5 The geometry of the earth

5.1 Overview

This chapter consists of three parts.

Part I: Global reference systems after GPS
A fundamental task is to define a global reference system based on a reference ellipsoid which is a good global representation of the earth and which should ideally be characterized by the following properties: Its center coincides with the geocenter (the earth's center of mass), its z-axis represents a suitably defined mean rotation axis of the earth, and the xz-plane is parallel to a mean plane close to the Greenwich meridian. The reference ellipsoid itself is defined to be an ellipsoid of revolution that globally approximates the geoid best in some global sense.

Actually, such a geometric or physical definition cannot be absolutely accurately and unambiguously realized; the final definition will always contain an arbitrary conventional element.

To make things even more complicated, the earth is not a completely rigid body. It can (again approximately!) be regarded as an elastic body with a liquid core. It undergoes small more or less periodic changes. So it must be referred to a mean ellipsoid that does not change with time.

All this will be taken for granted in the present introductory treatment. We shall assume a well-defined geocentric reference ellipsoid with rigid dimensions, a fixed origin, and a time-invariant orientation – close to reality but, in principle, conventionally adopted. For temporal changes in the earth's body and rotation, the reader may be referred to Moritz and Mueller (1987).

Before the advent of satellite geodesy, a geocentric reference system could not be realized. Thus, we had to work with a local geodetic system displaced with respect to the geocenter by an unknown amount on the order of up to a few hundred meters. Therefore, we must take into account a translation (parallel shift) of the local reference ellipsoid with respect to a geocentric system. This implies three translation parameters.

Note that "local" here is used in the sense of "regional", i.e., for a country, territory, or region, in contrast to "global".

Usually, the orientation of a local reference system is accurately known since the direction of the xyz-axes was accessible by astronomical measurements quite accurately at least for the last two centuries. Thus, the orientation of a local geodetic datum is known to the order of $0.1''$ (arc seconds).

Today, we can readily determine the deviation of a local system or *local datum* from a global reference system. We have the deviation of

- size and shape of the reference ellipsoid (a, f),
- translation (x_0, y_0, z_0), and
- orientation (three very small Euler angles $\varepsilon_1, \varepsilon_2, \varepsilon_3$).

Since GPS is very well established (cf. Hofmann-Wellenhof et al. 2001), we assume a general knowledge for granted and recapitulate in this book only some basic facts.

Part II: Three-dimensional geodesy: a transition
This part considers how the concepts of geodesy in the modern sense of Molodensky, Marussi, and Hotine would look shortly before the advent of satellites, but already including electronically measured spatial distances (trilateration). We work with local Cartesian coordinates rotated in a known way by the astronomically measurable quantities Φ, Λ, A (astronomical latitude, longitude, azimuth), considered as Eulerian angles of rotation of the local with respect to the global axes. However, we have no means to determine the geocenter. So the situation is somewhat more complicated but still geometrically well defined and transparent. "Local" here means "strictly local", varying from point to point together with their plumb lines defined by (Φ, Λ).

The main problem with this approach is the impossibility of measuring precise zenith angles because of atmospheric refraction. We may say that the vertical dimension is much worse defined than the horizontal dimension.

Finally, we shall consider how terrestrial and GPS data can be combined.

Part III: Local geodetic datum
The way out of the dilemma of the worse vertical dimension is a complete separation of horizontal and vertical and determining the latter by the differential method of astrogeodetic geoid determination. This was a "2+1-dimensional" rather than a three-dimensional approach, logically more complicated but practically more accurate. In fact, the former (and present) astrogeodetic methods can be understood much better by deriving them from the global situation. Thus, today with GPS we are in a much better position practically as well as theoretically: the classical local datums can be understood best by their relation with the global geometry. "Local geodetic system" or "local geodetic datum" is again meant in the sense of "regional", e.g., the North-American Datum or the European Datum.

GPS permits to separate the geometry from the gravity field, which continues to be a challenge for physical geodesy to be solved by a combination of terrestrial and satellite data.

Part I: Global reference systems after GPS

5.2 Introduction

Geodesy, as the theory of size and shape of the earth, is not a purely geometrical science since the earth's gravity field, a physical entity, is involved in many geodetic measurements, especially terrestrial ones.

The gravimetric methods are usually considered to constitute physical geodesy in the narrower sense. The measurements of triangulation, leveling, and geodetic astronomy, all make essential use of the plumb line, which, being the direction of the gravity vector, is no less physically defined by nature than its magnitude, that is, the gravity g. All determinations of the geoid by various methods and its use as well as the use of deflections of the vertical belong to physical geodesy, quite as well as the gravimetric methods.

Even in the age of GPS, we have many previous geodetic data which continue to be useful and have to be understood in order to be optimally combined with the new satellite data. In precise operations of engineering geodesy such as tunnel surveying, the plumb line and deflections of the vertical must be taken into account.

For an optimal understanding and use of local (or rather regional) geodetic datums, we must know their relation to a global geodetic system as used in GPS. Therefore, it is appropriate to start with global geometry in a rather elementary way.

A few introductory ideas may help in comprehending this subject. To fix the position of a point in space, we need three coordinates. We can use, and have used, a rectangular Cartesian coordinate system. This is the basic geometric coordinate system. It may be easily converted computationally to ellipsoidal coordinates φ, λ, h referred to any given reference ellipsoid.

For many special purposes, however, it is preferable to take what we have called the *natural coordinates*: Φ (astronomical latitude), Λ (astronomical longitude), and H (orthometric height), which directly refer to the gravity field of the earth (Sect. 2.4). The height H may be obtained by geometric leveling, combined with gravity measurements, and Φ and Λ are determined by astronomical measurements.

As long as the geoid can be identified with an ellipsoid, the use of these coordinates for computations is very simple. Since this identification is sufficient only for results of rather low accuracy, the deviations of the geoid from an ellipsoid must be taken into account. As we have seen, the geoid has rather disagreeable mathematical properties. It is a complicated surface with discontinuities of curvature. Thus, it is not suitable as a surface on which to perform mathematical computations directly, as on the ellipsoid.

To repeat, the ellipsoidal coordinates φ, λ, h are defined such as to refer to the ellipsoid exactly as the natural coordinates refer to the geoid, hence their names.

Since the deviations of the geoid from the ellipsoid are small and computable, it is convenient to add small reductions to the original coordinates Φ, Λ, H, so as to get values which refer to an ellipsoid. In this way we shall find in Sect. 5.12:

$$\begin{aligned} \varphi &= \Phi - \xi\,, \\ \lambda &= \Lambda - \eta \sec\varphi\,, \\ h &= H + N\,; \end{aligned} \tag{5–1}$$

φ and λ are the ellipsoidal coordinates on the ellipsoid, sometimes also called *geodetic latitude* and *geodetic longitude* to distinguish them from the *astronomical latitude* Φ and the *astronomical longitude* Λ. Astronomical and ellipsoidal coordinates differ by the deflection of the vertical (components ξ and η). The quantity h is the *geometric height* above the ellipsoid; it differs from the *orthometric height* H above the geoid by the geoidal undulation N.

Geodetic measurements (angles, distances) are treated similarly. The principle of *triangulation* is well known: historically, distances were obtained indirectly by measuring the angles in a suitable network of triangles; only one baseline was necessary in principle to furnish the scale of the network. Triangulation was indispensable in former times, because angles could be measured much more easily than long distances.

Nowadays, however, long distances can be measured directly just as easily as angles by means of electronic instruments, so that triangulation, using angular measurements, is often replaced or supplemented by *trilateration*, using distance measurements. The computation of triangulations and trilaterations on the ellipsoid is easy. It is, therefore, convenient to reduce the measured angles, baselines, and long distances to the ellipsoid, in much the same way as the astronomical coordinates are treated. Then the ellipsoidal coordinates φ, λ obtained (1) by reducing the astronomical coordinates and (2) by computing triangulations or trilaterations on the ellipsoid can be compared; they should be identical for the same point.

Today, of course, GPS is the best method for determining φ, λ, and h directly.

5.3 The Global Positioning System

The following sections on the Global Positioning System (GPS) are extracted from Hofmann-Wellenhof et al. (2003: Sect. 9.3) which in return is based on

Hofmann-Wellenhof et al. (2001: Chap. 2). For details supplementing the compact description here, the reader is referred to these books.

5.3.1 Basic concept

GPS is the responsibility of the Joint Program Office (JPO), a component of the Space and Missile Center at El Segundo, California. In 1973, the JPO was directed by the U.S. Department of Defense (DOD) to establish, develop, test, acquire, and deploy a spaceborne positioning system. The present navigation system with timing and ranging is the result of this initial directive. GPS was conceived as a ranging system from known positions of satellites in space to unknown positions on land, at sea, in air, and in space. The original objectives of GPS were the instantaneous determination of position and velocity on a continuous basis, and the precise coordination of time (i.e., time transfer).

Based on code or carrier phase measurements, GPS uses pseudoranges derived from the broadcast satellite signal.

Using the code measurements, the pseudorange is derived from measuring the travel time of the coded signal and multiplying it by its velocity. Since the clocks of the receiver and the satellite are never perfectly synchronized, a clock error must be taken into account. Consequently, each equation of this type comprises four unknowns: the three point coordinates contained in the true range and the clock error. Thus, four satellites are necessary to solve for the four unknowns. Indeed, the GPS concept assumes that – without obstruction – four or more satellites are in view at any location on or near the earth 24 hours a day.

Using carrier phase measurements, ambiguities must be taken into account as additional unknowns. For more details see Hofmann-Wellenhof et al. (2001: Sect. 6.1.2).

5.3.2 System architecture

Space segment

Constellation
The GPS satellites have nearly circular orbits with an altitude of about 20200 km above the earth, i.e., they are mean earth orbit (MEO) satellites, yielding a period of nominally 12 sidereal hours. The nominal constellation consists of 24 operational satellites deployed in six evenly spaced planes (A to F) with an inclination of 55° against the equator and with four satellites per plane. Furthermore, active spare satellites for replenishment may

be operational. See http://tycho.usno.navy.mil/gpscurr.html for the current status.

With the nominal constellation, the space segment provides global coverage with four to eight simultaneously observable satellites above 15° elevation angle at any time of day. If the elevation mask is reduced to 10°, occasionally up to 10 satellites will be visible; and if the elevation mask is further reduced to 5°, occasionally 12 satellites will be visible.

Satellites categories

Essentially, the GPS satellites provide a platform for radio transceivers, atomic clocks, computers, and various ancillary equipment. The electronic equipment of each satellite allows the user to measure a pseudorange to the satellite, and each satellite broadcasts a message which allows the user to determine the spatial position of the satellite for arbitrary instants. The auxiliary equipment of each satellite, among others, consists of solar panels for power supply and a propulsion system for orbit and stability control.

There are several classes or types of GPS satellites. These are the Block I, Block II, Block IIA, Block IIR, Block IIR-M, and the future Block IIF and Block III satellites. An up-to-date description is difficult because new notations are introduced in a rather arbitrary way; an example is the recently introduced notation Block IIR-M.

Eleven Block I satellites were launched in the period between 1978 to 1985. Today, none of them is in operation anymore.

The essential difference between Block I and Block II satellites is related to U.S. national security. Block I satellite signals were fully available to civilian users. Starting with Block II, satellite signals may be restricted for civilian use. The Block II satellites are equipped with mutual communication capability. Some of them carry retroreflectors and can be tracked by laser ranging.

The Block IIR satellites ("R" denotes replenishment or replacement) have a design life of 10 years. They are equipped with improved facilities for communication and intersatellite tracking. Block IIR-M satellites incorporate two new military signals and a second civil signal. The first Block IIR-M was launched on September 25, 2005.

Currently (April 2006), the first launch of a Block IIF satellite ("F" denotes follow on) is scheduled for 2008 (instead of the previously projected dates mid of 2006 and 2007). These satellites will broadcast a third civil signal on L5 (see Sect 5.3.5).

Presently, the DOD undertakes studies for the next generation of GPS satellites, called Block III satellites. Preliminary dates (likely to change) are 2011/12 for first launches and on-orbit tests (Civil GPS Service Interface

Committee 2002). These satellites will be characterized by an assured and improved level of integrity without the need of augmentation.

Satellite signal
The key to the accuracy of the system is the fact that all signal components are precisely controlled by atomic clocks. These highly accurate frequency standards of GPS satellites produce the fundamental frequency of 10.23 MHz. Coherently derived from this frequency are (presently) two signals in the L-band, the L1 and the L2 carrier waves generated by multiplying the fundamental frequency by 154 and 120, respectively, yielding

$$\mathrm{L1} = 1575.42 \text{ MHz}\,,$$
$$\mathrm{L2} = 1227.60 \text{ MHz}\,.$$

These dual frequencies are essential for eliminating the major source of error, i.e., the ionospheric refraction.

The pseudoranges that are derived from measured travel times of the signal from each satellite to the receiver use two pseudorandom noise (PRN) codes that are modulated onto the two carriers.

The C/A-code (coarse/acquisition-code) is available for civilian use. Each C/A-code is a unique sequence of 1023 bits, called chips, which is repeated each millisecond. The duration of each C/A-code chip is about 1 μs. Equivalently, the chip length – denoted also as wavelength or chip width (Misra and Enge 2001: Sect. 2.3.1) – is about 300 m. The C/A-code is presently modulated upon L1 only and is purposely omitted from L2. This omission allows the JPO to control the information broadcast by the satellite and, thus, denies full system accuracy to nonmilitary users.

The P-code (precision-code) has been reserved for U.S. military and other authorized users. This is achieved by using the W-code to encrypt the P-code to the Y-code (anti-spoofing). The P-code has an effective chip length of about 30 m. The P-code is modulated on both carriers L1 and L2.

In addition to the PRN codes, a data message is modulated onto the carriers consisting of status information, satellite clock bias, and satellite ephemerides. The orbit data are given as Kepler-like elements and are denoted as broadcast ephemerides. The full set of elements is given in, e.g., Montenbruck and Gill (2001: Sect. A.2.2). It is worth noting that the present signal structure will be improved in the near future (see Sect. 5.3.5).

Control segment
The operational control system (OCS) consists of a master control station, monitor stations, and ground control stations. The main tasks of the OCS

are tracking of the satellites for the orbit and clock determination and prediction, time synchronization of the satellites, and upload of the data message to the satellites.

Master control station
The master control station is located at the Consolidated Space Operations Center (CSOC) at Shriver Air Force Base, Colorado Springs, Colorado. CSOC collects the tracking data from the monitor stations and calculates the satellite orbit and clock parameters by a Kalman estimator. These results are then passed to one of the three ground control stations for eventual upload to the satellites. The satellite control and system operation is also the responsibility of the master control station.

Monitor stations
There are five monitor stations located at Hawaii, Colorado Springs, Ascension Island in the South Atlantic Ocean, Diego Garcia in the Indian Ocean, and Kwajalein in the North Pacific Ocean. Each of these stations is equipped with a precise atomic time standard and receivers which continuously measure pseudoranges to all satellites in view. Pseudoranges are measured every 1.5 seconds and, using ionospheric and meteorological data, they are smoothed to produce 15-minute interval data which are transmitted to the master control station.

Ground control stations
These stations collocated with the monitor stations at Ascension, Diego Garcia, and Kwajalein are the communication links to the satellites and mainly consist of the ground antennas. The satellite ephemerides and clock information, calculated at the master control station and received via communication links, are uploaded to each GPS satellite via S-band radio links.

User segment
The diversity of the military and civilian users is matched by the type of receivers available today.

On the basis of the type of observables (i.e., code pseudoranges or phase pseudoranges) and of the availability of codes (i.e., C/A-code, P-code, or Y-code), GPS receivers can be classified. For the majority of navigation applications, C/A-code pseudorange receivers will suffice. With this type of receiver, only code pseudoranges using the C/A-code on L1 are measured. Typical devices output the three-dimensional position either in latitude, longitude, and height or in some map projection systems, e.g., universal transverse Mercator (UTM) coordinates and height.

5.3.3 Satellite signal and observables

Components of the signal

The official description of the GPS signal is given in the GPS Interface Control Document ICD-GPS-200, available at www.navcen.uscg.gov. Details may also be found in Spilker (1996).

The (current) components of the signal are summarized in Table 5.1. Note that the nominal fundamental frequency f_0 is intentionally reduced by about 0.005 Hz to compensate for relativistic effects.

The navigation message essentially contains information about the satellite health status, the satellite clock, the orbit, and various correction data.

The parameters in the block of orbit information are the reference epoch, six parameters to describe a Kepler ellipse at the reference epoch, three secular correction terms and six periodic correction terms.

Observables

In concept, the GPS observables are ranges which are deduced from measured time or phase differences based on a comparison between received signals and receiver-generated signals. As mentioned earlier, the ranges are biased by satellite and receiver clock errors and, consequently, they are denoted as pseudoranges. Essentially, pseudoranges differ from distances by an unknown additive constant.

Apart from the satellite and the receiver clock bias, further error sources can be classified into three groups, i.e., satellite-related errors (e.g., orbital errors), signal propagation medium-related errors (e.g., ionospheric and tropospheric refraction), and receiver-related errors (e.g., antenna phase center variation, multipath), but are omitted in the subsequent simplified models. Extended models are given in Hofmann-Wellenhof et al. (2003: Sect. 10.2.2).

Table 5.1. Components of the satellite signal

Component	Frequency or code chipping rate [MHz]		Wavelength
Fundamental frequency	f_0	$= 10.23$	
Carrier L1	$154\, f_0$	$= 1575.42$	19.0 cm
Carrier L2	$120\, f_0$	$= 1227.60$	24.4 cm
P-code	f_0	$= 10.23$	
C/A-code	$f_0/10$	$= 1.023$	
Navigation message	$f_0/204\,600$	$= 50 \cdot 10^{-6}$	

Code pseudoranges

The measured time difference Δt is affected by the satellite clock error δ_S and the receiver clock error δ. The error δ_S of the satellite clock can be modeled by a polynomial with the coefficients being transmitted in the navigation message. Assuming the δ_S correction is applied, the time interval Δt multiplied by the speed of light c yields the code pseudorange R and, hence,

$$R = c\,\Delta t\,. \tag{5–2}$$

Assuming a common time reference for satellite and receiver, e.g., GPS time, the term Δt may be decomposed into the run time $\Delta t(\mathrm{GPS})$ and the receiver clock errors δ leading to

$$R = c\,\Delta t(\mathrm{GPS}) + c\,\delta = \varrho + c\,\delta\,, \tag{5–3}$$

where ϱ is the geometric range between the satellite and the receiver. The receiver module responsible for code pseudorange measurements is denoted as delay lock loop (DLL). Details on the DLL functionality are given in Misra and Enge (2001: Sect. 9.5).

Phase pseudoranges

Assuming again that the satellite clock error correction is applied, the phase pseudorange Φ is modeled by

$$\lambda\,\Phi = \varrho + c\,\delta + \lambda\,N\,, \tag{5–4}$$

where the carrier wavelength λ has been introduced. The range ϱ represents the distance between the satellite at emission epoch t and the receiver at reception epoch $t + \Delta t$. Phase measurements are ambiguous, since the initial integer number N of cycles between satellite and receiver is unknown. As long as the tracking of a satellite is not interrupted, the ambiguity remains constant within the tracking loop of the receiver. The responsible receiver hardware is denoted as phase lock loop (PLL). Compared to (5–3), the phase pseudorange differs from the code pseudorange only by the phase ambiguity term $\lambda\,N$. Dividing the above equation by λ scales the phase to cycles.

As mentioned previously, the majority of navigation applications does not need carrier phase measurements. Only for increased accuracy requirements (e.g., relative positioning; see below), phase measurements become relevant.

Doppler data

Some of the first solution models proposed for GPS were to use the Doppler observable. Considering Eq. (5–4), the equation for the observed Doppler

shift scaled to range rate is given by

$$D = \lambda \dot{\Phi} = \dot{\varrho} + c\,\dot{\delta}\,, \tag{5–5}$$

where the derivatives with respect to time are indicated by a dot. The raw Doppler shift is less accurate than integrated Doppler.

The Doppler shift is measured in the carrier tracking loop of a GPS receiver (Misra and Enge 2001: Sect. 9.6). Assuming a known satellite velocity, the Doppler shift can be used to estimate the velocity of the user.

5.3.4 System capabilities and accuracies

Two operational capabilities are distinguished: firstly, the initial operational capability (IOC) and, secondly, the full operational capability (FOC).

IOC was attained in July 1993, when 24 (Block I/II/IIA) GPS satellites were operating and were available for navigation. Officially, IOC was declared by the DOD on December 8, 1993.

FOC was achieved when 24 Block II/IIA satellites were operational in their assigned orbits and the constellation was tested for operational military performance. Even though 24 Block II and Block IIA satellites were available since March 1994, FOC was not declared before July 17, 1995 which indicates an extensive testing phase.

The selection of the GPS observation technique depends upon the particular requirements of the project; especially the desired accuracy plays a dominant role.

Point positioning

When using a single receiver, usually point positioning with code pseudoranges is performed. The concept of point positioning is simple (Fig. 5.2). Without clock errors, trilateration in space (i.e., using three ranges) solves the task to determine the point coordinates. Using pseudoranges, four observations are necessary to account for the three coordinate components and the receiver clock error. For point positioning, GPS provides two levels of service: the standard positioning service (SPS) with access for civilian users and the precise positioning service (PPS) with access for authorized users.

SPS performance standards are based on signal-in-space performance. Contributions of ionosphere, troposphere, receiver, multipath, topography, or interference are not included. Furthermore, SPS is provided on the L1 signal only; the L2 signal is not part of the SPS (Department of Defense 2001). The global average positioning domain accuracy amounts to 13 m horizontal error (95% probability level) and 22 m vertical error (95% probability level).

The PPS has access to both codes and provides accuracies down to the meter level.

Differential GPS

Selective availability (SA), the deliberate degradation of the point positioning accuracy by "dithering" (i.e., distorting on purpose) the satellite clock (called δ-process) and manipulating the ephemerides (called ε-process), has led to the development of differential GPS (DGPS). Only the basic idea is explained here.

DGPS is based on the use of two (or more) receivers, where one (stationary) reference or base receiver is located at a known point and the position of the (mostly moving) remote receiver is to be determined. Using code pseudoranges, at least four common satellites must be tracked simultaneously at both sites. The known position of the reference receiver is used to calculate corrections to the observed pseudoranges. These corrections are then transmitted via telemetry (i.e., controlled radio link) to the roving receiver and allow the computation of the rover position with far more accuracy than for the single-point positioning mode.

Using DGPS based on C/A-code pseudoranges, real-time accuracies at the 1–5 m level can be routinely achieved. Phase-smoothed code ranges yield the submeter level (Lachapelle et al. 1992). Even higher accuracies can be reached by the use of carrier phases (precise DGPS). For ranges up to some 20 km, accuracies at the subdecimeter level can be obtained in real time (DeLoach and Remondi 1991). To achieve this accuracy, the ambiguities must be resolved "on the fly" and, therefore, (generally) dual-frequency receivers are required. Furthermore, five satellites per epoch are required.

After the deactivation of SA in May 2000, DGPS must be seen from a different viewpoint. The increased point positioning accuracy achieved with a single receiver may suffice for some kinds of applications.

Relative positioning

At present, highest accuracies are achieved in the relative-positioning mode with observed carrier phases. Relative positioning is associated with baselines, i.e., the three-dimensional vector between a known reference station and the location to be determined. Processing a baseline requires that the phases are simultaneously observed at both baseline endpoints (Fig. 5.1). Originally, relative positioning was only possible by postprocessing data. Today, (near) real-time data transfer over short baselines is routinely possible, which enables real-time computation of baseline vectors and has led to the real-time kinematic (RTK) technique.

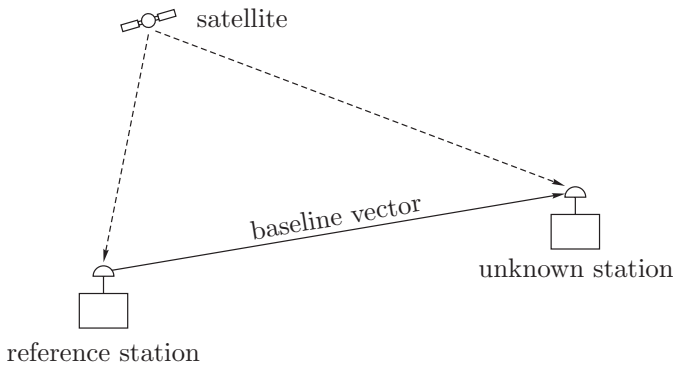

Fig. 5.1. Concept of relative positioning

Static relative positioning

The reference station and the unknown station are static, i.e., no motion occurs between the two points of the baseline. When highest accuracy is an issue, then this is the preferred method. Fully depending on the application and on the length of the baseline, the observation time may amount from several tens of minutes to many hours. Referring to navigation, where usually motion is involved, static relative positioning is of minor importance. The reader is referred to Hofmann-Wellenhof et al. (2001: Sect. 7.1.2) for details.

Kinematic relative positioning

The kinematic method is very productive because the greatest number of points can be determined in the least time.

The drawback is that after initialization a continuous lock on at least four satellites must be maintained.

The semikinematic or stop-and-go technique is characterized by alternatively stopping and moving one receiver to determine the positions of fixed points along the trajectory. The most important feature of this method is the increase in accuracy when several measurement epochs at the stop locations are accumulated and averaged. This technique is often referred to simply as kinematic method. Relative positional accuracies at the centimeter level can be achieved for baselines up to some 20 km.

The kinematic technique requires the resolution of the phase ambiguities by initialization which can be performed by static or kinematic techniques. Currently available commercial software (for dual-frequency receivers) only requires 1–2 minutes of observation for baselines up to 20 km to resolve the ambiguities kinematically ("on the fly").

5.3.5 GPS modernization concept

In January 1999, the USA announced the GPS modernization concept, a $400 million initiative. The key feature is the implementation of a new signal structure in future satellites.

Future GPS satellites

The Block IIR satellites increase their presence in the GPS constellation. A new effort will bring modernized functionality to IIR satellites. These modernized satellites, denoted as IIR-M (replenishment-modernization), will provide new services to military and civilian users. New signals and increased L-band power will significantly improve the navigation performance (Marquis 2001).

The Block IIF and the Block III satellites are the next generations.

These next generations of satellites will have many improvements over the present satellites. It is planned to include the capability to transmit data between satellites to make the system more independent. The autonomous navigation (auto-nav) capability via intersatellite cross-link ranging will allow the satellites to essentially position themselves without extensive ground tracking. In summary, the future satellites will have the following mainly military advantages:

- Navigation accuracy will be maintained for six months without ground support and control.
- Uplink jamming concerns will be minimized.
- One upload per spacecraft per month instead of one or even more per day will be performed.
- Need for overseas stations to support navigation uploads will be reduced.
- Improved navigation accuracy will be achieved.

New signal structure

Referring to codes, presently civil users have unlimited access only to the C/A-code on the carrier L1. The modernization will provide new signals: implementing military codes (M-codes) on L1 and on L2 and a civilian code on L2 (abbreviated as L2c). The M-code will provide the authorized users with more signal security, improved acquisition options, and more jamming resistance. The new civilian L2c signal will provide nonauthorized users dual-frequency operation to perform ionospheric error correction. In addition to these codes, a new L5 frequency will be provided for civilian users to enhance aviation applications. The notation L5 is chosen because, actually,

the satellites transmit additional signals at frequencies referred to as L3 and L4. These signals are classified and for military purposes only (Misra and Enge 2001: Sect. 2.3).

According to the modernization initiative released in 1999, the Interagency GPS Executive Board concept will be realized with the following specifications. Future GPS signals will be transmitted by three carriers where L1 and L2 remain unchanged, and the new carrier L5 is specified as

$$\mathrm{L5} = 115 f_0 = 1176.45 \text{ MHz}\,,$$

where $f_0 = 10.23\,\mathrm{MHz}$ denotes the basic GPS frequency. The carrier L5, placed in a protected aeronautical radio navigation service band, was recently allocated by the World Radio Conference organized regularly by the International Telecommunication Union (Vorhies 2000).

Note that both new civil GPS signals will have two codes. L5 will not share with military signals and use two equal-length codes in phase quadrature, each clocked at 10.23 MHz. L2 is shared between civil and military signals. The new L2c signal provides two codes by time multiplexing. The two codes are of different length (Fontana et al. 2001). The existing military Y-code will be replaced by new (split) M-codes.

The linear carrier phase combination of L2 with L5 results in a signal with a wavelength of about 5.9 m. Long wavelengths facilitate ambiguity resolution. By contrast, the linear combination of L1 with L5 will be used as ionosphere-free combination because large frequency differences are advantageous for calculating ionospheric corrections. The common processing of phase data from all three carriers will be performed in the three-carrier ambiguity resolution approach (Vollath et al. 1999).

A perspective for the implementation is given in the 2001 Federal Radionavigation Plan: IOC (18 satellites in orbit with the new L2c signal and M-code capability) is planned for 2008 and FOC (24 satellites in orbit) is planned for 2010. At least one satellite is planned to be operational with the new L5 capability no later than 2005, with IOC planned for 2012 and FOC planned for 2014.

5.4 From GPS to coordinates

So far, we have got an introductory GPS overview. Now we are interested in applying elementary GPS approaches to demonstrate how coordinates are obtained. Two examples, as simple as possible, are selected: point positioning and relative positioning.

5.4.1 Point positioning with code pseudoranges

The situation is shown in Fig. 5.2. The coordinates of A are to be determined by using GPS. As we know from Sect. 5.3.4, four pseudoranges to different satellites are necessary to determine the three coordinate components of A and the receiver clock error. Generalizing (5–3), we obtain

$$R_A^j(t) = \varrho_A^j(t) + c\,\delta_A(t)\,. \tag{5–6}$$

This is the code pseudorange at an epoch t, where $R_A^j(t)$ is the measured code pseudorange between the observing site A (as indicated in Fig. 5.2) and the satellite j, and $\varrho_A^j(t)$ is the geometric distance between the satellite and the observing point, and c is the speed of light. The last item is the receiver clock error $\delta_A(t)$. Note that we assume the simplest possible model, thus, we do not consider ionospheric and tropospheric influences, other biases and errors.

Examining Eq. (5–6), the desired point coordinates to be determined are implicitly comprised in the distance $\varrho_A^j(t)$, which can explicitly be written as

$$\varrho_A^j(t) = \sqrt{(X^j(t) - X_A)^2 + (Y^j(t) - Y_A)^2 + (Z^j(t) - Z_A)^2}\,, \tag{5–7}$$

where the WGS 84 (World Geodetic System 1984, see Sect. 2.11) coordinates $X^j(t)$, $Y^j(t)$, $Z^j(t)$ are the components of the geocentric position vector of the satellite at epoch t, and X_A, Y_A, Z_A are the three unknown WGS 84 coordinates of the observing site, which might be denoted $(X_A,\, Y_A,\, Z_A)_{\text{WGS 84}}$ or, which means the same, $(X_A,\, Y_A,\, Z_A)_{\text{GPS}}$.

How many unknowns are involved? Note that the satellite coordinates $X^j(t)$, $Y^j(t)$, $Z^j(t)$ may always be assumed known (more precisely, are cal-

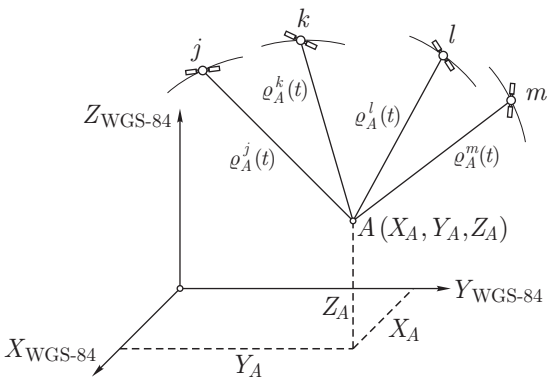

Fig. 5.2. Point positioning

culable) from the information broadcast by the satellite. Therefore, there remain the three unknown station coordinates X_A, Y_A, Z_A and the unknown receiver clock error $\delta_A(t)$. In other terms, at least four satellites are required to set up four equations of type (5–6). Denoting the satellites by j, k, l, m, the corresponding system of equations

$$\begin{aligned} R_A^j(t) &= \varrho_A^j(t) + c\,\delta_A(t)\,, \\ R_A^k(t) &= \varrho_A^k(t) + c\,\delta_A(t)\,, \\ R_A^l(t) &= \varrho_A^l(t) + c\,\delta_A(t)\,, \\ R_A^m(t) &= \varrho_A^m(t) + c\,\delta_A(t) \end{aligned} \tag{5–8}$$

is obtained or, by substituting (5–7) accordingly,

$$\begin{aligned} R_A^j(t) &= \sqrt{(X^j(t) - X_A)^2 + (Y^j(t) - Y_A)^2 + (Z^j(t) - Z_A)^2} + c\,\delta_A(t)\,, \\ R_A^k(t) &= \sqrt{(X^k(t) - X_A)^2 + (Y^k(t) - Y_A)^2 + (Z^k(t) - Z_A)^2} + c\,\delta_A(t)\,, \\ R_A^l(t) &= \sqrt{(X^l(t) - X_A)^2 + (Y^l(t) - Y_A)^2 + (Z^l(t) - Z_A)^2} + c\,\delta_A(t)\,, \\ R_A^m(t) &= \sqrt{(X^m(t) - X_A)^2 + (Y^m(t) - Y_A)^2 + (Z^m(t) - Z_A)^2} + c\,\delta_A(t) \end{aligned} \tag{5–9}$$

results. This system of equations comprises only the previously mentioned four unknowns X_A, Y_A, Z_A and the unknown receiver clock error $\delta_A(t)$ and may, thus, be solved. We do not consider linearization, possible redundant measurements, etc. We just intended to demonstrate the principle. The clock error is a by-product, but the desired result obtained from (5–9) are the GPS coordinates X_A, Y_A, Z_A; this means, the resulting coordinates are obatined in the WGS 84.

As described in Sect. 5.3.4, the accuracy of the point positioning method based on code ranges may be expected to amount some 10 m (nominally). A much higher accuracy is achieved by relative positioning treated in the next section.

5.4.2 Relative positioning with phase pseudoranges

The objective of relative positioning is to determine the coordinates of an unknown point with respect to a known point. In other words, relative positioning aims at the determination of the vector between the two points which is often called the baseline vector or simply baseline (Fig. 5.3). Let now A denote the known reference point, B the unknown point, and $\mathbf{b}_{AB}$ the baseline vector. Introducing the corresponding position vectors $\mathbf{X}_A$, $\mathbf{X}_B$,

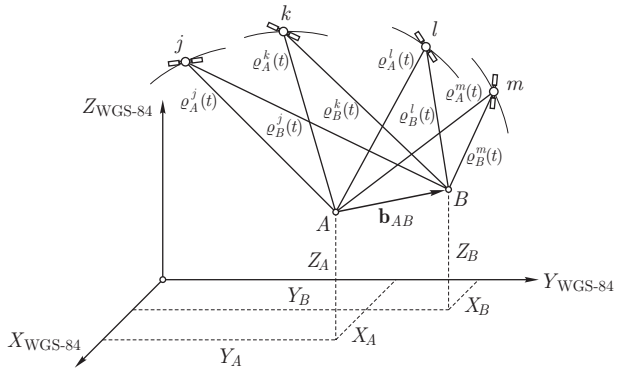

Fig. 5.3. Relative positioning

the relation

$$\mathbf{X}_B = \mathbf{X}_A + \mathbf{b}_{AB} \tag{5–10}$$

may be formulated, and the components of the baseline vector $\mathbf{b}_{AB}$ are

$$\mathbf{b}_{AB} = \begin{bmatrix} X_B - X_A \\ Y_B - Y_A \\ Z_B - Z_A \end{bmatrix} = \begin{bmatrix} \Delta X_{AB} \\ \Delta Y_{AB} \\ \Delta Z_{AB} \end{bmatrix} . \tag{5–11}$$

The coordinates of the reference point must be given in the WGS 84 and are usually approximated by a code pseudorange solution. Relative positioning can be performed with code pseudoranges (cf. Eq. (5–3)) or with phase pseudoranges (cf. Eq. (5–4)). Subsequently, only phase pseudoranges are explicitly considered. We repeat (5–4),

$$\lambda\,\Phi = \varrho + c\,\delta + \lambda\,N\,, \tag{5–12}$$

where we have already explained the wavelength λ, the phase Φ, the distance ϱ (which is the same as for the code pseudorange model), the speed of light c, the receiver clock error δ, and the ambiguity N in Sect. 5.3.3.

Introducing f, the frequency of the corresponding satellite signal, and taking into account the relation $f = c/\lambda$, we may divide (5–12) by λ obtaining

$$\Phi = \frac{1}{\lambda}\varrho + f\,\delta + N\,. \tag{5–13}$$

This may be generalized to

$$\Phi_i^j(t) = \frac{1}{\lambda}\varrho_i^j(t) + f\,\delta_i(t) + N_i^j\,, \tag{5–14}$$

where $\Phi_i^j(t)$ is the measured carrier phase expressed in cycles referred to station i and satellite j at epoch t. The time-independent phase ambiguity N_i^j is an integer number and, therefore, often called integer ambiguity or integer unknown or simply ambiguity.

Relative positioning requires simultaneous observations at both the reference and the unknown point. This means that the observation time tags for the two points must be the same. Assuming such observations (5–14) at the two points A and B to satellite j and another satellite k simultaneously at epoch t, the following measurement equations may be set up:

$$\begin{aligned}
\Phi_A^j(t) &= \frac{1}{\lambda}\varrho_A^j(t) + f\,\delta_A(t) + N_A^j\,,\\
\Phi_A^k(t) &= \frac{1}{\lambda}\varrho_A^k(t) + f\,\delta_A(t) + N_A^k\,,\\
\Phi_B^j(t) &= \frac{1}{\lambda}\varrho_B^j(t) + f\,\delta_B(t) + N_B^j\,,\\
\Phi_B^k(t) &= \frac{1}{\lambda}\varrho_B^k(t) + f\,\delta_B(t) + N_B^k\,.
\end{aligned} \tag{5–15}$$

Introducing the short-hand notations

$$\begin{aligned}
\Phi_{AB}^{jk}(t) &= \Phi_B^k(t) - \Phi_B^j(t) - \Phi_A^k(t) + \Phi_A^j(t)\,,\\
\varrho_{AB}^{jk}(t) &= \varrho_B^k(t) - \varrho_B^j(t) - \varrho_A^k(t) + \varrho_A^j(t)\,,\\
N_{AB}^{jk} &= N_B^k - N_B^j - N_A^k + N_A^j\,,
\end{aligned} \tag{5–16}$$

we form the double-difference model which is defined as

$$\Phi_{AB}^{jk}(t) = \frac{1}{\lambda}\,\varrho_{AB}^{jk}(t) + N_{AB}^{jk}\,. \tag{5–17}$$

Note that the receiver clock biases have canceled; this is the reason why double-differences are preferably used. This cancellation resulted from the assumptions of simultaneous observations and equal frequencies of the satellite signals (which is justified for GPS).

Assuming A as reference station with known coordinates, the remaining unknowns of the double-difference model are the desired coordinates X_B, Y_B, Z_B – which are comprised in $\varrho_B^j(t)$ and $\varrho_B^k(t)$ – and the ambiguities. To solve for these unknowns, we need more satellites (to set up additional double-differences) and also more epochs.

We do not consider linearization, possible redundant measurements, etc. We just intended to demonstrate the principle. The desired result obtained from (5–17) is the baseline vector $\mathbf{b}_{AB}$ with the components ΔX_{AB}, ΔY_{AB}, ΔZ_{AB} or, finally, the GPS coordinates X_B, Y_B, Z_B derived from (5–10) via

the known station A to achieve the high accuracy. Note that the resulting coordinates are obtained in the WGS 84.

This concludes the short introduction how the user of GPS gets WGS 84 coordinates, i.e., geocentric rectangular coordinates X, Y, Z or, computed from them, ellipsoidal coordinates φ, λ, h; see Sect. 5.6.1.

5.5 Projection onto the ellipsoid

Let us establish the position of a point P by means of the natural coordinates Φ, Λ, H. Then we may project it onto the geoid along the (slightly curved) plumb line. The orthometric height is the distance between P and its projection P_0 onto the geoid, measured along the plumb line (Fig. 5.4). Although this mode of projection is entirely natural, the geoid is not suited for performing computations on it directly; the point P_0 is, therefore, projected onto the reference ellipsoid by means of the straight ellipsoidal normal, thus getting a point Q_0 on the ellipsoid. In this way, the earth's surface point P and the corresponding point Q_0 on the ellipsoid are connected by a double projection, that is, by two projections which are performed one after the other and which are quite analogous, the orthometric height $H = PP_0$ corresponding to the geoidal undulation $N = P_0Q_0$. This double projection is called *Pizzetti's projection*.

It is much simpler to project the point P from the physical surface of the earth directly onto the ellipsoid through the straight ellipsoidal normal, thus obtaining a point Q. The distance $PQ = h$ is the ellipsoidal height, i.e., the height above the ellipsoid. The earth's surface point P is then determined by the ellipsoidal height h and the ellipsoidal coordinates φ, λ of Q on the ellipsoid so that the *ellipsoidal coordinates* φ, λ, h take the place of the *natural coordinates* Φ, Λ, H. This is called *Helmert's projection*.

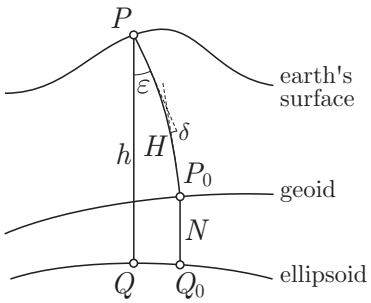

Fig. 5.4. The projection of Helmert and of Pizzetti

The practical difference between Pizzetti's and Helmert's projection is small. The ellipsoidal height h is equal to $H + N$ within a fraction of a millimeter. The ellipsoidal coordinates φ and λ, with respect to the two projections, are related by the equations

$$\begin{aligned} \varphi_{\text{Helmert}} &= \varphi_{\text{Pizzetti}} + \frac{H}{R}\,\xi\,, \\ \lambda_{\text{Helmert}} &= \lambda_{\text{Pizzetti}} + \frac{H}{R}\,\eta\,\sec\varphi\,, \end{aligned} \tag{5–18}$$

which can be read from Fig. 5.4, since $QQ_0 \doteq H\,\varepsilon$; $R = 6371$ km is the mean radius of the earth. Even if $\varepsilon = 1$ arc minute and $H = 1000\,\text{m}$, the distance QQ_0 is only about 30 cm and the ellipsoidal coordinates differ by less than $0.01''$, which is below the accuracy of astronomical observations. For most purposes, we may, therefore, neglect the difference between the two projections.

Pizzetti's projection is better adapted to the geoid, because there is an exact correspondence between a geoidal point P_0 and an ellipsoidal point Q_0. Helmert's projection has overwhelming practical advantages, notably the straightforward conversion of the ellipsoidal coordinates φ, λ, h into rectangular coordinates x, y, z; it is also simpler in other respects. The decisive advantage of Helmert's projection is its direct relation to GPS. It is, therefore, exclusively used now in practice.

5.6 Coordinate transformations

5.6.1 Ellipsoidal and rectangular coordinates

We now derive the relation between the ellipsoidal coordinates φ, λ, h and the corresponding rectangular coordinates x, y, z.

The equation of the reference ellipsoid in rectangular coordinates is

$$\frac{x^2 + y^2}{a^2} + \frac{z^2}{b^2} = 1\,. \tag{5–19}$$

The representation of this ellipsoid in terms of ellipsoidal coordinates is given by

$$\begin{aligned} x &= N\,\cos\varphi\,\cos\lambda\,, \\ y &= N\,\cos\varphi\,\sin\lambda\,, \\ z &= \frac{b^2}{a^2}\,N\,\sin\varphi\,, \end{aligned} \tag{5–20}$$

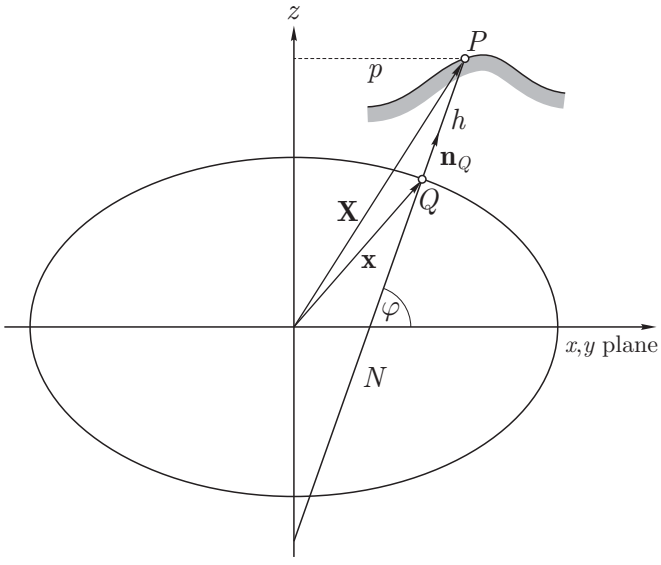

Fig. 5.5. Ellipsoidal and rectangular coordinates

where N is the normal radius of curvature (2–149):

$$N = \frac{a^2}{\sqrt{a^2 \cos^2\varphi + b^2 \sin^2\varphi}}\,. \tag{5–21}$$

These equations are known from ellipsoidal geometry; it may also be verified by direct substitution that a point with xyz-coordinates (5–20) satisfies the equation of the ellipsoid (5–19) and so lies on the ellipsoid. The components of the unit normal vector $\mathbf{n}$ are

$$\mathbf{n} = \left[\cos\varphi\,\cos\lambda\,,\ \ \cos\varphi\,\sin\lambda\,,\ \ \sin\varphi\right], \tag{5–22}$$

because φ is the angle between the ellipsoidal normal and the xy-plane, which is the equatorial plane (Fig. 5.5). Now let the coordinates of a point P outside the ellipsoid form the vector

$$\mathbf{X} = [X\,,\ Y\,,\ Z\,]; \tag{5–23}$$

similarly we have, for the coordinates of the point Q on the ellipsoid,

$$\mathbf{x} = [x\,,\ \ y\,,\ \ z\,]\,. \tag{5–24}$$

From Fig. 5.5, we read

$$\mathbf{X} = \mathbf{x} + h\,\mathbf{n}\,, \tag{5–25}$$

that is

$$\begin{aligned} X &= x + h\,\cos\varphi\,\cos\lambda\,, \\ Y &= y + h\,\cos\varphi\,\sin\lambda\,, \\ Z &= z + h\,\sin\varphi\,. \end{aligned} \tag{5–26}$$

By (5–20), this becomes

$$\begin{aligned} X &= (N+h)\cos\varphi\,\cos\lambda\,, \\ Y &= (N+h)\cos\varphi\,\sin\lambda\,, \\ Z &= \left(\frac{b^2}{a^2}\,N + h\right)\sin\varphi\,. \end{aligned} \tag{5–27}$$

These equations are the basic transformation formulas between the ellipsoidal coordinates φ, λ, h and the rectangular coordinates X, Y, Z of a point outside the ellipsoid. The origin of the rectangular coordinate system is the center of the ellipsoid, and the z-axis is its axis of rotation; the x-axis has the Greenwich longitude 0° and the y-axis has the longitude 90° east of Greenwich (i.e., $\lambda = +90°$).

A possible source of confusion is that the normal radius of curvature of the ellipsoid and the geoidal undulation are both denoted by the symbol N; in (5–27), N is, of course, the normal radius of curvature. Generally, let the context decide between quantities of such different magnitude (6000 km and 60 m).

Equations (5–27) permit the computation of rectangular coordinates X, Y, Z from the ellipsoidal coordinates φ, λ, h.

The inverse procedure, the computation of φ, λ, h from given X, Y, Z, is frequently performed iteratively, although a solution in closed form exists. A possible iterative procedure is as follows.

Denoting $\sqrt{X^2 + Y^2}$ by p, we get from the first two equations of (5–27) or from Fig. 5.5

$$p = \sqrt{X^2 + Y^2} = (N+h)\cos\varphi\,, \tag{5–28}$$

so that

$$h = \frac{p}{\cos\varphi} - N\,. \tag{5–29}$$

The third equation of (5–27) may be transformed into

$$Z = \left(N - \frac{a^2 - b^2}{a^2}\,N + h\right)\sin\varphi = (N + h - e^2 N)\sin\varphi\,, \tag{5–30}$$

where $e^2 = (a^2 - b^2)/a^2$. Dividing this equation by the above expression for p, we find

$$\frac{Z}{p} = \left(1 - e^2 \frac{N}{N+h}\right) \tan\varphi\,, \tag{5–31}$$

so that

$$\tan\varphi = \frac{Z}{p}\left(1 - e^2 \frac{N}{N+h}\right)^{-1}. \tag{5–32}$$

Given X, Y, Z, and hence p, Eqs. (5–29) and (5–32) may be solved iteratively for h and φ. As a first approximation, we set $h = 0$ in (5–32), obtaining

$$\tan\varphi_{(1)} = \frac{Z}{p}\,(1 - e^2)^{-1}\,. \tag{5–33}$$

Using $\varphi_{(1)}$, we compute an approximate value $N_{(1)}$ by means of (5–21). Then (5–29) gives $h_{(1)}$. Now, as a second approximation, we set $h = h_{(1)}$ in (5–32), obtaining

$$\tan\varphi_{(2)} = \frac{Z}{p}\left(1 - e^2 \frac{N_{(1)}}{N_{(1)} + h_{(1)}}\right)^{-1}. \tag{5–34}$$

Using $\varphi_{(2)}$, improved values for N and h are found, etc. This procedure is repeated until φ and h remain practically constant.

The result for λ is immediately obtained from the first two equations of (5–27):

$$\lambda = \arctan\frac{Y}{X}\,. \tag{5–35}$$

Many other computation methods have been devised. One example for the transformation of X, Y, Z into φ, λ, h without iteration but with an inherent approximation is

$$\begin{aligned} \varphi &= \arctan\frac{Z + e'^2\, b\, \sin^3\theta}{p - e^2\, a\, \cos^3\theta}\,, \\ \lambda &= \arctan\frac{Y}{X}\,, \\ h &= \frac{p}{\cos\varphi} - N\,, \end{aligned} \tag{5–36}$$

where

$$\theta = \arctan\frac{Z\,a}{p\,b} \tag{5–37}$$

is an auxiliary quantity and

$$e^2 = (a^2 - b^2)/a^2\,, \quad e'^2 = (a^2 - b^2)/b^2 \tag{5–38}$$

are first and second numerical eccentricity. As introduced in (5–28), $p = \sqrt{X^2 + Y^2}$. Actually, there is no reason why these formulas are less popular than the iterative procedure since there is no significant difference between the two methods. Computation methods with neither iteration nor approximation are, e.g., given by Sünkel (1977) and Zhu (1993).

5.6.2 Ellipsoidal, ellipsoidal-harmonic, and spherical coordinates

Even if we have several times pointed out the different definitions, it is very important to stress once more the need not to confuse the following coordinate triples (see Fig. 5.6):

- ellipsoidal coordinates: φ, λ, h;
- ellipsoidal-harmonic coordinates: β, λ, u, alternatively: $\vartheta_{\text{ellipsoidal-harmonic}}, \lambda, u$;
- spherical coordinates: $\bar{\varphi}, \lambda, r$, alternatively: $\vartheta_{\text{spherical}}, \lambda, r$.

The longitude λ is the same in all triples. The ellipsoidal coordinates latitude φ and longitude λ are sometimes also denoted *geodetic latitude* and *geodetic*

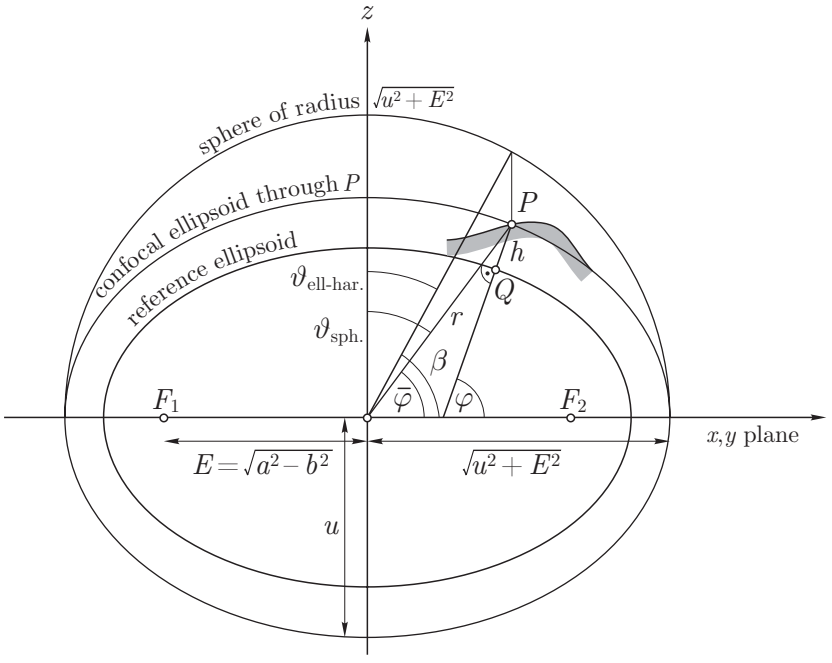

Fig. 5.6. Ellipsoidal, ellipsoidal-harmonic, and spherical coordinates

longitude. The ellipsoidal-harmonic coordinate β is the *reduced* latitude, and the spherical coordinate $\bar{\varphi}$ is the *geocentric* latitude.

The latitude φ refers to the *reference ellipsoid*. The reduced latitude β refers to the *coordinate ellipsoid* $u = \text{constant}$ (confocal ellipsoid through P in Fig. 5.6).

So far so clear. Real attention is necessary when using the coordinate ϑ, which has been introduced as complement of the spherical coordinate $\bar{\varphi}$ *and* as the complement of the ellipsoidal harmonic β as well.

Therefore, a correct but clumsy notation would be

$$\begin{aligned} \vartheta_{\text{ellipsoidal-harmonic}} &= 90^\circ - \beta\,, \\ \vartheta_{\text{spherical}} &= 90^\circ - \bar{\varphi}\,. \end{aligned} \tag{5–39}$$

Note, however, that we did not use these indications to distinguish between the spherical and the ellipsoidal-harmonic ϑ! Thus, the reader is challenged to attentively distinguish between these quantities. Wherever possible, we tried to avoid conflicts.

Some examples: we used the spherical coordinates r, ϑ, λ in Sects. 1.4, 1.11, 1.12, 1.14, 2.5, 2.6, 2.13, 2.18, etc. We used the ellipsoidal-harmonic coordinates u, ϑ, λ in Sects. 1.15, 1.16; we used the ellipsoidal-harmonic coordinates u, β, λ in Sects. 2.7, 2.8, and we used the spherical coordinates r, ϑ, λ as well as the ellipsoidal-harmonic coordinates u, β, λ in Sect. 2.9.

The following equations express the rectangular coordinates in these three systems:

$$\begin{aligned} X &= (N+h)\cos\varphi\,\cos\lambda = \sqrt{u^2+E^2}\,\cos\beta\,\cos\lambda = r\,\cos\bar{\varphi}\,\cos\lambda\,, \\ Y &= (N+h)\cos\varphi\,\sin\lambda = \sqrt{u^2+E^2}\,\cos\beta\,\sin\lambda = r\,\cos\bar{\varphi}\,\sin\lambda\,, \\ Z &= \left(\frac{b^2}{a^2}\,N+h\right)\sin\varphi \;\; = u\,\sin\beta \qquad\qquad\qquad = r\,\sin\bar{\varphi}\,. \end{aligned} \tag{5–40}$$

These relations, which follow from combining Eqs. (1–26), (1–151), and (5–27), can be used if we wish to compute u and β from h and φ or from r and $\bar{\varphi}$, etc.

5.7 Geodetic datum transformations

5.7.1 Introduction

First we define a *geodetic datum* or a *geodetic reference system*. It is defined by (1) the dimensions of the reference ellipsoid (semimajor axis a and

flattening f) and (2) its position with respect to the earth or the geoid. This relative position is most simply defined by the coordinates x_0, y_0, z_0 of the center of the reference ellipsoid with respect to the geocenter. Since the geocenter was not accessible to classical geodetic measurements before the satellite era, a *fundamental* or *initial point* P_1 on the earth surface was chosen, such as Meades Ranch for North America and Potsdam for Central Europe. It turns out that a convenient but conventional choice of the ellipsoidal coordinates $\varphi_1, \lambda_1, h_1$ of the fundamental point P_1 is equivalent to x_0, y_0, z_0 of the geocenter.

Thus, we have 5 defining parameters:

- 2 parameters a (semimajor axis) and f (flattening) as *form parameters*, and
- 3 parameters x_0, y_0, z_0 or $\varphi_1, \lambda_1, h_1$ as *position parameters*.

Later on we shall also admit a scale factor and small rotations around the three coordinate axes.

A (geodetic) datum transformation defines the relationship between a global (geocentric) and a local (in general nongeocentric) three-dimensional Cartesian coordinate system; therefore, a datum transformation transforms one coordinate system of a certain type to another coordinate system of the same type. This is one of the primary tasks when combining GPS data with terrestrial data, i.e., the transformation of geocentric WGS 84 coordinates to local terrestrial coordinates. The terrestrial system is usually based on a locally best-fitting ellipsoid, e.g., the Clarke ellipsoid or the GRS-80 ellipsoid in the U.S. and the Bessel ellipsoid in many parts of Europe. The local ellipsoid is linked to a nongeocentric Cartesian coordinate system, where the origin coincides with the center of the ellipsoid.

5.7.2 Three-dimensional transformation in general form

Consider two arbitrary sets of three-dimensional Cartesian coordinates forming the vectors $\mathbf{X}$ and $\mathbf{X}_T$ (Fig. 5.7). The 7-parameter transformation, also denoted as Helmert transformation or similarity transformation in space, between the two sets can be formulated by the relation

$$\mathbf{X}_T = \mathbf{x}_0 + \mu\,\mathbf{R}\,\mathbf{X}\,, \tag{5–41}$$

where $\mathbf{x}_0$ is the translation (or shift) vector, μ is a scale factor, and $\mathbf{R}$ is a rotation matrix.

The components of the shift vector

$$\mathbf{x}_0 = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} \tag{5–42}$$

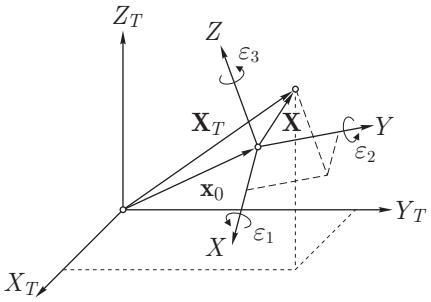

Fig. 5.7. Three-dimensional transformation

account for the coordinates of the origin of the $\mathbf{X}$ system in the $\mathbf{X}_T$ system. Note that a single scale factor is considered. More generally (but with GPS not necessary), three scale factors, one for each axis, could be used. The rotation matrix is an orthogonal matrix which is composed of three successive rotations

$$\mathbf{R} = \mathbf{R}_3\{\varepsilon_3\}\,\mathbf{R}_2\{\varepsilon_2\}\,\mathbf{R}_1\{\varepsilon_1\}\,. \tag{5–43}$$

Explicitly,

$$\mathbf{R} = \begin{bmatrix} \cos\varepsilon_2\cos\varepsilon_3 & \cos\varepsilon_1\sin\varepsilon_3 & \sin\varepsilon_1\sin\varepsilon_3 \\ & +\sin\varepsilon_1\sin\varepsilon_2\cos\varepsilon_3 & -\cos\varepsilon_1\sin\varepsilon_2\cos\varepsilon_3 \\ -\cos\varepsilon_2\sin\varepsilon_3 & \cos\varepsilon_1\cos\varepsilon_3 & \sin\varepsilon_1\cos\varepsilon_3 \\ & -\sin\varepsilon_1\sin\varepsilon_2\sin\varepsilon_3 & +\cos\varepsilon_1\sin\varepsilon_2\sin\varepsilon_3 \\ \sin\varepsilon_2 & -\sin\varepsilon_1\cos\varepsilon_2 & \cos\varepsilon_1\cos\varepsilon_2 \end{bmatrix} \tag{5–44}$$

is obtained.

In the case of known transformation parameters $\mathbf{x}_0$, μ, $\mathbf{R}$, a point from the $\mathbf{X}$ system can be transformed into the $\mathbf{X}_T$ system by (5–41).

If the transformation parameters are unknown, they can be determined with the aid of common (identical) points, also denoted as control points. This means that the coordinates of the same point are given in both systems. Since each common point (given by $\mathbf{X}_T$ and $\mathbf{X}$) yields three equations, two common points and one additional common component (e.g., height) are sufficient to solve for the seven unknown parameters. In practice, redundant common point information is used and the unknown parameters are calculated by least-squares adjustment.

Since the parameters are mixed nonlinearly in Eq. (5–41), a linearization must be performed, where approximate values $\mathbf{x}_{0\text{approx}}$, μ_{approx}, $\mathbf{R}_{\text{approx}}$ are required.

5.7.3 Three-dimensional transformation between WGS 84 and a local system

In the case of a datum transformation between WGS 84 and a local system, some simplifications will arise. Referring to the necessary approximate values, the approximation $\mu_{\text{approx}} = 1$ is appropriate and the relation

$$\mu = \mu_{\text{approx}} + \delta\mu = 1 + \delta\mu \tag{5–45}$$

is obtained. Furthermore, the rotation angles ε_i in (5–44) are small and may be treated as differential quantities. Introducing these quantities into (5–44), setting $\cos\varepsilon_i = 1$ and $\sin\varepsilon_i = \varepsilon_i$, and considering only first-order terms gives

$$\mathbf{R} = \begin{bmatrix} 1 & \varepsilon_3 & -\varepsilon_2 \\ -\varepsilon_3 & 1 & \varepsilon_1 \\ \varepsilon_2 & -\varepsilon_1 & 1 \end{bmatrix} = \mathbf{I} + \delta\mathbf{R}\,, \tag{5–46}$$

where $\mathbf{I}$ is the unit matrix and $\delta\mathbf{R}$ is a (skewsymmetric) differential rotation matrix. Thus, the approximation $\mathbf{R}_{\text{approx}} = \mathbf{I}$ is appropriate. Finally, the shift vector is split up in the form

$$\mathbf{x}_0 = \mathbf{x}_{0\text{approx}} + \delta\mathbf{x}_0\,, \tag{5–47}$$

where the approximate shift vector

$$\mathbf{x}_{0\text{approx}} = \mathbf{X}_T - \mathbf{X} \tag{5–48}$$

follows by substituting the approximations for the scale factor and the rotation matrix into Eq. (5–41).

Introducing Eqs. (5–45), (5–46), (5–47) into (5–41) and skipping details which can be found, for example, in Hofmann-Wellenhof et al. (1994: Sect. 3.3) gives the linearized model for a single point i. This model can be written in the form

$$\mathbf{X}_{T_i} - \mathbf{X}_i - \mathbf{x}_{0\text{approx}} = \mathbf{A}_i\,\delta\mathbf{p}\,, \tag{5–49}$$

where the left side of the equation is known and may formally be considered as an observation. The design matrix $\mathbf{A}_i$ and the vector $\delta\mathbf{p}$, containing the unknown parameters, are given by

$$\begin{aligned} \mathbf{A}_i &= \begin{bmatrix} 1 & 0 & 0 & X_i & 0 & -Z_i & Y_i \\ 0 & 1 & 0 & Y_i & Z_i & 0 & -X_i \\ 0 & 0 & 1 & Z_i & -Y_i & X_i & 0 \end{bmatrix}, \\ \delta\mathbf{p} &= \begin{bmatrix} \delta x_0 & \delta y_0 & \delta z_0 & \delta\mu & \varepsilon_1 & \varepsilon_2 & \varepsilon_3 \end{bmatrix}. \end{aligned} \tag{5–50}$$

Recall that Eq. (5–49) is now a system of linear equations for point i. For n common points, the design matrix A is

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_n \end{bmatrix} . \tag{5–51}$$

In detail, for three common points the design matrix is

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & X_1 & 0 & -Z_1 & Y_1 \\ 0 & 1 & 0 & Y_1 & Z_1 & 0 & -X_1 \\ 0 & 0 & 1 & Z_1 & -Y_1 & X_1 & 0 \\ 1 & 0 & 0 & X_2 & 0 & -Z_2 & Y_2 \\ 0 & 1 & 0 & Y_2 & Z_2 & 0 & -X_2 \\ 0 & 0 & 1 & Z_2 & -Y_2 & X_2 & 0 \\ 1 & 0 & 0 & X_3 & 0 & -Z_3 & Y_3 \\ 0 & 1 & 0 & Y_3 & Z_3 & 0 & -X_3 \\ 0 & 0 & 1 & Z_3 & -Y_3 & X_3 & 0 \end{bmatrix} , \tag{5–52}$$

which leads to a slightly redundant system. Least-squares adjustment yields the parameter vector $\delta\mathbf{p}$ and the adjusted values by (5–45), (5–46), (5–47). Once the seven parameters of the similarity transformation are determined, formula (5–41) can be used to transform other than the common points.

For a specific example, consider the task of transforming GPS coordinates of a network, i.e., global geocentric WGS 84 coordinates, to (three-dimensional) coordinates of a (nongeocentric) local system indicated by the subscript LS. The GPS coordinates are denoted by $(X, Y, Z)_{\mathrm{GPS}}$ and the local system coordinates are the plane coordinates $(y, x)_{\mathrm{LS}}$ and the ellipsoidal height h_{LS}. To obtain the transformation parameters, it is assumed that the coordinates of the common points in both systems are available. The solution of the task is obtained by the following algorithm:

1. Transform the plane coordinates $(y, x)_{\mathrm{LS}}$ of the common points into the ellipsoidal surface coordinates $(\varphi, \lambda)_{\mathrm{LS}}$ by using the appropriate mapping formulas.

2. Transform the ellipsoidal coordinates $(\varphi, \lambda, h)_{\mathrm{LS}}$ of the common points into the Cartesian coordinates $(X, Y, Z)_{\mathrm{LS}}$ by (5–27).

3. Determine the seven parameters of a Helmert transformation by using the coordinates $(X, Y, Z)_{\mathrm{GPS}}$ and $(X, Y, Z)_{\mathrm{LS}}$ of the common points.

4. For network points other than the common points, transform the coordinates $(X, Y, Z)_{\mathrm{GPS}}$ into $(X, Y, Z)_{\mathrm{LS}}$ via Eq. (5–41) using the transformation parameters determined in the previous step.

5. Transform the Cartesian coordinates $(X, Y, Z)_{\mathrm{LS}}$ computed in the previous step into ellipsoidal coordinates $(\varphi, \lambda, h)_{\mathrm{LS}}$, e.g., by the iterative procedure given in (5–28) through (5–34).

6. Map the ellipsoidal surface coordinates $(\varphi, \lambda)_{\mathrm{LS}}$ computed in the previous step into plane coordinates $(y, x)_{\mathrm{LS}}$ by the appropriate mapping formulas.

The advantage of the three-dimensional approach is that no a priori information is required for the seven parameters of the similarity transformation. The disadvantage of the method is that for the common points ellipsoidal heights (and, thus, geoidal heights) are required. However, as reported by Schmitt et al. (1991), incorrect heights of the common points often have a negligible effect on the plane coordinates (y, x). For example, incorrect heights may cause a tilt of a 20 km × 20 km network by an amount of 5 m in space; however, the effect on the plane coordinates is only approximately 1 mm.

For large areas, the height problem can be solved by adopting approximate ellipsoidal heights for the common points and performing a three-dimensional affine transformation instead of the similarity transformation.

5.7.4 Differential formulas for other datum transformations

Now we consider simplified cases. Suppose that the geocenter does not coincide with the center of the reference ellipsoid, but that *the geocentric axes and the ellipsoidal axes are parallel*. Such a parallel shift is also called a *translation* (Fig. 5.8). Assume a rectangular coordinate system XYZ whose origin is the geocenter, the axes being directed as usual. Let the coordinates of the center of the ellipsoid with respect to this system be x_0, y_0, z_0, as stated previously. Then Eqs. (5–27) must obviously be modified so that they become

$$\begin{aligned} X &= x_0 + (N + h) \cos\varphi \cos\lambda \,, \\ Y &= y_0 + (N + h) \cos\varphi \sin\lambda \,, \\ Z &= z_0 + \left(\frac{b^2}{a^2} N + h \right) \sin\varphi \,. \end{aligned} \tag{5–53}$$

These equations form the starting point for various important differential formulas of coordinate transformation.

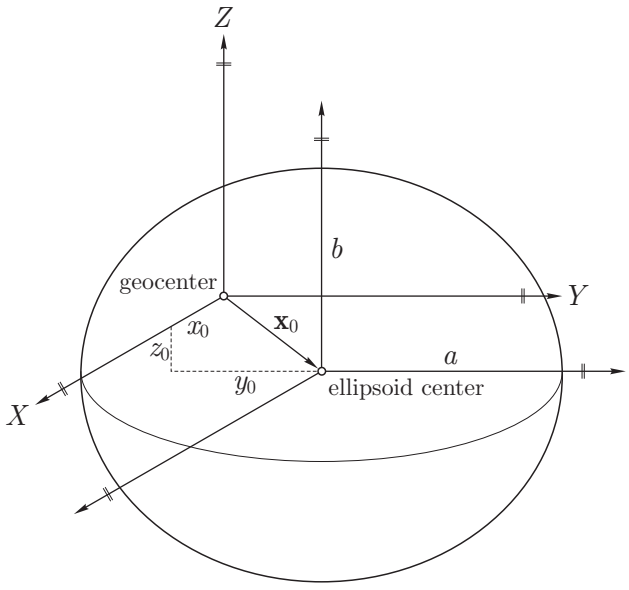

Fig. 5.8. Translation problem

First we ask how the rectangular coordinates X, Y, Z change if we vary the ellipsoidal coordinates φ, λ, h by small amounts $\delta\varphi, \delta\lambda, \delta h$ and if we also alter the geodetic datum, namely, the reference ellipsoid a, f and its position x_0, y_0, z_0, by $\delta a, \delta f$ and $\delta x_0, \delta y_0, \delta z_0$. Note that $\delta x_0, \delta y_0, \delta z_0$ correspond to a small translation (parallel displacement) of the ellipsoid, *its axis remaining parallel to the axis of the earth.*

The solution of this problem is found by differentiating (5–53):

$$\begin{aligned}
\delta X &= \delta x_0 + \frac{\partial X}{\partial a}\,\delta a + \frac{\partial X}{\partial f}\,\delta f + \frac{\partial X}{\partial \varphi}\,\delta\varphi + \frac{\partial X}{\partial \lambda}\,\delta\lambda + \frac{\partial X}{\partial h}\,\delta h\,,\\
\delta Y &= \delta y_0 + \frac{\partial Y}{\partial a}\,\delta a + \frac{\partial Y}{\partial f}\,\delta f + \frac{\partial Y}{\partial \varphi}\,\delta\varphi + \frac{\partial Y}{\partial \lambda}\,\delta\lambda + \frac{\partial Y}{\partial h}\,\delta h\,,\\
\delta Z &= \delta z_0 + \frac{\partial Z}{\partial a}\,\delta a + \frac{\partial Z}{\partial f}\,\delta f + \frac{\partial Z}{\partial \varphi}\,\delta\varphi + \frac{\partial Z}{\partial \lambda}\,\delta\lambda + \frac{\partial Z}{\partial h}\,\delta h\,,
\end{aligned} \tag{5–54}$$

since, according to Taylor's theorem, small changes can be treated as differentials.

In these differential formulas we shall be satisfied with an approximation.

Since the flattening f is small, we may expand (2–149) as

$$\begin{aligned} N &= \frac{a^2}{b}\,(1+e'^2\cos^2\varphi)^{-1/2} = \frac{a^2}{b}\,\left(1-\tfrac{1}{2}\,e'^2\cos^2\varphi\,\cdots\right) \\ &= a\,(1+f\,\cdots)(1-f\cos^2\varphi\,\cdots) = a(1+f-f\cos^2\varphi\,\cdots) \end{aligned} \tag{5–55}$$

yielding

$$N \doteq a\,(1+f\sin^2\varphi) \tag{5–56}$$

and

$$\frac{b^2}{a^2}\,N = (1-2f\,\cdots)\,a\,(1+f\sin^2\varphi\,\cdots) \doteq a\,(1-2f+f\sin^2\varphi) \tag{5–57}$$

and

$$b = a\,(1-f)\,, \qquad e'^2 = 2f\,\cdots\,. \tag{5–58}$$

Thus, Eqs. (5–53) are approximated by

$$\begin{aligned} X &= x_0 + (a + af\sin^2\varphi + h)\cos\varphi\,\cos\lambda\,, \\ Y &= y_0 + (a + af\sin^2\varphi + h)\cos\varphi\,\sin\lambda\,, \\ Z &= z_0 + (a - 2af + af\sin^2\varphi + h)\sin\varphi\,. \end{aligned} \tag{5–59}$$

Now we can form the partial derivatives in (5–54), for instance,

$$\frac{\partial X}{\partial a} = (1+f\sin^2\varphi)\cos\varphi\,\cos\lambda \doteq \cos\varphi\,\cos\lambda\,, \tag{5–60}$$

since we may neglect the flattening in these coefficients. This amounts to using for the coefficients, and only for them, a spherical approximation analogous to that of Sect. 2.13. Similarly, all coefficients are easily obtained as partial derivatives, and Eqs. (5–54) become

$$\begin{aligned} \delta X &= \delta x_0 - a\,\sin\varphi\,\cos\lambda\,\delta\varphi - a\,\cos\varphi\,\sin\lambda\,\delta\lambda \\ &\quad + \cos\varphi\,\cos\lambda\,(\delta h + \delta a + a\,\sin^2\varphi\,\delta f)\,, \\ \delta Y &= \delta y_0 - a\,\sin\varphi\,\sin\lambda\,\delta\varphi + a\,\cos\varphi\,\cos\lambda\,\delta\lambda \\ &\quad + \cos\varphi\,\sin\lambda\,(\delta h + \delta a + a\,\sin^2\varphi\,\delta f)\,, \\ \delta Z &= \delta z_0 + a\,\cos\varphi\,\delta\varphi + \sin\varphi\,(\delta h + \delta a + a\,\sin^2\varphi\,\delta f) \\ &\quad - 2a\,\sin\varphi\,\delta f\,. \end{aligned} \tag{5–61}$$

These formulas give the changes in the rectangular coordinates X, Y, Z in terms of the variation in the position (x_0, y_0, z_0) and the dimensions (a, f) of the ellipsoid and in the ellipsoidal coordinates φ, λ, h referred to it.

Transformation of the ellipsoidal coordinates

Several important formulas for the transformation of coordinates may be derived from Eqs. (5–61). First, let the position of P in space remain unchanged; that is, let

$$\delta X = \delta Y = \delta Z = 0\,. \tag{5–62}$$

Determine the change of the ellipsoidal coordinates φ, λ, h if the dimensions of the reference ellipsoid and its position are varied. Geometrically, this is illustrated by Fig. 5.9. The problem is, thus, to solve equations (5–61) for $\delta\varphi, \delta\lambda, \delta h$, the left-hand sides being set equal to zero. To get $\delta\varphi$, multiply the first equation of (5–61) by $-\sin\varphi\cos\lambda$, the second equation of (5–61) by $-\sin\varphi\sin\lambda$, and the third equation of (5–61) by $\cos\varphi$ and add all equations obtained in this way. For $\delta\lambda$, the factors are $-\sin\lambda$, $\cos\lambda$, and 0; for δh, they are $\cos\varphi\cos\lambda$, $\cos\varphi\sin\lambda$, and $\sin\varphi$. The result is

$$\begin{aligned}
a\,\delta\varphi &= \sin\varphi\cos\lambda\,\delta x_0 + \sin\varphi\sin\lambda\,\delta y_0 - \cos\varphi\,\delta z_0 + 2a\sin\varphi\cos\varphi\,\delta f\,,\\
a\cos\varphi\,\delta\lambda &= \sin\lambda\,\delta x_0 - \cos\lambda\,\delta y_0\,,\\
\delta h &= -\cos\varphi\cos\lambda\,\delta x_0 - \cos\varphi\sin\lambda\,\delta y_0 - \sin\varphi\,\delta z_0 - \delta a + a\sin^2\varphi\,\delta f\,.
\end{aligned} \tag{5–63}$$

These formulas express the variations $\delta\varphi, \delta\lambda, \delta h$ at an arbitrary point in terms of the variations $\delta x_0, \delta y_0, \delta z_0$ at a given point and the changes δa and δf of the parameters of the reference ellipsoid. Thus, they relate two different systems of ellipsoidal coordinates, provided these systems are so close to each other that their differences may be considered as linear. Mathematically, Eqs. (5–63) are infinitesimal coordinate transformations (essentially but not exclusively orthogonal transformations); to the geodesist, they give the effect

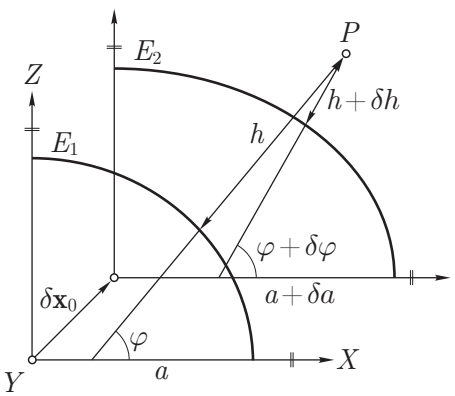

Fig. 5.9. A small change of the reference ellipsoid together with a small parallel shift

of a change in the geodetic datum.

Remark. The differential formulas could also be replaced by a successive application of the original finite formulas. Try!

Part II: Three-dimensional geodesy: a transition

5.8 The three-dimensional geodesy of Bruns and Hotine

The idea of a computation of a triangulation network in space dates back to Bruns (1878). On the basis of his ideas, Hotine (1969), and earlier in 1959, developed extensively the concept of a classical (pre-satellite) geodetic network in a rigorous three-dimensional way. For a comparison, see Levallois (1963).

Consider the polyhedron formed by triangulation benchmarks on the surface of the earth and the straight lines of sight connecting them (Fig. 5.10). Another set of straight lines – one through each corner – represents the plumb line at the stations.

In order to determine this figure, we need five parameters for each station – three coordinates and two parameters defining the direction of the plumb line. The main terrestrial observational data for this purpose are

1. horizontal angles and zenith angles, obtained by theodolite observations;

2. straight spatial distances, obtained by electronic distance measurements; and

3. astronomical observations of latitude and longitude to fix the direction of the plumb line, and of azimuth to determine the orientation of the polyhedron.

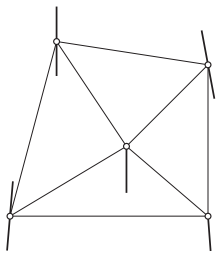

Fig. 5.10. Bruns' polyhedron

We may use a rectangular coordinate system; then the three coordinates to be determined will be X, Y, Z. The parameters defining the direction of the plumb line are conveniently taken to be Φ and Λ, astronomical latitude and longitude. We can express the astronomical azimuth A, the measured zenith angle z', and the spatial distance s in terms of these five parameters. This will be the scope of Sect. 5.9.

This information is purely "geometric". We need the terrestrial measurements (especially Φ, Λ, A) in order to link this geometry to the gravity field as represented by the plumb lines. The Bruns polyhedron is the best way to show this geometrically.

Today, GPS is the best way to determine global rectangular coordinates X, Y, Z or ellipsoidal coordinates φ, λ, h directly.

5.9 Global coordinates and local level coordinates

We shall use a Cartesian coordinate system XYZ introduced in Sect. 5.6.1, global but not necessarily geocentric. The coordinates X, Y, Z form a vector $\mathbf{X}$. Thus, the vectors $\mathbf{X}_i$ and $\mathbf{X}_j$ represent two terrestrial points P_i and P_j. We define the vector between these two points in the global coordinate system by $\mathbf{X}_{ij} = \mathbf{X}_j - \mathbf{X}_i$.

In addition, we introduce a "local level system" referred to the tangential plane to the level surface at a point P_i and to the local vertical, which is the tangent at P_i to the natural plumb line defined by the astronomical coordinates Φ and Λ, see Sect. 2.4. The axes $\mathbf{n}_i$, $\mathbf{e}_i$, $\mathbf{u}_i$ of this local (tangent plane) coordinate system at P_i corresponding to the north, east, and up

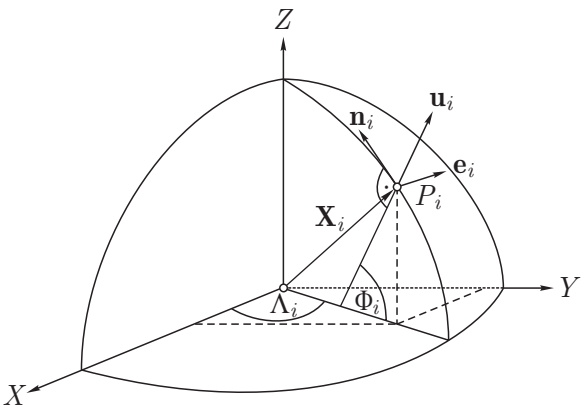

Fig. 5.11. Global and local level coordinates

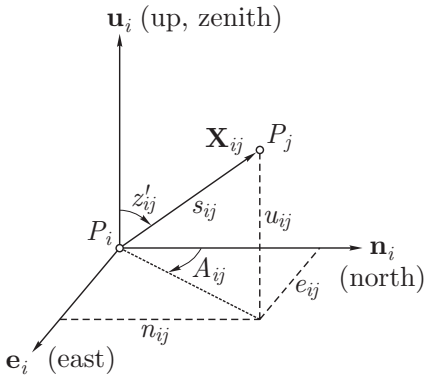

Fig. 5.12. Measurement quantities in the local level system

direction, are thus represented in the global system by

$$\mathbf{n}_i = \begin{bmatrix} -\sin\Phi_i \cos\Lambda_i \\ -\sin\Phi_i \sin\Lambda_i \\ \cos\Phi_i \end{bmatrix}, \quad \mathbf{e}_i = \begin{bmatrix} -\sin\Lambda_i \\ \cos\Lambda_i \\ 0 \end{bmatrix}, \quad \mathbf{u}_i = \begin{bmatrix} \cos\Phi_i \cos\Lambda_i \\ \cos\Phi_i \sin\Lambda_i \\ \sin\Phi_i \end{bmatrix}, \tag{5–64}$$

where the vectors $\mathbf{n}_i$ and $\mathbf{e}_i$ span the tangent plane at P_i (Fig. 5.11). The third coordinate axis of the local level system, i.e., the vector $\mathbf{u}_i$, is orthogonal to the tangent plane and has the direction of the plumb line.

Now the components n_{ij}, e_{ij}, u_{ij} of the vector $\mathbf{x}_{ij}$ in the local level system are introduced. These coordinates are sometimes denoted as ENU (east, north, up) coordinates. Considering Fig. 5.12, these components are obtained by a projection of vector $\mathbf{X}_{ij}$ onto the local level axes $\mathbf{n}_i$, $\mathbf{e}_i$, $\mathbf{u}_i$. Analytically, this is achieved by scalar products. Therefore,

$$\mathbf{x}_{ij} = \begin{bmatrix} n_{ij} \\ e_{ij} \\ u_{ij} \end{bmatrix} = \begin{bmatrix} \mathbf{n}_i \cdot \mathbf{X}_{ij} \\ \mathbf{e}_i \cdot \mathbf{X}_{ij} \\ \mathbf{u}_i \cdot \mathbf{X}_{ij} \end{bmatrix} \tag{5–65}$$

is obtained. Assembling the vectors $\mathbf{n}_i$, $\mathbf{e}_i$, $\mathbf{u}_i$ of the local level system as columns in an orthogonal matrix $\mathbf{D}_i$, i.e.,

$$\mathbf{D}_i = \begin{bmatrix} -\sin\Phi_i \cos\Lambda_i & -\sin\Lambda_i & \cos\Phi_i \cos\Lambda_i \\ -\sin\Phi_i \sin\Lambda_i & \cos\Lambda_i & \cos\Phi_i \sin\Lambda_i \\ \cos\Phi_i & 0 & \sin\Phi_i \end{bmatrix}, \tag{5–66}$$

relation (5–65) may be written concisely as

$$\mathbf{x}_{ij} = \mathbf{D}_i^T \, \mathbf{X}_{ij} \, . \tag{5–67}$$

The components of $\mathbf{x}_{ij}$ may also be expressed by the spatial distance s_{ij}, the azimuth A_{ij}, and the zenith angle z'_{ij}, which is assumed to be corrected for refraction. The appropriate relation is

$$\mathbf{x}_{ij} = \begin{bmatrix} n_{ij} \\ e_{ij} \\ u_{ij} \end{bmatrix} = \begin{bmatrix} s_{ij} \sin z'_{ij} \cos A_{ij} \\ s_{ij} \sin z'_{ij} \sin A_{ij} \\ s_{ij} \cos z'_{ij} \end{bmatrix}, \tag{5–68}$$

where the terrestrial measurement quantities s_{ij}, A_{ij}, z'_{ij} refer to P_i, i.e., the measurements were taken at P_i. Inverting (5–68) gives the measurement quantities explicitly:

$$\begin{aligned} s_{ij} \quad &= \sqrt{n_{ij}^2 + e_{ij}^2 + u_{ij}^2}\,, \\ \tan A_{ij} &= \frac{e_{ij}}{n_{ij}}\,, \\ \cos z'_{ij} \ &= \frac{u_{ij}}{\sqrt{n_{ij}^2 + e_{ij}^2 + u_{ij}^2}}\,. \end{aligned} \tag{5–69}$$

Substituting (5–65) for n_{ij}, e_{ij} and u_{ij}, the measurement quantities may be expressed by the components of the vector $\mathbf{X}_{ij}$ in the global system.

A note on azimuth and zenith distance

Since the local level coordinates refer to the local plumb line defined by the astronomical coordinates Φ, Λ (Sect. 2.4), A and z' are called astronomical azimuth and astronomical zenith distance (zenith angle). They will also play a basic role in Part III.

A final word on the zenith distance. The measured ("astronomical") azimuth is denoted by A, and the corresponding ellipsoidal azimuth is denoted by α. Since the ellipsoidal zenith distance is conventionally denoted by z, it would be consistent to indicate the measured ("astronomical") zenith distance by Z. This symbol, however, is firmly reserved for the third axis of the XYZ system, so we exceptionally, but consistently with the rest of the book, use the symbol z'. (Both A and z' will return in the following sections.)

5.10 Combining terrestrial data and GPS

5.10.1 Common coordinate system

So far, GPS and terrestrial networks have been considered separately with respect to the adjustment. The combination, for example, by a datum transformation, was supposed to be performed after individual adjustments. Now

the common adjustment of GPS observations and terrestrial data is investigated. The problem encountered here is that GPS data refer to the three-dimensional geocentric Cartesian system WGS 84, whereas terrestrial data refer to the individual local level (tangent plane) systems at each measurement point referenced to plumb lines. Furthermore, terrestrial data are traditionally separated into position and height, where the position refers to an ellipsoid and the (orthometric) height to the geoid.

For a joint adjustment, a common coordinate system is required to which all observations are transformed. In principle, any arbitrary system may be introduced as common reference. One possibility is to use two-dimensional (plane) coordinates in the local system as proposed by Daxinger and Stirling (1995). Here, a three-dimensional coordinate system is chosen. The origin of the coordinate system is the center of the ellipsoid adopted for the local system, the Z-axis coincides with the semiminor axis of the ellipsoid, the X-axis is obtained by the intersection of the ellipsoidal Greenwich meridian plane and the ellipsoidal equatorial plane, and the Y-axis completes the right-handed system. Position vectors referred to this system are denoted by $\mathbf{X}_{\mathrm{LS}}$, where LS indicates the reference to the local system.

After the decision on the common coordinate system, the terrestrial measurements referring to the individual local level systems at the observing sites must be represented in this common coordinate system. Similarly, GPS baseline vectors regarded as measurement quantities are to be transformed to this system.

5.10.2 Representation of measurement quantities

Distances

The spatial distance s_{ij} as function of the local level coordinates is given in (5–69). If n_{ij}, e_{ij}, u_{ij}, the components of $\mathbf{x}_{ij}$, are substituted by (5–65), the relation

$$\begin{aligned} s_{ij} &= \sqrt{n_{ij}^2 + e_{ij}^2 + u_{ij}^2} \\ &= \sqrt{(X_j - X_i)^2 + (Y_j - Y_i)^2 + (Z_j - Z_i)^2} \end{aligned} \tag{5–70}$$

is obtained, where (5–64) has also been taken into account, namely, the fact that $\mathbf{n}_i$, $\mathbf{e}_i$, $\mathbf{u}_i$ are unit vectors. Obviously, the second expression arises immediately from the Pythagorean theorem. Differentiation of (5–70) yields

$$ds_{ij} = \frac{X_{ij}}{s_{ij}}\,(dX_j - dX_i) + \frac{Y_{ij}}{s_{ij}}\,(dY_j - dY_i) + \frac{Z_{ij}}{s_{ij}}\,(dZ_j - dZ_i)\,, \tag{5–71}$$

where

$$\begin{aligned} X_{ij} &= X_j - X_i\,, \\ Y_{ij} &= Y_j - Y_i\,, \\ Z_{ij} &= Z_j - Z_i \end{aligned} \tag{5–72}$$

have been introduced accordingly. The relation (5–71) may also be expressed as

$$\delta s_{ij} = \frac{X_{ij}}{s_{ij}}\,(\delta X_j - \delta X_i) + \frac{Y_{ij}}{s_{ij}}\,(\delta Y_j - \delta Y_i) + \frac{Z_{ij}}{s_{ij}}\,(\delta Z_j - \delta Z_i) \tag{5–73}$$

if the differentials are replaced by differences.

Azimuths

Again the same principle applies: the measured azimuth A_{ij} as a function of the local level coordinates is given in (5–69). If n_{ij}, e_{ij}, u_{ij}, the components of $\mathbf{x}_{ij}$, are substituted by (5–65), the relation

$$\begin{aligned} \tan A_{ij} &= e_{ij}/n_{ij} \\ &= \frac{-X_{ij}\sin\Lambda_i + Y_{ij}\cos\Lambda_i}{-X_{ij}\sin\Phi_i\cos\Lambda_i - Y_{ij}\sin\Phi_i\sin\Lambda_i + Z_{ij}\cos\Phi_i} \end{aligned} \tag{5–74}$$

is obtained. After a lengthy derivation, the relation

$$\begin{aligned} \delta A_{ij} = {} & \frac{\sin\varphi_i\cos\lambda_i\sin\alpha_{ij} - \sin\lambda_i\cos\alpha_{ij}}{s_{ij}\sin z_{ij}}\,(\delta X_j - \delta X_i) \\ & + \frac{\sin\varphi_i\sin\lambda_i\sin\alpha_{ij} + \cos\lambda_i\cos\alpha_{ij}}{s_{ij}\sin z_{ij}}\,(\delta Y_j - \delta Y_i) \\ & - \frac{\cos\varphi_i\sin\alpha_{ij}}{s_{ij}\sin z_{ij}}\,(\delta Z_j - \delta Z_i) \\ & + \cot z_{ij}\sin\alpha_{ij}\,\delta\Phi_i \\ & + (\sin\varphi_i - \cos\alpha_{ij}\cos\varphi_i\cot z_{ij})\,\delta\Lambda_i \end{aligned} \tag{5–75}$$

is obtained. Approximate values are sufficient *in the coefficients*, denoted by $\varphi, \lambda, \alpha, z$ instead of Φ, Λ, A, z'.

Directions

Measured directions R_{ij} are related to azimuths A_{ij} by the orientation unknown o_i. The relation reads

$$R_{ij} = A_{ij} - o_i\,, \tag{5–76}$$

and the expression

$$\delta R_{ij} = \delta A_{ij} - \delta o_i \tag{5–77}$$

is immediately obtained.

Zenith angles

The zenith angle z'_{ij} as function of the local level coordinates is given in (5–69). If n_{ij}, e_{ij}, u_{ij}, the components of $\mathbf{x}_{ij}$, are substituted by (5–65), the relation

$$\begin{aligned} \cos z'_{ij} &= u_{ij}/s_{ij} \\ &= \frac{X_{ij}\cos\Phi_i\cos\Lambda_i + Y_{ij}\cos\Phi_i\sin\Lambda_i + Z_{ij}\sin\Phi_i}{\sqrt{X_{ij}^2 + Y_{ij}^2 + Z_{ij}^2}} \end{aligned} \tag{5–78}$$

is obtained, where (5–70) and (5–72) have been used. After a lengthy derivation, the relation

$$\begin{aligned} \delta z'_{ij} = &\frac{X_{ij}\cos z_{ij} - s_{ij}\cos\varphi_i\cos\lambda_i}{s_{ij}^2\sin z_{ij}}\,(\delta X_j - \delta X_i) \\ &+ \frac{Y_{ij}\cos z_{ij} - s_{ij}\cos\varphi_i\sin\lambda_i}{s_{ij}^2\sin z_{ij}}\,(\delta Y_j - \delta Y_i) \\ &+ \frac{Z_{ij}\cos z_{ij} - s_{ij}\sin\varphi_i}{s_{ij}^2\sin z_{ij}}\,(\delta Z_j - \delta Z_i) \\ &- \cos\alpha_{ij}\,\delta\Phi_i - \cos\varphi_i\sin\alpha_{ij}\,\delta\Lambda_i \end{aligned} \tag{5–79}$$

is obtained.

It is presupposed that the zenith angles are reduced to the chord of the light path. This reduction may be modeled by

$$z'_{ij} = z'_{ij_{\text{meas}}} + \frac{s_{ij}}{2R}\,k\,, \tag{5–80}$$

where $z_{ij_{\text{meas}}}$ is the measured zenith angle, R is the mean radius of the earth, and k is the coefficient of refraction. For k either a standard value may be substituted or the coefficient of refraction is estimated as additional unknown. In the case of estimation, there are several choices, e.g., one value for k for all zenith angles or one value for a group of zenith angles or one value per day. (It is known that measured zenith angles are "weaker" than other observations, which can be taken into account by giving them lower weights.)

Ellipsoidal height differences

The "measured" ellipsoidal height difference is represented by

$$h_{ij} = h_j - h_i \,. \tag{5–81}$$

The heights involved are obtained by transforming the Cartesian coordinates into ellipsoidal coordinates according to (5–36) or by using the iterative procedure given in Sect. 5.6.1. The height difference is approximately (neglecting the curvature of the earth) given by the third component of $\mathbf{x}_{ij}$ in the local level system. Hence,

$$h_{ij} = \mathbf{u}_i \cdot \mathbf{X}_{ij} \tag{5–82}$$

or, by substituting $\mathbf{u}_i$ according to (5–64), the relation

$$h_{ij} = \cos\Phi_i \cos\Lambda_i \, X_{ij} + \cos\Phi_i \sin\Lambda_i \, Y_{ij} + \sin\Phi_i \, Z_{ij} \tag{5–83}$$

is obtained. This equation may be differentiated with respect to the Cartesian coordinates. If the differentials are replaced by the corresponding differences,

$$\begin{aligned} \delta h_{ij} = {} & \cos\Phi_j \cos\Lambda_j \, \delta X_j + \cos\Phi_j \sin\Lambda_j \, \delta Y_j + \sin\Phi_j \, \delta Z_j \\ & - \cos\Phi_i \cos\Lambda_i \, \delta X_i - \cos\Phi_i \sin\Lambda_i \, \delta Y_i - \sin\Phi_i \, \delta Z_i \end{aligned} \tag{5–84}$$

is obtained, where the coordinate differences were decomposed into their individual coordinates.

Baselines

From relative GPS measurements, baselines $\mathbf{X}_{ij_{(\mathrm{GPS})}} = \mathbf{X}_{j_{(\mathrm{GPS})}} - \mathbf{X}_{i_{(\mathrm{GPS})}}$ in the WGS 84 are obtained. The position vectors $\mathbf{X}_{i_{(\mathrm{GPS})}}$ and $\mathbf{X}_{j_{(\mathrm{GPS})}}$ may be transformed by a three-dimensional (7-parameter) similarity transformation to a local system indicated by LS. According to Eq. (5–41), the transformation formula reads

$$\mathbf{X}_{\mathrm{LS}} = \mathbf{x}_0 + \mu \, \mathbf{R} \, \mathbf{X}_{\mathrm{GPS}} \,, \tag{5–85}$$

where the meaning of the individual quantities is the following:

$\mathbf{X}_{\mathrm{LS}}$	...	position vector in the local system ,
$\mathbf{X}_{\mathrm{GPS}}$	...	position vector in the WGS 84 ,
$\mathbf{x}_0$	...	shift vector ,
$\mathbf{R}$	...	rotation matrix ,
μ	...	scale factor .

Forming the difference of two position vectors, i.e., the baseline $\mathbf{X}_{ij}$, the shift vector $\mathbf{x}_0$ is eliminated. Using (5–85), there results

$$\mathbf{X}_{ij_{(\mathrm{LS})}} = \mu \, \mathbf{R} \, \mathbf{X}_{ij_{(\mathrm{GPS})}} \tag{5–86}$$

for the baseline. Similar to (5–49), the linearized form is

$$\mathbf{X}_{ij_{(\mathrm{LS})}} = \mathbf{X}_{ij_{(\mathrm{GPS})}} + \mathbf{A}_{ij}\,\delta\mathbf{p}\,, \tag{5–87}$$

where now the vector $\delta\mathbf{p}$ and the design matrix $\mathbf{A}_{ij}$ are given by

$$\begin{aligned} \delta\mathbf{p} &= [\delta\mu \quad \varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3]^T, \\ \mathbf{A}_{ij} &= \begin{bmatrix} X_{ij} & 0 & -Z_{ij} & Y_{ij} \\ Y_{ij} & Z_{ij} & 0 & -X_{ij} \\ Z_{ij} & -Y_{ij} & X_{ij} & 0 \end{bmatrix}_{(\mathrm{GPS})} . \end{aligned} \tag{5–88}$$

Note that the rotations ε_i refer to the axes of the system used in GPS. If they should refer to the local system, then the signs of the rotations must be changed, i.e., the signs of the elements of the last three columns of matrix $\mathbf{A}_{ij}$ must be reversed.

The vector $\mathbf{X}_{ij_{(\mathrm{LS})}}$ on the left side of (5–87) contains the points $\mathbf{X}_{i_{(\mathrm{LS})}}$ and $\mathbf{X}_{j_{(\mathrm{LS})}}$ in the local system. If these points are unknown, then they are replaced by known approximate values and unknown increments

$$\begin{aligned} \mathbf{X}_{i_{(\mathrm{LS})}} &= \mathbf{X}_{i0_{(\mathrm{LS})}} + \delta\mathbf{X}_{i_{(\mathrm{LS})}}\,, \\ \mathbf{X}_{j_{(\mathrm{LS})}} &= \mathbf{X}_{j0_{(\mathrm{LS})}} + \delta\mathbf{X}_{j_{(\mathrm{LS})}}\,, \end{aligned} \tag{5–89}$$

where the coefficients of these unknown increments ($+1$ or -1) together with matrix $\mathbf{A}_{ij}$ form the design matrix.

The vector $\mathbf{X}_{ij_{(\mathrm{GPS})}}$ in (5–87) is regarded as measurement quantity. Thus, finally,

$$\mathbf{X}_{ij_{(\mathrm{GPS})}} = \delta\mathbf{X}_{j_{(\mathrm{LS})}} - \delta\mathbf{X}_{i_{(\mathrm{LS})}} - \mathbf{A}_{ij}\,\delta\mathbf{p} + \mathbf{X}_{j0_{(\mathrm{LS})}} - \mathbf{X}_{i0_{(\mathrm{LS})}} \tag{5–90}$$

is the linearized observation equation.

In principle, any type of geodetic measurement can be employed if the integrated geodesy adjustment model is used. The basic concept is that any geodetic measurement can be expressed as a function of one or more position vectors $\mathbf{X}$ and of the gravity field W of the earth. The usually nonlinear function must be linearized where the gravity field W is split into the normal potential U of an ellipsoid and the disturbing potential T, thus, $W = U + T$. Applying a minimum principle leads to the collocation formulas (Moritz 1980 a: Chap. 11).

Many examples integrating GPS and other data can be found in technical publications. For example, there are attempts to detect earth deformations from GPS and terrestrial data.

Part III: Local geodetic datums

5.11 Formulation of the problem

As we have remarked several times, the weak point of the Bruns–Hotine method is the insufficient accuracy of the zenith angle measurement precluding the practical use of this method for larger triangulations. The trigonometric heights obtained in this way are significantly less accurate than the horizontal positions.

A practical solution of this problem was to separate positions and heights. The horizontal position was calculated on the reference ellipsoid in the way we shall see later. Accurate heights were obtained by leveling referred to the "actual" level surfaces, in particular to the geoid.

Thus, this theoretically and practically unsatisfactory procedure used two different reference surfaces: the ellipsoid for horizontal position and the geoid for heights. The mutual position of these two surfaces was not even known because of lack of knowledge of the geoidal height N. It has been rightfully ridiculed as "2+1-dimensional geodesy".

There is a way out of this dilemma even for local (or rather regional) geodetic systems. The trigonometric height h is not determined by zenith-angle measurements but by using the simple formula

$$h = H + N \tag{5–91}$$

from leveled orthometric heights H by adding the geoid height N!

But how do we get the geoid? Even before the satellite era, there existed two methods:

1. the *astrogeodetic method*, determining N from deflections of the vertical ξ and η;

2. the *gravimetric method*, using for this purpose gravity anomalies Δg.

The theories of both methods were known as early as 1850, but what was lacking were data, especially gravimetric ones. Serious practical applications started not much before 1950, a hundred years later, just before the advent of satellites. This will be discussed in detail later in this book.

A reasonable measuring accuracy was achievable, but another difficulty appeared. Both methods require the evaluation of integrals of the data (ξ and η, or Δg) as continuous functions. The data, however, are always measured at discrete points only. Interpolation is necessary and introduces additional errors. If the data are distributed uniformly and densely, resulting errors may be kept small. The fundamental problem exists, however.

Summarizing, we may say: (1) The method of zenith angles is theoretically rigorous but not in general sufficiently accurate; (2) the astrogeodetic method using integration of vertical deflections is not theoretically rigorous in this sense but still may be accurate enough.

Method 1 has been treated in Part II of this chapter, so method 2 warrants detailed considerations in the present Part III.

5.12 Reduction of the astronomical measurements to the ellipsoid

Now we establish the relation between the natural coordinates Φ, Λ, H and the ellipsoidal coordinates φ, λ, h referring to an ellipsoid according to Helmert's projection.

The ellipsoidal height h and the orthometric height H have been considered, e.g., in Sect. 4.6 (see also Fig. 5.4 and Eq. (5–91)). They are related by $h = H + N$.

Thus, there remains the *reduction of the astronomical coordinates* Φ *and* Λ *to the ellipsoid* and, if we also include the astronomical observation of the azimuth, the astronomical azimuth A to the ellipsoid in order to obtain the ellipsoidal coordinates φ and λ and the ellipsoidal azimuth α.

We introduce the auxiliary quantities

$$\begin{aligned} \Delta\varphi &= \Phi - \varphi\,, \\ \Delta\lambda &= \Lambda - \lambda\,, \\ \Delta\alpha &= A - \alpha\,. \end{aligned} \tag{5–92}$$

The reduction of Φ and Λ to the corresponding ellipsoidal coordinates φ and λ is implicitly contained in Eq. (2–230):

$$\begin{aligned} \xi &= \Phi - \varphi = \Delta\varphi\,, \\ \eta &= (\Lambda - \lambda)\,\cos\varphi = \Delta\lambda\,\cos\varphi\,, \end{aligned} \tag{5–93}$$

where we have substituted the respective auxiliary quantities. Thus, the conversion formulas from natural coordinates Φ, Λ, H to ellipsoidal coordinates φ, λ, h are

$$\begin{aligned} \varphi &= \Phi - \xi\,, \\ \lambda &= \Lambda - \eta/\cos\varphi\,, \\ h &= H + N\,. \end{aligned} \tag{5–94}$$

Now we turn to the reduction of the azimuth. Thus, the question is which $\Delta\alpha$ arises from $\Delta\varphi$ and $\Delta\lambda$. The answer is found in Eq. (5–75), where we

only consider the last two terms on the right-hand side (i.e., we do not take into account changes of the point coordinates). Omitting all subscripts and introducing the auxiliary quantities of (5–92), we immediately get

$$\Delta\alpha = \cot z \, \sin\alpha \, \Delta\varphi + (\sin\varphi - \cos\alpha \, \cos\varphi \, \cot z) \, \Delta\lambda \tag{5–95}$$

or, using $\Delta\varphi = \xi$ and $\Delta\lambda \cos\varphi = \eta$, yields

$$\Delta\alpha = \xi \sin\alpha \, \cot z + \sin\varphi \, \Delta\lambda - \eta \cos\alpha \, \cot z \, . \tag{5–96}$$

This equation may be rearranged to

$$\Delta\alpha = \sin\varphi \, \Delta\lambda + (\xi \sin\alpha - \eta \cos\alpha) \cot z \, . \tag{5–97}$$

Alternatively, by using $\Delta\lambda = \eta / \cos\varphi$, we get

$$\Delta\alpha = \eta \tan\varphi + (\xi \sin\alpha - \eta \cos\alpha) \cot z \, . \tag{5–98}$$

In first-order triangulation, the lines of sight are usually almost horizontal so that $z \doteq 90^\circ$, $\cot z \doteq 0$. Therefore, the corresponding term can in general be neglected and we get

$$\Delta\alpha = \eta \, \tan\varphi = \Delta\lambda \, \sin\varphi \, . \tag{5–99}$$

This is *Laplace's equation* in its usual simplified form. It is remarkable that the differences $\Delta\alpha = A - \alpha$ and $\Delta\lambda = \Lambda - \lambda$ should be related in such a simple way. Laplace's equation is fundamental for the classical astrogeodetic computation of triangulations (Sect. 5.14).

For later reference we note that the total deflection of the vertical – that is, the angle ϑ between the actual plumb line and the ellipsoidal normal – is given by

$$\vartheta = \sqrt{\xi^2 + \eta^2} \tag{5–100}$$

and that the deflection component ε in the direction of the azimuth α is

$$\varepsilon = \xi \, \cos\alpha + \eta \, \sin\alpha \, . \tag{5–101}$$

It is clear that ϑ in (5–100) has nothing to do with the two different ϑ used for spherical and ellipsoidal-harmonic coordinates (polar distances).

Returning to the reduction of astronomical to the corresponding ellipsoidal quantities, we have (5–94) for the reduction of Φ, Λ, H to φ, λ, h and, finally, the formula

$$\alpha = A - \eta \, \tan\varphi \tag{5–102}$$

reduces the astronomical azimuth A to the ellipsoidal azimuth α.

For the application of these formulas, we need the geoidal undulation N and the deflection components ξ and η with respect to the reference ellipsoid used. Two points should be noted:

1. The vertical axis of the reference ellipsoid is parallel to the earth's axis of rotation, but it need not be in an absolute position, its center coinciding with the earth's center of gravity. To repeat the reason: the earth axis is accessible to (astronomical) observation, whereas the geocenter is physically defined and inaccessible to direct geometrical observation.

2. The geocenter is accessible in two physically defined ways: (1) gravimetrically through Stokes' formula and (2) physically by the first Kepler law applied to satellite motion and responsible for the geocentricity of GPS orbits.

Note that unless otherwise stated, we always assume that our observations are made at sea level. This is not so unnatural for an inhabitant of a large plain region but causes headache to a geodesist working in the Alps or in the Rocky Mountains. We have already been confronted with this situation before, in gravity reduction, and will meet it repeatedly later, most prominently under the heading of Molodensky's problem.

It should also be mentioned that the ellipsoidal azimuth α in (5–102) refers to the actual target, which does not in general lie on the ellipsoid. For the conventional method of computation on the ellipsoid, one wishes the azimuth to refer to a target on the ellipsoid, which is the point at the foot of the normal through the actual target. Furthermore, α refers to what is called a normal section of the ellipsoid, rather than to a geodesic line, which is used in computation. In either case very small azimuth reductions are necessary; since these reductions are purely problems in ellipsoidal geometry, the reader is referred to any appropriate textbook.

Effect of polar motion

The direction of the earth's axis of rotation is not rigorously fixed, neither in space nor with respect to the earth, but undergoes very small, more or less periodic variations. Astronomers know it by the name of *nutation* (with respect to inertial space), geodesists know it by the name of *polar motion* (with respect to the earth's body). This phenomenon arises from a minute difference between the axes of rotation and of maximum inertia, the angle between these axes being about $0.3''$, and is somewhat similar to the precession of a spinning top. This motion of the pole has a main period of about 430 days, the Chandler period, but is rather irregular, presumably because of the movement of masses, atmospheric variations, etc. (Fig. 5.13).

The International Earth Rotation Service (IERS), initially International Latitude Service and then Polar Motion Service, which is maintained by the International Astronomical Union and by the International Union of

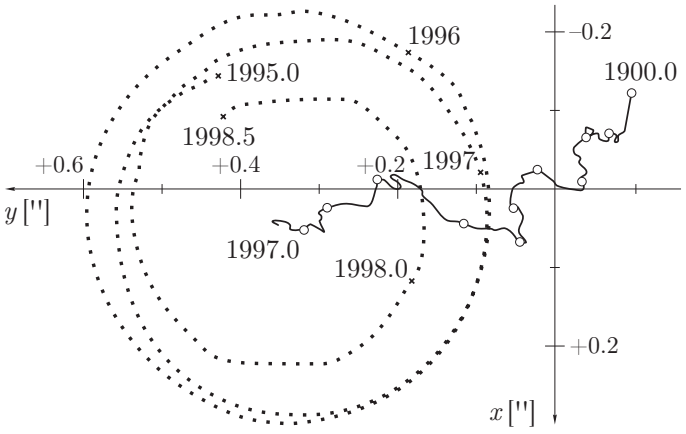

Fig. 5.13. Polar motion: mean pole displacement 1900–1997 (solid line), detailed polar motion 1995–1998 (dotted line)

Geodesy and Geophysics, continuously observes the variation of a number of parameters at a considerable number of stations distributed over the whole earth. Thus, it monitors variations of the earth's axis (polar motion) and of its angular speed of rotation.

The results are published as the rectangular coordinates of the instantaneous pole P_N with respect to a mean pole P_N^0. The astronomically observed values of Φ, Λ, and A naturally refer to the instantaneous pole P_N and must, therefore, be reduced to the mean pole, using the published values of x and y.

This is accomplished by means of the equations

$$\begin{aligned} \Phi &= \Phi_{\text{obs}} - x\,\cos\lambda + y\,\sin\lambda\,, \\ \Lambda &= \Lambda_{\text{obs}} - (x\,\sin\lambda + y\,\cos\lambda)\,\tan\varphi + y\,\tan\varphi_{\text{Gr}}\,, \\ A &= A_{\text{obs}} - (x\,\sin\lambda + y\,\cos\lambda)\,\sec\varphi\,. \end{aligned} \tag{5–103}$$

Now Φ, Λ, A are referred to the mean pole; these values are used in geodesy because they do not vary with time. Longitude, throughout this book, is reckoned positive to the east, as is usual in geodesy; it should be mentioned that in the past literature these formulas are often written for west longitude, according to the former practice of astronomers. Since the correction terms containing x and y are extremely small (of the order of $0.1''$), we may use either the ellipsoidal values φ and λ or the astronomical values Φ and Λ in these terms. The term containing φ_{Gr} (the latitude of Greenwich) in the formula for Λ is usually omitted, so that the mean meridian of Greenwich remains fixed as the conventional *zero meridian*, rather than the astronomical

longitude of Greenwich itself.

These formulas (5–103) are Eqs. (7-13), (7-14), and (7-15) of Moritz and Mueller (1987: pp. 419–420). It is interesting to note the close similarity between the azimuth reduction (5–98) because of the "zenith variation" – that is, the deflection of the vertical – and the longitude reduction of (5–103) because of the polar variation. Actually, the geometry for both cases is the same. The quantities $\xi, \eta, 90^\circ - z, \varphi$ correspond to $x, y, \varphi, \varphi_{\mathrm{Gr}}$; the difference in sign of $\sin\alpha$ and $\sin\lambda$ is due to the fact that, when viewed from the zenith, azimuth is reckoned clockwise and, when viewed from the pole, east longitude is reckoned counterclockwise.

5.13 Reduction of horizontal and vertical angles and of distances

Horizontal angles

To reduce an observed horizontal angle ω to the ellipsoid, we note that every angle may be considered as the difference between two azimuths:

$$\omega = \alpha_2 - \alpha_1 \,. \tag{5–104}$$

Hence, we can apply formula (5–98). In the difference $\alpha_2 - \alpha_1$, the main term $\eta \tan\varphi$ drops out, so that for nearly horizontal lines of sight the whole reduction may be neglected.

Vertical angles

The relation between the measured zenith angle z' and the corresponding ellipsoidal zenith angle z may be given as

$$z = z' + \varepsilon = z' + \xi \cos\alpha + \eta \sin\alpha \,, \tag{5–105}$$

where α is the azimuth of the target.

Spatial distances

Electronic measurement of distance yields straight spatial distances l between two points A and B (Fig. 5.14). These distances may either be used directly for computations in the ellipsoidal coordinate system φ, λ, h, as in "three-dimensional geodesy" (see Sect. 5.9), or they may be reduced to the surface of the ellipsoid to obtain chord distances l_0 or geodesic distances s_0.

We again approximate the ellipsoidal arc A_0B_0 by a circular arc of radius R that is the mean ellipsoidal radius of curvature along A_0B_0. By applying the law of cosines to the triangle OAB, we find

$$l^2 = (R + h_1)^2 + (R + h_2)^2 - 2(R + h_1)(R + h_2) \cos\psi \,. \tag{5–106}$$

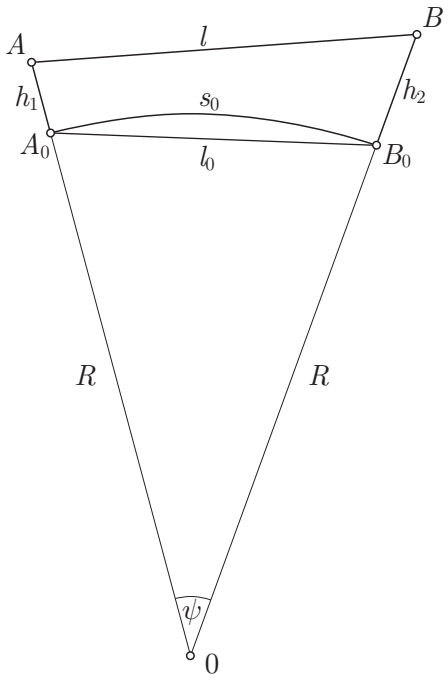

Fig. 5.14. Reduction of spatial distances

With

$$\cos\psi = 1 - 2\sin^2\frac{\psi}{2}\,, \tag{5–107}$$

this is transformed into

$$l^2 = (h_2 - h_1)^2 + 4R^2\left(1 + \frac{h_1}{R}\right)\left(1 + \frac{h_2}{R}\right)\sin^2\frac{\psi}{2}\,; \tag{5–108}$$

and with

$$l_0 = 2R\sin\frac{\psi}{2} \tag{5–109}$$

and the abbreviation $\Delta h = h_2 - h_1$, we obtain

$$l^2 = \Delta h^2 + \left(1 + \frac{h_1}{R}\right)\left(1 + \frac{h_2}{R}\right)l_0^2\,. \tag{5–110}$$

Hence, the chord l_0 and the arc s_0 are expressed by

$$l_0 = \sqrt{\frac{l^2 - \Delta h^2}{\left(1 + \frac{h_1}{R}\right)\left(1 + \frac{h_2}{R}\right)}}\,; \tag{5–111}$$

$$s_0 = R\,\psi = 2R\,\sin^{-1}\frac{l_0}{2R}\,. \tag{5–112}$$

Ellipsoidal refinements of these formulas may be found in Rinner (1956).

As a matter of fact, spatial distances are independent of the vertical. Therefore, the reduction formula (5–111) does not contain the deflection of the vertical ε.

5.14 The astrogeodetic determination of the geoid

Helmert's formula

The shape of the geoid can be determined if the deflections of the vertical are given. *Helmert's formula*

$$dN = -\varepsilon\,ds \tag{5–113}$$

as given in (2–372) is the basic equation (Fig. 5.15). Integrating this relation, we get

$$N_B = N_A - \int_A^B \varepsilon\,ds\,, \tag{5–114}$$

where

$$\varepsilon = \xi\,\cos\alpha + \eta\,\sin\alpha \tag{5–115}$$

is the component of the deflection of the vertical along the profile AB, whose azimuth is α (see Eq. (5–101)).

Formula (5–114) expresses the geoidal undulation as an integral of the vertical deflections along a profile. Since N is a function of position, this integral is independent of the form of the line that connects the points A and B. This line need not necessarily be a geodesic on the ellipsoid, and α may in the general case be variable. In practice, north-south profiles ($\varepsilon = \xi$) or east-west profiles ($\varepsilon = \eta$) are often used. The integral (5–114) is to be evaluated

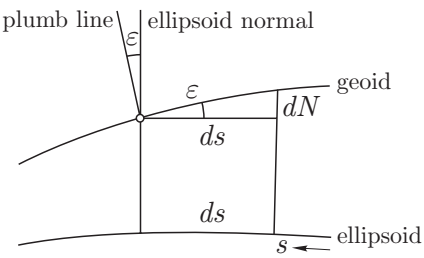

Fig. 5.15. Relation between geoidal undulation and deflection of the vertical

by a numerical or graphical integration. The deflection component ε must be given at enough stations along the profile such that the interpolation between these stations can be done reliably. Sometimes a map of ξ and η is available for a certain area. Such a map is constructed by interpolation between well-distributed stations at which ξ and η have been determined (Grafarend and Offermanns 1975). Then the profiles of integration may be suitably selected; loops may be formed to obtain redundancies which must be adjusted.

If the deflection components ξ and η are obtained directly from the equations

$$\xi = \Phi - \varphi\,, \qquad \eta = (\Lambda - \lambda)\cos\varphi\,, \tag{5–116}$$

that is, by comparing the astronomical and ellipsoidal (or geodetic) coordinates of the same point, then this method is called the *astrogeodetic determination of the geoid*.

The astronomical coordinates are directly observed; the ellipsoidal coordinates are obtained in the following way.

Determination of a local astrogeodetic datum

This is of historic interest only, but indispensible for an understanding of the present classical triangulation system. In agreement with Part I, but in contrast to Part II, "local" again means "regional", referring to a country (e.g., France) or even a continent (e.g., European Datum or North-American Datum). In a larger triangulation system, a certain "initial point" P_1 is chosen for which the undulation N_1 and the components ξ_1 and η_1 of the deflection of the vertical are prescribed. Here ξ_1, η_1, and N_1 may be assumed arbitrarily in principle; the position of the reference ellipsoid with respect to the earth is thereby fixed. For the sake of definiteness let us consider the case that has been of greatest practical importance, that is, the case in which $\xi_1 = \eta_1 = N_1 = 0$. In this case, because $\xi_1 = \eta_1 = 0$, the geoid and the ellipsoid have the same surface normal so that, because $N_1 = 0$, the ellipsoid is tangent to the geoid below P_1 (Fig. 5.16). The condition that the axis of the reference ellipsoid be parallel to the earth's axis of rotation finally determines the orientation of the triangulation net because Laplace's equation (5–99) then gives $\Delta\alpha_1 = \eta_1 \tan\varphi_1 = 0$, so that $\alpha_1 = A_1$; that is, at the initial point the ellipsoidal azimuth is equal to the astronomical azimuth.

Now we can reduce the measured distances and angles to the ellipsoid and compute on it the position of the points of the triangulation net (their ellipsoidal coordinates φ and λ) in the usual way. After measuring the coordinates Φ and Λ astronomically at the same points, we can then compute the deflection components ξ and η by (5–116). Starting from the assumed value

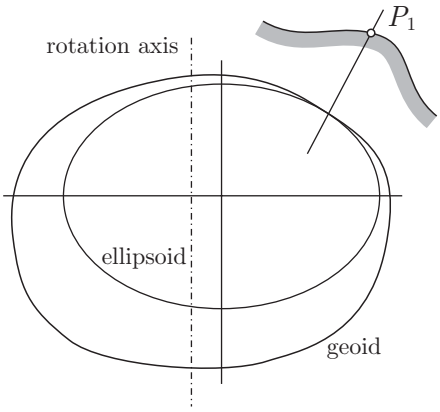

Fig. 5.16. The reference ellipsoid is tangent to the geoid at P_1

N_1 at the initial point P_1 (in our case, $N_1 = 0$), we can finally compute the geoidal heights N of any point of the triangulation net by repeated application of (5–114). These geoidal heights refer to the ellipsoid that was fixed by prescribing ξ_1, η_1, N_1, and, of course, its semimajor axis a and its flattening f. To employ a frequently used term, they refer to the given *astrogeodetic datum* $(a, f; \xi_1, \eta_1, N_1)$.

By means of N and the orthometric height H, the height h above the ellipsoid is obtained via $h = H + N$, so that the rectangular spatial coordinates X, Y, Z can be computed by (5–27). But unless ξ and η are absolute (geocentric) deflections, the origin of the coordinate system will not be at the center of the earth (see Sect. 5.7).

A flaw in the procedure described above apparently is that N, ξ, η are already needed for the reduction of the measured angles and distances to the ellipsoid. However, for this purpose approximate values of N, ξ, η are sufficient. These are obtained by performing the process just explained with unreduced angles and distances. We can also get suitable values for N, ξ, η in other ways, for instance, by Stokes' formula.

Use and misuse of Laplace's equation

It should be mentioned that in practice the component η has been often obtained from azimuth measurements using (5–102) in rearranged form, that is,

$$\eta = (A - \alpha) \cot \varphi \,, \tag{5–117}$$

because astronomical measurements of azimuth are simpler than those of longitude. This is a misuse which may lead to a systematic distortion of the

net. Longitude and azimuth are often measured at the same point. Then Laplace's condition

$$\Delta\alpha = \Delta\lambda \sin\varphi \tag{5–118}$$

furnishes an important check on the correct orientation of the net and forces the axis of the ellipsoid to be parallel with the earth's axis of rotation. Thus it may be used for adjustment purposes. Astronomical stations with longitude and azimuth observations are, therefore, called *Laplace stations*. For these purposes, the measuring accuracy of astronomical field observations is sufficient, in contrast to the use for directly determining horizontal positions by $\varphi = \Phi - \xi$, etc. in Sect. 2.21.

The astrogeodetic determination of the geoid, also called *astronomical leveling*, was known to Helmert (1880) and even before.

Comparison with the Stokes method

It is quite instructive to compare Helmert's formula

$$N = N_A - \int_A^B \varepsilon \, ds \tag{5–119}$$

for the astrogeodetic method with Stokes' formula

$$N = \frac{R}{4\pi\,\gamma_0} \iint_\sigma \Delta g \; S(\psi) \; d\sigma \tag{5–120}$$

for the gravimetric method. Both methods use the gravity vector $\mathbf{g}$. It is compared with a normal gravity vector $\boldsymbol{\gamma}$. The components $\xi = \Delta\varphi$ and $\eta = \Delta\lambda\cos\varphi$ of the deflection of the vertical represent the differences in *direction*, and the gravity anomaly Δg represents the difference in *magnitude* of the two vectors. Helmert's formula determines the geoidal undulation N from ξ and η, that is, by means of the direction of $\mathbf{g}$, and Stokes' formula determines N from Δg, that is, by means of the magnitude of $\mathbf{g}$. Both formulas are somewhat similar: they are integrals which contain ε, or ξ and η, and Δg in linear form.

Otherwise, the two formulas show marked differences which are characteristic for the respective method. In Helmert's formula, the integration is extended over part of a profile; thus, it is sufficient to know the deflection of the vertical in a limited area. The position of the reference ellipsoid with respect to the earth's center of gravity is unknown, however, and can be determined only by means of the gravimetric method or, more practically, the analysis of satellite orbits (Sect. 7.2). Furthermore, the astrogeodetic method can be used only on land, because the necessary measurements are impossible at sea.

In Stokes' formula, however, the integration should be extended over the whole earth. The gravity anomaly Δg must be known all over the earth; however, accurate gravity measurements at sea are possible. The gravimetric method yields, for the whole earth, absolute geoidal undulations: the center of the reference ellipsoid coincides with the center of the earth. Nowadays, this is only a theoretical possibility because the required complete coverage of the whole earth is not available; again, GPS helps. Nevertheless, the gravimetric method is still basic: it furnishes, not the geocenter, but details of the geoid, together with the astrogeodetic method!

The astrogeodetic method has often been applied to the determination of geoidal sections. We mention, because of its pioneering character and its romantic title, "Das Geoid im Harz" by Galle (1914). In the years following 1970 it is becoming rare to use Helmert's integral formula in its original form, and deflections of the vertical are more and more combined with other data (gravity, GPS, and other satellite data) for a uniform determination of geoid and gravity field (see Chaps. 10 and 11).

Adjustment of nets of astrogeodetic geoidal heights

With a sufficiently dense net of astrogeodetic stations (preferably Laplace points) with an average station distance of 10–20 km, the Helmert integral (5–119) can be approximated by

$$\Delta N_{AB} \equiv N_B - N_A = -\int_A^B \varepsilon \, ds = -\frac{\varepsilon_A + \varepsilon_B}{2} \int_A^B ds \qquad (5\text{–}121)$$

or

$$\Delta N_{AB} = -\frac{\varepsilon_A + \varepsilon_B}{2} \, s_{AB} \, . \qquad (5\text{–}122)$$

Thus, the undulation difference can be computed for the line AB, and similarly for other lines BC and CA in the triangle ABC (Fig. 5.17). The closure condition

$$\Delta N_{AB} + \Delta N_{BC} + \Delta N_{CA} = 0 \qquad (5\text{–}123)$$

must be satisfied and imposed as a condition in the least-squares adjustment

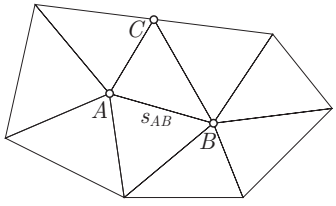

Fig. 5.17. Triangular net for an astrogeodetic geoid

of the net. Accordingly, the other triangles can be computed as in any other height network (e.g., leveling net).

It is curious that it may be shown that such closures are mathematically equivalent to the well-known relation

$$\frac{\partial^2 N}{\partial x\, \partial y} = \frac{\partial^2 N}{\partial y\, \partial x}\,. \tag{5–124}$$

See also Sect. 4.5.

5.15 Reduction for the curvature of the plumb line

Motivation

The astronomical coordinates Φ and Λ, as observed on the surface of the earth, are not rigorously equal to their corresponding values at the geoid because the plumb line, the line of force, is not straight, or in other words, because the level surfaces are not parallel. Thus, if we wish our astronomical coordinates to refer to the geoid, we must reduce our observations accordingly.

Examples of such cases are the following:

1. The gravimetric deflections have usually been computed by Vening Meinesz' formula for the geoid, so that either the gravimetric deflections must be reduced upward to the ground point or the astronomical observations must be reduced downward to the geoid, in order to make the two quantities comparable.

2. If astronomical observations are used for the determination of the geoid, the same reduction, in principle, must be applied.

Important remark

The principle of reduction of the plumb line is of fundamental theoretical importance for understanding the geometry of the earth's gravity field. In practice, it is usually disregarded if the topography is sufficiently flat, or replaced by more sophisticated methods in mountainous areas, as we shall see later (Sects. 8.12 and 8.13). The present section may be skimmed at first reading, except for the *normal curvature of the plumb line* at its very end.

Principles

Consider the projection of the plumb line onto the meridian plane. According to the well-known definition of the curvature of a plane curve, the angle

between two neighboring tangents of this projection of the plumb line is

$$d\varphi = -\kappa_1 \, dh \,, \tag{5–125}$$

where the minus sign is conventional and the curvature κ_1 is given by (2–50):

$$\kappa_1 = \frac{1}{g} \frac{\partial g}{\partial x} \,. \tag{5–126}$$

The x-axis is horizontal and points northward. Hence, the total change of latitude along the plumb line between a point on the ground, P, and its projection onto the geoid, P_0, is given by

$$\delta\varphi = \int_{P_0}^{P} d\varphi = -\int_{P_0}^{P} \kappa_1 \, dh \tag{5–127}$$

or

$$\delta\varphi = -\int_{P_0}^{P} \frac{1}{g} \frac{\partial g}{\partial x} \, dh \,. \tag{5–128}$$

Using κ_2 of (2–51), we similarly find for the change of longitude

$$\delta\lambda \, \cos\varphi = -\int_{P_0}^{P} \frac{1}{g} \frac{\partial g}{\partial y} \, dh \,, \tag{5–129}$$

where the y-axis is horizontal and points eastward.

Alternative formulas

There is a close relationship between the curvature reduction of astronomical coordinates and the orthometric reduction of leveling, considered in Sect. 4.3.

The orthometric correction $d(\mathrm{OC})$ has been defined as the quantity that must be added to the leveling increment dn in order to convert it into the orthometric height difference dH:

$$d(\mathrm{OC}) = dH - dn \,. \tag{5–130}$$

From Fig. 5.18, we see that, for a north-south profile, the curvature reduction and the orthometric correction are related by the simple formula

$$\delta\varphi = \frac{\partial(\mathrm{OC})}{\partial x} \,. \tag{5–131}$$

Similarly, we find

$$\delta\lambda \, \cos\varphi = \frac{\partial(\mathrm{OC})}{\partial y} \,. \tag{5–132}$$

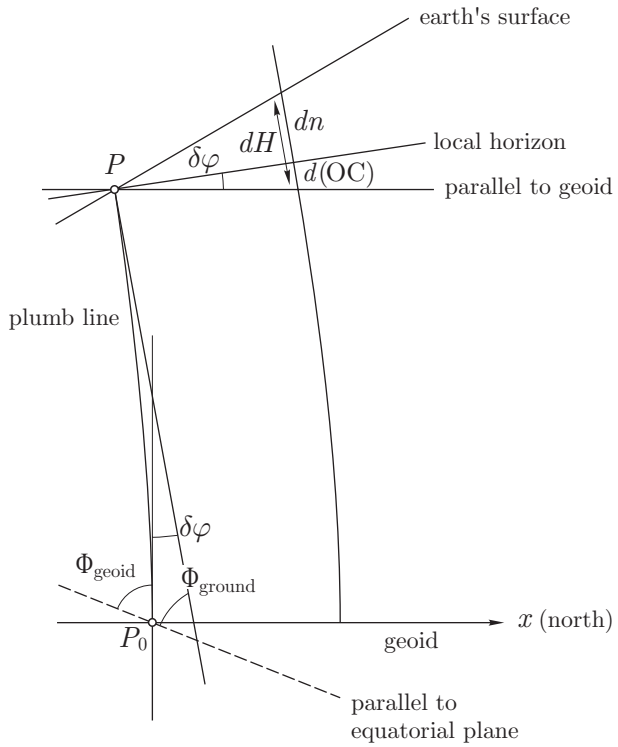

Fig. 5.18. Plumb-line curvature and orthometric correction

According to Sect. 4.3, we have

$$dC = g\,dn = -dW\,, \qquad H = \frac{C}{\bar{g}}\,. \tag{5–133}$$

Hence, (5–130) becomes

$$d(\mathrm{OC}) = dH - \frac{1}{g}\,dC = dH + \frac{1}{g}\,dW\,, \tag{5–134}$$

so that

$$\begin{aligned} \delta\varphi &= \frac{\partial H}{\partial x} + \frac{1}{g}\,\frac{\partial W}{\partial x}\,, \\ \delta\lambda\,\cos\varphi &= \frac{\partial H}{\partial y} + \frac{1}{g}\,\frac{\partial W}{\partial y}\,. \end{aligned} \tag{5–135}$$

These equations relate the reduction for the curvature of the plumb line to the orthometric height H and the potential W. In view of the irregular shape of the plumb lines, it is remarkable that such simple general relations as (5–131), (5–132), and (5–135) exist.

These relations may be used to find computational formulas for the curvature reductions $\delta\varphi$ and $\delta\lambda$. We have

$$\begin{aligned} d(\mathrm{OC}) &= dH - \frac{dC}{g} = d\left(\frac{C}{\bar{g}}\right) - \frac{dC}{g} \\ &= \frac{dC}{\bar{g}} - \frac{C}{\bar{g}^2}\,d\bar{g} - \frac{dC}{g} = -\frac{C}{\bar{g}^2}\,d\bar{g} + \frac{g-\bar{g}}{\bar{g}}\,\frac{dC}{g} \end{aligned} \tag{5–136}$$

or

$$d(\mathrm{OC}) = -\frac{H}{\bar{g}}\,d\bar{g} + \frac{g-\bar{g}}{\bar{g}}\,dn\,. \tag{5–137}$$

By substituting this into (5–131) and (5–132), we obtain

$$\begin{aligned} \delta\varphi &= -\frac{H}{\bar{g}}\,\frac{\partial\bar{g}}{\partial x} + \frac{g-\bar{g}}{\bar{g}}\,\tan\beta_1\,, \\ \delta\lambda\,\cos\varphi &= -\frac{H}{\bar{g}}\,\frac{\partial\bar{g}}{\partial y} + \frac{g-\bar{g}}{\bar{g}}\,\tan\beta_2\,, \end{aligned} \tag{5–138}$$

where we have set

$$\tan\beta_1 = \frac{\partial n}{\partial x}\,, \qquad \tan\beta_2 = \frac{\partial n}{\partial y}\,, \tag{5–139}$$

so that β_1 and β_2 are the angles of inclination of the north-south and east-west profiles with respect to the local horizon; $\bar{g}$ is the mean value of gravity between the geoid and the ground. In these formulas, we need only this mean value $\bar{g}$, together with its horizontal derivatives, and the ground value g, whereas in (5–128) and (5–129), we must know the horizontal derivatives of gravity all along the plumb line. The detailed shape of the plumb lines does not directly enter into (5–138) as it does into (5–128) and (5–129).

The mean value $\bar{g}$ is found by a Prey reduction of the measured gravity g. In order that the numerical differentiations $\partial g/\partial x$ and $\partial g/\partial y$ give reliable results, a dense gravity net around the station is necessary, and the Prey reduction must be performed carefully. The inclination angles β_1 and β_2 are taken from a topographical map.

The sign of these corrections may be found in the following way. If g decreases in the x-direction, then formulas (5–128) and (5–138) give $\delta\varphi > 0$ and Fig. 5.18 shows that Φ at P_0 is then greater than at P. The same holds for Λ, so that we have

$$\begin{aligned} \Phi_{\mathrm{geoid}} &= \Phi_{\mathrm{ground}} + \delta\varphi\,, \\ \Lambda_{\mathrm{geoid}} &= \Lambda_{\mathrm{ground}} + \delta\lambda\,. \end{aligned} \tag{5–140}$$

Integrated form

In formula (5–114), the deflection components ξ and η refer to the geoid. This means that the astronomical observations of Φ and Λ must be reduced to the geoid.

It is also possible and often more convenient to apply this correction for plumb-line curvature not to the astronomical coordinates Φ and Λ but to the geoidal height differences computed from the unreduced deflection components.

These N values, denoted by N', are obtained by using in (5–116) the directly observed Φ and Λ, which define the direction of the plumb line at the station P (Fig. 5.19). The notation N will be reserved for the correct geoidal heights. Then we read from Fig. 5.19:

$$dh = dN + dH = dN' + dn\,, \tag{5–141}$$

where h is the geometric height above the ellipsoid. Thus, we see that the difference between the unreduced and the correct element of geoidal height,

$$dN' - dN = dH - dn = d(\mathrm{OC})\,, \tag{5–142}$$

is equal to the difference between the element dH of orthometric height and

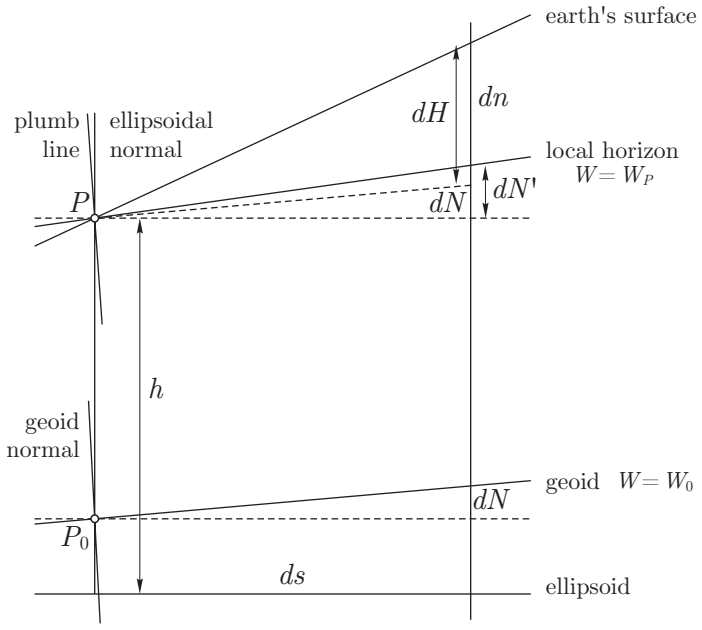

Fig. 5.19. Reduction of astronomical leveling

the leveling increment dn, which is the orthometric reduction $d(\mathrm{OC})$. Thus,

$$N_B - N_A = N'_B - N'_A - \mathrm{OC}_{AB}\,, \tag{5–143}$$

so that we can immediately apply Eq. (4–46):

$$N_B - N_A = -\int_A^B \varepsilon\, ds - \int_A^B \frac{g-\gamma_0}{\gamma_0}\, dn + \frac{\bar{g}_B - \gamma_0}{\gamma_0}\, H_B - \frac{\bar{g}_A - \gamma_0}{\gamma_0}\, H_A\,, \tag{5–144}$$

where γ_0 is our usual constant γ_{45°; the deflection components ε are computed from the observed ground values Φ and Λ by (5–116) and (5–115). These ideas go back to Helmert, but they are hardly used anymore.

Curvature of the normal plumb line

If, instead of the actual gravity g, the normal gravity γ is applied for the computation of the plumb-line curvature, we find, using

$$\gamma = \gamma_a \left(1 + f^* \sin^2\varphi - \frac{2}{a}\, h\, \cdots\right)\,, \tag{5–145}$$

that

$$\begin{aligned} &\frac{\partial\gamma}{\partial x} \doteq \frac{1}{R}\,\frac{\partial\gamma}{\partial\varphi} \doteq \frac{2\gamma_a}{R}\, f^* \sin\varphi\, \cos\varphi \doteq \frac{2\gamma}{R}\, f^* \sin\varphi\, \cos\varphi\,, \\ &\frac{\partial\gamma}{\partial y} \doteq \frac{1}{R\cos\varphi}\,\frac{\partial\gamma}{\partial\lambda} = 0\,. \end{aligned} \tag{5–146}$$

Hence, the integrand $(1/\gamma)(\partial\gamma/\partial x)$ in (5–128) does not depend on h, so that the integration can be performed immediately. We find

$$\begin{aligned} &\delta\varphi_{\mathrm{normal}} = -\frac{f^*}{R}\, h\, \sin 2\varphi = -0.17''\, h_{\,[\mathrm{km}]}\, \sin 2\varphi\,, \\ &\delta\lambda_{\mathrm{normal}} = 0\,. \end{aligned} \tag{5–147}$$

The curvature of the normal plumb line in the east-west direction is zero, owing to the rotational symmetry of the ellipsoid of revolution. The *normal reduction* (5–147) *is very simple and practically important*, see especially Sect. 8.13.

5.16 Best-fitting ellipsoids and the mean earth ellipsoid

We define the mean earth ellipsoid physically as that ellipsoid of revolution which shares with the earth the mass M, the potential W_0, the difference

between the principal moments of inertia $G(C - \bar{A})$, where $\bar{A} = (A + B)/2$, and the angular velocity ω.

It is also possible to define the mean earth ellipsoid geometrically as that ellipsoid which approximates the geoid most closely. This definition is perhaps more appealing to the geodesist; it may, for instance, be formulated by the condition that the sum of the squares of the deviations N of the geoid from the ellipsoid be a minimum:

$$\iint_{\sigma} N^2 \, d\sigma = \text{ minimum} \tag{5–148}$$

(this integral is to be considered the limit of a sum). The condition of closest approximation may also be expressed in terms of the deflections of the vertical:

$$\iint_{\sigma} (\xi^2 + \eta^2) \, d\sigma = \text{ minimum}\,, \tag{5–149}$$

minimizing the sum of the squares of the total deflection of the vertical

$$\vartheta = \sqrt{\xi^2 + \eta^2}\,. \tag{5–150}$$

Many other similar definitions of closest approximation are possible.

The first definition, based on the condition (5–148), is the most plausible and the most appropriate intuitively, as has been already noted by Helmert; in principle, however, all definitions are more or less conventional and are equivalent theoretically as we shall see below.

The second definition, based on the condition (5–149), uses deflections of the vertical and is, thus, particularly well adapted to the astrogeodetic method. However, since this method can be applied only over limited areas, at most spanning the continents, the integral (5–149) must be replaced by a sum covering the astronomical stations of a restricted region:

$$\sum (\xi^2 + \eta^2) \, d\sigma = \text{ minimum}\,. \tag{5–151}$$

In this way, we can get only the best-fitting ellipsoid for the region considered, rather than a general earth ellipsoid. As Fig. 5.20 indicates, a *locally best-fitting ellipsoid* may be quite different from the mean earth ellipsoid, which can be considered a best-fitting ellipsoid for the whole earth.

If a reasonably good approximation of the earth ellipsoid by a local best-fitting ellipsoid is desired, it is advisable to subtract the effect of the topography and of its isostatic compensation from the astrogeodetic deflections of the vertical before the minimum condition (5–151) is applied. The purpose of this procedure is to smooth the irregularities of the geoid. In this way,

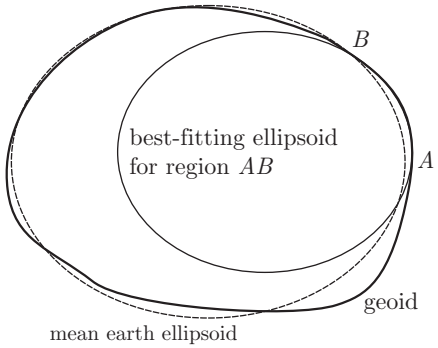

Fig. 5.20. A locally best-fitting ellipsoid and the mean earth ellipsoid

Hayford computed the international ellipsoid as ellipsoid that best fits the isostatically reduced vertical deflections in the United States. Rapp (1963) made an interesting recomputation.

Please note: Don't use formula (5–151) in spite of its historical importance: the determination of local best-fitting ellipsoids is hopelessly obsolete now!

The previously described method is impaired by unknown density anomalies and by the lack of complete isostatic compensation. Therefore, it is better to go still one step further and subtract the gravimetrically computed values ξ^g, η^g from the astrogeodetic deflections ξ^a, η^a. Then the minimum condition

$$\sum \left[(\xi^a - \xi^g)^2 + (\eta^a - \eta^g)^2 \right] = \text{ minimum} \qquad (5\text{–}152)$$

results. Thus, we may say that Hayford's method is equivalent to the use of (5–152), the gravimetric values ξ^g, η^g being approximated by deflections that represent the effect of topography and of its isostatic compensation only. If the isostatic compensation were complete, and if we had perfect knowledge of the density above the geoid, both methods would give exactly the same result if applied properly.

Equivalence of different definitions of the earth ellipsoid

It is quite remarkable that the minimum definitions (5–148) or (5–149) and a similar definition due to Rudzki, using the condition

$$\iint_\sigma (\Delta g)^2 \, d\sigma = \text{ minimum}\,, \qquad (5\text{–}153)$$

yield results which, to the usual spherical approximation, are identical with each other and with the physical definition in terms of M, W_0, $C - \bar{A}$, and

ω. This can be seen as follows. We write the spherical-harmonic expansion of the disturbing potential in the form

$$T = \frac{G\,\delta M}{R} + \sum_{n=1}^{\infty}\sum_{m=0}^{n}\left[a_{nm}R_{nm}(\vartheta,\lambda) + b_{nm}S_{nm}(\vartheta,\lambda)\right]. \tag{5–154}$$

Then, according to Sect. 2.17, Eqs. (2–351) and (2–359) or (2–363), we have

$$N = \frac{G\,\delta M}{R\,\gamma_0} - \frac{\delta W}{\gamma_0} + \frac{1}{\gamma_0}\sum_{n=1}^{\infty}\sum_{m=0}^{n}\left[a_{nm}R_{nm}(\vartheta,\lambda) + b_{nm}S_{nm}(\vartheta,\lambda)\right] \tag{5–155}$$

and

$$\Delta g = -\frac{G\,\delta M}{R^2} + \frac{2\delta W}{R} + \frac{1}{R}\sum_{n=1}^{\infty}\sum_{m=0}^{n}\left[(n-1)\,a_{nm}R_{nm}(\vartheta,\lambda) + (n-1)\,b_{nm}S_{nm}(\vartheta,\lambda)\right]; \tag{5–156}$$

remember that γ_0 denotes a global mean value of gravity. The condition of equal masses, $\delta M = 0$, is very natural and will be assumed. If we square the formulas for N and Δg and integrate over the whole earth, then all the integrals of products of different harmonics R_{nm} and S_{nm} will be zero, according to the orthogonality property (1–83), and the remaining integrals will be given by (1–84). Thus, we find

$$\iint_{\sigma} N^2\,d\sigma = \frac{4\pi}{\gamma_0^2}\,\delta W^2 + \frac{4\pi}{\gamma_0^2}\sum_{n=1}^{\infty}\frac{1}{2n+1}\left[a_{n0}^2 + \sum_{m=1}^{n}\frac{(n+m)!}{2(n-m)!}\left(a_{nm}^2 + b_{nm}^2\right)\right], \tag{5–157}$$

$$\iint_{\sigma} (\Delta g)^2\,d\sigma = \frac{16\pi}{R^2}\,\delta W^2 + \frac{4\pi}{R^2}\sum_{n=1}^{\infty}\frac{(n-1)^2}{2n+1}\left[a_{n0}^2 + \sum_{m=1}^{n}\frac{(n+m)!}{2(n-m)!}\left(a_{nm}^2 + b_{nm}^2\right)\right]. \tag{5–158}$$

By a more complicated derivation, which we omit here but which can be found in Molodenskii et al. (1962: p. 87), one gets the similar formula

$$\iint_{\sigma} (\xi^2+\eta^2)\,d\sigma = \frac{4\pi}{R^2\,\gamma_0^2}\sum_{n=1}^{\infty}\frac{n\,(n+1)}{2n+1}\left[a_{n0}^2 + \sum_{m=1}^{n}\frac{(n+m)!}{2(n-m)!}\left(a_{nm}^2 + b_{nm}^2\right)\right]. \tag{5–159}$$

Varying the size and shape of the reference ellipsoid and its position with respect to the earth changes only the coefficients δW, a_{10}, a_{11}, b_{11}, and a_{20}, leaving the other coefficients practically invariant. Thus, the minimum of any of the integrals (5–157), (5–158), (5–159) is obtained if all these coefficients are equal to zero. Now, $\delta W = 0$ means equal potential $U_0 = W_0$; $a_{10} = a_{11} = b_{11} = 0$ means absolute position (coincident centers of gravity); and $a_{20} = 0$ means equality of J_2 or of $C - (A + B)/2$.

Therefore, the equivalence of the physical definition by means of M, W_0, $C - \bar{A}$, ω and of the condition of closest approximation in any of the forms (5–148), (5–149), or (5–153) has been established. (It may be noted that (5–158) contains no first-degree term, because of the factor $(n - 1)^2$, and that (5–159) contains no term of degree zero, so that these equations do not determine the missing terms.)

Best-fitting ellipsoid and World Geodetic System

It should be remembered, however, that the mean earth ellipsoid, defined in this manner, is not necessarily the best reference surface for practical geodetic purposes. It is essentially defined empirically by means of empirical determinations of GM, W_0, etc. Its parameters will change with every improvement in the quality or the number of the relevant measurements (gravity, distances, etc.). Since an enormous amount of numerical data is based on an assumed reference ellipsoid, it would be highly impractical to change it very often, for this would involve repeated transformations of all the data. It is much better to use a fixed reference ellipsoid with rigidly assumed parameters, which can be more or less arbitrary if only they give a reasonably good approximation. In this respect, the Geodetic Reference System 1980 is still (2005) perfectly acceptable.

A certain amount of conflict exists between the interests of geodesists and astronomers regarding the earth ellipsoid. The geodesist needs a permanent reference surface, whereas the astronomer wants the best approximation of the earth by an ellipsoid. A good compromise is to use a fixed geodetic reference ellipsoid, but from time to time to compute the "best" corrections to the assumed parameters for astronomical and other purposes. This has been the practice of the IAG since 1974.