0$. So, the equation for s_z is:

$$\partial_t s_z(x,0,t) = \left[\frac{-g\rho}{\tau_0(2\mu k + g\rho)}\right] s_z(x,0,t).$$

And, the initial condition is: $s_z(x,0,0) = s_z$ given by Equation 7.2.22. The solution is:

$$s_z(x,0,t) = \left[\frac{g}{2k\mu + \rho g} - \frac{1}{\rho}\right] M_0 e^{ikx} \exp\left[\frac{-g\rho t}{\tau_0(2k\mu + g\rho)}\right]. \tag{7.2.23}$$

7.3 Numbers

For the earth, $\rho \approx 5$ gm/cm^3, $g \approx 10^3$ cm/s^2, and $\mu \approx 10^{12}$ dyne/cm^2. So: $\rho g/\mu \approx 5 \times 10^{-9}$ cm^{-1}. For wavelengths much smaller than the radius of the earth — which is a reasonable approximation for the ice load — $k \gg 2\pi/6.4 \times 10^8$ cm$^{-1} \approx 10^{-8}$ cm^{-1}. So: $\mu k \gg \rho g$. So, for our case, Equation 7.2.23 reduces to

$$\begin{aligned} s_z(x,0,t) &\approx \left[\frac{g}{2k\mu} - \frac{1}{\rho}\right] M_0 e^{ikx} \exp\left[\frac{-g\rho}{2k\mu}\,\frac{t}{\tau_0}\right] \\ &\approx -\left[\frac{1}{\rho}\right] M_0 e^{ikx} \exp\left[\frac{-g\rho}{2k\mu}\,\frac{t}{\tau_0}\right]. \end{aligned}$$

7.4 Interpretation

The original depression is $-(1/\rho)M_0 e^{ikx}$. The elastic rebound at $t = 0$ is

$$\frac{gM_0 e^{ikx}}{2k\mu + \rho g} \approx \frac{gM_0 e^{ikx}}{2k\mu} \ll \frac{M_0 e^{ikx}}{\rho}.$$

So, the initial elastic uplift is much smaller than the original depression.

Most of the uplift occurs through the later visco-elastic relaxation. As $t \to \infty$, $s_z \to 0$, and so the surface is flat again. The decay time for the uplift is $\tau_0(2k\mu/g\rho)$ which is $\gg \tau_0$. Thus, the decay time $\gg$ relaxation time for the solid. Note, also, that the decay time depends on k. Consequently, for a more realistic ice load distribution, where there are contributions from lots of k's, the depression changes shape as the ground uplifts, with longer wavelengths uplifting more quickly. The result that the long wavelengths decay more quickly can be understood as follows. There is a balance between the buoyancy force acting on the deformed outer surface, and the viscous shearing stresses within the earth. The buoyancy force depends on the amplitude of the displacement field. The shear stress depends on the strain rate; which, in turn, depends not only on the displacement amplitude, but also on the spatial wavelength and the rate of motion: the strain rate varies as (displacement amplitude)/(wavelength $\times$ decay time) . Since the buoyancy force does not depend on either the wavelength or the decay time, then in order for the shear stress to balance the buoyancy force the decay time must vary inversely with the wavelength. Thus, large wavelength terms have larger rates of decay.

When this half-space model is used to interpret observed decay times, it gives mantle viscosities that are of the same order as those inferred from more realistic models. Nonetheless, it obviously uses a pretty simple representation of the earth. The most recent models extend this representation to include:

1. the earth's spherical shape.

2. compressibility and self-gravitation.

3. radially-dependent values for the earth's material properties. Seismic results for ρ, λ, and μ are used. And a layered profile is adopted for the viscosity, usually consisting of an elastic lithosphere and core, and a 2-layer viscosity model: one layer above the seismic 660 km discontinuity (the upper mantle), and the other layer below that discontinuity (the lower mantle), with both layers assumed to have uniform viscosities. Data can then be inverted to find the upper and lower mantle viscosities, and the lithospheric thickness. Sometimes people include more than two viscous layers. In fact, there have even been several recent studies of the effects of lateral variations in viscosity.

4. people sometimes also consider the effects of the melting ice on global sea level, and then use the change in sea level as an additional load on the earth.

5. a reasonably realistic model of the surface ice distribution.

The best ice sheet models are reasonably complicated. Their ice boundaries are largely determined from end moraine positions. The total ice volume comes from geological data for changes in sea level. The idea is that if you are far away from the ice, the uplift is not affected much by the viscous rebound, but mostly only by changes in the ocean volume caused by the melting ice. Thus, by geologically determining the sea level change during the melting, you can infer the total ice mass as a function of time. To estimate the more detailed temporal and spatial distribution of the ice, people use parameterized time- and space-dependent functions, and then fit the parameters in those functions to the postglacial rebound observations, *in addition* to fitting the viscosity.

7.5 The Data

People have fit their model results to various types of data, to infer mantle viscosity, etc. Many of these data types are ambiguous, in that they may at least partially reflect the effects of other geophysical phenomena.

1. geological mapping of ancient raised beaches, both from areas under the ice sheets and from areas around the edges of the ice sheets. This is probably the most unambiguous type of data, and is particularly useful for constraining lithospheric thickness and the viscosity profile within, say, the upper 1000 km of the earth.

2. the geoid and long wavelength free-air anomalies in the regions surrounding the ice loads. For example, there are large negative g_{FA} and geoid anomalies over Hudson Bay, with the g_{FA} anomaly reaching about -50 mgal. These anomalies have sometimes been interpreted as the effects of postglacial rebound: the earth has still not fully adjusted to the removal of the ice loads, and so there is a mass deficiency under those regions. When interpreted in this manner, the observed anomalies have been found to provide tight constraints on the deep mantle viscosity. However, it has been argued that a significant fraction — maybe most — of the

observed gravity anomalies in these regions may be due to causes not related to postglacial rebound (for example, to the effects of deep continental roots; or maybe to the effects of mantle convection). In fact, people have recently become nervous enough about this possibility, that they have pretty much stopped using these types of observations in their postglacial rebound inversions.

3. LAGEOS orbit solutions have allowed for the determination of the secular (i.e. linear in time) change in J_2, commonly denoted as $\dot{J}_2$. The observed $\dot{J}_2$ is approximately -2×10^{-11} yr^{-1} to -3×10^{-11} yr^{-1} (for comparison, $J_2 \approx 10^{-3}$). This result has been interpreted as due to the on-going re-distribution of mass associated with postglacial rebound, and has provided a strong constraint on lower mantle viscosity. However, there are other possible causes of $\dot{J}_2$. For example, present-day changes in the ice volumes of Greenland or Antarctica are not at all well known, and their effects on $\dot{J}_2$ could well be as large or larger than the effects of postglacial rebound. So, people who interpret the observed $\dot{J}_2$ as due to postglacial rebound, do so at their peril.

4. secular changes in the earth's rotation rate (as inferred from ancient eclipse data) and in the pole position (as observed with telescopes over the last 100 years or so). Both of these motions will be discussed in more detail in Section 9 below. There are probably important contributions to this secular variability due to changes in the earth's inertia tensor caused by postglacial rebound. When interpreted in this manner, the observations have provided tight constraints on lower mantle viscosity. Except that, as with $\dot{J}_2$, there could also be large contributions from other mechanisms that involve mass re-distribution, including changes in polar ice volumes.

5. on-going vertical and, to a lesser extent, horizontal displacements of the earth's surface; and, similarly, on-going changes in surface gravity. Rates of uplift can be as large as 1 cm/yr close to Hudson Bay and in Scandinavia, causing gravity at the earth's surface to decrease by about 3 μgals/yr. People are conducting campaigns to

observe these changes using GPS and absolute gravimeters. They should eventually provide reasonably unambiguous constraints on viscosity in the upper 1000 km, or so, of the mantle. Understanding the on-going vertical motion along coastlines is also important in trying to interpret observed changes in sea level as recorded by tide gauges. Tide gauge data have been used to infer the global rise in sea level over the past century (probably 1.5–2.0 mm/yr). But, those studies require that first the effects of postglacial rebound be removed from the data, so that the remainder can be interpreted in terms of sea level. For example, if a tide gauge indicates a rise in sea level, it may be because sea level *did* rise; or it may be because the land under the tide gauge subsided. Vertical motion along the North American and Northern European coastlines (where a disproportionally large percentage of the world's tide gauges are located) caused by postglacial rebound is likely to be as large as several mm/yr in places. Thus, errors in the postglacial rebound estimates could well corrupt the global sea level rise results. Conversely, people have used tide gauge data along the eastern coast of North America to try to identify those rebound errors and, in particular, to learn about mantle viscosity.

7.6 Conclusions

There is still some disagreement as to what the postglacial rebound studies imply for the earth's viscosity profile. Different studies have reached somewhat different conclusions. Estimates of upper mantle viscosity (above 670 km depth) seem to have converged to values of between about $\eta \approx 0.5 \times 10^{21}$ Pascal-second and 1.0×10^{21} Pascal-second (1 Pascal-second = 1 Newton–sec/m^2). (A 10^{21} Pascal-second viscosity corresponds to a Maxwell relaxation time of a few centuries, and to decay times that are much longer than that.)

As for the jump in viscosity in going from the upper to the lower mantle (from above 670 km depth, to below that depth), estimates range from maybe only a factor of 4–5 jump in viscosity, to a lower mantle η that is one to two orders of magnitude greater than the upper mantle η.

The disagreements probably reflect the fact that postglacial rebound is presently an underdetermined problem. There are not enough unambiguous observations of the rebound to reolve all the conceivable unknown variables associated with the earth's visco-elasticity and with the ice sheet. The possibility of significant lateral variability in viscosity and lithospheric thickness complicates this situation even further.

Chapter 8

Earth Tides

The gravitational attraction of the sun and moon causes tides both in the ocean and in the solid earth. The tides in the ocean are, of course, familiar to anyone who has visited the sea shore. The solid earth tides are not as well known to most people. This is not because the earth tides are small. In fact, peak-to-peak vertical tidal displacements are typically many tens of centimeters. Instead, the difficulty is that when you try to observe the solid earth tides you are standing on the earth and moving right along with it. As a result the motion is not readily apparent and cannot be detected without sensitive instruments. Analogously, you are not aware of ocean tides if you are in the middle of the ocean — you must be able to observe both the ocean and the land. In fact, the ocean tides you observe at the sea shore are really the *difference* between the tidal displacements of the ocean and the solid earth.

8.1 A Qualitative Description of the Tidal Force

The gravitational force from the sun and moon causes orbital motion of the earth. This motion is by far the largest effect of that force on the earth. But the gravitational force also acts to deform the earth and oceans, and those deformations are the tides. In fact, it is usual to separate the luni-solar gravitational force into a part that is constant over the earth and which causes the orbital motion, and a small remainder which causes the tides. That remainder is called the tidal force.

For example, the total force from, say, the moon is represented by the arrows in Figure 8.1 (the following description works equally well for the force from the sun). Ω

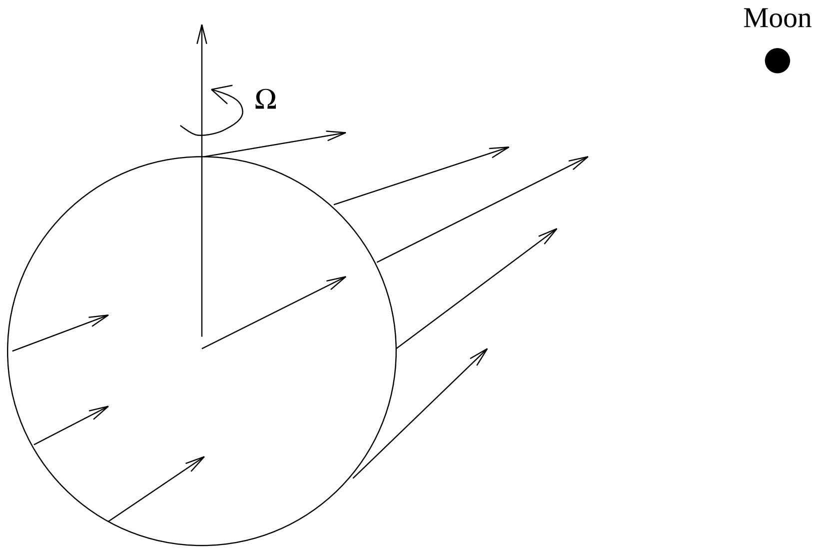

Figure 8.1:

represents the earth's rotation. Note that the total force at every point is directed toward the moon, and that the arrows closest to the moon are the largest (the difference in arrow length is greatly exaggerated in the figure).

The orbital force on the earth is the average of all the arrows. To a high degree of approximation, the average force equals the force acting at the center of the earth. If we subtract the arrow at the center of the earth from all the other arrows, we are left with the tidal force, which is the force that tends to deform the earth (see Figure 8.2). The tidal force, by its definition, causes no net force on the earth and so does not affect the earth's orbital motion.

Note from Figure 8.2, that the force is radially outward on the sides toward and away from the moon, and is radially inward on the other sides. Looking down from above the North Pole we get the force pattern shown in Figure 8.3. This pattern remains fixed with respect to the moon, and the earth rotates relative to it. This causes the tidal force at a fixed point on the earth to be variable with time. For example, suppose the moon were

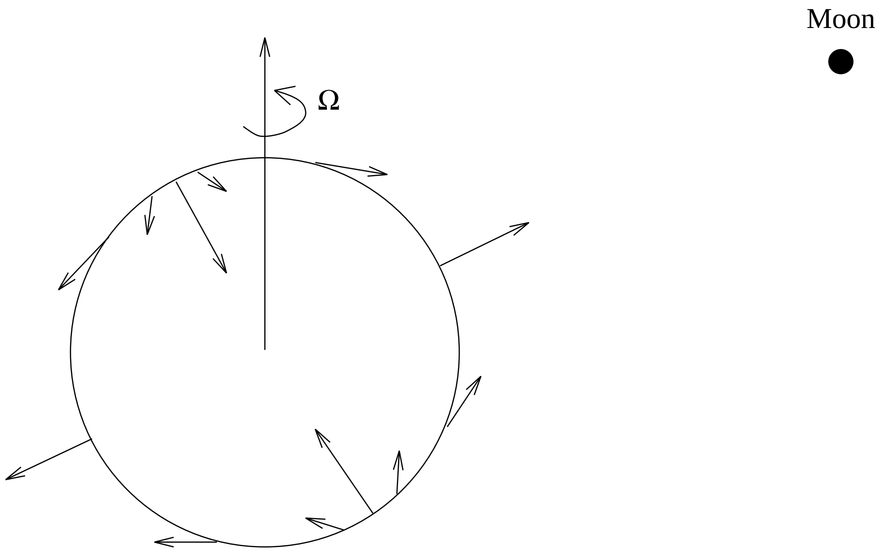

Figure 8.2:

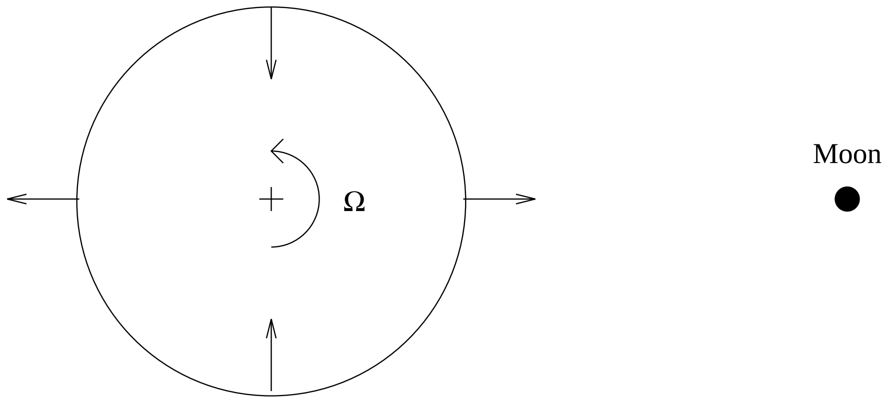

Figure 8.3:

on the equator, as in Figure 8.4. Every point in the earth would then travel through two outward bulges and two inward depressions during one day (see Figure 8.5). So, the frequency of the tidal force as seen from the earth would be $\Omega/2 = 2$ cycles/day, corresponding to a 12 hour (semi-diurnal) period.

As another example, suppose the moon were above the North Pole, as in Figure 8.6. Then any point on the earth remains always in the same bulge or depression as the earth rotates. So, the period in this case would be ∞: there is no time dependence.

As a final example, suppose the moon were inclined at 45° to the equator as in

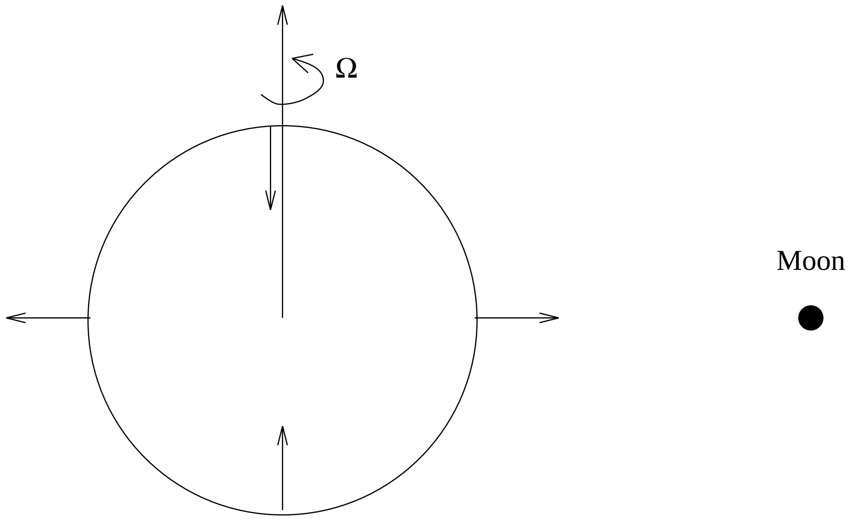

Figure 8.4:

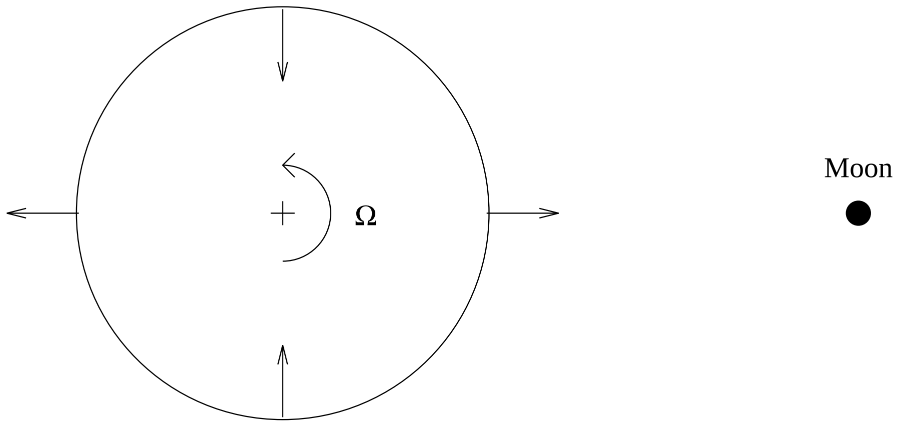

Figure 8.5:

Figure 8.7. Consider point A, initially in an outward bulge. After 12 hours that point has moved to an inward depression (point A'), and after 24 hours it is back where it started. So the period that this point (or any point in the earth) sees in this case is 1 day, and the frequency $= \Omega = 1$ cycle/day. The tide, in this case, is diurnal. (This description is a little oversimplified. There is some 12 hour and ∞-period power even at $45°$.)

In reality, the moon is not fixed at any one of these three spots. Instead, it orbits the earth. So at any given time, depending on where the moon is, you get a linear

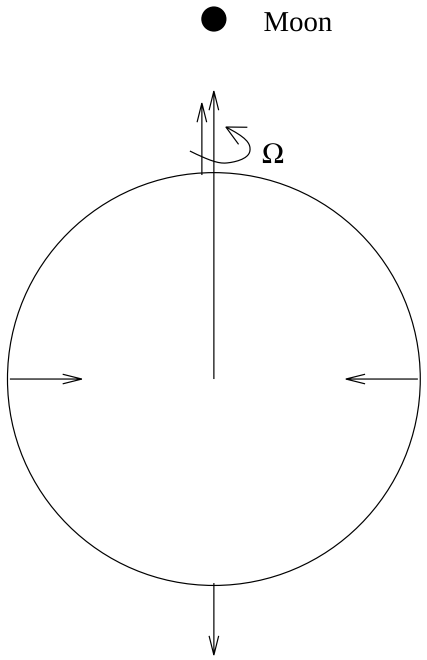

Figure 8.6:

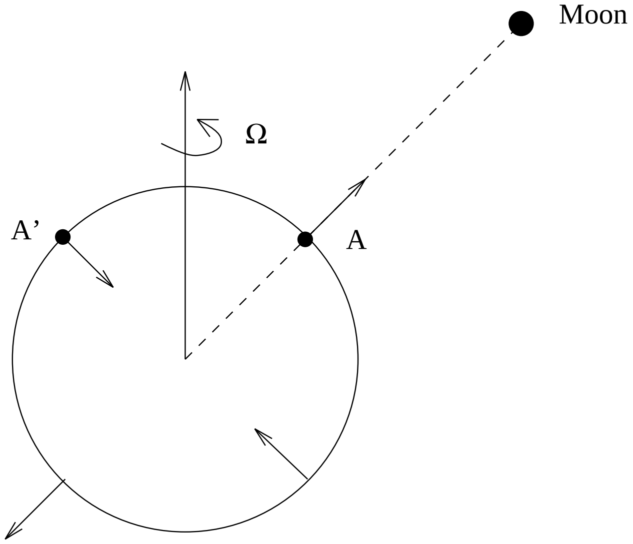

Figure 8.7:

combination of all three frequencies (12 hour, 24 hour, ∞). The moon moves between $\pm 23.5^\circ$ of the equator (with a monthly period), so most of its time is spent at low latitudes where the 12 hour term is most important. So, the semi-diurnal tides are the largest tides at most points. The next biggest are the diurnal tides (the moon gets closer to 45° than to 90°). The ∞-period terms are the smallest.

Furthermore, because the moon moves in its orbit, these three frequencies are actually split into three bands of frequencies centered about 0, 1, and 2 cycles per day. These bands are referred to as the long period, diurnal, and semi-diurnal tidal bands, respectively. The modulating frequencies vary between 1 cycle/18.6 years and 1 cycle/10 days, and are all frequencies of the orbital motions of the moon about the earth and of the earth about the sun. So the long period tides have frequencies between 1 cycle/18.6 years and about 1 cycle/10 days; the diurnal tides between about $(1+\frac{1}{10})$ cycle/day and $(1-\frac{1}{10})$ cycle/day; and the semi-diurnal tides between about $(2+\frac{1}{10})$ cycle/day and $(2-\frac{1}{10})$ cycle/day. The most important (largest) modulating frequencies are 1 cycle/year, 2 cycle/year, 1 cycle/month, 2 cycle/month, and 1 cycle/18.6 years.

8.2 Tidal Potential

The tidal force causes displacements in the earth and ocean. These are the earth and ocean tides. The deformation will, of course, occur at the same frequencies as the tidal force. But, before discussing the deformation, I want to provide a more qualitative description of the tidal force. I will do this by defining what is called the tidal potential.

First, we note that the total gravitational potential at a point $\overline{x}$ in the earth, due to the moon is:

$$V(\overline{x}) = \frac{GM}{|\overline{x} - \overline{R}|}$$

where $\overline{R}$ is the position vector of the moon (assumed to be a point mass), and where M = lunar mass (V = *negative* of the potential energy, as usual). Assume the origin of our coordinate system is located at the center of the earth. Let (θ, ϕ) and (θ', ϕ') be the spherical angles of $\overline{x}$ and $\overline{R}$, respectively. We can expand $1/|\overline{x} - \overline{R}|$ in terms of Y_l^m, as discussed back in Section 3.3.6, to obtain:

$$V(r,\theta,\phi) = \frac{GM4\pi}{R} \sum_{l=0}^{\infty} \left(\frac{1}{2l+1}\right) \sum_{m=-l}^{l} \left(\frac{r}{R}\right)^l Y_l^{m*}(\theta',\phi') Y_l^m(\theta,\phi) \tag{8.2.1}$$

where r and R are the radial coordinates of $\overline{x}$ and $\overline{R}$, and where I have used $R > r$ (so that $r_> = R$, $r_< = r$ in the appropriate Section 3.3.6 equation).

The $l = 0$ term in Equation 8.2.1 is a constant, independent of (r, θ, ϕ), and so has no physical significance (i.e. the gravitational force is the gradient of V, and the gradient of a constant is zero). This term can be ignored.

The $l = 1$ terms are proportional to $rY_1^m(\theta, \phi)$; and this is proportional to z for $m = 0$, and to $(x \pm iy)$ for $m = \pm 1$. ($Y_1^0 \approx \cos\theta$, $Y_1^{\pm 1} \approx \sin\theta e^{\pm i\phi}$.) The force is the gradient of $V(r, \theta, \phi)$. So, for the $l = 1$ terms the corresponding force is a constant in the $\hat{e}_z$ direction for $m = 0$, and a constant in the $\hat{e}_x \pm i\hat{e}_y$ "directions" for $m = \pm 1$. So, the $l = 1$ terms give constant forces throughout the earth. These cause orbital motion but do not deform the earth.

Furthermore, you can show that for a spherically symmetric earth, the $l > 1$ terms in Equation 8.2.1 cause no net force on the earth, and so do not affect the earth's orbital motion. You can infer this by using Equation 8.2.1 to find $\int_{\text{earth}} \rho \overline{\nabla} V d^3\overline{x} =$ (net force on the earth), and showing that only the $l = 1$ terms contribute to the result. I won't do that here.

So, we now define the tidal potential, V_T, so that its gradient is the tidal force = (gravitational force) – (net force on earth). Then V_T will be given by Equation 8.2.1, except that the sum over l starts at $l = 2$ instead of at $l = 0$.

$R =$ mean earth/moon distance $\approx 3.84 \times 10^5$ km. And, $r \leq$ earth's radius $= 6.371 \times 10^3$ km. So, $(r/R) \leq 1/60$. So, every time you increase l by 1 in Equation 8.2.1, you decrease the corresponding contribution to V by 60. So, to an accuracy of 1 part in 60, we need only consider the $l = 2$ terms (the leading terms) in our spherical harmonic expansion of V_T:

$$V_T \approx \frac{GM4\pi}{5R} \left(\frac{r}{R}\right)^2 \sum_{m=-2}^{2} Y_2^{m*}(\theta', \phi') Y_2^m(\theta, \phi). \tag{8.2.2}$$

Occasionally, and depending on the application, people also include the $l = 3$ terms in V_T. We will not.

It is also usual to eliminate the negative m contributions by noting that:

$$\begin{aligned} Y_2^0 &= \text{real} \\ Y_2^{m*}(\theta',\phi')Y_2^m(\theta,\phi) + Y_2^{-m*}(\theta',\phi')Y_2^{-m}(\theta,\phi) &= 2\mathbf{Re}\left[Y_2^{m*}(\theta',\phi')Y_2^m(\theta,\phi)\right] \end{aligned}$$

where '**Re**' = real part, and where we have used $Y_2^{-m} = (-1)^m Y_2^{m*}$. Then Equation 8.2.2 reduces to:

$$V_T \equiv \mathbf{Re}\left[\frac{r^2}{a^2}\sum_{m=0}^{2} c_m Y_2^{m*}(\theta',\phi')Y_2^m(\theta,\phi)\right]$$

where a = earth's mean radius, and

$$c_m = \frac{GMa^2}{5R^3}4\pi \begin{cases} 1 & \text{if} \quad m=0 \\ 2 & \text{if} \quad m=1,2. \end{cases}$$

The quantity c_m is a measure of the strength of the tidal potential at the earth's surface (where $r = a$).

There is a similar description for the solar tides. The only approximation we used here was that $r/R \ll 1$, which is an even better approximation for the sun than for the moon. To get a feeling for whether solar or lunar tides are larger, we note that $M_{\text{sun}}/R^3_{\text{sun}} \approx 0.46(M_{\text{moon}}/R^3_{\text{moon}})$. So, c_m for the moon is roughly twice as large as c_m for the sun. So, lunar tides are about twice as large as solar tides.

The quantities R, θ', and ϕ' all depend on time, due to the motion of the moon about the earth (or of the earth about the sun, for the solar tides). The point $\overline{x}$ is also moving in space due to the earth's rotation. This motion gets mixed together to produce the characteristic semi-diurnal, diurnal, and long period tidal bands. Let's see how this works.

First, we need to be more precise about our coordinate system. We use a system that's fixed to the earth, so that it is rotating with the earth. Then (r,θ,ϕ), which are the coordinates of the point $\overline{x}$, are fixed, time-independent numbers. Let the $\hat{e}_z$ axis of the coordinate system be along the rotation axis. Then R and θ' depend on the lunar orbit, and ϕ' depends on the lunar orbit *and* on the earth's rotation. See Figure 8.8.

The angle ϕ' (the "mean lunar longitude") changes both because the moon is moving

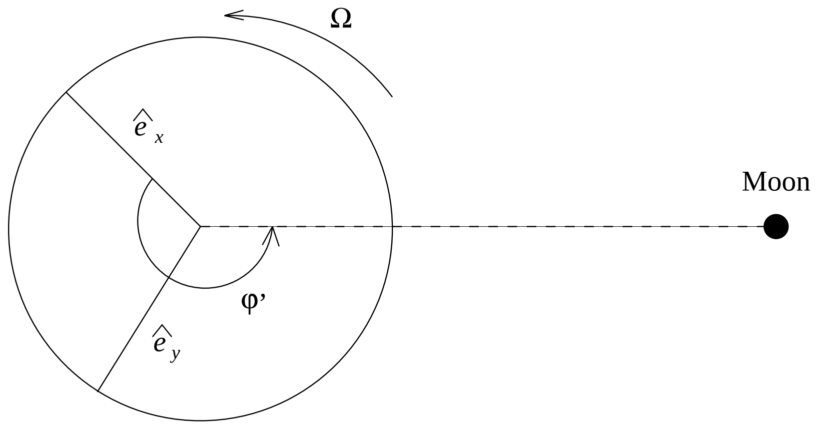

Figure 8.8:

in space, and because the $\hat{e}_x$ axis is rotating in space:

$$\phi' = -\Omega t + \alpha(t) \tag{8.2.3}$$

where $\alpha(t)$ = "moon's right ascension" is the contribution to ϕ' from the lunar motion in space. (To be technically correct, ϕ' in Equation 8.2.3 should also include a constant, which depends on our choice for $t = 0$. That is, the contribution from the earth's rotation should really be $-\Omega(t - t_0)$, where t_0 is determined by the angular position of the earth at $t = 0$. Here, though, we are absorbing the Ωt_0 term into $\alpha(t)$.) Note:

$$Y_2^{m*}(\theta', \phi') = Y_2^{m*}(\theta', \alpha)e^{im\Omega t}.$$

So:

$$V_T = \mathbf{Re}\left[\frac{r^2}{a^2}\sum_{m=0}^{2} c_m Y_2^{m*}(\theta', \alpha)e^{im\Omega t}Y_2^m(\theta, \phi)\right]. \tag{8.2.4}$$

Here, c_m (through R), θ', and α all depend on time because of the lunar orbital motion. The periods of each of these three variables are long compared to 1 day.

So, V_T separates into three frequency bands, each characterized by a different value of m:

$\boldsymbol{m = 2}$**:** The dominant time dependence is $e^{i2\Omega t}$, so these terms are *semi-diurnal* (frequencies $\approx 2\Omega$).

$m = 1$: The dominant time dependence is $e^{i\Omega t}$, so these terms are *diurnal* (frequencies $\approx \Omega$).

$m = 0$: The $e^{im\Omega t}$ term $= 1$, so the time dependence in this case comes entirely from the $c_0 Y_2^0(\theta', \alpha)$ term. These tides are *long period*.

The factor $c_m Y_2^{m*}(\theta', \alpha)$ in Equation 8.2.4 can be expanded as a discrete Fourier series in time:

$$c_m Y_2^{m*}(\theta', \alpha) = \sum_f H_m(f) e^{i(ft+\phi_f)} \tag{8.2.5}$$

where the sum over f is over all frequencies which contribute to the time dependence, the $H_m(f)$ are real constants, and the ϕ_f are phases. The frequencies, f, are linear combinations of the frequencies used to describe R, θ', and α. I won't bother to find those frequencies, or the $H_m(f)$. Those are all determined by the orbital motion of the moon and earth. The frequencies, f, range from 1/18.6 cycles/year to 1/10 cycles/day, and are all $\ll \Omega$. So:

$$V_T = \mathbf{Re}\left[\frac{r^2}{a^2}\sum_{m=0}^{2}\left(\sum_f H_m(f)e^{i(\omega t+\phi_f)}\right)Y_2^m(\theta, \phi)\right] \tag{8.2.6}$$

where $\omega = f + m\Omega$ is close to $m\Omega$. So, there are three disjoint frequency bands, each with frequencies close to $m\Omega$, and each with a spatial dependence proportional to $r^2 Y_2^m(\theta, \phi)$.

8.3 Tidal Response of the Earth

V_T consists of a sum of terms of the form

$$V_T' = \frac{r^2}{a^2} Y_2^m(\theta, \phi) e^{i\omega t} \tag{8.3.1}$$

where $\omega \approx m\Omega$. Each of these terms causes the earth to deform. If we can model the earth's response to each one of these terms individually, then by adding together our results for the different m's and ω's, we can find the response to the total V_T. That follows because the differential equation that describes the deformation is linear.

To find the deformation caused by one of these V_T' terms, we start by labeling each particle inside the earth using the position vector, $\overline{x}$, the particle would occupy if there were no tides. Let $\overline{s}(\overline{x},t)$ denote the time-dependent displacement of the point at $\overline{x}$. The differential equation for $\overline{s}(\overline{x},t)$ comes from Newton's Second Law of Motion ($F = ma$), and has the form: $\rho\partial_t^2\overline{s} =$ (internally-generated forces) $+ \rho\overline{\nabla}V_T'$ where "(internally-generated forces)" includes the effects of internal stresses ($\overline{\nabla}\cdot \overleftrightarrow{\tau}$) and of gravitational self-interaction. The details of the differential equation are complicated and so I won't show them here. But, in the frequency domain (where $\partial_t^2 \to -\omega^2$), the equation has the form:

$$-\rho\omega^2\overline{s} = H \cdot \overline{s} + \rho\overline{\nabla}V_T' \tag{8.3.2}$$

where H is a complicated differential operator. The solution must also satisfy traction-free boundary conditions at the outer surface:

$$\hat{n}\cdot \overleftrightarrow{\tau}\Big|_{\text{surface}} = 0. \tag{8.3.3}$$

Note that $\overleftrightarrow{\tau}$ must be written in terms of $\overline{s}$, to obtain boundary conditions on $\overline{s}$. (There are other boundary conditions that must be satisfied, but that are not given here, that involve the continuity of the gravitational potential and of its radial derivative across the outer surface.) Remember that ω and V_T' are the tidal frequency and tidal potential, respectively, and are known quantities in this problem.

Incidentally, the differential equation (Equation 8.3.2) and the boundary condition (Equation 8.3.3) are also used to find the earth's seismic free oscillations, except that for the free oscillations $V_T' = 0$. In that case the differential equation is homogeneous, and so will have non-zero solutions for $\overline{s}$ only for certain values of ω: values where ω^2 is an eigenvalue of the operator $-H/\rho$. In that case, ω is the free oscillation eigenfrequency and $\overline{s}$ is the eigenfunction. Seismologists are usually more interested in ω than in $\overline{s}$.

For the tidal case $V_T \neq 0$, and the resulting inhomogeneous equation has a unique, non-zero solution for any ω. Actually, that's not quite true. Problems would arise if $\omega =$ free oscillation eigenfrequency, since then the sun and moon would be forcing at a resonance period of the earth. Luckily, the tidal frequencies ω do not coincide with any free oscillation frequencies. The shortest tidal periods we are considering here are close to 12 hours (though there are $l = 3$ and $l = 4$ terms at 8-hour and 6-hour periods, respectively), whereas the longest free oscillation periods are less than one hour.

The easiest way to solve the differential equation is to expand $\overline{s}$ as a sum of spherical harmonics, Y_l^m. Except that because $\overline{s}$ is a *vector*, not a scalar, it must be expanded in terms of *vector* spherical harmonics:

$$\hat{r}Y_l^m(\theta,\phi), \qquad \overline{\nabla}Y_l^m(\theta,\phi), \qquad \hat{r}\times\overline{\nabla}Y_l^m(\theta,\phi). \tag{8.3.4}$$

There are three vector spherical harmonics for each (l,m) because a vector has three components. It turns out that any vector field can be expanded in terms of these functions (just as any scalar field can be expanded in terms of the scalar $Y_l^m(\theta,\phi)$'s), where the expansion coefficients are functions of r. So:

$$\overline{s} = \sum_{l,m}\left[s_{lm}^1(r)\hat{r}Y_l^m(\theta,\phi) + s_{lm}^2(r)\overline{\nabla}Y_l^m(\theta,\phi) + s_{lm}^3(r)\hat{r}\times\overline{\nabla}Y_l^m(\theta,\phi)\right].$$

The idea is to put this expansion for $\overline{s}$ into the differential equation and the boundary conditions, and then to solve for the $s_{lm}^i(r)$. The resulting differential equations for $s_{lm}^i(r)$ will be ordinary differential equations with r as the independent variable. (This method is also used to find free oscillations.)

The reason that a spherical harmonic expansion is useful for this problem, is that to a first approximation the earth is spherically symmetric and non-rotating. Earth tides (and free oscillations) computed for a spherical, non-rotating earth model turn out to be pretty accurate (I'll mention, later, what happens when rotation and ellipticity — the most important departure from spherical symmetry — are included in the tidal model.) And for a spherical, non-rotating earth model, vector spherical harmonics *separate* the differential equation and boundary conditions. That is, H acting on any one of the three harmonics in Equation 8.3.4, gives a result that has an angular dependence that can be described by harmonics with the same (l,m) values.

More specifically, H acting on $s_{lm}^1\hat{r}Y_l^m$ gives a result that can be written as a linear combination of $\hat{r}Y_l^m$ and $\overline{\nabla}Y_l^m$, where the coefficients depend on r and are related to $s_{lm}^1(r)$ (they are either radial derivatives of $s_{lm}^1(r)$, or simple multiplication). The result has no $Y_{l'}^{m'}$ dependence for $l'\neq l$ or $m'\neq m$, and, in fact, no $\hat{r}\times\overline{\nabla}Y_l^m$ dependence.

Similarly, H acting on $s^2_{lm}(r)\overline{\nabla} Y_l^m$ gives $\hat{r} Y_l^m$ and $\overline{\nabla} Y_l^m$ terms. And, H acting on $s^3_{lm}(r)\hat{r} \times \overline{\nabla} Y_l^m$ gives only a $\hat{r} \times \overline{\nabla} Y_l^m$ term.

This separation should not be surprising. You undoubtedly encountered situations when studying electricity/magnetism or quantum mechanics, where you found that Y_l^m's separated spherically symmetric, *scalar* differential equations. Vector problems are no different.

The implication of this separation for tides is useful. Note that V_T' is proportional to $r^2 Y_2^m$, so that:

$$\overline{\nabla} V_T' \propto \frac{2}{r} \hat{r} Y_2^m + r^2 \overline{\nabla} Y_2^m.$$

So, the forcing term in the differential equation can be written as a sum of two vector spherical harmonics, both with $l = 2$. If you trace back through the discussion of what H does to vector spherical harmonics, you will see that the only coefficients in the expansion for the tidal solution $\overline{s}$ which are non-zero, are $s^1_{2m}(r)$ and $s^2_{2m}(r)$, so that:

$$\overline{s} = s^1_{2m}(r)\hat{r} Y_2^m(\theta, \phi) + s^2_{2m}(r)\overline{\nabla} Y_2^m(\theta, \phi).$$

So, it takes only two radially-dependent functions to describe tides for a spherically symmetric earth.

Incidentally, the fact that spherical harmonics separate the equations, is also important for seismic free oscillations. It implies that a free oscillation eigenfunction is described by harmonics with a single (l, m). There may, of course, be lots of different eigenfunctions for that single (l, m). This result is analogous to the result for the hydrogen atom eigenfunctions in quantum mechanics: each eigenfunction has a single (l, m), but you need a third index, n, to label the different eigenfunctions corresponding to a given (l, m). For seismic free oscillations you get a further separation between the $\hat{r} Y_l^m, \overline{\nabla} Y_l^m$ solutions, and the $\hat{r} \times \overline{\nabla} Y_l^m$ solution. A free oscillation eigenfunction is either described by a $\hat{r} \times \overline{\nabla} Y_l^m$ term, in which case it is called a *toroidal* free oscillation; or it's described by $\hat{r} Y_l^m$ and $\overline{\nabla} Y_l^m$ terms, in which case it is called a *spheroidal* free oscillation.

But, to get back to the tidal solution, note that $s^1_{2m}(r)$ represents radial displacements, and $s^2_{2m}(r)$ represents horizontal displacements ($\overline{\nabla} Y_l^m$ is in the $\hat{e}_\theta$ and $\hat{e}_\phi$ directions). You can only observe displacements at the earth's outer surface: $r = a$. So, you only observe $s^1_{2m}(a)$ and $s^2_{2m}(a)$. And that implies that for a given m and ω, you need only model and/or measure two numbers to describe displacements: $s^1_{2m}(a)$ and $s^2_{2m}(a)$. It is usual to define two *Love numbers*, h and l, which are non-dimensional normalizations of $s^1_{2m}(a)$ and $s^2_{2m}(a)$. Specifically, suppose $V'_T = (r/a)^2 Y_2^m(\theta,\phi)e^{i\omega t}$. Then, h and l are defined so that at the outer surface (where $r = a$):

$$\begin{array}{rrcccl}
\text{radial displacement:} & s_r(a) & = & \dfrac{h}{g}Y_2^m(\theta,\phi)e^{i\omega t} & = & \dfrac{h}{g}V'_T(r=a) \\
\text{southward:} & s_\theta(a) & = & \dfrac{l}{g}\partial_\theta Y_2^m(\theta,\phi)e^{i\omega t} & = & \dfrac{l}{g}\partial_\theta V'_T(r=a) \\
\text{eastward:} & s_\phi(a) & = & \dfrac{l}{g\sin\theta}\partial_\phi Y_2^m e^{i\omega t} & = & \dfrac{l}{g\sin\theta}\partial_\phi V'_T(r=a)
\end{array}$$

where g is the gravitational acceleration at the earth's surface.

It turns out you need to define a third Love number, k, to describe gravity observations. Once you know $\overline{s}(\overline{x},t)$ everywhere inside the earth, you can compute the perturbation in the earth's gravity field caused by $\overline{s}$ ($\overline{s}$ represents deformation — and so can be used to find the perturbation in the earth's density distribution). We define $\phi(\overline{x},t)$ as the perturbation in the earth's gravitational potential. ϕ is a scalar, so it can be expanded in terms of scalar $Y_l^m(\theta,\phi)$'s. It turns out that whenever $\overline{s}$ can be written as a sum of $\hat{r}Y_2^m$ and $\overline{\nabla}Y_2^m$ for a fixed value of m, then ϕ is proportional to Y_2^m, with the same value of m:

$$\phi(\overline{x},t) = \phi(r)Y_2^m(\theta,\phi)e^{i\omega t}.$$

We thus define a third dimensionless Love number k, so that when $V'_T = (r/a)^2 Y_2^m(\theta,\phi)$:

$$\phi(r=a,t) = kY_2^m(\theta,\phi)e^{i\omega t} = kV'_T(r=a). \tag{8.3.5}$$

All tidal observations can be described with linear combinations of the three numbers h, l and k. (As an example, I will describe the tidal signal in surface gravity in the following section.) You might expect that the Love numbers would depend on m and ω. But it turns out that for a spherical, non-rotating earth, the numbers are totally independent of

m (analogous to the result for the hydrogen atom that the radial eigenfunctions are independent of m) and are nearly the same at all tidal frequencies. So, you can completely describe the earth tides with just three numbers — valid (approximately) for all m and ω. You can determine these numbers from observations, and then compare with results obtained by solving the differential equation, to try to constrain, for example, models of the material properties (ρ, μ, λ) in the earth's interior.

Although people have used tidal observations to learn about the earth's interior (see the discussion below), tides have not turned out to be as useful as people had originally hoped — at least for learning about the earth's material properties. There are a number of reasons for this. The most obvious is that tidal observations can provide, at most, only three observational constraints: h, k and l. By comparison, over 1000 degenerate free oscillation frequencies have been cataloged, along with countless body- and surface-wave observations, and these have proven far more useful than tides for learning about the earth's material properties.

On the other hand, tides occur at longer periods than do seismic disturbances. Anelasticity in the mantle is apt to cause the apparent μ and λ to depend on frequency. Thus, tides offer the possibility of constraining that frequency dependence, and thus learning about anelasticity. So far, ocean loading uncertainties (see below) and other errors have limited what tidal observations have been able to say about anelasticity. But recent improvements in ocean tide models that use, for example, satellite altimeter data, may overcome this problem.

It is useful to be able to understand and model earth tides for another reason. Namely, tidal signals in such things as tilt, strain, surface displacements, and surface and satellite gravity, can be very large, and, if not removed, can easily obscure other signals that people might be interested in.

8.3.1 An Example: Surface Gravity Tides

I claimed above that every tidal observation can be written as a linear combination of Love numbers. As an example, consider the tidal effect on gravitational acceleration as measured by a gravimeter placed on the earth's surface. The meter records a change in gravity if:

1. there is a change in the gravitational potential; or

2. the surface is displaced vertically through the earth's original gravitational field.

Let's consider these two contributions separately.

1. The gravitational acceleration at any fixed point in space is g (positive downwards) $= -\partial_r V'$ where V' = the gravitational potential. We are interested in the tidal contributions to V'. There are two such contributions. One, V_T, comes from the direct attraction of the sun and moon. The other, ϕ, is caused by tidal perturbations of the earth's internal mass distribution.

   We consider the contribution from V_T, first. Since we are only including the $l = 2$ terms in our spherical harmonic expansion of V_T, the radial dependence of V_T is r^2 — see, for example, Equation 8.2.6. Thus, the effect on g at the outer surface (where $r = a$), is:

$$\Delta g = -\left.\partial_r V_T\right|_{r=a} = -\left.\frac{2}{a}V_T\right|_{r=a}$$

   Next, we include the contribution from ϕ. Consider any one of the $\left(\frac{r}{a}\right)^2 Y_2^m(\theta,\phi)e^{i\omega t}$ terms in V_T (see Equation 8.2.6). These terms cause tidal displacements in the earth, which lead to changes in the earth's internal mass distribution. The gravitational potential at the surface of the earth caused by this mass redistribution, is:

$$\left.\phi\right|_{r=a} = kY_2^m(\theta,\phi)e^{i\omega t}. \tag{8.3.6}$$

   (see Equation 8.3.5). To find the radial derivative of this potential, we need to know $\phi(r)$ for $r \geq a$. (The gravimeter sits outside the earth, where $r \geq a$.) That's a boundary value problem: solve $\nabla^2\phi = 0$ for $r > a$, where $\left.\phi\right|_{r=a}$ is given by Equation 8.3.6. The solution that is bounded at $r = \infty$ is:

$$\phi = k\left(\frac{a}{r}\right)^3 Y_2^m(\theta,\phi)e^{i\omega t} \qquad \text{for } r \geq a$$

where we have used the result that the radial dependence of a Y_l^m solution to Laplace's equation, is $r^{-(l+1)}$. So, outside the earth:

$$-\partial_r \phi = \frac{3}{r}\phi.$$

So, at $r = a$:

$$\Delta g = -\partial_r \phi|_{r=a} = \frac{3}{a}\phi\Big|_{r=a} = \frac{3}{a}kV_T\Big|_{r=a}$$

So, the total contribution to Δg caused by V_T and ϕ, is:

$$\Delta g = \left[-\frac{2}{a} + \frac{3}{a}k\right] V_T|_{r=a}.$$

2. The unperturbed gravitational acceleration outside the earth is $g(r) = GM/r^2$. The tidal force causes the outer surface to be displaced vertically by the distance $(h/g)\, V_T|_{r=a}$. So, the perturbation in g as observed by the gravimeter, is:

$$\Delta g = \frac{h}{g}V_T\Big|_{r=a} \partial_r g_0|_{r=a} = (\frac{h}{g}\, V_T|_{r=a})(-\frac{2}{a}g) = -\frac{2}{a}h\, V_T|_{r=a}.$$

So, adding the contributions from 1 and 2 gives the total tidal gravity signal recorded at the gravimeter:

$$\Delta g = -\frac{2}{a}\left[1 - \frac{3}{2}k + h\right] V_T|_{r=a}.$$

The factor $1 - \frac{3}{2}k + h$ is called the gravimetric factor, and is written as:

$$\delta \equiv 1 - \frac{3}{2}k + h.$$

Then:

$$\Delta g = -\delta\frac{2}{a}\, V_T|_{r=a}.$$

8.3.2 Numerical Results

Typical results (models and observations agree pretty well) are:

$$\begin{aligned} h &\approx 0.6 \\ l &\approx 0.085 \\ k &\approx 0.3 \\ \delta &\approx 1.16. \end{aligned}$$

Notice:

- $h > l$, implying that vertical tidal displacements are larger than horizontal tidal displacements at most locations;

- δ is close to 1, which is the value of δ you would obtain for a rigid earth. This implies that the gravitational effects of the vertical displacement nearly cancel the effects of the perturbation in the earth's gravity field.

You can go through a similar exercise relating tilt and strain to the Love numbers. I won't do that. Typical effects of tides on various measurement types are:

$$\begin{array}{ccll} \text{gravity} & \approx & 60\ \mu\text{gal} & \\ \text{strain} & \approx & 10^{-8} & \left(\text{strain} = \frac{\text{change in length}}{\text{length}}\right) \\ \text{tilt} & \approx & 40\ \text{msec of arc} & \\ \text{surface displacements} & \leq & 1\ \text{meter.} & \end{array}$$

8.3.3 What the simple model ignores.

What are the other reasons (besides "only three constraints") that have made it hard to learn about the earth's deep interior from tidal observations? Or, in other words, what have we ignored in the simple model above that can affect tidal observations? Here is a partial list:

8.3.3.1 The Oceans

The tidal force, of course, also causes ocean tides. The ocean tides act on the underlying solid earth and deform it. They force the solid earth in two ways:

1. They cause pressure at the surface of the earth: positive pressure at high tide, negative pressure at low tide.

2. The excess water mass at high tide (and the decreased water mass at low tide) acts gravitationally on the earth.

The induced deformation of the solid earth is called the ocean load tide. It occurs at the same frequencies as does the solid earth tide, since they are both caused by the same tidal forcing from the sun and moon. The effects of the load tide on gravity and on surface motion are typically 5%, or so, of the effects of the solid earth tide. The effects on tilt and strain, though, can be 100% or more of the body tide effects if the instruments are near the coast.

Thus, to learn about the solid earth from observations of the body tide, you must first somehow model and remove the effects of the ocean load tide. This is a two step process.

First, you need a model for the global ocean tide. Ocean tide models do exist, and have been greatly improved by the availability of TOPEX/Poseidon altimeter data. Those data can be used all by themselves to solve for the ocean tides. Or, they can be used to help constrain a hydrodynamical model.

Second, you must construct a geophysical model to predict the earth's response to surface loading. This is done by solving differential equations for the earth which are identical to the body tide differential equations described above. Except that in this case $\hat{n}\cdot \overset{\leftrightarrow}{\tau} \neq 0$ at the outer surface, because there is a non-zero surface traction caused by the load. These models allow you to compute load Green's functions, which describe the deformation of the earth caused by a point load on the surface. The Green's functions can then be convolved with an ocean tide model to predict the total oceanic loading.

Most of the uncertainty in this process comes from uncertainties in the ocean tide model. Ocean tide solutions based on TOPEX/Poseidon data are believed to have errors in the open ocean that are on the order of 2-3 cm, when summed over all tidal frequencies. Results near coasts are probably somewhat worse than this, because the relatively coarse spatial sampling of the altimeter makes it difficult to resolve the large spatial gradients that can occur in the tides for those regions. Coastal and shallow water tides are generally better obtained by using hydrodynamical models. In fact, TOPEX/Poseidon deep water solutions are sometimes assimilated into global hydrodynamical models to try to improve the tidal solutions both in the open ocean and near coasts. Errors of 2-3 cm are equivalent to relative errors of about 2-5% of the total tidal solution. Ocean tide errors of this magnitude cause errors of a few tenths of a percent in gravity body tide observations.

8.3.3.2 Local Effects

Tilt and strain tidal amplitudes can be very sensitive to local things, like topography, geology, and even the shape of the hole in the ground that you put your meter in. (You often place strainmeters and tiltmeters underground to get them away from even more serious surface environmental effects: such as fluctuations in atmospheric pressure, temperature, ground water, etc.) To learn about the earth's deep interior you must model all these local effects, and that can be difficult. On the other hand, if you can remove the effects of the global body tide on your observations by using global models, then you might be able to use the residuals to learn about the local effects; specifically, to learn about the local geological structure. This is one application where tides have proven useful.

8.3.3.3 Effects of rotation and ellipticity

For a rotating, elliptical earth the Y_l^m do not exactly separate the differential equations. For example, a Y_2^m term in the tidal potential causes a displacement field, $\overline{s}$, that has contributions from Y_l^m with $l \neq 2$. This effect on tidal observations is relatively small. It affects things at about the 1/300 level, which is the size of the earth's ellipticity, and of the ratio of the centrifugal force to the gravitational force.

Instead, the most important effect of rotation and ellipticity is to introduce significant frequency dependence into the Love numbers h, k and l. It causes the Love numbers to be resonant in the diurnal band. The resonance is caused by a normal mode of the earth with a period of $1 - 1/n$ days. This mode involves a relative rotation between the core and mantle, and "n" depends on the shape of the core/mantle boundary. For a hydrostatic shape, $n \approx 460$. People have looked for this resonance effect in the amplitudes of diurnal earth tides (the resonance also shows up in earth rotation observations — see Section 9.5, below), and have concluded that n is probably closer to 430. This has been used to constrain the non-hydrostatic core/mantle boundary shape.

There is also an observed phase lag between the earth tides and the tidal potential that is associated with the core-mantle rotational resonance in the tides. The size of this phase lag has implications for core-mantle dissipative processes.

Chapter 9

Earth Rotation

The earth does not rotate at a constant rate about a fixed axis. Instead:

- The rotation rate is variable.
- The rotation axis moves.

Consider the observational consequences of these two things, one at a time.

9.1 Variable Rotation Rate

A variable rotation rate is often referred to in terms of its effect on the length of a day (lod). For example, an increase in the rotation rate causes a decrease in the lod. Back in Chapter 2, I described how Δlod (the change in the lod) is observed. Until 20 years or so ago, the method was to record the transit times of stars. Today, people use VLBI, LLR, LAGEOS, and GPS observations, instead.

Until the advent of atomic clocks, people kept time by watching the stars. For example, every time star A was overhead, you might say it was 1:00 AM. The time determined this way is called Universal Time (UT). Or, to be more precise, UT is the time at Greenwich, England, determined in this manner.

When atomic clocks were developed in the 1950's, people found that UT didn't agree with atomic clock time (AT). You expect AT to be regular, in the sense that AT should be equal to the quantity t that enters into the classical dynamical equations of physics. The inference, then, is that UT is not constant. Or, in other words, that the earth's rotation rate varies with time. As an alternative to using Δlod, people often refer to variations in rotation rate in terms of the accompanying effects on UT: defined as δUT.

I want to find the relations between δUT, Δlod, and the change in the earth's rotation rate. UT can be determined from the transit times of objects in space; including stars, quasars, and artificial satellites. Because UT is observed to vary with respect to AT, we write UT = UT(t), where t is atomic time. Let $\Omega(t)$ be the angular velocity of rotation for the earth, so that Ω has units of rad/sec. We choose some arbitrary orientation of the earth as the initial epoch of UT. Suppose we are trying to determine UT by using a telescope to observe the locations of stars. We define $\phi(t)$ as the angle swept out by the telescope after time t. This angle is illustrated in Figure 9.1, where we are looking down on the earth from above the North Pole.

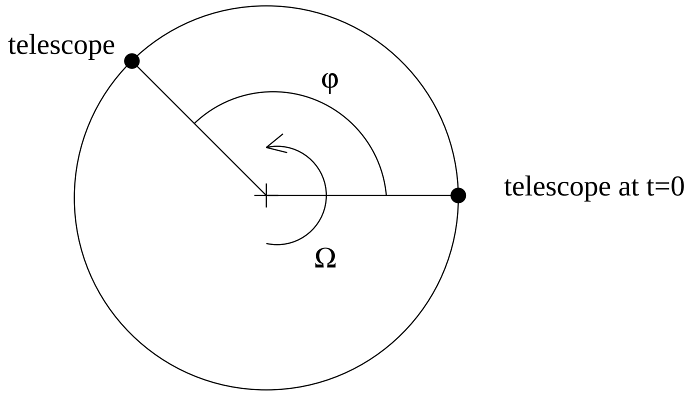

Figure 9.1:

The rotation rate is the time derivative of ϕ:

$$\frac{d\phi}{dt} = \Omega(t)$$

or:

$$\phi(t) = \int_0^t \Omega(t')\, dt'.$$

UT is defined by assuming (erroneously) that the earth is rotating at a constant rate of 1 revolution per 24 hours, so that $\Omega_0 \times \mathrm{UT} = \phi$, where Ω_0 is the average value of $\Omega(t)$ (Ω_0

= constant). So:

$$\mathrm{UT}(t) = \frac{1}{\Omega_0} \int_0^t \Omega(t')\, dt'.$$

Write $\Omega(t) = \Omega_0 + \delta\Omega(t)$, where $\delta\Omega(t)$ represents the variability in rotation rate. Then the variability in UT is:

$$\delta\mathrm{UT}(t) = \frac{1}{\Omega_0} \int_0^t \delta\Omega(t')\, dt'.$$

So if you observe $\delta\mathrm{UT}(t)$, then you can determine

$$\delta\Omega(t) = \Omega_0 \frac{d}{dt}\, \delta\mathrm{UT}(t).$$

Once you know $\delta\Omega(t)$, you can determine Δlod. The rotation rate, $\Omega = \Omega_0 + \delta\Omega$, is the number of radians per second subtended by the earth. The number of seconds per radian is then $1/(\Omega_0 + \delta\Omega)$. The number of seconds per revolution is defined as the lod. So:

$$\mathrm{lod} \approx \frac{2\pi}{\Omega_0 + \delta\Omega} \approx \frac{2\pi}{\Omega_0}\left[1 - \frac{\delta\Omega}{\Omega_0}\right]$$

assuming $\delta\Omega/\Omega_0 \ll 1$. So, the change in the lod is:

$$\Delta\mathrm{lod}(t) = -\left(\frac{2\pi}{\Omega_0}\right)\left(\frac{\delta\Omega(t)}{\Omega_0}\right)$$

where $2\pi/\Omega_0 = \mathrm{lod}_0$ = number of seconds per day if there were no variation in rotation rate. So:

$$\frac{\Delta\mathrm{lod}}{\mathrm{lod}_0} = -\frac{\delta\Omega(t)}{\Omega_0}. \tag{9.1.1}$$

9.2 Motion of the Rotation Axis

The rotation axis can undergo two types of motion.

1. The axis can move with respect to inertial space (as determined by the apparent positions of stars). This sort of motion is called *precession* and *nutation*.

2. The axis can move with respect to the earth. Imagine going to the North Pole and putting a stake in the ground. If you then stand back and watch the rotation axis, you discover that it moves with respect to the stake. You call this sort of motion

polar motion. It looks as though the rotation axis is moving, but actually that axis does not move much in inertial space. Instead, it is the earth (i.e. the stake) that moves. The earth can be described as wobbling about the rotation axis. For this reason, polar motion is sometimes also referred to as *wobble.*

Whenever there is nutation there is also wobble, and vice versa. For example, if the rotation axis moves in inertial space (i.e. if there is nutation), then the earth-fixed axis must also move in inertial space, since the rotation axis is the rotation axis *of the earth.* But, the earth-fixed axis won't move in quite the same way as the rotation axis, so there is relative motion between the two axes. This is wobble.

Still, it is useful to differentiate, conceptually, between nutation and wobble. Neither of them is directly measured with any technique. Instead, what *is* observed (by watching stars or satellites or the moon or quasars) is the inertial space motion of the *earth-fixed axis.* You never directly observe the rotation axis — and both wobble and nutation refer to motion of the rotation axis. Nevertheless, once you determine the motion of the earth-fixed axis, you can calculate the motion of the rotation axis.

What you find when you do all this is that you can characterize the motion as either mainly nutation or mainly wobble, depending on the frequency of the motion. If the earth-fixed axis moves at periods long compared to 1 day as seen from inertial space, then the motion is mainly a nutation: the rotation axis moves in inertial space but remains nearly coincident with the earth-fixed axis. If the earth-fixed axis moves around approximately diurnally, as seen from inertial space, then the motion is mainly wobble: the rotation axis doesn't move much in inertial space, but the earth-fixed axis does. I will not show this here.

If the motion occurs at some other period, you can't usefully identify it as either wobble or nutation. It will be a combination of the two. As it happens, though, little significant motion has been observed at other periods. That's evidently because there are few processes which can efficiently excite motion at other periods. Long-period inertial space motion (nutation) is excited by the gravitational attraction of the sun and moon. (The motion of the sun and moon in inertial space is long period.) And, diurnal inertial space motion is excited by processes that occur on and within the earth (changes in winds, or atmospheric pressure, or ocean currents, or fluid flow in the core, etc.). These processes involve slow changes with respect to the earth. Thus, since the earth is rotating, these processes appear to vary diurnally as seen from inertial space.

This leads to another point. Motion which is long period as seen from inertial space is approximately diurnal as seen from earth. And, motion that is diurnal as seen from space is long period as seen from the earth (at least it is if the motion as seen from space is prograde: in the same direction as the earth's rotation; and wobble *is* prograde). For example, during wobble the earth-fixed axis moves around diurnally in inertial space — and it moves in the same sense as the earth's rotation. So, the axis moves slowly with respect to points fixed in the earth, and thus the motion is long period as seen from the earth.

It is the period with respect to the earth that you should remember:

$$\begin{aligned} \text{wobble} &= \text{long period} \\ \text{nutation} &= \text{diurnal.} \end{aligned}$$

Let's discuss these things (Δlod, precession/nutation, wobble) in more detail.

9.3 Changes in the lod

There are variations in the lod at many different time scales.

- First, there is a linear increase in the lod of 1–2 msec/century (1 day = 86,400 sec, so 1 msec $\cong$ one part in 10^8).

- Second, there are irregular "decadal fluctuations" of 4–5 msec over 20–30 years.

- Third, there are variations of 2–3 msec at periods of less than 5 years. Most of this short-period variability occurs at distinct periods: 2 weeks, 1 month, 6 months, 1 year.

9.3.1 Linear change in the lod

You can't separate the linear change from the decadal fluctuations unless you have a data span that is significantly longer than a century. The way to obtain such a long data record is to use ancient Greek, Babylonian, etc., eclipse records. You note the recorded time of day when the eclipse occurred. You then compare with what you predict, given the present positions of the sun and moon together with the present-day rotation rate. The discrepancy is a measure of the integrated change in rotation rate since the time the eclipse occurred. Specifically, you find the change in UT between the time of the eclipse ($t = t_1$) and the present-day ($t = t_2$):

$$\delta \mathrm{UT} = \frac{1}{\Omega_0} \int_{t_1}^{t_2} \delta\Omega(t')\, dt'.$$

For a linear change in rotation rate: $\delta\Omega = -\delta\dot{\Omega} t$, where $\delta\dot{\Omega} = \text{constant}$. (I have inserted a negative sign into this equation for $\delta\Omega$, because then $\delta\dot{\Omega} > 0$: i.e. a linear increase in the lod implies that $\delta\Omega < 0$.) So:

$$\delta \mathrm{UT} = -\frac{\delta\dot{\Omega}}{\Omega_0} \frac{t_2^2 - t_1^2}{2}.$$

Equation 9.1.1 implies that:

$$\begin{aligned} -\frac{\delta\dot{\Omega}}{\Omega_0} &= \frac{\partial_t(\Delta \mathrm{lod})}{\mathrm{lod}_0} \\ &\approx \frac{2 \text{ msec/century}}{86,400 \text{ sec}} \\ &= 2 \times 10^{-8}/\text{century}. \end{aligned}$$

So, after one century:

$$\delta \mathrm{UT} = \left(\frac{2 \times 10^{-8}}{\text{century}} \right) \times \frac{1}{2} \text{ century}^2 = 10^{-8} \text{ century} \approx 30 \text{ seconds}.$$

After 2000 years:

$$\delta \mathrm{UT} = \left(\frac{2 \times 10^{-8}}{\text{century}} \right) \times \frac{400}{2} \text{ century}^2 \approx 3.3 \text{ hours}.$$

This means that an eclipse 2000 years ago would have occurred 3.3 hours earlier (as measured by looking at, for example, the sun or moon) than would be predicted using today's rotation rate to extrapolate backward in time. A discrepancy this large would be clearly evident in the ancient eclipse data.

People also have used planetary occultation data (for example, when a planet disappears behind the sun) from the 1600's and later, to look for lod fluctuations. There the record is more precise, but the secular change in UT is not as large. For example, after 400 years: $\delta\mathrm{UT} \approx 8$ minutes. The ancient data is thus more useful for inferring the secular increase in the lod.

Most of the secular change in the lod is due to tidal friction in the earth and oceans. The moon (for example) causes a tidal bulge in the earth and oceans, which is oriented towards the moon. If there was no energy dissipation, the bulge would be oriented *exactly* towards the moon. But, because there *is* dissipation, the earth and oceans take a short time to fully respond to the tidal force. The maximum uplift of the surface occurs a short time after the moon is overhead, and so the tidal bulge leads the earth-moon vector by the angle λ (shown greatly exaggerated in Figure 9.2). The angle λ is about 0.4 degrees, corresponding to about a 10 minute lag time in the earth. The moon's gravitational force

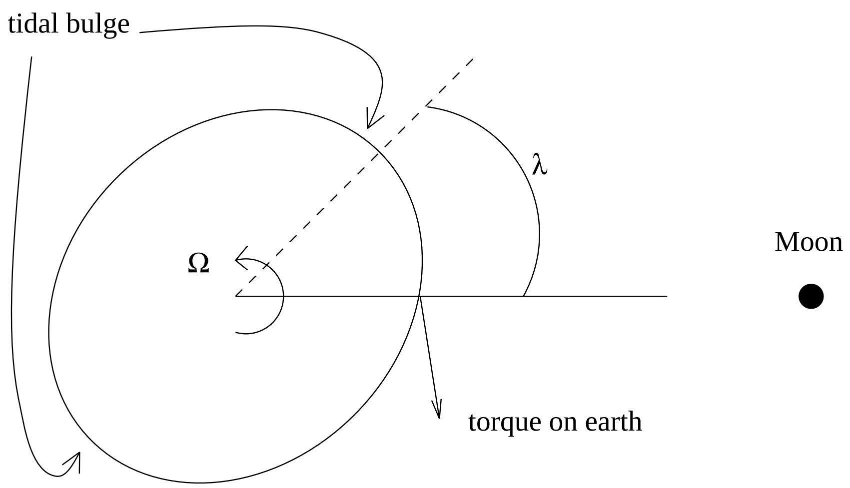

Figure 9.2:

can then act on this bulge to torque the earth in a direction opposite to the rotation. Thus the earth slows down, and so the lod increases. If you break the tidal force into its individual tidal constituents (i.e. the $e^{i\omega t}$ terms in Equation 8.2.6), you find that all the diurnal and semi-diurnal terms contribute to the torque, but that the lunar semi-diurnal terms contribute the most. About 80% of the dissipation comes from the lunar tides; the remaining 20% is from the solar tides.

Incidentally, a similar process presumably occurred on the moon in the past. Dissipation of tidal energy in the moon (the tides in that case are caused primarily by the earth) slowed the moon's rotation, so that now the moon always keeps the same side to the earth.

But, returning to the earth: Most of the energy dissipation in the earth occurs in the oceans, rather than in the solid earth. That is, it is mostly the tidal bulge in the *oceans* that leads the earth-moon vector. This is because frictional effects are simply more important in the oceans than in the earth. Ocean tide models now do a pretty good job of reproducing the dissipation required by the lod observations.

You can independently determine the lag angle, λ, using satellite ranging data, since the tidal bulge affects satellite orbits. Once you know λ you can predict the decrease in rotation rate. What you find is that the earth should be slowing down by about 25% more than the value inferred from the ancient eclipse data.

And there's still another, and even better, way to obtain an estimate of the effects of tidal friction. The earth's tidal bulge acts gravitationally on the moon and causes it to accelerate in the direction of the earth's rotation (counter-clockwise in Figure 9.2, above). This can also be understood in terms of the conservation of angular momentum. If the earth loses angular momentum due to its decreasing rotation, the moon must gain an equal amount of angular momentum, through changes in its orbit. What you find is that the moon moves further away from the earth and increases its orbital period about the earth.

People have determined the increase in orbital period using LLR data. It's easier to detect that than to detect the increase in the orbital radius, because an offset in the orbital period builds up over many orbits to give a large change in the lunar position. The results are then used to estimate the decrease in the earth's rotation rate. The result obtained in this manner is consistent with the satellite estimates of λ: the effect of tidal friction is about 25% larger than the secular change inferred using the ancient eclipse data.

Incidentally, if you use the LLR and satellite results — or even the eclipse results — to extrapolate backwards in time, you conclude that the moon was so close to the earth 1.5 billion years ago that it would have been torn apart by gravitational tidal forces from the earth. Yet the moon is known to be at least 4 billion years old. The explanation is that the current oceanic dissipation rates are apparently anomalously large. The oceanic dissipation depends, among other things, on the shape of the ocean basins. And the basins have changed with time due to continental drift.

Anyway, although the 25% discrepancy could simply be the result of noisy ancient eclipse data, it is generally believed to be real. It implies that there is some other mechanism causing the earth's rotation rate to *increase* secularly with time. It is likely that this increase in rotation rate is caused by a decrease in the earth's moment of inertia. The situation is similar to a spinning figure skater raising her hands over her head. This decreases her moment of inertia, and so she spins faster to conserve angular momentum.

The reason people believe that the 25% discrepancy is real and is due to a decrease in the moment of inertia, is that this interpretation is consistent with the LAGEOS $\dot{J}_2$ results. The geoid coefficient J_2 is related to the earth's polar moment of inertia. So the LAGEOS $\dot{J}_2$ observations can be used to directly infer the secular change in the lod, due to the re-distribution of mass in or on the earth. And the results predict an increase in the lod that is consistent with the 25% discrepancy between the eclipse data and the effects of tidal friction. Assuming this interpretation is correct, it also implies that whatever has been causing $\dot{J}_2$ over the past couple decades (the time span of the LAGEOS data), has been going on at pretty much the same rate over the past few thousand years.

What is causing this secular change in the moment of inertia? Or, equivalently, what is causing the observed change in J_2? As described in Section 7.5, the result has been interpreted as due to postglacial rebound. But it may also partially reflect on-going changes in the mass of the polar ice caps (Antarctica and Greenland). In fact, it is likely that there are significant contributions from *both* these (and other) mechanisms.

9.3.2 Decadal Fluctuations in the lod

The decadal fluctuations are believed to be due to the transfer of angular momentum between the core and the mantle. If, for example, the core loses angular momentum, then the mantle must gain angular momentum in order to conserve the total angular momentum of the earth. This would cause the rotation rate of the mantle to increase, so that the observed lod decreases.

This suggests that there might be a correlation between the decadal lod fluctuations and decadal variations in the magnetic field. The situation, though, is complicated by the fact that magnetic field variations generated in the core will be attenuated and delayed as they pass up through the conducting mantle. So, the time dependence of the magnetic field at the outer surface may not look too much like the time dependence of the magnetic field at the top of the core. And, it is the field in the core that ought to be correlated with the fluid velocity.

On the other hand, suppose you assume the lod time dependence *does* look like the time dependence of the magnetic field at the top of the core. And suppose you can somehow identify the corresponding time-dependent signal in the magnetic field observed at the outer surface. You are then in a position to compare the time signature of the magnetic field at the outer surface, with that of the magnetic field at the top of the core (as inferred from the lod data), and this tells you something about the conductivity of the intervening mantle. People have used this approach to place bounds on mantle conductivity.

This, though, tells you nothing about the change in angular momentum of the core, and whether it could be large enough to explain the lod fluctuations. Recently there has been some significant progress on that problem. Estimates of fluid flow at the top of the core can be made from surface magnetic field data. Those estimates have been used, together with assumptions about how the fluid flow inside the core might be organized, to obtain estimates for the change of core angular momentum. The agreement with the lod data is remarkably good. At the very least, the results suggest that the core is, indeed, the likely source of excitation for the decade-scale fluctuations.

As for the core-mantle coupling mechanism responsible for this exchange of angular momentum, there is as yet no consensus. One possible source of the coupling is fluid pressure acting against topography on the core-mantle boundary. Another possibility is electro-magnetic forcing. The idea there is that the earth's magnetic field is caused by currents in the core. If the currents change with time, the field will change. The mantle is an electrical conductor, so the changes in the magnetic field induce currents in the mantle. These currents interact with the original magnetic field from the core, through the Lorentz Force, and the result can be a rotation of the mantle. It may be that both topographic and electro-magnetic coupling are important.

9.3.3 Short period lod fluctuations

There are a variety of lod fluctuations at periods of a few years and shorter, that are due to: earth and ocean tides; the atmosphere; and, to a lesser extent, wind-driven oceanic circulation.

First, there are variations in the lod at monthly and fortnightly periods. The amplitudes at these periods are approximately 0.2 to 0.3 msec. These are caused by the long period tides. The $l = 2$, $m = 0$ tidal deformation leads to a change in the earth's polar moment of inertia. But there is no net torque on the earth about the polar axis, and so there can be no change in polar angular momentum. Thus, the change in moment of inertia must be accompanied by a change in rotation rate.

It is not hard to understand why this variability occurs only at long periods, and not at diurnal or semi-diurnal periods. The polar moment of inertia varies as mass moves toward or away from the poles. So, it is sensitive to the orientation of the tidal bulge with respect to the equator. See Figure 9.3. That orientation is determined by the position of the moon (and sun) with respect to the equator. And so it varies only with the moon's orbital motion. It does *not* vary with the earth's rotational period. Thus, it is only the long period tides which can affect the rotation rate.

To model these variations in rotation, you must compute the tidal variations in the earth's moment of inertia. This means that you must model the long period earth tides.

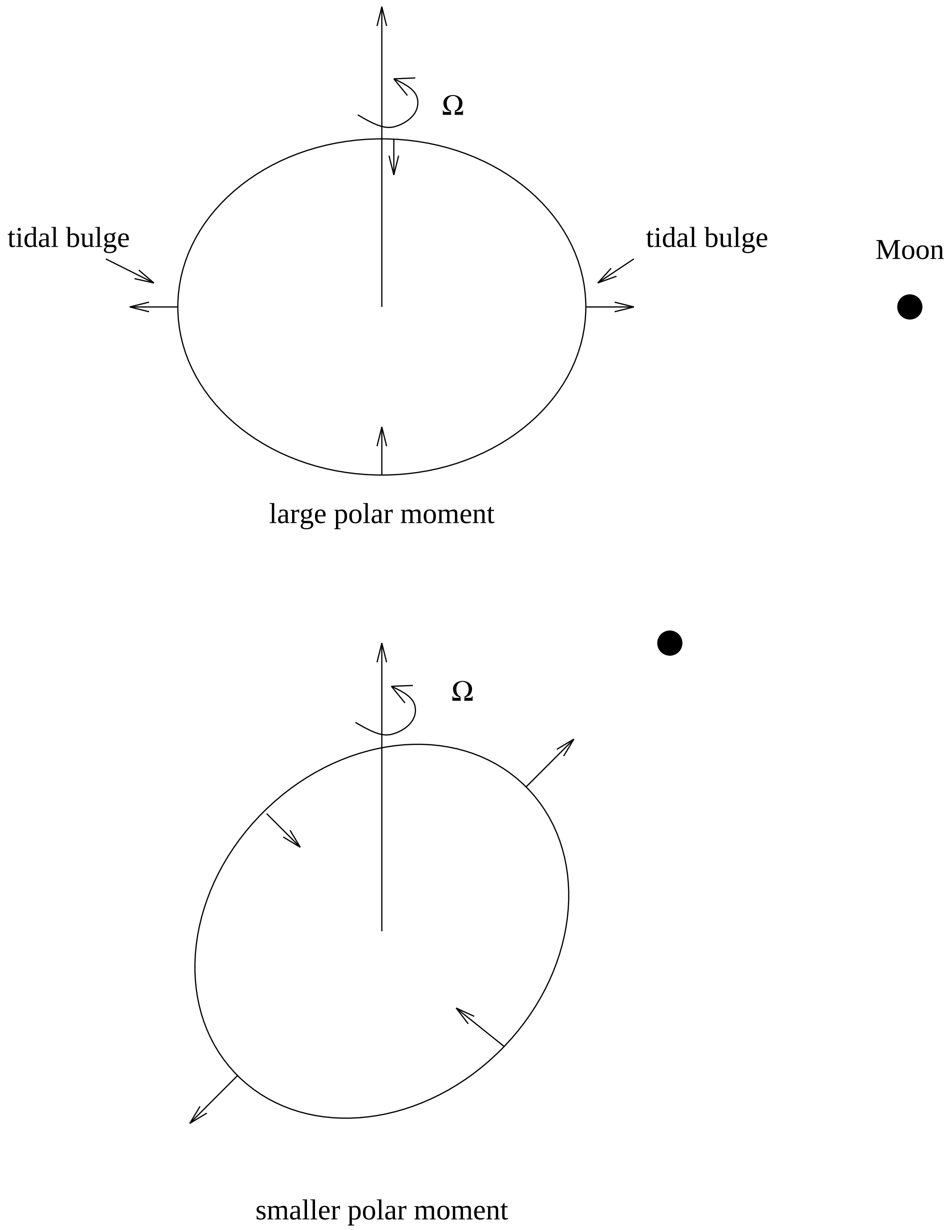

Figure 9.3:

You must also model the long period — monthly and fortnightly — ocean tides, since they, also, will contribute to the moment of inertia. The ocean tide effects are approximately 10% of the earth tide effects.

People have obtained good agreement between theory and observations of these lod terms. They have used the results to estimate the effects of mantle anelasticity at the fortnightly and monthly periods.

Long period tidal variations in lod occur at all long period tidal frequencies — not just at the monthly and fortnightly frequencies. But there are only four frequencies where the effects are large enough to be observable: fortnightly, monthly, 6 months, 12 months. (There is also a reasonably large effect at 18.6 years, but it is overwhelmed by the decadal fluctuation signal.) The problem at 6 months and 1 year is that there are even larger effects at those periods due to atmospheric and non-tidal oceanic perturbations. The tidal deformation mechanism accounts for only 10% of the annual and 30% of the semi-annual variability. About 5% of the annual and semi-annual variability comes from seasonal variations in ocean currents — most of that from variations in the circum-polar current around Antarctica. When that current weakens, for example, it loses angular momentum to the earth, and that lost angular momentum shows up as a change in the earth's rotation rate.

But, most of the 6 month and 1 year variability is caused by seasonal atmospheric forcing: particularly, the exchange of angular momentum between the solid earth and atmospheric winds at seasonal periods. For example, when the winds increase in strength from west to east, the earth slows down. This exchange of angular momentum is caused by a combination of surface friction torques (due to viscous drag as the winds blow over the surface) and mountain torques (caused by higher pressure on one side of a topographic feature than on the other — although you can also think of it as the force caused by winds blowing against mountains). The friction and mountain torques contribute about equally to the seasonal coupling.

Variations in atmospheric pressure contribute about 10% of the 6 month and 1 year lod amplitudes, so they contribute maybe 1/6 to 1/7 of the effects of winds. Here's a way to understand how changes in atmospheric pressure are related to changes in the lod:

A change in pressure means there is a change in the total atmospheric mass above that point. So, with enough pressure data, it is possible to determine variations in the atmosphere's moment of inertia. Because of conservation of angular momentum of the combined earth + atmosphere (+ oceans), seasonal variations in the atmosphere's moment of inertia cause variations in the lod. (Incidentally, why don't variations in the atmosphere's moment of inertia cause variations in the *atmosphere's* rotation, instead of in the rotation of the solid earth? Well, they might. In that case, you would have added an effect to the predicted lod which isn't really there. But, presumably, you would take that effect out again when you find the effects of winds, since the wind data would include this extra atmospheric rotation.)

People have used wind and pressure (and ocean current) data to estimate the effects of the atmosphere and oceans on the seasonal lod variations. The agreement is remarkable (after first removing the 6 month and 1 year tidal effects). In fact, the atmosphere (and, to a far lesser extent, the oceans) appear to cause essentially all of the observed short-period (less than 5 years), non-tidal lod variations. These include a 2 year term, a 50 day term, and irregular fluctuations associated with the El Niño event in the Southern Pacific.

9.4 Wobble

The observed wobble spectrum is much simpler than the lod spectrum. There is some evidence of a linear drift of the rotation pole with respect to the earth's surface: about 3 or 4 milli-arc-seconds/year in a direction somewhat between Greenland and Hudson Bay. This has been interpreted as due to post-glacial rebound. The idea is that as material flows into the region beneath Hudson Bay, the earth's inertia tensor changes, and that causes the pole to move. Interpreted in this way, the linear drift provides a good constraint on the lower mantle viscosity. On the other hand, present-day melting of ice in Greenland or Antarctica could also have important contributions to the observed polar drift, as could tectonic effects (i.e. mantle convection).

There is also evidence of a long period wobble, with about a 30 year period. The amplitude is on the order of 30 milli-arc-seconds. Nobody knows what is causing this motion.

9.4.1 Annual Wobble

Otherwise, there are just two features in the wobble spectrum that are obviously significant: a 12 month term and a 14 month term. The 12 month "annual wobble" is mostly caused by an annual variation in the inertia tensor of the atmosphere. To conserve angular momentum of the earth/atmosphere system, the rotation axis shifts in response to the inertia tensor perturbation. The amplitude of the annual wobble is approximately 0.1 arc-seconds of motion of the pole with respect to the earth. That translates to about 3 meters of displacement at the earth's surface.

You can estimate this effect from atmospheric pressure data. You find that most of the contributions come from the pressure variations associated with the Asian monsoon: high pressure (more mass) over Asia in winter, low pressure (less mass) in summer.

This pressure-related effect is responsible for most, but not all, of the annual wobble. An additional 25% of the observed amplitude is evidently due to seasonal variations in the distribution of water: in snow and ice, in soil moisture and ground water, in rivers and lakes, and in the ocean. These variations in water storage can cause perturbations in the inertia tensor at seasonal periods. Their contributions are hard to evaluate, so the estimate of 25% is only approximate.

The effects of winds and ocean currents on wobble are apparently negligible. Why are the effects of pressure dominant for the annual wobble, while being relatively unimportant for the annual variation in the lod? To explain the answer, we need to consider the possible coupling mechanisms that can act between between the atmosphere and the solid earth. The only way the atmosphere can cause either wobble or fluctuations in the lod, is through friction or mountain torques. Mountain torques can act on any surface topographic feature, including on the earth's enormous elliptical bulge (remember that the earth's equatorial radius is approximately 20 km larger than its polar radius). There can thus be pressure torques acting against this bulge. But, the axis of symmetry for the ellipse is the earth's mean rotation axis. Thus, the pressure torque in Figure 9.4 can cause wobble, but *not* variations in the lod. It turns out that the effects of this pressure torque are exactly equivalent to the effects of a change in the atmospheric inertia tensor.

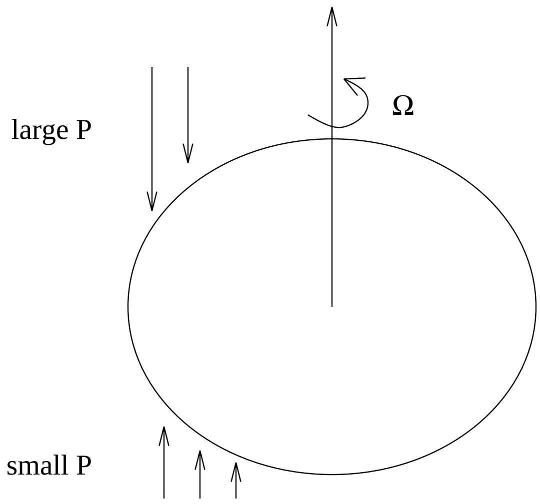

Figure 9.4:

Or, perhaps more clearly, you can think of the atmospheric effects on wobble as either:

$$\begin{pmatrix} \text{change in inertia tensor} \\ \text{(inferred from pressure data)} \end{pmatrix} + \begin{pmatrix} \text{change in angular} \\ \text{momentum of winds} \end{pmatrix}$$

or

$$(\text{elliptical bulge torque}) + \begin{pmatrix} \text{other mountain torques} \\ + \\ \text{friction torque} \end{pmatrix}.$$

The inertia tensor and elliptical bulge contributions are equivalent; and the wind and "other mountain torques and friction torque" contributions are equivalent. The elliptical bulge torque dominates, and so the pressure effects are much larger than the wind effects. Because there is no elliptical bulge torque along the polar axis, the pressure effects on the lod are much less important.

9.4.2 Chandler wobble

The 14 month term is called the Chandler wobble, after the man who discovered it around 1890. It is a normal mode of the earth. It is analogous to the free nutation of a top. And, just as for a top, the period depends on the earth's rotation rate (Ω), and on how non-spherical the earth is. The relevant parameter describing the earth's aspherical shape is $(C-A)/A$, where C and A are the polar and equatorial moments of inertia, respectively. For the earth, $(C-A)/A \approx 1/300$, and so the expected period would be about 300 days (frequency $= \Omega(C-A)/A \approx 10$ months).

Instead, Chandler found a 14 month (about 430 day) period. Its amplitude is variable, but is on the same order as - maybe even a little larger than - the amplitude of the annual wobble. The four-month discrepancy is due to deformation inside the earth caused by the change in the centrifugal force associated with the wobble. There are also effects from the fluid core and from the oceans which tend to cancel: the oceans increase the period by 1 month, and the core decreases it by 1 month.

The Chandler wobble period, and the observed Chandler wobble damping time of approximately several decades, are now well modeled. In fact, the observations have been used to estimate the effects of mantle anelasticity at a period of 14 months.

But people don't yet know what excites the Chandler wobble. Atmospheric and oceanic effects appear to provide no more than about 25% of the necessary power. Earthquakes (which perturb the inertia tensor) provide less than 10%. People have hypothesized that pressure torques on the mantle due to the fluid core might be a viable excitation mechanism, though it is very difficult to assess this idea quantitatively. Other mechanisms for core/mantle torques (such as electro-magnetic coupling) don't appear to be important.

9.5 Nutations and Precession

These are caused by the gravitational attraction of the sun and moon. The sun and moon torque the earth, as shown in Figure 9.5, through their gravitational force on the earth's elliptical bulge. Because the earth is rotating, it responds to that torque as a top would: it moves out of the page. In fact, the rotation axis precesses about the earth-moon vector. But the earth-moon vector changes its orientation as the moon orbits the earth. So the precession occurs about a moving axis. You can think of the resulting motion of the rotation axis as a precession about the normal to the ecliptic (the ecliptic is the

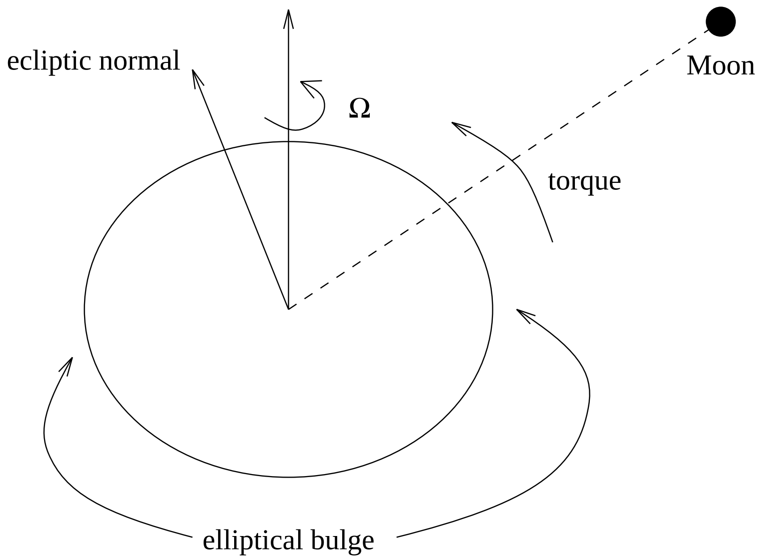

Figure 9.5:

plane of the moon's orbit) *plus* a series of higher frequency wiggles due to the motion of the moon (and sun). The precession is called the *precession of the equinoxes*. It has an angular amplitude of 23.5°, and a period of 26,000 years as seen from inertial space. The wiggles are called *nutations*. They have inertial space periods equal to the orbital periods of the sun and moon: 13.7 days, 27.6 days, 6 months, 1 year, 18.6 years, etc. The 18.6 year motion is the largest: about 20 arc-seconds (or 1/2 km) motion of the rotation axis. The other nutation terms have much smaller amplitudes. For example, the 6 month term has an amplitude of about 1 arc-second, which is equivalent to $\approx$ 30 m of motion at the earth's surface.

This motion is long period as seen from space, but it is diurnal as seen from points fixed on the earth. For example, the annual nutation appears from the earth to have periods of $(1 + (1/365))$ cycles/day and $(1 - (1/365))$ cycles/day (an annual modulation of 1 cycle/day gives two frequencies).

What makes the nutations interesting from a geophysical perspective, is that the earth has a normal mode, called the *free core nutation*, with a nearly diurnal eigenfrequency. I briefly described this mode when discussing earth tides on a rotating, elliptical earth (see Section 8.3.3.3). The situation for nutations is similar to that for earth tides. Namely, the nutation amplitudes are resonant at the free core nutation eigenfrequency. The eigenfrequency depends on the ellipticity of the core-mantle boundary. If the earth was hydrostatically pre-stressed, so that the boundary ellipticity was equal to the value predicted by solving Clairaut's equation, then the nutations would be resonant at $(1+(1/460))$ cycles/day (the free core nutation eigenfrequency for the hydrostatic case). However, the nutation observations show that the resonance actually occurs at a frequency closer to $(1+(1/430))$ cycles/day, a result consistent with the conclusions from earth tide studies. The difference between the observed resonance and the hydrostatic prediction, suggests that the core-mantle boundary ellipticity is about 10% larger than the hydrostatic value. This corresponds to about 1/2 km of non-hydrostatic Y_2^0 boundary topography.

Acknowledgments

Samizdat Press gratefully acknowledges the generous support of the Institute for Resource and Environmental Geosciences at the Colorado School of Mines, Unocal, Amerada Hess, the Society of Exploration Geophysicists (SEG) and Dr. J. Bee Bednar.