Computational methods for contact stress problems with normal and tangential loading

Christopher R. McGoldrick

Follow this and additional works at: http://scholarworks.rit.edu/theses

Recommended Citation

This Thesis is brought to you for free and open access by the Thesis/Dissertation Collections at RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritscholarworks@rit.edu.
Computational Methods for Contact Stress Problems with Normal and Tangential Loading

Christopher R. McGoldrick

Master of Science Thesis
1991

Rochester Institute Of Technology
College of Engineering
Department of Mechanical Engineering

Approvals:

Joseph Torok - Thesis Advisor

Richard Budynas

Richard Hetnarski

Mark Kempski

Charles Haines
Computational Methods for Contact Stress Problems
with Normal and Tangential Loading

I
Christopher R. McGoldrick
hereby grant permission
to
Wallace Memorial Library
of
Rochester Institute of Technology
to reproduce my thesis
in whole or part.
Any such reproduction will not be for
commercial use or for profit.
Thesis Abstract:

Computational Methods for Contact Stress Problems with Normal and Tangential Loading.

Christopher R. McGoldrick
Master of Science Mechanical Engineering
Rochester Institute Of Technology
Rochester, New York

January, 1991

This investigation contains an historical overview of contact stresses including various classifications of the problem and their associated solutions. The solution for the normal loading elliptical contact problem is reviewed, as formulated using classic methods of linear elasticity theory. An efficient computational method is developed to evaluate the elliptic integrals that arise. The limitations of this solution are investigated in detail and it is shown how the method could be extended to the sliding elliptical contact problem. Practical applications of contact stresses are included with an emphasis on mechanical design. Differences between the continuum model and the behavior of real materials are discussed.
# Contents

1. **Introduction** ......................................................... 1  
   1.1. What are contact stresses? ........................................... 1  
   1.2. Why they are important ............................................. 1  
   1.3. Nature of the Applied Loading .................................... 1  
   1.4. The Continuum Model .............................................. 3  
   1.5. The Elastic Foundation Model ................................. 5  
   1.6. Coordinate System For Non-conforming Single Point Contact 5  
   1.7. Milestones in Contact Stress Development .................. 7  
   1.8. The Proposed Computational Method .......................... 8  

2. **Classification of Contact Stress Problems** .......... 10  
   2.1. Relevant Factors .................................................. 10  
   2.2. The Rigid Punch and Other Stiff Geometries ................. 10  
   2.3. Elastic Half-Spaces .............................................. 11  
   2.4. Rolling Contact ................................................... 12  
   2.5. Plastic Deformation .............................................. 13  
   2.6. Conforming Contact ............................................. 14  
   2.7. Numerical Methods .............................................. 15  
   2.8. Hertzian Contact .................................................. 16  
      2.8.1. Elliptical Contact: "The General Case" ................. 16  
      2.8.2. Limitations or Bounds of the Hertz Theory .......... 17  
      2.8.3. Extension of Hertzian Contact .......................... 17  

3. **The Geometry of Contact** ....................................... 19  
   3.1. Local Surface Geometry .......................................... 19  
   3.2. Combinations of Bodies ......................................... 22  
   3.3. Magnitude Checks .............................................. 23  

4. **Elasticity Methods** ............................................... 27  
   4.1. Equilibrium ....................................................... 27  
   4.2. Strain at a Point ................................................ 28  
   4.3. Compatibility Equations for Linear Strain ................. 29  
   4.4. Hooke's Law ...................................................... 30  
   4.5. Stress Functions ................................................ 30  
   4.6. Determination of Displacements ............................... 31
5. Axisymmetric Solids of Revolution
   5.1. Stresses and Deformations in a Solid of Revolution ..................... 33
   5.2. Concentrated Normal Force ........................................... 36
   5.3. Distributed Normal Loads ............................................. 38
   5.4. Spherical Contact ..................................................... 41
   5.5. Cylindrical Contact - Another Two-Dimensional Case .................. 48

6. Elliptical Contact
   6.1. Surface Geometry for Elliptical Contact ................................ 51
   6.2. Development of Equations for Elliptical Contact ...................... 55
   6.3. Elliptical Solution Overview ......................................... 57
   6.4. An Elliptic Integral Primer ........................................... 59
       6.4.1. Complete Elliptic Integrals .................................... 60
       6.4.2. Incomplete Elliptic Integrals ................................... 62
   6.5. Explicit Numerical Procedures for Normally Loaded Elliptical
        Contact. .................................................................. 63
   6.6. Application of the 10-Step Method to Determine Maximum Stress
        Values. .................................................................. 69
   6.7. Contact Stresses 101 (Benchmark Testing) ................................ 73
   6.8. The Real Contact Problem ................................................ 76
       6.8.1. Surface Displacements ............................................. 77
       6.8.2. Surface Stresses ...................................................... 77

7. Practical Applications
   7.1. Gaussian Profiles and the Rusty Nail ..................................... 79
       7.1.1. On Surfaces and Their Statistical Descriptions .................. 79
       7.1.2. Measurement Techniques ............................................ 81
       7.1.3. Mathematical Surface Models ...................................... 82
       7.1.4. True Contact Area ..................................................... 84
   7.2. The Real Structure of the Surface Layers of Metallic Bodies ........ 84
   7.3. Friction Between Real Surfaces .......................................... 85
       7.3.1. The Classic Laws of Friction ....................................... 85
       7.3.2. Theories of Friction .................................................. 86
       7.3.3. Rough Surfaces and Friction ....................................... 87
       7.3.4. Design Tips for Friction Calculations ............................. 88
   7.4. Lubrication and Pitting Failure .......................................... 89
   7.5. Impact on a Worn Surface ................................................ 90
   7.6. Rolling Contact and Shakedown .......................................... 91
   7.7. Allowable Stresses ....................................................... 92
   7.8. Heat Treating and Contact Stresses ..................................... 92
       7.8.1. Nitriding ............................................................... 93
       7.8.2. Flame Hardening ..................................................... 93
7.9. Gears ................................................................. 95
7.10. Camrollers .................................................... 96
7.11. Wheel and Rail .............................................. 97

8. **Summary** .......................................................... 98
   8.1. On the use of the Continuum Model to Describe Real Materials .... 98
   8.2. On the Elliptical Solution Implemented in this Investigation ......... 98
   8.3. Tangential Loading for Elliptical Contact ............................. 99
   8.4. Direction for Future Work ....................................... 99

9. **A Personal Note** ............................................... 101

10. **References** .................................................... 105
1. Introduction

1.1. What are contact stresses?

Contact stresses are caused by the pressure of one solid on another over limited areas of contact. They are merely the localized stress distributions at and near the surface of the body. Contrast this to the axial stress on a cylindrical tensile test specimen in which the engineering stress represents the bulk behavior. Even there, the stress distribution normal to the cross-section is not uniform when necking starts to occur. The stresses on the surface of a body outside the contact area(s), where there are no applied tractions, must be zero for equilibrium. This is a boundary condition for some solutions. Stresses inside the contact area are merely the internal reaction of one body accommodating the intrusion of another.

The point where the force is applied moves a small amount in response to the load. This occurs because both the supporting and contacting materials are not rigid or fixed in space, but have finite stiffnesses associated with them. As a result, the bodies are not necessarily fixed in their geometry over time. Contact stresses are localized reactions, so that Saint Venant’s Principle applies to their analyses. The study of contact stresses is often avoided because the mathematics is difficult and does not lead to solutions in terms of formulas which can conveniently be used by designers.

1.2. Why they are important.

The premature failure of many machine elements can be attributed to excessively high contact stresses. In some cases, the maximum stress due to contact between members is regarded as a limiting criterion of design. Inaccurate calculation or neglect of their effects is a common design flaw. Some classic contact stress problems involve ball bearings, cylindrical rollers, gear teeth, and wheels on a track. The loading in these cases is often cyclic. Repeated loading is related to microscopic failure phenomena such as cracking, pitting, and subsurface fracture.

1.3. Nature of the Applied Loading

The method used to determine contact stresses depends on the initial contact geometry and the applied load. Limited cases of closed-form solutions are available, some of which will be reviewed in this investigation. Two-dimensional solutions can often be reduced to simple formulas whereas integral representations are required for the general three-dimensional case.
Special consideration is given to computation of stress inside the area of contact. The normal and tangential stress distributions are typically calculated independently and then superimposed. The normal force \( N \) is traditionally assumed to be directly proportional to the friction or tangential force \( F \), with the constant of proportionality being the coefficient of friction \( \mu \), that is

\[
F = \mu N. \tag{1.1a}
\]

The symbols \( P \) for normal force and \( Q \) for tangential force are commonly used in the contact mechanics literature and will be used here. Equation 1.1a which follows the notation found in introductory physics and statics textbooks, is an oversimplification of the relation between normal and tangential loading. In reality the normal and tangential loads, and the coefficient of friction, are distributed functions over the contact area so the relationship between them is more adequately represented by

\[
Q(x,y) = \mu(x,y) P(x,y). \tag{1.1b}
\]

If adhesive forces at the interface are neglected, the presence of any tangential load must be associated with a normal load. For analysis, it is assumed that there is no interaction or coupling between the normal and tangential force distributions other than that given by Equation 1.1b. Including tangential loading in the analysis is necessary for a realistic model of the contact between machine elements, but it significantly increases the complexity of the solutions such that a computer is required.

The contact geometry also affects the form of the pressure distribution between the bodies as in the cases of line, spherical, and elliptical contact which have uniform, spherical, and ellipsoidal distributions, respectively. Contact stresses, like all stresses, have a three-dimensional character (See Figure 1.1). Spherical and cylindrical contact are often approximated as two-dimensional problems to simplify the analysis and yield insight into the overall behavior. Long rollers in contact (plane strain) or thin disks (plane stress), will in reality experience end-effects that will create three-dimensional stress distributions. The presence of friction, which breaks any symmetry of the problem that was due to the geometry, also necessitates a three-dimensional treatment.

\[\text{Figure 1.1 Differential stress element for three dimensions.}\]
1.4. The Continuum Model

![Diagram of material models and relationships](image)

Figure 1.2 Family of material models. Branches of mechanics are not a clearly defined hierarchical tree but a more nebulous cluster of relationships and interconnections that overlap, like the neurons that created them.

The material model for the contact stress problems discussed throughout this investigation is based on linear elasticity. It borders however on plasticity theory in determining the onset and definitions of yield, which is of critical importance to designers. The three-dimensional nature of contact stresses therefore requires an understanding of stress as a tensor quantity and its associated principal values (See Figure 1.3).

The principal stresses are used in conjunction with material properties to determine if a given geometry can withstand the applied loading without yielding, or if yielding does occur, its extent. Every material has different deformation response characteristics to applied loads, as

![Mohr's circle for stress (3D)](image)

Figure 1.3 Mohr’s circle for stress (3D)
illustrated by a stress-strain diagram. This data is typically obtained from the tensile test of a circular rod. One of the tasks of the stress analyst is to relate the state of stress for this simple configuration to a more complex three-dimensional situation. The effect of attaining a certain stress level depends on the model of the material. Figure 1.4 illustrates a few of the commonly used idealizations of stress-strain curves used for calculating plastic behavior. The final stress state is path-dependent if the stress levels go beyond the material's elastic limit. This means that the sequence or order in which the loads are applied that cause plastic deformation determines the final stress state.

The framework for this investigation is the realm of continuum mechanics and theory of elasticity (see Figure 1.2). The determination of contact stresses has its roots in linear elasticity theory which is based on the continuum model of material behavior. The continuum model assumes the material to be continuous, homogeneous, and usually isotropic. These assumptions allow field quantities such as stress and displacement to be expressed as piecewise continuous functions. The earliest models of contact problems were analyzed as concentrated loads applied to an infinite elastic medium. These solutions provided insight into the overall behavior of concentrated loads, but contained singularities. Singularities and the infinite stresses and deformations associated with them are not possible in real materials.

Most materials exhibit a strong temperature dependence, usually softening with increasing temperature. The rate at which the load is applied (strain rate) can also have an effect on the stress-strain curve. A category of materials exists called visco-elastic, that exhibit time-dependent responses to stress. The stress-strain curve is a useful idealization, but in reality each material possesses a stress response function that is a multi-dimensional surface described by the variables strain, strain rate, temperature, and time, to name a few. The computational method developed in this investigation does not account for plastic deformations, but could be adapted to do so by evaluating the elastic modulus from a piece-wise linear approximation of the stress-strain model in the plastic region.
1.5. The Elastic Foundation Model

One of the early models developed to analyze contact phenomena treated the material like a bed of springs as shown in Figure 1.5. This was done to avoid having to solve an integral equation for pressure. The bed of springs is presumed to be penetrated by a rigid indenter. The contact pressure at any point depends only on the compression of the spring at that point. Point contacts or indentation by bodies with very small radii are not modeled adequately because there is no support from adjacent springs. The model gives useful first approximations to problems in which the bodies cannot be locally represented by principal curvatures. This model has proved useful in analyzing rolling contact with tangential loading and problems involving visco-elastic bodies [1]. Elastic foundation type models are sometimes used in concrete slab design in order to calculate the required thickness to withstand building column loads.

1.6. Coordinate System For Non-conforming Single Point Contact

Before we can discuss any particular method of solution, it is useful to define a coordinate system which can serve as a framework to survey the entire field of contact mechanics. The definitions and figures in this section are from Johnson [1].

When brought into contact, non-conforming surfaces will initially touch at a point or line if one of the bodies is cylindrical. The contact area is generally assumed to be small compared to the dimensions of the bodies themselves. The origin of the coordinate system is taken to be the point of contact. The Z axis is taken to coincide with the common normal to the two surfaces. The X-Y plane is called the tangent plane. The undeformed shapes of the two surfaces are specified by the functions:

\[ z_1 = f_1(x,y) \quad \text{and} \quad z_2 = f_2(x,y). \]

The separation between them before loading is given by: \( h = z_1 - z_2 = f(x,y). \)
In applications, contact may involve complex motions such as those that occur between gears and self-aligning bearings. The motion of a body at any instant in time is described by its linear and angular velocity vector with respect to the contact point. The following terms are used frequently in the literature, so they are defined here with reference to the general coordinate system shown in Figure 1.6.

**Sliding:** relative linear velocity between the two surfaces.

**Rolling:** relative angular velocity between the two bodies about an axis lying parallel to the tangent plane.

**Spin:** relative angular velocity about the common normal.
Figure 1.7 shows that the net tangential force can be resolved into two components $Q_x$ and $Q_y$. The force transmitted at the point of contact has the effect of compressing the deformable solids so they effectively make contact over an area of finite size. This makes it possible to transmit a resultant moment in addition to a force which could not occur if contact was truly at a single point. The components $M_x$ and $M_y$ are defined as rolling moments and are part of what causes "rolling resistance". The component $M_z$ accounts for friction inside the contact area and is called the spin moment.

1.7. Milestones in Contact Stress Development

The following section highlights some of the major stepping stones to the work in this investigation. Significant developments are identified in the mathematical formulation and treatment of the problem, but it is by no means an exhaustive summary of all the work in the field of contact mechanics. Most developments in the theory did not appear until the beginning of this century, when needed by the railway, gear, and bearing industries. Theoretical developments seemed to have stopped in the mid nineteen-sixties, due to lack of computational power to carry out the integral solutions. Research in numerical methods continues however.

1881 Heinrich Hertz [2] invented the classical theory of contact stresses in his paper "On the Contact of Elastic Solids". At age 24, Hertz was working as a research assistant to Helmholtz at the University of Berlin. He was studying the optical interferences in glass lenses and the effect of clamping forces on the fringe patterns. He made an analogy between contact pressure and electrostatic potential theory. His solution gave only the principal stresses in the contact area but has been extended over the years to cover a wide variety of physical cases, at least as a first approximation and benchmark.

1930 H.R. Thomas and V.A. Hoersh [3] transformed the Hertzian solution for stresses on the axis of symmetry into standard elliptic integrals and discovered that the shearing stress distribution along the axis has a maximum at some distance beneath the contact area. This showed that a method of failure which initiated below the surface, such as pitting, could be explained by the existence of these maximum shear stresses.

1939 G. Lundberg [4] considered the effect of a tangential load between arbitrary surfaces. He introduced three potential functions corresponding to the components of loading along the axes of a Cartesian coordinate system.
(1949) R.D. Mindlin [5] showed that the solution of the tangential and normal pressure distributions can be decoupled without introducing significant error. The degree of approximation is of the same order as neglecting the small shear tractions in the Hertzian problem with dissimilar materials. He developed the theory of slip between the contact surfaces. An undesirable characteristic of his solution is that the traction approaches infinity at the edges, which is not physically possible. Presumably, the tangential component of traction cannot exceed the product of the coefficient of friction $f$ and the normal load. For the case of circular contact under spin, it is shown how slip penetrates from the outer radius $a$, to an inner radius $a'$, and that in the included annulus, the traction remains at the greatest possible value of $f p$. With increasing torque, slip is eventually initiated over the entire contact area. Where there is no slip, displacements on the two bodies must match.

(1951) Cheng Kang Liu [6] and also in an oft-cited reference with J.O. Smith [7] investigated both normal and tangential loads that were assumed to have an ellipsoidal (Hertzian) distribution over the contact area. The magnitude of the intensity of the tangential load is assumed to be linearly proportional to that of the normal load when sliding motion is impending. The stresses in the body are presented in closed form for both the plane stress and plane strain problems. A numerical value of $1/3$ was assumed for the coefficient of friction.

They discovered that when the tangential loads are applied at the contact area, the maximum shearing stress may be at the surface instead of beneath it. This is a most important result in that it explains real failure phenomenon such as surface crack propagation.

1.8. The Proposed Computational Method

In his thesis [6], Liu formulates the solution to the three-dimensional problem with normal and tangential loads in terms of integrals. He states the need for a feasible way of evaluating the integrals. The computational power now exists to perform these calculations. The main portion of this investigation resulted in computer programs to evaluate these integrals within the framework of the contact stress problem, making this a feasible course of action.

The method developed in this investigation is based on the notation and equations from Seely and Smith [8], which were originated by Liu. Seely was Liu’s thesis advisor and a colleague of Smith. Determination of stresses involves the evaluation of elliptic integrals at each step of the solution procedure. Elliptic integrals can only be
evaluated by numerical methods, which makes the calculation of contact stresses an inherently numerical problem. The program treats stress as a tensor quantity to reinforce its three-dimensional nature, which is so often overlooked. **The method of determining contact stresses by evaluation of the elliptic integrals, and its implementation as a computer program, forms the central work of this investigation.** The computer programs used to implement these calculations could be developed into a tool to be used in product development, failure analysis, and as a pedagogical tool in the classroom as well.

---

The calculation of contact stresses is an inherently numerical problem.
2. Classification of Contact Stress Problems

2.1. Relevant Factors

There are many factors to consider when modeling a contact problem. Problems do not always fit into neat categories and there will be cross-over between classifications. Some of the relevant factors are as follows:

- **Geometry**: Is the problem two or three-dimensional? If two-dimensional, is it plane stress or plane strain? Is there single, multiple, or conforming (close-fit, nearly contacting) contact? What is the condition of each surface? Are they rough or smooth?

- **Loading**: Are there normal or tangential loads, or both? Is the loading cyclic? What is its magnitude? Is it a point or distributed load?

- **Relative motion**: Does rolling, sliding, or torsional motion occur between the bodies, or is there impact?

- **Deformation**: Are large strains present, as is likely with elastomers? Is there plastic deformation?

- **Thermal effects**: Is it a problem concerned primarily with heat transfer through the interface or does a thermal stress field need to be superimposed on the mechanically imposed stresses?

- **Materials**: Are the materials isotropic, anisotropic, or possibly visco-elastic?

- **Relative stiffness between the bodies**: Should it be treated as a rigid punch problem?

All of the above factors are important in determining which solution methods are applicable to a specific problem. For the purposes of this investigation, the following classifications have been found to be useful.

2.2. The Rigid Punch and Other Stiff Geometries

Problems in which one body has a significantly higher elastic modulus than the other can be treated as a "rigid punch" type problem, see Figure 2.1. The distinguishing feature of these problems that allows them to be solved in a closed-form manner is that the shape of the contact region is known for any given penetration of the indenter. Many indenter
shapes have been studied such as blunt ends, tapered wedges, cones, and spheres (See Barber [9], Galin [10], Gladwell [11], and Nadai [12]). Tangential loading is not usually associated with this type of problem.

![Figure 2.1 Rigid punch indentation.](image)

Another group of problems in which one body is considered rigid compared to the other is the rolling of metals. These involve thermal effects, as in the case of hot rolling and plasticity with work-hardening for cold rolling operations.

### 2.3. Elastic Half-Spaces

An elastic half-space is defined as a semi-infinite elastic solid bounded by a plane surface. It is one half of a Cartesian space, usually identified with the plus or minus $z$ regions. However, it is often convenient to use polar coordinates with the origin centered at the point of contact. Deformations are taken to be linearly proportional to the forces. The differential equations of equilibrium and the compatibility relations from elasticity are assumed to apply. The surface outside the contact region is free of stresses, within the region it is loaded by the *prescribed* normal and tangential pressure distributions. At large distances from the loading zone, the stresses must tend to zero.

The elastic half-space idealization amounts to ignoring the effects of one of the contacting bodies with the purpose of simplifying the boundary conditions. Solutions typically give infinite stresses at the point of contact, which cannot physically exist, but are a property of the solution. The overall stress patterns have been verified by photoelastic and finite element methods. The idealization gives some useful results and in fact forms the basis for more general methods. Tangential and normal loads for line and point contact have been investigated for two and three-dimensional cases. **Figure 2.2** illustrates arbitrary normal and tangential loading profiles for the two-dimensional case. Classical elasticity problems including uniform and linearly varying force profiles have been investigated by Timoshenko & Goodier [13], Johnson [1], and Lur'ë [14].
2.4. Rolling Contact

Rolling contact problems are treated as separate phenomena. Rolling (contact) is defined as relative angular motion between two bodies about an axis parallel to their common tangent plane. The frame of reference as defined in section 1.6 is considered to move with the point of contact. If the velocities $V_1$ and $V_2$ are unequal, the rolling motion is accompanied by sliding. If the angular velocities $\omega_1$ and $\omega_2$ are unequal, it is accompanied by spin. The terms free rolling and tractive rolling are used to describe motions in which the net tangential force $Q$ is zero and non-zero, respectively [1].

In steady rolling, the strain field does not change with time. The resultant tangential traction must not exceed its limiting value which is the product of the coefficient of friction and the resultant normal force. Freely rolling bodies having dissimilar elastic properties will develop different tangential strains. A special case involves two elastic bodies which are geometrically identical and have the same elastic properties. That is, they are completely symmetrical about their interface. When rolling under the action of a purely normal force, no tangential traction or slip can occur, so the contact stresses and deformations can be approximated by the static Hertz distribution [1].

Sliding is not as straightforward since some portions of the contact area may slip or possess relative motion, while the remainder does not and sticks together. A difference between the tangential strains in the two bodies in the sticking area leads to a small apparent slip, called creep. Sticking and micro-slip zones form in the contact area in relative proportions and locations that are determined by the interaction between friction forces and elastic deformations.

To determine the stresses in the body, an Eulerian point of view is taken in which the material is considered to move past the point of contact. The equilibrium of surface elements is formulated in terms of their velocity vectors which have components from the strain rates and the velocity of the bodies as a whole. This technique allows the differences in surface strains to be accounted for, which give rise to slip. The boundary conditions for the strain-derived differential equations of equilibrium are, for steady rolling contact:

1) In regions where the bodies stick together, the relative surface velocities are zero.
2) Tangential traction is limited by the coefficient of friction.
3) Direction of tangential traction must oppose slip.

These conditions are modified (meaning increased complexity) to account for tractive rolling, and differing amounts of slip from complete slip to none. They become quite complex for three-dimensional bodies with traction, spin, and transient behavior occurring during start-up. For visco-elastic rolling, the problems of compression and rolling-sliding must be distinguished, even when friction is absent, due to the time-dependence of the stress-strain law. Rolling contact will not be covered further in this investigation. It was mentioned for completeness and to help illustrate the complexities associated with contact problems.

2.5. Plastic Deformation

Problems involving plastic deformation require special treatment because the final stress state is path-dependent. This state depends on the material model used, i.e. perfectly plastic, elasto-plastic, etc., and on the work hardening model used (if any). See Mendelson [15] for theories on kinematic and isotropic hardening. Follansbee and Sinclair [16] have studied the indentation of a ball well into the fully plastic state. Such work could be used to develop an accurate model of hardness testing. The initial phases of a standard hardness test can be treated as a classic contact stress problem up to the point of plastic deformation. An experimental test stand shown in Figure 2.3 could be built to measure the force and deflection of an indenter as it was pressed into a material. It would be set up for automatic data acquisition enabling the test data to be correlated to the theoretical model. Such a system could also be used to test various indenter shapes and sizes as well as for the analysis of rigid punch phenomena by using sharp-edged indenters. A holographic stress analyzer could record the fringe patterns of surface deflection in real time and they could then be correlated with theoretical predictions.
2.6. Conforming Contact

Conforming contact arises from the geometric condition in which the two bodies are separated by a small distance over appreciable portions of their surfaces. Here, the geometry before and after loading is known ahead of time. For light loads, the contact area will be small in relation to the radii of the bodies. As the load becomes larger, the contact area grows rapidly to become a significant fraction of the total surface area, which violates the Hertzian assumptions. Such a condition can be more suitably treated by methods appropriate to conforming contact. This is an example of how the magnitude of the load can determine the applicable solution method. Examples of conforming contact are a cylinder in a tight fitting hole, initially contacting along a line, and a sphere in a spherical cavity where the magnitude of the radii of the bodies are very close. For these simple geometries at light loads, the stresses and deformations predicted by Hertzian theory are accurate. As the load is increased, the stresses in the conforming case will be lower because the supporting area increases at a faster rate than predicted by Hertzian theory. The deformations will be smaller because of the additional support offered by the second body.

Numerical methods are needed to solve conforming contact problems. Paul and Hashemi [17] have studied contact pressures on closely conforming elastic bodies and have developed a numerical method for frictionless surfaces. Their work includes a technique for automatically generating meshes that overlay the load-dependent contact patches. The following is an outline of the method:

A Cartesian coordinate system is set up, with the initial contact point as the common origin. The initial separation between points on the two bodies is known. The effect at each point resulting from the deformation of neighboring points, is weighted by an influence function and summed up. Using indenters of prescribed finite curvature, a fictitious interpenetration is generated. The bodies are discretized into subregions and a system of linear equations is set up that determines the deformations from the penetration at neighboring points by superposition. Stress evaluation points are taken to be at the centroids of the cells. If boundary conditions are not satisfied, a new penetration curve or contact boundary is generated. This is done iteratively until satisfactory convergence occurs. Finite element programs are used to generate the interpenetration curves.

The restrictions of the method include the assumptions that the contact region is assumed to be symmetrical about one axis in the tangent plane, the initial contact is at a single point, and that there is no friction. They claim to have achieved the first reliable solutions for the wheel and rail problem.
2.7. Numerical Methods

Numerical methods typically involve discretization of the domain. Figure 2.4 is an example of a numerical method by Singh and Paul [18] in which the pressure function \( p(x,y) \) is replaced by a piecewise constant pressure field. Note the large gradients near the edges of the grid. They report that the simply-discretized method is numerically unstable in the general case. Instability arises because the solution vector to the set of linear algebraic equations is very sensitive to small perturbations in the elements of the coefficient matrix. Large oscillations in the solution vector correspond to small variations in the elements of the coefficient matrix. Since perturbations are unavoidable in the process of creating the grid, convergence to the solution vector tends to be very erratic. This discrete method is incapable of predicting the proper stress distribution in non-symmetric problems. For non-symmetric cases, stabilizing techniques were introduced which they call the "Redundant Point Field Method" and the "Functional Regularization Method". Both of these significantly increased the amount of computation required. As a refinement of the procedure techniques need to be found to eliminate the singularities occurring near the edges of the grid. On the other hand, the large scale solution is very good and friction effects are accounted for.

Other numerical solutions include finite element methods. A commercially available code that handles three-dimensional contact has been recently developed [19]. The drawbacks to the finite element method, based on the amount of memory required for available commercial programs to run and the size of their data files, are that it:

- is computationally extensive.
- does not readily lead to parameterization studies, although parametrically generated FEA models could be generated.
- calculates stresses and deformations throughout the entire field when, during initial design phases, only critical design parameters are sought such as orders of magnitude or maximum values of stress at a particular location etc.

FEA may predominate as the preferred analysis method in the long run however, due to its ability to handle thermal stresses, non-linear effects (plasticity and large strain), anisotropic materials, and multi-point contact.
2.8. Hertzian Contact

The pioneering work of Hertz [2] includes the classic contact stress solution and was the first theoretical work done in this area. Elastic contact stress problems are classified as Hertzian if they satisfy the following (five) conditions:

1) The bodies are homogeneous, isotropic, obey Hooke's Law, and experience small strains and rotations, i.e. small strain elasticity.

2) The contacting surfaces are frictionless.

3) The dimensions of the deformed contact patch remain small compared to the principal radii of the undeformed surfaces.

4) The deformations are related to the stresses in the contact zones as predicted by the linear theory of elasticity for half spaces.

5) The contacting surfaces are continuous and may be represented by second-degree polynomials (quadratic surfaces) prior to deformation.

The majority of work on contact stress is based on the assumption that the contact region is the ellipse predicted by Hertzian analysis. The equations for circular contact areas will be developed later in this investigation to familiarize readers with the problem. These solutions can be expressed as simple formulas so the general concepts and behavior can be illustrated before proceeding to the more general case of elliptical contact areas which is the central topic of this investigation.

2.8.1. Elliptical Contact: "The General Case"

The most general case has no symmetry about the surface normal to the tangent plane. Tangential loading also destroys the symmetry of the problem. Major results for the general case of Hertzian contact are as follows:

- The contact area is bounded by an ellipse which can be calculated from the geometric configuration: the principal radii of the bodies and the relative angular orientation of the principal planes.

- The dimensions of the ellipse, a and b, increase directly as the cube root of the load P.

- The normal pressure distribution is ellipsoidal with the maximum pressure $p_0$ occurring at the center of the ellipse and having a value of
Regions of the solids which are remote from the contact zone approach each other by an amount which varies as $P^{2/3}$.

The principal stresses are all compressive. The maximum shearing stress occurs below the surface.

Note the non-linear dependence on the load of the contact area and deflection. The same formulas hold for both solids using the appropriate value of Poisson's ratio, $\nu$. See section 5.2 on Spherical Geometries for general numerical conclusions. Thomas and Hoersh [3], and Lundberg [4] show that variations of Poisson's ratio in the usual engineering range do not have a strong influence on the important features of the stress pattern. It may still affect the magnitude of the results, so the proper values should be used for design problems. Special cases of the general Hertzian problem are those of spherical and cylindrical contact. The equations in these simple cases do not involve elliptic integrals, however the presence of friction greatly increases their complexity.

2.8.2. Limitations or Bounds of the Hertz Theory

The Hertz theory has been confirmed by numerous experiments. Haines and Ollerton [20] report that the Hertz theory accurately predicts the shear stresses for elliptical contact where the semi-major axis of the ellipse of contact is as large as one-half the smallest radius of curvature of either of the contacting solids. The geometric predictions of the theory cannot be expected to be accurate when the deviations from the original hypothesis are this large. Beyond the elastic limit, some of the relations continue to be approximately valid, but the deviation increases with the degree of departure from linear elasticity.

2.8.3. Extension of Hertzian Contact

Several departures from the original hypotheses have been developed, including higher-order polynomials to model the initial geometry. The polynomial corrections are on the same order of magnitude as the assumptions of the linear elasticity theory itself so this is not a productive avenue, except for cases of conforming contact. Dissimilar materials have also been studied and some parameters developed to account for the relative stiffness between the two bodies. Contact of rough surfaces has been modeled by assuming that the bodies are covered with an array of hemispherical protrusions called asperities. This is an attempt to determine the true contact area and relies heavily on statistical methods. For more on this, see Section 7.0. Hydrodynamic lubrication problems are based on the assumption that the initial fluid pressure distribution is Hertzian in nature.
Attempts have been made to incorporate friction in the Hertzian results. The so-called Smith-Liu Equations [7] solve the two-dimensional cylindrical problem with tangential loading. The solution does not contain elliptic integrals, but the computations are still very tedious. Analytical solutions have been developed for sliding spherical contact by Goodman and Hamilton [21] which are more like explicitly-specified procedures rather than simple formulas. In this investigation a numerical method is defined as one in which the domain is decomposed into discrete regions or substructures, versus a computational one which evaluates the stress field at any continuous coordinate based on the continuum equations.
3. The Geometry of Contact

3.1. Local Surface Geometry

Two bodies in contact are presumed to touch initially at a single point. The curvature of each body is modeled by two radii, the principal and secondary, which are associated with the maximum and minimum surface curvatures of the body at the contact point. The planes containing the maximum and minimum curvatures are orthogonal. When two bodies are considered, the principal planes of each will not necessarily coincide so the angle between the planes of principal curvature must be considered. The complete geometry can therefore by specified by four radii and the angular orientation between the principal axes. The angle of orientation \( \alpha \) is meaningless if either one of the bodies is axisymmetrical, in which case it is assigned a value of 0° for calculation purposes. Negative radii indicate that the center of curvature for that principal plane lies away from the body, i.e. it is a concave surface. The classification of a given geometry and its subsequent mathematical treatment depends on the sign and relative magnitude of the four radii.

![Figure 3.1 Body with positive principal and secondary radii of curvature.](image)

The following is a rationale for the use of two radii to describe the local surface geometry. In modeling a continuum, it can be assumed that the surface has a continuous curvature, that is, there are no discontinuities in the region near the contact point. Mathematically, this means that the slopes must be continuous at all points on the surface. The simplest continuous functions that satisfy this requirement are second order polynomials, since by taking their second derivative with respect to displacement, a value which represents the curvature exists. The use of higher-order functions to
locally describe a surface leads to expressions having similar form but containing more terms. This approach is used in problems involving conforming contact [22].

There is a branch of mathematics called differential geometry in which physical quantities such as length take on infinitesimal values. Differential geometry is the study of geometric figures using the methods of calculus. A well-known theorem states that any surface can be locally characterized by two principal radii. This abstraction is acceptable for contact stress analysis, since we are only concerned with local effects near the point of application of the load. A surface in $\mathbb{E}^3$ (three-dimensional Euclidean space) is uniquely determined by certain local invariant quantities called the first and second fundamental forms. These quantities depend only on the surface and not the particular representation (coordinate system). Another differential geometry theorem states that: For every point $P$ on a surface, there exists a coordinate patch containing $P$ such that the directions of the $u$ and $v$ parameter curves are principal directions. See Lipschutz [23] for more on this.

Gaussian curvature, another invariant, is the product of the principal and secondary curvatures of the body. While it is not used in the determination of stresses and displacements, it could be a useful characterization of the surface for possible use as a variable in a dimensional analysis study of ball bearing failure rates. The theorems and derivations of differential geometry do not play a role in the actual determination of contact stresses. The point of mentioning them is to show that a surface described locally by two curvatures is not just a physical convenience introduced to solve the problem, but also has a sound mathematical basis.

Figure 3.2 illustrates that the bodies themselves can be quite arbitrary and irregular. It is easier to visualize the contact if you project or extend the local radii which characterize the surface to create a complete object. Various combinations of numeric sign and magnitude of the principal and secondary radii can be thought of as representing different bodies such as spheres and cylinders. Eight types of these bodies can be distinguished as defined in Table 3.1. These are not necessarily the overall shape of the bodies in contact, although they could be. If the radii that characterize the surface at any given point are extended, they would form whole bodies in these shapes. The curvature may be completely different at another point on the surface of the body, but it can not be drastically different if it is in the vicinity of the contact point. The classification and naming of the "Body Types" is for the purposes of visualization only.
The need for this classification arose in the computer implementation of the analysis of contact stresses. Certain cases or combinations of these "bodies" require special mathematical treatment to avoid division by zero and other computer-related peculiarities. The need to treat certain geometries as special cases is not mentioned anywhere in the literature. Many authors develop integral equations to solve for contact stresses but mention nothing about how to implement them, which introduces doubt as to whether or not they were ever used to actually calculate anything.
The most general body, one with unequal positive principal curvatures, is termed "Body Type I" and is illustrated in Figure 3.3. The inverse of this is both curvatures being negative as seen in Figure 3.5. Body Type II shown in Figure 3.4 is characterized by one positive and one negative radius. Figure 3.6 shows Body Type IV, the convex sphere. The simplest body type is a sphere with a radius of one unit. Body Type V, the concave sphere, is realized when both radii are equal and negative. This type of surface could be produced by a ball end mill. The torus shown in Figure 3.7 is not one of the body types itself, but the three points A, B, and C can each be represented by a different one. The torus was included to show that the local curvature should not be confused with the geometry as a whole. Body Type VIII, shown in Figure 3.8, is "The Plane" and is characterized by both radii being infinite. When one of the bodies is a plane or spherical type, the angle $\alpha$ becomes irrelevant due to symmetry.

### 3.2. Combinations of Bodies

Although there are no restrictions on the values of the radii on an individual body, other than being non-zero, there are restrictions when two bodies are considered. Certain combinations of radii and angular orientation produce geometries that are physically impossible for a single-point contact problem. The following figures provide a graphic illustration of some of the combinations of body types.

When two cylinders are placed in contact, they may produce elliptical or line contact depending on the relative angular orientation. Figure 3.9 shows the case of line contact between cylinders in which case $\alpha$ is $0^\circ$. A special case occurs when two cylinders of equal diameter are crossed at $90^\circ$, producing a circular contact area. An experimental apparatus such as this can be used to test coefficients of friction and wear-rates between various materials because the area can be easily predicted and the sliding velocity controlled by the rotational speed of the cylinders. The stress distribution for
this case is not the same as the spherical case of equal area and normal force, because the secondary radii are infinite. Consider an arbitrary point on the surface of the solid near the contact point. A sphere contacting another sphere or a plane will both produce circular contact areas. Now move some distance outside the contact area, keeping a constant \( z \) elevation. In the first case of two spheres, this location is outside either of the bodies while, in the second case it remains on the surface of the plane. The stress distributions cannot be identical because there is not a one-to-one mapping of the geometries. The maximum normal pressure and stresses on the centerline should be similar between the two cases however.

Line contact can occur if a cylindrical or planar body type is explicitly selected or the eccentricity of the elliptical contact area is so large that contact effectively becomes a rectangle. The circular and line contact cases are simpler than the general case in that they do not depend on elliptic integrals. The fact that the elliptical solution asymptotically approaches the closed-form circular and line contact solutions makes for a good benchmark test of a "general method".

A feature of the line contact problem is that it cannot be normalized (i.e. mapped onto the range from \( 0 < R < 1 \)) because of the infinite radius defining a cylinder or plane. To normalize a problem, all the radii are divided by the magnitude of the largest radius which then itself becomes unity. In line contact, the infinite radius would map all the other radii to zero.

Certain values of the four radii and the angular orientation can be specified which correspond to physically impossible geometries. For example, a body with a positive radius cannot be brought into single point contact with a body having a negative radius of smaller magnitude than the positive one. This would be a force fit. A mathematical solution might be obtained, but it would be incorrect to apply the results to a design problem. Geometries may also be specified that are physically possible but violate the assumptions of the theory, such as the case of conforming and multi-point contact.

### 3.3. Magnitude Checks

Potentially invalid geometries can occur when any one of the four radii takes on a negative value, i.e. possesses concave curvature. Negative curvatures occurring in the same principal plane on both bodies simultaneously produces a concave surface trying to contact another at a single point, which is impossible. This is like putting a square peg in a round hole, it doesn’t fit. A concave surface and a plane is not a valid combination for the same reason. If only one of the four radii is negative, but \( \alpha > 0 \), a
cross check of the fit between the principal radii of body one and the secondary radii of body two must be performed. Figure 3.10 illustrates what happens when the magnitude of the secondary radius of one body is larger than the magnitude of the principal radius of the other, and the bodies are oriented at any angle besides 0°. The result is contact in two points which is not permissible for this analytical idealization.

![Figure 3.10 Check if bodies fit together](image)

Following the same reasoning, Figure 3.11 illustrates how certain combinations of bodies can exist only at specific angular orientations, in this case 90°. This same type of restriction exists if the first body is either body type II or VI, and the second possesses negative curvature. Any point below the tangent plane would come into contact with the other body if rotation were to occur about the z axis, thus violating the assumption of single point contact on which the solution is based.

It is possible to specify physical dimensions that violate the assumptions of the theory and still obtain numerical results, but no confidence should be placed in them. To avoid this pitfall, it is necessary to examine the "fit characteristics" of various combinations of body types before proceeding with what could be an extensive amount of meaningless computation. As part of this investigation, a computer program has
been written which compares the magnitude of the four radii and their relative angular orientation to detect invalid configurations. In addition to the magnitude checks, certain combinations of body types represent cases that cannot be solved by the method involving elliptic integrals. These are the cases of axisymmetric and line contact, which can be solved by simpler closed-form methods anyway. A general purpose contact program requires the ability to detect when invalid geometries and cases requiring different solution methods have been specified. Table 3.2 shows the allowable combinations and other information about symmetry.

The sign and relative magnitude checks must be performed if any one of the four radii are negative.
Combinations of Bodies

Table 3.2 Shows valid combinations of "Body Types" for single point contact. The case number is by the row of body 1 and the column of body 2. All the possible cases are shown in the table on the left. A checkmark indicates that the situation characterized by the column heading is true.

NR  Indicates negative radii are present.
LC  Line contact.
α  Angular orientation must be specified.
IV  Invalid combination of body types.

Of the 36 possible combinations of the body types, if no distinction is made between which is body 1 and 2, only 20 are considered valid or physically possible for single point contact. Three of the cases represent line contact. The angle α is present in 8 cases, although restricted to a single value (indicated in the table) in 2 cases. A negative radius check must be done in 11 cases. A logic table similar to this is used by the computer programs developed in this investigation.

The point of this section was to show that the specification of the contact geometry is more complicated than it first appears. Elaborate validation checks must be performed on the input geometry if any engineering decisions are to be based on the results of the calculations.
4. Elasticity Methods

Here the fundamentals of three-dimensional elasticity and the classical solution technique using stress functions are reviewed.

4.1. Equilibrium

A three-dimensional differential cubic element is isolated for a free-body diagram. A force summation is taken for each coordinate direction resulting in three equilibrium equations. This summation includes the normal and shearing forces on each face of the element in addition to body forces. The resulting differential equations of equilibrium are:

\[ \frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{xz}}{\partial z} + X = 0 \]  
\[ \frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} + Y = 0 \]  
\[ \frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \frac{\partial \sigma_z}{\partial z} + Z = 0 \]

where \(X, Y,\) and \(Z\) are the distributed body forces which could result from gravity or electromagnetic forces. These equations must be satisfied at all points throughout the body. This system of equations is indeterminate since there are six unknown components of stress and only three equations. To solve for the stresses, assumptions must be made about the deformations to provide the remaining equations. These additional equations constitute the six equations of compatibility and are discussed in a following section.

The stresses at the surface must be in equilibrium with the external forces. These can be determined by force summation on a tetrahedron bounded by an inclined plane (the surface) and the coordinate directions yielding:

\[ \bar{X} = \sigma_x \ l + \tau_{xy} \ m + \tau_{xz} \ n \]  
\[ \bar{Y} = \tau_{xy} \ l + \sigma_y \ m + \tau_{yz} \ n \]  
\[ \bar{Z} = \tau_{xz} \ l + \tau_{yz} \ m + \sigma_z \ n \]
where $X$, $Y$, and $Z$ are the components of surface force per unit area and $l$, $m$, and $n$ are the direction cosines of the surface normal at the point with respect to the coordinate system. By considering the equilibrium of a tetrahedron internal to the material, if the stresses at that point are known, the components on any inclined plane can be determined. It is in this manner that the principal stresses are also determined, for use in failure theories.

4.2. Strain at a Point

Strain is a measure of the intensity of deformation in the vicinity of a point (meaning fixed location in space). The analysis of strain is a geometric problem and is unrelated to material properties. If the relative position of any two points in a continuous body is changed, then the body is strained.

The displacements of the body are assumed to be single-valued functions possessing up to the third continuous derivative. The six components of strain can be derived by considering a small linear element at some arbitrary position with respect to the axes in which each end is subject to a displacement vector. The distance between the end points of the initial and final configurations is used as a measure of the deformation. Part of the displacement vector contains translation and rotation as well, so to determine the strain, these components must be subtracted. Linear elasticity assumes that the difference between the configurations is small compared to unity. Two types of strain can occur. Linear strains are the unit elongations in each coordinate direction and shearing strains are the angular distortions between faces of the differential cubic element.

The strain components can be derived from geometry by considering the deformations, or mathematically by simply expanding the displacements in a Taylor series about the point in which higher order terms neglected. These components in the linear non-torsional case are:

\[
\begin{align*}
\epsilon_x &= \frac{\partial u}{\partial x} \quad & (a) \\
\epsilon_y &= \frac{\partial v}{\partial y} \quad & (b) \\
\epsilon_z &= \frac{\partial w}{\partial z} \quad & (c) \\
\gamma_{xy} &= \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \quad & (e) \\
\gamma_{xz} &= \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \quad & (e) \\
\gamma_{yz} &= \frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \quad & (f)
\end{align*}
\]
4.3. Compatibility Equations for Linear Strain

The six components of strain at each point are completely determined by the three functions $u, v, w$ representing the components of displacement, so an additional relationship must exist between them. Imagine that a body is divided into infinitesimal cubes stacked together. If each cube is subjected to arbitrary strains, they will no longer fit together as a continuous body, i.e. there will be gaps and overlap. There must be some other relationship between the strains so that the body will remain continuous after strain. These are sometimes called the continuity relations (which makes more sense) and were first derived by Saint-Venant. The components of strain cannot be taken as arbitrary functions of $x, y, z$, but are subject to the "compatibility equations". There are six differential relations between the components of strain. The first three equations are the in-plane dependence, the last three are the out-of-plane dependence. To obtain the first of these, equation (4.4(a)), differentiate twice with respect to $y$, $e_y$ twice with respect to $x$, and $7xy$ once with respect to $x$ and once with respect to $y$ to reveal the equality. It is used to determine if the strain field is compatible with the displacements in the $xy$ plane. Similar procedures are used to obtain the other forms (b) through (f).

\[
\begin{align*}
\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \\
\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial z^2} = 0 \\
\frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2} = 0 \\
\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 v}{\partial x \partial z} = 0 \\
\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 w}{\partial x \partial z} = 0 \\
\frac{\partial^2 v}{\partial x \partial y} + \frac{\partial^2 w}{\partial x \partial z} = 0
\end{align*}
\]
Elasticity

4.4. Hooke's Law

Hooke's Law is a relationship between stress and strain. The form for linear, homogeneous, isotropic materials is expressed by equations 4.5. It contains the two material constants $E$ and $v$ which are the modulus of elasticity and Poisson's ratio. These represent the material stiffness and lateral contraction effects, respectively. It is a subset of the Generalized Hooke's Law which may contain more material constants, the number depending on the material model. In non-isotropic materials there may be a different elastic modulus and Poisson ratio for each independent direction. These directions may not even be orthogonal as in the case of layered composite materials.

\[
6X = -[ax - v{0y} + aj], \quad (4.5a)
\]
\[
ey = -[ay - v\oz + ax], \quad (4.5b)
\]
\[
\z = K - \phi{1} + j). \quad (4.5c)
\]

4.5. Stress Functions

A common technique for solving the system of equations consisting of equilibrium, compatibility, and the boundary conditions, is to introduce what is called a stress function. This technique was pioneered by G.B. Airy in 1862 [24]. The stress components are defined to be partial derivatives of the stress function $\phi$. These components (ignoring body forces) are:

\[
a^2\phi \quad a^2\phi \quad a^2\phi \quad \frac{\partial^2\phi}{\partial x^2} \quad \frac{\partial^2\phi}{\partial y^2} \quad \frac{\partial^2\phi}{\partial x\partial y} \quad (4.6)
\]

The stress function and its appropriate derivatives are substituted into a form of the compatibility equation that is expressed in terms of stresses (having been transformed from strains using Hooke's law). For plane problems, a function can be developed which automatically satisfies the equilibrium and compatibility equations. The result is that in two dimensions the stress function must satisfy:

\[
W = \frac{1}{2} \left(\frac{\partial^2 \phi}{\partial x^2} \right) \quad \frac{1}{2} \left(\frac{\partial^2 \phi}{\partial y^2} \right) \quad (4.7)
\]

Equation 4.7 is called the biharmonic equation and $\phi$ is sometimes called the Airy stress function. The advantage of using an Airy stress function is that if one is found,
then it automatically satisfies equilibrium. A complete solution can then be found by satisfying the boundary conditions. The disadvantage of the stress function approach is that the search for the appropriate stress function is a trial and error process. Certain types of functions produce known stress fields however, so the search is not totally futile. Stress functions can also be combined by superposition.

Polynomials containing x and y terms can be used as stress functions to represent uniform, linearly and parabolically varying loads, for both normal and shearing force distributions. For polynomials of 2nd or 3rd degree there are no constraints in choosing the magnitude of the coefficients which are selected to represent different load cases. Higher order polynomials will yield systems of equations when substituted into the expressions for stress, but there will be relationships between the coefficients. The trick is to find suitable coefficients. Perhaps a symbolic algebra program applied to this problem (searching for suitable stress functions and evaluating the corresponding physical configuration) would yield some interesting results.

Stress functions expressed as Fourier Series have also been used to approximate discontinuous loading. Muskhelishvili’s method of complex potentials can also be used to deduce the potentials directly from the boundary conditions which has proved useful for curved boundaries such as elliptic holes (crack studies) and hyperbolic boundaries (knotches). Other stress functions and solution methods have been developed for torsional problems.

4.6. Determination of Displacements

The displacements u, v, and w of a point in the body are continuous functions of x, y, and z. Once the stresses are determined, the strains can be determined by Hooke’s Law. The displacements can now be determined making use of the strain definitions found in equations 4.3. Taking the derivative with respect to x, y, and z of each of the strain components yields 18 equations containing the second derivatives of displacement. The expressions derived from the u displacements are shown in equations 4.8 and the other 12 corresponding to v and w are similarly obtained.

\[
\frac{\partial^2 u}{\partial x^2} = \frac{\partial \varepsilon_x}{\partial x} \tag{4.8a}
\]

\[
\frac{\partial^2 u}{\partial y^2} = \frac{\partial \gamma_{xy}}{\partial y} \frac{\partial \varepsilon_y}{\partial x} \tag{4.8b}
\]

\[
\frac{\partial^2 u}{\partial z^2} = \frac{\partial \gamma_{xz}}{\partial z} - \frac{\partial \varepsilon_z}{\partial x} \tag{4.8c}
\]
Second derivatives must be used to ensure continuous slopes. The linear strain component \( \epsilon_x \) cannot simply be integrated once with respect to \( x \) to find the displacement \( u \) since it may be a function of \( x, y, \) and \( z \) and unknown functions of \( y \) and \( z \) would remain undetermined. The displacements \( u, v, \) and \( w, \) which are obtained by double integration of the strain field, are not entirely determined by the stresses and strains. Rigid body displacements and rotations can occur (from the constants of integration), but these can be calculated based on the constraints.
5. Axisymmetric Solids of Revolution

In the following sections, the theory of contact stresses is developed for the class of problems involving axisymmetric solids of revolution. By restricting analysis to this type of geometry, the theory and methods of solution can be developed with simple formulas before proceeding to the more general non-symmetric case which requires numerical integration.

5.1. Stresses and Deformation in a Solid of Revolution

A circular cylinder under uniform pressure, a rotating circular disk, torsion of a circular cylinder, and some contact stress problems can be classified together due to the similarities in their analyses. Although the classic Hertzian contact problem is considered torsionless, torsional contact has been investigated by Mindlin [5]. These problems are usually handled in cylindrical coordinates and the resulting stress components are independent of \( \theta \). In the axisymmetric case, the general strain-displacement relations in cylindrical coordinates reduce to:

\[
\begin{align*}
\epsilon_r &= \frac{\partial u}{\partial r} \\
\epsilon_\theta &= \frac{u}{r} \\
\epsilon_z &= \frac{\partial w}{\partial z} \\
\gamma_{rz} &= \frac{\partial u}{\partial z} + \frac{\partial w}{\partial r}
\end{align*}
\]

The equations of equilibrium become

\[
\begin{align*}
\frac{\partial \sigma_r}{\partial r} + \frac{\partial \tau_{rz}}{\partial z} + \frac{\sigma_r \sigma_\theta}{r} &= 0 \\
\frac{\partial \tau_{rz}}{\partial r} + \frac{\partial \sigma_z}{\partial z} + \frac{\tau_{rz}}{r} &= 0
\end{align*}
\]

and the stress tensor can be represented as

\[
\tau_{ij} = \begin{pmatrix}
\sigma_r & \tau_{r\theta} & \tau_{rz} \\
\tau_{r\theta} & \sigma_\theta & \tau_{\theta z} \\
\tau_{rz} & \tau_{\theta z} & \sigma_z
\end{pmatrix}.
\]

The compatibility equation can be written in terms of stresses. The stress components can then be expressed in terms of a stress function. Such a procedure yields the following expressions for the stress components:

\[
\sigma_r = \frac{\partial}{\partial z} \left( \nu \nabla^2 \Phi - \frac{\partial^2 \Phi}{\partial r^2} \right)
\]

(5.3a)
\[
\sigma_\theta = \frac{\partial}{\partial z} \left[ \nu \nabla^2 \Phi - \frac{1}{r} \frac{\partial \Phi}{\partial r} \right]
\]

(5.3b)

\[
\sigma_\theta = \frac{\partial}{\partial z} \left[ \nu \nabla^2 \Phi - \frac{1}{r} \frac{\partial \Phi}{\partial r} \right]
\]

(5.3b)

\[
\tau_{rz} = \frac{\partial}{\partial r} \left[ \left(1 - \nu\right) \nabla^2 \Phi - \frac{\partial^2 \Phi}{\partial z^2} \right]
\]

(5.3d)

The expressions for stress may be verified by substitution into the equilibrium equations 5.2. The above relations hold true provided that the chosen stress function \( \Phi \) satisfies the biharmonic equation:

\[
\nabla^2 \nabla^2 \Phi = 0
\]

(5.4)

Many problems can be conveniently solved using stress functions. Some yield solutions that take the same form as solutions to Laplace's equation, which is given by

\[
\Phi_n = R^n \Psi_n
\]

(5.5)

where \( \Psi_n \) is a function of \( \theta \) only. It is also possible to write the stress function as the product of two Legendre polynomials \( P_n(x) \) and \( Q_n(x) \). Timoshenko [13] showed that solutions can be obtained when the stress functions are polynomials in the form:

\[
\Phi_0 = A_0
\]

(5.6a)

\[
\Phi_1 = A_1 \ z
\]

(5.6b)

\[
\Phi_2 = A_2 \left[ z^2 - \frac{1}{3} \ (r^2 + z^2) \right]
\]

(5.6c)

\[
\Phi_3 = A_3 \left[ z^3 - \frac{3}{5} \ z \ (r^2 + z^2) \right]
\]

(5.6d)

\[
\Phi_4 = A_4 \left[ z^4 - \frac{6}{7} \ z^2 \ (r^2 + z^2) + \frac{3}{35} \ (r^2 + z^2)^2 \right]
\]

(5.6e)

\[
\Phi_5 = A_5 \left[ z^5 - \frac{10}{9} \ z^3 \ (r^2 + z^2) + \frac{5}{21} \ z \ (r^2 + z^2)^2 \right]
\]

(5.6f)
Other solutions can be obtained in the form of:

\[
\Phi_2 = B_2 (r^2 + z^2) \quad \text{(5.7a)}
\]

\[
\Phi_3 = B_3 z(r^2 + z^2) \quad \text{(5.7b)}
\]

\[
\Phi_4 = B_4 (2z^2 - r^2)(r^2 + z^2) \quad \text{(5.7c)}
\]

\[
\Phi_5 = B_5 (2z^3 - 3r^2 z)(r^2 + z^2) \quad \text{(5.7d)}
\]

The stress functions listed above are used in the next section dealing with a concentrated load. Each of these functions, and any linear combination of them, can also be taken as a stress function. Solutions to various problems may be found by adjusting the constants to match the boundary conditions. These polynomials are utilized for axisymmetric problems, where the number of variables has been reduced by symmetry. Another set of stress functions would have to be developed for more general non-symmetric cases, such as asymmetrically placed concentrated loads, but these are not easily found. For the non-symmetric cases and irregular geometries, numerical methods such as the finite difference or finite element methods can be used to solve for stresses and displacements. Polynomial solutions of higher order than six can be used to solve problems with non-uniformly distributed loads.

The above solutions could be applied to a thick lens, a variable thickness mirror, and shells of revolution. Other classic axisymmetric problems that can be solved using stress functions include the bending of a circular plates, rotating disk problems, spherical containers subjected to internal or external pressure, and contact stress problems such as contacting spheres. Solutions to some of the classic elasticity problems will be reviewed in the following sections. These solutions were developed in the early history of contact mechanics and form the building blocks of more recent solutions.
5.2. Concentrated Normal Force

The problem involving a concentrated normal force on an elastic hlf-plane was originally solved by J. Boussinesq [25] in 1885 and more recently by Love [26]. The case of a point or concentrated load acting on a plane is the simplest idealization of a contact stress problem and is the starting point of the analysis. It is a building block for solutions having distributed loads. The solution uses the basic methods of elasticity and can be developed in cartesian or polar coordinates, the latter leading to simpler expressions. The results are similar to the two-dimensional line loading problem. The solution to this problem is not derived here, but the results, which can be obtained by several methods, are as follows.

Timoshenko and Goodier [13] solved this problem using the fact that a point load is axisymmetric, which allows the use of the stress function technique. The stress function they used is

\[ \Phi = B(r^2 + z^2)^{1/2} \]  

where B is a constant yet to be determined. This stress function is substituted into equations (5.3) to obtain expressions for the stress components. These expressions contain the constant B which is evaluated by equating the resultant of the surface forces on a spherical cavity centered around the origin, to the applied load. To simplify the results, let \( \rho \) be the distance from the point of application of the external force to the point of interest. This is illustrated in Figure 5.1.

\[ \rho = (r^2 + z^2)^{1/2} \]

The resulting stresses are:

\[ \sigma_r = \frac{P}{2\pi} \left[ (1 - 2\nu) \left( \frac{1}{r^2} - \frac{z}{\rho r^2} \right) - \frac{3zr^2}{\rho^5} \right] \]  

\[ \sigma_\theta = -\frac{P}{2\pi} (1 - 2\nu) \left( \frac{1}{r^2} - \frac{z}{\rho r^2} - \frac{z}{\rho^3} \right) \]
\[
\sigma_z = -\frac{3P}{2\pi} \frac{z^3}{\rho^5}
\]  
(5.10c)

\[
\tau_r = -\frac{3P}{2\pi} \frac{rz^2}{\rho^5}
\]  
(5.10d)

On a global scale, the stresses are inversely proportional to the square of the distance from the origin. It can be seen from \textbf{equations 5.10} that at the origin, where \(\rho\) is zero, there is a singularity, i.e. the stress components are infinite. This is a physical impossibility, so the mathematical model needs refinement if more accurate calculation of the localized effects are to be determined.

The displacement components can be determined by substituting the stress components into the general strain-displacement relationships. The radial displacement is obtained by direct substitution into

\[
u = \epsilon_r r = \frac{r}{E} [\sigma_\theta - \nu(\sigma_r + \sigma_z)]
\]  
(5.11)

The vertical displacement requires integration with the result that:

\[
w = \frac{P}{4\pi G} \left[\frac{z^2}{\rho^3} + \frac{2(1 - \nu)}{\rho}\right]
\]  
(5.12)

which at the surface of the solid where \(z=0\) reduces to:

\[
w = \frac{(1 - \nu)}{2\pi G} \frac{P}{r}
\]  
(5.13)

Thus the shape of the deformed surface is a rectangular hyperboloid, which is asymptotic to the undeformed surface far from the origin, and at the origin, the displacement is infinite. \textbf{Figure 5.2} illustrates the overall stress patterns by photoelastic contours.
5.3. Distributed Normal Loads

The stresses and deflections resulting from distributed forces can be obtained using the results of the concentrated force solution. For most engineering purposes, it is desired to find the deflection of the free surface \( w(r,0) \) and the stresses (particularly the maxima) at any interior point \( A(r,z) \). Consider a general point on the surface \( B(r,0) \) and the effect on it from a distributed load \( p(\xi,\eta) \) as shown in Figure 5.3. The equations are simplified if the point of interest on the surface is taken as the origin of a coordinate system having polar coordinates \((s,\phi)\). The magnitude of the force acting on a differential area is \( psd\phi ds \). The displacement of the surface at point \( B \) arising from this differential force can be written using the results of the concentrated force case. The displacement at \( B \) due to the entire distributed load is then obtained by integration over the entire contact area and is given by:

\[
w = \frac{1}{\pi E} \frac{\nu^2}{2} \int \int p(s,\phi)sdsd\phi \tag{5.14}\]

The displacement components at any point on or beneath the surface can be obtained by integrating the stress components deduced from the concentrated force case. This general expression for displacement (and ones for stresses) have been investigated for a number of different force distributions, including normal and tangential loading, over polygonal, circular, and elliptical areas. Some of these are discussed in the following paragraphs. A common feature of all these problems is that they lead to elliptic integrals or very lengthy expressions whose only practical method of analysis involves the use of computers.
Polygonal regions subjected to uniform and triangular pressure distributions as shown in Figure 5.4 are the basis of strictly numerical methods for arbitrary shapes. Any polygonal region can be decomposed into triangular areas. The solution for a 2a x 2b rectangular area has been found by A.E.H. Love [26]. Love's results however have infinite values of the shearing stress $\tau_{xy}$ at the corners of the rectangles. All other stress components are finite. This type of problem has applications to the design of base plates for columns on an elastic foundation. Johnson and Bentall [27] have worked out the deflection of the surface acted on by a pyramidal force distribution on a hexagonal area. A polynomial pressure distribution acting on a triangular area has been solved by Svec and Gladwell [28].

Cases involving circular areas, even though the geometry is simple, still lead to integral solutions. For circular regions of radius $a$, solutions can be found if pressure distributions take the form:

$$p = p_0 \left(1 - \frac{r^2}{a^2}\right)^n$$  \hspace{1cm} (5.15)

in which $p_0$ is the maximum surface pressure and the exponent $n$ assumes different values. The solution is obtained by integrating over the contact area utilizing the point loading results. Points inside and outside the contact boundary are treated as separate cases, since the integrals take on different forms for each region. The geometry for such an integration is shown in Figure 5.5. The resulting expressions involve elliptic integrals, which play a larger role in the general theory of contact between two bodies which is the primary subject of this investigation.

Several representative load cases can be constructed by choosing the appropriate value of $n$ in equation 5.15. Uniform pressure is represented by setting $n=0$. Uniform normal displacement, associated with the rigid circular punch problem, is represented by $n = -\frac{1}{2}$. The Hertzian pressure distribution is realized by setting $n = \frac{1}{2}$. This case is developed in section 5.2 because of its relevance to the general solution.
As with circular contact, a number of pressure distributions can be specified over an ellipse, leading to elliptic integral solutions. The elliptic integrals are not the result of the elliptic contact area as might be supposed, but arise from integrating over the area using the polar coordinates which locate an arbitrary point. Pressure distributed over elliptical areas forms the basis of the numerical work developed later in this investigation.

Solutions exist for tangential point and distributed loads, axisymmetrical tractions, and torsional loading following similar methods [1]. The case of Hertzian contact can be expressed in terms of simple formulas for cases involving spherical and cylindrical geometries and appears in many mechanics and machine design textbooks, so many engineers have had exposure to them. These other solutions, which are not widely published, are formulated in terms of elliptic integrals with no explicit solution procedures mentioned.
5.4. Spherical Contact

The case of the sphere, which is described by its radius, is a special case of the general contact problem where the surfaces are described by two (the principal and secondary) radii. This case includes a sphere contacting another sphere, a spherical seat, or a plane. The symmetry around the z axis leads to closed-form solutions which makes the development of the theory easier to follow. This section presents a general development of the problem.

The geometry of the undeformed surfaces can be represented by the functions $z=f(r)$:

\[
 z_1 = \left( \frac{1}{2R_1} \right) r^2 \quad \text{and} \quad z_2 = \left( \frac{1}{2R_2} \right) r^2 \tag{5.16}
\]

where $R_1$ and $R_2$ are the radii of the bodies and $z$ is the distance to a point on each body from its side of the tangent plane. The mutual distance between a corresponding point on the two bodies is equal to the difference $z_1 - z_2$ and is sometimes referred to as "the approach of centers". Zero radii are not permitted, as that would be point contact. A plane surface is represented by an infinite radius, leading to $z$ being equal to zero everywhere. Negative radii, shown in Figure 5.6, produce negative $z$ values with the physical interpretation that the center of curvature lies away from the body, as in the case of a spherical seat.

If there is no pressure, the two bodies contact at a single point. This fact should be used to test the contact area reported by computer programs as a limiting case. As they move towards each other, there will be a local deformation around the initial point of contact. Since the bodies and the loading are symmetrical, the shape of the deformed surface will also be symmetrical. The contact area will assume a circular shape of radius $a$, which is assumed to be small in relation to the size of the bodies themselves. Let $w_1$ be the displacement of a point on body 1 due to the local deformation and $w_2$ be the deformation of the corresponding point (meaning the same $r$ but different $z$ coordinate) on the second body. The tangent plane is considered to be a fixed reference plane for this discussion.

Consider a point on each body, on the axis of symmetry, that is far enough from the contact point that the localized deformation can be ignored. These two points will approach each other by an amount $\delta$ and the distance between them will be reduced by $\delta$ -

![Figure 5.6 Geometry for spherical contact](image-url)
Figure 5.7 Shows the approach of two bodies.

\((w_1 + w_2)\). As the compression progresses, points not originally touching come into contact resulting in

\[
\delta - (w_1 + w_2) = z_1 + z_2 = \beta r^2 \quad (5.17)
\]

where \(\beta\) is a constant which depends on \(R_1\) and \(R_2\). So from geometrical considerations, for any point of the surface of contact

\[
(w_1 + w_2) = \delta - \beta r^2 \quad (5.18)
\]

The displacement of any point on body 1 inside the circular contact area is given by the point load solution:

\[
w_1 = \frac{1 - \nu_1^2}{\pi E_1} \int \int p(r) ds d\phi
\]

so that

\[
w_1 + w_2 = (k_1 + k_2) \int \int p(r) ds d\phi \quad (5.20)
\]

In which \(k_1\) and \(k_2\), sometimes called the "effective compliances", are given by
Combining equations 5.18 and 5.20 yields an expression which relates the material properties, pressure distribution, and the deformation:

\[(k_1 + k_2) \int \int p(r) ds dp = \delta - \beta r^2 \quad (5.22)\]

An expression for the pressure distribution \(p(r)\) must be found that satisfies equation 5.22. From the conditions of symmetry, the pressure distribution must be symmetrical about the center of the contact area. A hemispherical pressure distribution satisfies all the boundary conditions and can be shown to result in a unique solution. Assume that the pressure distribution takes the form:

\[p(r) = \frac{P_0}{a} \left( a^2 - r^2 \right)^{1/2} \quad (5.23)\]

The total load \(P\) can be obtained by summation of the assumed pressure distribution over the circular area of radius \(a\), with the result that

\[P = \frac{2}{3} P_0 \pi a^2 \quad (5.24)\]

The area and coordinate system must be defined to be able to determine the displacement at a point from the distributed load. Refer to the geometry in Figure 5.5. Point B is an arbitrary point within the loaded region and is taken as the origin of a polar coordinate system.

By the law of cosines

\[t^2 = r^2 + s^2 + 2rs \cos \phi \quad (5.25)\]

which allows the pressure distribution to be written as

\[p(s, \phi) = p_0 a \left( \alpha^2 - 2\beta s - s^2 \right)^{1/2} \quad (5.26)\]

where

\[\alpha^2 = a^2 - r^2 \quad \text{and} \quad \beta = r \cos \phi. \quad (5.27)\]

Using the result from equation 5.14 and the form of the pressure distribution expressed by equation 5.26, normal displacements inside the circle are obtained by integrating
Axisymmetric Solids of Revolution

\[ w(r) = \frac{1 - \nu^2}{\pi E} \left( \frac{P_0}{a} \right) \int_0^{2\pi} d\phi \int_0^{s_1} (\alpha^2 - 2\beta s^2)^{1/2} ds \]  
(5.28)

Now

\[ \int_0^{s_1} (\alpha^2 - 2\beta s^2)^{1/2} ds = \frac{1}{2} \alpha \beta + \frac{1}{2} (\alpha^2 + \beta^2) \left[ \frac{\pi}{2} - \tan^{-1}(\beta/\alpha) \right] \]  
(5.29)

When \( \phi \) is integrated from 0 to \( 2\pi \), \( \alpha \beta \) and \( \tan^{-1}(\beta/\alpha) \) vanish so

\[ w(r) = \frac{1 - \nu^2}{\pi E} \left( \frac{P_0}{a} \right) \frac{2\pi}{4} (a^2 - r^2 + r^2 \cos^2 \phi) \]  
(5.30)

Completing the integration, the vertical displacement at a given radius is given by:

\[ w(r) = \frac{1 - \nu^2}{E} \left( \frac{\pi P_0}{4a} \right) (2a^2 - r^2) \quad ; r \leq a \]  
(5.31)

For compactness let the compliances in series be represented by

\[ \frac{1}{E^*} = k_1 + k_2 \]  
(5.32)

and

\[ \frac{1}{R^*} = \frac{1}{R_1} + \frac{1}{R_2} \]  
(5.33)

The quantity \( E^* \) can be thought of as the effective material stiffness and \( R^* \) the effective radius of the combined geometry. Substituting the expression for vertical displacements, equation 5.31, into equation 5.18 yields:

\[ \frac{\pi P_0}{4aE^*} (2a^2 - r^2) = \delta - \left[ \frac{1}{2R^*} \right] r^2 \]  
(5.34)

When this equation is evaluated with \( r=0 \) and \( r=a \), two simple expressions for the deflection and radius of the contact area are obtained. By treating these two expressions and equation 5.24 as three equations and three unknowns, the following results are obtained which are in a more useful form for design purposes:

Contact Radius:  
\[ a = \left( \frac{3PR}{4E^*} \right)^{1/3} \]  
(5.35)
Axisymmetric Solids of Revolution

Compression: \[ \delta = \frac{a^2}{R} \]  \hspace{1cm} (5.36)

Maximum Pressure: \[ p_0 = \frac{3}{2} \frac{P}{\pi a^2} \]  \hspace{1cm} (5.37)

The radius of the circular contact area and the contact pressure increase as the cube root of the load. The approach of the two bodies is proportional to the load raised to the two-thirds power.

An integration procedure similar to the one performed above to determine the vertical displacement can be performed to determine the radial displacement. From the displacements, the strains and then the stresses can be determined. The expressions for stress take on different forms in five distinct regions of the contact geometry. These regions are: on the surface inside, and outside the contact area, beneath the surface on the z axis, within the loaded circle, and at all other interior points. Stresses at arbitrary interior points have been obtained by Huber [29] and Morton & Close [30]. *The applicability of expressions to a particular region is typical of closed form solutions.* This is because the initial geometry used to set up the integrations takes slightly different forms. On the surface within the contact area, the stresses are all compressive, except at the very edge where the radial stress is tensile and at its maximum value:

\[ \sigma_r(a,0) = \frac{(1-2\nu) p_0}{3} \]  \hspace{1cm} (5.38)

This component of stress is responsible for the ring cracks that appear when brittle materials are pressed into contact. Recall that brittle materials are notoriously weak in tension. Designers using ceramics should take note.

Beneath the surface, on the z axis:

\[ \frac{\sigma_r}{p_0} = \frac{\sigma_\theta}{p_0} \]

\[ \frac{\sigma_r}{p_0} = (\nu - 1)\left\{1 - \frac{z}{a}\tan^{-1}\left(\frac{a}{z}\right)\right\} + \frac{1}{2}(1 + \frac{z^2}{a^2})^{-1} \]  \hspace{1cm} (5.39)

and

\[ \frac{\sigma_z}{p_0} = - \left(1 + \frac{z^2}{a^2}\right)^{-1} \]  \hspace{1cm} (5.40)

In these equations, the value of \( \nu \) is associated with the body that the stresses are being calculated in. The maximum shearing stress lies beneath the surface on the z axis and is given by
\[ \tau_{\text{max}} = \frac{1}{2} |\sigma_z - \sigma_r| \]  

(5.41)

The stresses on the z axis are principal stresses. Since \( \sigma_\theta = \sigma_r \) on the z axis, the expression for von Mises stress in the general triaxial stress state can be reduced to:

\[ \sigma_0 = \sqrt{\sigma_z^2 - \sigma_r \sigma_z + \sigma_r^2} \]  

(5.42)

**Figure 5.8** is a plot of equations 5.39 through 5.42. It is intended to illustrate the general behavior of the spherical contact problem. The stress components have been normalized with respect to the maximum value in the stress field which is the normal stress at the point of contact. The depth below the surface has been normalized to the radius of the contact area which can be considered a characteristic length of the problem. The contact radius is still small with respect to the radii of the bodies and the stresses have dropped off to less than 10% of the maximum value at a distance of 3a below the surface. This illustrates that the effect of contact stresses is confined close to the surface. Note that \( \sigma_r \) goes tensile when \( z/a > 1.5 \).

The geometry used to create **Figure 5.8** is that of two 1mm diameter steel spheres being pressed together by a 100 N load. The specific values and the shape of the curves will vary if the data were different, but the same general behavior will be exhibited. The figure contains an interesting feature from a practical design point of view. The maximum value of the shearing stress occurs below the surface. The depth at which this maximum occurs is where plastic yielding is first expected to occur. This is based on the Maximum Shear Stress Theory and assuming that the material is ductile. For a design problem, the magnitude of the maximum and its actual location (in mm or inches) below the surface will likely be sought.
The following example is included to reveal the limitations of the spherical solution just presented. Consider a nylon ball compressed between two steel plates. In this problem there is slip at the surface. This is due to two factors: the low coefficient of friction between nylon and steel, and the relatively low elastic modulus of nylon compared to steel which allows large strains to accumulate very rapidly. This situation will exhibit plastic deformation at low loads. It is assumed in Hertzian contact that there is no slip between the solids in the contact area, consequently this problem can not be solved exactly by Hertzian methods, although they will provide a good first approximation.

A no-slip solution has been found for unequal spheres by L.E. Goodman [31] in 1962. The case of tangential force and twisting couple has been investigated by R.D. Mindlin [5]. If this were some type of rapid mass production manufacturing operation, the effects of temperature and strain rate would also have to be investigated. The Hertzian equations can still be used to model the initial stages of deformation. Follansbee and Sinclair [16] studied ball indentation of a strain-hardening solid well into the fully-plastic state. Their work could be used to model the Brinnel hardness test, which uses a 10 mm diameter ball and a 3000kg load. The depth of penetration (plastic deformation) in the Brinnel test corresponds to the hardness of the material.

| Note that although the theory of spherical Hertzian contact has its limitations, there are no singularities in the stress field. From a practical design point of view, this is very important in that it is a closer model of reality than the concentrated load solutions which predict infinite stresses at the point of application of the load. How can something be designed to withstand infinite stresses?! |

5.5. Cylindrical Contact, Another Two-Dimensional Case

Cylindrical contact is of great importance in practical design. Closed-form solutions are available for normal loading that do not involve elliptic functions. Results for normal loading are similar to the spherical case in that maximum shearing stress occurs below the surface etc. Methods are available for tangential loading as well, but are computationally intensive.

Cylindrical contact is one of the limiting cases of elliptical contact. As the ratio of a/b becomes large enough, the contact ellipse becomes a rectangle. An alternate approach is to recognize the two-dimensional nature of the problem from the outset and make use of the results from elasticity for a line-loaded half-space. The problem can be viewed as one of plane stress or plane strain, depending on the physical length of the cylinder. A cylinder is defined by one of the principal radii having an infinite value. Cylindrical contact can arise from a cylinder on a plane, a cylinder in contact with another cylinder when the principal planes are coincident, or a cylinder in a cylindrical trough.
The Smith-Liu [7] equations predict the stresses at any point in the body for cylindrical bodies in line contact with tangential loading. These equations show that when friction is present, the maximum shearing stress can appear at the surface. The maximum stress values are shown to increase in direct proportion to the coefficient of friction. Practical application of these equations include cam rollers, spur gears, cams and followers, and wheels in rolling contact, etc. Rolling contact possesses a few features in addition to those associated with the stationary case. Sliding or tangential loading also introduces additional complications. For problems involving rolling contact, the range of stress amplitude, as the roller moves past a specific location, is usually tracked for fatigue calculations.

Real cylinders have a finite length and significant deviations from the two-dimensional stress distributions occur near the ends. End effects fall into two main categories: rigid punch behavior and overhang of an edge. The degree to which the problem approaches a rigid punch depends on end conditions, i.e. does the cylinder have rounded or square ends? To reduce the stress concentration at the ends, the axial profile of the roller should be slightly barreled. It is theoretically possible to create a tapered roller profile such that the deflection under loading would result in uniform loading along the length. This would be difficult to manufacture and would only provide uniform contact for the calculated load and materials. A study of end effects has been done by Ahmadi, Keer, and Mura [32]. The literature contains extensive references to cylindrical contact problems, so it will not be pursued further here.
6. Elliptical Contact

The local geometry of an arbitrarily shaped body at the point of contact can be approximated by two radii, the principal curvatures, which are the minimum and maximum curvatures of the surface at that point. This approximation is sufficient since the resulting stress distributions are also localized. Elliptical contact areas can result from such geometries. For these reasons, the results of elliptical contact analysis are often called the "general solution". Elliptical contact results cannot be used (accurately) when there is contact at more than one point or the bodies closely conform to each other. The development of the theory parallels the spherical contact problem. The main difference between them being that the geometry is no longer axisymmetric making the resulting equations more complex.
In this section the geometric basis of the elliptical contact problem is developed. It is similar in many respects to the development of the spherical case. The equations in this section are taken from the work of Seely and Smith [8].

In the elliptical case, as with the spherical case, a functional representation of the surface must be assumed in order to develop expressions for distance between the bodies which can then be used to determine displacements. For the reasons described in Section 3, a function of at least second order is necessary to ensure continuous curvature of the surface. The equation used to represent the surface is assumed to take the form:

$$z = Ax^2 + By^2$$

The contact area (at $z = 0$) is represented by the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

An expression must be determined for the distance $z_l$ from the tangent plane to any point on the surface of body 1 at the initial point of contact with no load. Refering to Figure 6.1 let $u_x$ and $v_x$ be the axes in the tangent plane that lie in the planes of principal curvature of body 1.
Elliptical Contact

\[ z_1 = u_1 \tan \frac{1}{2} \beta = \frac{1}{2} u_1 \beta \]  

(6.3)

From triangle HKD where \( \beta \) is a small angle

\[ \tan \beta = \beta = \frac{KD}{HK} = \frac{u_1}{R_1}, \]  

(6.4)

so

\[ z_1 = \frac{1}{2} u_1 \frac{u_1}{R_1'} = \frac{1}{2} \frac{u_1^2}{R_1'}, \]  

(6.5)

Now in the plane containing the principal radius \( R_1 \), by similar analysis

\[ z_1 = \frac{v_1^2}{2R_1} \]  

(6.6)

The distance to any point \( G \) not lying in either plane of principal curvature may be approximated by

\[ z_1 = \frac{u_1^2}{2R_1'} + \frac{v_1^2}{2R_1} \]  

(6.7)

Note that if \( u \) or \( v = 0 \), then equation 6.7 reduces to equations 6.5 or 6.6. Also, if \( z_1 \) is a constant then the equation represents an ellipse. For the second body

\[ z_2 = \frac{u_2^2}{2R_2'} + \frac{v_2^2}{2R_2} \]  

(6.8)

where \( u_2 \) and \( v_2 \) are coordinates in the principal planes of body 2 and \( R_2 \) is the largest principal radii of body 2. The distance between points on the two surfaces is given by

\[ z = z_1 + z_2 \]  

(6.9)

\[ z = \frac{u_1^2}{2R_1'} + \frac{v_1^2}{2R_1} + \frac{u_2^2}{2R_2'} + \frac{v_2^2}{2R_2} \]  

(6.10)

Now this expression for \( z \) can be transformed to eliminate the coordinates \( u_2 \) and \( v_2 \) using the following relations (see Figure 6.2):

\[ u_2 = u_1 \cos \alpha + v_1 \sin \alpha \]  

(6.11a)

\[ v_2 = -u_1 \sin \alpha + v_1 \cos \alpha \]  

(6.11b)
This yields

\[ z = A'u_1^2 + 2H'u_1v_1 + B'v_1^2 \]  \hspace{1cm} (6.12a)

where

\[ 2A' = \frac{1}{R'_1} + \frac{1}{R'_2} \cos^2 \alpha + \frac{1}{R_2} \sin^2 \alpha \]  \hspace{1cm} (6.12b)

\[ 2H' = \left[ \frac{1}{R'_2} + \frac{1}{R_2} \right] \sin \alpha \cos \alpha \]  \hspace{1cm} (6.12c)

\[ 2B' = \frac{1}{R_1} + \frac{1}{R_2} \sin^2 \alpha + \frac{1}{R_2} \cos^2 \alpha \]  \hspace{1cm} (6.12d)
To find the equation of the ellipse with respect to the x-y axes, another coordinate transformation must be performed to eliminate the product term $u_1 v_1$. Let

$$u_1 = x \cos \lambda - y \sin \lambda \quad (6.13a)$$
$$v_1 = x \sin \lambda - y \cos \lambda \quad (6.13b)$$

which after a significant amount of algebra yields:

$$z = Ax^2 + By^2 \quad (6.14)$$

where A and B are the roots of a quadratic equation and are given by

$$A = \frac{1}{4} \left[ \frac{1}{R_1} + \frac{1}{R'_1} + \frac{1}{R_2} + \frac{1}{R'_2} \right]$$

$$- \frac{1}{4} \sqrt{\left[ \left( \frac{1}{R_1} - \frac{1}{R'_1} \right) + \left( \frac{1}{R_2} - \frac{1}{R'_2} \right) \right]^2 - 4 \left( \frac{1}{R_1} - \frac{1}{R'_1} \right) \left( \frac{1}{R_2} - \frac{1}{R'_2} \right) \sin^2 \alpha} \quad (6.15a)$$

and

$$B = \frac{1}{4} \left[ \frac{1}{R_1} + \frac{1}{R'_1} + \frac{1}{R_2} + \frac{1}{R'_2} \right]$$

$$+ \frac{1}{4} \sqrt{\left[ \left( \frac{1}{R_1} - \frac{1}{R'_1} \right) + \left( \frac{1}{R_2} - \frac{1}{R'_2} \right) \right]^2 - 4 \left( \frac{1}{R_1} - \frac{1}{R'_1} \right) \left( \frac{1}{R_2} - \frac{1}{R'_2} \right) \sin^2 \alpha} \quad (6.16b)$$

The constants $A$ and $B$ depend only on the magnitudes of the principal curvatures and on the angle between the principal planes. The expression for the surface is needed to be able to determine the distance between the bodies. Let $w$ denote the displacement due to local compression, then

$$\delta \ (w_1 + w_2) = z_1 + z_2 = z \quad (6.17)$$

$$(w_1 + w_2) = \delta - Ax^2 \quad By^2 \quad (6.18)$$

**Equation 6.18** is obtained from geometrical considerations only and is analogous to **equation 5.18**. The main assumption is that the area of contact is small compared to the radii of curvatures.
6.2. Development of Equations for Elliptical Contact

The original solution for stresses due to normal loading over elliptical contact regions was developed by Hertz [2], but contained only the stresses on the z axis. Hertz noted that equation 6.18 has the same form as that of a Newtonian Potential Equation for the attraction of a homogenous mass M in the shape of an ellipsoid upon a unit of mass at a point P some distance away. This Newtonian potential function satisfies the same differential equations that are required to be solved by the theory of elasticity. The analogy is that the stresses at the contact surface correspond to mass. In 1930 Thomas and Hoerch [3] transformed the Hertzian solution into elliptic integrals. Consult Love [26] and Routh [33] for a development of potential function methods as applied to elasticity problems.

The equations for determining the stresses involving elliptical contact areas can be derived by a number of methods. Elliptical contact is qualitatively similar to the circular contact but the mathematical expressions are more complex. The development of the equations, based on potential theory, is quite lengthy and will not be shown here. The reader is referred to Seely and Smith [8], Johnson [1], and Liu [6]. Of these three, the notation and formulation of Seely and Smith is most easily implemented as a numerical procedure, and is the one used in this investigation.

A more physically intuitive approach for the development of equations to determine stresses in the elliptical problem is to follow a procedure analogous to the one used for spherical contact. The surface of the bodies are now represented by second order functions of x and y instead of a single variable. The expression for the deformation due to a concentrated load is again made use of. The assumed pressure distribution changes from hemispherical to one that is ellipsoidal and is given by:

\[ p(x,y) = p_0 \left[ 1 - \frac{(x/a)^2 - (y/b)^2}{1/2} \right] \]  \hspace{1cm} (6.19)

This technique of integrating the point load displacement results, with an assumed pressure distribution, over a given area, is is sometimes called "The Method of Singularities". The pressure distribution expressed in equation 6.19 is integrated over the elliptical contact area to obtain the vertical displacement. Two more integrations are required to obtain the x and y displacements. Once the displacements are known the strains can be determined, then from these the stresses. This is easier said than done because the integrations are quite complex.

The resulting vertical displacement within the ellipse (obtained by potential function methods) is given by equation 6.20 which is analogous to equation 5.31 in the spherical problem.

\[ w(x,y) = \frac{1 - \nu^2}{\pi E} (L - Mx^2 - Ny^2) \]  \hspace{1cm} (6.20)
\[ w(x, y) = \frac{1 - \nu^2}{\pi E} (L - Mx^2 Ny^2) \] (6.20)

where

\[ M = \frac{\pi p_0 b}{k' a^2} \left[ K(k') - E(k') \right] \] (6.21)

\[ N = \frac{\pi p_0 b}{k' a^2} \left[ \frac{a^2}{b^2} K(k') - E(k') \right] \] (6.22)

\[ L = \pi p_0 b K(k') \] (6.23)

\( K(k') \) and \( E(k') \) are complete elliptic integrals and are discussed in a subsequent section. The eccentricity of the contact ellipse, \( k' \), is independent of the load and depends only on the relative curvatures of the bodies and is given by

\[ k' = \left[ 1 - \frac{b^2}{a^2} \right]^{1/2} \] (6.24)

Finding the components of stress and displacement at a general point in the solid requires solving for the root of a cubic equation. This root is then used as an upper limit for one of the integrals of the potential function method. Expressions for the stresses at all points in the solid have been developed by Sackfield and Hills [2], [3], and [4]. The expressions for stresses are written in terms of integrals, which are not a form that is readily useful for design purposes. Equations are simplified when stresses sought are on the surface along the x and y axis, or on the z axis, where they are maximum (for the normal load problem).
6.3. Elliptical Solution Overview

The solution for stresses on the z axis, where the stress values are maximum in the body, is of practical interest for design problems. The expressions shown below can be found in Seely and Smith [8]. The principal stresses are obtained from:

\[
\sigma_1 = \frac{b}{\Delta} [M(\alpha_z - \nu\alpha'_z)]
\]

(6.25a)

\[
\sigma_2 = \frac{b}{\Delta} [M(\alpha_y - \nu\alpha'_y)]
\]

(6.25b)

\[
\sigma_3 = \frac{b}{\Delta} \left[ \frac{M}{2} \left( \frac{1}{n} - n \right) \right]
\]

(6.25c)

In which

\[
\alpha_z = k \frac{z}{b} \left[ F(\phi,k') - H(\phi,k') \right] - \frac{1-n}{2}
\]

(6.26a)

\[
\alpha_y = k \frac{z}{b} \left[ F(\phi,k') - H(\phi,k') \right] + \frac{1}{2n} + \frac{1}{2} - \frac{n}{k^2}
\]

(6.26b)

\[
\alpha'_z = k \frac{z}{b} \left[ \left( \frac{1}{k^2} \right) H(\phi,k') - F(\phi,k') \right] - \frac{n}{k^2} + 1
\]

(6.26c)

\[
\alpha'_y = k \frac{z}{b} \left[ \left( \frac{1}{k^2} \right) H(\phi,k') - F(\phi,k') \right] + n - 1
\]

(6.26d)

\[
M = \frac{2k}{k^2 E(k')}
\]

(6.27)

\[
\Delta = \left( \frac{1}{A+B} \right) \left[ \frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2} \right]
\]

(6.28)

\[
b = 3 \sqrt{\frac{3kE(k')}{2\pi}} (P\Delta)
\]

(6.29)
Elliptical Contact

\[ n = \sqrt{\frac{k^2 + k^2(z/b)^2}{1 + k^2(z/b)^2}} \]  \hspace{1cm} (6.30)

\[ \phi = \cot^{-1} \left[ \frac{z}{b} \right] \]  \hspace{1cm} (6.31)

The eccentricity of the ellipse \( k' \), can be determined from

\[ R(k') = \frac{B}{A} = \frac{(1/k^2)E(k') - K(k')}{K(k') - E(k')} \]  \hspace{1cm} (6.32)

The values \( B \) and \( A \) are obtained from the geometry. These equations are not organized in a manner that are readily usable for design calculations. The implementation of this solution is the topic of the following sections.
6.4. The Elliptic Integral Primer

Before proceeding with the implementation of the solution of the elliptical contact problem, the subject of elliptic integrals will be investigated further. This section introduces elliptic integrals and some of the methods used to evaluate them.

Elliptic integrals frequently arise in the study of contact problems and have numerous applications to other physical problems. The distinguishing feature of these integrals is the integrand

\[ \sqrt{1 - k'^2 \sin^2 \theta} \quad (6.33) \]

The parameter \( k' \) is called the modulus and typically varies from 0 to 1. Other values of the modulus including imaginary ones can exist mathematically but the associated integrals are usually transformed such that \( 0 < k' < 1 \). If \( k' \) has a value of 0, then the integrand has the value 1. If \( k' = 1 \), then the integrand becomes a standard trigonometric integral. Any intermediate value of \( k' \) necessitates numerical integration. Other forms the integrand exist that can be expressed as polynomials. A complementary modulus, \( k \), is sometimes used and is related to the modulus by

\[ k = \sqrt{1 - k'^2} \quad (6.34) \]

Only numerical methods of evaluation are known, although it has not been proven that closed-form expressions do not exist. Perhaps some future symbolic mathematics program will reveal one. They can be evaluated by any numerical integration procedure. There are fundamental integrals called Legendre's canonical elliptic integrals of the first, second, and third kinds. Other forms exist, but they can all be expressed as algebraic combinations of these three.

The Legendre integrals of the first and second kinds are of interest in contact problems. In addition they are further classified as Complete or Incomplete. The limits of integration of these integrals represent angles; the lower limit is typically zero, and the upper limit can range from zero to \( \pi/2 \). When the upper limit equals \( \pi/2 \), the integral is termed complete, anything less is incomplete. Each "kind" (1st or 2nd) of integral can be visualized as a surface that is a function of two variables, the modulus \( k' \) and the upper limit of integration \( \phi \), called the argument. In physical applications, the variables of the problem (or combinations of them) are transformed into values of \( k' \) and \( \phi \) that lie in the first quadrant, i.e. both having positive values. Three-dimensional plots of elliptic integrals and other mathematical identities, can be found in Jahnke and Emde [37]. The elementary elliptic integrals are often given single letters to represent them and are defined in the following section.
6.4.1. Complete Elliptic Integrals

Equations 6.35 and 6.36 are one form of the definition of the complete elliptic integrals.

1st kind: \[ K\left(\frac{\pi}{2}, k'\right) = \int_0^{\pi/2} \frac{d\theta}{\sqrt{1 - k'^2 \sin^2 \theta}} \quad (6.35) \]

2nd kind: \[ E\left(\frac{\pi}{2}, k'\right) = \int_0^{\pi/2} \sqrt{1 - k'^2 \sin^2 \theta} \, d\theta \quad (6.36) \]

Series representations for these integrals also exist which can be computed faster than by a numerical integration procedure such as Simpson's rule or Gaussian quadrature. A series has the advantage of easily being computed to a specified number of significant digits, by stopping calculation when the magnitude of individual series terms become smaller than the desired precision. The numerical routines to evaluate them can be written in assembly language for optimum speed. The series expansions for the complete elliptic integrals are as follows:

\[ E\left(\frac{\pi}{2}, k'\right) = E(k') = \frac{\pi}{2} \sum_{m=0}^{\infty} \frac{1}{1 - 2m} (m^{-1/2})^2 k'^{2m} \quad (6.37a) \]

which expands as

\[ E(k') = \frac{\pi}{2} \left[ 1 - \frac{1}{4} k'^2 - \frac{3}{64} k'^4 - \frac{5}{256} k'^6 - \frac{175}{16384} k'^8 - \frac{441}{65536} k'^{10} \ldots \right] \quad (6.37b) \]

and

\[ K(k') = \frac{\pi}{2} \sum_{m=0}^{\infty} \frac{\left(\frac{1}{2}\right)_m \left(\frac{1}{2}\right)_m}{m! m!} k'^{2m} \quad (6.38a) \]

\[ K(k') = \frac{\pi}{2} \left[ 1 + \frac{1}{4} k'^2 + \frac{9}{64} k'^4 + \frac{25}{256} k'^6 + \frac{1225}{16384} k'^8 + \frac{3969}{65536} k'^{10} \ldots \right] \quad (6.38b) \]
Graphs of the complete elliptic integrals of the first and second kinds are shown in Figure 6.3. Both complete integrals attain the value of $\pi/2$ as the modulus goes to zero.

![Figure 6.3 Complete Elliptic Integrals](image)

Other standard forms appear in contact problems that can be written as combinations of the fundamental integrals:

\[
B = \frac{E - (1 - k'^2)K}{k'^2} \quad (6.39)
\]

\[
C = \frac{(2 - k'^2)K - 2E}{k'^4} \quad (6.40)
\]

\[
D = \frac{K - E}{k'^2} \quad (6.41)
\]
6.4.2. Incomplete Elliptic Integrals

An incomplete elliptic integral is one in which the upper limit is not a multiple of \( \pi/2 \). The two fundamental incomplete elliptic integrals are defined as

1st kind: \[ F(\phi, k') = \int_{0}^{\phi} \frac{d\theta}{\sqrt{1 - k'^{2} \sin^{2}\theta}} \] (6.42)

2nd kind: \[ H(\phi, k') = \int_{0}^{\phi} \sqrt{1 - k'^{2} \sin^{2}\theta} \ d\theta \] (6.43)

The use of the symbols \( K, E, F, H, B, C, \) and \( D \) to represent the elliptic integrals is common in the literature, though not universal. For additional information on elliptic integrals, consult Byrd [38].
Explicit Numerical Procedures for Normally Loaded Elliptical Contact.

From the solution overview in section 6.3 it is not immediately apparent where or how to begin the solution. This section contains explicit procedures that are used to determine contact stresses. It is the basic "core analysis". Later it is shown how this procedure can be embedded in larger iterative procedures to solve real engineering problems.

The results of the solution are the principal stresses below the surface on the z axis. These three elements of the stress tensor are sufficient for the use of most failure theories. The required input variables for this solution method can be broken into three categories:

- **Geometry**: $R_1, R', R_2, R_2'$ and $a$.
- **Material properties**: $v_1, v_2, E_1, E_2$.
- **Applied Loading**: Resultant applied normal force $P$.

The solution, based on equations 6.25 through 6.32 can be implemented on the computer in 10 clearly defined steps. The 10 steps must be repeated at other physical locations to determine the maximum values occurring in the stress field. The locations of the maxima of stresses are most sensitive to the radii of the bodies. It is assumed that the geometry has been checked for validity as per the conditions discussed in Section 3.

Validation of the input geometry can be considered STEP #0. 

**STEP #1** Determine $A$ and $B$

$$A = j_1 + j_1' + \ldots + R_j R_j R_1 R_1 \ldots$$

$$B = 4 \left( R_1 + R_2 + R_1 R_2 \right)$$

where

$$j_k = \left( R_k \right)^2 \sin^2 a$$
A computationally more efficient expression is of course produced for the numerical evaluation. Note that the difference between the principal curvatures is a recurring factor.

**STEP #2 Determine the elliptic modulus**

This is one of the steps which makes the determination of contact stresses inherently numerical. Equation 6.45, designated as the Ratio function \( R(k') \), must be solved numerically for the value of \( k' \). A plot of this function is shown in Figure 6.4. The value of the function is known from the input geometry and is equal to \( B/A \). When \( k' \) equals 0, which corresponds to spherical contact, \( R(k') \) becomes undefined. A value of 1 corresponds to line contact. Both ends of the range of \( k' \) degenerate into two-dimensional cases which can be solved by closed-form methods.

\[
R(k') = \frac{B}{A} = \frac{(1/k^2)E(k') - K(k')}{K(k') - E(k')} \quad \text{(6.45)}
\]

The bisection method is chosen to solve equation 6.45 for \( k' \) because it is guaranteed to converge and easy to implement. A Golden Section method can be used in the interval subdivision. At each iteration of the solution, both of the complete elliptic integrals, \( E \) and \( K \), must be evaluated. The value of \( k' \) is determined by a subroutine called GET_K in which the series solutions are used to evaluate the integrals. The values \( k \), \( E \), and \( K \) are a by-product of this routine, which are used in other steps of the solution procedure. For extremely large values of \( B/A \), the eccentricity of the ellipse approaches 1, and it may be simpler (more expedient) to consider the problem to be two-dimensional and invoke the appropriate, less complicated, subroutines. The value of \( k' \) needs to be computed only once for each geometry. If this solution procedure were incorporated into an iterative non-linear solution, this step would have to be repeated each time the geometry was updated.

Figure 6.4 The Ratio Function.
STEP #3 Determine intermediate constant M

\[ M = \frac{2k}{k'^2 E(k')} \quad (6.46) \]

When the geometry is circular, \( k' \) is 0, and equation 6.46 becomes undefined. The consequence of this is that this method, which is billed as the "general solution", doesn't even work for the simple case of spherical contact. A robust computer program should parse the input data and detect these geometries, then invoke the appropriate subroutines to handle them. This constant is evaluated once for each geometric configuration.

STEP #4 Determine intermediate constant \( \Delta \)

\[ \Delta = \left[ \frac{1}{A+B} \right] \left[ \frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2} \right] \quad (6.47) \]

Delta can be thought of as a spring constant representing the stiffness of the contact. Poission's ratio and the elastic modulus combine into an effective material stiffness. A and B represent stiffness contributions from the geometry.

STEP #5 Determine ellipse dimension

The dimension of the minor axis ellipse of the contact area is given by

\[ b = \sqrt[3]{\frac{3kE(k')}{2\pi} (P\Delta)} \quad (6.48) \]

STEP #6 Determine intermediate constant \( n \)

\[ n = \sqrt{\frac{k^2 + k^2 (\frac{z}{b})^2}{1 + k^2 (\frac{z}{b})^2}} \quad (6.49) \]

STEP #7 Transform physical dimensions into the elliptic integral argument

A change of variables is used in the solution for stress. The variable \( z \) represents the depth below the surface and can assume values from zero to infinity. Remember that these equations have their roots in the elastic half-space world where the material is
assumed to extend to infinite dimensions. In this calculation, the depth $z$ is transformed or \textit{mapped} onto an angle $\phi$ having a value of $\pi/2$ radians at the surface, and zero, at an infinite depth into the material. Let the distance $z$ be called the physical depth and $\phi$ can be the mapped depth. The physical depth is divided by the minor axis ellipse dimension $b$ as part of its conversion into the mapped depth angle $\phi$ which is the argument of the incomplete elliptic integrals.

\[
\phi = \cot^{-1} \left( \frac{z}{b} \right)
\]  \hfill (6.50)

which is equivalent to

\[
\phi = \tan^{-1} \left( \frac{b}{z} \right)
\]  \hfill (6.51)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6_5.jpg}
\caption{Conversion of depth below the surface into the elliptic integral argument.}
\end{figure}

\textbf{Figure 6.5} is a plot of equation 6.51. A result of this mapping is that the depths close to the surface are well represented while those far below it, not as well. Consider depths spaced at a regular interval of the variable $\phi$ such as every 0.01 radian, or perhaps an even finer limit imposed by the discrete representation of the numbers by a computer. If these angles are converted back to physical depths below the surface, the spacing between the $z$ values becomes farther apart at increasing depths. When using $\phi$, near the surface, the resolution with which incremental changes in distance can be represented is increased. This is a desirable feature considering the localized nature of contact stresses. All of this is a consequence of the mapping; the intended purpose is to allow the use of trigonometric relations for integration. In the numerical procedures, it is more convenient to work with the variable $\phi$ to represent depth below the surface.
Conversion back to $z$ is typically done at the end of the solution to give the results physical dimensions that can be used for design purposes. Normalizing with respect to the contact area dimension $b$ can indicate to what degree a particular effect is localized.

**STEP #8 Evaluate the incomplete elliptic integrals**

The incomplete elliptic integrals of the first and second kind, $F(\phi, k')$ and $H(\phi, k')$ must be evaluated at each point where the stresses are to be determined, the location specified by the angle $\phi$. Numerical integration is performed by a modified Simpson's rule. The integrals (and other stress values) are computed in double precision and rounded to single precision for the final results. A step size is chosen such that the results of the numerical integration are accurate to the limit of single precision variables, i.e. 0.000001. The numerical values for the evaluated elliptic integrals duplicate exactly the tables found in Byrd [38].

**STEP #9 Evaluate intermediate omega constants**

\[ \omega_x = k \frac{z}{b} [F(\phi, k') - H(\phi, k')] - \frac{1-n}{2} \]  
\[ \omega_y = k \frac{z}{b} [F(\phi, k') - H(\phi, k')] + \frac{1}{2n} + \frac{1}{2} - \frac{n}{k^2} \]  
\[ \omega'_x = k \frac{z}{b} \left[ \left( \frac{1}{k^2} \right) H(\phi, k') F(\phi, k') \right] - \frac{n}{k^2} + 1 \]  
\[ \omega'_y = k \frac{z}{b} \left[ \left( \frac{1}{k^2} \right) H(\phi, k') F(\phi, k') \right] + n - 1 \]

Computational efficiency is increased by making use of recurring terms.

**STEP #10 Evaluate the Principal Stresses**

\[ \sigma_1 = \frac{b}{\Delta} [M(\omega_x - \nu \omega'_x)] \]  
\[ \sigma_2 = \frac{b}{\Delta} [M(\omega_y - \nu \omega'_x)] \]
\[
\sigma_3 = \frac{b}{\Delta} \left[ \frac{M}{2} \left( \frac{1}{n} - n \right) \right] \tag{6.53c}
\]

These are the expressions for principal stresses on the z axis. The value of Poisson’s ratio used depends on which body the stresses are being calculated in. The maximum shearing and Von-Mises stresses can be computed from these three values for use in failure theories. The stress invariants can also be determined for use in plasticity calculations. The VonMises Stress is given by

\[
\sigma_0^2 = \frac{1}{2} \left[ (\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 \right] \tag{6.54}
\]

Additional results or quantities not directly part of the 10 step method for determining the stresses, but still important for design purposes, are the deflection and contact area. The deflection or approach between remote parts of the bodies is given by:

\[
\delta = \frac{3kPK(k')}{2\pi} \left[ \frac{A}{B} \right] \left[ \frac{\Delta}{b} \right] \tag{6.55}
\]

This formula can be used in typical bearing and shaft design calculations. Deflections due to contact stresses are typically small, however in certain applications such as machine tools, deflections of 0.01 mm can be significant. Deflections of the individual roller elements inside of a bearing assembly can be obtained.

The contact area is given by

\[
A_c = \pi ab \tag{6.56}
\]

The area could be important in determining the heat transfer between a body and a contacting surface probe. The maximum normal stress at the surface is also an important design parameter.
Maximum shearing octahedral stress are elliptical contact with the axis depending on the location can differ. In this approach, the elliptical stress can be expressed as ellipse (0,0,z).

The ty(0,0,z) solution is determined in the axis of symmetry. The elliptical stress is evaluated at the principal stress. It requires a lot of depth of calculating the primary issue. The design involves the selection of the location of the elliptical stress.

The value of the stresses can be determined from the principal stress. Since the stress is slightly complicated, it is necessary to evaluate the stress at different points. The elliptical stress is specified in the body. The elliptical stress is evaluated by the elliptical stress through the elliptical stress.

The elliptical stress is specified in the body. The elliptical stress is evaluated by the elliptical stress through the elliptical stress.

The elliptical stress is specified in the body. The elliptical stress is evaluated by the elliptical stress through the elliptical stress. The elliptical stress is specified in the body. The elliptical stress is evaluated by the elliptical stress through the elliptical stress. The elliptical stress is specified in the body. The elliptical stress is evaluated by the elliptical stress through the elliptical stress.
Stress vs. Depth (Elliptical Case)

Figure 6.6
Elliptical Contact

From a practical point of view, an infinite depth into the material is meaningless and the stresses there would be of no interest anyway since the peak contact stresses are known to be at or near the surface. Consider a depth into the material that corresponds to the center of curvature of the contact patch (visualize the projected body). This point is presumably far enough away from the contact point that St. Venant’s Principal applies, i.e. the effects from contact at the surface, are negligible. A depth angle \( \phi_c \) exists that corresponds to a physical depth equal to the largest principal radii of the contacting bodies. For this reason it is not necessary to calculate the stresses beyond \( \phi_c \) to an infinite depth, especially since this involves a significant amount of computation.

The key feature of the elliptical contact (normal loading only) problem is that the peak octahedral shearing stresses occur below the surface. This has been attributed to the cause of failure in many contact problems involving ductile materials. Yielding first occurs below the surface, then propagates to the surface. The root depth of pitting failures correlates closely to the location of the peak subsurface octahedral shearing stress. This depth can be used in determining the minimum depth of heat treating. Hardness levels at or beyond the depth corresponding to the maximum octahedral shear stress must be increased to a level that can resist yielding, or to the endurance limit of the material (if one exists), depending on the application. To do this, the mapped depth corresponding to the peak subsurface shearing stress must be converted back to physical depth.

The question of how to best determine the location of the maximum octahedral shearing stress remains. One possible method would be to calculate the stresses over the full range of depth angles, then pick the maximum from the values generated. This brute-force technique requires an unnecessary amount of computation since the values of \( \phi \) that approach 0 radians, i.e. infinite physical depth, are far removed from the contact point. It doesn’t even guarantee a maximum, since the stresses are calculated on depth intervals that are specified ahead of time. The coincidence of a maximum existing on a pre-specified point is remote. Calculating stresses for the full range of depth angles is useful for creating graphs to illustrate the overall characteristics of the solution.

A second method involves calculating the derivative of the maximum shearing stress with respect to depth. Starting at the surface and calculating stresses at increasing depths, the derivative can be approximated as a straight line slope between the two most recent depths, or using a weighted average of the last few to alleviate round-off errors. When the derivative equals zero, or is within some specified tolerance, the maximum and its location have been found.

Since the value of shearing stresses is assumed to be a continuous function of depth below the surface, a form of the bisection technique can be used to locate the maximum. This third method has the advantage of being able to determine the actual location of the maximum and probably uses the least amount of computation overall.
The actual values of the stresses were found to vary due to round-off error and the binary representations stored by the computer. Small fluctuation in a supposedly continuous function makes it difficult to determine which of two values is larger. The apparent values can lead the bisection method to discard the wrong half of the interval for that iteration. To overcome this difficulty, it was found necessary to perform calculations in double precision and then truncate the results to single precision. In all cases, it is more convenient to work in terms of the mapped depth angle, then convert the depth back to physical depth for the final answers.

Seely and Smith [8], the original source for the equations implemented here, present an outline for this elliptical solution, then proceed to show how to obtain numerical values from tables and charts they have developed. Determining the value of stresses and deflections can only be as accurate as interpolation from a graph. The multi-step procedure is also very tedious to do by hand. The computer program(s) written for this investigation produce values nearly identical (to within 2 significant digits) to the example problems in their book. The small differences are attributable to more accurate evaluation of the elliptic integrals by computer. They may have also rounded their numbers for publication (to achieve clarity for the textbook example).

These 10 steps can also be used in other types of iterative procedures to determine design-related information such as: the maximum load the configuration can take before yielding occurs, fatigue life, and at what depth stresses exceed a particular value. Parametric studies could also be undertaken to determine the effect of changing the local radii or material properties. The problem-space could be investigated by examining the effects of varying two of the ten variables at a time and viewing a third variable of interest as three-dimensional graphs. This could illustrate the sensitivity of the third variable, which could be total deflection or maximum shearing stress, to the effects of changing the other two, perhaps the principal radii or a material property.
6.7. Contact Stresses 101 (Benchmark Testing)

The computer implementation of the elliptical solution as outlined in the two previous sections, duplicates the numerical results of the example problems found in Seely and Smith almost exactly. The numerical routines for evaluating the elliptic integrals duplicate tabulated results to six decimal places, and more accuracy could be obtained by reducing the integration step size and using a higher precision representation for numbers in the computer. These facts generate confidence in the numerical results produced by the computer program developed for this investigation. In a practical problem, the number of significant digits of the input data would probably be the limiting factor in the accuracy of the results. Any radius will have a plus or minus tolerance associated with it. What uncertainty is assigned to Poission's ratio? Elastic modulus data is derived from the average of numerous tests so what is the exact value of an individual material sample? What are the effects of surface roughness? There is a theoretical limit to the model since actual materials are not truly isotropic, but have a discrete structure when examined in close enough detail. The computer can produce numerically accurate values that exceed the limits to which it is reasonable to extrapolate the model on which the elliptical theory is based.

An additional approach is taken here to determine the validity of the solution by comparison of the computer results to a known solution. The method of computing stresses by evaluating elliptic integrals cannot be used for the simple case of circular contact due to a singularity in step #3 of the 10 step method. However, the elliptical solution should asymptotically approach circular solution. To test this, the case of spherical-to-spherical contact is compared to one that is nearly spherical, i.e. a ball and a "near" ball. Consider two 1mm diameter spherical ball bearings contacting each other. The stresses can be evaluated using the closed-form solution expressed by equations 5.39 and 5.40. Now suppose one of the balls deviates from being a sphere, having a principal radius of 1mm and a secondary radius of 1.01mm as illustrated in Figure 6.7. This could represent an imperfect ball bearing due to manufacturing processes. To accurately determine the stresses, the 10 step procedure described in equation 6.44 through 6.54 must be used. This test pushes the solution method to its limit. The

Figure 6.7 Geometry for benchmark evaluation of elliptical solution.
calculated values should agree closely with those associated with the simple spherical case. The benchmark data used is shown in Table 6.1.

The following are intermediate computed values from the elliptical case for those interested in the details of the calculation. The eccentricity of the contact ellipse $k'$, is 0.9967 which is within 1/2% of the value of 1.0 associated with circular contact. Thus this benchmark test produces contact as nearly circular as would be produced by the contact of two perfect spheres. The complete elliptic integral $E(k') = 1.568205$, which would equal $\pi/2$ (1.570796) for circular contact. The dimension of the minor axis of the ellipse is $b = .6997E-04$, which compares to the circular contact radius of .7003E-04 meters. The maximum shearing stress in the non spherical body occurs at a mapped depth of $\phi = 1.12$ radians.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Units</th>
<th>Spherical 1%</th>
<th>1%</th>
<th>5%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Shearing Stress</td>
<td>m²</td>
<td>.3040E+10</td>
<td>.3036E+10</td>
<td>.3021E+10</td>
<td>.3002E+10</td>
</tr>
<tr>
<td>Depth of Maximum Shear *</td>
<td>m</td>
<td>.000034</td>
<td>.000034</td>
<td>.000034</td>
<td>.000035</td>
</tr>
<tr>
<td>Contact Area</td>
<td>N/m²</td>
<td>.1541E-07</td>
<td>.1543E-07</td>
<td>.1553E-07</td>
<td>.1565E-07</td>
</tr>
<tr>
<td>Deflection</td>
<td>N/m²</td>
<td>.9809E-05</td>
<td>.9801E-05</td>
<td>.9769E-05</td>
<td>.9733E-05</td>
</tr>
<tr>
<td>Max Normal Stress</td>
<td>m</td>
<td>-.9733E+10</td>
<td>-.9718E+10</td>
<td>-.9657E+10</td>
<td>-.9585E+10</td>
</tr>
</tbody>
</table>

Table 6.2 Results of benchmark comparison. The spherical results were obtained by closed form solution. The Additional columns indicate elliptical contact in which the secondary radius of the body was increased by the percentage indicated over the equivalent spherical radius. * Note: an actual search via the bisection method for the maximum shearing stress below the surface was not done, instead a table of values was computed at a regular depth interval to obtain a direct comparison between the geometries. From these values the maximums were picked.

The values summarized in Table 6.2 correspond very well between the circular and limiting elliptical case, and could be made closer by assuming the secondary radius of the body to be 1.001 rather than the 1.01 mm diameter used for the illustration. The elliptical case yields slightly smaller stress values than the circular one, since the contact area with which to distribute the load is larger. While the difference in the magnitude of stresses is small for this benchmark case, it may be more significant than it first appears. Such a small discrepancy would not likely make a difference in a one time yielding, but may contribute to a change in the probabilistic failure rate or mean time between failures (MTBF) for a batch, lot, or group of precision aircraft quality bearings.
This benchmark test shows that the elliptical method can be used to calculate the case of spherical contact asymptotically, although it is not the most computationally efficient way of doing so. The numerical results of these benchmark cases are so close that when graphs of the spherical and elliptical methods are superimposed as an overlay, the curves are indistinguishable. Deviations in geometries with larger eccentricities can however be illustrated graphically.

In summary, the 101 benchmark test gives excellent results even at the limits of the range of applicability pertaining to this solution method. The stresses reported by the 101 problem and the closed-form spherical solutions are so close as to be indistinguishable when plotted on the same graph. A similar type of benchmark test could be performed by asymptotically approaching line contact for which the results would also be expected to agree very closely with the closed-form solutions. It is important to remember that graphs and charts can be misleading because they represent only one "data set instance" versus the entire class. Curves plotted in the literature typically show the particular data set generated using a value of Poisson's ratio of \( \nu = 0.3 \).

This solution for elliptical contact stress problems is often called the general solution in the literature, when in fact there is no such thing. It does provide some useful results but does not cover all cases. It is incomplete or limited in the following ways, because it is:

1) Not valid at the surface, at the very point of contact itself.
2) Does not give expressions for principal stresses off the z axis.
3) Has a singularity for circular geometries.
4) Doesn’t specify which Poisson’s ratio is to be used in calculating the stresses. It is assumed to be for the body in question.
5) Does not account for friction.
6.8. The Real Contact Problem

The analysis until now has assumed the loading at the contact surface contains no tangential components. In the contact between machine elements tangential loading is present in a large number of cases. Consider power transmitted at the start up of a rolling system, frictional drag, and sliding-braking effects. Think of gears, roller bearings, wheels, cams and followers etc. The normal loading elliptical solution does not address these problems thus, the real problem in contact stress analysis is to determine the stress distribution in the presence of tangential loading. Normal loading elliptical contact is typically used for the stationary or static loading cases. Consider the wheel of a railroad car at rest. If a design cannot withstand the static loading case, then it surely cannot endure the higher stresses associated with tangential and dynamic loading. In tangential loading, the normal component is still present, so the normal loading solution, though limited in scope, must still be obtained as a minimum requirement.

Seely and Smith [8] have developed equations for cylinders subjected to tangential loading. They do not involve elliptic integrals, but are elaborate and not suited for hand calculations. The most important result of their work is that, in the presence of friction, the patterns for stress become distorted from those of the normal load case, such that the maximum shearing stress can occur at the surface instead of below it on the z axis. Tensile stresses at the surface can lead to rapid failure by crack propagation and stress concentrations due to surface defects. Similar distorted behavior is to be expected from the tangentially loaded elliptical contact problem. In this case the stress patterns would be three-dimensional. Graphical illustration of the results could best be obtained by taking section views perpendicular to all three coordinate planes at regular spacings away from each of them. On each of these section planes any element of the stress tensor or combined stress, such as Von Mises, could be plotted. The search for the maximum shearing stress would also have to proceed in three-dimensions.

The stress at the surface for the case of sliding elliptical contact has been studied by Vermeulen and Johnson [39] and, not surprisingly, involves elliptic integrals. A summary of their analysis follows.

The equations are based upon incipient sliding under a tangential force less than limiting friction, i.e. $Q < \mu P$. A central elliptical region with no relative movement between the bodies is presumed to exist with an annular region of slip. The complete stress tensor at any point on the contact surface is explicitly determined. The calculations are straightforward, but require the use of a computer to be of any practical value. The article was written in 1964 when the required computational computational power was not widely available.
6.8.1. Surface displacements

The displacements on the surface can be expressed as:

\[ u(x,y,0) = Xpa f^2 G \]

and

\[ v(x,y,0) = 2L - 6 T - 9 ab x^2 v^2 a^2 b^2 \theta^2 b \]

The displacements are expressed in terms of a set of functions \( T, \phi, \), and \( \theta \) that depend on Poisson's ratio and the relative value of the coefficients of the contact ellipse \( a \) and \( b \) as shown in Table 6.3. \( K, E, B, C, \) and \( D \) are the complete elliptic integrals as defined in Section 6.4. When \( a = b \), the contact is circular and the elliptic integrals disappear.

\[ (6.61a) \]
\[ (6.61b) \]
\[ (6.61c) \]
\[ (6.61d) \]

6.8.2. Surface Stresses

The surface stresses within the ellipse of contact due to the tangential traction acting alone can be obtained from the displacements. In particular:

\[ 2G T du, \]
\[ 3v \]
\[ Xq \]

\[ a \]

\[ (6.62a) \]
\[ (6.62b) \]
\[
\tau_{xy} = G \left[ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right] = -X_0 \left[ \frac{a}{b} \psi - \frac{1}{2} \Theta \right] \frac{y}{b}
\] (6.62c)

These equations predict a high tensile stress at the end of the axis parallel to the tractive force, i.e. at the point (-a,0,0). This stress adds directly to the tensile stress around the periphery of the contact ellipse due to the normal load.

For purely normal loading, and small values of \( \mu \), yielding is initiated below the surface (by either the Von Mises or Tresca criterion). However, when \( \mu \) exceeds a value of about 0.3 (0.26 for long cylinders and 0.31 for spheres), yielding first occurs at the surface. For ellipses whose eccentricity is not extreme (\( k' < 0.85 \)), the point of initial yield is still located at (-a,0,0). When the eccentricity is large, the location of the maximum shear stress moves towards the center of contact.
7. Practical Applications

The following sections attempt to address how realistic the answers from the numerical elliptical contact solution are. The view taken is not that of numerical error analysis, but in questioning the assumptions of the continuum model. Some of the differences between theory and practice are highlighted. The goal is to show the relevance, and expose the limitations, of contact stress theory in relation to mechanical engineering design. Included are: areas of current research, statistical analysis of surfaces, models of frictional behavior, lubrication, wear, and design examples. It will also be shown that the subject of contact mechanics touches on plasticity theory, tribology, materials science, and manufacturing methods.

One aspect of engineering is to determine how phenomena occurring in nature can be simplified for solving problems. In mechanics, bodies have been modeled as continua whose mechanical properties can be characterized by a few material constants such as Poisson's ratio and the modulus of elasticity. The actual number of constants depends on the complexity of the model. The Generalized Hooke's Law covers a wide variety of linear phenomena, including anisotropic materials. Other more complex models consider whether the material is time and temperature-dependent (e.g. creep and relaxation), or even frequency-dependent as in the case of some elastomers. Through the use of these models many real engineering problems have been solved. Certain areas however, require further investigation.

7.1. Gaussian Profiles and the Rusty Nail

7.1.1. On Surfaces and their Statistical Descriptions

In reality, all surfaces are rough when compared to molecular dimensions. Views on the true nature of surface contact are influenced by the techniques used to measure surface topography and the methods used to generate the surfaces in manufacturing processes. When viewed close-up, surfaces have many peaks and valleys as illustrated in Figure 7.1. Manufactured surfaces are not truly random, due to tool marks or finishing processes which typically create periodic surface waveforms. A sharp tool bit can create a triangular waveform on a turned shaft. The statistical distributions of the height and spacings of the surface peaks are often assumed to have Gaussian

![Figure 7.1 Rough Surface](image)
profile. Surface measurements show the Gaussian assumption to be good, at least as a first approximation, thus it makes the problem more mathematically tractable. A fair example of a Gaussian surface is probably a rusty nail. Its surface is highly random, due to the chemical degradation from oxidation.

A long standing question has been to determine if there is there a relationship between surface finish, contact stress, and the life expectancy of load bearing surfaces. Statistical parameters are used to describe real surfaces and go by such exotic names as [40]:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Roughness</td>
<td>Peak Count</td>
</tr>
<tr>
<td>Core Roughness Depth</td>
<td>Profile Depth</td>
</tr>
<tr>
<td>Datum Center Line</td>
<td>Profile Length Ratio</td>
</tr>
<tr>
<td>Datum Mean Line</td>
<td>Roughness Depth</td>
</tr>
<tr>
<td>Leveling Depth</td>
<td>Skew Amplitude Density</td>
</tr>
<tr>
<td>Material Content Curve</td>
<td>Waviness Depth</td>
</tr>
<tr>
<td>Mean Roughness</td>
<td></td>
</tr>
</tbody>
</table>

There are over 50 accepted statistical parameters [41] by which to characterize surfaces. Common engineering practice and international agreements have resulted in the specification of surface finish in terms of height profile only. Surface roughness is usually measured in microinches in the United States. The most common height profile measurement is the RMS value more commonly referred to as the centerline roughness average, specified $R_a$. Recent (1974), dissatisfaction with the one-parameter model has led to a two-parameter model in which the average wavelength of surface features is combined with $R_a$. The most recent acceptable models use two variables. Consult Johnson [1] for a more detailed discussion.

The average roughness is an excellent starting point since it provides average height readings for peaks and valleys. It is not totally adequate however, because it provides no information as to the shape of the surface. An additional parameter, called the bearing-length ratio, is the comparison of the horizontal surface area (an indication of numerous flat peaks) to the overall surface within a given evaluation length. A part could have a good $R_a$ value and still be a poor load bearing surface due to the presence of pointed sharp peaks.
that can break through fluid lubricant layers. Table 7.1 shows the statistical parameters that are most prevalent from the manufacturing point of view. In production engineering, appropriate and meaningful methods for the definition and specification of surfaces has long been a subject of debate.

7.1.2. Measurement Techniques

The most common technique for measuring surface topography has been the stylus profilometer. Its use has been enhanced in recent years by the computer. More recent techniques include the use of lasers to non-destructively examine and create digital maps of a surface. Three-dimensional topographical maps, accurate to microns, can be created that are similar to the one shown in Figure 7.3. These measurement machines are expensive, but have the advantage of being able to resolve smaller features than possible by the mechanical contact of a stylus of fixed radius.

Automatic data acquisition makes it possible to obtain large and accurate quantities of surface height data. Computers are needed to take and analyze the data since manual collection and analysis would be too time consuming and error prone to be reliable due to the sheer volume of data. Statistical techniques are used to calculate or extract any of the surface parameters from the sample profilometer records. The effect of variation in these statistical surface parameters should be correlated to theoretical contact stress calculations. A set of dimensional analysis parameters $\pi_1$, $\pi_2$, $\pi_3$, etc. could be developed and used in conjunction with contact stress experiments designed to detect any such correlation. Bearing manufacturing companies may have done this and are keeping the knowledge proprietary.

The best way to determine which of the many parameters are most useful for a given situation would be to determine the correlation between each of them and the quantity in question. This quantity might be the maximum normal stress, maximum shearing stress, the failure rate of a product in service, or a quality control inspection part rejection rate.
Extracting all of the accepted surface parameters from the data and searching for correlations to the desired variables would take an extensive amount of computer time. Some economy of computation could be expected since many of the statistical parameters have common mathematical terms such as the summation of $x_i^2$, etc. This process might even be implemented as a parallel computation on dedicated hardware. As tolerances on products become tighter and we strive for zero defects, perhaps we should be measuring smoothness instead of roughness.

Attempts have been made to characterize surfaces through their profiles using the methods of random signal analysis. It is appropriate to regard the surface as a random profile that can be described in terms of its height and associated autocorrelation functions. The results of such characterizations depend critically on the sampling interval used in measuring the profile. The size of a surface feature is a matter of perspective. A sampling interval too small could miss important features, as would one that was too large. An upper limit for the sample length is imposed by the size of the specimen and a lower limit by the radius of the profilometer, although the lower bound can be reduced by using optical methods where the wavelength of light is the limiting factor. All of the following physical features can be described by random profiles if viewed from the proper perspective: a micro-ripple on the surface, a scratch visible to the naked eye, a roadside ditch, hills, mountains, etc. So what is the proper physical scale to sample on? A practical solution would be to pick wavelengths that can capture natural surfaces and man-made tool marks. In modeling a rough surface, the wavelengths of surface features will all be smaller than the characteristic length of the geometry, i.e. the principal radii. Even the rough jagged edges produced by a tool tend to be periodic, so they too could be represented by some type of Fourier waveforms. A double Fourier series could be used to curve fit profilometer data and model surfaces similar to the expressions used to describe plate deflections.

### 7.1.3. Mathematical Surface Models

Conventional contact theory is derived on the basis of smooth profiles in continuous contact. Since real surfaces do not fully comply with the theoretical assumptions, it is important to consider the effect of the discontinuous contact caused by surface roughness. The first models of rough surfaces were the so called "Regular Wavy Surfaces", which assume the surface to have a sinusoidal height distribution. More recent attempts have been made to model surfaces as an array of hemispherical domes called "asperities". In the simplest models, the asperities are assumed to be the same size, with a Gaussian distribution of heights from the tangent plane. The tangent plane can be defined as the tip of the outermost asperity when examining individual contacts or the mean height of all asperities when concerned with the overall behavior of the body. When two surfaces are brought together, only a few of the asperities come into contact. Thus the applied load is concentrated, causing excessively high local stresses. These asperities deform, bringing into contact an increasing number of them until the load can be supported without further deformation. The important questions to be answered to fully understand the situation are:
1) To what extent is the deformation elastic or plastic?
2) By what extent do the individual asperities grow in size rather than increase in number?

The heights of the asperities above the mean datum line have a statistical distribution $\phi(z)$. The three most common parameters for specifying the individual asperities are: the standard deviation of asperity heights, radius of curvature of the asperities, and the number of asperities per unit area. The load is divided by simultaneously creating new contacts and enlarging existing ones. However the area of contact is proportional to the load, regardless of the mechanism of deformation.

Singh, Paul, and Woodward [42] have applied the Hertzian results for spherical geometries to contacts between the individual asperities. In calculations involving simulated asperity contacts, each asperity is assumed to deform independently according to the Hertz theory. A Hertzian result for a single asperity contact is that plastic deformation will occur at a pressure of approximately $3S_Y$, where $S_Y$ is the yield point of the material. Strain hardening of the asperities will decrease this factor of three. This "flow pressure" is commonly assumed to correspond with the hardness of the material. At low loads, pressures are much lower than Hertzian and spread out over a much larger area. At high loads, the calculated and Hertz pressure distributions agree well. Similar calculations have been performed by Greenwood and Tripp [43] who have worked out an elaborate iterative technique. They find that the effect of surface roughness is to reduce the effective pressure at the center of contact and spread it out over a wider region, beyond the Hertz circle.

All surface theories make specific assumptions about the geometry. Wear of the surface over time modifies it, making the stress and deflection predictions based on the initial stable geometry less accurate over time. To account for the changing surface over time, some type of adaptive iterative model would have to be developed. Work has been done in this area by Knothe and Le-The [44] to predict the life of railroad wheels. The deterioration of the surface, thus the assumed geometry, would have to be evaluated as frequently as it was degraded. This requires a model of wear rates. Such a complex model would need to be compared to a physical model to validate the predictions. The theory should then be modified accordingly. The computer simulation could thus be based upon profiles taken from surfaces used in experiments. Dowson [45] states that unwarranted idealization of the surface topography or lubricant rheology can lead to erroneous predictions.
7.1.4. True Contact Area

A major emphasis in the literature is on the true area of contact. The distinction is made between the real and apparent contact areas. This is especially important in calculations involving electrical and thermal transfer through the contact region. A problem with determining the contact area by electrical resistance through the interface is that the resistance is sensitive to the amount of surface oxidation.

A more difficult situation presents itself when the surface irregularities are such that contact does not occur over one continuous area, but over several high spots. The resulting load distribution may be obtained by superposition of the appropriate solutions for each contact interval. Poritsky [46] in 1950 anticipated problems with this type of model:

"The load distribution which assumes contact also produces displacements over other neighboring potential contacts, affecting interference there. Therefore superposition is not necessarily a simple process but involves successive corrections both in the division of the total load between the contact intervals, the width of these intervals, and in the amount of interference and the load distribution over each interval."

An algorithm to perform the superposition calculation would have to be developed to handle the different cases of contacting geometries. This superposition could even be done down to the level of the interaction between individual asperities.

7.2. The Real Structure of the Surface Layers of Metallic Bodies

The true nature of materials deviates from the continuum model on which contact theory is based. Figure 7.4 shows a schematic representation of the structure of a typical metallic surface. Ferrous materials do not have a pure metallic surface, but are covered by a thin "absorption layer". This forms, together with the oxide or reaction layer, the "outer boundary layer". Layer thicknesses are based on rolling friction investigations on plain carbon steels [47]. The apparatus used in these experiments consists of disks rotating at different, but constant speeds, to induce a known slip rate. Similar layered structures exist for other materials, although the specific details will vary.

There is a relationship between crystal orientation and frictional behavior. Alison and Wilman [48] have determined that the coefficient of friction of metals with a face-centered cubic lattice structure is greater than those that are hexagonally close-packed. These tests and similar works done by others are carried out in high vacuum conditions to eliminate the effects of oxidation. Real crystal lattices can deviate considerably from the regular order of a perfect laboratory specimen. Similar differences would be expected in contact between polymers when sliding parallel or perpendicular to the direction of the molecular chains.
Mechanical stress creates additional property changes in the surface layers, particularly in the inner boundary layer where the chemical and physical properties deviate considerably from that of the base material. When metals are plastically deformed, almost all of the physical and chemical properties are altered. Plastic deformation can cause a strong chemical activation which produces a spontaneous and visible oxide growth sometimes referred to as *frictional oxidation*. These changes greatly affect the friction and wear behaviors. A refined model of contact between bodies should take into account this layered structure, especially since contact stresses are a localized surface phenomena.

### 7.3. Friction Between Real Surfaces

The goal of this section is to examine the deficiencies of simplified theories of friction. Modeling of tangential contact loading theories requires an understanding of the causes of frictional forces. The microscopic interaction between real materials in contact is quite complex and does not readily lend itself to simple calculations.

#### 7.3.1. The Classic Laws of Friction

The classic laws of friction which are attributable to Da Vinci, Amontons, and Coulomb, are summarized as follows:

1) The friction force is generally assumed to be proportional to the normal force.

\[ F = \mu N \]
in which $\mu$ is the coefficient of friction. The symbols $P$ and $Q$ are frequently used in contact stress literature for the normal and tangential loads, respectively.

2) The coefficient of friction is a constant whose magnitude is determined only by the materials in contact.

3) The coefficient of friction is supposedly independent of the contact area, normal force, and the relative velocity or slip.

4) The static coefficient of friction is greater than the kinetic coefficient.

Many authors [49] have been independently able to prove that these postulates are sometimes invalid, but they continue to be perpetuated and used for designing. Another quantity not generally taken into account is the length of time over which the stress is applied. The correctness of these factors must be considered if an accurate computational model of tangential loading is to be developed.

7.3.2. Theories of Friction

Figure 7.5 Factors comprising a Friction System.

In order to understand the physical situation which has been determined by experimental results and microscopic investigations, the contact should be thought of as a "friction system" comprised of many factors. Krause and Demirci [47] have divided these factors into five main groups according to their effect (Figure 7.5).

The introductory section in Moore [49] contains an excellent description of the molecular and micro-mechanical bases of friction. It includes some basic equations that
could be used to develop a computational model of surface friction, taking into account such factors as adhesion, ploughing, shearing and electrostatic forces. Figures 7.6 and 7.7 illustrate the level of detail that must be considered for such a computational model of friction.

Figure 7.6 Elastomeric friction

Figure 7.7 Metallic friction

### 7.3.3. Rough Surfaces and Friction

The actual contact area between two bodies is reduced considerably as the roughness is increased, since initial contact occurs at the tips of the asperities. The greater the surface roughness, the greater are the specific surface pressures and frictional forces. The high specific surface pressure (contact stress) causes plastic deformation of the asperities, which in turn leads to a high intensity of frictional oxidation. Tests [47] have shown that the effect of the formation of oxide surface films is an initial reduction of roughness. Temperature and humidity play a role in this process, but the water molecules themselves are not the direct cause of the reduction in the coefficient of friction, since some steels do not form such layers. The change in coefficient of friction can be explained by the dependence of the chemical process on air humidity. Because of the greater friction force, rough rolling surfaces wear faster and to a higher degree than smoother ones. Independent of different initial roughness characteristics, the surface roughnesses of the two bodies will become equal after a certain rolling length (if they are the same material and

Figure 7.8 Coefficient of friction as a function of rolling length.
of equal hardness). After the film has been destroyed, the coefficient of friction increases rapidly, to a maximum value, then tends towards a quasi-constant value as illustrated in Figure 7.8. This could be described as a polishing type of wearing in.

7.3.4. Design Tips for Friction Calculations

Due to the diversity of factors in the contact between two bodies, the friction force cannot be determined by one all-encompassing law. The lack of a comprehensive theory does not help the average designer when facing a contact stress problem that involves friction. For the practicing engineer, the following design tips are suggested for calculations involving friction:

- If the application is critical, do an experiment that simulates the operating conditions as closely as possible.

- Determine an appropriate margin of uncertainty in the calculated result. Note that the traditional term "factor of safety" is not used. The consequences of too large or small a value being assigned to the coefficient of friction does not necessarily mean immediate failure (that depends on the context), although failure due to excessive stress limits could be the eventual result. The design consequences might be as simple as specifying a larger motor or brake size than would otherwise be required due to the drag forces predicted by the assumed value for the coefficient of friction.

- Try to anticipate the conditions that will exist after the device has been in service for awhile or the variety of environmental conditions to which it will be exposed, i.e. try simulating throwing a little sand into the gears.

- Look up published values of the coefficient of friction for the pair of surfaces under consideration, pick the worst case conditions, then assume the value to be constant at the maximum or minimum value (depending on the application).

- Recognize that a non-constant μ could be the source of the difference between calculations and experiments results.
7.4. Lubrication and Pitting Failure

The introduction of lubrication greatly alters the patterns of stress distribution in contact problems and has more effect on tangential rather than normal loading. Lubrication can reduce contact stresses by as much as 20% below values predicted by the Hertzian equations. Stresses are lower because a hydrodynamic film develops that separates the bodies in contact and spreads the load over a larger area.

Lubrication can affect pitting failures. If a small crack is present on the surface, lubricant can be forced into it. As the two bodies move past each other through rolling or sliding, contact stresses raise the pressure of the oil in the crack equal to the level of compressive stress at the surface, causing high tensile stresses at the crack root. This stress causes the crack to grow, which eventually leads to pitting. Pitting can also occur by repeated loading at stresses above the material's endurance limit. The thickness or depth of the pits has been shown to correspond to the depth at which the maximum shear stress occurs.

Components in rolling contact have a statistically predictable life. The appearance of a small pit on the surface as shown in Figure 7.9 usually signals the end of a part's useful life, since continued operation results in rapid disintegration and eventual failure. Symptoms of pitting are excessive vibration and noise. Lubrication can carry hard metal particles from the pits to other locations and thus cause further damage through abrasion.

Figure 7.9  Example of pitting failure.

The possible methods of reducing the likelihood of pitting are either to use no lubrication and highly polished surfaces, or to increase the viscosity of the lubricant. Using no lubrication at all is impractical and polishing surfaces can be expensive. Changing the viscosity can be effective, but knowledge of other factors in lubrication design can be helpful to avoid problems that could arise such as excessive heat generation and increased
frictional drag. Scoring, which occurs at high speeds, is a sign that the lubricant quality and quantity are insufficient for the load.

Lubrication is often classified into three regimes based upon the degree to which the lubricant separates the contacting surfaces. In boundary lubrication, there is continuous contact over most of the area. Mixed-film lubrication has intermittent local contact of high points through the lubricant. In the case of hydrodynamic lubrication, the surfaces are completely separated by a thin fluid film.

There is an entire field of study called elasto-hydrodynamics. It is primarily concerned with supporting the load on a thin film of fluid and also takes into account the elastic deformations of the two bodies. Due to the complexity and coupling of the equations, elasto-hydrodynamics relies heavily on computational methods [50]. The initial pressure distribution in the fluid, not surprisingly, is usually assumed to be of a Hertzian nature.

7.5. Impact on a worn surface.

The mechanical response of impulsively-loaded contact is an important consideration in machine design. A large number of repeated load cycles must be sustained with a minimum degradation of the contact surface. In order to determine rational criteria for material selection, surface preparation, and lubrication, a quantitative theory for impact wear has been proposed by Engle and Bayer [51]. This reference, which is discussed in the following paragraphs, studies the effects of a single impact and its effects over repeated applications.

The local failure of materials is frequently associated with a key parameter of microscopic origin such as the maximum subsurface shearing stress. By wearing, the contact attempts to redistribute the load to reduce these stresses, so it is important to be able to estimate the stresses at any stage of wear. For any given configuration, a stress-related failure parameter should be selected. Simple mechanical models are used to estimate the wear path for rolling contact by using parabolic wear craters.

Axisymmetric impact of a ball indenting an elastic half space is the simplest analytical model of an impact problem. A numerical method of "point matching" is used to evaluate displacement and force-modifying functions when incrementally exceeding the elastic limit as predicted by the Hertzian solutions. The contact area is divided into concentric annuli of equal thickness. The total elastic displacement of the surface is expressed as a linear combination of the pressure from each ring. The displacements of the two bodies are made compatible at the boundary circles, which yields a linear system of algebraic equations. This formulation makes it possible to solve for the pressures. A similar technique could be applied to elliptical contacts as well. At each concentric ellipse, the elliptic integral solutions implemented in this investigation would have to be evaluated.
Smoothing must be done between the annular regions to obtain intermediate displacements.

The maximum radial stress occurs at the edge of contact and plays a key role in crater formation. As the penetration becomes deeper, the formerly tensile regions migrate inside the contact area where the loads are compressive. This sets up a reversal of the load which leads to fatigue failure. From physical intuition, the shape of the wear path for repeated loading will correspond to the fastest relief of the critical failure parameter. This can be found from a steepest descent search for the load in the stress-geometry space.

A more accurate technique could be developed, possibly involving less computation. It would have varying annular widths, which depended on the amount of wear. This work could be extended to determine the effect of shock loads on power transmission systems with shafts and bearings. The study of high-speed impacts however, such as the process of shot peening or the penetration of projectiles, is a field of study unto itself. In impact analysis, it is necessary to account for the kinetic energy of the bodies, wave propagation, and thermodynamic effects. Such elaborate analyses rely heavily on numerical methods.

7.6. Rolling Contact and Shakedown

Most practical applications of rolling contact involve the repeated application of the load. The deformation cycle in rolling contact involves rotation of the principal axes of strain, with very little change in the total strain energy. If in some pass the elastic limit is exceeded somewhere within the stress field, plastic deformation will take place and thus introduce residual stresses. In subsequent passes, the material is subjected to the combined effects of the load and the residual stresses. Generally, residual stresses are protective in that they make further yielding less likely.

It is possible that the residual stresses will build-up until a state is reached such that in all later passes, the deformation is entirely elastic. This is the process of shakedown, where under repeated loading, plastic deformation introduces residual stresses that make the steady-cyclic state purely elastic. Symonds [52] has shown that if any time-independent distribution of residual stresses can be found, which together with the elastic stresses due to the load, constitutes a system of stresses within the elastic limit, then the system will exhibit shakedown. If no such distribution can be found, then plastic deformation will occur in all subsequent passages. Johnson [1] shows that for a cylinder freely rolling on an elastic half-space, the load must be increased by more than 66% above first yield to produce continuous deformation with repeated rolling cycles.

Three-dimensional rolling bodies are more complicated, since all components of residual stress can arise. The stabilization of groove dimensions does not guarantee a true shakedown state since plastic shear parallel to the surface may still be taking place. Strain hardening can also lead to apparent shakedown.
7.7. Allowable Stresses

In machine design, the emphasis is often on the fatigue life in contact rather than a one-time yielding. Fatigue life data is difficult to establish for contact problems for a variety of reasons. Namely, because the state of stress is three-dimensional, the surface in contact is modified over its life, and the exact design conditions are difficult to duplicate. Due to the many unknowns, components are often designed with an expected finite lifetime often expressed in hours of operation rather than for infinite life. In rolling contact applications, there is a load reversal (fluctuation) as the roller or wheel passes over a fixed point. Another negative factor is the cumulative damage from occasional shock loading. As a starting point, it is highly recommended not to allow the maximum shear stress reversal to go above the material endurance limit, if one exists. The actual stress limits imposed on a contact stress design problem depend on many factors including the consequences of failure, so calculations should be accompanied by testing.

7.8. Heat Treating and Contact Stresses

Since contact stresses are a localized phenomenon, ways to combat their effects can also be of a localized nature. For design problems involving steels, this typically involves heat treating. Heat treating, which does not affect the elastic modulus, raises the material yield point. If the yield point is raised above the stress level induced by contact, plastic deformation can be avoided in the contact region. Heat treating is often done to a limited depth in the material on certain critical surfaces, or around the entire body through case-hardening, as is often done to gears.

Heat treating also has its disadvantages. Making the material too hard can make it brittle and susceptible to cracks. The loss of ductility also reduces the ability to withstand shock loads. Core crushing is a surface failure phenomenon that occurs in case-hardened materials. It can result when the case is too thin to support the compressive loads or the subsurface shearing stresses exceed the strength of the material below the case. Failure occurs from the core collapsing and being pressed into the material below it, or by the outer case cracking and breaking away.

It is recommended that the depth of the heat treat correspond to the location of the maximum subsurface shearing stress. The hardness should be increased to a level that corresponds to a yield point greater than the predicted shearing stresses. An approximate relation between hardness and tensile strength is given by:
in which \( Bn \) is the Brinell hardness and \( S_y \) is the tensile strength in lbs/in\(^2\). This relation does not apply to non-ferrous metals, with the possible exception of certain aluminum alloys [53]. Hardness conversion information such as that contained in Table 7.2 is provided by steel manufacturers in their material selection catalogs. In this table, Brinell hardness values above 500 are for a tungsten carbide ball, and those below 500 for a standard ball. Similar hardness conversion data in a graphical format is reported by Juvinal [54] and is shown in Figure 7.10. Using a hardness conversion table or an approximate relation, contact stress calculations can be related back to the yield strength determined in the simple tensile test. There are numerous methods of heat treating. The following two are commonly used in relation to contact problems.

### 7.8.1. Nitriding

Nitriding is a process of case hardening in which an iron-base alloy of special composition is created in an atmosphere of ammonia or in contact with some nitrogenous material. Surface hardening is produced by the absorption of nitrogen without quenching. Finished machined surfaces hardened by nitriding are subject to minimum distortion due to uniform heating in the bath.

### 7.8.2. Flame hardening

Flame hardening is a process in which the surface layer of an iron-base alloy is heated above the phase transformation temperature range by a high temperature flame. This heating is then followed by quenching. Care must be taken to prevent cracking. This process is used for selective hardening of large steel castings. Hardness ranges of 400 to 700 Brinell can be obtained. An oxy-acetylene torch is used for the heat and pressurized air is utilized for quenching. Tempilsticks, which are temperature sensitive crayons, are used to indicate the temperature. Two temperature ranges (the high and low of the recommended range) are wiped on the surface next to each other to indicate that the proper temperature has been achieved. They are inexpensive and available from any welding supply store.
<table>
<thead>
<tr>
<th>Brinell</th>
<th>Rockwell C</th>
<th>Tensile Strength x1000 psi</th>
</tr>
</thead>
<tbody>
<tr>
<td>745</td>
<td>65.3</td>
<td>...</td>
</tr>
<tr>
<td>712</td>
<td>63.4</td>
<td>...</td>
</tr>
<tr>
<td>682</td>
<td>61.7</td>
<td>...</td>
</tr>
<tr>
<td>653</td>
<td>60.0</td>
<td>...</td>
</tr>
<tr>
<td>627</td>
<td>58.7</td>
<td>323</td>
</tr>
<tr>
<td>601</td>
<td>57.3</td>
<td>309</td>
</tr>
<tr>
<td>578</td>
<td>56.0</td>
<td>297</td>
</tr>
<tr>
<td>555</td>
<td>54.7</td>
<td>285</td>
</tr>
<tr>
<td>534</td>
<td>53.5</td>
<td>274</td>
</tr>
<tr>
<td>514</td>
<td>52.1</td>
<td>263</td>
</tr>
<tr>
<td>495</td>
<td>51.0</td>
<td>253</td>
</tr>
<tr>
<td>477</td>
<td>49.6</td>
<td>243</td>
</tr>
<tr>
<td>461</td>
<td>48.5</td>
<td>235</td>
</tr>
<tr>
<td>444</td>
<td>47.1</td>
<td>225</td>
</tr>
<tr>
<td>429</td>
<td>45.7</td>
<td>217</td>
</tr>
<tr>
<td>415</td>
<td>44.5</td>
<td>210</td>
</tr>
<tr>
<td>401</td>
<td>43.1</td>
<td>202</td>
</tr>
<tr>
<td>388</td>
<td>41.8</td>
<td>195</td>
</tr>
<tr>
<td>375</td>
<td>40.4</td>
<td>188</td>
</tr>
<tr>
<td>363</td>
<td>39.1</td>
<td>182</td>
</tr>
<tr>
<td>352</td>
<td>37.9</td>
<td>176</td>
</tr>
<tr>
<td>341</td>
<td>36.6</td>
<td>170</td>
</tr>
<tr>
<td>331</td>
<td>35.5</td>
<td>166</td>
</tr>
<tr>
<td>321</td>
<td>34.3</td>
<td>160</td>
</tr>
<tr>
<td>311</td>
<td>33.1</td>
<td>155</td>
</tr>
<tr>
<td>302</td>
<td>32.1</td>
<td>150</td>
</tr>
<tr>
<td>293</td>
<td>30.9</td>
<td>145</td>
</tr>
<tr>
<td>285</td>
<td>29.9</td>
<td>141</td>
</tr>
<tr>
<td>277</td>
<td>28.8</td>
<td>137</td>
</tr>
<tr>
<td>269</td>
<td>27.6</td>
<td>133</td>
</tr>
<tr>
<td>262</td>
<td>26.6</td>
<td>129</td>
</tr>
<tr>
<td>255</td>
<td>25.4</td>
<td>126</td>
</tr>
<tr>
<td>248</td>
<td>24.2</td>
<td>122</td>
</tr>
<tr>
<td>241</td>
<td>22.8</td>
<td>118</td>
</tr>
<tr>
<td>235</td>
<td>21.7</td>
<td>115</td>
</tr>
<tr>
<td>229</td>
<td>20.5</td>
<td>111</td>
</tr>
</tbody>
</table>

Table 7.2 Hardness Conversion table.

Figure 7.10 Hardness relation to tensile strength.
7.9. Gears

The design and analysis of gears is an endeavor which utilizes the many methods of contact stress analysis. Figure 7.11 shows the stress patterns that develop as two spur gear teeth come into contact. Two areas of stress concentration are indicated by the photoelastic contours: at the root of the tooth and where teeth contact. Each location must be treated by different methods of analysis. Rolling occurs at and near the pitch line, while at other locations within the contact zone sliding takes place. In some cases, the slip velocity may be as high as 30% of the rotational velocity at the pitch radius. Generally, it is assumed that oil is squeezed out of the Hertz area, resulting in metal-to-metal contact, allowing the coefficient of friction to become fairly high. An accurate numerical method would have to account for the fact that the coefficient of friction is not constant.

Simplified analysis of stresses at the root is typically performed by assuming that the tooth acts as a cantilever beam. For the contact stress analysis, the principal radii equals the curvature of the involute profile of each tooth at the point of contact. The secondary radii for both bodies are infinite, i.e. the first approximation would assume line contact neglecting edge effects. As the gears rotate, the line of contact moves. Each rotational position of the gear set must be treated as a separate problem because the curvature of each tooth and load may vary. The equations for two-dimesnional contact with tangential loading developed by Liu [6] can be used for this calculation. Many possible tangential
Pressure distributions are possible for rolling contact if creep is allowed. The total drag force is obtained by integrating the tangential load distribution over the contact area:

$$F_d = \int \int Q(x,y) \, dx \, dy$$  \hspace{1cm} (7.2)

Bevel gears contact over a much larger area than spur gears and in a more complex manner that occurs in multiple planes. Numerical methods are needed to fully model this type of contact.

7.10. Camrollers

Figure 7.12 illustrates the internal construction of a camroller, a common device used to support rolling loads. There are two contact problems here. The design of the internal construction in terms of the size and number of needle bearings is a cylindrical contact problem. The outside is a contact problem from an application point of view. Some camrollers come with crowned radii, meaning there is a secondary curvature giving them a truncated ellipsoidal shape. While others, such as the one shown in Figure 7.12, are pure cylinders. Crowned rollers have the advantage of maintaining better self-alignment than do flat ones, which could cause high stresses if tipped up on edge. Crowned rollers have a smaller contact area than their flat-edged counterparts, which leads to slightly higher contact stresses. A pure cylinder would suffer from the rigid punch effect, so even the cylindrical rollers have a small radius at the edges of the cylinder. Stresses in materials in contact with cylindrical cam rollers under normal loading can be computed using the simple Hertzian equations. Crowned cam rollers require the use of the elliptic integral solutions.

Figure 7.12 Camroller
7.11. Wheel and Rail

The contact between a wheel and rail is a classic contact problem. The wheel, shown in Figure 7.13, has both a positive and negative radius. The rail head has one positive radius and the one in the plane of the rail centerline is infinite. Tangential loads from braking or acceleration cause the stress distributions to become complex. The stress state associated with a stationary wheel can be evaluated by the elliptical solution implemented in this investigation. This is in fact a good starting point in the design process of determining the required diameter of the wheel for the given wheel load and material. The dimensions of crane rails have been standardized, but there is sometimes leeway in the design process as to which size rail to choose. On curved sections of track or when a wheel is misaligned, the flanges can come into contact with the side of the rail head. This situation results in multi-point contact with tangential loading and can only be evaluated by numerical methods. The running surfaces of the wheels are usually hardened.

Figure 7.13 Crane wheel and rail
8. Summary

8.1. On the use of the Continuum Model to Describe Real Materials

In this investigation, some of the differences between real surfaces and their properties, and the idealized model of an elastic continuum were examined. Any theory of continuous, isotropic physical media eventually reaches a limit when trying to describe events that are on the same order of magnitude as the discrete molecular structure. Hence there is an ultimate limit of scale to which any model can be extrapolated. The continuum model is very useful, however, and can predict many physical phenomena.

Theoretical stress calculations are important in the design phase and in determining the cause of failure after it has occurred. In actual materials, discontinuities, particularly when they occur at points of high stress, may have a large influence on failure. The equations for stresses and deformation do not even begin to address slip dislocations between crystal planes, discontinuities from the manufacturing processes, surface finish, wear rates, work hardening, etc. Due to the number of variables involved, using more than three significant digits for practical design calculations involving contact stresses, except for specified physical dimensions, is questionable. For real problems, experiments should be done in conjunction with theoretical analyses.

All the mathematical models are simulations of physical problems. When analyzing a problem, it must be determined how closely the problem resembles one of the known solutions. It is important to know the assumptions and restrictions used in the derivation of each solution to prevent it from being applied or extrapolated where it is not valid. The field of contact stress analysis has advanced far beyond the original Hertz solutions, though have not invalidated those results. It is clear that there is no single all-encompassing solution for the contact between two bodies. The finite element method could be developed into such a tool with enough time. This will await the development of widely available one billion instruction per second computers having a parallel array processing ability which will exist in the next ten years.

8.2. On the Elliptical Solution Implemented in this Investigation

The process of implementation revealed a number of difficulties not discussed in the literature containing the derivations. Among these are the geometric validity checks that were discussed in the section on the geometry of contact. Other problems are singularities for certain geometric configurations.

This investigation has shown that direct evaluation of the integral formulation is a feasible and accurate method of determining contact stresses. The groundwork has been established for the three-dimensional case with friction. The computer programs developed
in this investigation could be used to map out the solution space for a range of specified design parameters (physical dimensions and choice of materials) in an exhaustive search. Third party libraries of 3D plotting routines are available that could be used in conjunction with the stress evaluation routines to make interpretation of the results more tractable. A fast computer is also a prerequisite.

The method involving direct evaluation of the integrals could be described as being closer to an analytical rather than numerical method. According to some mathematicians, numerical calculations are allowed in the proofs of theorems. There is a theorem in a branch of mathematics called group theory that is titled The Enormous Theorem [55]. It concerns whether or not the minimum number of simple groups has been proven to exist and is ten thousand pages long. Part of the logic chain requires the use of a computed result. Who checked this thing anyway? Purist mathematicians have a hard time accepting the use of a computer to prove a theorem, but engineers, who rely on the results of numerical calculations everyday, are probably not bothered by this. So you see, the distinction between numerical and analytical methods, is not so clear as you may have thought.

8.3. Tangential Loading for Elliptical Contact

Limited forms of solutions are available for tangential loading contact problems. The solution that was introduced in Section 6.8 has similarities to the normal loading elliptical problem in that it is formulated in terms of elliptic integrals. This and other analytical solutions for torsional loading at the surface could be implemented on computer as was done for the normal loading problem in this investigation. Some of the computational methods developed for normal loading could be applied to the other solutions as well. As is the case with the normal loading elliptical solution, the tangential loading solutions do not provide an all-encompassing answer to the contact stress problem. If implemented on a computer however, they would greatly extend the abilities of designers working on a wide range of practical problems. This would be the long sought practical solution for the sliding friction problem.

8.4. Direction for Future Work (or the Science Fiction Part)

The expressions for the stresses anywhere within the solid due to combined normal and tangential loading can be derived. Different expressions would be needed for determining stresses in various locations to side-step singularities that arise in analytical solutions. The stresses below the surface will likely depend on the incomplete elliptic integrals as they do in the normal loading case. The contact problem can be defined by a few simple parameters or variables versus the amount of information required to set up a finite element mesh.
Superposition applies, since the stresses are assumed to be linearly elastic. For each point in the material $p(x,y,z)$, an array is maintained for the elements of the stress tensor (actually only the six independent elements need to stored). The results for stresses at any point from different analytical solutions for normal, tangential, and torsional loading would be combined by matrix addition. Since contact stresses and their effects are localized, a program could determine how far away from the origin (perhaps in terms of the mapped depth angle) each of the solutions diminish to a level that does not make significant contributions to the stress tensor. The stress components would be computed only for points within this range. This would save a lot of computation, since checking a distance is much quicker than evaluating the elliptic integrals at every point. This could be called "intelligent point evaluation". Portions of these calculations might be implemented as a parallel process.

Determining the stresses or displacements at any single point within the solid requires a fairly significant amount of computation. The calculation for the stresses at a single point would then be embedded in a larger one to determine the magnitude and location of the local maxima of certain elements of the stress tensor. The elements of the tensor that are of interest, depends on the overall purpose of the calculation and the failure theory being used. A steepest descent type search could be used to locate the maxima. The search can be reduced somewhat by advance knowledge of where these maxima are likely to occur, i.e. behind the trailing edge of contact, or on the $z$ axis approximately $.3b$ below the surface, etc.

The whole point of this boils down to simulating the continuum model on the computer. A more accurate computational model of friction will be needed to model tangential loading. Also accounted for should be: surface roughness, wear rates, running-in of the surface, changes due to work hardening, and change of the coefficient of friction over time. Non-linear capabilities through successive linear approximation could be added to handle plasticity and contact between elastomers which can involve large strains. These capabilities should be the ultimate goal of a general purpose contact simulation program. Such a program could be used to verify the results of a finite element analysis or even break new ground. It is important to not rely too heavily on one method or solution strategy. Eventually, when computers become more powerful through sheer speed and parallel and multi-processing, it will be economically feasible to create models on the scale of the molecular level where the electro-magnetic interactions between individual atoms are accounted for. This will allow studies to be performed on crack propagation and crystallographic properties and imperfections. Such models would be useful in the design of solid-state devices and micro-miniature machines. It is not inconceivable that one day there will be a course in quantum mechanical FEA.
9. A Personal Note

The subject of contact stresses was never even mentioned in my undergraduate classes in strength of materials or machine design. I became interested in them at my first full time job. A machine designed by this company consisted of a three-stage translating cantilevered table called a shuttle. The load was rolled out to its extended position and supported on a set of cam rollers which traveled in an accurately machined groove in the sliding rails. The high contact stresses on these rollers would cause indentations at the full stroke position which allowed the tip of the device to deflect unacceptably. This affected the positioning accuracy and repeatability of the entire machine.

One version that I designed for a special customer order required an even longer cantilever extension, yet smaller cross-section, than the traditional design. The special slider rails were long and slender cold-rolled steel bars which frequently contain residual stresses from the cold working process. Forty-five assemblies were built for this project, each having eight rails for a total of 360 pieces. This is in contrast to a normal project having two to six assemblies with two slider rails each, totaling up to twenty four pieces. There were a lot of them.
The additional overhang of the load in this design caused the contact stresses to greatly exceed the yield point of the material normally used. A computer program had been written by the engineering staff to calculate contact stresses based on some handbook formulas from Roark's *Formulas for Stress and Strain*. This program was used to calculate the stresses using parameters for the custom design and it reported stresses that were very high. A material with a sufficient yield strength did not exist in its raw form to withstand the predicted stresses, which meant that we would have to heat treat the rails, or change the design parameters, or both.

Heat treating is a time consuming and expensive process. Long slender parts with residual stresses are notorious for warping when heat treated. The straightness of the slider rails was important to the positioning accuracy of the machine. Needless to say we had some warping problems. The company that performed the heat treating knew all the prior manufacturing stages and the raw material that was to be used before they quoted the job, so technically they were responsible for not holding the straightness tolerance specified on the drawings. They wouldn't accept full responsibility, so we ended up straightening them ourselves at the cost of extra man-hours. Due to schedule pressure, the parts were shipped in a lower quality condition than they should have been, which is a less than desirable outcome.

It was later revealed that the computer program had a known bug which no one had ever been told about, causing the program to give incorrect results. The calculated contact stresses were too high, misleading us into specifying more heat treatment than was necessary. They were not unreasonable however, so they went unchallenged. Being on the job for a just few months and under the relentless time pressure of the delivery schedule, you have no time to investigate such esoteric things as the correctness of a contact stress calculation program. The program was an established and accepted part of the design calculations. Heat treating was typically done to the standard designs as well, but they were larger cross-section pieces which were less subject to warping.

I went to the library to look up references on contact stresses. There I found *Mechanics of Materials* by Seeley and Smith which had the section on contact stresses which is the basis of this investigation. After introducing the formulation in terms of elliptic integrals, the book presents some simplified formulas and graphs of curves from which coefficients can be looked up to calculate stresses. This is not very accurate, due to the interpolation required and is also a time consuming series of calculations. I felt a better program should be written that incorporated the integral solution procedures. It should also be written as an interactive design tool to be used early in the design process. The
machines were built, but we went through much more expense and grief than necessary because of the lack of an available and accurate computational tool.

Another area where I have run into contact stress-limited design problems is the interaction between a crane wheel and rail. Although there is much more going on here than simple Hertzian contact, particularly from frictional effects and side loading, the basic equations can be used to size the components being designed. Taking into account the sliding friction will tell you something about the longevity of the part and the type of surface treatment that is required.

More recently I encountered contact stresses in a problem involving two concentric pistons in a pneumatic pump. The movement of the outer cylinder is accomplished by force transmitted through a set of locking balls. Sliding contact was destroying the device. Looking at the equations with no specific numbers substituted, revealed that in this case the stresses could be lowered by using softer materials. This allowed larger deformations (which were permissible in this case) to increase the size of the contact patch, thus decreasing the contact stresses. Switching from a steel construction to nylon balls and pistons gave positive longevity results. This problem would be classified as sliding spherical contact in a conforming groove, which is quite nasty mathematically, so we just built one.
The nylon pump is the opposite of the normal situation where increased hardness is sought to support the loads, while minimizing deflections. Examples of designs seeking very high hardness are ceramic bearings now being used in some turbochargers and aircraft quality bearings. The surfaces of these materials have yield points in excess of 200,000 psi. Small deformations (.0001 inch) are desired and very fine surface finishes are required. Deformations from the approach of centers of the contacting bodies in a turbine bearing running at 50,000 rpm could cause the blades to hit the outer housing, or induce unwanted vibrational modes.

In a number of machine design problems that I have worked on, contact stresses were the limiting factor. The solutions to these problems drove the overall form of the designs. Handbook formulas only cover the simple cases of Hertzian contact. I seem to run into a lot of contact stress problems and need a fast and accurate way to get answers. This is what motivated me to implement the integral formulation for stresses as a computer program.

Christopher R. McGoldrick, P.E.
Acknowledgements:

The most significant contributions to this investigation came from the following three references:

Seely and Smith [8]: The integral equations that were implemented in this investigation can be found here. The notation used in this text and the variable names used in the computer programs conform to this reference.

Johnson [1]: A very comprehensive reference of the field of contact mechanics with an emphasis on solved problems. It is reffered to many times in this investigation from the areas of theoretical development through practical applications.

Timoshenko and Goodier [13]: A classic reference in elasticity theory. The review of elasticity methods and development of the point load case is based on this work.


