Tribology on the Small Scale
A Bottom Up Approach to Friction, Lubrication, and Wear

C. Mathew Mate

MESOSCOPIC PHYSICS AND NANOTECHNOLOGY

OXFORD UNIVERSITY PRESS
Tribology on the Small Scale
MESOSCOPIC PHYSICS AND NANOTECHNOLOGY

Series Editors

Harold G. Craighead
Purusottam Jena
Charles M. Marcus
Allan H. MacDonald
Shaul Mukamel
Mark A. Reed
John C. Spence

1. Y. Imry: Introduction to Mesoscopic Physics
2. Y. Imry: Introduction to Mesoscopic Physics, Second Edition
4. P.A. Mello and N. Kumar: Quantum Transport in Mesoscopic Systems
5. D. Gatteschi, R. Sessoli and J. Villain: Molecular Nanomagnets
6. C. M. Mate: Tribology on the Small Scale – A Bottom Up Approach to Friction, Lubrication, and Wear
Tribology on the small scale

A bottom up approach to friction, lubrication, and wear

C. Mathew Mate
Hitachi San Jose Research Center
Some may wonder why the world needs another book on tribology. The importance of friction, lubrication, adhesion, and wear in technology and everyday life is well known; they are encountered whenever two surfaces come into contact, such as when you walk across a room, push a pencil across a piece of paper, or stroke your favorite pet. While many excellent books have been written on tribology, most have focused on analyzing the macroscopic aspects, with only slight attention paid to the rich interplay going on between the atoms and molecules at the contacting surfaces, as these have historically been poorly understood. In recent decades, however, many talented physicists, chemists, engineers, and materials scientists have finally begun to decipher the nanoscale origins of tribological phenomena. Given the tremendous progress and excitement generated by this endeavor, now seems the opportune time for a new book on tribology emphasizing how macroscopic tribological phenomena originate at the atomic and molecular level.

The goal of this book is to incorporate a bottom up approach to friction, lubrication, and wear into a versatile book on tribology. This is done by focusing on how these tribological phenomena occur on the small scale—the atomic to the micrometer scale. I hope to demonstrate that focusing on the microscopic origins leads to a superior understanding of tribology than typically achieved by tribology books that take a macroscopic empirical approach. It is also hoped that the reader becomes enthused with the same excitement as those working in the field have for unraveling the mysteries of friction, lubrication, and wear, as well as an appreciation for the many challenges that remain.

This book covers the fundamentals of tribology from the atomic scale to the macroscale. The basic structure—with chapters on topography, friction, lubrication, and wear—is similar to that found in conventional tribology texts. These chapters cover the microscopic origins of the macroscopic concepts commonly used to describe tribological phenomena: roughness, elasticity, plasticity, friction coefficients, and wear coefficients. Some macroscale concepts (like elasticity) scale down well to the micro- and atomic scale, while other macroscale concepts (like hydrodynamic lubrication) do not. In addition, this book also has chapters on surface energy and surface forces and covers topics not typically found in tribology texts, but which become increasingly important at the small scale: capillary condensation, disjoining pressure, contact electrification, molecular slippage at interfaces, and atomic scale stick–slip.

This book is intended to be suitable as a textbook for tribology courses taught at the advanced undergraduate and graduate level in many engineering programs. In terms of the scientific and mathematical background expected of the reader,
no special knowledge is assumed beyond that typically encountered by science and engineering students in their first couple of years at a university.

In addition to college students learning about tribology for the first time, this book is intended for several other audiences:

• Academics and scientists who wish to learn how friction, lubrication, and wear occur at the microscopic and atomic scales.
• Engineers and technicians who do not consider themselves tribologists, but who work with technologies (such as MEMS, disk drives, and nanoimprinting) where a good grasp is essential of how tribological phenomena occur on the small scale.

I would like to thank all those who provided help and encouragement during the writing of this book:

• Sonke Adlung and Oxford University Press for providing me with the opportunity to publish this book with them and for their encouragement and patience over the five year period that this book was written.
• Prof. Steve Granick and the members of his group for their hospitality during the six months that I spent on sabbatical at the University of Illinois. I would also like to thank Steve for several figures and his valuable comments on some of the draft chapters.
• My management at Hitachi for their support, particularly my immediate manager, Peter Baumgart, who carefully read and provided valuable comments on every chapter.
• My colleagues at Hitachi who were kind enough to comment on various draft chapters: Tsai-Wei Wu, Run-Han Wang, Robert Waltman, Bruno Marchon, Ferdi Hendriks, Bernhard Knigge, Qing Dai, Xiao Wu, Barry Stipe, Bing Yen, and Zvonimir Bandic.
• My wife, Shelli, who has always been a constant source of support and encouragement.
3.3 Plastic deformation 48
3.3.1 Basic relations 48
3.3.2 Hardness 49
3.4 Real area of contact 50
3.4.1 Greenwood and Williamson model 51
3.4.1.1 Example: TiN contacts 53
3.4.1.2 Real area of contact using the Greenwood and Williamson model 54
3.4.1.2.1 Example: Recording head on a laser textured disk surface 55
3.5 Inelastic impacts 59
3.6 References 61

4 Friction 63
4.1 Amontons’ and Coulomb’s laws of friction 63
4.2 Adhesion and plowing in friction 66
4.2.1 Adhesive friction 66
4.2.2 Plowing friction 68
4.2.3 Work hardening 70
4.2.4 Junction growth 70
4.3 Static friction 72
4.3.1 Stick–slip 74
4.3.1.1 Velocity-controlled stick–slip 75
4.3.1.2 Time-controlled stick–slip 77
4.3.1.3 Displacement-controlled stick–slip 78
4.4 References 81

5 Surface energy and capillary pressure 82
5.1 Liquid surface tension 82
5.2 Capillary pressure 85
5.2.1 Capillary pressure in confined places 87
5.2.2 The Kelvin equation and capillary condensation 90
5.2.2.1 Example: Capillary condensation of water in a nanosized pore 91
5.2.2.2 Example: Capillary condensation of an organic vapor at a sphere-on-flat geometry 91
5.3 Interfacial energy and work of adhesion 92
5.4 Surface Energy of Solids 93
5.4.1 Why solids are not like liquids 93
5.4.2 Experimental determination of a solid’s surface energy 95
5.4.2.1 Contact angles 96
5.4.2.1.1 Estimating interfacial energies 97
5.4.2.1.2 Zisman method for estimating surface energy for a solid 98
5.4.2.1.3 Types of wetting 101
6 Surface forces derived from surface energies
6.1 The Derjaguin approximation
6.2 Dry environment
6.2.1 Force between a sphere and a flat
   6.2.1.1 Example: Adhesion force between two polystyrene spheres
   6.2.1.2 Example: Adhesion force between a polystyrene sphere and a PTFE Flat
   6.2.1.3 Example: Adhesion force for an atomically sharp asperity
6.2.2 Adhesion-induced deformation at a sphere-on-flat contact
   6.2.2.1 The Johnson–Kendall–Roberts (JKR) theory
   6.2.2.2 The Derjaguin–Muller–Toporov (DMT) theory
   6.2.2.3 Adhesion deformation in nanoscale contacts
6.3 Wet environment
6.3.1 Force for a sphere-on-flat in a wet environment
   6.3.1.1 Example: Lubricant meniscus force on an AFM tip
   6.3.1.2 Solid–solid adhesion in the presence of a liquid meniscus
6.3.2 Water menisci in sand
6.3.3 Meniscus force for different wetting regimes at contacting interfaces
   6.3.3.1 Toe dipping regime
      6.3.3.1.1 Example: Toe dipping adhesion with exponential distribution of summit heights
   6.3.3.2 Pillbox and flooded regimes
   6.3.3.3 Immersed regime
6.3.4 Example: Liquid adhesion of a microfabricated cantilever beam
6.4 References
## 10 Lubrication in tight spots

10.1 Confined liquids 246
10.2 Boundary lubrication 255

10.2.1 Molecular mechanisms of boundary lubrication 256
10.2.2 Molecularly thin liquid boundary lubricant layers 260

10.2.2.1 Example of the importance of end-groups in a liquid lubricant film 262

10.3 Capillary and disjoining pressures 265

10.3.1 Disjoining pressure 265
10.3.2 Distribution of a liquid film around a pore opening 267

10.3.2.1 Example: Measurement of the disjoining pressure of a perfluoropolyether lubricant 269

10.3.3 Lubricant distribution between contacting surfaces 270
10.3.4 Meniscus force 272

10.3.4.1 Example: Stiction of a recording head slider 272
10.3.4.2 Calculating meniscus force 273

10.3.4.2.1 Example: Calculation of stiction force of disk drive sliders in the pillbox regime 275

10.3.4.2.2 Padded or stiction-free slider 276

10.3.5 Liquid menisci at high speeds 278

10.4 References 279

## 11 Atomistic origins of friction

11.1 Simple models for adhesive friction 284
11.2 Atomistic models for static friction 286

11.2.1 Frenkel–Kontorova model 287

11.2.1.1 Experimental realizations of ultra-low friction in incommensurate sliding systems 289

11.2.2 Tomlinson model 290

11.2.2.1 Example: An AFM tip sliding across an NaCl crystal at ultra-low loads 291

11.2.3 Molecular dynamic simulations 295

11.2.4 Example: Cold welding 295

11.2.5 Why static friction occurs in real-life situations 295

11.3 Atomic origins of kinetic friction 297

11.3.1 Sliding isolated molecules and monolayers across surfaces 297

11.3.2 Quartz crystal microbalance 299

11.3.2.1 Example: Xe on Ag(111) 300

11.3.3 Movement of a liquid film on a surface with the blow-off technique 301

11.3.3.1 Example: Wind-driven flow of perfluoropolyether lubricants on silicon wafers 302
11.3.4 Pinning of an absorbed layer 307
11.4 References 308

12 Wear 313
12.1 Simple model for sliding wear 314
12.2 Major influences on wear rates 317
  12.2.1 Wear maps 318
12.3 Mechanisms of wear 319
  12.3.1 Wear from plastic deformation 319
  12.3.2 Adhesive wear 320
    12.3.2.1 Example: An atomic level simulation of adhesive wear 321
  12.3.3 Abrasive wear 321
  12.3.4 Oxidative wear 325
    12.3.4.1 Metals 325
    12.3.4.2 Carbon overcoats 326
    12.3.4.3 Ceramics 326
12.4 Plasticity at the nanoscale 327
12.5 References 329
Index 331
This page intentionally left blank
Starting in childhood, we all acquired a sufficient working knowledge of tribology to lead happy, productive lives. Crawling as infants, we mastered the frictional forces needed to get us where we wanted to go. Eventually, we graduated from crawling to walking to school, where an occasional icy sidewalk, if one lived in a cold climate, provided a challenging lesson on slippage and traction. If our teacher at school asked us to move our chair backwards, we knew intuitively that the chair would be easier to slide if no one was sitting on it. Mastering friction was also a critical component in many of the games that we played, whether it was gripping a bat, maintaining traction when running and stopping, or putting a devilish spin on a ping-pong ball.

In addition to friction, we also encountered wear and adhesion at a young age. The detrimental aspects of wear may have been first learnt when our favorite toys wore out quicker than we felt they should. The more positive aspects of wear may have been first appreciated as we sanded our first woodworking projects and polished our first art sculptures into their final artistic shapes. We also quickly learnt that crayons and pencils become dull when rubbed against paper, but can be sharpened back up by grinding in a pencil sharpener. Our first awareness of adhesion may have been on a humid day when someone commented on how “sticky” it feels. Or perhaps it was when we first wondered why spiders and flies can walk on the ceiling, but dogs and cats can’t.

Once in college, science and engineering classes usually only cover the topics of friction, lubrication, adhesion and wear at a rudimentary level. For friction, all that is usually taught is that: static friction is greater than kinetic friction, friction is proportional to the normal contact force, and viscous friction in a fluid is proportional to velocity. The principles of thin film lubrication are covered only in an advanced class on fluid mechanics as an example of solving Reynolds’ equation. Wear is only briefly covered in specialized engineering classes on tribology, despite its pervasiveness as a failure mechanism; and adhesion is covered incidentally in courses on chemical bonding or polymer physics.

Considering how much we encounter the tribological phenomena of friction, lubrication, adhesion, and wear in our daily lives and the wide extent of these phenomena in technology, one might be puzzled why these topics are only marginally covered in our current education system. In the author’s opinion, the fault lies not with colleges and universities, but rather with the inherently complicated and interconnected physical origins of most tribological phenomena.
This multifaceted nature has made it difficult for scientists and engineers to develop predictive theories for most tribological phenomena. Instead, empirically derived trends (for instance that friction is proportional to the loading force) are often the only predictive tools available. These empirical approaches have the drawback of being predictive only over a limited range of parameters. Since the underlying physical mechanisms are not well understood, often one does not even know which are the important parameters or over what range the observed trends are valid. Similarly, if a purely analytical approach is attempted, the lack of knowledge of the relevant parameters often leads to inaccurate predictions of tribological behavior. This poor predictive power has led to the field of tribology being perceived in many scientific quarters as more of a “black art” than as a scientific discipline. This lack of predictive power may also be the reason why educators are reluctant to spend much time on tribology concepts whose application may be dubious in many situations.

For example, if one wanted to analyze the friction force acting on a chair sliding across a hardwood floor, the most expedient approach would be to take advantage of past empirical studies that have shown that friction is generally proportional to the loading force (the weight of the chair and the person sitting on it), with a proportionality constant called the coefficient of friction or $\mu$. Using this approach, the next step is to determine from experiment how $\mu$ depends on the parameters suspected of influencing friction: sliding velocity, hardness of the wood, type of floor wax, etc. After a few hours of experiment, one would begin to have a good idea how friction depends on these parameters, but, without further experimentation, would have trouble predicting how the friction might change if new parameters were introduced, for example, by adding felt pads to the bottom of the chair legs to prevent them from scratching the hardwood floor.

While many tribology problems are still best approached through empirical investigations, these types of investigations are not the focus of this book. Instead, the focus is on the physical origins of tribology phenomena and how understanding these can be used to develop analytical approaches to tribological problems. In essence, the goal is to make tribology less of a black art and more of a scientific endeavor. This will be done by emphasizing how the tribological phenomena of friction, lubrication, adhesion, and wear originate at the small scale; or, equivalently, how physical phenomena occurring at the atomic to micron scale eventually lead to macroscale tribological phenomena. The hope is that, once readers have gained a solid understanding of the nanoscale origins of tribological phenomena, they will be well equipped to tackle new tribology problems, either by applying analytical methods or developing better empirical approaches.

1.1 Why is it called tribology?

The pursuit of the microscopic origins of friction, lubrication, adhesion, and wear is not a recent scientific activity. Over the centuries, many have pondered on
these origins, and, in recent decades, it has become quite fashionable for leading scientists to take up the challenge. One of the early pioneers and champions of the microscopic approach was Prof. David Tabor (1913–2005) of Cambridge University. One of the frustrations faced by Tabor and others working in this field at that time was the lack of a scientific name for the area of study encompassing all the phenomena occurring between contacting objects. It was felt that this lack of terminology was depriving the field of a certain level of status and respect within the scientific community. [For example, some of Tabor’s colleagues at the Cavendish Laboratory in Cambridge would disparagingly refer to his research group as the “Rubbing and Scrubbing Department” (Hahner and Spencer 1998).] To counter this, Tabor coined the name tribophysics for the research group that he headed while investigating practical lubricants, bearings, and explosives at Melbourne University during the Second World War, which he derived from the Greek word tribos meaning rubbing.

In 1966 the Jost Report produced by a Committee of the British Department of Education and Science officially launched the word tribology to describe the entire field, which was derived from Tabor’s earlier word, tribophysics (Jost 1966). While the literal translation of tribology is “the science of rubbing,” in the Jost Report, a longer, somewhat less satisfying, definition was adopted: “The science and technology of interacting surfaces in relative motion and of associated subjects and practices.” After the Jost Report, the term tribology quickly became established as the field’s official name, and the word tribology now commonly appears in the titles of papers, books, journals, professorships, and institutions concerned with this topic.

While the name tribology has certainly increased the credibility of the field as a valid discipline of scientific and engineering, the term remains somewhat unknown outside the field. So, tribologists need to be prepared to explain the word to those who have not heard of it before and to counter those scientists who want to make fun of tribologists by claiming that tribology stands for “try-again-bology.”

1.2 Economic and technological importance of tribology

One of the goals of the 1966 Jost Report was to document the potential economic savings that could be achieved through the development and adoption of better engineering practices for minimizing the unnecessary wear, friction, and breakdowns associated with tribological failures. The possible savings within the United Kingdom were estimated to be roughly equivalent to 1% of its GNP. Since the Jost Report, other agencies have also evaluated possible savings, and the consensus view now is that 1.3% to 1.6% of GNP could be gained through proper attention to tribology (Jost 1990).

While these potential economic benefits have long been recognized, this has not always been followed up with the level of investment in tribology research and development felt to be warranted by many tribologists. Possibly a major factor
in the reluctance to back tribology projects comes from its historical ‘black art’ character, which tends to cast doubt on how successful a proposed tribology project will be. Hopefully, this book will help diminish these doubts by demonstrating a rational and scientific basis for approaching tribological problems. Another way of diminishing doubts about the science of tribology is through a few success stories.

1.2.1 Some tribology success stories

1.2.1.1 Reducing automotive friction

Improving the fuel economy of cars and trucks has long been a major technology goal of the automotive industry (Tung and McMillan 2004). An obvious way for improving fuel efficiency is reducing energy loss through friction. For a typical car, the energy produced by the engine gets consumed as follows (Pinkus and Wilcock 1978):

- 12% drives the wheels to overcome inertia, rolling resistance, air resistance, etc.: this is the energy that actually pushes the car forward.
- 62% is lost through engine heat.
- 16% is lost internally in the engine through friction from the motion of the pistons, valve train, bearings, etc. in the engine, as well as from drawing in air, expelling exhaust, and operating the various fans and pumps in the engine.
- 10% is lost through friction in the transmission, differential, and wheel bearings.

In principle, if one could build a car out of frictionless components, fuel consumption could be reduced by upwards of 15% (Tung and McMillan 2004). In the absence of such frictionless components, the automotive industry has struggled over its history to reduce friction wherever possible in order to improve fuel efficiency. Given the fairly mature nature of automotive technology, only small incremental improvements are achieved in each design cycle. The accumulative effect of these small improvements, however, has been quite substantial: from 1980 to 2005, the average fuel economy of cars sold in the U.S. increased from 20 to 25 miles per gallon, despite the average horsepower almost doubling from 100 to 190 (New York Times, March 30, 2006).

Figure 1.1 shows the impact that improved engine oils have had on fuel efficiency. From 1993 to 2005, changes in engine oil resulted in a 1.2% improvement in fuel efficiency through the reduction of non-viscous engine friction. While this may seem an inconsequential amount, it is enough to reduce the annual gasoline consumption in the U.S at the 1997 consumption levels by 1.5 billion gallons. Along with the improvement in fuel economy, these newer engine oils also achieve better wear protection capability, oxidation resistance, piston cleanliness, and sludge and varnish formation tendencies.

Over the same 1993–2005 period, an additional 1% improvement in automotive fuel efficiency was achieved by using lower viscosity transmission fluids and
Fig. 1.1. Improvement in fuel efficiency due to the reduction in engine friction achieved through changes in engine oils for gasoline fueled (GF) vehicles. GF-1 to GF-4 engine oil specifications correspond to changes in base oil and additive chemistry. Reducing the oil viscosity from 5W-30 to 5W-20 provides an additional 0.5% improvement. The baseline is Energy Conserving II engine oils available prior to 1993. Courtesy of Ron Jensen and Arup Gangopadhyay at Ford Research and Advanced Engineering.

another 1% by using lower viscosity gear lubricant. A substantial jump in fuel economy is expected with the adoption of continuously variable transmissions (CVT), which can provide ~10% improvement over a six-speed transmission through the reduction of friction (Gangopadhyay 2006).

1.2.1.2 MEMS and solving adhesion in Digital Micro-mirror Devices
It has long been realized that miniaturization of machines can result in major new technologies. In recent years, the most promising way for fabricating microscale devices has been to use processes originally developed for fabricating semiconductor electronic devices. By using these fabrication processes, mechanical functions (such as actuation, fluid flow, thermal response, etc.) can be integrated on a small area of a chip along with electronic signal processing. A big advantage of fabricating these *micro-electromechanical systems* (MEMS) with semiconductor processing techniques is that it achieves excellent economies of scale, since many devices are fabricated simultaneously onto a single chip. This low unit cost, along with the integration of mechanical and electrical functions into a small space, enables whole new types of technologies to become commercially viable. Currently, the most successful MEMS devices are the accelerometers used to trigger
airbags in cars and the Digital Micro-mirror Devices (DMD) used in Digital Light Processing (DLP) for video projection.

For those MEMS devices where surfaces rub against each other, however, this process of miniaturization often results in severe tribological problems (Williams and Le 2006). For example, much fanfare was made about the first working MEMS micro-motor in 1988 (Fan et al. 1989). While these micro-motors rotated as desired, they suffered from high friction and excessive wear, resulting in lifetimes too short for most applications. Despite intensive research efforts (Maboudian et al. 2002; Williams and Le 2006), no satisfactory lubrication scheme has been developed for MEMS micro-motors. The situation is similar for many other potential MEMS devices facing tribology hurdles. Consequently, MEMS tribology problems are avoided whenever possible by designing them so that the moving parts do not rub against each other.

Digital Micro-mirror Devices (DMD)—developed by Texas Instruments and used in DPL video projection devices such as large-screen televisions—provide, however, a success story where a microscale tribology problem had to be solved in order to achieve sufficient reliability for commercialization. In a DMD, an array of mirrors, each about 16 µm across, is used to project an image onto a video screen. As illustrated in Fig. 1.2, the intensity of each pixel is controlled by rotating the individual micro-mirrors through ±10° by using electrostatic attraction. Before shipping its first DMD product in 1996, Texas Instruments carried out extensive reliability engineering and testing (Douglass 1998; Van Kessel et al. 1998), and a number of reliability issues were addressed: stuck mirrors, fatigue of the mirror hinge, hinge memory, mirrors breaking as result of vibration and shock, and particles preventing mirrors from rotating. Here we focus on the sticking of the mirrors against their mechanical stop, which was a persistent tribology problem for which the TI engineers implemented a combination of clever solutions based on a good micro-understanding of the adhesive mechanism.

In a DMD, the individual micro-mirrors are rotated from the on to off positions. To ensure that each mirror has the correct angular position at the end of the rotation, the mirror yoke is designed to come to rest against a mechanical stop, as illustrated in Fig. 1.2. During the development of the DMD, it was found that adhesive forces acting at this contact would sometimes be large enough to result in the mirror sticking against the stop, making it non-functional. These adhesive forces originate from the meniscus force due to water vapor condensing around the contact (discussed in Chapter 6) and from van der Waals forces (discussed in Chapter 7). A number of design changes were implemented in the DMD to reduce the magnitude of these adhesive forces and to improve the release function:

- The DMD was hermetically sealed in a dry environment to minimize the capillary condensation of water.
- The contacting parts were covered with a low surface energy “anti-stick” material to minimize the van der Waals force.
• Miniature springs were added to the parts of the mirror yoke that makes contact—the “spring tips” shown in Fig. 1.2. These spring tips store elastic energy when the parts come into contact, that helps push the mirror away from the surface when the electrostatic attractive force is released.

These design modifications dramatically reduced the tendency of the micro-mirrors to stick, greatly improving the DMD operating margins (Douglass 1998; Van Kessel et al. 1998). From 1996 to 2006, Texas Instruments sold over 10 million DLP systems with DMDs, demonstrating that, when proper attention is paid to the microscale tribology issues, a reliable product can include a MEMS device with contacting components.

1.2.1.3 Slider–disk interfaces in disk drives

The disk drive industry may be the most striking example of where an exponential improvement of a technology over many decades has been sustained through the continual solving of tribological problems.

The first disk drive was introduced in 1956 as part of the IBM RAMAC computer. The RAMAC disk drive stored an impressive, for its day, 4.4 megabytes in a space the size of a small refrigerator; in 2006, the smallest drives, a little over an inch across, stored up to 10 gigabytes. This tremendous increase in storage capacity and reduction in size came largely about from an exponential growth in the areal storage density (the number of bits that can be stored per square inch on a disk surface): from 1956 to 2006, the areal storage density of disk drives increased by a factor of $10^5$, at annual growth rates
ranging from 30 to 100%. To sustain this tremendous growth in areal density over a period of time spanning decades, all aspects of disk drive technology had to be continually improved. Here we focus on the tribological challenges recently faced by the disk drive industry to sustain this rapid rise in areal storage density.

Inside a disk drive, a slider with read/write recording head flies over a rotating disk, as illustrated in Fig. 1.3. The information is stored as magnetic bits in a thin layer of magnetic material on the disk surface. Since the magnetic field from these bits decays rapidly away from the disk magnetic medium with a decay distance that scales with bit size, the magnetic spacing between the head sensor and the magnetic medium on the disk needs to scale in future disk drive products with the lateral size of the magnetic bit on the disk: as the areal density goes up, the spacing must go down. Fig. 1.4 plots the relationship between magnetic spacing and linear bit density.

**Fig. 1.3.** (Top) A recording head slider flying over a rotating disk surface in a disk drive. (Bottom) An enlarged view of the trailing edge of the slider where the recording head is located. The cross section illustrates the various components of the recording head and disk that contribute to the magnetic spacing, which is defined as the distance from the top of the disk magnetic medium to the bottom of the head sensor. The higher the areal storage density, the smaller the magnetic spacing needs to be.
Since the recording head slider flies at high speeds (typically around 10 m/s) with a clearance of just a few nanometers over the disk surface, careful attention needs to be paid during the product development and manufacture towards minimizing the risks from high speed contacts between the slider and disk. To minimize the number of these contacts, the slider is designed to fly on an air bearing generated by the disk pulling air underneath the slider and over a series of steps and pockets precisely fabricated onto the bottom surface of the slider. These surface features form an air bearing surface (ABS) that generates a lifting force, which balances the loading force from the suspension, enabling the slider to fly over the disk with its trailing edge a few nanometers above disk surface. To protect against occasional impacts, the disk and slider surfaces are coated (about 3–4 nm for 2005 drives) with a hard material, usually diamond-like carbon. To further ensure that the slider–disk interface is not damaged by the high speed impacts, a molecularly thin film of lubricant (typically about 1 nm thick) is applied over the carbon overcoated disk.

As is apparent from Fig. 1.4, the space available between the head sensor and the disk magnetic layer for these protective layers has been rapidly diminishing. The dramatic reduction in magnetic spacing from 96 nm to 16 nm during the decade from 1995 to 2005 was not achieved through any major technological innovations, but rather through careful attention to tribological detail; in particular:

- Improved fabrication processes and better tolerance control for slider air bearing surfaces (ABS), resulting in better control of the final slider–disk spacing.
• Introduction of denser carbon overcoats through changes in composition and deposition processes
• Design of lubricant systems that preserve at least a monolayer of lubricant on disk surfaces over a disk drive’s lifetime
• Careful control of disk topography to achieve sub-nanometer roughness

With these improvements, drives achieve excellent long term reliability while their sliders fly over the disk surfaces at high speeds with incredibly small clearances—only 3 to 5 nm for 2005 drives. By 2015, the disk drive industry hopes to increase the areal storage density to more than 1 Tbits/in$^2$, which will mean reducing the magnetic spacing to less than $<7$ nm (Wood 2000; Mate et al. 2005). This ultra-small magnetic spacing is unlikely to be achieved simply by evolving the current slider–disk interfaces, as the various constituents—overcoats, lubricant, roughness, and clearance—are approaching their physical limits (Mate et al. 2000; Mate et al. 2005). To address the tribology challenges being imposed by these ultra small spacings, a number of new approaches are being proposed:

• **Thermal fly-height control** – A small heater integrated into the recording head is used to actively control the clearance to less than a few nanometers (Suk et al. 2005).
• **Wear-in pad** – The recording head is located on a small pad that protrudes from the trailing edge of the slider. A few nanometers is burnished off the pad once the slider is mounted in the drive to remove the slider overcoat’s contribution to the magnetic spacing and to improve the flying height tolerances (Singh et al. 2004; Strom et al. 2004).
• **Contact recording** – The trailing edge of the slider is run in contact with the disk eliminating the contribution of clearance to the magnetic spacing (Mate et al. 2004; Mate et al. 2005).

Of these technologies, thermal fly-height control has already begun to be implemented in 2005 disk drives. For the wear-in pad and contact recording concepts, serious tribology issues need to be resolved: excessive wear, high friction, and corrosion of magnetic material through overcoats being damaged by wear (Mate et al. 2004; Singh et al. 2004; Mate et al. 2005).

1.3 A brief history of modern tribology
The practice of tribology goes as far back as the prehistoric humans, who first used wear to fashion tools and friction to start fires.

Scientific investigation of friction began with Leonardo da Vinci who recorded in his notebooks the observation that friction was proportional to load. This law of friction was rediscovered by the French physicist Guillaume Amontons, who published his findings, now referred to Amontons’ Laws of Friction, in 1699. Amontons’ results immediately provoked controversy about the microscopic
origins of friction, which even today has not been fully resolved. The possible origins of friction are discussed in Chapter 4 and 11.

With the industrial revolution, machinery of all sorts came into widespread use and with it a growing need for a better control of friction, lubrication, and wear. During this period, the principles of hydrodynamic lubrication were first discovered through the experimental work of Beauchamp Tower (1884) and the theoretical work of Osborne Reynolds (1886). The subsequent development of this theory of hydrodynamic lubrication enabled reliable bearings to be designed for lubricating the machinery of the modern age. Lubrication is discussed in Chapters 9 and 10.

During the twentieth century, enormous industrial growth and development of new technologies further fueled demand for better tribology. To meet this demand, numerous tribological engineering solutions were developed in the last century, notably:

- Hydrodynamic bearing design (Chapter 9)
- Theory of contact mechanics (Chapter 3)
- Synthetic lubricants
- Solid lubricants
- Wear-resistant materials

It has long been thought that a lack of understanding of the microscopic origins of tribological phenomena has impeded the development of the best tribology technology. In the middle of the twentieth century, several scientists did pioneering studies on the microscopic origins of friction, lubrication, and wear, producing:

- Bowden and Tabor’s theory of molecular adhesion for friction (Chapter 4 and 11)
- Greenwood and Williamson’s analysis of multi-asperity contact area (Chapter 3)
- Hardy’s studies of boundary lubrication (Chapter 10)
- Archard’s law for adhesive wear (Chapter 12)

The latter part of the twentieth century saw an accelerated effort to determine the micro- to atomic scale origins of tribology phenomena. Besides the desire to reduce unnecessary friction and wear in existing technological applications, two new factors have also driven this recent activity in micro- and nanotribology:

1. *Scientific* – New experimental and theoretical techniques for characterizing materials at the nanometer scale, facilitating the discovery of the atomic and molecular origins of friction, lubrication, adhesion, and wear.

2. *The promise of new micro- and nanoscale technologies* – Numerous breakthrough technologies with miniaturized moving components are expected to
become enabled once solutions are developed for their tribology problems occurring on the micro- and nanoscales.

1.3.1 *Scientific advances enabling nanoscale tribology*

The nature of contact makes its microscopic origins difficult to study. Generally, contact occurs at a multitude of contact zones at the apexes of small protrusions or asperities on opposing surfaces, as illustrated in Fig. 1.5(a). Since these contacts are sandwiched between two solids, they are inaccessible to most scientific characterization techniques. Adding to this difficulty, the contact occurs

![Fig. 1.5](image)

Fig. 1.5. (a) Due to surface roughness, contact between two solid surfaces occurs primarily at the summits of the surface asperities. (b) Vacuum techniques for analyzing surfaces before or after contact. Auger electron spectroscopy (AES) = impinging electrons, energy of ejected electrons analyzed; electron spectroscopy for chemical analysis (ESCA) = impinging x-rays, energy of ejected electrons analyzed; secondary ion mass spectroscopy (SIMS) = impinging ions, energy and charge-to-mass ratio measured of scattered and ejected ions. (c) Optical techniques such as Raman spectroscopy, Fourier transform infrared spectroscopy (FTIR), and sum frequency generation can now be used to analyze a buried contact interface. (d) An atomic force microscope (AFM) tip rubbing against a surface can be used to simulate a single asperity contact.
principally at the summits of the surface roughness, meaning that the material volume affected by contact tends to be nanoscopically small and difficult to detect. The push and pull of the asperities rubbing against each other further complicates matters as the contacting microstructures are not static, but evolve as the contact progresses.

Fortunately, over the past half century, numerous analytical techniques have been developed for characterizing surfaces and nano-amounts of materials. These newer techniques have led to a wealth of information on how the structure and chemistry of surfaces evolve during contact, and influence tribological phenomena.

For one category of surface analytical techniques [illustrated in Fig. 1.5(b)], the individual surfaces are placed in high vacuum and irradiated with electrons, ions, or x-rays and the kinetic energy of the ejected electrons or ions measured with an energy analyzer. Since only electrons or ions near the surface can be ejected toward the detector, these are very sensitive techniques for determining surface structure and chemistry, but the necessity of vacuum means these techniques are only used for analyzing surfaces before and after contact. Some vacuum techniques valuable for characterizing the chemical composition and molecular structure of tribological surfaces are:

- Auger electron spectroscopy (AES)
- electron spectroscopy for chemical analysis (ESCA)
- secondary ion mass spectrometry (SIMS)

Fuller descriptions of these techniques can be found in (Briggs and Seah (1990) and Somorjai 1994, 1998).

Recently, several optical and x-ray probes have been developed that allow for in situ characterization of the buried contacting surfaces as illustrated in Fig. 1.5(c):

- Fourier transform infrared spectroscopy (FTIR)
- Raman spectroscopy (Campion and Kambhampati 1998)
- sum frequency generation (Shen 1994)
- x-ray diffraction and x-ray reflectivity (Mate et al. 2000)

Since contact typically occurs at the summits of asperities, characterizing the topography of contacting surfaces has always been an important starting point for characterizing a tribological surface. Initially this was done with stylus profilometers that only measured profiles with micron resolution along single lines. Next came the scanning electron microscope (SEM), which is capable of imaging surfaces with nanometer lateral resolution, but is unable to quantify the heights of the surface features. More recently topography measurements have been dominated by scanning probe techniques—the scanning tunneling microscope (STM) and, more importantly, the atomic force microscope (AFM)—which
can generate 3D topography images with true atomic resolution. The use of AFM for topography measurements is discussed in Chapter 2.

In addition to topography measurements, the AFM is also able to measure the contact forces acting on a single asperity tip, as illustrated in Fig. 1.5(d). Consequently, since its invention in 1986, AFM has become one of the principal tools for investing nanoscale contact and friction forces. The use of AFM for measuring contact forces is discussed in Chapters 6, 7, 8, 10, and 11.

The surface force apparatus (SFA), which was developed more than a decade earlier than the AFM, is another important tool for measuring the forces between contacting surfaces; in an SFA, the forces are measured between two atomically smooth mica sheets in a variety of chemical environments. SFA force measurements are discussed in Chapters 7, 8, and 10.

The dramatic increase in computer performance over the past few decades has led to the prevalence of many computer based simulation techniques for predicting physical phenomena. Several of these simulation techniques have been adapted to studying tribological phenomena:

- Molecular dynamics simulations have been used to directly address the atomic origins of friction, lubrication, and adhesion, but are limited to small volumes and short time scales. Figure 1.6 shows an example of the behavior of lubricant molecules sandwiched between two asperities as determined by a molecular dynamics simulation (Thompson and Robbins 1990; He et al. 1999; Gao et al. 2004).
- Finite element analysis simulations are routinely used to analyze the contact mechanics of macro-, micro-, and nanoscale multi-asperity contacts.
- Most modern fluid film bearings are designed using computer aided design (CAD) software. In many instances, analysis programs have been developed to handle bearings with sub-micron dimensions.

1.3.2 Breakthrough technologies relying on tribology at the small scale

Perhaps the most exciting aspect of modern tribology’s push towards the nanoscale is the potential payoff that this research can have for enabling breakthrough micro- and nanoscale technologies (Perry and Tysoe 2005).

As device components miniaturize, they become more susceptible to the forces and atomic scale phenomena occurring at their contacting surfaces. As a consequence, established engineering solutions that work well for macro-tribology situations are generally inadequate for micro- and nanoscale devices.

For example, fluid film bearings provide lubrication for the moving parts of most macroscopic machines. The thickness of the lubricant film in these bearings is typically microns to millimeters; but, when devices are miniaturized to the sub-millimeter scale, the gap between the moving parts of the bearing’s lubricant film becomes sub-micron. This small gap leads to contact between the asperities on the moving surfaces as well as to confinement effects within the lubricant film.
that dramatically increase the lubricant’s viscosity; both friction and wear from contact and the enhanced viscosity can reduce the bearing’s effectiveness if it is not properly redesigned for the small scale.

Further, as the parts become micro-sized, the capillary and molecular adhesion forces begin to dominate over gravity and inertia forces. So, while loading forces and bulk hardness may be the major factors determining friction and wear of macro-sized objects running in dry contact, they are less important for minute objects where molecular adhesion forces are comparable to loading forces.

Previously, we discussed the two examples of the digital micro-mirror devices and disk drives where successful products were shipped once solutions were found to the problems associated with micro- and nanoscale tribology. In the next few sections, we briefly discuss exciting future technologies where implementation is being held up by problems of tribology on the small scale.
1.3.2.1 Nanoimprinting

Driven by the difficulties and high tooling costs of extending photolithography techniques to the production of sub-100 nm features, a tremendous research effort is being expended on developing alternatives to photolithography for fabricating nanoscale structures. Due to its low cost and apparent simplicity, the most promising alternative appears to be imprint lithography, which is called *nanoimprinting* when used to replicate surfaces patterned with features less than 100 nm in dimension (Chou et al. 1996; Scheer et al. 1998; Michel et al. 2001; McClelland et al. 2002; McClelland et al. 2005). The devil of course is always in the details, and, for nanoimprinting, a very critical detail is the high level of tribology related defects generated during the imprint process.

Figure 1.7 illustrates a typical imprint lithography process. In this technique, a transparent template (usually quartz) is first fabricated with the sub-100 nm

![Figure 1.7](image)

*Fig. 1.7. Schematic of imprint lithography process: (a) Alignment of template and deposition of a UV-curable resist material that will serve as the etch barrier. (b) Imprinting and exposure to UV. (c) Withdrawal of the template. (d) After etching and removal of the resist mask, the substrate has the inverse pattern of the template.*
features that are the inverse of the structures to be replicated. This expensive and time-consuming step of making these original nanoscale features is usually done with electron beam lithography. Next, the template is aligned over the substrate, and a low viscosity, UV-curable resist material is injected between them [Fig. 1.7(a)]. The template and substrate are pressed together, and the resist exposed to UV light through the template [Fig. 1.7(b)]. The template is withdrawn, leaving the inverse pattern replicated in the resist material [Fig. 1.7(c)]. Finally, this pattern is etched into the substrate to create a permanent inverse replica of the original pattern [Fig. 1.7(d)].

A major problem faced by nanoimprinting processes is minimizing defects to an acceptable level. For example, if features are spaced 40 nm apart, this corresponds to a density of $6.25 \times 10^{10}$ per square centimeter. Since the features are only a few tens of nanometers across, a few cubic nanometers of material going astray every now and then can easily add up to an unacceptable number of defects.

For a nanoimprinting process such as illustrated in Fig. 1.7, the two principal mechanisms for generating defects originate from contact phenomena during the imprint process, so are tribological in origin:

1. When the template and substrate are pushed together, the uncured resist material needs to flow around and wet all the nanoscale features on the template surface, as any uncovered portions become defect voids in the replicated pattern. The spreading of the uncured material into the nanoscale features on the template surfaces is governed by the flow of liquids in tight spaces, which is discussed in Chapters 9 and 10. Wetting is discussed in Chapter 5.

2. When the template is withdrawn, it has to separate cleanly from the resist material. If small bits of this material adhere to the template during withdrawal, not only does this create void defects for that particular replica, but also, if the material that adhered to the template is not cleaned off, this defect will propagate to future replicas.

A common way to prevent the cured resist from adhering to the template is to coat the template surface with a thin film, called a release layer, that has a low surface energy. This release layer is typically one molecule thick, as a thicker film would adversely impact the dimensions of the replicated nanoscale features and a thinner film would provide inadequate coverage. The release layer works by lowering the resist–template adhesion energy below the resist cohesion energy, reducing the probability of pulling out bits of the cured resist. The physical origins of and the interconnections between adhesion force and surface energy are discussed in Chapters 5, 6, and 7.

Even with a low surface energy release layer, proper release of the smallest features can be problematic as high friction forces on the sides of these features can dominate the relatively small cohesive force. Again this is a case where the surface area of the features does not scale down as fast the volume of the features as they are made smaller, leading to the surface forces of
friction and adhesion becoming larger than the internal cohesion energy as features are made nanoscopic. Hopefully, future research will result in a sufficient nanoscale understanding of the adhesion and material strengths to enable the development of nanoimprinting release steps that produce acceptably low defect levels.

1.3.2.2 IBM’s millipede for high density storage
Not only can scanning probes, such as STM and AFM, image surfaces with atomic resolution, but they can also create marks on a surface a few atoms across. It has long been realized that such a “marking” scheme has the potential to be an ultra-high density storage technology. One promising scanning probe storage device is the IBM millipede. The name “millipede” originates from the array of several thousand AFM cantilevers used in parallel to make small indents in a polymer film. The presence of an indent in a particular location serves as a logical bit “1,” while the absence of an indent provides a logical bit “0.” In 2004, the IBM research team demonstrated millipede devices reading and writing at an areal storage density of 641 Gbits/in\(^2\) (about seven times higher than a 2004 disk drive areal density) at an error rate better than \(10^{-4}\) (Pozidis et al. 2004).

A long term goal of the millipede project is to achieve indents that are only a few nanometers apart in order to produce areal densities of \(\sim 10\) Tbits/in\(^2\), two orders of magnitude greater than a 2005 disk drive.

Figure 1.8 illustrates the process whereby a millipede cantilever makes an indent as it scans across the surface of the polymer medium. The thermomechanical write process is as follows: the tip is heated by resistive heating of the cantilever; the hot tip is brought into contact using electrostatic attraction; and the contact with the hot tip softens the polymer enabling the light load on the tip to indent the polymer. Since the tips are very sharp, with radii on the order of nanometers, indents as small as a few nanometers across can be achieved; this is what enables the ultra-high storage densities. For the read process, the tip is scanned across an indent while the tip–polymer thermal conductivity is measured; when the tip goes into the indented hole, the thermal conductivity increases due to the increased contact area, revealing the indent location.

With indents only a few nanometers across, the millipede needs to be designed so that the constant rubbing of the tip against the polymer leaves the surface undamaged at the nanometer scale: The potential for nanoscale surface damage from contact is one of the most critical problems faced by the millipede project!

One type of surface damage faced by the millipede are ripples, 1 to 30 nm in height, induced in the polymer film by the frictional forces of the AFM tips rubbing against it (Schmidt et al. 2003). The IBM Research team has found that this rippling problem can be alleviated by having a cross-linked polymer underlayer, as shown in Fig. 1.8, but this rippling still remains a long term reliability concern for a millipede device (Knoll et al. 2006).

Another major surface damage problem faced by the millipede is the nanoscopic wear of the tips. To achieve its high storage density, a millipede
Fig. 1.8. How a millipede cantilever writes a bit “1” in the 50 nm thick poly(methyl methacrylate) or PMMA film. As the tip scans over the substrate, a current pulse is passed through the resistive heater on the cantilever, momentarily heating the cantilever tip to 400°C. A voltage pulse between the tip and substrate brings the hot tip into contact, with a slight loading force, softening the PMMA polymer film and allowing the tip to make a small indentation in the polymer film that serves at a “1” bit. Reprinted from Knoll et al. (2006) with permission from Elsevier.

depends on thousands of tips all having a radius on the order of a few nanometers. If one tip suffers wear damage that increases its effective radius by a few nanometers, this tip would then have difficulty reading the indents that it had previously written, and the data stored using that tip might be lost. As of 2006, no reliable method has been reported for preventing wear of AFM tips.

1.3.2.3 Nanotechnology

Much of the current excitement about nanotechnology and nanoscience is driven by the development of new techniques that allow not only the manipulation of matter at the atomic scale, but also the integration of these nanoscale structures into complicated systems or devices. If this bottom up approach is to become a viable technology, new manipulation methods are needed to move nanoscale objects reliably to places where they can be assembled into nanoscale devices. This will require a good understanding of how friction and adhesion occur on the atomic and molecular scale, in essence: nanotribology.

Many people, when promoting nanotechnology, like to quote the following passage from Richard Feynman’s famous 1959 article There’s Plenty Room at the Bottom (Feynman 1960): “The principles of physics, as far as I can see, do not speak against the possibility of maneuvering things atom by atom. It is not an attempt to violate any laws; it is something, in principle, that can be done; but, in practice, it has not been done because we are too big.”
What is usually neglected are Feynman’s caveats in that paper on how friction and adhesion at the atomic and molecular level present major challenges for developing atomic scale machines:

- “All things to do not simply scale down in proportion. There is the problem that materials stick together by molecular (Van der Waals) attractions.”
- “Lubrication involves some interesting points...Let the bearings run dry!”

Almost a half-century later, in the emerging field of nanotechnology, people are on the verge of fulfilling Feynman’s vision of building machines atom by atom. With our accumulated knowledge over the same last half-century of the atomic origins of friction and adhesion, we are also beginning to appreciate the difficulties and subtleties of manipulating nanoscopic pieces of matter: Even with his caveats, things are not as simple as Feynman suggested!

While Feynman was right that molecular adhesive forces dominate over external forces when things are scaled down to the molecular level, he neglected to mention the many other molecular forces that occur besides van der Waals attraction. As we will see in Chapters 6 and 7, in addition to van der Waals forces, other forces—electrostatic forces, hydrophilic, hydrophobic, structural, double-layer, and capillary adhesive forces—contribute to the adhesion between small objects. We also now appreciate better the major role that molecular adhesion plays in generating the friction as two materials slide over each other, as discussed in Chapter 11.

Feynman’s phrase “Let the bearings run dry!” was based on the assumption that friction originates from viscous forces between atoms acting incoherently across a sliding interface and, therefore, should be negligible at small sliding velocities. As discussed in chapter 10, we now know that molecules trapped in between two solids give rise to a static friction force by interlocking contacting surfaces. The negligible friction envisioned by Feynman has only been observed in a few carefully constructed experiments designed to bring into contact two solid surfaces with incommensurate atomic structures, with weak interactions across their sliding interface (Chapter 11). This phenomenon of superlubricity has given hope to the concept that nano-machines can be built with little or no friction between their moving parts.

1.4 References


CHARACTERIZING SURFACE ROUGHNESS

To begin to understand friction, lubrication, and wear, we need to first understand how two surfaces make contact with each other. Bowden (1950) pointed out that “Putting two solids together is rather like turning Switzerland upside down and standing it on Austria—the area of intimate contact will be small” (Bowden 1950). For most tribological surfaces, the slopes of the asperities are much smaller than the mountain slopes of Switzerland and Austria. For very smooth surfaces (for example the slider–disk interface of a disk drive) the situation is more like putting Kansas on top of Nebraska where much of the contact is just as likely to occur on the debris (such as cows) as with topographic features (such as farmhouses or the occasional hill). The widely different topography of Switzerland–Austria versus Kansas–Nebraska leads to quite different contact: interlocking jagged peaks versus mashing cows between two flat pastures. Likewise, as we examine what happens at the atomic scale when two solids touch, much depends on the nanoscale surface texture: For atomically smooth mating surfaces, any loose atom or molecule acts like debris rolling around between the two surfaces. For sharp asperities, strong molecular adhesive forces and high contact pressures result in extensive plastic deformation on the atomic scale when two solids touch. Understanding tribological behavior first depends on understanding the topography of the two contacting surfaces. In this chapter, we cover the general concepts of surface topography, how to apply them to tribological surfaces, and experimental techniques for measuring surface topography, principally, atomic force microscopy (AFM).

2.1 Types of surface roughness

While an infinite variety of surface topographies exist, only a small number of ways are used, in practice, to describe surface topographies. This can work well for those surfaces (which fortunately includes most practical surfaces) where the topography is a mixture of forms that each dominate at different length scales. For example, Fig. 2.1 illustrates a surface texture that is regular at large length scales and random at small length scales. The regular structure could have come from a grinding or machining process producing a wave-like pattern over the surface with a well defined ‘direction of lay.’ The line profile orthogonal to the direction of lay illustrates the waviness spacing (or wavelength) and waviness height. Further magnification reveals a finer irregular structure superimposed on the regular wavy structure. With further magnification the
roughness from the individual atoms is revealed. The revealing of different roughness scales with higher magnification is a common occurrence in topography measurements, leading one well known tribologist (Archard 1957) to describe surfaces as ‘protuberances on protuberances on protuberances.’

The geometric characteristics of the surface in Fig. 2.1, which are common to many types of surfaces, can be categorized from the smallest scale to the largest scale as follows:

- **Atomic scale roughness**: Roughness from the atoms on the surface of the solid.
- **Roughness**: Short wavelength surface irregularities characterized by hills (asperities) and valleys. These surface features typically have random spacings, heights and depths.
- **Waviness**: Surface irregularities with a much larger spacing than the roughness features and with a small statistical variation of spacing and amplitude (i.e., a regular structure). Waviness can result from heat treatment, clamping deflections, vibrations, warping strains, and, for machined parts, from the grinding or polishing process used to finish the surface.
- **Errors of form**: Flaws that occur during manufacture or later misuse of the part; gross deviation in shape resulting from the part not being made correctly or being damaged.

While the above characterizations of topographies are common to many surfaces, others have topographies that do not fit simply in the above categories. For example, the roughness may be regular at all length scales or random at all length scales. Surfaces with fractal topographies are particularly interesting...
cases, as further magnification brings out finer scale features that have the same topographical characteristics as the larger scale features.

2.2 Roughness parameters

Profilometry and atomic force microscopy (AFM) are two common ways to measure surface topography. In these techniques, a sharp stylus is placed at the end of a flexible cantilever and then dragged over the surface with a low contact force. As the stylus traces over the surface, its up and down motion is recorded to provide a 1-dimension profile of the surface topography as illustrated in Fig. 2.2. By repeatedly collecting traces separated by small distance in the orthogonal direction, a 3-dimensional representation of the topography can be generated. These profile traces can then be analyzed with a computer to determine parameters that describe the surface roughness.

2.2.1 Variations in Z-height

For a single profile trace, a center line or mean height is first found by determining the line through the trace for which the mean-square deviation of the trace about that line is a minimum. (For this line, the area above the line is equal to the area below the line.) Then $z$ is defined as the height of the profile relative to the mean height. For 3-D collections of traces, a plane defined in the same manner is used to determine the mean height or surface mean.

The first roughness parameter, the average roughness $R_a$, is defined by

$$R_a = \frac{1}{L} \int_0^L |z| \, dx. \quad (2.1)$$

The root-mean-square deviation of $z$ about the mean height (also called the rms, standard deviation $\sigma$ or the $R_q$) is a statistically more meaningful parameter and is defined by

$$\text{rms} = \sigma = R_q = \left( \frac{1}{L} \int_0^L z^2 \, dx \right)^{1/2}. \quad (2.2)$$

![Fig. 2.2. Single profile trace in the sample x-direction (Johnson 1985).](image)
For most surfaces, $R_a$ and $R_q$ are similar in magnitude; for example (Bhushan 1999), for a sinusoidal surface,

$$R_q = \left( \frac{\pi}{2\sqrt{2}} \right) R_a = 1.11 R_a \quad (2.3)$$

while for a surface with a random Gaussian distribution of surface heights,

$$R_q = \sqrt{\frac{\pi}{2}} R_a = 1.25 R_a. \quad (2.4)$$

Several other parameters exist for further characterizing the variation of $z$-height of a line profile:

- $R_t$ – The vertical height from the lowest valley to the highest asperity. Also known as the peak-to-valley (P–V) roughness.
- $R_p$ – The distance from the mean height to the highest asperity.
- $R_z$ – The distance from the average of the five lowest points to the five highest points of the profile.

Obviously, these parameters characterize only the variation of $z$-height about the surface mean and do not provide information of how neighboring $z$’s are correlated. They incompletely characterize a surface’s topography by not providing any information about lateral variations of topography: slopes, shapes, and sizes of surface features, or their spacing or regularity. For example, Fig. 2.3 shows surface profiles that all have the same value of $R_q$, but different frequencies and shapes that result in widely different contact mechanics and contact areas in tribological situations. These roughness parameters for $z$-height variations

![Fig. 2.3. Surface profiles that all have same value of $R_q$.](image)
are mainly useful for characterizing surfaces with a particular type of roughness, such as would be made by the same manufacturing method where lateral variations scale with height variations. A common mistake is to compare the z-height roughness parameters between surfaces textured by different methods and assume that surfaces with similar $R_q$ values, for example, will have similar tribological behavior.

2.2.2 Asperity summits roughness parameters

Since contact usually occurs at asperity summits, the most tribologically relevant topography parameters are often those characterizing these peaks:

- $n_s$ – The number of the asperity summits per unit area.
- $\sigma_s$ – The standard deviation of summit heights.
- $R_s$ – The mean radius of curvature of the summits. An individual asperity summit has associated with it two radii of curvatures, $R_x$ and $R_y$, measured in two orthogonal directions. The mean radius is then defined by

$$\frac{1}{R^i} = \frac{1}{2} \left( \frac{1}{R_x} + \frac{1}{R_y} \right)$$  \hspace{1cm} (2.5)

where the $i$-subscript refers to it being for the $i$-th asperity. $R_s$ is the average over all $R_i$’s.

Obtaining these summit roughness parameters generally requires that the 3-dimensional surface topography data be collected and appropriately analyzed. We can, however, estimate some of the summit parameters from those obtained from single line profiles (Greenwood 1984):

- For an isotropic textured surface with a Gaussian distribution of summit heights, the standard deviation of summit heights can be approximated by the standard deviation of $z$-heights,

$$\sigma_s \sim \sigma.$$  \hspace{1cm} (2.6)

- For randomly rough surfaces, the mean height of the asperities above the surface mean is usually in range of 0.5 to 1.5$\sigma$.

- The mean radius of curvature $R_s$ of the summit asperities can be approximated with the root-mean-square curvature $\sigma_\kappa$ of the profilometer trace,

$$R_s \sim \sigma_\kappa,$$  \hspace{1cm} (2.7)

where $\sigma_\kappa$ is determined from the profiles of $z$-height. In the computation of $\sigma_\kappa$, the curvature $\kappa$ is determined at discrete intervals along the profile: For three consecutive heights $z_{i-1}$, $z_i$, and $z_{i+1}$ with a lateral interval spacing $h$, the curvature at interval $i$ is

$$\kappa_i = \frac{(z_{i+1} - 2z_i + z_{i-1})}{h^2}.$$  \hspace{1cm} (2.8)
Then the rms curvature is determined by
\[ \sigma^2_{\kappa} = \left( \frac{1}{n} \right) \sum_{i=1}^{n} \kappa_i^2 \]  
(2.9)
where \( n = L/h \) is the number of intervals in the profile length \( L \).

- The summit density \( n_s \) can be estimated from the number of peaks per unit length \( n_p \) along a \( z \)-height profile. For a wavy surface made of two orthogonal wave patterns, \( n_s = n_p^2 \); while, for a random isotropic surface, \( n_s \sim 1.8 n_p^2 \) (Greenwood 1984).

### 2.3 Surface height distributions

The \( z \)-height roughness parameters \( R_a \), \( R_q \), and \( \sigma \) only begin to tell how the surface heights vary from the surface mean. By plotting the **height distribution** (also called the **probability distribution** or **height histogram**), one can obtain a much more detailed picture of how the surface heights vary. The height distribution is the probability \( \phi(z) \) that the height of a particular point on the surface is between \( z \) and \( z + dz \) and is calculated by determining the fraction of the profile line (see Fig. 2.4) that is between the lines for \( z \) and \( z + dz \).

Integrating the probability distribution \( \phi(z) \) gives the cumulative probability function \( \Phi(z) \):
\[ \Phi(z) = \int_{z}^{\infty} \phi(z') \, dz' \]  
(2.10)
generating the S-shaped curve on the right side of Fig. 2.4. \( \Phi(z) \) is also called the **bearing area curve** as it expresses, as a function of \( z \), the fraction of the area underneath the surface contour and above a particular \( z \)-height (see Fig. 2.4).

For many real surfaces, the height distribution is close to a Gaussian distribution, which is given by
\[ \phi(z) = \frac{\exp(-z^2/2\sigma^2)}{\sigma \sqrt{2\pi}}. \]  
(2.11)
A useful feature of a Gaussian distribution is that the standard deviation \( \sigma \) completely describes the distribution. For a Gaussian distribution, 68.3% of the

![Fig. 2.4. Height or probability distribution \( \phi(z) \) and cumulative probability distribution \( \Phi(z) \).](image-url)
surface heights lie within one standard deviation ($\pm \sigma$) of the surface mean, 95.4% within $\pm 2\sigma$, and 99.7% within $\pm 3\sigma$.

While many surfaces can be described by Gaussian distributions, it is important to remember that most surfaces are not. Typically, a ground surface has Gaussian height distribution, but subsequent polishing of that ground surface will remove the high asperities from the height distribution. As another example, consider a surface with a regular structure, such as the machined surface illustrated in Fig. 2.1: the height distribution for this surface would be poorly described by a Gaussian distribution as the surface heights are not random, but wavy with a characteristic wavelength and amplitude generated by the machining process. At length scales much shorter than the characteristic wavelength of the waviness, the $z$-height roughness may potentially be described by a Gaussian distribution.

### 2.4 Measuring surface roughness

Due to the importance of surface topographies to many scientific and engineering fields besides tribology, a wide variety of techniques have been invented over the years for measuring surface topographies. These techniques fall into two broad categories:

1. Contact methods, where part of the measuring instrument touches the surface of interest. Examples: stylus profilometry, scanning probe microscopy.

The choice of measuring tool often depends on the length scale for which topographical information is desired. For example, if it is desired to know the surface roughness over a large area of the sample without too much precision, an optical technique is typically used. If the area of interest is less than a few millimeters in size, a simple line scan is all that is needed, and a mechanical stylus profilometer might be used. If it is desirable to know the surface topography with the highest possible resolution in a relatively small area ($<100 \mu m$ across), a scanning probe such as the scanning tunneling microscope (STM) or atomic force microscope (AFM) is used.

As we are interested in this book mainly with what happens at the smallest possible scale, we only discuss in detail how the atomic force microscope measures surface topography, as this is now the technique of choice for measuring surface topographies at the small scale.

#### 2.4.1 Atomic force microscopy (AFM)

In atomic force microscopy, the force acting on the last few atoms of a sharp tip is measured as that tip moves gently over the sample surface, as illustrated in Fig. 2.5. This is accomplished by mounting the tip at the end of a cantilever with a known spring constant; the cantilever deflects by an amount proportional to the
force. To achieve atomic resolution, it is necessary to measure deflections smaller than the size of an atom; since atoms are typically a few ångstroms in diameter, the measurement sensitivity should be better than an ångstrom. Figure 2.6 shows several methods that have been developed for measuring the small deflections of AFM cantilevers: electron tunneling (Binnig, Quate et al. 1986); optical interference (Martin, Williams et al. 1987); optical deflection (Meyer and Amer 1988; Drake, Prater et al. 1989); and capacitance (Neubauer, Cohen et al. 1990). Today many AFM designs routinely measure deflections as small as $10^{-11}$ m (0.1 Å). Currently, the most common detection method used in commercial AFMs is optical deflection [Figs 2.6(c) and 2.7].

For determining the topography of surfaces, an AFM can use either the repulsive force when the tip is in contact with the surface or the attractive force when the tip is separated by a small distance from the sample. In the repulsive mode (Binnig, Quate et al. 1986), the tip is scanned across a hard surface while touching it with constant force, so that the tip moves up and down over the surface features in much the same way as a stylus moves in the groove of an old fashioned phonograph record. In this mode, very high resolution surface topographs can be made. In a typical AFM, a tip with radius on the order of 10 nm contacts with a repulsive force of a few nanonewtons achieving a lateral resolution of a few nanometers. To achieve true atomic resolution, it is necessary to have a tip sharper than 10 nm and to touch the surface with a repulsive force less than 0.1 nN, as any greater force would likely disturb the atoms at the end of the tip or on the sample surface.
Fig. 2.6. The different methods for measuring cantilever deflection in AFMs: (a) tunneling detection (STM, scanning tunneling microscopy); (b) capacitance detection; (c) optical deflection; and (d) optical interference.

For softer, more fragile surfaces, attractive mode imaging is used. In this mode (Martin, Williams et al. 1987), the tip is held 2 to 10 nm away from the sample, so that it only senses the attractive van der Waals force. The attractive van der Waals force, however, is much weaker in magnitude than the repulsive force. A sensitive way of measuring this weak force is to vibrate the cantilever a small amount at a frequency below its natural frequency and to monitor the vibration amplitude. When an attractive force acts the tip, the cantilever’s natural resonance frequency is shifted to a lower frequency, leading to increasing vibration amplitude as the natural frequency approaches the driving frequency. A feedback loop is used to maintain a constant shift in frequency by adjusting the height of the tip above the surface, as the tip is scanned over the surface. The attractive force acting on the tip can be increased by adding an electrostatic force to the van der Waals force by applying a bias voltage between the tip and sample (Martin, Williams et al. 1987; Hu, Xiao et al. 1995; Kim, Mate et al. 1999).

A third AFM imaging method, referred to as ‘tapping mode’ or ‘intermittent contact mode,’ combines some of the features of the repulsive and attractive force imaging methods. In intermittent contact mode, the cantilever is vibrated at its
natural frequency with a relatively large oscillation amplitude, typically tens of nanometers. When the tip touches the sample briefly during an oscillation cycle, the large amplitude will pull the tip far enough away from the sample to disengage it from adhesive forces that would tend to stick the tip to the sample. The intermittent contact between the tip and sample reduces the vibration amplitude by an amount proportional to the intermittent contact force. A feedback loop is used to adjust the $z$-height of the sample so as to keep the reduction of amplitude constant as the oscillating tip scans over the surface. The intermittent contact method generates a repulsive contact force less damaging than that for a contact imaging, while achieving higher resolution than possible for attractive force imaging.

2.4.2 Example: Disk surfaces in disk drives

In modern disk drives, the disks have to be incredibly smooth so that the recording head can fly as close as possible to the recording medium on the disk surface in order to store the highest density of bits in the medium. For example, in 2004 disk drives, the mean-to-peak roughness of less than or equal to 5 nm enabled storage densities greater than 70 Gbits/in$^2$. To achieve this level smoothness, disk manufacturers have developed methods for polishing and etching nearly atomically smooth finishes onto the surfaces of disk substrates. Later, the layers for
recording medium and protected overcoat are deposited onto these substrates, and these conform closely with the substrate topography, so the final disk surface roughness remains close to the substrate roughness.

Figure 2.8 shows an intermittent contact AFM image of a small area of 65 mm diameter glass disk substrate that represents the state-of-the-art in 2004 for the smoothest achievable glass surface. Over this half-micron sized area, the peak-to-valley roughness $R_t$ is 16.8 Å, the average roughness $R_a$ is 1.2 Å, and the rms roughness $R_q$ is 1.5 Å; since the rms roughness is somewhat less than a typical diameter of an atom, we can say that these surfaces are nearly atomically smooth. On the submicron scale of the AFM image, only an isotropic micro-roughness is apparent with no waviness. (These substrates do, however, have a long wavelength (> 5 µm) waviness, one to two nanometers in amplitude, that becomes apparent in AFM images over much larger areas.) The bottom of Fig. 2.8 shows that the $z$-height distribution for the AFM image is well fitted by a Gaussian distribution.

While very smooth disk surfaces are desirable so that the recording head can fly as close as possible, most disk substrates are actually roughened up a bit before the recording medium is deposited, to help reduce the real area of contact and the friction force in case the recording head accidentally touches the disk surface. Figure 2.9 shows an AFM image of a substrate that has been roughened by polishing with an abrasive in such a manner that small grooves, a few nanometers deep, are made on the substrate surface, oriented around the circumferential direction of the disk surface. The short wavelength waviness oriented in this direction also helps improves the recording performance as well as the tribology: When the magnetic layer for the longitudinal recording medium is sputter-deposited onto the substrate, the circumferential texture helps the magnetic grains grow with the right crystal orientation to obtain the best recording performance. Over the small area of the AFM image, the roughness parameters are, of course, higher than for the smooth substrate in Fig. 2.8. (Peak-to-valley roughness $R_t$ is 32.3 Å, the average roughness $R_a$ is 4.8 Å, and the rms roughness $R_q$ is 6.5 Å.) The bottom of Fig. 2.9 shows the distribution of $z$-heights for the AFM image of the textured substrate. Due to the polishing process removing the high points and creating grooves in the surface, this height distribution is not well fitted by a Gaussian distribution, but instead is significantly skewed with a broad tail toward negative $z$-heights. Skewing the distribution in the negative $z$-height direction helps to increase the roughness parameters like $R_t$, $R_a$, and $R_q$, which describe variation of $z$-heights about the mean and most strongly influence the tribology of head–disk contacts, without substantially increasing the mean-to-peak roughness $R_p$, which impacts the magnetic spacing between the recording head and medium.

Even though it has been deliberately roughened up, the disk substrate in Fig. 2.9 is still significantly smoother than disks used in drives only a few years ago. Figure 2.10 shows an example of a disk surface from around the year 2000, which has an isotropic rather than a circumferential textured surface.
Fig. 2.8. (Top) Intermittent contact AFM image of a 500 nm × 500 nm area of a glass disk substrate that is typical for the smoothest that can currently be achieved. \( R_t = 16.8 \, \text{Å}, \) \( R_a = 1.2 \, \text{Å}, \) and \( R_q = 1.5 \, \text{Å}. \) (Bottom) Bars show the distribution of \( z \)-heights from the AFM image, while the curve shows a Gaussian distribution with the same standard deviation. (AFM image courtesy of T.-W. Wu and C. M. Mate, Hitachi Global Storage Technologies.)
FIG. 2.9. (Top) Intermittent contact AFM image of a 500 nm × 500 nm area of a glass disk substrate, typical of that used in 2004 for manufacturing disks for disk drive applications. $R_t = 32.3 \, \text{Å}$, $R_a = 4.8 \, \text{Å}$, and $R_q = 6.5 \, \text{Å}$. (Bottom) Bars show the distribution of $z$-heights for the AFM image, while the curve shows a Gaussian distribution with the same standard deviation. (AFM image courtesy of T.-W. Wu and C. M. Mate, Hitachi Global Storage Technologies.)
For this earlier disk, the z-height roughness parameters ($R_t = 67.7 \, \text{Å}$, $R_a = 9.8 \, \text{Å}$, $R_q = 11.7 \, \text{Å}$) are considerably higher than those for the textured disk in Fig. 2.9, but still remarkably small compared to most engineered surfaces.

For the AFM image in Fig. 2.10, identifying the individual asperity peaks is much easier than for the AFM images in Figs 2.8 and 2.9. With suitable software, the heights and mean radius of curvature of these asperity summits can be measured and from these values, the asperity roughness parameters can be determined for this disk surface:

- $n_s$ (the number of the asperity summits per unit area) = 226/µm²
- $\sigma_s$ (the standard deviation of summit heights) = 9.0 Å
- $R_s$ (the mean radius of curvature of the summits) = 80 nm

### 2.5 References


When two materials are pressed against each other, initial solid–solid contact occurs at the surfaces’ high points. This area where these asperities touch is usually an extremely small fraction (<1%) of the total area covered by the surfaces, but the forces generated between the contacting atoms in this small area are responsible for most tribological phenomena—friction, wear, adhesion, etc. Consequently, understanding how forces acting on asperities distort the material around the points of contact provides an important basis for understanding tribology.

In this chapter, we study how the surface topography changes when two objects are brought into contact. In the previous chapter we learned how to characterize the topography of a non-contacting surface, but characterizing its topography when in contact is extremely difficult as these surfaces are inaccessible to most experimental techniques. Usually, we have to analyze the contact mechanics of the hidden contacting interface if we want to know the contact stresses, the extent of elastic and plastic deformation, and the real area of contact. With contact mechanics, we combine our knowledge of surface geometry (gained in the previous chapter) with our understanding of the mechanical properties of materials (to be gained in this chapter) to determine the stresses and deformations within the contacting materials.

3.1 Atomic origins of deformation

The mechanical properties of a material originate from how the individual atoms are bound to their neighboring atoms in the material. When two similar atoms—say of aluminum—are brought together, a bond will form between them. Figure 3.1 plots the forces bonding these two atoms together: an attractive force, which gradually decreases in magnitude over a range of 4 to 5 atomic diameters as the atoms move apart; a repulsive force, which increases rapidly as the atoms are squeezed together due to the overlap of the electron clouds that surround each atom; and the resultant force, which is the sum of the attractive and repulsive force. The attractive force can have several different origins: metallic bonding that occurs in metals, ionic bonding that occurs between the positive and negative charged ions like in crystals of table salt, and van der Waals bonding such as occurs between the organic compounds that might make up a piece of
The dashed lines show the attractive force that dominates at large distances and the repulsive force that dominates at short distances. The solid line shows the resultant sum of the attractive and repulsive forces. $OB$ is the separation distance at equilibrium when the resultant force equals zero. Pulling on the atoms with a force $f''$ increases the separation to $OB''$, while pushing the atoms together with a force $f'$ decreases the separation to $OB'$.

When the attractive and repulsive forces are balanced, the resultant force acting between the two atoms is zero, meaning the atoms are at their equilibrium separation distance (point $B$ in Fig. 3.1). Within a solid, the atoms have found positions where the sum of all the forces acting between them and the neighboring atoms is zero. To achieve this balance, the atoms frequently order in a lattice that minimizes the interaction energy between them, as illustrated in Fig. 3.2(a), with the atoms spaced at their equilibrium separation distance. If a tensile force that is not too large is applied to the crystal, as shown in Fig. 3.2(b), the bonds between the atoms stretch like little springs, increasing the spacing between the atoms so that the resultant force between each atomic pair becomes slightly negative (point $B''$ in Fig. 3.1). If a compressive force is applied, the spacing between the atoms decreases, and the resultant force becomes slightly positive (point $B'$ in Fig. 3.1). The solid deforms until the added attractive (repulsive) forces between the atoms compensate the externally applied tensile (compressive) forces.
For small applied forces, the atoms only move a small fraction of the bond length away from the equilibrium separation distance, and the restoring force is proportional to the distance moved. This is an elastic deformation. For elastic deformations, the atoms return their original positions if the applied force is removed. If a large enough tensile force is applied to the solid that the force on individual atoms exceeds the value of point C in Fig. 3.1, the attractive restoring force will be insufficient to hold the material together. Under these conditions, plastic deformation takes place, and many atoms are no longer at their original positions when the applied force is removed.

Plastic deformation can take place several ways: For brittle materials, the atoms do not slide easily past each other. So, when an externally applied tensile force. Similarly, if a shear force is applied to the solid as shown in Fig. 3.2(c), the atoms move slightly sideways until atomic level forces balance the applied shear force.

Fig. 3.2. Atoms in a crystal structure subjected to (a) no applied force, (b) to a small tensile force, (c) to a small shear force, and (d) to a tensile force large enough to cause slippage along a crystal plane.
force exceeds the attractive force holding the atoms together in the solid, the material just snaps. For ductile materials, atoms slide one atomic position along a slip plane as illustrated in Fig. 3.2(d). This slip motion repeats many times along a large number of slip planes and grain boundaries in the solid until all the forces on individual atoms are below the threshold value $C$ needed to initiate plastic deformation. Figure 3.3 illustrates how slippage at a series of crystal planes in a single crystal can lead to plastic elongation of the entire crystal.

The sliding of one plane of atoms all in unison over another plane of atoms is extremely difficult as the force required is the sum of the force that needs to be exerted on each atom in the plane to slide it over the activation barrier between adjacent lattice sites. The force for slippage is much less if the motion occurs only for a small number of atoms around a dislocation rather than all the atoms in a crystal plane. Consequently, most ductile plastic deformation processes occur via the motion of dislocations (Haasen 1996). Figure 3.4 illustrates how one type of dislocation, an edge dislocation, causes motion along a slip plane: A shear stress is applied that pushes the upper half of the crystal to the left and lower half to right. The compression on the upper half lattice causes several planes of atoms to move one lattice spacing to the left forming an extra plane (marked as hatched circles) dislocation within the crystal. The shear stress causes this extra plane to move through the crystal one lattice spacing at a time until it emerges at the left. The force to move this dislocation through the crystal is much smaller than trying to slide all the atoms at once through one lattice spacing, as only a small number of atoms associated with the dislocation are moved from their low energy equilibrium positions in the crystal lattice.

From these atomic origins of elastic and plastic deformations, we are now ready to discuss the some the basic relations of contact mechanics.
3.2 Elastic deformation

3.2.1 Basic relations

When a solid is subjected to a load $L$, like the one applied along the axis of the cylinder in Fig. 3.5(a), a stress is created in the solid that produces an elastic deformation of the object this is characterized by a strain:

\[ \sigma_1 = \text{stress in axial direction} = \frac{L}{\text{cross-sectional area}}, \quad (3.1) \]

\[ \varepsilon_1 = \text{strain in axial direction} = \frac{\text{change in length}}{\text{original length}}. \quad (3.2) \]

If the strain is small, so that the atoms move only a small fraction of their equilibrium separation distance, the strain is proportional to the applied stress:

\[ \varepsilon_1 = \frac{\sigma_1}{E} \]

\[ \sigma_1 = E\varepsilon_1 \quad (3.3) \]

where $E$ is the elastic constant or Young’s modulus of the material. A strain is considered elastic if it is reversible, i.e., the solid returns to its original shape when the stress is removed. Being reversible also means that all the work expended in elastically deforming the material can be recovered when the stress is removed.
When you squeeze an object (like the cylinder in Fig. 3.5(a)) in one direction, some of the material moves in the direction orthogonal to the applied stress. The elastic strain $\varepsilon_2$ in the transverse direction is proportional to the elastic strain $\varepsilon_1$ in the direction of applied stress:

$$\varepsilon_2 = -\nu \varepsilon_1$$

(3.4)

where $\nu$ is the Poisson ratio, which ranges from 0.10 to 0.5, but for most metals is between 0.25 and 0.30. (Cork is a somewhat exceptional material in that $\nu \sim 0$, so, when you apply an axial load to push a cork into a wine bottle, it doesn’t expand in the transverse direction to resist sliding into the bottle.)

When a solid is subjected to a shear stress as shown in Fig. 3.5(b), the shear strain $\gamma$ is proportional to the shear stress $\tau$:

$$\tau = G \gamma$$

(3.5)

where $G$ is the shear modulus. The shear modulus is related to the Young’s modulus for isotropic materials by

$$E = 2G(1 + \nu).$$

(3.6)

### 3.2.2 Elastic deformation of a single asperity

Next we will consider the contact mechanics of a single asperity being pushed against with a flat surface with some load $L$. Later we use these results for single asperity contact to derive the contact phenomena for multiple asperity contacts. Rather than considering all the possible shapes for the asperity assumption, we mainly consider the simplest possible shape.

#### 3.2.2.1 Approximating a single asperity contact

Physicists often tell the following joke about how they like to approximate many physical situations: The story is that a dairy farm wanted to modernize its operations with the latest scientific ideas, and someone suggested that the best idea
would be to hire the smartest science professor at the local university as a consultant. So inquiries were made and eventually a consulting contract was signed with a brilliant young physics professor. On the first day of the contract, the professor arrives at the farm where the dairymen are eagerly waiting to hear the latest scientific ideas that they can apply to their dairy farm. The professor proceeds to tell them all about his latest superstring theory for unifying the forces of nature. The farmers quickly become exasperated and say “Look we’re a dairy farm, we want to know about cows!” The physics professor, however, isn’t the least bit perturbed by this outburst; he says simply “To apply my theory, assume that a cow can be approximated as a sphere.”

In our analyses of tribological contacts, we often use the crude approximation that the contact geometry for single asperities can be modeled simply as a sphere on a flat. Our purpose for using this spherical approximation to is mainly to gain insights into the nature of a single asperity contact. For contacts with a non-spherical geometry one can either refer to an analytical treatment in a suitable text (Johnson 1985) or use an appropriate numerical method to analyze the contact mechanics. Usually, however, the “sphere on a flat” approximation provides reasonably accurate predictions of the quantities of most physical interest, like contact areas and stresses. (It certainly works better than the spherical cow approximation.)

3.2.2.2 Elastic contact area for a sphere on a flat

During his 1880 Christmas vacation, a 23-year-old Heinrich Hertz worked out an analysis of elastic deformation for an elliptical contact area (Hertz 1882; Johnson 1985). Frequently, we simplify the analysis of a contact mechanics problem by transforming it into an equivalent sphere on flat geometry.

When a spherically shaped summit is brought into contact with a flat surface with a load $L$ as shown in Fig. 3.6, the surfaces deform to create contact zone of radius $a$. According to Hertz’s equations for elastic deformation for a sphere on a flat, the radius of the contact zone is given by

$$a = \left(\frac{3RL}{4E_c}\right)^{1/3}$$

(3.7)

where $E_c$ is the composite elastic modulus of two contacting materials, with moduli $E_1$ and $E_2$, given by

$$\frac{1}{E_c} = \frac{1 - \nu_1^2}{E_1} + \frac{1 - \nu_2^2}{E_2}.$$ 

(3.8)

For this geometry, the real area of contact $A$ is given by

$$A = \pi a^2 = \pi \left(\frac{3RL}{4E_c}\right)^{2/3}$$

(3.9)
Contact pressure (normal stress)

\[ p_m = \frac{L}{A} = \frac{1}{\pi} \left( \frac{4E_c}{3R} \right)^{2/3} L^{1/3} \]  
(3.10)

The pressure is not uniform over the circular contact area, but has a maximum at the center and falls to zero at the edge of the contact zone as illustrated in Fig. 3.6(c). For the Hertz pressure distribution

\[ p = p_o \left\{ 1 - \left( \frac{r}{a} \right)^2 \right\}^{1/2} \]  
(3.11)

the maximum \( p_o \) is \( 3/2 \) times the mean pressure \( p_m \). This pressure distribution elastically deforms the two bodies in the normal direction by the amount

\[ \Delta z = \frac{a^2}{R} = \left( \frac{9L^2}{16RE_c^2} \right)^{1/3} \]  
(3.12)

3.2.2.2.1 Example: Spherical steel particle sandwiched between two flat surfaces  
As an example, let’s consider the situation illustrated in Fig. 3.7 where a small stainless steel particle (radius \( R = 1 \mu m \)) is sandwiched between two ideally smooth surfaces, also made of stainless steel. This situation is a simple model for a small wear or debris particle stuck between two contacting surfaces. For conditions where elastic contact occurs, we would like to determine the contact radius \( a \), the mean pressure \( P_m \), and the magnitude of the compression \( \Delta z \)
where the particle touches the two flats. For 304 stainless steel, elastic modulus \( E = 200 \) GPa, Poisson ratio \( \nu = 0.3 \), and hardness \( H = 1.5 \) GPa; so, the composite elastic modulus is

\[
\frac{1}{E_c} = \frac{1 - \nu_1^2}{E_1} + \frac{1 - \nu_2^2}{E_2} = 2 \left( \frac{1 - 0.3^2}{200 \text{ GPa}} \right),
\]

\[ E_c = 110 \text{ GPa} \]

Where the particle touches the flat surface, the contact radius \( a \), the mean pressure \( P_m \), and the magnitude of the compression \( \Delta z \) are given by

\[ a = \left( \frac{3RL}{4E_c} \right)^{1/3} = \left( \frac{3 \times 1 \mu m}{4 \times 110 \text{ GPa}} \right)^{1/3} L^{1/3} \]

\[ P_m = \frac{1}{\pi} \left( \frac{4E_c}{3R} \right)^{2/3} L^{1/3} = \frac{1}{\pi} \left( \frac{4 \times 110 \text{ GPa}}{3 \times 1 \mu m} \right)^{2/3} L^{1/3} \]

\[ \Delta z = \left( \frac{9L^2}{16RE_c^2} \right)^{1/3} = \left( \frac{9}{16 \times 1 \mu m \times (110 \text{ GPa})^2} \right)^{1/3} L^{2/3}. \]

Let’s consider the range of loads from 1 nN to 1 µN (1 nN corresponds roughly to the strength of a single atomic bond, while 1 µN is just over the threshold for plastic deformation in this contact geometry):

<table>
<thead>
<tr>
<th>load ( L ) (nN)</th>
<th>contact radius ( a ) (nm)</th>
<th>mean pressure ( P_m ) (GPa)</th>
<th>vertical compression per contact ( \Delta z ) (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.9</td>
<td>0.089</td>
<td>0.0036</td>
</tr>
<tr>
<td>10</td>
<td>4.1</td>
<td>0.19</td>
<td>0.017</td>
</tr>
<tr>
<td>100</td>
<td>8.8</td>
<td>0.41</td>
<td>0.077</td>
</tr>
<tr>
<td>1000</td>
<td>19</td>
<td>0.89</td>
<td>0.36</td>
</tr>
</tbody>
</table>

From these results, we notice that for this small particle undergoing elastic deformation, the contact zone is always small, usually only a few nanometers across. The elastic compression of the particle is also small, no more than a
few ångstroms. Since the particle radius is relatively small, the mean contact pressure is fairly high, and, for the largest load $1 \mu \text{N}$, pressure is comparable to the 1.5 GPa hardness value of steel, suggesting that, if we push much harder, the deformation will mainly be plastic rather than elastic and we will need a different analysis of the contact mechanics.

### 3.3 Plastic deformation

#### 3.3.1 Basic relations

When a piece of ductile material, such as the cylinder shown in Fig. 3.8, is subjected to tensile stress high enough such that the applied force on the individual atoms exceeds the attractive restoring force, the atoms start to slide past one another along slip planes as illustrated in Figs. 3.2(d), 3.3, and 3.4. These regions of the material undergo plastic deformation. The tensile (or compressive) force per unit area needed to initiate plastic deformation is called the yield stress $Y$.

One characteristic of plastic deformation is that the process is irreversible; i.e., when the stress is removed, atoms that underwent plastic deformation are quite happy to stay in their new locations, so the cylinder on the right side of Fig. 3.8 remains permanently elongated.

For a spherically shaped indenter contacting a flat, several other stresses act on the materials in addition to the normal stress plotted in Fig. 3.6(c): a stress radial to the axis of symmetry and a shear stress that is $0.5$ (normal stress − radial stress). As plastic deformation typically involves slippage of one atomic plane

---

**Fig. 3.8.** Left: A cylinder subjected to a tensile load too weak to cause plastic deformation. Right: A cylinder of ductile material, such as a metal, subjected to a tensile load large enough to cause plastic deformation.
over another, the magnitude of the shear stress determines whether plastic deformation occurs. If $\nu = 0.3$, the maximum shear stress occurs at a depth of $0.48a$ below the contact plane, with a value of $0.31p_o$ or $0.47p_m$ (Johnson 1985). Using the Tresca criterion (Hill 1950), plastic deformation is initiated at this depth when the maximum shear stress exceeds $Y/2$, which for the sphere on flat geometry occurs when $p_o = 1.6Y$ or $p_m = 1.1Y$. Using eq. (3.10), the load required to initiate plastic deformation for a sphere of flat geometry is

$$L_Y = \frac{1}{6} \left( \frac{R}{E_c} \right)^2 (1.6\pi Y)^3. \quad (3.17)$$

3.3.2 Hardness

When a spherical indenter is pushed against a flat so as to cause plastic deformation, the plastic zone is at first small [Fig. 3.9(a)], and the elastic deformation of the surrounding material works to contain the plastic deformation. As the load increases, the plastic zone expands to accommodate the increasing contact pressure. Eventually, it breaks out to the free surface, and material plastically flows around the sides of the indenter as illustrated in Fig. 3.9(b). The load where plastic break out occurs in metals is typically 50 to 100 times greater than the load needed to initiate plastic flow below the surface (Johnson 1985). By the time full plasticity occurs, the mean contact pressure has risen to $p_m \simeq 3Y$. Once material starts to flow plastically around the indenter, the mean contact pressure remains fairly constant with increasing load; the area of plastic deformation simply increases proportionally to load. For this situation, we can define the hardness $H = \text{load/area of permanent indentation}$. So, for the sphere on flat

![Fig. 3.9](image)

Fig. 3.9. Plastic deformation of a hard sphere indenting a flat surface of a softer material. (a) At the onset of plastic deformation, the plastic zone only occupies a small volume at the region of maximum pressure. $a_Y = 2.5YR/E_c$ is the contact radius when plastic deformation initiates. (b) At higher loads, the plastic volume flows around the sides of the indenter.
geometry where the material is fully plastic under the indenter [Fig. 3.9(b)],

\[ H \simeq 3Y. \]  \hspace{1cm} (3.18)

For other non-spherical–flat contact geometries—such as for conical, pyramid, and flat end-shaped indenters—the ratio of indentation pressure and yield stress \( Y \) is also constant with the proportionality constant always close to 3, so eq. (3.18) still applies (Johnson 1985). The measured hardness does not depend so much on the shape of the apex of the indenter, rather on the ability of the material to plastically flow around the sides of the indenter.

3.4 Real area of contact

Many tribological properties, such as friction and wear, tend to be proportional to the real area of contact. Since contacting surfaces touch only at the top of their asperities, the real area of contact is typically a small fraction of the apparent area of contact. As mentioned before, measuring the real contact area of the hidden contacting interface is usually extremely difficult. Instead, one typically has to resort to estimating the real area of contact by analyzing the contact mechanics of the contacting surfaces. As inputs to this analysis, one needs:

1. **Surface roughness.** As described in the previous chapter, an AFM or 3-D optical profilometer can be used to determine the surface topographies of the contacting surfaces and their roughness parameters. For rough surfaces, usually the most useful roughness parameters for a contact mechanics analysis are the mean radius of curvature \( R \) of the asperity summits and the standard deviation of summit heights \( \sigma \). These values are then used to approximate the contact geometry by estimating the individual single asperities with spherical asperities all having the same radius of curvature \( R \).

2. **Deformation constants.** As discussed in the previous section, an indenter can be used to determine the micro-hardness \( H \) of the near surface region of materials. The same indentation experiment can also be used to determine the Young’s modulus \( E \) of the material (Doerner and Nix 1986; Oliver and Pharr 1992; Pharr, Oliver et al. 1992; Oliver and Pharr 2004).

3. **Contact forces.** Contact forces are made up of the external forces applied normally to the interface (the load \( L \)) and tangential to the interface (the friction \( F \)), along with the internal adhesive forces (\( L_{\text{adh}} \)).

With these inputs, a computer running a finite element analysis code can be used to obtain a reasonably accurate estimate for the real area of contact for a given contact situation. While a finite element analysis can be fairly accurate, it is often difficult to see from such an analysis how the area of contact originates from the physical properties of the contacting interfaces. To better comprehend the interplay between the various roughness and deformation parameters, we next discuss some simple models for estimating the real area of contact.
Fig. 3.10. Contact geometry assumed in the Greenwood and Williamson model.

3.4.1 Greenwood and Williamson model

In 1966 Greenwood and Williamson (Greenwood and Williamson 1966) developed a model for elastic deformation of multi-point contact that illustrates how contact area depends on the roughness parameters of summit curvature and distribution of summit heights together with the elastic modulus. In this model, Greenwood and Williamson extended the Hertzian theory of elastic contact for a sphere on a flat to the situation illustrated in Fig. 3.10, with the following assumptions about the nature of the contacting interface:

1. A non-deformable flat surface contacts an elastically deformable rough surface.
2. The rough surface is covered with spherically shaped asperities, all with the same radius of curvature $R$.
3. The asperity heights vary randomly with either a Gaussian or exponential distribution of summit heights.

In the model, each asperity in contact (asperities with $z > d$, where $z$ is the height of asperity from the surface mean when not in contact and $d$ is the distance between the flat surface and the mean height of the rough surface) deforms elastically without interacting with its neighbors and supports a load (from eq. (3.12))

$$L_{\text{asperity}} = \frac{4}{3} E_c R^{1/2} (z - d)^{3/2}. \quad (3.19)$$

The total load supported is given by the integrating over the distribution of asperity heights $\phi(z)$ from $z = d$ to $\infty$:

$$L = \frac{4}{3} N E_c R^{1/2} \int_{d}^{\infty} (z - d)^{3/2} \phi(z) dz \quad (3.20)$$

where $N$ is the total of asperities.

As discussed in Chapter 2, the assumption of a Gaussian height distribution of spherically shaped summits best describes surfaces with an isotropic texture, such as produced by a grinding or random surface finishing process. Using a Gaussian distribution of summit heights, Greenwood and Williamson numerically calculated the real area contact for contact geometry in Fig. 3.10.
Using an exponential distribution of summit heights, they could derive analytical expressions for contact area and other contact parameters. We will discuss these analytical expressions in some detail in order to illustrate how the physical parameters influence contact.

An exponential distribution of summit heights can be expressed as

$$\phi(z) = \frac{C}{\sigma} \exp\left(-\frac{z}{\sigma}\right)$$

where $z$ is the summit height relative to the surface mean, $\sigma$ the standard deviation of summit heights, and $C$ a constant. The approximation of an exponential distribution is most likely to be valid if contact is limited to the highest tenth of asperities. For elastic contact between a flat surface and an exponential distribution of summit heights, Greenwood and Williamson showed that (Greenwood and Williamson 1966):

- The number of asperities in contact is

$$n = \frac{L}{\pi^{1/2} E_c \sigma^{3/2} R^{1/2}}$$

- The real area of contact is

$$A_r = \left(\frac{\pi R}{\sigma}\right)^{1/2} \frac{L}{E_c}$$

- The mean area per asperity is

$$\frac{A_r}{n} = \pi R \sigma$$

- The mean contact radius is

$$a = \sqrt{R \sigma}$$

- The mean contact pressure is

$$p_m = \frac{L}{A_r} = 0.56 E_c \left(\frac{\sigma}{R}\right)^{1/2}$$

From these analytical expressions, we learn some important aspects for the contact situation in Fig. 3.10:

- The number of asperities and the real area of contact are proportional to load.
- The mean size of contact area per asperity is independent of load.
- The mean contact pressure is independent of load.

In this analysis, as the load increases, a sufficient number of new asperities are brought into contact so that these mean values of contact area, radius and pressure stay constant.
Similar results are also obtained if a Gaussian distribution is used for asperity heights. For example, with a Gaussian distribution, the Greenwood and Williamson model determines the mean contact pressure to be

$$p_m = [0.3 \text{ to } 0.4] \times E_c \left(\frac{\sigma}{R}\right)^{1/2}.$$  

(3.27)

This expression has the same dependence on $E_c$, $\sigma$, and $R$ as that for an exponential distribution of heights, but the factor $[0.3–0.4]$ now depends on the degree of loading and is slightly less than the 0.56 factor in eq. (3.26).

3.4.1.1 Example: TiN contacts

To see how the Greenwood and Williamson model works, we will use eqs. (3.22)–(3.26) to determine the contact parameters $n$, $A_r$, $a$, and $P_m$ for a rough TiN surface contacting a flat TiN surface with a load $L=1 \text{ N}$. (Titanium nitride coatings are commonly deposited onto metals and ceramics by plasma vapor deposition in situations where a hard surface is desired.) The material parameters for TiN are: $E = 600 \text{ GPa}$, $\nu = 0.2$, $E_c = 313 \text{ GPa}$, and $H = 17 \text{ GPa}$ (Bhushan 1999). We will assume the following roughness parameters for the rough surface: standard deviation of summit heights $\sigma = 20 \text{ nm}$ and mean radius of curvature of summits $R = 10 \text{ \mu m}$.

$$n = \frac{L}{\pi^{1/2} E_c \sigma^{3/2} R^{1/2}} = \frac{1 \text{ N}}{\pi^{1/2} \times 313 \text{ GPa} \times (20 \text{ nm})^{3/2} (10 \text{ \mu m})^{1/2}} = 202$$  

(3.28)

$$A_r = \left(\frac{\pi R}{\sigma}\right)^{1/2} \frac{L}{E_c} = \left(\frac{\pi \times 10 \text{ \mu m}}{20 \text{ nm}}\right)^{1/2} \frac{1 \text{ N}}{313 \text{ GPa}} = 127 \text{ \mu m}^2$$  

(3.29)

$$a = \sqrt{R\sigma} = \sqrt{10 \text{ \mu m} \times 20 \text{ nm}} = 0.45 \text{ \mu m}$$  

(3.30)

$$p_m = 0.56E_c \left(\frac{\sigma}{R}\right)^{1/2} = 0.56 \times 313 \text{ GPa} \times \left(\frac{20 \text{ nm}}{10 \text{ \mu m}}\right)^{1/2} = 7.8 \text{ GPa}$$  

(3.31)

So, the Greenwood and Williamson model predicts that, even with the fairly substantial load of 1 N, only a relatively small number of asperities (202) will be in contact and that the average contact radius ($a=0.45 \text{ \mu m}$) is a very small fraction of the mean curvature of the asperities; thus to the area of solid–solid contact is a small fraction of the apparent contact area. This turns out to be a fairly general result, that, for any contacting materials with moderate to high hardness and elastic modulus, the number of asperities in elastic contact will be fairly small and with fairly small contact zones. Since the load is concentrated on these small contact zones, the contact pressures are fairly high and will induce plastic deformation if rubbed against a softer material. (Indeed, TiN is frequently used to coat cutting tool surfaces, so that the plastic deformation, which is the
first step in the cutting process, is more likely to happen in the material being cut than in the material doing the cutting.)

3.4.1.2 Real area of contact using the Greenwood and Williamson model

In order to use the appropriate analysis for determining the real area of contact, one should first determine whether the deformation at the contacting asperities is predominantly elastic or plastic. To achieve this, the Greenwood and Williamson model can be used to predict whether the average contact pressure is less than hardness (indicating mostly elastic deformations) or greater than hardness (indicating mostly plastic deformation). For this purpose, Greenwood and Williamson used their model to define a plasticity index, denoted $\psi$:

$$
\psi = \left( \frac{\sigma}{R} \right)^{1/2} \left( \frac{E_c}{H} \right)
$$

(3.32)

for characterizing contacts as elastic or plastic. If $\psi \geq 1$, the Greenwood and Williamson model predicts that the average contact pressure exceeds the hardness and the deformations are predominately plastic. If $\psi < 0.6$, the model predicts that the contact pressure is below the threshold for plasticity and the contacts deform elastically. From the definition of the plasticity index, we see that plastic deformation is more likely either when the material becomes softer (the hardness $H$ decreases relative to composite elastic modulus $E_c$) or when the surface becomes rougher ($\sigma$ increases relative to $R$).

**Plastic deformation.**

Under conditions for plastic deformation, the contacting asperities of the softer material will plastically flow so as to increase the contact area supporting the load until the mean contact pressure equals the hardness. So the contact geometry stabilizes when

$$
p_{\text{mean}} = \frac{L}{A_r} = H
$$

(3.33)

which can be rearranged to give an expression for the real area of contact $A_r$: in the situations when plastic deformation dominates:

$$
A_r \simeq \frac{L}{H}.
$$

(3.34)

Note that no assumptions were made about the contact geometry in deriving this expression other than the condition of plasticity, so it is a fairly general expression for real contact area for plastic contacts.

**Elastic deformation.**

Now for contacting situations that are primarily elastic, the simplest way to estimate the real area of contact is to use the Greenwood and Williamson model
with a Gaussian distribution of summit heights to obtain the following expression from eq. (3.27)

\[ A_r \simeq 3 \left( \frac{R}{\sigma} \right)^{1/2} \frac{L}{E_c} \quad (3.35) \]

3.4.1.2.1 Example: Recording head on a laser textured disk surface  As mentioned in Sections 1.2, 1.3 and 2.4.2, disk surfaces inside disk drives are made very smooth so that the recording head can fly within a few tens of nanometers of the disk surface in order to achieve good signal-to-noise during the read/write operations of the disk drive. The disk surface, however, is not made as smooth as possible, but rather a slight roughness is deliberately added to reduce the real area of contact between the recording head and disk when they occasionally touch. This is particularly true on that part of the disk surface next to the disk hub, referred to as the “landing zone,” where the recording head slider comes to rest when the disk drive is powered off (Johnson, Mate et al. 1996). A major design concern is minimizing the friction force between the slider and disk that needs to be overcome to start the disk spinning again when the drive is powered back on. For a given loading force on the slider, the amount of roughness controls the friction force, which is proportional to real area of contact. To ensure that the startup friction force is not large enough to prevent the disk from spinning, the landing zone has a texture that is rougher than the rest of the disk surface where the data is stored. A common way to texture the disk landing zone is to use a pulsed laser to melt the disk substrate in localized spots that resolidify to form small bumps 5 to 40 nm in height at well defined spacings on the disk surface (Baumgart, Krajnovich et al. 1995; Johnson, Mate et al. 1996; Tam, Pour et al. 1996). By controlling the energy and time duration of the laser pulse, the height and shape of the bumps can be accurately controlled along with the spacing between bumps; hence, laser texturing allows for the control of the contact area between the head and the disk.

The schematic in Fig. 3.11(a) shows a recording head slider resting on a laser textured landing zone. To a large extent this situation closely matches the idealized case of the Greenwood and Williamson model: An ideally smooth surface (the slider surface) contacts a rough surface (the laser textured disk surface) where the asperities have spherically shaped summits [or can be approximated as spherical (Gui, Kuo et al. 1997)] with nearly identical radii of curvatures and with a Gaussian distribution of summit heights.

For this example, we want to determine the real area of contact for the typical conditions for a slider parked on a laser texture zone: slider loading force \( L = 20 \text{ mN} \) and laser textured bumps with spherical summits \( R = 200 \mu \text{m} \), with a standard deviation of bump heights \( \sigma = 1.5 \text{ nm} \), and with a density of 2000 bumps per \( \text{mm}^2 \).

First, we need to determine the composite elastic modulus \( E_c \) from the elastic modulus of the disk surface (\( \sim 130 \text{ GPa} \)) and from the elastic modulus of the
Al₂O₃–TiC sintered material that makes up the slider body (∼450 GPa) and assuming a Poisson ratio \( \nu = 0.25 \):

\[
\frac{1}{E_c} = \frac{1 - \nu_1^2}{E_1} + \frac{1 - \nu_2^2}{E_2} = \frac{1 - 0.25^2}{130 \text{ GPa}} + \frac{1 - 0.25^2}{450 \text{ GPa}}.
\]

\[ E_c = 108 \text{ GPa} \]  

Next we need the hardness of the material where plastic deformation is most likely to take place. Since the disk has a layered structure near its surface [Fig. 3.11(b)], we need to determine the typical depth where the shear stress responsible for initiating plastic deformation reaches a maximum (Section 3.3.1); for each asperity contact zone, this maximum occurs at a distance approximately 0.5\( a \) away from the contact interface. Using eq. (3.25) to determine the mean contact radius \( a \), the mean distance \( d \) from the contacting interface where plastic deformation initiates is

\[ d \approx 0.5\sqrt{R\sigma} = 0.5\sqrt{(200 \mu\text{m})(1.5 \text{nm})} = 0.27 \mu\text{m}. \]  

For the disk, this depth is below the magnetic and chromium layers and is in the NiP overlayer that coats the AlMg substrates typically used for laser textured disks [Fig. 3.11(b)]. The NiP layer has a hardness \( H \sim 7 \text{ GPa} \), which is much less than the hardness of the slider material \( H \sim 22 \text{ GPa} \) for sintered Al₂O₃–TiC).
Now, we are ready to determine the plasticity index $\psi$ from eq. (3.32):

$$\psi = \left(\frac{\sigma}{R}\right)^{1/2} \left(\frac{E_c}{H}\right) = \left(\frac{1.5 \text{ nm}}{200 \mu m}\right)^{1/2} \left(\frac{108 \text{ GPa}}{7 \text{ GPa}}\right) = 0.08. \quad (3.41)$$

Since $\psi < 0.6$, the deformation is primarily elastic, and we need to use eq. (3.35) for determining the real area of contact:

$$A_r \simeq 3 \left(\frac{R}{\sigma}\right)^{1/2} \frac{L}{E_c} = 3 \left(\frac{200 \mu m}{1.5 \text{ nm}}\right)^{1/2} \left(\frac{20 \text{ mN}}{108 \text{ GPa}}\right) = 203 \mu m^2. \quad (3.42)$$

So, the actual contact area is a small fraction the apparent area of $1 \text{ mm}^2$ for a typical recording head slider sitting on a disk.

Equation (3.22) of the Greenwood and Williamson model also can be used to estimate the number of laser texture bumps in contact with the slider:

$$n = \frac{L}{\pi^{1/2} E_c \sigma^{3/2} R^{1/2}} = \frac{20 \text{ mN}}{\pi^{1/2} \times 108 \text{ GPa} \times (1.5 \text{ nm})^{3/2} (200 \mu m)^{1/2}} = 127. \quad (3.43)$$

So, underneath the $1 \text{ mm}^2$ area of the slider, only about 127 out of 2000 bumps contact the slider, which means the approximation of exponential height distribution used for deriving eq. (3.22) should be valid.

It is enlightening to examine the situations where the above analysis of a slider sitting on a laser textured disk begins to break down:

1. When the slider sits on the laser textured disk surface, the average separation scales with the mean bump height plus their standard deviation. Shorter bumps and a small standard deviation are preferred as this reduces the amount of interference once the slider starts to fly. If the average separation is too small, however, the adhesive force can be larger than the externally applied load. In particular, at separations smaller than $\sim 20 \text{ nm}$, the lubricant tends to flood the gap between the slider and the disk, dramatically increasing the adhesive force from the lubricant menisci around the contact points (Gui and Marchon 1995). To properly estimate the real area of contact, the adhesive force needs to be determined and added to the loading force. Ways of estimating adhesive forces will be discussed in later chapters.

2. The loading and adhesive forces compress the laser textured bumps [eq. (3.12)], reducing the mean slider–disk separation. For lower bump densities, the separation decrease from the bump compression can be a significant fraction of the mean separation due to the increase in the force per bump as the density decreases. The reduced separation leads to an increased adhesive force, which leads to further bump compression. At sufficiently low bump density and small initial bump heights, the stiffness of the laser texture bumps is not sufficient to prevent the increasing adhesive force during bump compression.
from collapsing the slider–disk separation, leading to catastrophically high friction (Gui and Marchon 1995; Gui, Kuo et al. 1997).

3. In our estimate of the real area of contact, we assumed that the top of the laser textured bumps and the slider surfaces are ideally smooth. As discussed in Section 2.4.2, the disk surface, while very smooth, still has a small amount of roughness, so potentially the real area of solid–solid contact may be less than that estimated by eq. (3.42). (Similarly, the slider surface also has a slight roughness, though typically smaller than the disk roughness.) To re-estimate the contact area taking into account the disk roughness at the top of the of the laser textured bumps, we will assume, for the disk, the roughness shown in Fig. 2.10 (measured for a disk that has summits with a standard height deviation $\sigma = 0.9\, \text{nm}$ and a mean radius of curvature $R = 80\, \text{nm}$) and, for the slider, an ideally smooth surface. In this case, the plasticity index $\psi$ from eq. (3.32) is

$$\psi = \left( \frac{\sigma}{R} \right)^{1/2} \left( \frac{E_c}{H} \right) = \left( \frac{0.9\, \text{nm}}{80\, \text{nm}} \right)^{1/2} \left( \frac{108\, \text{GPa}}{5\, \text{GPa}} \right) = 2.3 \quad (3.44)$$

where we have used the hardness for the magnetic layer. Since most of the deformation is plastic (since $\psi > 1$), we now need to use eq. (3.34) to estimate the real area of contact $A_r$:

$$A_r = \frac{L}{H} = \frac{20\, \text{mN}}{5\, \text{GPa}} = 4\, \mu\text{m}^2. \quad (3.45)$$

How do we account for this new estimate of the contact area being nearly two orders of magnitude smaller than the previous estimate [eq. (3.42)]? Figure 3.12 illustrates what is going on: When a laser texture bump contacts the flat surface, the loading force elastically deforms the bump to create a contact zone

![Fig. 3.12](image-url)

Fig. 3.12. (Left) Flat surface contacting three “sombrero” shaped laser textured bumps. (Right) Expanded view of the summit of a laser textured bump contacting the flat surface. The dashed line shows the macroscopic radius $R$, which is flattened at the summit due to elastic deformation to form a contact zone with diameter $2a$. The nanoscale roughness is also shown with an exaggerated height scale. The gray shows the lubricant and contaminant that fill in the gaps between the nanoscale roughness and the flat.
with diameter $2a$, over which the nanoscale roughness makes contact. Since the nanoscale roughness is only a few nanometers in height, the gaps between the roughness invariably fill with lubricant or mobile contaminants drawn into the narrow space by the molecular level attractive forces. So, the contact area estimated by eq. (3.42) (using the Greenwood–Williamson model with the radius of curvature of the laser texture bumps and ignoring their nanoscale roughness) includes both the solid–solid contact area at the summits of the nanoscale roughness and the area filled with lubricant/contaminant between the nanoscale summit contacts. The estimate of eq. (3.45) is then only for the solid–solid contact area at the summits of the nanoscale roughness, which is only a small fraction of the contact zone formed by the elastic deformation of the summit of the laser textured bump.

3.5 Inelastic impacts

If one object collides into another as shown in Fig. 3.13, the kinetic energy is converted into work compressing the two bodies elastically. An inelastic impact (Johnson 1972; Johnson 1985) occurs if the threshold for plastic deformation is exceeded: $p_m > 1.1Y$. The velocity $V_Y$ at which the plasticity threshold is reached is given by

$$\left(\text{kinetic energy}\right) \frac{1}{2} m V_Y^2 \simeq \frac{53R^3Y^5}{E_c^4} \left(\text{work needed to reach threshold}\right) \quad (3.46)$$

For the simple case of a sphere of density $\rho$ striking a flat surface, the plasticity threshold velocity is given by

$$\frac{\rho V_Y}{Y} = 26 \left(\frac{Y}{E_c}\right)^4. \quad (3.47)$$

Example: For a hard sphere striking a soft metal ($Y = 0.3\,\text{GPa}$, $E_c = 70\,\text{GPa}$), $V_Y = 0.02\,\text{m/s}$. Since the threshold velocity is so low, most impacts between metallic bodies involve some plastic deformation.

Fig. 3.13. An object of mass $m$ and velocity $V$ colliding with a flat surface. During the collision, an asperity with radius $R$ elastically deforms the contacting surfaces a distance $\delta$. 
Fig. 3.14. (Top) A ruby ball dropped onto a disk made on an AlMg alloy substrate. All the kinetic energy of the ball goes into plastically deforming a zone of radius $a$. (Bottom) Data from the author’s lab of the measured area of plastic deformation versus impact velocity; theory from eq. (3.48) with $H = 0.9$ GPa, $E_c = 70$ GPa, and $R = 1.5$ mm.

If the impact velocity is much higher than $V_Y$, most kinetic energy goes into plastically deforming the materials:

$$\frac{1}{2} m V^2 = p_m \frac{\pi a^4}{4R}$$  \hspace{1cm} (3.48)

where $p_m$ is the mean contact pressure during the impact ($p_m \approx H$) and $\pi a^4/4R$ is the volume of material plastically deformed during impact.

As a practical example, Fig. 3.14 shows the results obtained in the author’s lab for dropping hard, ruby balls onto disks typically found inside disk drives in desktop computers. These disks are made by depositing the magnetic layers and protective overcoats onto a substrate made from an AlMg alloy, with $H = 0.9$ GPa, $E_c = 70$ GPa. As the radius of the ball is fairly large, the maximum
pressure occurs at a depth of $8 \, \mu m$, well within the disk substrate material [Fig. 3.11(b)]. From the above example, $V_Y = 0.02 \, m/s$, so dropping a ball onto the disk from a height greater than $20 \, \mu m$ will cause an inelastic impact. In the experiment, balls were dropped from heights greater than $2 \, mm$, and most of the kinetic energy was dissipated through plastically deforming the AlMg substrate (no elastic rebound). As shown in the figure, the area of plastic deformation, as measured optically, was found to be linear with the impact velocity, with the slope predicted by eq. (3.48).

During a disk drive’s operation, if the recording head impacts the disk surface, with causes an inelastic impact like those in Fig. 3.14, information stored in the magnetic layer may be lost. This can be a serious problem for drives in portable computers, which can experience much more mechanical shock than desktop computers. Consequently, all disks in laptop drives are fabricated from glass substrates where the ratio of $Y/E_c$ is typically a factor of ten larger than for metal substrates.

### 3.6 References


In the last chapter we considered contact between materials where a loading force acts perpendicular to the contacting surface. In this chapter we consider what happens when a tangential force is applied to slide one material over another. When materials move against each other, frictional forces opposing this motion are generated in the contact regions. Consider, for example, the forces acting on a stationary block sitting on the inclined plane, as shown in Fig. 4.1(a), where one can adjust the incline angle $\alpha$ by using a lab jack. Fig. 4.1(b) shows the three distinct forces acting on the block to keep it stationary:

1. The weight of the block due to gravity $L = mg$
2. A reactive force $L \cos \alpha$ that the incline exerts in the direction normal to the block’s surface in response to the weight, and
3. A static friction force $L \sin \alpha$ that the incline exerts in the direction tangential to the block’s surface.

For small $\alpha$, the reactive and friction forces adjust as the angle $\alpha$ changes so that their vector sum balances the weight of the block, keeping the block stationary. As $\alpha$ increases, the static friction force increases proportionally to $\alpha$ until reaching a maximum value at some critical angle $\alpha_s$. Above this angle, the static friction force is constant, so the tangential component of the block’s weight exceeds the opposing friction force, and the block slides down the incline.

In this chapter, we examine how a friction force originates at the microscale contacts between solid surfaces and begin to examine the atomic origins of this force. In later chapters, we examine the friction of lubricated surfaces and delve more deeply into the atomic and molecular origins of friction.

### 4.1 Amontons’ and Coulomb’s laws of friction

In many situations, sliding friction during sliding can be described by the following “laws”:

- **First law of friction** The friction force is proportional to the normal load.
- **Second law of friction** The friction force is independent of the apparent area of contact.

The first law can be expressed by the simple equation:

$$ F = \mu L $$

(4.1)
where $\mu$ is called the coefficient of friction. This expression also implies the second friction law that friction is independent of the apparent area of contact. For the block sliding on the incline in Fig. 4.1, $\mu = \tan \alpha_s$, where $\alpha_s$ is the incline angle where the block starts to slide.

While the friction force usually has some dependence on sliding velocity, this dependence is typically fairly small, leading to a third law of friction:

*Third law of friction*  
Kinetic friction is independent of sliding velocity.

As this law is never strictly true, it is best not thought of as a law but rather as an approximation for the weak dependence of friction during sliding on velocity. Once sliding stops, however, the force needed to initiate sliding (static friction) is greater than the force needed to sustain sliding (kinetic friction).

The first two laws are referred to as Amontons’ laws of friction after Guillaume Amontons (1663–1705) who published them in 1699 in the Proceedings of the French Royal Academy of Sciences (Dowson 1978). The third law, that kinetic friction has a weak dependence on velocity was provided by Coulomb (1736–1806) eighty years later and is often referred to as Coulomb’s law of friction. The first two friction laws had actually been deduced previously by Leonardo da Vinci (1452–1519), in the middle of the fifteenth century, who described them in his notebooks which have only recently been discovered. Since Amontons published his results while da Vinci didn’t, the glory of having his name associated with the laws of friction belongs to Amontons.

Amontons’ first law, that friction is proportional to load, was readily accepted by the French Royal Academy, as it fitted with the everyday experience of the members that heavy objects are more difficult to slide than light objects. The second law, that friction is independent of apparent area of contact, however, was hotly disputed as many thought that friction should somehow also scale with contact area. We now know that friction does indeed scale but with the real area of contact, which is usually proportional to load, rather than the apparent area. This dispute prompted De la Hire to repeat Amontons’ experiments, and he confirmed Amontons’ conclusions, further solidifying support within the Academy for Amontons’ Laws.

At that time, another major concern (which continues to the present day) was that these laws are not strict physical laws derived from fundamental principles,
but rather empirical laws based on experimental observations. Since these laws apply to many practical situations, they are a good starting point for most analyses of friction. Many situations, however, deviate from these laws. In order to know when these laws can be applied and when they cannot, it helps to understand their physical origins.

Amontons, like many people who have studied friction, realized that the surfaces that he worked with were not completely smooth, and he suspected that roughness is somehow responsible for friction. He proposed that friction originated from the surface roughness in two ways: (1) For rigid asperities, friction was due to the force needed to pull the weight up the slopes of the surface roughness as shown in Fig. 4.2(a). (2) For deformable asperities, the asperities act as flexible springs that are bent over during sliding so that friction increases with the amount of deflection as illustrated in Fig. 4.2(b). For both these explanations, friction depends only on load and not the size of the contacting objects. De la Hire also believed that friction originated from roughness, but added the concept that friction also involved the work that comes from asperities breaking off during the sliding process as illustrated in Fig. 4.2(c). Coulomb in his studies of friction found that the static friction force was substantially higher than kinetic friction and would increase the longer the two surfaces stayed in stationary contact. Since the difference between static and kinetic friction was higher for wood

![Diagram](image.png)

**Fig. 4.2.** Early concepts of how friction originated from roughness. (a) Amontons’ concept for rigid asperities that friction originates from the force needed to raise the surfaces up the slopes of the roughness. (b) Amontons’ concept for flexible asperities that the friction originates from the force bending the asperities. (c) De la Hire concept that some of the friction comes from breaking off the asperity summits.
than for metal surfaces, Coulomb proposed that the wood surfaces were covered by fibers oriented like bristles on a brush. When two surfaces were pressed into contact the bristles slowly penetrate each other, increasing the static friction over time. Once the surfaces start to slide, Coulomb proposed that the bristles folded back on themselves to reduce friction.

These initial ideas involving how friction originates from roughness have now evolved over the intervening 300 years to where friction between two solid surfaces is thought to originate via two mechanisms:

1. An adhesion force needed to shear the contacting junctions where adhesion occurs.
2. A plowing force needed to plow the asperities of the harder surface through the softer surface.

4.2 Adhesion and plowing in friction

4.2.1 Adhesive friction

While numerous researchers had pointed out that adhesion occurs between two materials when they are brought into contact and that these interatomic forces should contribute to friction, it was Bowden and Tabor who in the 1930s and 1940s developed a model to describe the connection between adhesive forces and friction forces and performed experiments to verify these concepts (Bowden and Tabor 1950). The model is fairly simple: When two surfaces touch, the contacting asperities undergo elastic and plastic deformation, as analyzed in the previous chapter and illustrated in Fig. 4.3. At each contacting junction, labeled \( i \), a small contact area \( A_i \) is generated. Within \( A_i \), the surface atoms are in intimate contact generating attractive and repulsive interatomic forces between the atoms as discussed in section 3.1. When a tangential force is applied to slide one object over the other, shear stresses develop over the junction interfaces to resist this force. At low shear stresses, the interatomic forces between the atoms are strong enough to prevent the atoms from sliding over each other, and the material around the contact junctions deforms elastically to counter the applied tangential force. At some critical shear stress, the applied force acting on the individual atoms exceeds the sum of all the interatomic forces that are trying to keep the atoms in their positions, and they start to slide over each other. The shear stress needed to start and sustain sliding is called the shear strength \( s \), and the force needed to shear the \( i \)th junction is \( F_i = A_i s \).

If we assume that all junctions have the same shear strength \( s \), the adhesive friction force \( F_{adh} \) to shear all the junctions and slide the object is given by:

\[
F_{adh} = A_r s
\]

(4.2)

where \( A_r \) is the total real area of contact (\( A_r = \sum A_i \)). In reality the shear strength is likely to vary from junction to junction, so \( s \) more appropriately corresponds to the average shear strength of the junctions. In this model, the adhesive friction is proportional to load (Amontons' law) if the real area of
contact is proportional to load and the shear strength of contact junctions is independent of contact pressure. (A key advantage of the model of adhesive friction is that the expression $F_{\text{adh}} = A_r s$ also applies in situations where $A_r$ is not proportional to load—situations where Amontons’ Law is not valid—such as when two atomically smooth surfaces are brought together so that real area of contact becomes similar to the apparent area of contact.)

As discussed in Chapter 3, the real area of contact between rough surfaces is generally proportional to load. In particular, the Greenwood and Williamson model shows that, if the summit heights of the surface roughness have a Gaussian distribution and the deformation at the contacting asperities is primarily elastic, $A_r$ can be approximated (eq. (3.35)) as

$$A_r \approx 3 \left( \frac{R}{\sigma} \right)^{1/2} \frac{L}{E_c}$$

where $R$ is the average radius of curvature of the summits, $\sigma$ the standard deviation of summit heights, $L$ the loading force, and $E_c$ the composite elastic modulus. If instead the contact regions mainly deform plastically, $A_r$ can be
approximated by eq. (3.34):
\[ A_r \approx \frac{L}{H} \]
where \( H \) is the hardness of the softer material. For both cases, the real area of contact varies linearly with the loading force \( L \) pushing the two surfaces together, so the friction force \( F_{\text{adh}} = A_r s \) is also proportional to \( L \). If the shear strength is independent of contact pressure, \( F_{\text{adh}} \) is also independent of the apparent area of contact.

For example, metal surfaces finished by grinding or polishing, even though smooth to the touch, still have asperities sharp enough such that most asperities plastically deform during initial contact (\( \psi > 1 \)). For this situation, \( A_r = L/H \) and \( F_{\text{adh}} = sL/H \), leading to
\[ \mu_{\text{adh}} = \frac{s}{H} \]  
(4.3)
for shearing junctions of contacting metals or other contacting materials where most of the deformation is plastic. If the strong adhesion occurs at the contacting surfaces, the junctions rupture during shear within the weaker of the two materials, and \( s \) equals the shear strength of the softer material. From the Tresca yield criterion discussed in section 3.2.1, \( s = Y/2 \), where \( Y \) is the yield stress of the softer material. Since \( H \simeq 3Y \) [eq. (3.18)],
\[ H \sim 6s \]  
(4.4)
and
\[ \mu_{\text{adh}} \simeq \frac{s}{H} \sim \frac{1}{6} \]  
(4.5)
From this equation, we see that friction is directly related to shear strength and hardness of the softer material in those situations where strong adhesion and plastic deformation occur at the asperity contacts.

**4.2.2 Plowing friction**

The contribution to friction force from plowing hard asperities through a softer surface is called the plowing friction or \( F_{\text{plow}} \). A simple estimate of \( F_{\text{plow}} \) can be made by multiplying the total projected area \( A_0 \) in the direction of motion of the contacting asperities by the pressure that must be exerted on the softer material in order to induce plastic flow (approximately, its hardness \( H \)):
\[ F_{\text{plow}} \approx A_0 H. \]  
(4.6)

This is illustrated in Fig. 4.4 for a single, cone shaped asperity plowing through a soft material. A cone shaped asperity is used to keep the geometry simple [Fig. 4.4(b)]. The single asperity in Fig. 4.4 exerts a plowing force \( F_{\text{plow}} \), to displace the softer material, that is \( H \) times the cross-sectional area of the groove \( ax = x^2 \tan \varphi \):
\[ F_{\text{plow}} = H x^2 \tan \varphi. \]  
(4.7)
Fig. 4.4. (a) A rigid cone shaped asperity plowing a groove through a softer material. (b) Schematic showing the cone’s semi-angle $\varphi$ and the groove depth $x$.

As the asperity plows through the softer material, the loading force $L$ is supported by the contact pressure $H$ of the softer material acting over an area $\pi a^2/2$ underneath the cone:

$$L = \frac{1}{2} H \pi a^2 = \frac{1}{2} H \pi x^2 \tan^2 \varphi.$$  \hspace{1cm} (4.8)

This leads to a coefficient of friction for plowing of

$$\mu_{\text{plow}} = \frac{F_{\text{plow}}}{L} = \frac{2}{\pi} \cot \varphi.$$  \hspace{1cm} (4.9)

As the slopes of most surfaces of practical interest are less than $10^\circ (\varphi > 80^\circ)$, we should expect $\mu_{\text{plow}}$ to be less than 0.1.

While the adhesive and plowing contributions to friction are not completely independent, it is convenient to treat them separately, as we have been doing, and express the total friction force $F$ as the sum

$$F = F_{\text{adh}} + F_{\text{plow}}.$$  \hspace{1cm} (4.10)

For the particular case where one of the contacting materials is much softer than the other and most contact junctions deform plastically, so that both eqs. (4.5) and (4.9) are valid, the total friction arising from the combination of adhesion and plowing can be expressed by the coefficient of friction:

$$\mu = \frac{s}{H} + \frac{2}{\pi} \cot \varphi$$

$$\mu \simeq \frac{s}{H} + \tan \theta$$  \hspace{1cm} (4.11)

where $\theta = 90^\circ - \varphi$ represents a typical slope of the plowing asperities. Since $s/H \sim 0.17$ (eq. (4.5)), eq. (4.11) predicts that $\mu$ should not exceed 0.2, for contacting metals with similar hardness (so that the plowing contribution is negligible), and $\mu$ should not exceed 0.3 for a hard metal sliding on a softer one. Experimentally, however, the friction coefficients for metals are found to be several times these values, indicating that other mechanisms are contributing to friction. The two main ones for sliding metals are work hardening and junction growth.
4.2.3 Work hardening

The previous analyses assumed that the yield strength of the materials remains constant during the sliding process. Ductile materials, however, undergo a process of work hardening or strain hardening where the material’s yield stress, hardness, and shear strength increase as the amount of plastic deformation increases.

This is how work hardening occurs on the atomic scale: Plastic deformation occurs when atoms slide past each other in the form of slip planes and dislocations (Section 3.1). Eventually, the slip planes and dislocations run into obstacles, such as impurities or grain boundaries, pinning them and making further deformation more difficult. Since a higher stress has to be applied to continue the plastic flow, a previously plastically deformed material has higher hardness and shear strength. Asasperities sheared during sliding are more plastically deformed near the surfaces of the contact zones than regions far away where the load is supported, the surface shear strength $s$ increases faster than the near surface hardness, leading to higher values of $\mu_{\text{adh}}$ than expected from eq. (4.5) or (4.11).

4.2.4 Junction growth

A more important contribution to higher values of $\mu_{\text{adh}}$ is junction growth (Tabor 1959). When we derived the contact area expression for plastically deforming asperities [eq. (3.34)], we assumed that only normal forces were acting on the contacting asperities. If a tangential force acts simultaneously on an asperity, plastic deformation occurs at a lower value of the normal contact pressure. To illustrate this consider the idealized asperity shown in Fig. 4.5. First, a loading force $L$ is applied so that the contact pressure $p_0$ in the asperity is at the onset of yield. Then, a tangential force $F$ is applied to the asperity to generate a shear stress $\tau$ within the asperity. The extra stress from $\tau$ causes plastic flow of the asperity until the normal pressure is reduced to $p_1$. The relationship between $p_1$, $\tau$, and $p_0$ is determined by the particular shape of the asperity and the yield criterion. For the cubic shaped asperity in Fig. 4.5,

$$p_1^2 + 4\tau^2 = p_0^2$$

(4.12)

Fig. 4.5. Model for plastic deformation in an idealized, cubic shaped asperity pressed against a counterface. The stresses are shown acting on an internal element: (a) with only a normal force $L$ and (b) with both a normal force $L$ and an applied tangential force $F$. 

which is a statement of the Tresca criterion for yield (Hill 1950). Converting pressure and stresses into the normal and tangential forces \( (p_1 = L/A, \tau = F/A) \), we obtain

\[
L^2 + 4F^2 = A^2 p_0^2. \tag{4.13}
\]

Since \( p_0 \), the compression yield stress, is a material property that remains constant, the contact area \( A \) will grow as the applied tangential force \( F \) is increased. In principle, the junction area growth could increase indefinitely but is usually limited by two factors: (1) the reduced ductility of the asperity by work hardening (increase of \( p_0 \)) and (2) the presence of a weak interfacial film that has a lower shear strength \( s_i \) than the bulk shear strength \( s_0 \). Lubricant films are examples of weak interfacial films that are deliberately applied to the surfaces to provide a low shear strength.

It is straightforward to extend the model to include the effect of a weak interfacial film: As the tangential force \( F \) increases, the junction area \( A \) grows until it reaches the value

\[
A_{\text{max}} = \frac{F_{\text{max}}}{s_i}. \tag{4.14}
\]

when sliding occurs at the interface film with a low shear strength \( s_i \). Tresca’s yield criterion can be used to relate the bulk shear strength to the bulk yield stress:

\[
p_0 = 2s_0. \tag{4.15}
\]

From eq. (4.13), the maximum contact area \( A_{\text{max}} \) of the junction before the onset of sliding is given by

\[
A_{\text{max}} = \frac{L}{2\sqrt{s_0^2 - s_i^2}}. \tag{4.16}
\]

Combining eqs. (4.13)–(4.16) we derive the maximum adhesive friction for an individual junction with junction growth:

\[
\frac{F_{\text{max}}}{L} = \frac{1}{2\sqrt{(s_0/s_i)^2 - 1}}. \tag{4.17}
\]

This expression for the \( F_{\text{max}}/L \) is plotted in Fig. 4.6.

From eqs. (4.16) and (4.17) and from Fig. 4.6, we can see that, as \( s_i \) tends to the bulk value \( s_0 \), both the maximum contact area and the friction of the junction tend to infinity. We also see that surface films with low interface shear strengths greatly reduce friction by inhibiting junction growth. In the limit where the interface film can be sheared almost effortlessly \( (s_i << s_0) \), no junction growth occurs: \( F = A s_i \) and \( L = A p_0 \), leading to

\[
\mu = \frac{F}{L} = \frac{s_i}{p_0}. \tag{4.18}
\]
4.3 Static friction

From our everyday experience in dealing with friction, most of us have learnt that the force required to start an object sliding is greater than the force needed to maintain sliding, i.e., static friction is greater than kinetic friction. What is less obvious is that the value of static friction typically increases the longer the two surfaces stay in stationary contact, as illustrated in Fig. 4.7(a), and the kinetic friction typically decreases as the sliding velocity increases [Fig. 4.7(b)]. Generally, the increase in static friction to an asymptote is so quick that we think of static friction as having constant value. Similarly, the drop in kinetic friction from its instantaneous rest time value to a lower value that changes slowly with velocity happens over such a small range of velocity near zero that kinetic friction is typically approximated as being independent of velocity (remember the third law of friction).

The increase in static friction with longer rest times can be understood at the atomic and molecular level as follows: When two surfaces are brought into contact, with an applied load $L$ at time $t = 0$, a contact area $A_c(t = 0)$ initially forms, and the lateral force needed to overcome the adhesive friction is $F_{adh} = sA_c(0)$. At $t = 0$, however, the atoms in each contact zone start experiencing the atomic forces acting across the contact [Fig. 4.8(a)]; The long range adhesive forces pull
Fig. 4.7. Typical plots of (a) static friction as function of rest time, and (b) kinetic friction as a function of sliding velocity.

Fig. 4.8. (a) Arrangement of atoms on an asperity at \( t = 0 \). The arrows indicate the directions that the atoms move to reach their new equilibrium positions in (b).

the atoms and molecules surrounding the contact junction into the junction, and the short range repulsive forces push atoms out from the center of the contact junctions. Over time, the atoms move to more energetically favorable locations at the junction edges [Fig. 4.8(b)]. As the solid–solid contact area \( A_c(t) \) grows with time, the adhesive friction \( F_{adh} \) grows.

In addition to the movement of solid atoms, molecules from the surrounding environment may also move into the crevices around the contacting asperities, drawn by the attractive interatomic forces, as illustrated in the inset in Fig. 4.9. As these molecules collect around the contacting asperities, they form liquid menisci that exert an adhesive force on each asperity due to the capillary pressure inside the meniscus. The net adhesive force \( L_{adh} \) from the sum of all the menisci forces adds to the applied load \( L \). In accordance with Amontons’ law, the static friction force \( F_s(t) \) increases as the net normal force \( (L + L_{adh}(t)) \) increases:

\[
F_s(t) = \mu(L + L_{adh}(t)). \tag{4.19}
\]

A common example of this: When the humidity is high, people complain about it feeling “sticky” as the water in their sweat, which is less likely to evaporate at high humidity, exerts meniscus forces on contacting parts of the body. Similarly, at high humidity, water vapor condenses around the contact points between
Fig. 4.9. (a) Two surfaces before they touch. The bottom surface has a thin film of mobile molecules, such as a thin film of liquid lubricant or of adsorbed molecules from the air. (b) After the two surfaces touch, the molecules are pulled into form menisci around the contact asperities, as illustrated in the inset.

objects with hydrophilic surfaces, such as dust particles on glass windows, causing them to stick more strongly.

Another example: A thin film of lubricating oil applied to reduce kinetic friction, such as illustrated in Fig. 4.9, often makes static friction worse. The reason is that when the surfaces stay in stationary contact, lubricant is drawn into and accumulates in the narrow gaps between the contacting surfaces, leading to a higher meniscus adhesion force, resulting in higher friction.

The interatomic forces that result in the junction contact areas increasing over time, of course, also exist when the two surfaces slide over each other. During the sliding, contact junctions are constantly being formed and ruptured. During the brief period that a junction exists, the atoms are pushed and pulled toward new equilibrium positions, increasing the contact area of the junction. This leads to the kinetic friction of solids decreasing with increasing sliding speed as illustrated in Fig. 4.7(b). This phenomenon has a similar physical origin to the increase in static friction with rest time: For slower sliding speeds, the contact junctions exist for longer, allowing for more time for their contact areas to grow, leading to higher friction. As discussed in a later chapter on lubrication, when a liquid film is sheared between two solid surfaces, however, the opposite trend is observed: friction increases with velocity, as the frictional forces are generated by viscous dissipation within the liquid.

4.3.1 Stick–slip

One consequence of static friction’s being higher than kinetic friction and of kinetic friction’s decreasing with increasing sliding speed is the phenomenon of stick–slip: rather than sliding at a constant speed, sliding occurs as a sequence of sticking and slipping or as an oscillation at a resonance frequency of the system. Stick–slip can manifest itself as an unpleasant squealing and chattering noise, such as a squeaky door hinge or the squeak of nails on a blackboard. It is also frequently a sign of a malfunctioning device; for example, as windshield
wipers are exposed to sunlight, they degrade, and the static friction increases relative to the kinetic friction until eventually they stick–slip ineffectively across the windshield, instead of smoothly wiping water droplets away as when new. In a few cases stick–slip may actually be desirable, for example, in the action of a bow rubbing across the strings of a violin, resulting in the string’s producing musical tone.

4.3.1.1 Velocity-controlled stick–slip

Stick–slip arises from the interplay of friction with the dynamics of the mechanical system. To better see how stick–slip comes about, we consider the mechanical system shown in Fig. 4.10. In this system, a block with mass \( m \) sits on a belt moving with velocity \( V \). Restraining the motion of the block is a spring element,

![Fig. 4.10.](image)

(a) A block sitting on a belt moving with velocity \( V \) and restrained by a spring element and dashpot. (b) How friction force \( F \) acting on the block depends on relative velocity between the block and the belt.
with stiffness $k$ and dashpot damper $\eta$, that is attached to the fixed support on the left. A load $L$ (the weight of the block $mg$ plus the externally applied load and adhesive forces) pushes the block in the normal direction against the belt, and a friction force $F = \mu L$ acts in the tangential direction to oppose relative motion between the block and belt. When the displacement $x$ of the block away from the spring’s equilibrium position is sufficiently small so that $F < \mu_s L$, the block moves with the same velocity as the belt ($\dot{x} = V$). In this case, the tangential forces acting on the block are

$$|\eta V + kx = F| < \mu_s L.$$  \hspace{1cm} (4.20)

As displacements become larger, the spring elongates and eventually exerts a force exceeding the static friction $\mu_s L$, at which point the block starts to slide against the belt with a relative velocity $V_{rel} = V - \dot{x}$. When the block slides over the belt, its equation of motion is

$$m\ddot{x} + \eta \dot{x} + kx = F.$$  \hspace{1cm} (4.21)

Here we consider velocity-controlled stick–slip that originates from a friction force decreasing with increasing velocity. To simplify the analysis, we assume that friction decreases linearly with increasing relative velocity:

$$F = (\mu_s - \alpha V_{rel})L \quad \text{if } V_{rel} > 0$$  
$$F = -(\mu_s + \alpha V_{rel})L \quad \text{if } V_{rel} < 0$$  \hspace{1cm} (4.22)

where $\alpha$ is the slope of friction as a function of velocity. The assumption that kinetic friction is linear with velocity can be valid for a sufficiently small range of the friction vs. velocity curve in Fig. 4.7(b); for example, Fig. 4.10(b) shows the friction for a small range of sliding speeds near zero velocity.

Combining eqs. (4.21) and (4.22) leads to

$$m\ddot{x} + (\eta - \alpha L)\dot{x} + kx = (\mu_s - \alpha V)L \quad \text{if } V_{rel} > 0$$  
$$m\ddot{x} + (\eta - \alpha L)\dot{x} + kx = (-\mu_s - \alpha V)L \quad \text{if } V_{rel} < 0$$  \hspace{1cm} (4.23)

The interesting feature of eq. (4.23) is the term $(\eta - \alpha L)$ on the left side of the equations, which represents the effective damping coefficient of the mechanical system. When this term is negative $[(\eta - \alpha L) < 0]$, the energy being supplied by the friction force to the mechanical system is greater than the energy being dissipated in the dashpot damper, driving an oscillatory or stick–slip motion. This occurs at high load $L$ or at high $\alpha$, the slope of the $\mu_k - V_{rel}$ curve. Otherwise, for $(\eta - \alpha L) > 0$, the effective damping term is positive, indicating that any oscillation in the spring system is eventually damped out and no steady-steady oscillation or stick–slip motion occurs.

Figure 4.11 shows solutions for the motion of the block initially sitting on the belt at $t = 0, x = 0$, and $\dot{x}(0) = V = 1 \text{ m/s}$ using the differential equation (4.23) for when $V_{rel} \neq 0$ and using the condition (4.20) to determine when $V_{rel} = 0$. At first, the block moves with the belt until the static friction is exceeded.
Fig. 4.11. Motion of the block on the moving belt illustrated in Fig. 4.10. For this analysis, \( m = 1 \text{ kg}, k = 100 \text{ N/m}, \mu_s = 0.5, \alpha = 0.2, V = 1 \text{ m/s}, \eta/2m = 10 \text{ s}^{-1} \) (\( \eta \) chosen to provide critical damping of the mass–spring system.) Different loads are used to generate different values of the effective damping coefficient \((\eta - \alpha L)\): dot-dash line, \( L = 50 \text{ N}, (\eta - \alpha L) = +10 \); dotted line, \( L = 100 \text{ N}, (\eta - \alpha L) = 0 \); and dashed line, \( L = 150 \text{ N}, (\eta - \alpha L) = -10 \).

- For the low load case, where the effective damping coefficient is positive, once the static friction is exceeded, the displacement of the spring quickly settles to a constant value corresponding to a constant friction force acting on the block as it slides smoothly against the belt.

- If the load \( L \) or the slope \( \alpha \) is increased, the damping coefficient becomes less positive. For the special case where \((\eta - \alpha L) = 0\), once the static friction is exceeded the block undergoes a sinusoidal motion about the displacement corresponding to the average friction during sliding. During this oscillation, the energy received by the block from friction while traveling in the same direction of the belt is lower than the energy dissipated through friction when moving in the opposite direction of the belt, as the friction is lower for the higher relative velocity. The extra energy from the difference in friction drives the oscillation through the cycle before being dissipated in the spring damping.

- At higher loads, where the damping coefficient is negative, the kinetic friction is high enough that the block stops moving relative to the belt or “sticks” to the belt for part of the oscillation; once the static friction is exceeded, the block “slips” to the next part of the oscillation where it sticks again.

4.3.1.2 Time-controlled stick–slip

In the previous section, we assumed that the static friction force did not increase significantly as the stick time increased. As discussed earlier, static friction tends to rise the longer two surfaces are in stationary contact, as illustrated in Fig. 4.7(a). If the increase in static friction during a stick phase...
of the oscillation is greater than the variation in kinetic friction during the slip phase of the oscillation, the stick–slip process is said to be time-controlled rather than velocity-controlled. To understand how a higher value of static friction leads to stick–slip, we will follow the approach originally developed by Rabinowicz (Rabinowicz 1958). We again consider the block on the moving belt in Fig. 4.10(a), but this time with the friction increasing with stick time as shown in Fig. 4.12. If the block is initially at rest at $t = 0$, the force exerted by the spring on the block during the first stick phase increases as $kVt$. At the point A, the spring force exceeds the static friction, and the block slips across the belt surface until coming to rest at point B. Since kinetic friction $F_k$ is assumed constant (represented by the straight line in Fig. 4.12), the friction force at point B is as much below $F_k$ as point A is above $F_k$. The block undergoes a second stick phase from C to D, followed by a second slip to E before settling into a steady state stick–slip oscillation. From the diagram in Fig. 4.12, we see that increasing the spring constant $k$ or the belt velocity $V$ increases the slope $kV$ of the stick phase, leading to a smaller amplitude of the stick–slip oscillation. If $kV > (dF/dt)_{t=0}$, smooth sliding rather occurs than stick–slip.

4.3.1.3 Displacement-controlled stick–slip

In addition to being velocity-controlled and time-controlled, stick-slip can also be displacement-controlled, where stick–slip is caused by the friction force varying as function of position over the sliding surface. For example, if the belt in Fig. 4.10(a) has a non-uniform surface so that the friction force $F$ varies along

---

**Fig. 4.12.** Static friction increasing with time for generating time-controlled stick–slip. Reproduced from (Rabinowicz 1958).
Fig. 4.13. (a) Geometry of an atomic force microscope (AFM) tip contacting a mica surface. The tip was made by etching the end of a tungsten wire to a sharp point (tip radius $R \sim 300$ nm) and bending it toward the sample so that the cantilever spring constant is the same orthogonal to the surface and in the $x$-direction. (b) Friction loop for an AFM tip sliding over a cleaved mica surface. The deflection of the AFM cantilever parallel to surface is plotted vs. the sample position in the weak direction of the cantilever, with a spring constant $= 100$ N/m and a load $L = 10^{-6}$ N. Reprinted with permission from Erlandsson et al. (1988). Copyright 1988, American Institute of Physics.

Displacement stick–slip occurs not only for macroscale systems like the block on the belt in Fig. 4.10, but also at the atomic scale. This type of atomic scale stick–slip process can be readily observed in an atomic force microscope (AFM) such as the one shown in Fig. 4.13(a), which has a cantilever that flexes easily in the direction of the friction force. Figure 4.13(b) shows a “friction loop” generated by such an AFM tip sliding back and forth over a freshly cleaved mica surface. When the direction of sample motion changes, the tip initially moves with the sample, and the wire cantilever deflection is proportional to the static friction force. At point A, the cantilever wire exerts enough lateral force on the tip to overcome the static friction, and the tip starts to slide across the surface. The sliding process is not uniform, but rather occurs as a series of slips with each having a distance corresponding to the 5.2 Å unit cell size of the hexagonal SiO$_4$ layer of the cleavage plane of muscovite mica.

This displacement-controlled stick–slip process occurs when the spring constant of the AFM cantilever parallel to the surface is less than the spatial...
derivative of the friction force that the AFM tip experiences as it slides over the surface. When the tip slides over a crystal surface with a periodic arrangement of surface atoms, the tip experiences a friction force with the same periodicity as the crystal lattice. When the slope of the decreasing portion of the periodic force is greater in magnitude than the cantilever spring constant, the tip slips to the next stable position. Figure 4.14 illustrates how this occurs.

A remarkable thing about the atomic scale stick–slip process shown in Fig. 4.13 is that the contact area is much larger than the unit cell area of the mica surface. For a tungsten tip rubbing against mica, the composite elastic modulus $E_c = 19$ GPa; then, using eqs. (3.7) and (3.9), one can estimate the contact radius and area as $a = 23$ nm and $A = 1633 \text{ nm}^2$ when $L = 10^{-6}$ N and $R = 300$ nm. This estimate for contact area implies that the contact zone covers approximately 6000 mica surface unit cells. Even though the friction force on each atom at the end of the tungsten tip is periodic as the tip slides over the surface, these forces are randomly out of phase with each other due to the inhomogeneous and disordered nature of the atoms on the AFM tip surface. Since the atoms should be incommensurate across the sliding interface, the net friction force after summing over all the tip atoms should have only a small periodic component, which gets smaller as the contact area becomes larger. The relatively large periodic force observed in Fig. 4.13, which becomes larger as the contact area increases at higher loads, implies that the tip atoms that slide over the surface are somehow coherent with those of the mica surface.

Since these large friction forces with atomic scale periodicity are observed on layer compounds like mica and graphite, a likely explanation is that a small flake...
of either mica (Erlandsson et al. 1988) or graphite (Mate et al. 1987) becomes attached to the end of the tip (Pethica 1986; Dienwiebel et al. 2004). As this flake is dragged over the surface, the atoms on the flake coherently go in and out of phase with the sample lattice, resulting in a friction force with periodicity of the lattice. This hypothesis predicts that the friction force is sensitive to orientation of the flake with the friction force being highest at those angles where the flake atoms line up with the substrate surface atoms. (Dienwiebel et al. 2004) have measured the friction of a tip sliding over graphite sample as a function of the rotation angle of the sample underneath the tip; they found indeed that high friction with atomic scale stick–slip is observed at every 60 degrees of crystal orientation, corresponding to the angles where the atomic lattice on the graphite flake impaled on the tip lines up with the lattice on the graphite sample surface. At small angles away from these high friction angles, the friction force drops to a very low value since any misorientation of the flake results in the atoms being incommensurate across the sliding interface. From the width of the friction peak as function of angle, Dienwiebel et al. were able to estimate the diameter of the graphite flake attached to the tip to be 7 to 12 lattice spacings across.

4.4 References


SURFACE ENERGY AND CAPILLARY PRESSURE

To most people, it is not initially obvious that surfaces should have an energy associated with them. Not only are significant amounts of energies associated with the surfaces and interfaces of solids and liquids, these surface energies are responsible for driving many surface phenomena. For example, surface energies account for how detergents clean our clothes, how paints spread on surfaces, and how insects are able to walk on water. Consequently, the concept of surface energy underlies much of our understanding of the surface phenomena that make up tribology. In this chapter, we introduce surface energy and several related concepts: work of adhesion, capillary pressure, and adhesion hysteresis. The following chapters build on these concepts to understand the forces between surfaces and the relationships between surface energy, lubrication, and friction.

First, let’s cover a few points of nomenclature. The extra energy needed to form a material surface exposed to gas or vacuum is called either the surface free energy, the excess surface free energy, or the surface tension. In this book, we use the terms surface energy when describing solids and surface tension when describing liquids. For a surface that forms an interface between two condensed phases, the extra energy associated with this energy is called the interfacial energy. The energy associated with surfaces or interfaces is represented by the symbol $\gamma$ and has units of energy per unit area ($\text{ergs/cm}^2 = \text{mJ/m}^2$) or, equivalently, force per unit distance ($\text{mN/m} = \text{dynes/cm}$, equivalent to $\text{ergs/cm}^2$, which are the units for surface tension).

The terms surface and interface are often used interchangeably to describe the boundary between two material media or phases. The term surface, however, is used generally when a condensed (solid or liquid) phase is on one side and a rarified (gas or vacuum) phase is on the other side of the interface; while, the term interface is used, when condensed phases are present on both sides of the interface.

5.1 Liquid surface tension

You have probably noticed that when water droplets form they have a spherical shape rather than a cubic or irregular shape. The reason for this spherical shape is that when a droplet forms, extra energy is needed to form the droplet’s surface; to minimize this surface energy, the liquid rearranges to have the minimum surface area, which occurs with the spherical shape.
Figure 5.1 illustrates the forces acting on molecules in a liquid that give rise to surface energy. A liquid molecule in the bulk feels the attractive cohesive forces from its neighbors uniformly distributed around it, while the cohesive forces acting on a molecule at the surface are unbalanced since this surface molecule lacks neighbors on the surface side. So, if a molecule starts to move away from the surface, the attractive cohesive forces pull it back into the liquid, creating a tension that pulls the liquid surface flat, hence preference for the name *surface tension* when discussing liquid surfaces. Thermodynamically, the surface tension $\gamma$ is the reversible work $W$ required to create a unit area $A$ of a liquid surface:

$$\gamma = \frac{dW}{dA}. \quad (5.1)$$

In other words, the surface tension is the work done per unit area against the cohesive forces to bring molecules from the bulk to the surface. Surface tensions of typical liquids in air at room temperature range from 20–40 mN/m for hydrocarbons (for example, $\gamma = 25 \text{ mN/m}$ for paraffin wax), to 73 mN/m for water, and up to 485 mN/m for mercury. The increasing surface tension from hydrocarbons to water to mercury indicates increased bond strength (higher cohesive energy) between the atoms in liquid mercury compared with the bond strength between the liquid hydrocarbon molecules.

An easy way to visualize that tension really exists at liquid surfaces is to sprinkle some pepper on the surface of water in your kitchen sink and then dip a bar of soap into the water. The pepper floating on the water moves rapidly away from the location where the soap touches the water. What is happening? As the water surface becomes covered with a monolayer of soap, the surface tension drops from that of bare water (73 mN/m) down to that of soap (30 mN/m). When the boundary of the spreading soap film encounters the grains of pepper,
the 43 mN/m difference in surface tension pulls the grains across the surface in front of the spreading soap, since, at the boundary between the covered and bare water surfaces, the soap surface molecules exert smaller cohesive forces on grain than the water surface molecules.

A simple way of measuring the surface tension is to use a Wilhelmy plate setup as illustrated in Fig. 5.2, which converts the surface tension into a measurable force. In this measurement, a plate, with width \( L \) and thickness \( dL \), is lowered into the liquid, whose surface tension \( \gamma \) is being measured. The material of the plate is chosen so that the liquid wets the plate, i.e., the meniscus contact angle = 0°. As discussed later in Section 5.4.2.1.3, this condition occurs when the work of adhesion between liquid and solid is greater than liquid internal cohesion energy; that way, when the plate touches the liquid surface, the attraction of the plate material for the liquid pulls it up to form a meniscus with a zero contact angle, enabling the liquid surface tension to act parallel to the plate surface, i.e., in the vertical direction. At the contact line at the top of the meniscus, the liquid exerts on the plate the same tension as occurs on the liquid surface. The total force \( F_{\text{men}} \) exerted on the plate by the meniscus equals the contact line perimeter times the liquid surface tension \( \gamma \):

\[
F_{\text{men}} = 2(L + dL)\gamma. \tag{5.2}
\]
By measuring the force acting on the plate in excess of the weight of the plate \(F_{\text{men}} = W_{\text{total}} - W_{\text{plate}}\), the liquid surface tension \(\gamma\) can be determined by solving eq. (5.2) for \(\gamma\).

Numerous other methods exist for measuring liquid surface tensions which are detailed in the book by Adamson and Gast (1997).

### 5.2 Capillary pressure

As said earlier, the liquid rearranges inside a water droplet to form a sphere as this shape minimizes the surface energy. If the droplet is initially non-spherical, the liquid needs to move around within the droplet to achieve the low energy spherical configuration; this liquid movement is driven by pressure differences between regions of the surface that have different curvature; these pressure differences originate from the surface tension pushing and pulling the liquid toward the minimum energy geometry.

The pressure associated with a curved liquid surface is called the *capillary* or *Laplace pressure* and is described by the equation developed by Young and Laplace in 1805. A simple derivation (Adamson and Gast 1997) of this equation can be obtained by considering a soap bubble (Fig. 5.3), with an equilibrium radius \(r\) and pressure difference \(\Delta P\) between air inside and outside the bubble; the goal of the derivation is to determine the relationship between \(r\) and \(\Delta P\). For this soap bubble, the total surface free energy is \(4\pi r^2 \gamma\), where \(\gamma\), in this case, refers to the combined surface tension from the two sides of the soap film. If the bubble shrinks by an amount \(dr\), the total surface energy becomes

\[
4\pi (r - dr)^2 \gamma = 4\pi r^2 \gamma - 8\pi r dr \gamma + 4\pi dr^2 \gamma \\
\approx 4\pi r^2 \gamma - 8\pi r \gamma dr.
\]  

(5.3)

The decrease in surface energy \(8\pi r \gamma dr\) is converted into the work \(4\pi r^2 \Delta P dr\) compressing the air inside the bubble:

\[
8\pi r \gamma dr = 4\pi r^2 \Delta P dr. 
\]  

(5.4)

---

**Fig. 5.3.** Soap bubble with an initial radius \(r\) that shrinks by an amount \(dr\).
Rearranging eq. (5.4), we arrive at the equation for the equilibrium pressure difference between the inside and outside of the bubble:

$$\Delta P = \frac{2\gamma}{r}. \quad (5.5)$$

This derivation for the pressure difference can be extended to the more general case of an arbitrarily curved liquid surface. This pressure difference $\Delta P$, the Laplace pressure or the capillary pressure ($P_{cap}$), is given by the equation of Young and Laplace:

$$P_{cap} = \gamma \left( \frac{1}{r_1} + \frac{1}{r_2} \right) \quad (5.6)$$

where $r_1$ and $r_2$ are the two principal radii of curvature that define the curved surface. If the surface is spherical in shape, $r_1 = r_2 = r$ and capillary pressure is given by eq. (5.5).

The sign of $r_1$ or $r_2$ depends on which way the surface is curved: If the liquid meniscus curves inward when looking from the outside of the liquid, $r > 0$, as for the droplet illustrated in Fig. 5.4(a). If the surface curves outward, then $r < 0$, as for the meniscus that wets the inside of a narrow capillary tube like that illustrated in Fig. 5.4(b). From the equations for capillary pressure [eqs. (5.5) and (5.6)], we see that the signs of the radii of curvature governs whether the

![Diagram of liquid droplet and capillary rise](image)

**Fig. 5.4.** (a) Liquid droplet with radius $r > 0$, resulting in the pressure inside the droplet being higher than in the air outside. (b) The capillary rise of a liquid in a narrow cylindrical tube where the liquid wets the wall of the tube, for example, water in a glass tube with a hydrophilic surfaces. The meniscus has a radius $r < 0$, resulting in the liquid pressure around the meniscus being lower than the air outside, which pulls the liquid up the tube a distance $h$. 
pressure in the liquid is greater than or less than the surrounding environment. For example, for the liquid droplet shown in Fig. 5.4(a), the pressure in the liquid is higher than that of the air outside the droplet. For the example in Fig. 5.4(b), having $r < 0$ leads to the liquid rising up the capillary tube. The “capillary rise” occurs since the pressure of the liquid around the meniscus is less than that of the outside air; the liquid is therefore pulled up the tube until an equilibrium height $h$ is reached where the weight of the liquid in the tube exerts a pressure equal and opposite to the capillary pressure: equilibrium occurs when $h g \rho = -2 \gamma / r$, where $g$ is the acceleration due to gravity and $\rho$ is the liquid density. This capillary rise to the same height $h$ occurs even if the air is removed so that a vacuum (pressure equal to zero) exists outside the liquid, indicating that the absolute liquid pressure can be less than zero.

One might be tempted to think that, since the pressure is higher inside the liquid droplet than outside it, the droplet should expand until the pressure difference becomes zero. Likewise, for the capillary rise geometry in Fig. 5.4(b), one might think that, since the pressure is higher above the meniscus than below it, the meniscus would be pushed down below the liquid reservoir surface, in the same manner that the high pressure behind a piston pushes it towards the low pressure end of a tube. Two analogies to other physical systems can help clarify one’s understanding: (1) Within the Earth, the weight of the Earth’s outer layers pushes down on the interior generating very high pressures; this positive pressure, however, does not cause the Earth’s surface to move toward outer space as this pressure is counterbalanced by the gravitational forces. (2) A balloon has significantly higher pressure inside than out, but does not expand as the air pressure pushing against the balloon’s interior surface is counterbalanced by the elastic stress generated by stretching the balloon. Similarly, a liquid droplet does not swell due to the higher pressure inside, as this internal pressure is counterbalanced by the stress generated from the surface tension that pulls on the curved liquid surface. In fact, the droplet’s internal pressure is the physical consequence of the energy and tension associated with the liquid surface.

5.2.1 Capillary pressure in confined places

In tribology, we are frequently interested in what happens in confined spaces, such as in the narrow gap between two solids when they are brought into contact. In these confined spaces, liquids are pulled in by capillary action or condensation and, once there, have a tremendous influence on adhesion, friction, and lubrication. In a later chapter, we discuss directly how the capillary pressure influences adhesion forces; in this section, we outline some of the possible ways that liquids distribute themselves within confined geometries, and the associated capillary pressures.

First let’s consider the example where a drop of water is placed on a glass microscope slide and another slide is placed on top of it. As the two slides come together, the droplet spreads out within the gap between the slides to form a pillbox shaped meniscus as shown in Fig. 5.5.
Fig. 5.5. Drop of water between two microscope slides with either (a) hydrophilic or (b) hydrophobic surfaces. For both cases, \( P_{\text{cap}} \sim \gamma/r \) but it is negative for the hydrophilic case (a) and positive for the hydrophobic case (b).

If the slides have clean, hydrophilic surfaces [Fig. 5.5(a)], the water contact angle \( \theta \) is typically small, certainly \(<90^\circ\), so the meniscus has a concave surface with \( r<0 \) at the water–air interface. Using eq. (5.6) with \( r_1 = r \) and \( r_2 = R \), we see that

\[
P_{\text{cap}} = \gamma \left( \frac{1}{R} - \frac{\cos \theta}{d/2} \right) \approx -\frac{2\gamma \cos \theta}{d}.
\] (5.7)

The negative value of the capillary pressure \( P_{\text{cap}} \), means that the meniscus force \( F_{\text{men}} = P_{\text{cap}} \times \text{(meniscus area)} \) acts to pull the two slides together. Since the capillary pressure increases as the separation \( d \) decreases, the two slides will continue to be pulled together by the meniscus force, until either, trapped particles or surface roughness prevent them from coming any closer, or the meniscus reaches the edges of the glass microscope slides. If the meniscus is still within the edges of the slides, the final capillary pressure can be extremely high. For example, if the final separation distance is \( d = 1\,\mu\text{m} \) and \( \theta \sim 0^\circ \), then

\[
P_{\text{cap}} = -2\gamma/d \\
= -2 \times (73 \text{ mN/m})/10^{-6} \text{ m} \\
= -146 \text{ kPa} = -14.6 \text{ N/cm}^2
\] (5.8)

So, for every square centimeter of the gap filled with the water, a force of 14.6 N in the normal direction would be needed to separate the slides. (The glass usually breaks when this much force is applied.) An easier way to separate slides is to place them underwater so that the gap completely fills with water, eliminating the air–water meniscus surface and the force derived from its capillary pressure.

Figure 5.5(b) shows the case where the surfaces of the microscope slides are hydrophobic, which can be done by coating the slides with wax or bonding organic monolayers to their surfaces (Fadeev and McCarthy 2000). When the contact angle \( \theta \) is greater than \( 90^\circ \), \( r>0 \) and \( P_{\text{cap}}>0 \); since the capillary pressure is positive, the meniscus force \( F_{\text{m}} \) acts to repel the two slides, rather than to pull them together as for the hydrophilic case, and the final separation distance \( d \) occurs when the meniscus force balances the weight of the top slide. Another way to observe a repulsive meniscus is to use a hydrophobic slide as the plate in the Wilhelmy plate experiment shown in Fig. 5.2, with water as the liquid; since the water contact angle \( >90^\circ \), the meniscus will bend down rather than up, so
the normal force on the hydrophobic slide/plate acts to push it out of the water rather than pull it in as for the hydrophilic slide/plate. Water strider insects provide a more elegant example of this effect: the high contact angle of water against these insect legs generates a positive capillary pressure that provides an upward force sufficient to lift the insect body above the water surface thus enabling them to walk on water. Water pollution, like a thin oil film that lowers the surface tension, can reduce the insect’s buoyancy to the point where drowning rather walking on water occurs.

Figure 5.6 shows examples of liquid menisci in few other confined geometries along with the associated capillary pressures ($P_{\text{cap}}$):

- Figure 5.6(a) shows a liquid meniscus inside a narrow capillary tube or a cylindrically shaped pore with a circular cross section. The geometry is the same as for the capillary rise geometry shown in Fig. 5.4(b): a hemispherical meniscus with radii of curvature $r_1 = r_2 = r$, where $r$ equals inner tube radius when the liquid wets the walls of the tube (contact angle $\theta = 0^\circ$). From eq. (5.6), the capillary pressure $P_{\text{cap}} = 2\gamma/r$ and acts to pull the liquid along the capillary.

- The left side of Fig. 5.6(b) shows a liquid film coating the walls of a capillary tube or pore. For this geometry, $r_1 = r$, $r_2 = \infty$, so $P_{\text{cap}} = \gamma/r$. As the liquid film becomes thicker, for example, by increasing the vapor partial pressure to increase condensation, the liquid film becomes unstable, tending to jump the gap across the pore and form a meniscus as shown on the right side of Fig. 5.6(b). The capillary pressure (now $P_{\text{cap}} = 2\gamma/r$) of the meniscus pulls in liquid that was coating the pore surfaces to leave a much thinner film.

- Figure 5.6(c) shows a spherical particle sitting on a surface covered with a thin liquid film. The liquid from the film wicks up around the sphere to make a meniscus.

\[ P_{\text{cap}} = \frac{2\gamma}{r} \]  
\[ P_{\text{cap}} = \frac{\gamma}{r} \]  
\[ P_{\text{cap}} = \frac{\gamma}{r_1} \]

**Fig. 5.6.** (a) Liquid meniscus in a capillary or pore. (b) Left: Liquid film coating the walls of a capillary tube. Right: For thicker films, the liquid surfaces become unstable and condense to form a meniscus. (c) Spherical particle sitting on a surface with liquid film that wicks up around the particle to form a meniscus.
meniscus with radii of curvature $r_1$ and $r_2$, as defined in Fig. 5.6(c). (Note: for this geometry, the radii are opposite in sign: $r_1 < 0$ and $r_2 > 0$.) If $r_1$ and $r_2$ are comparable in size, the capillary pressure is given by eq. (5.6). For the more common situation where $|r_2| \gg |r_1|$, the capillary pressure can be approximated $P_{\text{cap}} \simeq \gamma/r_1$. This negative pressure pulls liquid into the meniscus until the capillary pressure equals the pressure in the film. The pressure in the film, which originates from the molecular interactions between the solid surface and the liquid, is called the *disjoining pressure* and is discussed in detail in Chapters 7 and 10.

### 5.2.2 The Kelvin equation and capillary condensation

One consequence of the pressure difference across a curved liquid surface is that the vapor pressure over a liquid surface depends on the degree of curvature of the surface. This leads to the phenomenon of *capillary condensation*, where vapors condense into small cracks and pores at vapor pressures significantly less than the saturation vapor pressure, as illustrated in Fig. 5.7. Capillary condensation frequently has a major impact on the tribology of contacting surfaces as, even at fairly low vapor pressures, the water and organic vapors condense into the nanometer size gaps and crevices between contacting surfaces, greatly influencing

![Fig. 5.7. Capillary condensation of water vapor between two hydrophilic rough contacting surfaces. At low humidity (bottom case), small water menisci (gray), with a small meniscus radius $r_a$ given by the Kelvin equation, form around the contacting asperities. As the humidity increases, the meniscus radius increases, causing the menisci to grow and to fill in some of the spaces between the contacting asperities (middle case). At sufficiently high humidity (top case), the meniscus radius $r_c$ is sufficiently large that the condensed water floods all the gaps between the contacting surfaces.](image-url)
their adhesion and lubrication characteristics (Binggeli and Mate 1994; Binggeli and Mate 1995).

The expression describing how vapor pressure depends on curvature is called the **Kelvin equation** and is easily derived as follows:

For the liquid, the excess Gibbs free energy of a curved liquid surface relative to a flat liquid surface is given by

\[
\Delta G = \int V_m \, dP = 2\gamma V_m / r
\]  

(5.9)

where \(V_m\) is the molar volume and using \(\Delta P = 2\gamma / r\).

For the vapor, the excess Gibbs free energy of the vapor pressure \(P\) over the curved liquid surface relative to the saturation vapor pressure \(P_s\) over the flat surface is given by

\[
\Delta G = RT \ln \left( \frac{P}{P_s} \right).
\]  

(5.10)

Equating eqs. (5.9) and (5.10) yields the **Kelvin equation**:

\[
2 \left( \frac{1}{r_1} + \frac{1}{r_2} \right)^{-1} = r_k = \frac{2\gamma V_m}{RT \ln \left( \frac{P}{P_s} \right)}
\]  

(5.11)

where \(r_k\) is called the **Kelvin radius** and corresponds to the mean radius of curvature of the condensed meniscus. Eq. (5.11) has been verified to hold for \(r_k\) as small as 4 nm for condensed organic liquids (Fisher and Israelachvili 1981) and for \(r_k\) as small as 5 nm for condensed water (Kohonen and Christenson 2000). The Kelvin equation shows that the saturation vapor pressure \(P\) over liquid surfaces that curve inward [like those shown in Figs. 5.4(b), 5.5(a), and 5.6] is lower than the bulk saturation vapor pressure \(P_s\); for those that are curved outward [Figs. 5.4(a) and 5.5(b)], condensation occurs for vapor pressures higher than \(P_s\).

5.2.2.1 Example: Capillary condensation of water in a nanosized pore

Let’s consider water condensing in a narrow pore, such as shown in Fig. 5.6(a), to form a spherical meniscus that wets the walls of the pore and has a radius of curvature \(r\). For water \(\gamma = 73 \text{ mN/m}, V_m = 18 \text{ cm}^3\), and the Kelvin equation predicts \(r = \infty\) at \(P/P_s = 1\), \(r = -10 \text{ nm}\) at \(P/P_s = 0.9\), and \(r = -1 \text{ nm}\) at \(P/P_s = 0.34\). So, for relative humidities in the range of 34 to 90%, condensation of water vapor only forms a meniscus in nanometer sized pores, that is, in those with diameters ranging 2 to 20 nm.

5.2.2.2 Example: Capillary condensation of an organic vapor at a sphere-on-flat geometry

On hydrophobic surfaces or at low relative humidities, capillary condensation of water becomes energetically unfavorable. Organic vapors, which always exist to
a certain extent in the atmosphere, more readily condense onto surfaces due to their lower surface tensions. Here, we examine the condensation of the organic vapor, hexane, around the sphere–flat contact geometry [for example, the particle on flat illustrated in Fig. 5.6(c) and for one of single asperity contacts in Fig. 5.7]. For hexane $\gamma = 18\text{ mN/m}$ and $V_m = 130\text{ cm}^3$, and assuming $|r_2| \gg |r_1|$, then the Kelvin equation predicts $r = -11\text{ nm}$ at $P/P_s = 0.9$ and $r = -1\text{ nm}$ at $P/P_s = 0.3$. So, just as for water condensation in pores, hexane and, likewise, other organic vapors condensed only in nanometer sized crevices at contacting interfaces for partial pressures much below saturation.

5.3 Interfacial energy and work of adhesion

Before discussing solid surface energies and how they differ from liquid surface tensions, it is convenient to introduce first the concepts of interfacial energy and work of adhesion.

When two materials contact each other, like materials A and B shown in Fig. 5.8, the interface between them has an energy $\gamma_{AB}$ associated with it. To better understand this energy, we first consider the work of adhesion for this interface, labeled $W_{AB}$. The work of adhesion is defined to be the change in free energy or the reversible work done per unit area to separate the two contacting materials to infinity in vacuum. For separating at a hypothetical interface within a single component material ($A = B$), the interfacial energy is defined to be zero ($\gamma_{AA} = 0$); in this case, the work of adhesion $W_{AA}$ is called the work of cohesion. Since the energy associated with creating a unit of surface area of material exposed to vacuum is $\gamma_A$ or $\gamma_B$, the work of cohesion of material A can be related to its surface energy by

$$\gamma_A = \frac{1}{2} W_{AA}. \quad (5.12)$$

For dissimilar materials contacting across an interface, the energy needed to create this interface can be divided into two steps: the energy first needed to create two surfaces A and B separated in vacuum [Fig. 5.8(a)], then the energy gained when A and B are brought into contact [Fig. 5.6(b)]. Therefore the change

![Fig. 5.8. (a) Two materials A and B when separated have surface energies $\gamma_A$ and $\gamma_B$. (b) When joined together, they have an interfacial energy $\gamma_{AB}$.](image-url)
in total free energy $\gamma_{AB}$ to create this interface is
\[
\gamma_{AB} = \frac{1}{2} W_{AA} + \frac{1}{2} W_{BB} - W_{AB} = \gamma_A + \gamma_B - W_{AB}, \tag{5.13}
\]
which is called the Dupré equation. The Dupré equation can be extended to the separation of materials A and B in a third medium, say a liquid or vapor C:
\[
\gamma_{AB} = \gamma_{AC} + \gamma_{BC} - W_{ACB} \tag{5.14}
\]
where $W_{ACB}$ is the work of adhesion between materials A and B in medium C.

Equations (5.13) and (5.14) are fairly general, where materials A and B can be either solids or liquids. Next, we consider the special case where one is solid, which we label S, and the other is a liquid, which we label L. For the case of a solid–liquid interface in the presence of a vapor V, the Dupré equation (5.14) is written as
\[
\gamma_{SL} = \gamma_{SV} + \gamma_{LV} - W_{SVL}. \tag{5.15}
\]
For future convenience, we will drop the V subscripts so that eq. (5.15) becomes
\[
\gamma_{SL} = \gamma_S + \gamma_L - W_{SL}. \tag{5.16}
\]
The vapor is assumed still to be present and can have a major impact on these surface energies as, over time, vapor molecules like water and hydrocarbons may absorb on to otherwise clean surfaces, lowering their surface energies.

5.4 Surface Energy of Solids

5.4.1 Why solids are not like liquids

For solid surfaces, the concept of surface energy presents some difficulties. First, solid surfaces are rarely uniform or equilibrated to their lowest energy state. Second, when a force is applied to change a solid’s shape then released, plastic deformation or viscous forces often prevent it from rapidly reverting back into its original shape.

For liquids, the concept of surface energy is a fairly straightforward: When the surface area of a liquid increases, a certain amount of work has to be expended to move liquid atoms or molecules from the bulk to their new equilibrium positions on the surface; the amount of work divided by the newly created surface area equals the surface energy. This work is reversible: when the liquid surface area is reduced back to its original area, the liquid does work on its surroundings equal to the energy gained from the loss of surface area; the energy comes from the liquid atoms or molecules moving from the higher energy surface sites to lower energy bulk configurations. For most liquids, the movement back and forth between the surface and bulk positions happens over a very short distance, a few atomic or molecular diameters; consequently, for low viscosity liquids, the expenditure or release of surface energy happens quickly. Therefore, the surface tension of liquids does not change as the surface area changes ($\partial \gamma / \partial A = 0$).
For solids, the situation is quite different, as, by definition of a solid, the atoms and molecules do not flow when a stress is applied, unless the stress exceeds the threshold for plastic yield. So, if we apply a tensile stress to a solid bar, the atomic bonds stretch, the surface area becomes slightly larger. If we stay in the elastic regime, the atoms stay in their positions relative to each other, and the work expended stretching the atomic bonds is recovered when the atoms move back to their original positions when the stress is released (the work is reversible). If the applied stress exceeds the solid’s plasticity threshold, the solid atoms “flow” through the movement of dislocations and the sliding of slip planes as described in section 3.1. During plastic flow, some of the energy applied to deforming the solid goes to moving atoms to and from the surface, changing the surface area. For solids, unlike liquids, the surface energy does not remain constant as we change surface area ($\partial \gamma / \partial A \neq 0$).

In general, the energy needed to increase the surface area of a solid depends greatly on how it was created and the history of the surface. For example, the energy needed to cleave a crystalline solid along one of its crystal planes to create a unit surface area is much lower than creating the same surface area by plastically deforming the crystal. For the cleaved surface, the energy expended simply breaks the bonds across the cleaved crystal plane so that the surface atom positions terminate the bulk crystal structure. However, when plastic flow increases the surface area (for example, when a metal bar is stretched past its yield point as illustrated in Fig. 3.8), the movement of atoms takes place mainly through the movement of dislocations. These dislocations first have to be created, which requires energy to move the atoms out of their energetically favorable positions in the crystal lattice. In addition to the energy expended breaking bonds to generate surface atoms, energy has also to be expended to create and move the dislocations responsible for plastic flow. A surface formed by plastic deformation has a significant number of dislocations, steps, and other defects left over from the deformation process, which have higher formation energy than the surface atoms that simply terminate the bulk crystal structure. The fraction of the surface covered with these higher energy sites depends on the history of the deformation process, so the surface energy also depends on this history.

Once a solid surface has been created, by cleavage, plastic deformation, or some other process, it is not necessarily stable and may relax to a lower energy configuration, if one exists, as follows.

For cleaved surfaces, the termination of the bulk structure is not the lowest energy configuration for the surface atoms (Somorjai 1994). Instead, the spacing in the first and second layer is found to be significantly smaller than the bulk spacing, as the loss of atom bonding on one side of these surface atoms leads to stronger bonding and shorter bond lengths for the atoms on the other side. These same forces often further reconstruct by moving surface atoms laterally relative to the bulk positions. This is especially true for covalently bonded solids like semiconductors, as the dangling bonds created when the surface is cleaved
can not be easily satisfied except through a drastic rearrangement of the surface atoms.

For dislocations and defects, these high energy surface sites often anneal out at high temperatures through the diffusion of surface and bulk atoms. For two rough surfaces pressed together, the solid–solid contact area will slowly grow as atoms diffuse to new low energy sites (discussed in section 4.3 and illustrated in Fig. 4.8).

The surface energy also changes due to the absorption of molecules from the environment. Solid surfaces bare of any adsorbed molecules occur only in very specialized situations, such as inside an ultra-high vacuum chamber or in outer space, where a low probability exists for molecules striking the surface. Most surfaces are exposed to gas or liquid environments where molecules are impinging all the time. Those impinging species that lower the surface energy through the formation of a chemical bond are energetically favorable to adsorb onto the surface.

5.4.2 Experimental determination of a solid’s surface energy

Experimental determination of a solid’s surface energy is problematic. Since the surface energy depends on the history of the surface and can potentially change over time, one can not simply measure it once and log it in a table for future reference, as for liquid surfaces. A key difference is that solids do not readily flow when a stress is applied; instead elastic stresses are generated within the solid bulk in addition to those from the surface tension. Consequently, surface energy measurements involving stretching of the surface (which works well for liquids) are difficult to apply to solids unless done at high enough temperatures to promote rapid atomic diffusion and by stretching the surface area at an exceptionally slow rate to relieve the bulk stress. The following are some of the experimental methods that are used for estimating the surface energies of solids (Bikerman 1970; Overbeek 1971–1974; Adamson and Gast 1997).

- **Cleavage of brittle crystals** – Used for brittle materials. The value of $\gamma$ is determined using eq. (5.12), from the work done cleaving a single crystal to expose the two crystal faces on opposite sides of the cleavage plane. One way of measuring the work of adhesion is illustrated in Fig. 5.9. A more sophisticated way of doing this is to use the surface force apparatus (SFA), discussed in a future chapter. For example, the surface free energy of mica surfaces has been measured using SFA to be a few thousand ergs/cm$^2$ in vacuum (Obreimoff 1930; Bryant 1962), 413 ergs/cm$^2$ in dry nitrogen, and 185 ergs/cm$^2$ in laboratory air (Frantz et al. 1997; Frantz and Salmeron 1998), which shows the effect that airborne contaminations, such as water (Christenson 1993), have on lowering surface energies.

- **Calorimetry** – Used for powders. When a crystal dissolves, the heat of solution is determined from the temperature rise. When a powder of small crystallites is dissolved, an additional heat is given off from the loss of surface energy. The
surface energy is determined by dividing the extra heat by the powder surface area, which has to be determined by a separate method.

- **Contact angles** – Used for those low energy solids, where the solid surface energy is less than the surface tension of a test liquid. Discussed in detail in the next section.

Experimental values for solid surface energies should be viewed as more qualitative than quantitative, given all the difficulties in achieving equilibrated, non-hysteretic results, and since surface energies of solids can change over time.

### 5.4.2.1 Contact angles

A common way of estimating solid surface energies is to measure the contact angle that a liquid makes with its surface. This technique is generally applied only to low energy solid surfaces, as most available liquids like water and hydrocarbons have relatively low surface tensions and wet solids with higher surface energies. The contact angle method uses the ability of a liquid to flow in response to the tension generated by the differences in surface energies at the solid–liquid contact line. Our discussion of contact angles will also help illustrate the concepts of surface energy, work of adhesion, and energy hysteresis.

Let’s start by considering the edge of a liquid droplet shown in Fig. 5.10 advancing over a flat solid surface in the presence of its own vapor. When the droplet moves a distance $ds$, it covers, per unit length of the droplet edge, an area $ds$ on the solid and creates an area $ds \cos \theta$ of fresh liquid surface. As the droplet edge advances, the contact angle $\theta$ decreases, due to the finite volume of the droplet, and the edge stops advancing when $\theta$ is such that the surface energy lost at the solid–vapor interface equals the energy expended to create the new areas at the liquid–vapor and liquid–solid interfaces. This corresponds to an equilibrium contact angle $\theta$ satisfying the Young’s equation:

$$\gamma_{SL}ds + \gamma_Lds \cos \theta - \gamma_Sds = 0$$ (5.17)
Fig. 5.10. Edge of liquid droplet on a solid surface, with a contact angle $\theta$.

$$\cos \theta = \frac{\gamma_S - \gamma_{SL}}{\gamma_L}. \quad (5.18)$$

So, if the liquid surface tension $\gamma_L$ is higher than the difference between the solid surface energy and the interfacial energy ($\gamma_L > \gamma_S - \gamma_{SL}$), the liquid droplet has a finite contact angle; while, if $\gamma_L < \gamma_S - \gamma_{SL}$, the liquid spreads over the solid surface. Rearranging Young’s equation yields an expression for the solid–vapor surface energy:

$$\gamma_S = \gamma_L \cos \theta + \gamma_{SL}. \quad (5.19)$$

Therefore, $\gamma_S$ can be determined from eq. (5.19) if the three quantities $\gamma_L$, $\theta$, and $\gamma_{SL}$ are known. The liquid surface tension $\gamma_L$ can be determined using the Wilhelmy plate method. The contact angle $\theta$ can be measured by placing a droplet on the solid surface as described below. The challenge comes in determining or estimating the solid–liquid interface energy $\gamma_{SL}$.

5.4.2.1.1 Estimating interfacial energies Before discussing ways of estimating $\gamma_{SL}$, we need to first notice that the two unknowns in eq. (5.19)—$\gamma_S$ and $\gamma_{SL}$—can be reduced to one unknown by combining eq. (5.19) with the Dupré equation (5.15) to form the Young–Dupré equation:

$$W_{SL} = \gamma_L (1 + \cos \theta). \quad (5.20)$$

Therefore, the work of adhesion $W_{SL}$ between a liquid and solid can be obtained by measuring the contact angle if the value of $W_{SL}$ is between 0 and $2\gamma_L$.

5.4.2.1.1 Good–Girifalco method Using combining rules developed from intermolecular interactions (Israelachvili 1991), the solid–liquid work of adhesion $W_{SL}$ can be estimated in terms of $\gamma_S$ and $\gamma_L$, after which eq. (5.20) can be used to solve for $\gamma_S$. One combining rule approach developed by Good and Girifalco is to estimate $W_{SL}$ as a geometric mean of $\gamma_S$ and $\gamma_L$ (Girifalco and Good 1957; Girifalco and Good 1960):

$$W_{SL} = 2\Phi \sqrt{\gamma_S \gamma_L} \quad (5.21)$$
where $\Phi$ is a correction factor related to the mismatch in intermolecular interactions. If the intermolecular interactions are similar (for example, a hydrocarbon liquid against a hydrocarbon solid interacting solely via dispersion forces), $\Phi$ is close to unity. If the interactions are dissimilar (for example, a hydrogen bonding liquid like water against a pure hydrocarbon solid), $\Phi$ is significantly less than unity, but usually in the range of 0.5 to 1.

5.4.2.1.2 Fowkes’ method Another combining rule approach, developed by Fowkes, is first to express the surface energy as a sum of components corresponding to specific types of intermolecular interactions (Fowkes 1964):

$$\gamma = \gamma^d + \gamma^p + \gamma^i + \cdots$$  \hfill (5.22)

where $d$, $p$, and $i$ superscripts stand for, respectively, dispersion, polar, and induction components, and the dots represents other, less significant terms (these molecular interactions are discussed in more detail in Chapter 7). Since dispersion and polar terms typically dominate over the other surface energy terms, the work of adhesion can be expressed [using eq. (5.21) with $\Phi = 1$] as

$$W_{SL} \simeq 2\sqrt{\gamma^d_S \gamma^d_L} + 2\sqrt{\gamma^p_S \gamma^p_L}.$$  \hfill (5.23)

The procedure (Waltman et al. 2003) for using the Fowkes analysis to determine a solid’s surface energy $\gamma_S$ is first to determine its dispersive component $\gamma^d_S$. This is done by using a liquid, such as an alkane, that interacts with other materials predominately via the dispersive van der Waals forces ($\gamma_{\text{alkane}} \simeq \gamma^d_{\text{alkane}}$) to measure the contact angle $\theta_{\text{alkane}}$ of that liquid against the solid in question. This contact angle and surface tension are then used in a combination of eqs. (5.20) and (5.23) to yield $\gamma^d_S$:

$$\gamma^d_S = \gamma^d_{\text{alkane}} \left( 1 + \cos \theta_{\text{alkane}} \right)^2.$$  \hfill (5.24)

Next, the contact angle is measured for a liquid with known values of $\gamma^d_L$ and $\gamma^p_L$ where $\gamma^p_L$ is a substantial fraction of the total liquid surface tension, such as for water where $\gamma^d_L = 20$ mN/m and $\gamma^p_L = 53$ mN/m. Using this contact angle for the polar liquid, $\gamma_S$ can be determined by substituting into Young’s equation the following expression for the solid–liquid interfacial energy (combination of eqs. (5.16) and (5.23)):

$$\gamma_{SL} = \gamma_S + \gamma_L - 2\sqrt{\gamma^d_S \gamma^d_L} - 2\sqrt{\gamma^p_S \gamma^p_L}$$  \hfill (5.25)

and solving for $\gamma_S$.

5.4.2.1.2 Zisman method for estimating surface energy for a solid Another widely used method for estimating a solid–vapor surface energy from contact angle measurements is the empirical method developed by Zisman and
co-workers (Shafrin and Zisman 1960; Fowkes and Zisman 1964). They noticed that, for a homologous series of liquids, the cosine of the contact angle could be described by

\[ \cos \theta = a + b \gamma_L = 1 - \beta (\gamma_L - \gamma_c) \]  

(5.26)

and that the intercept \( \gamma_c \) (where \( \cos \theta = 1 \)) is fairly independent of the series used. Zisman called \( \gamma_c \) the critical surface tension as \( \gamma_c \) characterizes the surface tension of a liquid that just wets the solid surface. Using the Good–Girifalco analysis, \( \gamma_c \) is related to \( \gamma_S \) by \( \gamma_c = \Phi^2 \gamma_S \); so, the critical surface tension equals the solid surface energy when \( \Phi = 1 \), such as when the liquid–solid interactions are primarily dispersive. Since alkanes interact with other materials primarily through dispersive forces, they are frequently chosen as the homologous series so that \( \gamma_c \approx \gamma_S \). Figure 5.11 shows an example of the Zisman analysis: \( \cos \theta \) is plotted versus \( \gamma_L \) for a series of alkanes on the surface of a disk from a hard disk drive. The line through the data points extrapolates to \( \cos \theta = 1 \) at \( \gamma_c = 17.9 \text{ mN/m} \) for the disk surface indicating a value \( \gamma_S \approx 17.9 \text{ mN/m} \) for this surface.

One of the key findings of Zisman’s work is that the critical surface tension varies in a systematic way with chemical composition or structure of solid

![Fig. 5.11. Zisman plot showing the linear regression of contact angle cosine versus surface tension \( \gamma_L \) of alkane droplets on a disk from a hard drive. The disk has a 3.8 Å thick film of perfluoropolyether lubricant. Reprinted with permission from Wu and Mate (1998). Copyright 1998 American Chemical Society.](image-url)

<table>
<thead>
<tr>
<th>Surface constitution</th>
<th>$\gamma_c$ at 20° C (mN/m)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fluorocarbon surfaces</strong></td>
<td></td>
</tr>
<tr>
<td>$-$CF$_3$</td>
<td>6</td>
</tr>
<tr>
<td>$-$CF$_2$H</td>
<td>15</td>
</tr>
<tr>
<td>$-$CF$_3$ and $-$CF$_2$-</td>
<td>17</td>
</tr>
<tr>
<td>$-$CF$_2$-</td>
<td>18</td>
</tr>
<tr>
<td>$-$CH$_2$–CF$_3$</td>
<td>20</td>
</tr>
<tr>
<td>$-$CF$_2$–CFH–</td>
<td>22</td>
</tr>
<tr>
<td>$-$CF$_2$–CH$_2$–</td>
<td>25</td>
</tr>
<tr>
<td>$-$CFH–CH$_2$–</td>
<td>28</td>
</tr>
<tr>
<td><strong>Hydrocarbon surfaces</strong></td>
<td></td>
</tr>
<tr>
<td>$-$CH$_3$ (crystal)</td>
<td>20–22</td>
</tr>
<tr>
<td>$-$CH$_3$ (monolayer)</td>
<td>22–24</td>
</tr>
<tr>
<td>$-$CH$_2$-</td>
<td>31</td>
</tr>
<tr>
<td>$-$CH$_2$– and $=$CH– phenyl ring edge</td>
<td>33</td>
</tr>
<tr>
<td><strong>Chlorocarbon surfaces</strong></td>
<td></td>
</tr>
<tr>
<td>$-$CClH–CH$_2$–</td>
<td>39</td>
</tr>
<tr>
<td>$-$CCl$_2$–CH$_2$–</td>
<td>40</td>
</tr>
<tr>
<td>$=$CCl$_2$</td>
<td>43</td>
</tr>
<tr>
<td><strong>Nitratated hydrocarbon surfaces</strong></td>
<td></td>
</tr>
<tr>
<td>$-$CH$_2$ONO$_2$ (crystal)</td>
<td>40</td>
</tr>
<tr>
<td>$-$C(NO$_2$)$_3$ (monolayer)</td>
<td>42</td>
</tr>
<tr>
<td>$-$CH$_2$NHNO$_2$ (crystal)</td>
<td>44</td>
</tr>
<tr>
<td>$-$CH$_2$ONO$_2$ (crystal)</td>
<td>45</td>
</tr>
</tbody>
</table>

surfaces. Table 5.1 shows how the critical surface tensions vary with the chemical structure of low energy surfaces. For example, $-$CF$_2$- covered surfaces are found to have $\gamma_c = 18$ mN/m, the same value as was found for the disk surface in Fig. 5.11, indicating that the disk surface may be covered predominately with $-$CF$_2$- units; this is to be expected since the disk is covered by a 3.8 Å thick film of perfluoropolyether lubricant consisting of mostly $-$CF$_2$- units. Why some molecular surface structures have lower surface energy than others is still a very active research topic, and the reader is referred to Chaudhury (1996) for a review of the current understanding.

Such a correlation between the surface chemical structure and surface energy is very useful when it comes to designing surfaces for tribological applications. For example, if lowest surface energy is desired, to help minimize adhesive and friction forces, then the results of Table 5.1 indicate that it is better to cover the surface
Fig. 5.12. When a droplet (above) is placed on a surface three types of wetting can occur: (a) wetting, (b) partial wetting, and (c) non-wetting—a subcategory of partial wetting where $\theta > 90^\circ$.

with fluorinated molecules, such as the perfluoropolyether lubricants used for disk drives, rather than hydrocarbons, chlorocarbons, and nitrated hydrocarbons.

5.4.2.1.3 Types of wetting  When a liquid droplet is placed on a surface, as shown in Fig. 5.12, three types of wetting can be defined:

- **Wetting** – The contact angle $\theta$ between the liquid and solid is zero or so close to zero that the liquid spreads over the solid. Spreading occurs when the work of adhesion between the liquid and solid is equal to or greater than the liquid’s internal cohesion energy.
- **Partial wetting** – Contact angle $\theta > 0$.
- **Non-wetting** – A subset of partial wetting where the contact angle $\theta > 90^\circ$. While, thermodynamically, a contact angle $\theta$ of $90^\circ$ has no particular significance, above this angle, liquids tend to ball up and easily roll off the surface if the contact angle hysteresis is low. Also, for $\theta > 90^\circ$, $\gamma_{SL} > \gamma_S$ indicating that it takes more energy to create a solid–liquid interface than is gained from the loss of solid–vapor interface.

Surfaces may be described as:

- **Hydrophilic** – The surface has an affinity for water and a low water contact angle.
- **Hydrophobic** – The surface has a repulsion for water and a high water contact angle.

5.4.2.1.4 Contact angle measurements  The most common way to measure a contact angle $\theta$ is to examine the edge of a droplet resting on a flat surface, as shown in Fig. 5.13. Typically, a syringe is used to place the droplet on the surface, and a short working distance telescope is used to provide a magnified view of the droplet edge. The contact angle is measured either by visual examination through
the telescope eyepiece with a goniometer scale or by capturing the image with a video camera and determining $\theta$ with computer software. Sometimes the droplet is enclosed so that the air around the droplet becomes saturated with the liquid’s vapor; this serves two purposes: (1) it prevents the droplet from shrinking due to evaporation, and (2) it ensures that the $\gamma_S$ being measured is at equilibrium with the liquid vapor.

For the droplet method in Fig. 5.13, the simplest way of placing the droplet on the surface is to let it drop a small height from the syringe. Another common approach is to position the syringe with its end in contact with the droplet, as illustrated in Fig. 5.14; this has the advantage that both the *advancing* and *receding* contact angles can be measured by, respectively, swelling and shrinking the droplet on the surface.
Table 5.2. Advancing and receding contact angles measured by the author for water on popular windshield coatings deposited onto glass microscope slides.

<table>
<thead>
<tr>
<th>Coating</th>
<th>Water contact angle (°)</th>
<th>$\gamma_{water} \times (\cos \theta_r - \cos \theta_a)$ (mN/m)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Advancing</td>
<td>Receding</td>
</tr>
<tr>
<td>Bare glass</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>Rain Slick</td>
<td>64</td>
<td>38</td>
</tr>
<tr>
<td>AccuVision</td>
<td>85</td>
<td>48</td>
</tr>
<tr>
<td>Rain Shield</td>
<td>87</td>
<td>58</td>
</tr>
<tr>
<td>Rain-X</td>
<td>91</td>
<td>77</td>
</tr>
</tbody>
</table>

5.4.2.1.5 Contact angle hysteresis If the droplet shape is in true thermodynamic equilibrium with the solid surface, the advancing and receding contact angles have the same value given by Young’s equation (5.18). For most situations, however, the advancing angle is significantly larger than the receding angle. A practical consequence of this hysteresis is that droplets are pinned to the surface until a sufficient shear force is applied to the droplet to initiate motion. For example, rain droplets remain pinned to a car’s windshield due to contact angle hysteresis until a sufficient wind shear force is generated to move the droplet across the windshield. (At most safe speeds, windshield wipers are still needed to remove droplets, particularly the smaller ones.) Table 5.2 shows the advancing and receding water contact angles on several popular windshield coatings sold to reduce the total pinning force for rain droplets on windshields. The last column of Table 5.2 gives the force per unit length of the drop perimeter needed to initiate motion, calculated using

$$F = \gamma_L (\cos \theta_r - \cos \theta_a)$$  \hspace{1cm} (5.27)

where $\theta_r$ and $\theta_a$ are, respectively, the receding and advancing contact angles (Wolfram and Faust 1978). While the force per unit length of perimeter actually goes up when the treatments are applied to the glass surfaces, these coatings are effective at reducing the total pinning force since the increased water contact angle leads to a much smaller droplet footprint on the glass and a corresponding smaller total perimeter length, which results in a lower total force.

When the advancing and receding angles differ, many researchers prefer to use the advancing contact for determining solid surface energies, as the solid surface is not yet been altered by having liquid flowing over it in this case. However, the larger the difference between the advancing and receding contact angles becomes, the more questionable is the use of contact angle measurements for determining surface energies, as the magnitude of the hysteresis measures to
what extent the solid–liquid interface does not retrace its original path when the droplet recedes, i.e., the extent to which the advancing and receding processes are not thermodynamically reversible.

Typically, hysteresis implies that a physical system is trapped in a metastable, non-equilibrium state. For contact angles, the origin of hysteresis is unclear in most situations, which has led to extensive debates in the scientific literature (Miller and Neogi 1985; Israelachvili 1991; Adamson and Gast 1997). Multiple mechanisms have been suggested to explain why the edge of a liquid droplet becomes trapped in a metastable state (Fig. 5.15):

1. **Surface roughness** – Even though the microscopic contact angle is constant, the macroscopic contact angle can vary depending on the slope of the local roughness [Fig. 5.15(a)]. For advancing liquids, slopes that provide large contact angles are favored, while, for receding liquids, slopes that provide small angles are favored. (Miller and Neogi 1985; Chen et al. 1999)

2. **Chemical heterogeneity** – The solid–liquid interface is not chemically uniform, but instead has a work of adhesion $W$ that varies substantially on length scales much smaller than the droplet diameter. From the Young–Dupré equation (5.20), regions with a relatively low work of adhesion have higher microscopic contact angles. An advancing edge of a liquid droplet catches on these regions while a receding edge catches on the regions of high work of adhesion [Fig. 5.15(b)].

3. **Molecular rearrangement** – Molecules on a solid surface may adopt a different orientation or conformation depending on whether they are exposed to liquid or vapor. If the timescale for the molecule to change its orientation is slow compared with the rate at which the edge of the liquid moves over the surface, and the adhesion energy varies with molecular orientation, this will result in contact angle hysteresis [Fig. 5.15(c)]. For example, Hennig et al. (2004) have measured a $35^\circ$ difference between advancing and receding contact angles for water on polyimide surfaces, which they attribute to water swelling the polymer.

4. **Contamination** – Solids with high surface energies tend to become contaminated with a layer of lower surface energy material such as adsorbed hydrocarbon vapors. The lower surface energy layer results in a higher contact angle for an advancing liquid than for the initially clean surface. If the liquid dissolves the contamination layer as it advances, the contact angle will be lower when the liquid recedes on the now cleaner, higher energy surface [Fig. 5.15(d)].

### 5.4.3 Adhesion hysteresis

Contact angle hysteresis is a subcategory of the more general phenomenon of adhesion hysteresis: Under most realistic conditions, the work expended to separate two surfaces is greater than the work originally gained when bringing them
together. Besides contact angle hysteresis, adhesion hysteresis manifests itself in other tribology phenomena where energy is dissipated during adhesion, such as during loading–unloading cycles and friction.

When we introduced the concept of work of adhesion in section 5.3, a basic assumption was that the contact process occurs under ideal equilibrium and thermodynamically reversible conditions, meaning that the work done to bring two surfaces together equals the work needed to separate them. When real
surfaces are brought together, however, equilibrium conditions are rarely initially achieved; instead, the atoms and molecules at the contacting interfaces slowly diffuse and reorient themselves in an effort eventually to achieve an equilibrium configuration. One such way that molecules move into the contacting interface is by the capillary condensation of water and organic vapors (as discussed in Section 5.2.2). Figure 5.16 illustrates some of the potential movement of atoms and molecules around a contacting asperity between two rough surfaces. By the time the surfaces are pulled apart, the atoms and molecules have moved or condensed to energetically more favorable sites, requiring more work to disrupt the new bonding geometries formed during contact, increasing the work of adhesion. In Section 4.3, we discussed how this same process leads to friction force increasing with the contact time between two surfaces.

Adhesion hysteresis can also occur for perfectly flat and chemically homogeneous surfaces supported by elastic materials. This type of adhesion hysteresis contributes to “rolling” friction and elastoplastic adhesive contacts during loading–unloading cycles (Bowden and Tabor 1967; Greenwood and Johnson
Figure 5.17 illustrates how the movement, reorientation, interdigitation etc. of atoms and molecules can, over time, transform an initially smooth, homogeneous interface into a rougher interface with chemical heterogeneity.

For surfaces covered with polymer or surfactant films, interdigitation of molecules on the opposing surfaces is suspected of being a major cause of the adhesion hysteresis observed for these surfaces, even if they are nominally smooth and chemically homogeneous. For example, Israelachvili and co-workers have studied the connection between adhesion hysteresis and interdigitation for thin surfactant layers on atomically smooth mica surfaces brought together in a surface force apparatus (SFA) (Chen et al. 1991; Yoshizawa et al. 1993; Yoshizawa et al. 1993; Israelachvili et al. 1994; Yoshizawa and Israelachvili 1994). They have observed that the adhesion hysteresis is strongly correlated to the phases of the surfactant layers, which are classified as crystalline, amorphous glass, or liquid-like, depending on the degree of mobility and order within the film. For both crystalline and liquid-like monolayers, negligible hysteresis was observed by the Israelachvili group, while, for monolayers with an amorphous glass phase, significant hysteresis was observed. To explain these results, Israelachvili has proposed that the alkyl chains within a crystalline surfactant layer are too rigid to interdigitate, so exhibit no hysteresis. For liquid-like layers, the alkyl chains interdigitate quickly during loading and unloading cycles, enabling them to be nearly always at equilibrium conditions, resulting in negligible hysteresis. For the amorphous glassy layers, the time scale for molecular reorganization of the alkyl chains is intermediate compared with that of the crystalline and liquid-like layers, so the molecular interdigitation process has a significant “relaxation” time after coming into contact and after separation. Figure 5.18 illustrates the degree of interdigitation for these three types of surfaces.

Chaudhury and Owen have also found that adhesion hysteresis depends on the phase of monolayer surfactant films deposited onto a PDMS polymer substrate: hysteresis is significant for crystalline monolayers, negligible for liquid-like monolayers, and significant for submonolayer films (Chaudhury and Owen 1993). Ruths and Granick have studied the rate dependence of the adhesion hysteresis for various polymers adsorbed on the smooth mica surfaces (Ruths and Granick 1998a; 1998b). They found that the difference in the work of adhesion measured on approach and separation increases logarithmically with increasing contact time or decreasing separation rate; they suggest that more interdigitation occurs between the polymers on the opposing surfaces, the longer the surfaces remain in contact, leading to more hysteresis.

While chemical effects, like those described above, are a major source of adhesion hysteresis, the mechanics of adhesion measurements is also a major source of hysteresis in experimental values for the work of adhesion, as pointed out by Israelachvili (Chen et al. 1991). In a typical experiment (illustrated in Fig. 5.19), the work of adhesion is determined by integrating the force \( F \) acting normally
Fig. 5.18. Schematic of the three classifications of molecular surface layers and the resulting influence on adhesion. (A) Solid-like or crystalline. (B) Amorphous. (C) Liquid-like. Reprinted with permission from Yoshizawa et al. (1993). Copyright 1993 American Chemical Society.

to the surfaces over the separation distance $D$:

$$W = \int F dD$$

where the integral is over the path that the atoms take during approach and separation. Since the force $F$ is never measured at the surfaces themselves ($S$ in Fig. 5.19), but rather some distance away ($S'$), the force is transmitted through the elastic material in between points $S$ and $S'$, with effective spring constant $K_s$ [Fig. 5.19 (Middle, left)]. As the two surfaces approach along the force vs. distance curve shown in Fig. 5.19 (Bottom), at some separation distance $D_A$, the surfaces suddenly jump into contact when the gradient of the attractive force is greater than the material spring constant $K_s$. As one tries to separate the surfaces, one has to apply a force large enough to overcome the attractive adhesive force. Again the material elasticity creates an instability on separation at $D_s$, and the surfaces jump apart out to $D_R$. On the atomic level, the sudden separation of two surfaces involves the spontaneous breaking of bonds as illustrated in Fig. 5.19 (Middle, right).
Fig. 5.19. Origin of adhesion hysteresis during approach and separation for two solid surfaces. (Top; and Middle, left) The force is not measured at the surfaces $S$, but rather at some distance back, at $S'$ with the elasticity of the material in between providing an effective spring constant $K_s$. (Middle, right) Bonds breaking at the atomic and molecular level during separation. (Bottom) Force versus separation distance during approach and separation of two attracting surfaces. On approach, the attractive force overwhelms the spring constant at point $D = D_A$, and the two surfaces jump into contact. A similar instability occurs at $D = D_s$ when the surfaces are pulled apart. Reprinted with permission from Chen et al. (1991). Copyright 1991 American Chemical Society.
From the force vs. distance curve in Fig. 5.19, we see that the surfaces do not follow the same path when they separate as when they approach: the shaded area in Fig. 5.19 (Middle, left) corresponds to extra work that needs to be added to the system complete the cycle and separation, i.e., the adhesion hysteresis $\Delta W$.

5.5 References


SURFACE FORCES DERIVED FROM SURFACE ENERGIES

While surface energy is a valuable concept for understanding surfaces, in tribology, it is often more important to know the forces between contacting surfaces rather than the interaction energy. For macroscopic objects, forces like gravity tend to be more significant than the attractive forces between contacting surfaces arising from molecular level interactions. At the micro- or nanoscale, however, the situation becomes reversed, as the attractive surface forces scale with the radius of curvature, while gravity and inertia forces scale with the object’s volume. So, these attractive surface forces, which may seem inconsequential in our everyday experience, dominate phenomena occurring at the small scale.

Since force and energy are often related, it is natural to expect that forces at surfaces will be related to surface energies. In the last chapter, we discussed briefly several examples of forces related to surface energy and surface tension:

- Wilhelmy plate technique – Fig. 5.2
- capillary force – Fig. 5.5
- work of adhesion – Fig. 5.19

In this chapter, we go into greater detail on these surface forces derived from surface energies.

6.1 The Derjaguin approximation

The Derjaguin approximation provides an elegant way to relate the force between surfaces to the interaction energy. It can also be used to relate the forces encountered in one contact geometry to other types of contact geometries.

In the Derjaguin approximation, the force law $F(D)$ between two curved surfaces is expressed in terms of the interaction energy per unit area $W(D)$ of two planar surfaces. The approximation being made is that the separation distance $D$ and the range of the force interaction are both much less than the radii of curvature of the two surfaces. Within this approximation, it is straightforward to show that (Derjaguin 1934; Israelachvili 1991 pp. 161–164), for two spheres with radii $R_1$ and $R_2$ and for any force law that is a function of separation distance:

$$F(D)_{\text{sphere/sphere}} = 2\pi \left( \frac{R_1 R_2}{R_1 + R_2} \right) W(D).$$

(6.1)
So, if we know $W$, the interaction energy per unit area in a planar geometry with particular separation $D$, either from calculation or some energy measurement, we can determine the force $F$ between two spheres with radii $R_1$ and $R_2$ with the same separation distance $D$.

A special case of eq. (6.1) is the sphere-on-flat geometry ($R_1 = R$, $R_2 = \infty$), where $F(D)$ becomes

$$F(D)_{\text{sphere/flat}} = 2\pi RW(D). \quad (6.2)$$

Similarly, for two cylinders with radii $R_1$ and $R_2$ and crossed at an angle $\theta$ to each other:

$$F(D)_{\text{cylinder/cylinder}} = 2\pi \sqrt{R_1 R_2} W(D)/\sin \theta. \quad (6.3)$$

Eq. (6.3) is frequently used with the surface force apparatus (SFA) technique to determine, as a function of their separation distance, the interfacial energy $W(D)$ between two surfaces from the measured force $F(D)$. In a typical SFA experiment, two cylindrically shaped surfaces are crossed at a right angle ($\theta = 90^\circ$) and $R_1 \sim R_2 = R$, so eq. (6.3) reduces to eq. (6.2), indicating that the SFA cross cylinder geometry is equivalent to the sphere-on-flat geometry.

The Derjaguin approximation greatly facilitates the comparison of theories of interaction energy between parallel surfaces with experiments involving interaction forces between curved surfaces. While a geometry with two flat, parallel surfaces is convenient for theoretical analysis, this geometry is generally impractical from an experimental point of view as it is extremely difficult to align two surfaces so that they are perfectly parallel. Instead, it is far easier to measure forces between curved surfaces, such as between two spheres, a sphere and a flat, or crossed cylinders, all of which generate a well defined point of contact, and then use the Derjaguin approximation to relate the experimental forces to theoretical energies for parallel surfaces.

### 6.2 Dry environment

#### 6.2.1 Force between a sphere and a flat

As an example of the use of the Derjaguin approximation to determine a force from surface energy, let’s consider a rigid sphere of material A sitting on a rigid flat surface of material B, as shown in Fig. 6.1(a). For this case, the sphere and flat are in a dry environment (no liquids are present to form a meniscus around the contact point). If we assume that the surface roughness is negligible (i.e., $D \sim$ an atomic distance), then $W(D)$ becomes the work of adhesion $W_{AB}$. Using the Derjaguin approximation as expressed in eq. (6.2), the adhesion force $L_{\text{adh}}$ acting between the sphere and flat is related to the work of adhesion by

$$L_{\text{adh}} = 2\pi RW_{AB}. \quad (6.4)$$

The symbol $L_{\text{adh}}$ is used here for the adhesion force as $L_{\text{adh}}$ acts like a loading force pushing the sphere toward the flat. If the sphere and flat have the same
surface energy ($\gamma_S = \gamma_A = \gamma_B$), then $W_{AB} = 2\gamma_S$ and the adhesive force becomes

$$L_{\text{adh}} = 4\pi R\gamma_S.$$  \hfill (6.5)

6.2.1.1 Example: Adhesion force between two polystyrene spheres

Let’s calculate the adhesion force between two 1 $\mu$m diameter polystyrene spheres contacting each other. Using the Derjaguin approximation, eq. (6.1) the force between two spheres becomes, using $R_1 = R_2 = R$ and $W = \gamma_{\text{polystyrene}},$

$$L_{\text{adh}} = 2\pi R\gamma_{\text{polystyrene}}.$$  \hfill (6.6)

so the sphere–sphere adhesion is half that for a sphere-on-flat. With $\gamma_{\text{polystyrene}} = 33$ mN/m,

$$L_{\text{adh}} = 2\pi (0.5 \mu m) \times (33 \text{ mN/m})$$

$$= 0.1 \mu N.$$  \hfill (6.7)

6.2.1.2 Example: Adhesion force between a polystyrene sphere and a PTFE Flat

Next, let’s examine the adhesion force for a sphere-on-flat when they are made of dissimilar materials, by considering the special case of a 1 $\mu$m diameter polystyrene sphere sticking to a surface of polytetrafluoroethylene (more commonly called PTFE or Teflon), where $\gamma_{\text{PTFE}} = 18$ mN/m. To determine the work of adhesion between the dissimilar materials, we will use the Good–Girifalco method described in section 5.4.2.1.1.1. Since both Teflon and polystyrene are polymeric materials, the interaction between them is dominated by dispersion forces, i.e., $\Phi = 1$, and eq. (5.21) becomes

$$W_{\text{PTFE-polystyrene}} = 2\sqrt{\gamma_{\text{PTFE}}\gamma_{\text{polystyrene}}}.$$  \hfill (6.8)
The adhesion force for this sphere-on-flat situation is
\[ L_{\text{adh}} = 4\pi R \sqrt{\gamma_{\text{PTFE}} \gamma_{\text{polystyrene}}} \]
\[ = 4\pi (0.5 \mu m) \sqrt{(18 \text{ mN/m})(33 \text{ mN/m})} \]
\[ = 0.15 \mu N \] (6.9)

Even though PTFE or Teflon is often called “non-stick” because of its very low surface energy of 18 mN/m compared to other solid surfaces, the adhesive force is still not negligible. For this example of a 1 µm diameter spherical particle, the adhesive force is much greater than gravity (∼ 5 x 10^{-15} N).

So how big does the polystyrene particle have to be for the force of gravity to equal the adhesive force?

\[ \frac{4}{3} \pi g \rho R^3 = 4\pi R \gamma_S \]
\[ R = \sqrt{\frac{3\gamma_S}{g\rho}} \] (6.10)

where \( g \) is the acceleration due to gravity, and \( \rho \) is the sphere density. For the case of a polystyrene sphere of PTFE flat, \( \gamma_S = (\gamma_{\text{PTFE}} \gamma_{\text{polystyrene}})^{1/2} = 24 \text{ mN/m} \);

so, if the polystyrene sphere has a radius less than

\[ R = \sqrt{\frac{3 \times 24 \text{ mN/m}}{\left(9.8 \text{ m/s}^2\right) \times \left(1050 \text{ kg/m}^3\right)}} \]
\[ = 2.7 \text{ mm} \] (6.11)

the adhesive force from surface energy will be greater than the force of gravity acting on the sphere. Eq. (6.11) predicts that a fairly large ∼0.5 cm plastic particle should stick to a nominally low energy surface like Teflon, indicating that the contribution of surface energy to the adhesion force can be quite strong. In practice, it is rare for such large particles to stick, as microroughness tends to reduce the surface energy contribution. (We will come back to this point shortly.)

6.2.1.3 Example: Adhesion force for an atomically sharp asperity

Finally, we look at the adhesion force acting on a single asperity contacting a flat surface, where the radius of curvature is so small that the apex is made up of only a few atoms or molecules. Such sharp asperities are desirable as tips in atomic force microscopes (AFMs) and scanning tunneling microscopes (STMs), when one wants to resolve atomic and molecular surface features. A major problem, however, is keeping these tips sufficiently sharp to achieve atomic resolution, since adhesive forces may pull off those few atoms at the apex responsible for the atomic resolution.

Let’s consider the example of a tip with a 1 nm radius of curvature, which is small enough to achieve atomic resolution, but large enough that it is still reasonable (just barely) to apply the concept of surface energy. To take advantage of
PTFE’s low surface energy, to reduce the adhesive force, one might be tempted to use a PTFE tip. For a PTFE tip scanning over polystyrene, we can calculate the adhesive force by substituting $R = 1\, \text{nm}$ into eq. (6.9) resulting in $L_{\text{adh}} = 0.3\, \text{nN}$. While this is indeed a low adhesive force, if we assume that it is acting over an area with a diameter the same as the chain diameter of a PTFE molecule ($\sim 6\, \text{Å}$), then this would correspond to an applied tensile stress of 1 GPa, a hundred times higher than the yield strength of bulk PTFE (9 MPa). Consequently, we would expect that adhesive forces would plastically deform an initially atomically sharp PTFE tip, increasing its contact area to the point where the adhesive stress is below the plasticity threshold. So, while AFM tips are sometimes coated with PTFE to reduce meniscus adhesive forces, the tip atoms that do the atomic imaging need be of a much harder material.

One such hard material commonly used for AFM tips is silicon nitride (hardness = 20 GPa). Using a surface energy of $70\, \text{mN/m}$ for silicon nitride in eq. (6.9), $L_{\text{adh}} = 0.6\, \text{nN}$ for a 1 nm tip radius in contact with a polystyrene surface. Again assuming that this adhesive force acts over an area of 6 Å in diameter, the stress from this adhesive force corresponds to 2 GPa, a factor of ten less than the hardness of silicon nitride.

### 6.2.2 Adhesion-induced deformation at a sphere-on-flat contact

#### 6.2.2.1 The Johnson–Kendall–Roberts (JKR) theory

When we first considered the sphere-on-flat geometry in Chapter 3, we used the Hertz analysis to describe how an external loading force elastically deforms the contacting geometry. The Hertz analysis assumes that no adhesive forces act between the bodies, which leads to consideration of only compressive stresses within the solids. In a major advance, Johnson et al. (1971) extended the Hertz analysis to include the elastic deformations arising from the adhesive forces associated with surface energy. Within the Johnson–Kendall–Roberts (JKR) theory, a short range adhesive force results in a tensile stress being added to the Hertzian compressive stress, so the total normal stress $p(r)$ at the interface becomes

$$p(r) = p_0(1 - r^2/a^2)^{1/2} + p'_0(1 - r^2/a^2)^{-1/2}$$

(6.12)

where $a$ is the contact radius and $r$ is the distance from the center of the contact (Fig. 6.2). The first term in eq. (6.12), which is positive, is the compressive stress from eq. (3.11), while the second term, which has a negative value of $p'_0$, is the tensile stress induced by the adhesion. Eq. (6.12) indicates that the net stress is compressive towards the center of the contact zone and tensile toward the outer edges of the contact zone. As illustrated in Fig. 6.2, the tensile stress elastically deforms the edges of the contact zone, increasing its size so that the actual contact radius $a$ becomes larger than the $a_H$ predicted by the Hertz analysis.

While eq. (6.12) predicts that the stress should go to infinity at the edge of the contact zone, this theoretical divergence only comes about as JKR theory implicitly assumes that the adhesion forces act over an infinitesimally small distance exactly at the sphere–flat interface. Once the theory is modified to have an
Without adhesion (Hertz)

With adhesion

Fig. 6.2. Deformation and contact radius of an elastic sphere contacting a rigid flat. Solid line: Johnson–Kendall–Roberts (JKR) contact with adhesion. Dashed line: Hertzian contact without adhesion.

attractive force with a finite range, such as by using a Lennard–Jones potential (Muller et al. 1980; Muller et al. 1983; Fogden and White 1990), the infinite tensile stress disappears. An infinite stress would of course never occur in actual solids, as once the stress exceeds the threshold for plastic flow, the atoms move to more stable bonding configurations.

The basis of the JKR theory (Johnson 1985 pp. 125–129) is the simple concept that the total surface energy is reduced by extending the contact area through elastic deformation, which is opposed by the increase in strain energy due to the deformation; the equilibrium contact radius $a$ corresponds to the balance of the two energies. The total energy $U_T$ of the sphere-on-flat geometry is the sum of surface energy $U_S$ lost due to the formation of a finite contact area plus the energy $U_E$ stored in the elastic deformations:

$$U_T = U_S + U_E. \quad (6.13)$$

If the overall elastic deformation $\Delta z$ of the two objects in the normal direction is kept constant, the equilibrium contact radius $a$ occurs when

$$\left[ \frac{\partial U_T}{\partial a} \right]_{\Delta z} = \left[ \frac{\partial U_E}{\partial a} \right]_{\Delta z} + \left[ \frac{\partial U_S}{\partial a} \right]_{\Delta z} = 0. \quad (6.14)$$

The variation in strain energy with contact radius is given by Johnson [1985 eq. (5.47)]

$$\left[ \frac{\partial U_E}{\partial a} \right]_{\Delta z} = \frac{\pi^2 a^2}{E_c} p_0^{'2} \quad (6.15)$$

where $E_c$ is the composite elastic modulus [eq. 3.8]. The variation in surface energy loss with contact radius is

$$U_S = -W_{AB} \pi a^2$$

$$\frac{\partial U_S}{\partial a} = -2W_{AB} \pi a \quad (6.16)$$
where $W_{AB}$ is the work of adhesion between materials A and B and the minus sign comes from surface energy being lost as the contact area grows. Using eqs. (6.15) and (6.16) in eq. (6.14), and with some manipulation, we obtain the following expression for $p'_0$:

$$p'_0 = -\left(\frac{2W_{AB}E_c}{\pi a}\right)^{1/2}. \quad (6.17)$$

The net force normal to the contacting interface (the externally applied loading force $L_{\text{ext}}$) is found by integrating the $p(r)$ over the contact area

$$L_{\text{ext}} = \int_0^a 2\pi rp(r)dr = \left(\frac{2}{3}p_0 + 2p'_0\right)\pi a^2. \quad (6.18)$$

Substituting for $p_0$ [eqs. (3.10) and (3.11)] and for $p'_0$ [eq. (6.17)] and rearranging gives the relationship between the normal force $L_{\text{ext}}$ and the contact radius $a$:

$$\left(L_{\text{ext}} - \frac{4E_c a^3}{3R}\right)^2 = 8\pi W_{AB}E_c a^3. \quad (6.19)$$

If no external loading force is applied, only the attractive forces derived from the work of adhesion pull the sphere toward the flat; at $L_{\text{ext}} = 0$, eq. (6.19) predicts that the stresses from this attractive interaction generate a finite contact radius

$$a_0 = \left(\frac{9\pi W_{AB}R^2}{E_c}\right)^{1/3}. \quad (6.20)$$

The contact radius for non-zero loading forces $L_{\text{ext}}$ can be determined by manipulating eq. (6.19) to yield

$$a^3 = \frac{3R}{4E_c} \left(L_{\text{ext}} + 3\pi RW_{AB} + \sqrt{6\pi RW_{AB}L_{\text{ext}} + (3\pi RW_{AB})^2}\right), \quad (6.21)$$

which reduces when $W_{AB} = 0$ to eq. (3.7) from the Hertz analysis.

Figure 6.3 plots how the JKR contact area varies with the external loading force $L_{\text{ext}}$. (Figure 6.3 also plots the contact areas from the Derjaguin–Muller–Toporov (DMT) theory and an intermediate theory, both of which will be discussed shortly.) The JKR contact area is finite when the $L_{\text{ext}} = 0$, and this area continues to decrease as the loading becomes more negative, until the contact becomes unstable and the surfaces separate; at this critical point, the release rate for stored elastic energy exceeds the creation rate for surface energy. At separation, the magnitude of the external loading force corresponds to an adhesive force:

$$L_{\text{adh}} = \frac{3}{2} \pi W_{AB}R. \quad (6.22)$$
For the case where the sphere and flat are the same material \( W_{AB} = 2\gamma_S \), eq. (6.22) becomes

\[
L_{adh} = 3\pi\gamma_S R. \tag{6.23}
\]

At separation, the contact radius predicted by JKR theory is still finite and given by

\[
a_s = \frac{a_0}{4^{1/3}} = 0.63a_0. \tag{6.24}
\]

A major shortcoming of JKR theory is that it assumes perfectly smooth surfaces. For materials with a very low elastic modulus, such as rubber or silicone elastomers, this is not much of a problem, as the force required to flatten the surface roughness is less than the attractive forces. (This is why rubber and elastomer surfaces often feel tacky to the touch.) Indeed, Johnson, Kendall, and Roberts verified their theory by pressing together rubber spheres with a low elastic modulus of 0.8 MPa (dramatically lower than a material such as glass with a modulus of \( 10^5 \) MPa) and measuring the resulting contact radius as a function of applied load. Their data agree well with eq. (6.21) when contacting in air and immersed in water. From these experiments, they determined that \( \gamma_S = 35 \text{ mJ/m}^2 \) and \( \gamma_{SL} = 3.4 \text{ mJ/m}^2 \) for rubber in air and water, respectively. From these values, Young’s equation [eq. (5.18)] predicts a contact angle of 64° for water on rubber, close to the experimental value of 66° of Johnson et al.
(1971), providing the first experimental confirmation of Young’s equation where all the variables are measured independently.

Many subsequent studies with low elasticity surfaces have further verified JKR theory (Roberts and Othman 1977; Lee 1978; Chaudhury and Whitesides 1991; Chaudhury and Whitesides 1992). One particular situation where JKR theory applies, that deserves special mention, is the contact geometry of the surface force apparatus (Israelachvili et al. 1980; Horn et al. 1987; Frantz et al. 1997); in this apparatus, the low elasticity is provided by a soft glue used to attach the thin mica sheets to a rigid cylindrical substrate.

6.2.2.2 The Derjaguin–Muller–Toporov (DMT) theory
Another shortcoming of the JKR theory is that it only works well for materials with low elasticity, since only for these soft materials does the elastic response mimic that expected for an infinitesimally short ranged attractive force. For more realistic attractive forces with an extended range, the stresses induced are spread out over a larger volume leading to a lower elastic strain than predicted by the JKR theory. In the limit where the materials are perfectly rigid, a different theory developed by Derjaguin et al. (1975) is used. In the Derjaguin–Muller–Toporov (DMT) theory, the stress profile is assumed to be Hertzian, i.e., described by eq. (3.11), so the stress is always compressive within the contact zone, but the normal loading force is the external loading force plus the total adhesive force, such as from surface energy as given by eq. (6.4) or from the meniscus force discussed later. How the DMT theory predicts the contact area of a sphere-on-flat to vary with load is plotted in Fig. 6.3.

The JKR and DMT theories represent limiting cases where one extreme is strong adhesion coupled with highly compliant materials and the other extreme is low adhesion coupled with rigid materials. Many contact situations are intermediate between these two extremes, and Maugis (1992) has developed a comprehensive analysis to handle these intermediate cases.

6.2.2.3 Adhesion deformation in nanoscale contacts
In a few cases, researchers have been able to verify that adhesive forces lead to an increased contact area for nanoscale asperity contacts. These experiments involve sliding sharp atomic force microscope (AFM) tips, with radii of a few tens of nanometers, over surfaces with different loads and determining the contact area by applying a suitable contact mechanics model. Using this method, JKR adhesion has been verified for a platinum-coated tip sliding over a mica surface in ultrahigh vacuum (Carpick et al. 1996) and for a silicon tip sliding over highly stepped NaCl(001) in dry vacuum (Meyer et al. 1996), while DMT adhesion was verified for a silicon nitride tip in contact with mica in humid air (Carpick et al. 1997). These results show that, even for micro-asperity contacts only a few nanometers in size, adhesive forces can lead to contact areas greater than what would be expected from a Hertzian analysis.
6.3 Wet environment

When surfaces have a micro-roughness not readily flattened by the adhesive forces, JKR theory is not expected to apply. Often in these situations, however, the gaps between the contacting asperities become filled with liquid, and the adhesion force from the resulting liquid meniscus can be readily predicted.

As most environments have some degree of humidity or wetness associated with them, surfaces tend not to be completely dry. Possible environmental sources for liquid that forms menisci around the contact points include:

1. *Thin films of lubricants and contaminants (Fig. 4.9)* – Even a sub-monolayer coverage of a contaminant can result in a meniscus a few nanometers in height if that contaminant has enough surface mobility to diffuse to the contact points.

2. *Capillary condensation of water and organic vapors (Fig. 5.7)* – As discussed in Section 5.2.2, for partial vapor pressures significantly below saturation, capillary condensation results in menisci a few nanometers in height. These nanoscale menisci can be difficult to detect, except by the significant adhesion force they generate.

6.3.1 Force for a sphere-on-flat in a wet environment

Let’s consider the extra loading force \( L_{\text{men}} \) that a liquid meniscus adds to a sphere sitting on a flat [Fig. 6.1(b)]. The meniscus force \( L_{\text{men}} \) is the sum of two component forces: a surface tension force \( L_{\text{ten}} \) and a capillary pressure force \( L_{\text{cap}} \).

The surface tension force \( L_{\text{ten}} \) in the axial direction comes from the vertical projection of the liquid surface tension pulling on the solid surfaces along the edges of the contact line of the meniscus. At the contact line of the meniscus on the flat, the vertical component of the surface tension is

\[
L_{\text{ten}}/ \text{unit perimeter length} = 2\gamma_L \sin \theta_2
\]

\[
L_{\text{ten}} = 4\pi r_2 \gamma_L \sin \theta_2.
\]  

(6.25)

In the sphere-on-flat geometry, \( L_{\text{ten}} \) tends to be a much smaller component of the total meniscus force than the capillary pressure component \( L_{\text{cap}} \). Two effects lead to \( L_{\text{ten}} \) having a smaller value:

1. If the liquid wets the flat, \( \sin \theta_2 \) is small, and \( L_{\text{ten}} \) is then small from eq. (6.25).
2. If \( \theta_2 \) is large, the radius \( r_1 \) is large, and \( r_2 \) needs to be small in order for the overall radius of curvature to stay constant at the Kelvin radius [eq (5.11)].

Consequently, when a meniscus is small and subtends a small angle \( \varphi \), we can approximate the meniscus force \( L_{\text{men}} \) as being equal to the capillary pressure force \( L_{\text{men}} \approx L_{\text{cap}} \).

The capillary pressure force \( L_{\text{cap}} \) originates from the capillary pressure \( P_{\text{cap}} \) generated by the curvature of the meniscus surface acting over the area of the meniscus. For the geometry shown in Fig. 6.1(b) where \( r_1 \ll r_2 \) and \( \gamma_L \) is the
surface tension of the liquid,

\[ P_{\text{cap}} = \gamma_L \left( \frac{1}{r_1} + \frac{1}{r_2} \right) \simeq \frac{\gamma_L}{r_1}. \]  
(6.26)

To determine the capillary force acting along the axial direction, we multiply the capillary pressure times the area \( A \) of the meniscus parallel to the flat surface,

\[ A \simeq \pi r_2^2 \simeq 2\pi Rd \]  
(6.27)
since \( r_2^2 \simeq 2Rd \) when \( \varphi \) is small. So, the capillary pressure force is

\[ L_{\text{cap}} = A \times P_{\text{cap}} \simeq (2\pi Rd) \left( \frac{\gamma_L}{r_1} \right). \]  
(6.28)

When \( \varphi \) is small,

\[ d \simeq r_1 (\cos \theta_1 + \cos \theta_2) \]  
(6.29)

and eq. (6.28) becomes

\[ L_{\text{men}} \simeq L_{\text{cap}} \simeq 2\pi R \gamma_L (\cos \theta_1 + \cos \theta_2). \]  
(6.30)

So, for a small meniscus, the meniscus force does not depend on the size of the meniscus.

If the liquid wets both surfaces \((\theta_1 \simeq 0 \text{ and } \theta_2 \simeq 0)\), eq. (6.30) reduces to

\[ L_{\text{men}} \simeq 4\pi R \gamma_L. \]  
(6.31)

Conveniently, eq. (6.31) has the same form as eq. (6.5), so one only has to remember one formula for the sphere-on-flat geometry, and to remember to use the surface energy of the solid, in a dry environment, and the surface tension of the liquid, in a wet environment.

6.3.1.1 Example: Lubricant meniscus force on an AFM tip

Figure 6.4 shows the force acting on an AFM tip as samples with thin liquid lubricant films are brought into contact with the tip and then withdrawn. Before a sample makes contact with the lubricant, the force is close to zero; when contact occurs, the lubricant wicks up around the tip generating a sudden attractive meniscus force of \( \sim 2 \times 10^{-8} \text{ N} \). Since the liquid perfluoropolyether polymer used as a lubricant has a low surface tension of \( \gamma_L = 24 \text{ mN/m} \), it wets the sample and tip surfaces with a contact angle near zero, and we can use eq. (6.31) to estimate the mean macroscopic tip radius from the measured jump in meniscus force:

\[
R \simeq \frac{L_{\text{men}}}{4\pi \gamma_L} \\
\simeq \frac{2 \times 10^{-8} \text{ N}}{4\pi (24 \text{ mN/m})} \\
\simeq 66 \text{ nm}.
\]  
(6.32)
As the tip pushes through the thin liquid film, the meniscus force increases linearly with penetration distance \( u \). This increase in force occurs as the meniscus area scales with \( u \), which can be seen in a modified eq. (6.27):

\[
A \simeq \pi r_2^2 \\
\approx 2\pi Rd \\
\approx 2\pi R[r_{\text{eff}} (1 + \cos \theta) + u]
\]

(6.33)

where \( r_{\text{eff}} \) is the effective capillary radius:

\[
r_{\text{eff}} = \left( \frac{1}{r_1} + \frac{1}{r_2} \right)^{-1}.
\]

(6.34)
When $\theta \sim 0^\circ$, as it is for the liquid films in Fig. 6.4, the dependence of meniscus force on penetration distance $u$ is given by

$$F_{\text{men}} \simeq 4\pi R \left(1 + \frac{u}{2r_{\text{eff}}} \right). \quad (6.35)$$

So, the effective capillary radius—and, consequently, the capillary pressure in the meniscus—can be determined by measuring the slope of the force vs. distance curve when the tip is in contact with the liquid (Mate et al. 1989; Mate and Novotny 1991). From the force vs. distance curves in Fig. 6.4, we can see that the capillary radius is larger for the thicker films. In a later chapter we will discuss the reason for this, which relates to the balance of disjoining and capillary pressures.

Once the tip is pushed all the way through the liquid films in Fig. 6.4, it contacts the hard wall repulsion of the underlying solid surface. When the tip is withdrawn, an extra force, approximately $3 \times 10^{-8}$ N in excess of the meniscus force, has to be applied to overcome the solid–solid adhesive force.

6.3.1.2 Solid–solid adhesion in the presence of a liquid meniscus

Eqs. (6.30), (6.31), and (6.35) give only the capillary pressure component to the adhesive force. To get the total adhesive force $L_{\text{adh-tot}}$ acting on the sphere in contact, one needs to add the solid–solid adhesive force to $L_{\text{men}}$ [i.e., combining eqs. (6.4) and (6.30)]:

$$L_{\text{adh-tot}} = 2\pi R \gamma_L (\cos \theta_1 + \cos \theta_2) + 2\pi R W_{\text{ALB}} \quad (6.36)$$

where $W_{\text{ALB}}$ is the work of adhesion between the two solids A and B in the presence of a liquid L. So, in Fig. 6.4, where $R = 66$ nm, the $3 \times 10^{-8}$ N extra force to overcome the solid–solid adhesion implies that $W_{\text{tip-liquid-sample}} \sim 72$ mN/m for this situation.

When the sphere and flat are the same material ($W_{\text{ALA}} = 2\gamma_{\text{SL}}$ and $\theta_1 = \theta_2 = \theta$), eq. (6.36) becomes

$$L_{\text{adh-tot}} = 4\pi R (\gamma_L \cos \theta + \gamma_{\text{SL}}), \quad (6.37)$$

which using Young’s equation (5.19) becomes

$$L_{\text{adh-tot}} = 4\pi R \gamma_S \quad (6.38)$$

indicating that, if the sphere makes good contact with a flat of the same material, the adhesion force is not influenced by the presence of the liquid meniscus.

Experimentally, however, the presence of liquid around a contact point is frequently found to alter the adhesive force even though eq. (6.38) suggests that it shouldn’t. There are several reasons for this:

1. The loading and adhesive forces may not be sufficient to squeeze all the liquid molecules out from the region where the two solids come into contact, as illustrated in Fig. 6.5(b). In such a situation, the solid surfaces are separated
by one or two layers of liquid molecules rather than making intimate contact. In this case, the work of adhesion is smaller than $W_{ALB}$, due to the separation of the surfaces by the molecular layer. In the next chapter, we discuss how the attractive force between solid surfaces varies with separation distance. In this situation, the adhesive force becomes more dominated by the meniscus force.

2. Most solid surfaces have a microscale roughness that greatly reduces the expected solid–solid adhesive force. The liquid meniscus force is less likely, however, to be reduced by surface roughness, an effect perhaps best illustrated by the example of water adhering sand particles to surfaces or the formation of sandcastles.

### 6.3.2 Water menisci in sand

If you look at sand underneath a microscope, you notice that the particles are not spherical, but instead have an irregular shape with a fair amount of micro-roughness. So, a sand particle sitting on a flat, rather being a sphere-on-flat geometry, is more like the situation shown in Fig. 6.6, where solid–solid contact occurs only at the summits of the irregular protrusions.

At low humidity, capillary condensation causes minute menisci of water to form around the particle–flat contact points as shown in Fig. 6.6(a). Since these protrusions or asperities have radii of curvatures much smaller than the mean macroscopic radius $R$ of the particle, the net adhesive force acting on the particle is much smaller than that predicted by $4\pi R\gamma_S$. Consequently, if we sprinkle dry sand onto a flat surface and turn the surface over, most of the sand falls off since the adhesive force on each particle is less than the force of gravity.

The situation, however, is quite different at high humidity, such as might be experienced at an ocean beach. At high humidity, the large Kelvin radius leads to the condensation of water menisci large enough to flood the gaps between the contacting asperities, as illustrated in Fig. 6.6(b). For this flooded situation, the meniscus force is approximated by eq. (6.31), with $R = $ the mean radius of the sand particle. So if you spend a day at the beach, you will notice that sand
sticks to most surfaces that it comes into contact with, but eventually falls off when moved to a drier environment like the inside of your home.

Let’s estimate the meniscus force for a typical sand particle with a diameter of 0.4 mm sitting on a flat with a flooded contact region, and $\theta_1$ and $\theta_2 = 0^\circ$:

$$L_{\text{men}} = 4\pi R \gamma_L$$

$$= 4\pi \times (0.2 \text{ mm}) \times (73 \text{ mN/m})$$

$$= 0.18 \text{ mN}$$

(6.39)

which is much larger than the weight of the sand particle:

$$Weight = \frac{4}{3} \pi g \rho R^3$$

$$= \frac{4}{3} \pi \times (9.8 \text{ m/s}^2) \times (2400 \text{ kg/m}^3) \times (0.0002 \text{ m})^3$$

(6.40)

$$= 0.0008 \text{ mN}$$

So, if the flat is turned upside down, the particle does not fall off.

We can also understand why, at the beach, the partially wet strip between the water and the sand is the easiest place to walk on. Since the water only partially covers the sand, the meniscus forces are strongest, making it relatively firm. Further away from the water, the sand is only slightly wet with only a few menisci forming between the particles, so the sand is loose, soft, and difficult to walk on. Right at the water’s edge, the gaps between the sand particles are mostly flooded, so very few menisci are present to hold the sand together, again making it difficult to walk on. [This explanation is attributed to Einstein (Mermin 2005).]

To summarize: For dry sand, the micro-roughness reduces the adhesive forces to the point where they are not enough to stick the sand particles together,
and the particles collapse in a pile as shown in Fig. 6.7(a). Add a little water, however, and the sand becomes quite sticky and can be used to build a sandcastle as shown in Fig. 6.7(b). If you keep adding water to the sand, eventually all the space between the particles fills up with water, and menisci no longer exist to hold the sand together; now, all you have is unattractive silt.

6.3.3 Meniscus force for different wetting regimes at contacting interfaces
The trend discussed in the previous section for water in sand can be generalized to other situations when liquid menisci form around the contact points. Fig. 6.8 illustrates the different ways liquids can wet two contacting surfaces separated by surface roughness. The meniscus force between the surfaces depends on which of the four regimes—\textit{toe dipping, pillbox, flooded} and \textit{immersed}—occurs at the contacting interface:

6.3.3.1 Toe dipping regime
The \textit{toe dipping} regime occurs when the amount of liquid is sufficiently small that menisci form only around the summits of the asperities in contact with the opposing surface [Fig. 6.8(a)]. The summits are the “toes” that dip slightly into the liquid menisci. For contact angles $\sim 0^\circ$, the toe dipping regime occurs when $r < d/2$, where $r$ is the outer meniscus radius and $d$ is the separation distance between the two surfaces. If the $i$th summit has a mean radius of curvature $R_i$, 

---

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig6_7}
\caption{(a) Dry sand collapses into an unattractive pile. (b) Add a little water to sand and one can form a beautiful sandcastle.}
\end{figure}
Fig. 6.8. Four wetting regimes for a liquid at the interface of a smooth slider surface contacting a rough surface (Mate and Homola 1997; Mate 1998). The surface roughness separates the surfaces by a distance \( d \), and the liquid contact angles are assumed \( \sim 0^\circ \). Following the nomenclature introduced by Matthewson and Mamin (1988), these wetting regimes are labeled toe dipping, pillbox, flooded and immersed. Reproduced from Mate and Homola (1997) with permission from Springer Science and Business Media. Copyright 1997 Kluwer Academic Publishers.

then from eq. (6.31) the meniscus force \( L_{\text{men-i}} \) at this asperity contact is

\[
L_{\text{men-i}} = 4\pi R_i \gamma_L
\]  

(6.41)

and the total meniscus force acting between the two surfaces = \( \sum L_{\text{men-i}} \). So, if all the asperities have the same height and radius of curvature as shown in Fig. 6.8, the total meniscus force is

\[
L_{\text{men-tot}} = n4\pi R\gamma_L
\]

(6.42)

where \( n \) is the number of asperities in contact.

6.3.3.1.1 Example: Toe dipping adhesion with exponential distribution of summit heights  The case illustrated in Fig. 6.8 is unrealistic in that asperity summits on real surfaces do not all have the same height, but rather a range of heights described by some distribution, such as a Gaussian distribution
In Section 3.4.1.2.1, we discussed the example of laser textured surfaces in disk drives, where laser pulses are used to make bumps in disk surfaces, with a well controlled height. When the recording head slider sits on the laser textured surface, one might expect that this situation could be modeled by assuming that the slider makes contact with all the bumps as in Fig. 6.8. However, the laser textured bumps are not all exactly the same height, but rather have a slight variation in height with a standard deviation $\sigma$ of a few nanometers. Consequently, at low to moderate loading forces, the slider only contacts the highest bumps, a small fraction of the bumps underneath the slider surface.

In current example, we determine the magnitude of the meniscus force for a recording head slider sitting on a laser textured disk surface, with the same parameters as discussed in Section 3.4.1.2.1. (i.e., $L_{\text{ext}} = 20 \text{ mN}$, $R = 200 \mu\text{m}$, and $\sigma = 1.5 \text{ nm}$). In this situation, contact is only made with the highest bumps, and we can assume that these bumps have a height distribution approximated by an exponential. Using eq. (3.22) to estimate the number of bumps in contact:

$$n = \frac{L_{\text{ext}} + L_{\text{men-tot}}}{\pi^{1/2} E_c \sigma^{3/2} R^{1/2}}$$

where the total load $L$ now is expressed as the sum of the externally applied load $L_{\text{ext}}$ and a load $L_{\text{men-tot}}$ from the total meniscus force acting on those asperities in contact. Using eq. (6.42) for $L_{\text{men-tot}}$ and solving for $n$,

$$n = \frac{L_{\text{ext}}}{\pi^{1/2} E_c \sigma^{3/2} R^{1/2} - 4\pi \gamma_L R}.$$  \hspace{1cm} (6.43)

Using $\gamma_L = 24 \text{ mN/m}$ for the perfluoropolyether liquid applied to disk surface as a thin lubricant film:

$$n = \frac{20 \text{ mN}}{\pi^{1/2}(108 \text{ GPa})(1.5 \text{ nm})^{3/2}(200 \mu\text{m})^{1/2} - 4\pi(24 \text{ mN/m})(200 \mu\text{m})} = 206$$  \hspace{1cm} (6.44)

considerably higher than the $n = 127$ determined in Section 3.4.1.2.1 for the situation with no meniscus force. The total meniscus force can then be determined using eq. (6.42):

$$L_{\text{men-tot}} = n4\pi R \gamma_L$$

$$= 206 \times (6.0 \times 10^{-5} \text{ N})$$  \hspace{1cm} (6.45)

$$= 12 \text{ mN}$$

For this example, the total loading force $L = L_{\text{ext}} + L_{\text{men-tot}}$ is 60% higher than the externally applied loading force $L_{\text{ext}}$. So, we can see that the existence of meniscus forces contacting asperities increases substantially the total loading force acting at a contacting interface.
6.3.3.2 Pillbox and flooded regimes

As the amount of liquid between the surfaces increases, the liquid menisci start to flood the gaps between the asperity contacts [Fig. 6.8(b)]. When this occurs, the menisci around the contact points become very short, wide cylinders—a shape reminiscent of the old fashioned, cylindrically shaped boxes in which pills used to be dispensed—leading to the name pillbox to describe this regime.

As the amount of liquid between the surfaces further increases, the pillbox menisci eventually run together, flooding the areas between the contact points: the flooded regime in Fig. 6.8(c). As the pillbox and flooded regimes correspond to low and high fill levels of the gaps between the contacting surfaces, they represent the different extremes of a single wetting regime: the pillbox–flooded regime.

If the liquid has a contact angle $\theta = 0^\circ$, the radius of curvature of the edge of the menisci in the pillbox–flooded regime is $r = d/2$; this sets the capillary pressure $P_{\text{cap}} = -2\gamma/d$. Consequently, the total meniscus force is given by

$$L_{\text{pillbox–flooded}} = -A_{\text{men}}P_{\text{cap}} = A_{\text{men}}\frac{2\gamma_L}{d}$$

(6.46)

where $A_{\text{men}}$ is the total flooded area.

A drop of water sandwiched between two glass slides [Fig. 5.5(a)] provides a simple example of an interface in the pillbox–flooded regime: The capillary pressure pulls to the two slides together, and the water floods the gap to form one large pillbox shaped meniscus. As discussed in Section 5.1, the slides end up with a finite separation distance determined either by the surface roughness or by the heights of the particles trapped between the slides. If this separation distance is $d = 1 \mu m$, the capillary pressure within the pillbox meniscus is $-14.6 \text{ N/cm}^2$, making it extremely difficult to separate the slides.

Besides placing a volume of liquid between the contacting surfaces, liquid can also be introduced several other ways:

- A thin lubricant film can be present on one the surfaces before contact; we will discuss this in more detail in a later chapter on lubrication.
- Another common way is through the capillary condensation of vapors.

Since the boundary between the toe dipping and the pillbox–flooded regimes occurs when

$$r = d/(\cos \theta_1 + \cos \theta_2)$$

(6.47)

where $\theta_1$ and $\theta_2$ are the contact angles of the liquid on the two contacting surfaces, we can use the Kelvin equation [eq. (5.10)] to determine the boundary between the two regimes when the liquid is introduced by condensation. Fig. 6.9 plots the boundaries for condensing water and hexane vapors for the situation where these liquids wet the contacting surfaces ($\theta_1$ and $\theta_2 = 0^\circ$). For this wetting
132 SURFACE FORCES DERIVED FROM SURFACE ENERGIES

Toe dipping regime

Hexane

Water

Fig. 6.9. Boundaries between the toe dipping and pillbox–flooded regimes for the capillary condensation of water and hexane vapors between contacting surfaces when these liquids wet the solid surfaces ($\theta_1$ and $\theta_2 = 0^\circ$). $P$ is the vapor partial pressure and $P_s$ the vapor saturation pressure.

Pillbox–flooded regime

Separation, $d$ (nm)

$P/P_s$

$0.1 \ 0.2 \ 0.3 \ 0.4 \ 0.5 \ 0.6 \ 0.7 \ 0.8 \ 0.9 \ 1$

$10^{-1} \ 10^0 \ 10^1$

$d = 2r = \frac{2\gamma L V_m}{\ln(P/P_s)}$

where $\gamma$ is the liquid surface tension, $V_m$ the molar volume, $P$ the vapor partial pressure, and $P_s$ the saturation vapor pressure. From Fig. 6.9, we can see, for example, that, at a relative humidity of 90%, toe dipping occurs for separations $d > 10$ nm and pillbox-flooding occurs for separations $d < 10$ nm. So, capillary condensation of water in the pillbox–flooded wetting regime only occurs at very small separations (a few nanometers) or at very high relative humidities (over 90%). Similarly, organic vapors such hexane with partial pressures less than 85% only condense as pillbox–flooded menisci for separations less than 10 nm.

6.3.3.3 Immersed regime

When the liquid between contacting surfaces comes from capillary condensation of vapors, the pillbox–flooded regime is thermodynamically unstable since the equilibrium Kelvin radius is greater than the meniscus radius $d/2$. This imbalance drives more vapors to condense, growing the flooded area until the entire contact area becomes flooded, and a meniscus forms on the sides of the contacting bodies as shown in Fig. 6.8(d). This is the immersed regime. Since the gaps between contacting bodies tend to be very small, it can take an exceptionally long time (days, months, even years) for vapor molecules to diffuse into these gaps and
condense. During this time, the meniscus force contribution to the adhesive force, as given by eq. (6.46), slowly increases; this is a major reason why static friction increases with rest time as described in Section 4.3. When the meniscus starts to grow around the sides of the slider as shown in Fig. 6.8(d), the capillary pressure and the meniscus force start to decrease. So, the transition from the flooded to immersed wetting regimes corresponds to a maximum in the meniscus force between the contacting surfaces. Eventually, the meniscus radius in the immersed regime equilibrates with the Kelvin radius, and the meniscus force reaches an equilibrium value.

For non-volatile liquids in the pillbox–flooded regime, the menisci grow because of the capillary pressure pulling in from the molecularly thin liquid films that coat the solid surfaces. This continues until the disjoining pressure associated with the liquid film equilibrates with the capillary pressure in the menisci. The disjoining pressure and the mechanism of equilibration are discussed further in Chapter 10.

6.3.4 Example: Liquid adhesion of a microfabricated cantilever beam

In recent years, fabrication of micromechanical structures from thin films has become a commercially viable method for fabricating microelectromechanical systems (MEMS) such as accelerometers for airbag deployment and digital mirror displays for video projection. As the gaps that separate the free structures are typically only a few microns, adhesion of microstructures to nearby surfaces is an inherent problem in MEMS.

An adhesion problem often first appears during the final stages of fabrication of MEMS devices, where an etchant is used to free the structure and then the etchant is rinsed away with deionized water. If the rinse water is removed by simply drying through evaporation, the capillary force exerted by a drying meniscus can be sufficient to bend the flexible microstructures so that they contact neighboring surfaces. This effect is illustrated in Fig. 6.10 for a simple microfabricated cantilever beam structure. During the early stages of drying, the immersed regime occurs [Fig. 6.10(b)], and the capillary force is fairly low. As the liquid volume around the cantilever decreases, the meniscus radius decreases, increasing the capillary pressure and causing the cantilever to bend toward the substrate until it eventually touches [Fig. 6.10(c)]. This situation corresponds to the flooded regime where the capillary force reaches a maximum.

Several methods have been developed either to avoid meniscus formation or to alleviate the resulting adhesion or “stiction” (Maboudian and Howe 1997; Kim et al. 1998):

- **Supercritical drying** – Liquid CO₂ is used for the final rinse step, followed by pressurizing and heating above the supercritical point, and then venting at constant temperature so that no meniscus forms during drying.
Fig. 6.10. Side view of a microelectromechanical system (MEMS) cantilever beam. (a) A free standing beam. (b) The beam during the evaporation of the rinse liquid. (c) The beam when the capillary force leads to adhesion of the beam to the substrate.

- **Low surface-tension solvents** – Rather than drying after the water rinse step, the water is rinsed away with a low surface-tension solvent such as methanol to reduce the magnitude of the capillary force in the final drying phase.
- **Polymer standoffs** – The microstructures are fabricated with polymer standoffs that hold them apart during the drying process and that are removed with a final ashing step.
- **Surface roughening** – The microstructure and substrate surfaces are roughened to increase the separation of the surface means during contact, so as to reduce the capillary pressure when in the flooded pillbox wetting regime. Increased roughness also helps reduce the number of surface asperities in contact, and, correspondingly, the meniscus force for the toe dipping phase of drying.
- **Low-energy surface monolayers** – The microstructure and substrate surfaces are coated with a monolayer film that provides for a high water contact angle to reduce the meniscus force.

The last two methods are also effective for reducing the likelihood that the microstructures will become stuck after fabrication and during use, a potential problem in high humidity conditions.

Now one might think that if one did a thorough job of drying the microstructure, say by heating it above the boiling point for water, that those structures
stuck together by the meniscus forces would become unstuck once the meniscus forces are gone. Generally though, once stuck, the structures do not become unstuck since solid–solid adhesion at the contacting asperity summits, along with the van der Waals and electrostatic forces acting across the narrow gap separating the contacting surfaces, are sufficient to hold them together. These forces are discussed in the following chapter.

6.4 References


In the preceding two chapters, we discussed the concept of surface energy, and how some forces between surfaces arise from surface energy and surface tension. Ultimately though, surface energy must also originate from the forces between the atoms; so, if we truly want to understand surface forces, we need to understand the various forces that can act between pairs of atoms and the ways that we can sum up these forces to obtain the surface energy or the net force acting between surfaces. This approach also lets us examine a wider range of surface forces than just those related to surface energies, as well as allowing us to examine how surface forces, such as adhesion, depend on separation distance and other physical parameters.

In this chapter, we focus on those forces originating from atomic potentials or contact electrification acting in a normal direction between two surfaces, leaving for later viscous and tangential (frictional) forces. Later chapters will also delve into how the various forces acting between atoms contribute to friction, adhesion, and wear.

### 7.1 Normal force sign convention

Before beginning this discussion of forces acting in the normal direction to surfaces, we should first realize that the choice of sign for this force is somewhat arbitrary, with the choice usually depending on the context. For the forces between atoms and molecules, the typical convention is for an attractive force to have a negative sign and a repulsive force a positive sign (as used, for example, in Fig. 3.1). Since many forces between solids originate from the atomic and molecular potentials, it is natural to continue this sign convention when discussing those forces between surfaces originating from these potentials. (This is the convention adopted in this chapter.) This convention results in an adhesive force having a negative sign.

When discussing macroscopic tribological phenomena, however, the normal force acts as a **loading force**, which, by convention, has a positive sign when pushing two surfaces towards each other. This contrasts with the convention of using a negative sign for an adhesive force even when the adhesive force acts to push two surfaces into contact. So, when considering an adhesive force’s contribution to the loading force in a macroscopic situation, one typically switches the sign of the adhesive force from negative to positive so that it matches the sign of the externally applied loading force. This was the approach utilized in previous chapters,
where a positive sign convention was used for the adhesive force, and the nomenclature $L_{adh}$ used to represent an adhesive force contributing to the loading force.

A similar sign convention issue occurs when energy is considered. When discussing the interaction between atoms and molecules, an attractive interaction results in a negative interaction energy. When the attractive interactions between atoms, however, are summed to obtain the interaction energy between surfaces, this interaction energy is often referred to as the work of adhesion (Section 5.3), which, by convention, has a positive sign.

### 7.2 Repulsive atomic potentials

Back in Section 3.1, we discussed how the forces between atoms lead to the mechanical properties of materials. Now, we discuss how these atomic and molecular forces combine to become the forces acting between solid surfaces. Figure 3.1 illustrated how the force between two atoms has two components: a short range repulsive force and a long range attractive force. We start by considering the repulsive atomic potential (Israelachvili 1991 Section 7.2).

When two atoms are brought close enough together, their electron clouds overlap; as this overlap is energetically unfavorable, a repulsive force is generated to minimize the amount of overlap. This repulsive force is fundamentally quantum mechanical in nature, arising from the Pauli exclusion principle that two electrons can not occupy the same quantum state. This atomic repulsion is called either exchange repulsion, hard core repulsion, steric repulsion, or, for the case of ions, Born repulsion. The essential characteristic of an atomic repulsion potential is that it rises steeply with decreasing separation distance $r$. A number of expressions are frequently used for describing a repulsive potential:

- **Hard wall potential**
  \[ w(r) = \begin{cases} \infty & \text{where } r \leq r_0 \\ 0 & \text{where } r > r_0 \end{cases} \]
  (7.1)
  and $r_0$ represents the hard wall radius

- **Inverse power-law potential**
  \[ w(r) = \frac{A}{r^n} \]
  (7.2)
  where $n$ is usually in the range of 9 to 16. A common choice for the value of $n$ is 12, which, when combined with the attractive potential for a van der Waals force, gives the well known Lennard–Jones potential between atoms:
  \[ w_{L–J}(r) = \frac{A}{r^{12}} - \frac{B}{r^6}. \]
  (7.3)

- **Exponential potential**
  \[ w(r) = c e^{-r/r_0} \]
  (7.4)
  where $c$ and $r_0$ are adjustable parameters, with $r_0$ on the order 0.2 Å
These potentials are empirical relations with no physical theory underlying them. All three potentials, however, do a good job of describing the steeply rising character of the actual repulsive potentials. The inverse power and exponential potentials are more realistic potentials than the hard wall potential as they allow for some compressibility or softness of the potential; while the hard wall potential has the advantage of simplicity.

7.3 Van der Waals forces

7.3.1 Van der Waals forces between molecules

Van der Waals describes the class of interactions between atoms and molecules where the interaction energy $w(r)$ varies as $1/r^6$ and the force varies as $1/r^7$ with separation distance $r$. Three types of van der Waals forces occur, which are illustrated in Fig. 7.1:

1. **Orientation force (Keesom interaction)** – The electrostatic interaction between two polar molecules, such as two water molecules, with both having freely rotating permanent dipoles, averaged over the thermally excited probability distribution of rotation angles. Since the antiparallel alignment

![Fig. 7.1. Schematic of the three types of van der Waals interactions between molecules: (a) Orientation interaction between two freely rotating polar molecules. (b) Induction interaction between a polar and a non-polar molecule. (c) Dispersion interaction between two non-polar molecules. (d) Illustrates how the electric field $E$ of a polar molecule induces a dipole in a non-polar molecule.](image-url)
of the dipoles is more energetically favorable, an attractive force is generated when they adopt the antiparallel arrangement and a repulsive force when they adopt a parallel arrangement. Higher temperatures disorder this alignment, reducing the interaction force.

2. **Induction force (Debye interaction)** – The electrostatic interaction between a polar molecule with a freely rotating permanent dipole moment and a non-polar molecule, averaged over the thermally excited probability distribution of rotation angles. As illustrated in Fig. 7.1(d), the electric field from the permanent dipole moment induces a dipole moment in the non-polar molecule \( \mu_{\text{ind}} = E\alpha \) where \( \alpha \) is the polarizability of the non-polar molecule; the two dipoles then interact electrostatically to generate the induction force.

3. **Dispersion force (London interaction)** – The induced dipole–dipole interaction between two non-polar molecules that is quantum mechanical in origin. A simple explanation is that a non-polar atom such as helium has an instantaneous dipole moment, which naively can be associated with the “instantaneous positions” of the electrons around its nucleus. The electric field from this instantaneous dipole polarizes any nearby atom inducing in it a dipole moment. The electrostatic interaction between these two induced dipoles corresponds to an exchange of virtual photons, generating an instantaneous dispersion force.

Even though the dispersion force or London force is the weakest of the three van der Waals interactions, it is often considered the most important for the following reasons:

- The dispersion force is always present between materials.
- The dispersion force does not decrease with temperature, unlike the orientation force.
- The dispersion force is usually the largest contributor to the van der Waals interaction, except for the most polar of molecules, such as water.

For two dissimilar polar molecules interacting across a vacuum, the van der Waals potential \( w_{\text{VDW}}(r) \) can be expressed as the sum of these three interactions (Israelachvili 1991 eq. (6.17)):

\[
w_{\text{VDW}}(r) = -\frac{C_{\text{VDW}}}{r^6} = -\frac{C_{\text{ind}} + C_{\text{orient}} + C_{\text{disp}}}{r^6}
= -\left[ (u_1^2\alpha_{02} + u_2^2\alpha_{01}) + \frac{u_1^2u_2^2}{3kT} + \frac{3\alpha_{01}\alpha_{02}h\nu_1\nu_2}{2(\nu_1 + \nu_2)} \right] / (4\pi\varepsilon_0)^2 r^6
\]

(7.5)

where \( u_1 \) and \( u_2 \) are the permanent dipole moments of the two molecules, \( \alpha_{01} \) and \( \alpha_{02} \) are their polarizabilities, and \( h\nu_1 \) and \( h\nu_2 \) are their first ionization potentials. Eq. (7.5) indicates that the van der Waals force \( F_{\text{VDW}} \) is always attractive across a vacuum (i.e., \( F_{\text{VDW}} = -\partial w_{\text{VDW}}/\partial r < 0 \)).
When a third medium (say a liquid solvent) fills the space between the two interacting molecules, the interaction of the molecular dipole moments within this medium must be taken into account when computing the van der Waals force. McLachlan (1963) developed a generalized theory for the van der Waals interaction that describes the effect of an intervening medium. From the McLachlan theory, an expression can be derived for the van der Waals interaction between dissimilar materials across an intervening third medium, in terms of the dielectric constants $\varepsilon(0)$ at zero frequency and refractive indices $n$ determined at optical frequencies of the materials involved (Israelachvili 1991 pp. 99–103). The van der Waals potential $w_{\text{VDW}}(r)$ is expressed as the sum of a “zero frequency” dipolar contribution $w(r)_{\nu=0}$ and “finite frequency” dispersion force contribution $w(r)_{\nu>0}$:

$$w_{\text{VDW}}(r) = w(r)_{\nu=0} + w(r)_{\nu>0}. \quad (7.6)$$

For two dissimilar spherical molecules or small particles labeled 1 and 2, with radii $a_1$ and $a_2$ with an intervening medium labeled 3, and assuming the electromagnetic adsorption frequency $\nu_e$ is the same for all three media, the zero and finite frequency contributions to the van der Waals interaction are

$$w(r)_{\nu=0} = -\frac{3kT a_1^3 a_2^3}{r^6} \left( \frac{\varepsilon_1(0) - \varepsilon_3(0)}{\varepsilon_1(0) + 2\varepsilon_3(0)} \right) \left( \frac{\varepsilon_2(0) - \varepsilon_3(0)}{\varepsilon_2(0) + 2\varepsilon_3(0)} \right) \quad (7.7)$$

and

$$w(r)_{\nu>0} = -\frac{\sqrt{3}h\nu_e a_1^3 a_2^3}{2r^6} \frac{(n_1^2 - n_3^2)(n_2^2 - n_3^2)}{(n_1^2 + 2n_3^2)^{1/2}(n_2^2 + 2n_3^2)^{1/2} \left[ (n_1^2 + 2n_3^2)^{1/2} + (n_2^2 + 2n_3^2)^{1/2} \right]^2}. \quad (7.8)$$

The important points to notice about eqs. (7.7) and (7.8) are that:

- Since $h\nu_e \gg kT$, the finite frequency dispersion force contribution $w(r)_{\nu>0}$ is typically much greater than the zero frequency dipolar contribution $w(r)_{\nu=0}$, especially when the intervening medium 3 is a gas or vacuum ($n_3 \sim 1$).
- The van der Waals force between two molecules is greatly reduced by having another material, such as a liquid, intervening between the two molecules. For the special case where all three media have the same refractive index ($n_1 = n_2 = n_3$), then $w(r)_{\nu>0} = 0$ (van der Waals dispersive component is zero).
- For dissimilar molecules, $w(r)_{\nu>0}$ is repulsive if $n_3$ is intermediate between $n_2$ and $n_1$. So, while the van der Waals force is always attractive when medium 3 is a gas or vacuum ($n_3 \sim 1$) and usually attractive when $n_3 > 1$, it can be repulsive (depending on the magnitude of $w(r)_{\nu=0}$) in the special situation of $n_1$ and $n_2$ being unequal and $n_3$ being intermediate between these values.
7.3.1.1 Retardation effects for dispersion forces

In the above expressions for the van der Waals potential energies [eqs. (7.5), (7.7), and (7.8)], the interaction decays at the rate of \(-1/r^6\) with separation distance. At sufficiently large distances, however, the dispersion force component decays at a faster rate of \(-1/r^7\). This comes about from the finite amount of time that it takes an electric field to travel between the molecules, which is determined by the speed of light in the intervening medium. At larger separations, the time that it takes for the electric field to travel from the first atom to the second atom and back again becomes comparable to the oscillation period of the fluctuating instantaneous dipole of the first atom. So, when the electric field from the induced dipole moment of the second atom returns to the first atom, the direction of the first atom’s dipole moment has changed from where it originally started, weakening the interaction.

For two atoms in free space, this retardation effect begins at separations greater than 5 nm, and the potential approaches a \(-1/r^7\) dependence at \(r > 100\) nm. Since the dispersion force is already very weak at 5 nm separation, retardation effects are usually of little interest. When an intervening medium is present, however, the retardation manifests itself at smaller separations as the speed of light is slower. Only the dispersion force suffers from retardation; so, while the dispersion force may be the dominant van der Waals force at small separations, the induction and orientation contributions to the van der Waals force ultimately dominate at larger separation distances.

7.3.2 Van der Waals forces between macroscopic objects

In the 1930s, Hamaker showed that the van der Waals forces between molecules could be extended to macroscopic objects by performing a pairwise sum over the atoms in these objects (Hamaker 1937). This approach requires the assumption that the van der Waals interaction is additive, which is valid if no intervening material exists between the two objects (i.e., a gap of vacuum or gas exists between them). In the next couple of sections, we will use this approach of summing the molecular pair potentials, which is called the microscopic approach, to derive [following Israelachvili (1991, chapters 10 and 11)] the van der Waals interaction potential and force for several geometries.

7.3.2.1 Molecule–flat surface interaction

If an atom or molecule is a distance \(D\) above a flat solid surface as shown in Fig. 7.2, it experiences an attractive van der Waals force between it and all the atoms in the solid. To evaluate the net interaction energy, we sum the individual pair potentials [Fig. 7.2(a)]:

\[
 w_{\text{net}}(r) = -\sum_i \frac{C_{\text{VDW}}}{r_i^6}.
\] (7.9)

Rather than summing up over all the individual atoms in the solid, it is simpler, of course, to convert this sum into an integral. For this integral, we take
Fig. 7.2. (a) Interaction of an atom above a surface represented as the sum of the individual pair potentials between that atom and atoms in the solid. (b) Ring shaped integration element with radius \( x \) for determining the interaction energy of a molecule above a planar surface. Courtesy of J. N. Israelachvili.

The element of integration to be a circular ring with radius \( x \), as illustrated in Fig. 7.2(b), where each part of the ring is a distance \( r = \sqrt{z^2 + x^2} \) from the molecule, the cross-sectional area of the ring is \( dx \, dz \), and the ring volume is \( 2\pi x \, dx \, dz \). If \( \rho \) is the number density of the atoms in the solid, the number of atoms in the ring is \( 2\pi \rho x \, dx \, dz \), and the net potential energy for a molecule a distance \( D \) above the surface is

\[
\begin{align*}
w_{\text{molecule-flat}}(D) &= -2\pi C_{\text{VDW}} \rho \int_{z=D}^{z=\infty} dz \int_{x=0}^{x=\infty} dx \frac{dx}{(\sqrt{z^2 + x^2})^6} \\
&= \frac{\pi C_{\text{VDW}} \rho}{2} \int_{D}^{\infty} \frac{dz}{z^4} \\
&= -\frac{\pi C_{\text{VDW}} \rho}{6D^3}
\end{align*}
\]

(7.10)

and the net van der Waals force acting on the molecule is

\[
F_{\text{molecule-flat}} = -\frac{\partial w_{\text{molecule-flat}}(D)}{\partial D} = -\frac{\pi C_{\text{VDW}} \rho}{2D^4}.
\]

(7.11)

The force is directed along the \( z \)-axis, and the negative sign indicates that it directed towards the surface, i.e., attractive. (In doing this integral, we also made the approximation of an \( r^{-6} \) potential even at distances \( r \) where retardation effects may be significant; this approximation is valid when \( D \) is much less than the distance where the interaction transitions from non-retarded to retarded.)
7.3.2.2 Flat–Flat interaction

Now that we have the interaction energy between an atom or molecule above a planar surface [eq. (7.10)], it is straightforward to derive the expression for the van der Waals energy between two planar surfaces. Rather than considering the case of two surfaces of infinite lateral extent (which would result in an infinite interaction energy), we consider the case illustrated in Fig. 7.3(a) where the top surface has a unit area, so as to determine the van der Waals potential per unit area, while the bottom surface has a much larger area, effectively extending to infinity. For an atom in the top material (with a number density of atoms \( \rho_2 \)) at a distance \( z \) away from the bottom surface (with a number density of atoms \( \rho_1 \)), the interaction energy per atom is given by eq. (7.10). Multiplying this interaction energy by the number of atoms in a thin sheet of thickness \( dz \) gives the interaction energy for the sheet, and integrating from \( z = D \) to infinity gives the van der Waals energy per unit area

\[
W_{\text{flat-flat}}(D) = -\frac{\pi C_{\text{VWD}} \rho_1 \rho_2}{6} \int_D^\infty \frac{dz}{z^3}.
\]

(7.12)

\[
= -\frac{\pi C_{\text{VWD}} \rho_1 \rho_2}{12D^2}.
\]

Fig. 7.3. (a) Geometry for calculating the interaction energy between a planar surface with unit area and second planar surface of much larger area with \( D \ll b \). (b) For sphere and planar surface with \( D \ll R \). Courtesy of J. N. Israelachvili.
While eq. (7.12) was derived for one surface with unit area interacting with an infinitely larger one, it is also valid for the interaction between two unit area surfaces provided the separation distance $D$ is much smaller than the lateral dimensions of the surfaces ($D \ll b$).

From eq. (7.12) the van der Waals force per unit area (or pressure) acting between two surfaces is readily determined:

$$F_{\text{flat-flat}}(D) = -\frac{\partial W_{\text{flat-flat}}(D)}{\partial D} = -\frac{\pi C_{\text{VDW}} \rho_1 \rho_2}{6D^3}. \quad (7.13)$$

### 7.3.2.3 Sphere–flat interaction

For the sphere-above-flat geometry shown in Fig. 7.3(b), the van der Waals potential again can be calculated by doing an integral as done previously for the molecule–surface and surface–surface potentials (Israelachvili 1991 p. 157). For situations where $D \ll R$, a simpler method, however, is to use the Derjaguin approximation expressed in eq. (6.2) together with the van der Waals energy between two flat surfaces [eq. (7.12)] to determine the van der Waals force for a sphere above a flat surface:

$$F(D)_{\text{sphere-flat}} = 2\pi RW_{\text{flat-flat}}(D) = -\frac{\pi^2 R C_{\text{VDW}} \rho_1 \rho_2}{6D^2}. \quad (7.14)$$

Integrating this force, we obtain the van der Waals interaction energy for the sphere-above-flat geometry:

$$W_{\text{sphere-flat}}(D) = \int_D^\infty F_{\text{sphere-flat}}(D') dD'$$

For $D \ll R$,

$$W_{\text{sphere-flat}}(D) = -\frac{\pi^2 C_{\text{VDW}} \rho_1 \rho_2}{6D}. \quad (7.15)$$

where the constant of integration is determined by setting $W_{\text{sphere-flat}}(\infty) = 0$.

For the other extreme of a sphere far away from the flat ($D \gg R$), the van der Waals potential can be approximated by multiplying the potential per atom [eq. (7.10)] times the number of atoms in the sphere $(4\pi R^3 \rho_2 / 3)$:

For $D \gg R$,

$$W_{\text{sphere-flat}}(D) = -\left(\frac{4\pi R^3 \rho_2}{3}\right) \frac{\pi C_{\text{VDW}} \rho_1 \rho_2}{6D^3}. \quad (7.16)$$

### 7.3.3 The Hamaker constant

Common to the above expressions for the van der Waals interaction between macroscopic objects is the combination of parameters $C_{\text{VDW}} \rho_1 \rho_2$, which together describe the strength of the van der Waals interaction between materials 1 and 2.
across an empty gap. It is convenient to rewrite these expressions with this combination of parameters being replaced with a new parameter called the *Hamaker constant*, in recognition of Hamaker’s contributions to the field. Within the microscopic approach described above, the relationship between the Hamaker constant $A$ and the molecular level parameters is

$$A = \pi^2 C_{\text{VDW}} \rho_1 \rho_2. \quad (7.17)$$

From eq. (7.17), we see that the Hamaker constant or the strength of the van der Waals interaction is proportional to the number densities and the polarizabilities of the atoms and molecules in the two interacting bodies. [Eq. (7.5) shows that $C_{\text{VDW}}$ is proportional to the polarizabilities].

Figure 7.4 summarizes the expressions of the non-retarded van der Waals interaction potentials and forces, between macroscopic objects across an empty gap in terms of the Hamaker constant. While these expressions were derived in the previous sections using the microscopic approach, they still remain valid when the more rigorous quantum electrodynamics (QED) approach described in the next section is used.

### 7.3.3.1 Determining Hamaker constants from Lifshitz’s theory

Rather than summing over point-to-point interaction potentials, a more fundamental (and consequently more precise) approach to calculating the van der Waals interaction is to use quantum electrodynamics (QED). This was first done by Casimir (1948) who analyzed the retarded attractive dispersion force between two perfectly conducting plates, a force now referred to as the *Casimir force*. Lifshitz subsequently developed the general QED solution for the van der Waals interaction between two materials (which could be either conductors or non-conductors) separated by a third intervening material (Lifshitz 1956; Dzyaloshinskii et al. 1960; Dzyaloshinskii et al. 1961).

In Lifshitz’s theory, the materials are treated as a continuum, and the van der Waals force arises from fluctuations in the electromagnetic field between the two objects, modified by the intervening material due to the boundary conditions imposed by the geometry. The fluctuations or virtual photons take the form of standing waves that only occur at certain optical frequencies, with other frequencies from the continuum of virtual photons being excluded from the gap between the two materials. This exclusion results in a pressure that pushes the two materials together.

Using Lifshitz’s theory, the van der Waals force between macroscopic objects can be calculated from the geometry of the interacting objects and from the frequency-dependent dielectric response function $\varepsilon(\omega)$ of the two interacting materials and any intervening materials. Typically though, the calculation is only done for a planar geometry, such as the one illustrated in Fig. 7.5, and the result used to determine the Hamaker constant for the particular combinations
Van der Waals forces

<table>
<thead>
<tr>
<th>Geometry</th>
<th>Van der Waals interaction energy</th>
<th>Van der Waals interaction force</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two flat surfaces</td>
<td>$W = - \frac{A}{12 \pi D^2}$</td>
<td>$- \frac{\partial W}{\partial D} = - \frac{A}{6 \pi D^3}$ per unit area</td>
</tr>
<tr>
<td>Sphere above flat $D &lt;&lt; R$</td>
<td>$W = - \frac{AR}{6D}$</td>
<td>$- \frac{\partial W}{\partial D} = - \frac{AR}{6D^2}$ per unit area</td>
</tr>
<tr>
<td>Particle above flat $D &gt;&gt; R$</td>
<td>$W = - \frac{2AR^3}{9D^3}$</td>
<td>$- \frac{\partial W}{\partial D} = - \frac{2AR^3}{3D^4}$ per unit area</td>
</tr>
</tbody>
</table>

Fig. 7.4. Non-retarded van der Waals interaction energies and forces for selected geometries.

of materials involved. The van der Waals force for other geometries is then computed by using the determined Hamaker constant in the appropriate expression (some of which are listed in Fig. 7.4).

A major advantage of Lifshitz’s theory over the earlier microscopic approach of Hamaker is that the van der Waals interaction can be calculated when an intervening material is present; Figure 7.5(a) illustrates such a case, where materials 1 and 3 interact across a gap filled with material 2. For these cases, we append subscripts to the Hamaker constant to identify the order of the materials across the gap; in the situation illustrated in Fig. 7.5, for example, the Hamaker constant is labeled $A_{123}$.
The Lifshitz analysis is also readily extendable to the case where multiple layers of materials exist in the gap between two objects (Ninham 1970; Israelachvili 1972; White et al. 2005). Figure 7.5(b) illustrates such a case, where a material 1 with an overlayer 2 interacts across a gap filled with medium 3 with another half space of material 5 with overlayer 4. For this situation, where the film thicknesses $h_2$ and $h_4$ are constant, the Lifshitz analysis can be used to calculate an effective Hamaker constant $A_{12345}$, but now $A_{12345}$ is a function of the gap separation distance $D$ (White et al. 2005). In the limit of very small or large separation distances, the Hamaker constant for the multilayer system in Fig. 7.5(b) can be approximated by the Hamaker constant for the appropriate three material system: for $D \ll h_2, h_4$, $A_{12345}(D) \rightarrow A_{234}$; for $D \gg h_2, h_4$, $A_{12345}(D) \rightarrow A_{135}$ (though retardation effects often become important at these larger separation distances).

As the accuracy of the Lifshitz approach depends on the accuracy of the dielectric spectra obtained from experiments and of their arithmetical representations, extensive work must first be done to obtain good representations of these optical spectra, before the Lifshitz analysis can be applied. Recent advances in vacuum ultraviolet spectroscopy and valence-electron-energy-loss spectroscopy have led
to improved spectra over the wide range of optical frequencies needed for such an optical analysis (French et al. 1995; French et al. 1998; Palik and Ghosh 1998). With good mathematical representations of optical spectra now available, precise calculations of Hamaker constants can now be done (certainly more precise than those from the McLachlan theory mentioned previously). Several tabulations of Hamaker constants can be found in the literature (Hough and White 1980; Bergstrom 1997; French 2000).

Tables 7.1, 7.2, and 7.3 give values for the Hamaker constant for material combinations sometimes encountered in tribology. We can see that:

- The Hamaker constants across a vacuum or air gap are all within an order of magnitude of each other. For the same materials acting across a gap, these range at the low end from $3.8 \times 10^{-20}$ J, for the low surface energy material PTFE ($\gamma = 18$ mN/m), to $2.96 \times 10^{-19}$ J for the high density material diamond at the high end of the range ($\rho = 1.76 \times 10^{23}$ cm$^{-3}$).

- The Hamaker constants for polymers ($3.80$ to $6.26 \times 10^{-20}$ J) are at the lower end of this range, due to the lower density and polarizability of these materials. The lower Hamaker constant helps to account for polymers having weaker interactions and lower surface energies than other materials.

- Materials with high density and polarizability are at the upper end of the range ($1$ to $3 \times 10^{-19}$ J). Metals are examples of highly polarizable materials, which are not discussed in the tables. Interacting across a vacuum, metals typically have a Hamaker constant of $\sim 4 \times 10^{-19}$ J (Israelachvili 1991 Section 11.4).

- The Hamaker constants are greatly reduced when water fills the gap between the two materials. This is a fairly general phenomenon: introducing a liquid into the gap between materials typically reduces the van der Waals interaction

| Table 7.1. Non-retarded Hamaker constants for inorganic materials. Two identical materials separated by water or a vacuum (or air) gap, i.e., $A_{121}$, material 2 is vacuum or water. Calculated from Lifshitz theory (Bergstrom 1997). |
|---|---|---|
| Material | Hamaker constant ($10^{-20}$ J) | Across vacuum or air | Across water |
| $\alpha$-Al$_2$O$_3$ | 15.2 | 3.67 |
| C (diamond) | 29.6 | 13.8 |
| Mica | 9.86 | 1.34 |
| $\beta$-SiC | 24.6 | 10.7 |
| Si$_3$N$_4$ (amorphous) | 16.7 | 4.85 |
| SiO$_2$ | 6.50 | 0.46 |
Table 7.2. Non-retarded Hamaker constants for polymer materials. Two identical materials separated by water or a vacuum (or air) gap, i.e., $A_{121}$, and material 2 is vacuum or water. Calculated from the Lifshitz theory.

<table>
<thead>
<tr>
<th>Polymer</th>
<th>Hamaker constant ($10^{-20}$ J)</th>
<th>Across vacuum or air</th>
<th>Across water</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly(methyl methacrylate) PMMA$^a$</td>
<td>5.84</td>
<td>0.147</td>
<td></td>
</tr>
<tr>
<td>ET-MAA$^a$</td>
<td>4.77</td>
<td>0.216</td>
<td></td>
</tr>
<tr>
<td>Polyester (PEST)$^a$</td>
<td>6.09</td>
<td>0.405</td>
<td></td>
</tr>
<tr>
<td>(PIMI)$^a$</td>
<td>6.26</td>
<td>0.523</td>
<td></td>
</tr>
<tr>
<td>Polycarbonate (PCARB)$^a$</td>
<td>5.08</td>
<td>0.350</td>
<td></td>
</tr>
<tr>
<td>Polystyrene (PSTY)$^a$</td>
<td>5.56</td>
<td>0.316</td>
<td></td>
</tr>
<tr>
<td>Polytetrafluoroethylene (PTFE)$^b$</td>
<td>3.80</td>
<td>0.333</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ French 2000

$^b$ Hough and White 1980

Table 7.3. Non-retarded Hamaker constants ($10^{-20}$ J) for inorganic materials interacting against four different types of materials across vacuum/water. Calculated from Lifshitz theory (Bergstrom 1997).

<table>
<thead>
<tr>
<th>Material</th>
<th>Silica</th>
<th>Silicon nitride</th>
<th>Alumina</th>
<th>Mica</th>
</tr>
</thead>
<tbody>
<tr>
<td>C (diamond)</td>
<td>13.7/1.71</td>
<td>22.0/7.94</td>
<td>21.1/7.05</td>
<td>17.0/4.03</td>
</tr>
<tr>
<td>Mica</td>
<td>8.01/0.69</td>
<td>12.8/2.45</td>
<td>12.2/2.15</td>
<td>9.86/1.34</td>
</tr>
<tr>
<td>6H-SiC</td>
<td>12.6/1.52</td>
<td>20.3/7.22</td>
<td>19.2/6.05</td>
<td>15.5/3.54</td>
</tr>
<tr>
<td>β-Si3N4</td>
<td>10.8/1.17</td>
<td>17.3/5.13</td>
<td>16.5/4.43</td>
<td>13.3/2.61</td>
</tr>
<tr>
<td>SiO2 (quartz)</td>
<td>7.59/0.63</td>
<td>12.1/2.07</td>
<td>11.6/1.83</td>
<td>9.35/1.16</td>
</tr>
</tbody>
</table>

by an order of magnitude below the vacuum/air value. This should not be surprising, as it was pointed out during the discussion of the McLachlan theory [eqs. (7.7) and (7.8)].

Tables 7.1 and 7.2 tabulate the Hamaker constants for two identical materials acting across a gap. From these Hamaker constants (which have the form $A_{121}$ or $A_{1v1}$, where the subscript $v$ represents a vacuum or vapor gap), we can approximate the Hamaker constant for pairs of dissimilar materials by using the combining relations (Israelachvili 1991 Sec. 11.9):

$$A_{121} = A_{212}$$  \hspace{1cm} (7.18)

$$A_{123} \simeq \pm (A_{121}A_{323})^{1/2}$$  \hspace{1cm} (7.19)
\[ A_{1v3} \simeq (A_{1v1}A_{3v3})^{1/2} \quad (7.20) \]
\[ A_{121} \simeq A_{1v1} + A_{2v2} - 2A_{1v2} \quad (7.21) \]

### 7.3.3.2 Example: Van der Waals force on a polystyrene sphere above a Teflon flat

In Section 6.2.1.2, we calculated the adhesive force, arising from surface energy, for a polystyrene sphere sitting on a PTFE (Teflon) flat. Since the interactions between the molecules within and between polystyrene and PTFE are mainly van der Waals, we can also calculate this adhesive force directly from the van der Waals force, if we know the appropriate value to use for the separation distance \( D \). (Simply using \( D = 0 \) would result in the unphysical value of \( F_{VDW} = -\infty \).) Our approach here will be to calculate how the van der Waals adhesive force varies with separation distance \( D \), and then determine at what separation distance the adhesive force is the same as that determined in Section 6.2.1.2 using surface energies.

First, we estimate the Hamaker constant using eq. (7.20):

\[
A_{PTFE-v-PSTY} = (A_{PTFE-v-PTFE}A_{PSTY-v-PSTY})^{1/2} \\
= \left[ (3.8 \times 10^{-20} \text{ J}) \times (5.56 \times 10^{-20} \text{ J}) \right]^{1/2} \\
= 4.6 \times 10^{-20} \text{ J} \quad (7.22)
\]

The van der Waals force is then calculated using the equation for a sphere above a flat (Fig. 7.4, assuming \( D \ll R \)):

\[
F_{VDW} = -\frac{A_{PTFE-v-PSTY}R}{6D^2} \\
= -\frac{(4.6 \times 10^{-20} \text{ J}) \times (0.5 \mu \text{m})}{6D^2} \\
F_{VDW} (\mu \text{N}) = -\frac{0.38}{D (\text{Å})^2} \quad (7.23)
\]

Figure 7.6 plots the van der Waals force between this 1µm polystyrene sphere and a PTFE flat as function of separation distance \( D \).

In Section 6.2.1.2, we found that \( L_{adh} = 0.15 \mu \text{N} \) (implying that \( F_{VDW} = -0.15 \mu \text{N} \), when the sphere contacts the flat). From Fig. 7.6 and eq. (7.23), we see that this value of adhesive force occurs at \( D_0 = 1.65 \text{ Å} \); as discussed in the next section, this is generally the van der Waals separation distance where the atoms on two opposing surfaces are in intimate contact.

As mentioned in the previous chapter, surface roughness can substantially reduce the magnitude of the adhesive force. One simple approach for estimating how surface roughness reduces the van der Waals adhesive force is to set the van der Waals separation distance \( D \) equal to the separation of surface means when the two rough surfaces are in contact plus \( D_0 \). For example, if we assume that
roughness separates the surface means by 8.35 Å (so $D = 10$ Å) for this 1 μm diameter particle sitting on a flat, we see from Fig. 7.6 that $F_{\text{VDW}}$ decreases to $-3 \mu N$ or a factor of fifty smaller magnitude than what is expected for perfectly smooth surfaces.

### 7.3.4 Surface energies arising from van der Waals interactions

For van der Waals interactions it is straightforward to see how the surface energy relates to the forces between atoms:

In prior sections, we summed the van der Waals interaction energies between the atoms in one material with all the atoms in the other material to obtain the object–object interaction energy. Summing instead over the interactions between all atoms in both materials determines, this time, the total van der Waals interaction energy. For two parallel surfaces of identical materials separated by $D$, the total van der Waals interaction energy per unit area can be expressed as

$$W_{\text{total-per-unit-area}}(D) = (\text{bulk cohesive energy}) + 2\gamma - \frac{A}{12\pi D^2} \quad (7.24)$$

where the first term is the bulk cohesive energy, the second term is the surface energy of the two surfaces arising from van der Waals energy, and the last term is the flat–flat van der Waals interaction energy. If $D_0$ is the separation distance where the atoms on the two surfaces come into intimate contact such that
\[ W_{\text{total-per-unit-area}}(D_0) = \text{(bulk cohesive energy)}, \text{ this implies that} \]

\[ \gamma = \frac{A}{24\pi D_0^2}. \quad (7.25) \]

At first, one might be tempted to equate \( D_0 \) with the interatomic distance \( a \); this approach, however, ends up underestimating the surface energy \( \gamma \). The problem occurs as eq. (7.25) is essentially derived from a continuum theory where the surfaces are treated as ideally smooth with none of the lumpiness associated with the atomic structure of actual surfaces; consequently, eq. (7.25) is not really valid for separation distances on the order of an atomic diameter. It can be shown, however, that \( D_0 = a/2.5 \) provides a good estimate for the value of \( D_0 \) (Israelachvili 1991 p. 203). If we use \( a = 4 \, \text{Å} \) as a typical atomic spacing, \( D_0 = 1.65 \, \text{Å} \) (also the value determined in Section 7.3.3.2), and eq. (7.25) becomes

\[ \gamma = \frac{A}{24\pi (1.65 \, \text{Å})^2}. \quad (7.26) \]

Eq. (7.26) can be rewritten to estimate the Hamaker constant if we know the material’s surface energy \( \gamma \):

\[ A (\text{J}) \approx 2.1 \times 10^{-21} \, \gamma \left( \text{mJ/m}^2 \right). \quad (7.27) \]

### 7.3.5 Van der Waals adhesive pressure

If we imagine an arbitrary plane dividing a solid, the van der Waals force acting across this plane is an adhesive pressure that holds the solid together and is balanced by the atomic level repulsive forces. This adhesive pressure is given by

\[ P = \frac{A}{6\pi D_0^3}. \quad (7.28) \]

So, picking a lower end value of the Hamaker constant, \( A = 5 \times 10^{-20} \, \text{J} \), we find that, in weakly bound materials, with \( D_0 = 1.65 \, \text{Å} \), the adhesive pressure is

\[ P = \frac{5 \times 10^{-20} \, \text{J}}{6\pi \left(1.65 \times 10^{-10} \, \text{m}\right)^3} \]

\[ = 590 \, \text{MPa} \]

\[ = 5823 \, \text{atm} \quad (7.29) \]

So, the adhesive pressure holding these materials together is impressively large!

In principle, when two separate bodies are brought into contact, similarly large adhesive pressure occurs over the nominal contact area. However, it is only
for materials like elastomers, that have elastic moduli low enough for the van der Waals adhesive forces deform the surfaces to the extent that intimate contact occurs over the whole area, that such high pressures can develop over the whole contact area. (These are also those situations where the JKR theory of Section 6.2.2.1 works best for estimating the adhesive force from surface energies.) Typically, the stiffness of the surface roughness prevents intimate contact from occurring except at asperity summits, so the high adhesive pressures are localized to these small contact zones (where they can still dramatically influence the resulting tribology). As discussed in the previous chapter, however, the gaps between the solid–solid contacts can become filled with soft materials or liquids from the condensation of vapors or from preexisting lubricant or contamination layers; for this situation, the adhesive pressure exists over the entire contact area, but is the average of the adhesive pressures at solid–solid contacts and the soft material between the contacts.

The next section discusses in detail how surface roughness influences the average van der Waals adhesive pressure in the absence of intervening material between the solid–solid contacts.

7.3.6 Van der Waals interaction between contacting rough surfaces

So far the expressions for the van der Waals interaction between surfaces (such as those in Fig. 7.4) have assumed that the surface roughness is much smaller than the separation distance \( D \). Often, though, we need to know the van der Waals contribution to the adhesive force acting between rough contacting surfaces. Intuitively, we can see that increasing roughness lowers the van der Waals adhesive force as it results in a larger fraction of the opposing surfaces being farther away from each other. The challenge is to come up with a suitable analytical or numerical method for estimating the van der Waals interaction between these rough surfaces. Since the nature of roughness can vary dramatically from one situation to the next, a single modeling approach is unlikely to be applicable for all cases. In what follows, we discuss only three different approaches out of the many possible ways of handling the van der Waals forces between rough surfaces; fortunately, these approaches apply to a wide range of rough contacting interfaces.

1. Separation of surface means approach – The simplest approach is to model the surfaces as being relatively smooth except for a few exceptionally high asperities that set the average separation distance \( D_{\text{ave}} \) when contact occurs. Figure 7.7(a) illustrates this situation where the rms surface roughness \( \sigma \ll D_{\text{ave}} \) and the van der Waals force across the non-contacting area dominates over the force acting at the contacting asperities. The adhesion energy per unit area can then be approximated by using the expression for the van der Waals energy between two parallel surfaces:

\[
W_{\text{ad}} = \frac{A}{12\pi D_{\text{ave}}^2}
\]  

(7.30)
where we assume that $D_{\text{ave}} < 5 \text{ nm}$ so we can use the non-retarded expression. For this situation, the average adhesive pressure is given by

$$P_{\text{ad}} = -\frac{\partial W_{\text{ad}}}{\partial D} = -\frac{A}{6\pi D_{\text{ave}}^3}$$  \hspace{1cm} (7.31)

2. Greenwood–Williamson approach – The next approach applies to rough surfaces with spherical or elliptically shaped asperities summits whose heights follow approximately a Gaussian distribution. For this situation illustrated in Fig. 7.7(b), we can use the Greenwood–Williamson (G–W) method (described in Section 3.4.1) to model the two rough surfaces in terms of an equivalent rough surface contacting a flat surface. For the equivalent rough surface, the asperities are all assumed to have the same radius of curvature $R_{\text{ave}}$, which is the average radius of curvature of asperities on the top and bottom surfaces, and the distribution of asperity heights is assumed to be Gaussian with a standard deviation given by

$$\sigma = \left(\sigma_{\text{top}}^2 + \sigma_{\text{bottom}}^2\right)^{1/2}.$$

The next step is to determine the mean separation $D_{\text{ave}}$ where the repulsive forces due the elastic or plastic deformation of the contacting asperities balances the adhesive pressure from van der Waals forces and meniscus forces, plus any externally applied load. This is accomplished by determining the repulsive contact and adhesive pressures as function of average separation distance $D_{\text{ave}}$ through numerical integration over the distribution of asperity heights.
heights. The van der Waals force acting on an asperity with a height \( z \) from the surface mean is given by using the sphere-above-flat expression in Fig. 7.4 with \( D_0 = 1.65 \text{Å} \) used for those asperities in contact and \((D_{\text{ave}} - z + D_0)\) used for those not in contact. Typically, \( D_{\text{ave}} \) is found to be in the range of 3 to 5σ, depending on the strength of the adhesive pressures, the externally applied load, and the contact stiffness.

Once the \( D_{\text{ave}} \) has been determined, the total van der Waals interaction energy per unit area (the work of adhesion \( W_{\text{ad}} \)) can be determined by integrating the individual sphere-above-flat energies over the asperity height distribution. This approach can be extended to account for a multitude of other possible effects at the contacting asperities such as elastic and plastic deformations and the capillary effects of lubricant menisci (Polycarpou and Etsion 1998).

3. The Comprehensive approach – The last approach is a more comprehensive, numerical method that can be applied to most types of surface roughness (Delrio et al. 2005). First, a representative sample of the surface topographies of the opposing surfaces is collected using AFM images with sufficiently high resolution to capture the details of individual asperities. Next, pairs of these images are placed facing each other in computer software with their surfaces means separated by varying \( D_{\text{ave}} \) distances. For each \( D_{\text{ave}} \), the local separation distance \( d_{\text{local-i}} \) for each pixel \( i \) is computed by subtracting from \( D_{\text{ave}} \) the sum of the \( z \)-heights measured relative to the respective surface means:

\[
d_{\text{local-i}} = D_{\text{ave}} - (z_{\text{substrate-i}} + z_{\text{cantilever-i}}).
\]

(7.32)

[For those cases where contact occurs \((z_{\text{substrate-i}} + z_{\text{cantilever-i}}) > D_{\text{ave}}\), \( d_{\text{local-i}} \) is set equal to zero.] Then, the work of adhesion per unit area is determined as a function of \( D_{\text{ave}} \) by treating each pixel location as a parallel surface geometry and averaging over the image area:

\[
W_{\text{ad}} (D_{\text{ave}}) = \frac{1}{N_{\text{pixels}}} \left( \sum_i \frac{A}{12\pi (d_{\text{local-i}} + D_0)^2} \right)
\]

(7.33)

where \( N_{\text{pixels}} \) is the number of pixels in an image and \( D_0 \) is the effective separation distance at contact. Differentiating eq. (7.33) with respect to \( D_{\text{ave}} \) determines the average van der Waals adhesive pressure as a function of \( D_{\text{ave}} \); where it balances the repulsive contact pressure (which can be determined, for example, by the Greenwood–Williamson method) determines the equilibrium average separation distance.

7.3.6.1 Example: Stuck microcantilevers

Now let’s apply the previous analysis to the practical example of a microfabricated cantilever beam that was initially free standing as shown in Fig. 6.10(a), but which, for some reason, has become stuck to the substrate surface, as shown in Fig. 7.7. This stuck cantilever is similar to the one discussed in Section 6.3.4,
but now we consider the case where no liquid is present to generate a meniscus force and only van der Waals forces act across the contacting interface. Figure 7.8 shows the results of Delrio et al. who have experimentally realized this van der Waals stuck cantilever geometry by first coating the surfaces with a hydrophobic monolayer to prevent the capillary condensation of water vapors (Delrio et al. 2005).

As well as being of technical importance to the design of MEMS devices, the stuck cantilever geometry also has scientific value in that the work of adhesion can be experimentally measured from the lengths of the unstuck and stuck portions, $s$ and $L - s$, respectively, making it possible to compare theory with experiment
An equilibrium adhered length $L - s$ occurs when the sum of the elastic energy $U_E$ stored in the unattached portion and the adhesion energy $U_S$ stored in the adhered portion is minimized, i.e., when

$$\frac{d}{ds}(U_E + U_S) = 0.$$  \hfill (7.34)

For the case shown in Fig. 7.7 where the attached portion lies flat on the substrate, the elastic energy stored in the bent cantilever can be expressed as (de Boer and Michalske 1999):

$$U_E = \frac{Ewt^3d^2}{2s^3}.$$  \hfill (7.35)

where $w$ is the width of the cantilever beam and $E$ its elastic modulus. The adhesion energy stored in the attached portion is simply the work of adhesion per unit area $W_{ad}$ times the contact area:

$$U_S = -W_{ad}w(L - s)$$  \hfill (7.36)

Using eqs. (7.35) and (7.36) in eq. (7.34) and solving for the work of adhesion in terms of the length of the unattached portion and the beam parameters yields

$$W_{ad} = \frac{3}{2} \left( \frac{Et^3d^2}{s^4} \right).$$  \hfill (7.37)

Next, we would like to understand the role of surface roughness in influencing the work of adhesion determined experimentally using eq. (7.37).

Figure 7.8 shows that the calculated results (using the comprehensive approach) agree well with experimentally determined values of the work of adhesion for silicon oxide cantilever beams stuck on silicon oxide substrates with varying amounts of surface roughness (Delrio et al. 2005). Delrio et al. also compare the results from the comprehensive approach with those of the separation of surface means approach and a single-asperity model. An important thing to notice is that the values for the van der Waals interaction energy obtained from the separation of surface means approach are much lower than those from the comprehensive approach and experimental values. It is always the case that the separation of surface means approach underestimates the van der Waals interaction energies and forces. This is because, due to the $D^{-2}$ dependence of the van der Waals interaction energy, those parts of a rough surface with surface heights above the surface mean contribute much more to the interaction energy than what is lost from those portions with surface heights below the surface mean.

### 7.3.7 Example: Gecko adhesion

Geckos are lizards that have the remarkable ability of being able to run up smooth walls and across ceilings with relative ease. Scientists and many casual observers have long puzzled over how geckos, weighing up to 50 grams, are able
successfully to scale such a wide variety of surfaces. Many insects also have the ability to cling to surfaces, by secreting liquids from glands in their feet so as to generate adhesion through capillary forces. Since these glands are absent in geckos, another mechanism must be at work.

Recently, Kellar Autumn and co-workers have determined that geckos exploit the weak, but always present van der Waals forces, to adhere to surfaces (Autumn et al. 2000; Autumn and Peattie 2002; Autumn et al. 2002). Geckos are able to do this because of extraordinary nanoscale structures on the pads of their toes that work together to engage and disengage the van der Waals adhesive force. As illustrated in Fig. 7.9, these toe pads consist of ridges of lamellae that are covered with arrays of hair-like beta-keratin bristles, called setae. For a typical tokay gecko, an individual seta has about a 100 µm length and a 5 µm diameter. Each seta branches into hundreds of tiny endings, called spatulae, that touch the surface to engage the van der Waals forces. The individual spatulae ends, where adhesion occurs, have a flattened shape, about 200 nm across and need to be properly oriented to achieve maximum adhesion.

Measurements with whole tokay gecko indicate that the two front feet can generate a shear force of 20.1 N (Irschick et al. 1996), which corresponds to an average force of 6.2 µN per seta. Measurements by (Autumn et al. 2000) on

![Fig. 7.9. (a) Schematic of a gecko foot illustrating the lamellar structure of setae hairs on the toes (Autumn and Peattie 2002). (b) An individual seta hair as it comes close to contacting a flat surface. (c) The ends of the spatulae that must become oriented towards the flat surface if strong van der Waals adhesion is to occur.](image-url)
individual seta indicate that the key to the achieving the maximum adhesion and shear forces is properly orienting the seta and its spatulae ends relative the contacting surface. First the seta stalk needs to be oriented within a few degrees of the angles that align its end parallel with the contacting surface. Next, to maximize the contact area, the seta is first pushed toward the surface then dragged a short distance across it so as to bring the flat ends of the spatulae uniformly flush against the surface. With this procedure, a maximum shear force can be achieved for an individual seta of 200 µN, 32 times larger than the average obtained by measuring with the whole animal. Geckos are thought to use a similar procedure when they plant their toes on a surface to obtain the best traction and adhesion; then they disengage the forces, rotating their toes so as to peel off the setae ends off the contacting surface. The maximum adhesive force is also found to be 200 µN. If all 6.5 million setae of a 50 gram tokay gecko achieved this maximum value, the total adhesion force would be 1.3 kN, enough to support the weight of two humans.

The van der Waals force acting on an individual spatula can be estimated using the equation for two planar surfaces:

$$F_{ad\text{-}spatula} = -A_{spatula} \frac{A}{6\pi D^3}$$

(7.38)

where $A_{spatula} = 2 \times 10^{-14}$ m$^2$ is the typical area of an end of a spatula. If we assume that the Hamaker constant $A = 1 \times 10^{-19}$ J and that the separation distance is as small as possible ($D = D_0 = 1.65 \text{ Å}$), then

$$F_{ad\text{-}spatula} = -\left(2 \times 10^{-14} \text{ m}^2\right) \frac{10^{-19} \text{ J}}{6\pi (0.165 \text{ nm})^3}$$

$$= -24 \mu \text{N}$$

(7.39)

So, a single seta with a hundred contacting spatulae could potentially achieve an adhesion force as high as 2.4 mN. Since this an order of magnitude higher than the measured value for an individual seta, it is likely that a small amount of roughness between the contacting surfaces leads to an average spacing $D_{ave} > D_0$. For example, if we assume 100 spatula per seta make contact, a 200 µN adhesion force for an individual seta can be achieved when $D_{ave} = 0.38 \text{ nm}$, suggesting that spatula ends come fairly close to achieving the most intimate possible contact.

7.3.8 Van der Waals contribution to the disjoining pressure of a liquid film

So far we have been considering only the van der Waals interaction between two solid bodies separated either by a vacuum, a gas, or a liquid. The situation described in Fig. 7.5 is fairly general, however, and materials 1, 2, and 3 can be any combination of gas, liquid, or solid. Consequently, the Hamaker constant can be calculated for many other situations besides the interaction between two solid bodies. One combination frequently encountered in tribology is the situation illustrated in Fig. 7.10, where a liquid lubricant film with thickness $h$ covers a solid surface. Since the air–liquid interface and the liquid–solid interface are two
parallel flat surfaces, the same expression derived in Section 7.3.2 for two solid objects can be used for this situation. Therefore, the van der Waals interaction between the lubricant film and the solid in excess of the cohesive and interfacial energies is

$$\pi(h) = \frac{A_{SLV}}{6\pi h^3}$$  \hspace{1cm} (7.40)

where $\pi(h)$ is called the *disjoining pressure* of the liquid film and $A_{SLV}$ is the Hamaker constant for the solid–liquid–vapor geometry. Since the term “disjoint” comes from the combination of “dis” and “join”, which means to pull apart or separate, so the disjoining pressure can be considered as the pressure acting to separate the solid–liquid and liquid–air interfaces.

In eq. (7.40), the sign convention is chosen so that the Hamaker constant $A_{SLV}$ is positive when an attractive van der Waals interaction exists between the liquid film and the solid surface so that an attractive interaction leads to a positive value for the disjoining pressure $\pi(h)$. Such an attractive interaction works to thicken the liquid film. This convention differs from that used for two solids separated by a gap, where an attractive interaction between the solids, which also provides for a positive Hamaker constant, leads to a negative force that acts to thin the gap separating the solids.

One shortcoming of eq. (7.40) is that $\pi(h) \to \infty$ as $h \to 0$. This potentially unphysical situation of infinite disjoining pressure can be remedied by invoking the concept of a closest approach distance $D_0$ between the liquid film and solid substrate, as was done for calculating the van der Waals contribution to surface energy in Section 7.3.4. In this case eq. (7.40) is rewritten

$$\pi(h) = \frac{A_{SLV}}{6\pi (h + D_0)^3}$$  \hspace{1cm} (7.41)

where $D_0$ is on the order of the sum of the van der Waals radii of the substrate and liquid atoms. Since $D_0$ is on the order of a few ångströms, the difference between eq. (7.40) and eq. (7.41) only becomes significant for film thicknesses $h$ less than a few nanometers. For example, much better agreement has been
found between experiment and theory by using eq. (7.41) and $D_0 = 3.17 \, \text{Å}$ for the spreading of nanometer thick perfluoropolyether liquid films (Marchon and Karis 2006).

7.4 Liquid-mediated forces between solids

The presence of a liquid between two solids can greatly change the nature of the forces between the opposing surfaces. In Chapter 6, we discussed the forces arising from the surface tension of the liquid between the surfaces. In this section, we discuss some of the forces acting directly across the gap due to the presence of a liquid.

Previously in this chapter, we discussed how the presence of a liquid typically reduces the size of the van der Waals force by an order of magnitude, decreasing its significance. We also showed that the van der Waals force remains attractive, if the value of the liquid refractive index is not intermediate between those of the solid refractive indexes (a fairly rare event). Consequently, one might suspect that all dissolved particles would tend to stick together due to the van der Waals attraction and precipitate out as a solid mass; this does not generally happen as various repulsive and oscillatory forces occur in liquids to separate surfaces. This is fortunate: as our bodies are 75% water, we would be subjected to a rather unpleasant fate if all our suspended solids precipitated out.

7.4.1 Solvation forces

When a flat solid surface is introduced into a liquid, the liquid molecules adjacent to the surface rearrange to pack well against it, as this helps lower the overall free energy. As illustrated in Fig. 7.11, this packing leads to several features in the liquid density profile near the surface:

- The layer of molecules in contact with the surface has a density $\rho_s$ that is higher than the bulk liquid density $\rho_{\text{bulk}}$.
- The liquid structure transitions from being ordered next to the solid to being disordered in the bulk liquid, with each subsequent layer being less well ordered than those closer to the surface.
- The liquid density profile has peaks spaced by a molecular diameter $a$, and the amplitude of these density oscillations decays exponentially away from the surface.

It is important to realize that this ordering of liquid molecules next to a surface does not require any liquid–liquid or liquid–wall interaction, but instead is determined primarily by the geometry of the molecules and how they pack around the constraining boundary. Since a similar structural ordering also occurs for solvent molecules around a particle or solute molecule in a solution, forces arising from any disruption of this ordering are referred to as solvation forces.

When a second solid surface is brought near this first surface so as to sandwich a film of liquid between the two surfaces, as shown in Fig. 7.12, the ordering
Fig. 7.11. (a) Illustration of how packing of spherical liquid molecules next to a solid surface leads to layering of molecules near the surface. (b) Density profile of the liquid near the surface.

Fig. 7.12. (a) Ordering of spherical liquid molecules between two parallel surfaces of finite extent; the molecules are free to exchange with those in the bulk liquid. The molecules are able to form well defined layers when the separation distance $D$ is a multiple of the molecular diameter $a$. (b) Solvation pressure as expressed by eq. (7.42). Reproduced from Israelachvili (1991, Fig. 13.2). Copyright 1991, with permission from Elsevier.
at the individual surfaces is disrupted, and the molecules rearrange to find the most energetically favorable packing geometry. Figure 7.12(a) illustrates how the molecular packing for spherical molecules between two parallel surfaces varies with the separation distance $D$. Since liquid molecules interact via some potential, such as a Lennard–Jones potential [eq. (7.3)], it takes energy to disrupt the original ordering of the liquid; therefore, this disruption caused by the second surface generates an oscillatory solvation pressure $P(D)$ with attractive maxima when $D$ is a multiple of the molecular diameter $a$. To a good approximation, the solvation force can be described by an exponentially decaying cosine function (Tarazona and Vicente 1985):

$$P(D) \approx -kT \rho_s \cos(2\pi D/a)e^{-D/a}$$

which is shown in Fig. 7.12(b). Eq. (7.42) is just the expression for the oscillating solvation force contribution, which is superimposed on the other forces acting between solids such as the van der Waals and the electrostatic double-layer forces. It should be noted that these contributions are not necessarily additive; for example, within Lifshitz’s theory, the van der Waals force depends on the density of the intervening liquid, and this density is, in turn, influenced by the solvation force.

One might think that it would be necessary to have two parallel surfaces to observe the oscillatory solvation force, as any curvature of the surfaces would lead to a range of separation distances and an averaging out of the oscillatory component of the solvation force. Somewhat surprisingly, however, surface curvature does not wipe out the solvation force, which can be understood as follows: If the ordering of liquid molecules between two parallel surfaces results in an interaction energy $W(D)$ that oscillates with separation distance $D$, then, by the Derjaguin approximation, the force $F(D)$ between two spheres, a sphere-on-flat, or cross cylinders will also be oscillatory. For example, for a sphere-on-flat geometry, $F(D)_{\text{sphere–flat}} = 2\pi RW(D)$ [eq. (6.2)]; so the solvation force simply scales with the radius of curvature of the surface.

7.4.1.1 Example: Squalane between smooth mica surfaces
Figure 7.13 shows the force measured by Zhu and Granick to push together or to separate two curved mica surfaces in the presence of liquid squalane molecules (Zhu and Granick 2004). As the surfaces are pushed together, squalane drains out of the gap without measurable resistance until $D$ is reduced to about 4 nm. For separations <4 nm, an oscillatory force of alternating attraction and repulsion appears, with the repulsive maxima corresponding to the pressure needed to squeeze out each layer during compression and with the attractive minima corresponding to the suction pressure needed to pull in each layer when the surfaces are separated. The minimum separation achieved during squeezing is $D = 9\text{ Å}$, corresponding to the molecular diameter of a balled up squalane molecule [situation $B$ in Fig. 7.12(a)].
The oscillatory character of the force in Fig. 7.13 is quite striking, tempting one to think that this phenomenon is quite easy to observe. While many examples of oscillatory forces have been documented for liquids confined between mica surfaces in surface force apparatus (SFA) experiments, it is actually fairly difficult to observe an oscillatory force in other experimental situations, even though the liquid molecules may be forming ordered layers between contacting surfaces. A major reason is that most surfaces are not as atomically smooth as mica; when the surfaces are randomly rough, the oscillatory character of the force is smeared out and it appears as a monotonic solvation force. Typically, a roughness on the order of the molecule dimension $a$—generally just a few ångstroms—is all that it takes to make the oscillatory force disappear (Gee and Israelachvili 1990; Gao et al. 2000; Granick et al. 2003; Samoilov et al. 2004).

The results in Fig. 7.13 for liquid squalane help illustrate how roughness can interfere with measurements of an oscillatory solvation force. Prior to this work,
surface force apparatus experiments with alkanes had observed oscillatory forces for linear alkanes but no oscillation when a single methyl group was added to the linear chain (Gee and Israelachvili 1990). For squalane, a highly branched alkane with six symmetrically placed methyl groups along a C$_{24}$ backbone, only one shallow minimum had been observed in SFA force vs. distance curves (Granick et al. 1995; Drummond and Israelachvili 2001) prior to Zhu and Granick’s work; therefore it was thought that adding the branching methyl groups was sufficient to “roughen” the surfaces of the molecules themselves and thereby disrupt the ordering between the surfaces. What Zhu and Granick found, however, was that using the correct preparation method to obtain the smooth mica surfaces was key to observing the oscillatory forces in Fig. 7.13. They used the preparation method developed by Frantz and Salmeron, where the mica surfaces are recleaved in the SFA just before the liquid is placed between them so as to minimize the opportunity for airborne contaminants to land on the surfaces and generate a molecular scale roughness (Frantz and Salmeron 1998). While earlier preparation methods also provided surfaces with large percentages of atomically smooth areas, AFM topography measurements confirm that the Frantz–Salmeron method does generate surfaces with fewer sub-nanometer size contaminants. This subtle change of surface roughness at the atomic and molecular scale appears to have a dramatic influence on whether or not the oscillatory solvation forces are observed. This may not be the whole story, however, as Gourdon and Israelachvili have been unable to reproduce the oscillations with mica surfaces prepared using nominally the same Frantz and Salmeron method, suggesting that there may be additional reasons why the oscillations are so pronounced in Fig. 7.13 (Gourdon and Israelachvili 2006). One possibility is the improved detection scheme used by Zhu and Granick where the separation distance $D$ is measured over a smaller area to obtain a higher resolution of its value (Wong et al. 2006).

7.4.1.2 Oscillatory solvation forces at sharp AFM contacts

Another approach for experimentally measuring the solvation force is to use a sharp atomic force microscope (AFM) tip contacting a flat surface in the presence of the liquid. Since a typical AFM tip has a radius of curvature of only a few tens of nanometers, this gets around the roughness issue somewhat, as the opposing surfaces need only be smooth over their effective contact area, typically less than 10 nm across. One major challenge, however, is measuring the small amplitude of the oscillatory force since it scales with tip radius, though this has been done in several instances using modulation techniques (Lim et al. 2002; Lim and O’Shea 2002). For example, if we take $\Delta F/R = 12$ mN/m for the oscillation amplitude of squeezing out the next to last squalane layer shown in Fig. 7.13, we should expect a $\Delta F$ of 0.6 nN for the strongest oscillation when $R = 50$ nm; Lim and O’Shea measure the value of $\Delta F = 1.5$ nN for the maximum force oscillation for squalane between an AFM tip and a smooth graphite surface (Lim and O’Shea 2002). (The higher than expected value may indicate a higher affinity of squalane for graphite than mica.)
The observations of oscillatory solvation forces at sharp AFM contacts indicates that such forces should also occur at individual asperity contacts separated by liquids. For contacting rough surfaces, however, these weak oscillatory forces are almost always averaged out over the aggregate of all asperity contacts.

7.4.2 Forces in aqueous medium
The high dielectric constant and high degree of hydrogen bonding of water leads to some distinct forces when surfaces are separated by liquid water, that is, electrostatic double-layer force, hydration repulsion, and hydrophobic attraction. These forces have been extensively studied due to their importance to biological systems, colloidal suspensions, and other aqueous-based systems. In the next few sections, we provide a short overview of these forces specific to an aqueous medium.

7.4.2.1 Electrostatic double-layer force
When immersed in water, a solid surface develops a surface charge by one of two mechanisms:

1. Ions dissolved in the water absorb onto the surface.
2. Chemical groups on the surface ionize or dissociate into ions. For example:
   a. Organic surfaces frequently have carboxylic groups that dissociate in the presence of water (–COOH → −COO− + H+).
   b. Glass surfaces hydroxylate in the presence of water, and these surfaces become negatively charged when the resulting silanol groups dissociate (SiOH → SiO− + H+).

Once these charges are present on a surface, they are balanced, so that the net charge is zero, by counterions dissolved in the region next to the surface as illustrated in Fig. 7.14. A few of these counterions bind to the surface to form what is called the Stern or Helmholtz layer. The remaining counterions form a diffuse electric double layer in the liquid close to the charged surface. In this electric double layer, two competing forces act on the counterions: the attractive electrostatic force pulling the ions toward the oppositely charged surface and the entropic force or thermal agitation pushing these ions apart from each other and away from the surface to increase entropy.

When two surfaces with similar charges are brought together in a liquid, as shown in Fig. 7.14, their electric double layers overlap, increasing the concentration of counterions near the surfaces. Since this increase in concentration reduces entropy, it generates an entropic repulsive interaction that is greater than attractive electrostatic interaction. The resulting repulsive force is referred to as the electric or electrostatic double-layer force, even though the repulsion arises from entropic confinement. The contact value theorem expresses the repulsive pressure due the electrostatic double-layer force in terms of the increased counterion
density at the solid surfaces [for a derivation of this theorem see Israelachvili (1991, section 12.7)]:

\[ P(D) = kT \left( \rho_s(D) - \rho_s(\infty) \right). \]  

(7.43)

At large distances, the electrostatic double-layer force decays exponentially due to screening of the electric field by the ion concentration; the decay length equals the Debye length:

\[ \lambda_D = \sqrt{\frac{\varepsilon_r \varepsilon_0 k_B T}{2ce^2}} \]  

(7.44)

where \( \varepsilon_0 \varepsilon_r \) is the dielectric permeability of the liquid and \( c \) is the concentration of a monovalent salt in mol/L. If higher valency ions are present, \( 2c \) is replaced with \( \sum c_i Z_i^2 \) where \( c_i \) is the concentration of the \( i \)th ion species, \( Z_i \) its valence, and the sum is over all the types of ions present.

We can see from eq. (7.44) that the higher the dielectric constant, the longer the Debye length. So while the electrostatic force can occur in all liquids, only
for high dielectric constant liquids does it extend for significant distances. With a dielectric constant \( \varepsilon_r = 78.4 \) at 25\(^\circ\)C, water has one of the highest values for dielectric constants, resulting in a very long range electrostatic double-layer force at low ion concentration. For water, eq. (7.44) reduces to \( \lambda_D = 3.04 \text{ Å}/\sqrt{c} \) and, at pH = 7\((c = 10^{-7} \text{ M})\), \( \lambda_D = 960 \text{ nm} \sim 1 \mu\text{m} \), a pretty large decay distance. Adding salt to water increases the concentration of the dissolved ions, resulting in the electrostatic interaction being more effectively screened (i.e., a shorter Debye length). For example, the ion concentration (predominately from dissolved NaCl or KCl) in most fluids in animal bodies (including our own bodily fluids) is about 0.2 M, resulting in a Debye length \( \lambda_D = 0.7 \text{ nm} \) for these aqueous fluids: much smaller than for pure water, but still large enough for electrostatic double-layer repulsion to play a major role in our bodily functions.

For colloids in aqueous medium, the electrostatic double-layer repulsion keeps particles in suspension. Adding salt, however, will cause many colloidal systems to coagulate. This behavior was first quantitatively explained by Derjaguin and Landau (1941) and Verwey and Overbeek (1948) in what is now called the DLVO theory of colloidal stability. In this theory, the interaction between two particles is assumed to consist of two parts: a van der Waals attraction and an electrostatic double-layer repulsion. At low salt concentrations, the double-layer repulsion keeps the particles separated. Increasing the salt concentration leads to increased screening of the double-layer repulsion and, at a critical concentration, the height of the repulsive barrier is reduced to the point that the van der Waals attraction causes the particles to coagulate.

### 7.4.2.2 Hydration repulsion and hydrophobic attraction

Within bulk water, a strong network of hydrogen bonding exists between the water molecules. Anytime a foreign molecule or surface is introduced into water, however, the water molecules in the immediate vicinity reorient themselves to obtain the most favorable interaction in their new situation, even if the reorientation interferes with the hydrogen bonding between water molecules. When two surfaces are brought within a few molecular diameters of each other in water, the hydrogen bonding network becomes further disrupted, and this leads to either repulsive or attractive forces between the surfaces depending on whether the surfaces are hydrophilic or hydrophobic, respectively.

The precise nature of hydration repulsion between two hydrophilic surfaces is still unclear, but is thought to be due to water molecules adjacent to the surface forming strong hydrogen bonds with hydrophilic surface groups such as hydroxyl (–OH) groups and with hydrated surface ions. The repulsion then corresponds to the energy needed to disrupt the hydrogen bonding network sufficiently to squeeze the hydrated water out from between the surfaces. Empirically, the interaction energy \( W \) for hydration repulsion is found to decay exponentially with the surfaces’ separation distance \( D \):

\[
W = Ae^{-D/\lambda_H} \tag{7.45}
\]
where $\lambda_H$ ranges from 0.2 to 1.4 nm and $A$ ranges from $10^{-3}$ to 10 J/m$^2$. This short range makes hydration repulsion readily distinguishable from other types of interactions with longer range (electrostatic, van der Waals, and electrostatic double-layer for salt concentrations below 0.1 M). For molecularly smooth surfaces, the hydration repulsion may also exhibit an oscillatory component due to the solvation force of layered water molecules. Since the hydration force extends to separations of more than two molecular layers of water, more is contributing to the force than just the hydration bonding of the water layer adjacent to the hydrophilic surfaces. One theory is that the adjacent, hydrated water molecules are more ordered than in bulk liquid and that this higher degree of ordering extends away from the surface for several water layers (Marcelja and Radic 1976). Further information about hydration forces can be found in the reviews by Cevc (1991), Leikin et al. (1993), Israelachvili and Wennerstrom (1996), and Israelachvili (1991, Section 13.5).

By definition, hydrophobic surfaces are those that repel water (water contact angles $\geq 90^\circ$); so we should not be surprised that, when two hydrophobic surfaces immersed in water are brought together, water is spontaneously repelled from between the surfaces when the gap is sufficiently small. Associated with this exodus of the water is an attractive force acting between the two hydrophobic surfaces, called the hydrophobic attraction. When the two surfaces come into contact, we know that the work of adhesion equals $2\gamma_i$, where $\gamma_i$ is the hydrophobic surface–water interfacial energy. In SFA experiments, the hydrophobic attraction is found to decay exponentially with separation distance $D$ with a decay length $\lambda_o = 1–2$ nm; so we can write its interaction energy as

$$W = -2\gamma_i e^{-D/\lambda_o}. \quad (7.46)$$

While eq. (7.46) describes well the strength of hydrophobic attraction as a function of separation distance, it does not enlighten us as to how the interaction originates at the molecular level (which, as with the hydration force, is still largely unclear). One approach is to look at the similar hydrophobic attraction that occurs between hydrophobic molecules dissolved in water. In this case, an attraction occurs because water molecules have difficulty hydrogen bonding to a dissolved hydrophobic molecule, so instead they form a hydrogen bonding network around the isolated hydrophobic molecule that maximizes the hydrogen bonding among the adjacent water molecules. Since this hydrogen bonding network has a more ordered structure compared to liquid water, it comes at a price in free energy due to the reduction in entropy. When two hydrophobic molecules, surrounded with their structured water shells, come together, the free energy is reduced due to the gain in entropy when some of the structured water is released from the shells and returned to the more disordered bulk liquid state. A similar effect is thought to occur when two hydrophobic surfaces are brought together (Eriksson et al. 1989; Wallqvist and Berne 1995; Forsman et al. 1996; Lum et al. 1999). Evidence that water molecules adjacent to a hydrophobic surface reorient themselves away from the surface to increase hydrogen bonding with nearby
Another possible mechanism for hydrophobic attraction is as follows. Dissolved gas forms nanometer sized bubbles or “nanobubbles” on the hydrophobic surfaces; these bubbles coalesce to bridge the gap when the surfaces are brought together; since the water contact angle is greater than 90°, the bridging bubbles act like a meniscus that exerts an attractive force described by eq. (6.30). This nanobubble mechanism leads to a different, much longer range type of hydrophobic attraction than described by eq. (7.46), and can appear at tens of nanometers separation (Tyrrell and Attard 2001). Some have also suggested that dissolved gas can spontaneously form bridging bubbles between the hydrophobic surfaces without having to go through the intermediate phase of first forming bubbles on the individual surfaces (Rabinovich et al. 1982; Yaminsky and Ninham 1993; Andrienko et al. 2004).

If hydrophilic and hydrophobic surfaces are brought together in water, neither hydration repulsion nor hydrophobic attraction is observed. These “Janus” type interfaces, however, have other interesting tribological properties. For example, water films sheared in an SFA between a hydrophilic and a hydrophobic surface are observed to have much lower friction than water sheared between two hydrophilic or two hydrophobic surfaces (Zhang et al. 2002).

### 7.5 Contact electrification

The ancient Greeks were the first to document the phenomenon of contact electrification or triboelectricity. One of the Seven Sages of Greece, Thales of Miletus (624–547 B.C.), described experiments involving rubbing amber with cat’s fur; the amber could then be used to pick up small objects. We now know that this phenomenon is due to the transfer of electric charges during contact, leaving the surfaces with net electrical charges that electrostatically attract other objects. While some make the distinction between triboelectricity, from rubbing, and contact electrification, from simple contact, here we treat the two situations as the same contact electrification phenomenon.

Contact electrification is easily observed in everyday life. For example, when the humidity is low, walking across a carpeted floor generates a net charge on your body, that discharges with a slight shock when you touch a conductor like a metal doorknob. Or your plastic comb will start to attract hair after a single pass through dry hair. If the hair is wet, water conducts away the surface charges, neutralizing the effect.

Since contact electrification occurs frequently during contact of dissimilar materials, its potential impact needs to be taken into account when designing tribological systems. For example, an adhesive electrostatic force can contribute to a loading force on rubbing surfaces, leading to high friction and wear and perhaps premature failure. Also, the electrostatic charge built up by contact
electrification can suddenly discharge (an event called electrostatic discharge or ESD), damaging surrounding materials and electrical circuits.

Contact electrification is not always detrimental, however, as evidenced by the several important technologies developed to exploit the phenomenon:

- Photocopiers and laser printers use contact electrification to charge toner particles, which are then manipulated with electrostatics.
- Some mining processes use contact electrification to separate the ore.
- In electrostatic spray painting, contact electrification gives the paint droplets a net charge before they are directed towards a grounded conducting surface to be painted.
- A van de Graaff generator uses contact electrification between a roller and an insulating belt to generate a high electric field that is then used to generate positively charged ions and to deposit them onto the belt. These charges are later removed from the belt to create a constant current source capable of achieving extremely high voltages—as high as 20 million volts.

7.5.1 Mechanisms of contact electrification
7.5.1.1 Conductor–conductor contact

The charge density $\sigma$ on a conductor surface such as a metal is related to the electric field $E$ at the surface by $E = \sigma / \varepsilon_0 \varepsilon_r$, where $\varepsilon_0 \varepsilon_r$ is the dielectric permeability of material outside the conductor. On a conductor, the electrostatic potential is constant over the surface, implying that the electric field is oriented normally to a conductor surface. To calculate the electrostatic force between two conductors, we first note that the stored electrostatic energy equals $CV^2/2$, where $V$ is the electrostatic potential difference between the two conductors and $C$ their capacitance. The gradient of the electrostatic energy with respect to the separation distance $D$ gives the electrostatic force:

$$F_{el} = \frac{V^2}{2} \frac{\partial C}{\partial D}.$$  \hfill (7.47)

With the sign convention used in this chapter, the $F_{el}$ is negative as the electrostatic force between two conductors is always attractive.

Figure 7.15 shows the simplest geometry for calculating the electrostatic force acting between conductors: two parallel plates. If $D \ll \sqrt{A}$ where $A$ is the area of the plates, the electric field between the plates is $E = V / D$, so the charge density at the conductor surfaces is $\sigma = \pm \varepsilon_0 \varepsilon_r V / D$. Since the capacitance of two parallel conducting plates is

$$C_{plates} = \frac{\varepsilon_0 \varepsilon_r A}{D}.$$  \hfill (7.48)
Fig. 7.15. Two parallel plate conductors $\alpha$ and $\beta$ with bias voltage $V_{\text{bias}}$ applied between them. The electrostatic potential from the contact potential and the bias voltage generates electrostatic forces $F_{\text{el}}$ that act to decrease the separation $d$. The plates are held in equilibrium by mechanical forces $F_{\text{mech}}$.

The electrostatic force $F_{\text{el}}$ pulling the plates towards each other is given by

$$F_{\text{el-plates}} = \frac{-\varepsilon_0 \varepsilon_r A}{2D^2} V^2. \quad (7.49)$$

The electrostatic potential difference $V$ is the sum of the external applied bias voltage $V_{\text{bias}}$ and the internal contact potential $V_c$:

$$V = V_{\text{bias}} + V_c. \quad (7.50)$$

The contact potential $V_c$ is the potential difference between plates arising from contact electrification of the conducting surfaces.

Figure 7.16 illustrates how two metals with different work functions develop a contact potential between them, in terms of electron energy levels (Harper 1967; Lowell and Rose-Innes 1980). If initially, as in Fig. 7.16(a), the metals are not electrically connected and have no potential difference between them, the electron energy is constant in the gap between the two metals. When an electrical connection is made between the two metals as shown in Fig. 7.16(b), electrons flow to bring the system into thermodynamic equilibrium, which occurs when the metals’ Fermi levels equilibrate to the same energy. This exchange of electrons leads to the surface with a higher work function having a net negative charge and the surface with a lower work function having a net positive charge; the opposite surface charge densities then generate an electrostatic potential difference $V_c$.
Fig. 7.16. Energy of an electron inside and outside two metals with different work functions $\varphi_A$ and $\varphi_B$. (a) When no electrical connection or potential difference exists between the metals. (b) When electrical connection is made between the two metals, the Fermi level energies $E_F$ become equal generating a potential energy difference $eV_c = (\varphi_B - \varphi_A)$ between the vacuum level energies of the two metals. One way that this electrical connection can be made is by bringing the metals into contact or close enough (a few nanometers) for electrons to tunnel through the energy barrier. (c) Applying a bias voltage $V_{\text{bias}} = -V_c$ cancels the potential difference and the electrostatic force between the two metals.

between the two metals proportional to the difference in work functions:

$$\sigma = \varepsilon_0 \varepsilon_r E = \varepsilon_0 \varepsilon_r \left( \frac{V_c}{D} \right)$$

(7.51)

$$V_c = \left( \varphi_A - \varphi_B \right) / e$$

(7.52)

Work functions are difficult to predict in practical situations as they depend not only on the type of metal, but also on the presence of oxides, lubricants, and other surface contaminants.

It should be remembered that the act of bringing two conductors into contact establishes an electrical connection, where electrons suddenly discharge across the interface at the point of contact. During this discharge, the contact potential and electrostatic attractive force are created. This new electrostatic force now needs to be overcome, along with the non-electrostatic adhesive forces, during separation of the surfaces, resulting in another component of adhesion hysteresis (previously discussed in Section 5.4.3).

Applying an external bias voltage equal and opposite to the contact potential ($V_{\text{bias}} = -V_c$), as shown in Fig. 7.16(c), can be used to cancel the contact potential and the resulting surface charge density. When $V_{\text{bias}} \neq -V_c$, an attractive electrostatic force exists between the conducting surfaces. MEMS and other microactuator devices often use electrostatic force to actuate the motion of small mechanical transducers, so the magnitude of the contact potential needs to be taken into account when designing the bias voltage circuitry.
Example: Recording head slider flying over a disk in a disk drive

During read-write operations of a disk drive, a recording head slider flies on a stiff air bearing over a rotating disk surface. While it is desirable to have the recording head fly as closely as possible to the disk surface to achieve the highest possible recording density, disk drives are designed so that the head–disk spacing is always greater than a minimum value to ensure that high speed contacts do not damage the recording head or the disk. In current drive designs, this minimum value is typically just a few nanometers, while the mean spacing between the disk and the portion of the slider closest to the disk is typically between 7 and 11 nm. At these small separations, van der Waals and electrostatic forces can significantly reduce the flying height from the value predicted from the air bearing simulation calculations used to design the air bearing surface of the slider.

The recording head slider and disk are both made of reasonably well conducting materials: in the slider, sintered Al$_2$O$_3$-TiC is used, while, in the disk, the magnetic layer is a cobalt based alloy. Also, both the slider and disk are overcoated with a few nanometers of amorphous carbon, and a monolayer of lubricant is deposited on the disk. Since the slider and disk materials are dissimilar, a significant contact potential is expected: indeed, a number of experiments have measured slider–disk contact potentials ranging between $-1$ V to $+1$ V (Kiely and Hsia 2002; Knigge et al. 2004; Feng et al. 2005). Some interesting questions arise: Under what conditions does the electrostatic force from the contact potential exceed the van der Waals force and to what extent do these forces reduce the expected slider–disk spacing? The answers to these questions can generally be extended to other situations involving the attractive forces between dissimilar materials.

To calculate the van der Waals force, we first need to know the Hamaker constant for a slider flying over a disk. White et al. have evaluated Hamaker constants for the multilayer system of the slider–disk geometry: in the range of spacings from 2 to 11 nm, they find that the Hamaker constant $A_{svd}$ for this slider–vapor–disk geometry is dominated by the overcoats and lubricant films and lies in the range from 0.9 to $1.5 \times 10^{-19}$ J (White et al. 2005; Crone et al. 2006). For our estimate here, we will use the mid-value: $A_{svd} = 1.2 \times 10^{-19}$ J.

For this example calculation, we consider a particular style of slider design where a small pad at the location of the recording head protrudes from the rest of the slider air bearing surface that faces the disk (see Fig. 7.17). This slider design is sometimes used to minimize the possible contact area of the slider with the disk surface (Singh et al. 2004). Since the height of this pad (typically $>10$ nm) is greater than the pad–disk spacing, the van der Waals and electrostatic forces acting on the pad are much greater than over the other portions of the slider; this greatly simplifies the calculations, as the total van der Waals and electrostatic forces acting on the slider can then be estimated by considering only their contribution over the pad area. We will make the further simplifying
assumption that the slight pitch angle of the slider can be ignored so that we can treat the pad–disk geometry as two parallel surfaces.

If we assume that the pad has a 50 µm × 50 µm area, the van der Waals force acting on the pad is

$$F_{VDW} = - (50 \mu m \times 50 \mu m) \frac{1.2 \times 10^{19} J}{6\pi D^3}$$

$$F_{VDW}(mN) = -\frac{16}{[D(nm)]^3}$$  (7.53)

Inside disk drives, the sliders and disks are typically grounded, so \(V_{\text{bias}} = 0 V\), and assuming a contact potential of \(V_c = 0.5 V\), then the electrostatic force acting on the pad becomes

$$F_{\text{el}} = - (50 \mu m \times 50 \mu m) \left(8.85 \times 10^{-12} \text{ F/m}\right) \frac{(0.5 V)^2}{2D^2}$$

$$F_{\text{el}}(mN) = -\frac{2.8}{[D(nm)]^2}$$  (7.54)

Figure 7.17 plots the forces acting on this pad area, as a function of separation distance, for the van der Waals force expressed by eq. (7.53), the electrostatic force with \(V_c = 0.5 V\) expressed by eq. (7.54), and the electrostatic force when
$V_c = 1.0 \text{ V}$. Since the electrostatic force falls off with a lower power ($D^{-2}$ versus $D^{-3}$ for van der Waals), it becomes the dominant force at larger distances. From the plot in Fig. 7.17, we see that for $V_c = 0.5 \text{ V}$ the crossover occurs at $D \sim 5 \text{ nm}$ while, for $V_c = 1.0 \text{ V}$, at the high end of the likely contact potentials, the crossover occurs at a very small separation: $D \sim 1 \text{ nm}$.

To get a feel for how these forces might affect the slider–disk spacing, we will use $10^6 \text{ N/m}$ as a typical value for the stiffness of the air bearing that keeps the pad and disk separated. At the crossover point of $D = 5 \text{ nm}$ for $V_c = 0.5 \text{ V}$, $F_{\text{el}}$ and $F_{\text{VDW}} = 0.12 \text{ mN}$ so the net force pulling the slider towards the disk is 0.24 mN. This leads to a reduction of the slider flying height of $\Delta D = (0.24 \text{ mN})/(10^6 \text{ N/m}) = 0.24 \text{ nm}$. In this example, the small pad area results in the relatively small change in fly height (5%), as the pad reduces the effective area over which the van der Waals and electrostatic forces act. For sliders without such a pad, several numerical simulations have shown that the van der Waals and electrostatic forces cause much larger spacing reduction (Thornton and Bogy 2003; Knigge et al. 2004; Crone et al. 2006). From Fig. 7.17, we can also see that $dF/dD$ increases rapidly with decreasing spacing and will exceed the air bearing stiffness at sufficiently small spacings; that is, the attractive forces suddenly pull the slider into the disk surface when the separation gets too small. This snap-in effect limits the practical minimum slider–disk spacing to just a few nanometers.

### 7.5.1.2 Metal–insulator and insulator–insulator Contacts

While the contact electrification of conductors is reasonably well understood, the situation is less clear when one or both of the materials are insulators. Charge transfer at insulator surfaces occurs not only from the exchange of electrons but also from the exchange of ions and charged clusters of atoms. The amount of charge present on an insulator surface tends to increase with each contact or with the amount of rubbing time before eventually reaching a maximum charge density. The continual accumulation of surface charges with repeated contacts or rubbing is thought to originate from the following mechanism. During each contact, the contact electrification is localized to those nanoscale areas where actual solid–solid contact occurs, and, for insulators, the surface charge is slow to leak to neighboring areas. Further surface charges develop when further contacts or rubbing bring fresh areas into solid–solid contact.

The surface state theory has been developed along lines similar to the metal–metal contact electrification theory for describing the charge transfer process of insulators (Davies 1967; Inculet and Wituschek 1967; Lowell and Rose-Innes 1980; Castle 1997). In this theory, an insulator is assumed, as illustrated in Fig. 7.18, to have states, which are localized at its surface and have energy levels within the band gap of the insulator. These states can be occupied by either electrons or ions, and the energy difference between the highest occupied state and the vacuum level corresponds to the insulator’s effective work function. During contact, charges transfer from the surface with the lower effective work function $\Phi_A$ to the surface with the higher effective work function $\Phi_B$, i.e., towards the
direction of equalizing the energy levels of the highest occupied states of the two materials.

If \( N \) is the density of surface states per unit energy per unit area, the maximum possible charge density \( \sigma_L \) that can occur from contact electrification is, in the limit of a low density of surface states,

\[
\sigma_L = -eN(\Phi_A - \Phi_B)
\]  

In the low density limit, \( N \) is so small that the electrostatic potential difference stays near zero and the energy levels of the surface states barely change after the charge transfer [Fig. 7.18(a)].

At the high density limit of surface states, the surface charge density generates enough electrostatic potential energy difference to shift the energy levels of the two materials to the point where the effective Fermi levels equalize as shown in Fig. 7.18(b), resulting in an electrostatic potential difference \((\Phi_A - \Phi_B)/e\). In this high density limit, the maximum possible charge density \( \sigma_H \) that can occur from contact electrification is determined by the electric field between the surfaces:

\[
\sigma_H = -\varepsilon_0 \frac{(\Phi_A - \Phi_B)}{eD}
\]  

where \( D \) is the effective separation distance between the two insulators when in contact and \( \varepsilon_0 \) the permittivity of free space. Eq. (7.56) predicts surface charge densities higher than ever observed in experiments, so represents an upper limit of possible charge transfer. For both the low density [eq. (7.55)] and high density [eq. (7.56)] limits, the surface state theory predicts that the amount of charge exchange depends linearly on the effective work function difference.
Typical maximum surface charge densities observed experimentally on insulators are on the order of $10^{-5} – 10^{-4}$ C/m$^2$. The electrostatic voltages generated by such a high surface charge density can be sufficient to break down air and even some materials. Since the electric fields are strongest when the separation distances are small, the discharges are most prevalent just after separation when the electric field is still high. Several types of discharge processes can occur during separation:

1. Tunneling through the narrow energy barrier when the surfaces are first separated and the gap is still less than a few nanometers.
2. At larger distances, field emission of electrons can occur if the electric field strength becomes comparable to typical field emission strengths $6 \times 10^9$ to $2.3 \times 10^9$ V/m (Horn and Smith 1992).
3. The combination of high electric fields and emitted electrons and charged particles during a tribo-event can lead to the breakdown of the gas separating the surfaces, as described by Paschen’s law. In this discharge process, an avalanche event occurs where ionized gas molecules emit electrons that ionize even more gas molecules; this process continues until these newly created charges neutralize the surface charges that are responsible for the high electric fields.

This last mechanism is one of the principal mechanisms for the very interesting phenomenon of *triboluminescence*, where light is emitted during fracture and rubbing of dielectric materials (Walton 1977). This is particularly strong in piezoelectric materials where the stresses required to fracture lead to polarization of the material before fracture and surface charges after fracture. (Since sugar is a piezoelectric material, triboluminescence can be readily observed by watching someone in a dark room chew on a brittle hard candy with their mouth open.) The high surface charge density then causes a discharge in the gas in the gap between the fractured surfaces, emitting a significant amount of UV and visible light.

Triboluminescence is also frequently observed with non-piezoelectric materials, though it is generally not as intense as for piezoelectric materials. For these materials, other mechanisms in addition to the gas discharge can occur (Chakravarty and Phillipson 2004). For example, rubbing two materials together frequently causes electrons and ions to be emitted and excite a light-emitting plasma in the surrounding gas (Nakayama and Hashimoto 1995).

### 7.5.2 AFM studies of contact electrification

Despite decades of study, the details of how contact electrification happens at the atomic level are still only poorly understood. From the early days of atomic force microscopy (AFM), it has been hoped that this technique, with its ability both to measure the electrostatic force and to image its gradient on the nanometer scale, would quickly clarify the nanoscale details. So far, however, AFM has made...
only a modest contribution to our understanding of contact electrification. Here are a few highlights of these AFM studies:

- Terris et al. have shown that bringing a metal nickel tip into contact with an insulating polymethyl methacrylate (PMMA) sample creates a surface charge that can be imaged with the same tip by sensing the gradient of the electrostatic force (Terris et al. 1989). Surprisingly, they found that the charged regions contain sub-areas with both positive and negative charges, which would appear to be at odds with the surface state theory prediction that surface charge should all have the same sign and have a magnitude scaling with the work function difference between the metal and insulator. This result can be squared with surface state theory if the PMMA surface is assumed to have an effective work function varying across the surface and ranging below and above the tip work function. Terris et al. also observed charged regions much larger than the solid–solid contact area, suggesting that the charge spreads out from the contact regions shortly after contact occurs.

- Gady et al. have measured the force acting on a polystyrene sphere attached to an AFM cantilever after contact with gold and graphite surfaces (Gady et al. 1998). After contacting the graphite surface, a strong electrostatic force was observed, indicating that a net surface charge was created on the sphere surface. For the gold surfaces, only the van der Waals force was observed after contact. Gady et al. claim that these results are consistent with the surface state theory of contact electrification.

- Saint Jean et al. have studied both analytically and experimentally the relative importance of van der Waals and electrostatic forces acting between a metal surface and a metal AFM tip (Saint Jean et al. 1999). As noted before, when a contact potential is present, the electrostatic force always dominates at larger separation distances and the van der Waals force at shorter distances. They claim that the crossover between van der Waals and electrostatic forces occurs at a separation distance equal to half of the tip radius of curvature. Also, a brief review of electrostatic forces between conducting AFM tips and surfaces can be found in the paper of Butt et al. (2005).

- Van der Oetelaar et al. have studied the tribocharging of disk surfaces covered with carbon overcoats and perfluoropolyether lubricants when a contacting AFM tip scans over a small area (van den Oetelaar et al. 2001). Since the decay time of the surface potential depends strongly on oxygen and water vapor, they propose that this transient component occurs due to the removal of oxygen bound to the surface during scanning. The non-transient component is proposed to originate from charges, that are created during the rubbing process, being trapped in states in the energy gap of the insulating carbon film.
7.6 References


MEASURING SURFACE FORCES

The basic concept for measuring the force acting in the normal direction between two solid surfaces is to convert the magnitude of the force $F$ into a measurable displacement. Figure 8.1 illustrates this approach. First, one of the solids is attached to a flexible spring element with a known spring constant $K_{spring}$. Then, the two surfaces are brought incrementally closer together by moving either the base of the other solid or the base of the spring a distance $\Delta D_0$. Measuring the spring element deflection $\Delta D_{spring}$ and applying Hooke’s law determines the change in force $\Delta F$ due to the displacement $\Delta D_0$:

$$\Delta F = K_{spring} \Delta D_{spring}.$$ (8.1)

Rather than working with changes in force $\Delta F$, it is more convenient to work with the total force $F$ measured relative to some reference. By convention, the force $F$ is set equal to zero at large separations ($D = +\infty$).

Once the force $F$ has been determined as a function of sample position, this can be converted into a plot of force versus separation distance ($F$ vs. $D$) by using the following relationships to determine the separation distance $D$:

$$\Delta D = \Delta D_0 - \Delta D_{spring}$$

$$D = D_0 - D_{spring}.$$ (8.2)

Figure 8.2 shows the typical shape of a $F$ vs. $D$ curve during approach and retraction. For the separation distance, the convention is to choose $D = 0$ when $F = 0$ while the two solids are in repulsive hard wall contact, as done in Fig. 8.2. The rationale for this choice is that the measurement zero for separation distance then lies within an atomic diameter of the theoretical zero for the separation distance of the surfaces’ outermost layer of atoms.

Fig. 8.1. Idealized setup for measuring the force acting between two surfaces.
Fig. 8.2. A typical force-versus-separation distance ($F$ vs. $D$) curve where the force is attractive at large separation distances and repulsive at small distances. As the two surfaces approach each other, the force gradient may exceed the spring constant ($dF/dD > K_{\text{spring}}$) and the upper surface snaps into contact. Similarly, when the surfaces are pulled apart, the surfaces will snap apart where $dF/dD > K_{\text{spring}}$. (See also Fig. 5.19.)

One of the principal uses of surface force measurements is to confirm the atomic origins of the adhesive and repulsive forces between surfaces. Since many of these atomic scale forces act over just few ångstroms, one needs to measure the distances—$\Delta D$, $\Delta D_0$, and $\Delta D_{\text{spring}}$—with a precision of an ångstrom or better to determine these forces accurately. Fortunately, current displacement measurement techniques can routinely measure sub-ångstrom displacements.

Figure 8.2 also illustrates a frequent complication with force measurements: The attractive force causes the two surfaces to snap in or out of contact during either the approach or the retraction. This also leads to adhesion hysteresis, where the maximum attractive force on during retraction is less than the maximum experienced during approach (Section 5.4.3).

Snap-in/snap-out occurs when the spring element is too soft. If $K_{\text{spring}}$ becomes smaller than the gradient of the attractive force, then, the rate of increase in attractive force during an approach becomes greater than the rate of increase of the counteracting force of the spring element, and a sudden snap into contact occurs where the gradient exceeds the spring constant $K_{\text{spring}}$. Since the surfaces jump over those separation distances where high gradients occur, the force cannot be measured over these separation distances. A similar snap-out occurs when the surfaces are retracted. The advantage, however, of using a soft spring element in force measurements is that this improves the sensitivity for small forces. In principle, the snap-in/snap-out problem could be avoided by using a suitably stiff spring element, but at the price of losing force sensitivity at other separations.
Next we discuss two of the techniques, SFA and AFM, that have been developed for measuring forces between surfaces with separations of atomic scale dimensions.

### 8.1 Surface force apparatus

Since the mid-1970s, the surface force apparatus (SFA) has been a major workhorse for measuring molecular scale forces.

Figs. 8.3 and 8.4 illustrate how an SFA works (Israelachvili and Adams 1978; Parker et al. 1989; Israelachvili and Mcguiggan 1990). At the heart of an SFA are two curved pieces of mica (radius of curvature \( \sim 1 \text{ cm} \)) that are held in a

![Schematic of the SFA](image)

**Fig. 8.3.** Schematic of the surface force apparatus (SFA) for measuring forces acting normally to crossed cylindrical surfaces. The top cylindrical lens is attached to a piezoelectric (PZT) tube that provides small changes \( \Delta D_0 \) in the vertical position of the top curved surface. The bottom lens is attached at the end of a double cantilever that bends a distance \( \Delta D_{\text{cantilever}} = \Delta F/K_S \) when a force \( \Delta F \) acts normal to the surface. (The double cantilever arrangement is used to ensure that the bottom lens does not tilt when the cantilevers bend.) An inchworm motor provides for course positioning of the bottom lens relative to the top lens. The changes in separation between top and bottom surfaces are measured by optical interference of white light that enters through a bottom view port and exits through a microscope objective focused on the contact region of the two curved surfaces. The enclosure can be filled with either vapor or liquid.
Fig. 8.4. Enlarged view of the cross-cylinder geometry of the surface force apparatus. Sheets of mica, about 2μm thick, are silvered on their backsides to be semi-reflective and then attached with epoxy to the crossed cylindrical lenses shown in Fig. 8.3. The sides of the mica sheets facing each other are cleaved to provide atomically smooth surfaces over most of their surface area. On the right: A cross-sectional view of the two mica surfaces together with a meniscus formed by adding a small liquid droplet. Those wavelengths of the white light that form standing waves between the surfaces are transmitted through to the spectrograph and charge-coupled device (CCD) camera. From the wavelengths $\lambda$ transmitted, as measured by the spectrograph, the distance between silvered surfaces can be determined as a function of lateral position $x$.

crossed cylinder geometry so as to form a circular or elliptical contact area when their surfaces touch. One mica sheet is attached with epoxy to a cylindrical lens mounted on a cantilever spring element, which deflects a distance $\Delta D_{\text{cantilever}}$ when a normal force acts between the two mica sheets. The other mica sheet is epoxied to another cylindrical lens mounted on a PZT tube that provides a few microns of motion in the vertical direction ($\Delta D_0$).

The separation distance $D$ between the two mica sheets is measured by multi-beam interferometry. With this technique slightly transparent silver films are deposited onto the backsides of the mica sheets; an intense source of white light
is focused onto the bottom lens; and a small fraction of the incident light enters
the optical cavity formed by the two silvered surfaces. Those wavelengths of
light that constructively interfere within this cavity emerge from the optical
cavity and enter the spectrograph. The spectrograph disperses these fringes of
equal chromatic order (FECO) onto the CCD camera, generating an output
of the sort shown schematically in Fig. 8.4. Algorithms can be used for deter-
mining, from the location of these fringes, the separation distance between the
silvered mica surfaces, refractive indexes of the material between them, and the
cross-sectional shape of the contact zone (Israelachvili 1973; Horn and Smith
1991). Sometimes optical coatings are used in place of the silver coatings, par-
ticularly, if other optical probes are to be used to further study the confined
liquids (Mukhopadhyay et al. 2003). In addition to the FECO interferometry
technique, two other techniques—capacitance (Tonck et al. 1988; Frantz et al.
1996) and optical fiber interferometry (Frantz et al. 1997)—have also been used
for measuring the changes in the separation distance between SFA mica surfaces.

In an SFA, zero separation of the two facing mica surfaces ($D=0$) is deter-
dined by bringing the surfaces into contact in the absence of liquid or adsorbed
layers, so that the atoms on the opposing mica surfaces are in intimate contact.
The distance between the silvered surfaces under this intimate contact condition
is measured first, using the FECO technique, and then this distance subtracted
from future measurements when liquid or polymer molecules are present between
the surfaces. Using eq. (8.2), the SFA cantilever deflection $\Delta D_{\text{cantilever}}$ is deter-
dined from the separation $\Delta D$ measured by the FECO technique and the
distance $\Delta D_0$ that the top surface is moved (i.e., $\Delta D_{\text{cantilever}} = \Delta D_0 - \Delta D$);
then, the normal force is determined from Hooke’s law [eq. (8.1)]. For most SFAs,
a typical sensitivity for $\Delta D$ is 1 Å; so, with a relatively soft spring constant of
$K_{\text{spring}} = 100 \text{ Nm}^{-1}$, the force sensitivity is 10 nN.

The mica sheets are a key feature of the SFA technique. These sheets are
formed from single crystals of mica that can easily be cleaved along a certain
crystal plane to form a surface atomically flat over centimeter sized areas. For
such surfaces, effectively no roughness is present, except for the periodic arrange-
ment of mica surface atoms (roughness $<1$ Å). Since the surface roughness is
negligible, this vastly simplifies the interpretation of SFA results. The lack of
surface roughness also means that, at $D=0$, the atoms on opposing surfaces are
uniformly in contact with a spacing approximately that of the distance between
the bulk atomic layers.

The SFA technique has been extended beyond mica surfaces by coating the
cleaved mica surfaces with other materials such as molecular monolayers, oxides
(Horn et al. 1989; Hirz et al. 1992; Ducker et al. 1994), and metals (Parker
and Christenson 1988; Levins and Vanderlick 1995). These materials are usually
deposited onto the mica in such a way that the resulting surfaces are nearly
atomically smooth. With such smooth surfaces, it has become fairly straightforward to observe molecular level force effects that would otherwise be smeared out
by roughness, such as the effects of solvation forces as discussed in Section 7.4.1.
Fig. 8.5. Comparison of the contact regions for three different techniques for probing molecular scale forces: (a) Surface force apparatus (SFA) with liquid molecules forming layers between the surfaces. The repulsive forces from the molecular layers sandwiched between the mica surfaces elastically deform the mica sheets and the glue attaching them to the cylindrical lenses, to form two parallel surfaces in the region of high contact pressure. In the figure, the scale normal to the surface is greatly exaggerated: the parallel region typically extends tens of microns across, while the gap between the surfaces is only a few nanometers. (b) Atomic force microscope (AFM) with an atomically sharp tip and no meniscus around the contact. Ultra-high vacuum conditions are typically required to achieve a contact without a meniscus of contaminants. (c) AFM with a spherical particle tip coated with a monolayer of molecules. This type of AFM, where the tip is coated with a well defined molecular entity, is referred to as chemical force microscopy.

Another feature of the SFA is illustrated in Fig. 8.5(a). When the mica sheets squeeze a liquid between them, they elastically deform so that the facing surfaces become parallel to each other. This flexing of the thin mica sheets is possible because of the easy compression of the soft epoxy used to
attach the mica to the cylindrical lenses. Most liquids generate a repulsive force sufficient to deform the mica when confined to a few nanometers of thickness between the mica sheets. This allows these liquids to be studied in a confined geometry between two atomically smooth, parallel surfaces, a feature that has been extensively used to study lubricating properties of molecular films.

For modest applied loads, this parallel region is typically tens of microns across. So, while the SFA has great force sensitivity to molecular forces at nanometer separations, measured with ångstrom precision, these results are averaged over large nominal contact areas. If these forces need to be measured with a high lateral resolution, typically some variation of an atomic force microscope is used, as described in the next section.

8.2 Atomic force microscope

Since its introduction in 1986 (Binnig et al. 1986), atomic force microscopy has evolved into a versatile technique for measuring the surface forces at small separation distances, in addition to its use in measuring surface topographies (Section 2.4.1). Figure 8.6 shows an older style atomic force microscope (AFM) developed at the IBM Almaden Research Center, where a wire serves as the cantilever spring element whose deflection $\Delta D_{\text{cantilever}}$ is measured by optical interference.

While many types of structures have been used over the years as AFM cantilevers, for most modern AFM instruments, cantilevers are made, using microfabrication techniques, to be very thin ($0.5–1 \mu m$) to provide for a low spring constant (0.1 to 50 N/m) and are made fairly short ($<200 \mu m$) to provide for a high resonance frequency (10 kHz to 1 MHz). For modern AFMs, the cantilever deflection is typically measured by the optical beam deflection technique [Figs 2.6(c), 2.7, and 8.5(b)].

Sharp tips are fabricated to protrude from the free end of the cantilever beam with ultra small radii of curvature (typically, $R < 10 \text{ nm}$) at the end. For AFM imaging, the tip sharpness ensures that, when the tip scans across the sample surface, the contact zone is only a few atoms across, resulting in a high—potentially sub-nanometer—lateral resolution. For force measurements, such a sharp tip means that only a small number of atoms and molecules at the end of the tip are involved in generating the forces at small separation distances. Indeed, the main advantage of using an AFM for studying contact forces is its ability to measure the forces operating when the first few atoms on the summit of a single asperity touch the atoms on the opposing surface. One disadvantage of these sharp tips is that force is much smaller than for blunter tips, making it difficult to measure. Another complication with working with such sharp tips is that they do not remain sharp after a few contacts with sample surfaces, due to plastic deformation of the tip; it is not uncommon for a tip with an initial radius $<10 \text{ nm}$ to blunt to a few tens of nanometers.
One clever solution to this blunting of AFM tips is to attach a narrow carbon nanotube onto the end of the tip (Dai et al. 1996; Wong et al. 1998). Since these tubes are fairly flexible, they can bend when the repulsive force exceeds the small threshold needed to buckle the tube. This limits the stresses applied to the tip during contact, minimizing the damage to the end of the nanotube. Since the diameter of a carbon nanotube is typically <10 nm, high resolution imaging of surface forces can be achieved with little risk of damage, and, even if the end of the nanotube breaks off, the tube diameter, which determines the resolution, remains constant.

One active area of research involves using tips covered with a well defined molecular entity (Noy et al. 1997). Then, the interactions force between this entity and that on the opposing sample surface can be measured and mapped.
over the sample surface. Force microscopy done with these functionalized tips is often referred to as chemical force microscopy (Noy et al. 1997). Functionalized tips can be purchased from several AFM cantilever vendors. The three most common ways of preparing tips with functionalized surfaces are:

1. Silanization of a tip made of either silicon nitride or silicon oxide
2. Coating the tip first with gold then self-assembling thiols onto the gold film
3. Functionalizing the end of a carbon nanotube tip (Wong et al. 1998).

Another approach for obtaining a tip with a well defined chemically active layer on its surface is to replace the standard sharp AFM tip with a smooth, spherical particle with a well defined diameter and then to coat that particle with the molecular entity that one wants to study. Figure 8.5(c) shows a spherical particle at the end of an AFM cantilever coated with a functionalized monolayer. Using a spherical particle at the end of the cantilever has several advantages:

1. Since the sphere radius is much larger than a standard AFM tip (microns rather than nanometers), the force is much larger, and, hence, easier to measure.
2. Since the sphere’s radius is known, the measured $F$ vs. $D$ curve can be converted into the energy per unit area versus separation distance curve by using the Derjaguin approximation (Section 6.1). For this method to work well at small separations, the spherical particles need to be as smooth as possible, as the Derjaguin approximation is only valid at separation distances significantly greater the surface roughness.

If high lateral resolution is needed for an AFM force measurement (for example if one wants to measure the force acting between a single atom protruding from the end of the tip and a single atom sitting on a sample surface) the force is typically too small to be measured simply by using the small cantilever deflection. For this type of measurement, a more sensitive force detection method is to oscillate the cantilever at its resonance frequency and measure the shift of the resonance frequency as a function of separation distance (Dürrig 2000; Giessibl 2003). This method determines the derivative of the force acting on the tip at its closest approach during the oscillation. Integrating this derivative over separation distance, one obtains the net force acting on the tip. This oscillating method also has the advantage that one can measure the attractive force for smaller separation distances than with the static deflection method, without snap-in, as most of the tip oscillation occurs at a safe distance away from the sample.

In a later section in this chapter, we present an example of the use of this oscillation method to measure the van der Waals force between a sharp silicon tip and a silicon surface.
8.3 Examples of forces acting on AFM tips

As discussed in Chapters 6 and 7, when two solids are brought in close proximity, numerous types of forces act between them. In the remaining sections of this chapter, we discuss a few examples of these forces as they act on AFM tips. The purpose here is

1. To illustrate the capabilities of AFM to measure and distinguish between the different surface forces
2. To illustrate under what conditions these different surface forces become important
3. To illustrate the magnitude of these forces.

The results discussed here represent only a miniscule fraction of the AFM results in the published literature. For a more extensive survey of what has been learned about surface forces using AFM, the reader is referred to several recent review papers (Carpick and Salmeron 1997; Cappella and Dietler 1999; Janshoff et al. 2000; Giessibl 2003; Butt et al. 2005). A similarly extensive literature also exists regarding surface force measurements by SFA; the literature up to 1991 was reviewed in the book by Israelachvili (Israelachvili 1991).

8.3.1 Van der Waals forces under vacuum conditions

As discussed in Chapter 7, van der Waals forces are always present between surfaces. Consequently, if an AFM tip is brought near a surface in vacuum or gas, it will always experience an attractive van der Waals force. For separation distances much smaller than the tip radius, we can approximate the tip–sample geometry as being equivalent to the sphere-above-flat geometry. Using the relationship for a sphere-above-flat geometry (Fig. 7.4), the van der Waals force on an AFM tip can be estimated as

\[ F_{vdw \ sphere-\ flat}(D) = -\frac{A_{tip-v-sample}R}{6D^2}. \]  

(8.3)

For example, if an AFM tip contacts a flat surface where

- tip radius \( R = 10 \text{ nm} \)
- Hamaker constant \( A_{tip-v-sample} = 10^{-19} \text{ J} \)
- separation distance in contact \( D = 0.165 \text{ nm} \) (Section 7.3.4)

then

\[ F_{vdw \ sphere-\ flat}(D_{contact}) = -\frac{(10^{-19} \text{ J})(10 \text{ nm})}{6(0.165 \text{ nm})^2} = -6.1 \text{ nN} \]  

(8.4)

This magnitude of van der Waals force (6.1 nN) is readily measurable with a wide range of AFM cantilevers; however, it quickly drops off as the tip is moved away from the sample and, by \( D = 1 \text{ nm} \), the van der Waals force acting on this
10 nm tip is reduced to only 0.17 nN. Consequently, the van der Waals force at a few nanometers’ separation can be difficult to measure unless one uses a low spring constant cantilever, but this will make the AFM measurement susceptible to the snap-in and snap-out behavior discussed in Fig. 8.2.

If a tip with a larger radius is used, the van der Waals force, which scales with radius, becomes easier to measure. Figure 8.7(a) shows the simulated $F$ vs. $D_{\text{sample}}$ curves for the van der Waals force acting at an AFM tip with radius $R = 30$ nm for two different cantilever spring constants. For the very weak cantilever spring constant of 0.01 N/m, the van der Waals force induces a substantial cantilever deflection (a few ångstroms at a separation distance of 20 nm),

![Diagram](a) $k = 0.1$ Nm$^{-1}$

![Diagram](b) $k = 0.01$ Nm$^{-1}$

**Fig. 8.7.** (a) Simulated van der Waals $F$ vs. $D_{\text{sample}}$ curves during approach (left) and separation (right) for a 30 nm radius AFM tip near a surface where the Hamaker constant $A_{\text{tip} - \text{v} - \text{sample}} = 10^{-19}$ J. Here, $z_F$ is the cantilever deflection ($D_{\text{cantilever}}$) and $z_p$ is the distance that the sample moves ($D_{\text{sample}}$). (b) Simulated van der Waals force $F$ vs. $d$ (separation distance) curve for a silicon tip interacting with a silicon surface. From Janshoff et al. (2000), with permission from Wiley-VCH Verlag GmbH and Co, KG.
but this weak cantilever also results a “snap-in” on approach at a separation of 10 nm and a “snap-out” on withdrawal to ~1100 nm separation. Increasing the cantilever spring constant by a factor of ten reduces the snap-in/snap-out distances, but also decreases the force sensitivity by a factor of ten.

While eq. (8.3) provides a reasonable estimate of the van der Waals force at small separations, once contact occurs, the adhesive forces distort the surfaces somewhat away from the sphere-on-flat geometry. As discussed in Section 6.2.2, the coupling of adhesive forces with elasticity increases the contact zone area, changing the magnitude of the expected adhesive force. These effects can be analyzed using the JKR (Johnson et al. 1971), DMT (Derjaguin et al. 1975), Maugis (Maugis 1992), or MYD (Muller et al. 1980) theories, with the choice of theory depending on the relative strengths of the adhesive force and elastic modulus. Figure 8.7(b) shows the calculated force, as a function of separation distance, from the combination of the van der Waals interaction and the repulsive elastic contact, for silicon tips of different radii as they come into contact with a flat silicon sample. The effect of the elasticity is taken into account with an MYD analysis (Janshoff et al. 2000).

As an example of what has been achieved experimentally in measuring the weak van der Waals force on AFM tips, Fig. 8.8 shows the $F$ vs. $D_{\text{sample}}$ curve for a sharp silicon AFM tip approaching a reconstructed silicon(111) surface under ultra-high vacuum conditions (Lantz et al. 2001). Doing this experiment in ultra-high vacuum not only ensured that an atomically clean silicon surface could be obtained, but also prevented the formation of a meniscus around the tip due to capillary condensation, at small separations.

Even though the force is very weak (only 40 pN at a distance of 20 Å and 3.8 nN at contact at $D \sim 2$ Å), the Dürig method (Dürig 2000) of oscillating the cantilever at its resonance frequency enables force to be measured with a good signal-to-noise ratio. Fitting eq. (8.3) to the data in Fig. 8.8 enabled Lantz et al. to determine that $A_{\text{si-v-si}} R = 9.1 \times 10^{-28}$ J·m. Using the calculated Hamaker constant for silicon $A_{\text{si-v-si}} = 1.865 \times 10^{-19}$ J (Senden and Drummond 1995), the tip radius for this experiment is estimated to be 4.9 nm. Even though this is a very sharp radius, it only represents the average curvature at the end of the tip; atomic scale roughness at the end of the tip can provide an individual protruding atom that enables the atomic lateral resolution in an AFM image.

8.3.2 Capillary condensation of contaminants and water vapor

Few tribological contacts occur between clean surfaces under vacuum conditions. More typically the surfaces are either immersed in a liquid (such as a lubricant or an aqueous environment) or surrounded by a gaseous environment. One might think that surface forces in a gaseous environment should be fairly similar those in a vacuum, as the gas itself does not have much influence on these forces. Unless the gases are very pure, however, the small amounts of contaminant vapors present in most gases can have a significant impact on surface forces either by
Adsorbing onto the surfaces to form a contamination layer or by condensing around the contact points.

For tribological contacts exposed to air, water vapor is the most common contaminant, either adsorbing directly onto surfaces or undergoing capillary condensation around the contacts. Consequently, relative humidity tends to have a major impact on surface forces. As an example of the effect that humidity can have on forces acting on an AFM tip near a sample, Fig. 8.9 shows the $F$ vs. $D$ curves as a tip first approaches and then withdraws from a silicon wafer in 0% and 92% relative humidity. At large tip–sample separations, the force is defined to be zero, and it stays near zero as the sample approaches the tip (the van der Waals force is too small to measure). At point A, a sudden jump in the attractive force is observed due to the formation of a meniscus, as illustrated in Fig. 8.10(a). This meniscus forms even at zero humidity, when no water is available to condense around the tip, indicating that instead trace amounts of some mobile contaminant are still able to condense around an AFM tip at this small separation distance. As the humidity is increased from 0 to 92%, the separation distance $A$ where the meniscus forms moves 30 Å further out as water vapor condenses more easily around the tip and swells the meniscus.
Fig. 8.9. Normal $F$ vs. $D$ curves for a tungsten AFM tip contacting a clean silicon wafer in nitrogen at low and high humidities. The AFM geometry shown in Fig. 8.6 was used for these experiments. The tip radius $\sim 100$ nm and the wire spring constant $K_{\text{spring}} = 50$ N/m. The zero force is defined to occur at large separations, and zero separation is defined to occur at point B when the repulsive force on the inward approach goes through zero. A indicates the separation where a meniscus forms around the tip during approach due to condensation of water vapor (high humidity case) or contaminants (low humidity case). C indicates the point where the meniscus breaks upon separation. Reprinted, with permission, from Binggeli and Mate (1995). Copyright 1995, American Institute of Physics.

As the sample is withdrawn, the meniscus eventually breaks at point C in Fig. 8.9, a process that is illustrated in Fig. 8.10(b). Intuitively, the meniscus should break when the separation distance is approximately twice the meniscus radius $r_1$ (Mate and Novotny 1991). For the high humidity case, the breakage point corresponds to a 350 Å meniscus height, indicating that $r_1 \sim 175$ Å. Since at 92% relative humidity the Kelvin radius is $-65$ Å [eq. (5.11)], this indicates that, at the meniscus breakage point, the meniscus diameter is $\sim 100$ Å.

For the 0% humidity curve in Fig. 8.9, the meniscus force is $F_{\text{men}} = -30$ nN. When the tip is withdrawn, the total adhesive force is $F_{\text{ad–tot}} = -125$ nN, which needs to be overcome to break solid–solid contact. (Notice that a “snap-out” event occurs at that point.) As discussed in Section 6.3.1.2, the total adhesive force is the sum of meniscus force $F_{\text{men}}$ and solid–solid adhesive force $F_{\text{solid–adh}}$ [eq. (6.36)]:

$$
F_{\text{adh-tot}} = F_{\text{men}} + F_{\text{solid-adh}} = -2\pi R\gamma_{\text{L}}(\cos \theta_{\text{tip}} + \cos \theta_{\text{sample}}) - 2\pi RW_{\text{tip–liquid–sample}}.
$$

(8.5)
The solid–solid adhesive force $F_{\text{solid-adh}}$ is due to the tip–sample adhesion energy $W_{\text{tip-liquid-sample}}$ in the presence of the unknown contamination liquid that makes up the meniscus. The tip radius for these experiments was estimated to be 100 nm from SEM imaging, allowing us to estimate $W_{\text{tip-liquid-sample}}$ in this case:

$$W_{\text{tip-liquid-sample}} = \frac{-F_{\text{solid-adh}}}{2\pi R} = \frac{95 \text{ nN}}{2\pi (100 \text{ nm})} = 150 \text{ mJ/m}^2$$

\[ (8.6) \]

8.3.3 Bonded and unbonded perfluoropolyether polymer films

Figure 8.11 shows examples of the normal force in air between a tungsten AFM tip and silicon wafers with either an unbonded liquid polymer film or a bonded soft-solid polymer film, the sort of films frequently used to lubricate surfaces.
Fig. 8.11. Normal $F$ vs. $D$ curves comparing the behavior of a clean Si(100) wafer, a 40 Å thick unbonded liquid perfluoropolyether polymer film, and an 15 Å thick bonded perfluoropolyether polymer film deposited onto a silicon wafer. Reprinted with permission from Blackman et al. (1990). Copyright 1990, American Physical Society.

For comparison, the $F$ vs. $D$ curve is also shown for a nominally clean silicon wafer (upper plot). As before, the sudden increase in attractive force at point A is due to water and other vapors in the air condensing around the tip to form a meniscus.

When the sample is covered with a thin film of liquid polymer (the middle $F$ vs. $D$ curve), a sudden attractive force is observed at point A as the tip approaches the sample. This is also due to a liquid meniscus forming around the tip, but in this case the condensation occurs because the liquid in the polymer film migrates across the sample surface to form the meniscus rather than from vapor condensation.

The lower $F$ vs. $D$ curve in Fig. 8.11 is for a tip contacting a bonded polymer film on a surface. This sample was originally covered with the same type of perfluoropolyether polymer film as for the middle curve, but the polymer was then bonded to the silicon wafer by chemically reacting the end groups of the polymer to the silicon oxide surface, which converts the liquid film into a robust soft-solid film on the surface. The structure of this bonded polymer is sketched in Fig. 8.10(c), and resembles the structure of spaghetti lying on a plate, but with both ends of each spaghetti strand attached to the plate surface. Since the bonded perfluoropolyether film has a low surface energy (17.9 mN/m, Fig. 5.11), it also suppresses the condensation of contamination and water vapor. When the
tip first contacts this bonded polymer film (point A), the attractive force does not increase sharply, as for the liquid film, but rather increases gradually as the tip penetrates into the film [Fig. 8.10(c)].

As the tip–sample separation distance nears zero, eventually the force turns repulsive when the molecules in the contact zone are squeezed out or compressed between the tip and sample surface. In this repulsive area, the slope of the force vs. separation curve gives the stiffness of the contact zone. For the clean silicon wafer, the stiffness is 100 N/m at point B, but rises quickly to 350 N/m when in hard-wall contact at point C. The stiffness $S$ is related to the elastic properties and the size of the contact zone by

$$S = kaE_c$$

(8.7)

where $k$ is a geometric factor between 1.9 and 2.4, $a$ is the radius of contact, and $E_c$ is the composite elastic modulus for the tip and sample (Pethica and Oliver 1987). If we estimate the composite elastic modulus $E_c = 2 \times 10^{11}$ N/m$^2$ and if $S = 110$ N/m, then eq. (8.7) shows that the contact radius is estimated at $a = 3$ Å, indicating that a contact zone a few atoms across forms when the tip first makes solid–solid contact.

To separate the tip and sample surfaces from contact, enough force needs to be applied to overcome the adhesive force. For the clean Si(100) surface (upper curve), the maximum solid–solid attractive force during retraction occurs at point D where tip and sample suddenly jump apart. For this tip–silicon wafer contact, the total adhesive force at pull-off is $F_{\text{tot-adh}} = -5 \times 10^{-8}$ N, of which $-3 \times 10^{-8}$ N is from the meniscus force and $-2 \times 10^{-8}$ N is from solid–solid adhesion. From the middle curve, we see that the addition of the perfluoropolyether polymer liquid film reduces the pull-off adhesive force to just the capillary force, indicating that, in the presence of this particular liquid, $W_{\text{tip-liquid-sample}} \sim 0$.

As the sample is further withdrawn in Fig. 8.11, the attractive forces gradually decrease as the tip tries to break free from the molecules that have gathered in the contact zone. For the liquid perfluoropolyether, the tip has to be retracted 400 Å before it breaks free from the rather large liquid meniscus that forms around the contact zone, while for the other surfaces the break-free distance is 40–50 Å. Breaking free from menisci that form around contact points is a major contributor to the adhesion hysteresis (Section 5.4.3).

8.3.4 Electrostatic double-layer force

Immersing the AFM in a liquid can of course eliminate the influence of capillary forces on an AFM force measurement. As discussed in Section 7.4, the presence of a liquid also leads to new types of forces between the two solid surfaces: solvation, electrostatic double-layer, hydration repulsion, hydrophobic attraction, etc. In this section, we discuss an example of an AFM measurement of the electrostatic double-layer force, a common force in aqueous solutions.

As discussed in Section 7.4.2.1, oxides covered with water develop a surface charge that depends on the water’s pH. This surface charge is balanced by the
counterions dissolved in the water, which segregate towards the surface. Bringing two of these surfaces into close proximity disrupts the counterion densities near the surfaces resulting in a repulsive force called the electrostatic-double layer force, which has a decay length given by the Debye length [eq. (7.44)].

Figure 8.12 plots the force divided by particle radius for a silica particle mounted on an AFM cantilever as it approaches a flat rutile titania crystal in different pH environments. According to the Derjaguin approximation (Section 6.1), normalizing the force on the particle by dividing it by the particle’s radius makes the ordinate axis in Fig. 8.12 equivalent to \( \frac{2\pi W_{\text{silica-water-titania}}}{D} \). This normalization is made easier by using a particle attached to the cantilever rather than a conventional sharp AFM tip, as the particle’s larger radius (2.5 \( \mu \)m) can be measured with a much smaller percent error than a sharp AFM tip (radius \( \sim 10 \) nm).

The density of charge on both surfaces is mainly determined by pH. At high pH, both materials are negatively charged resulting in electrostatic repulsion. In Fig. 8.12 this repulsion is readily observable for pH = 8.8 out to separations of 40 nm. As the pH is decreased below the isoelectric point of titania at pH 5.6, the titania surface becomes positively charged, and the electrostatic repulsion turns into an electrostatic attraction. The silica surface remains negatively charged until the pH is lowered below its isoelectric point at pH 3. The \( F \) vs. \( D \) curve in Fig. 8.12 for pH 3 is clearly attractive for all separation distances prior to solid–solid contact near zero separation. These repulsive and attractive forces at finite

![Fig. 8.12. \( F \) vs. \( D \) curve between a 5\( \mu \)m diameter SiO\(_2\) spherical particle and TiO\(_2\) crystal surface. The different curves are, from top to bottom, at pH = 8.8, 7.2, 6.3, 5.3, and 3.0 with 1 mM KNO\(_3\). Reprinted with permission from Larson et al. (1995). Copyright 1995, American Chemical Society.](image)
separation can be adequately described with the DLVO theory that accounts for both the electrostatic double-layer force and the van der Waals force (Larson et al. 1995).

8.4 References


LUBRICATION

When dry surfaces of metals, ceramics, and most polymers slide against each other, friction coefficients $\mu$ are typically 0.5 or higher. In many practical situations, however, such high values of $\mu$ can imply intolerably high frictional forces and energy losses. For those situations where high friction is not explicitly needed (like the brakes on your car) lubricants are used to reduce friction and wear to acceptable levels.

The benefits of applying liquids and greases to surfaces to reduce friction and wear have been known since antiquity (Dowson 1978). Traces of animal fats have been found on ancient wheel axles, indicating that a practical knowledge of lubrication goes back as far as the invention of the wheel. Mural paintings in ancient Egyptian tombs illustrate the method of applying oils and greases to wooden boards over which large stones, statues, and obelisks could then be slid with relative ease (pulled by only a few hundred slaves) during the construction of these magnificent monuments.

Only with the work of Tower and Reynolds in the late 1880s, however, did scientists and engineers begin to sort out how lubricants alleviate friction and wear, and only in the past few decades has progress been made at understanding lubrication mechanisms at the molecular level.

9.1 Lubrication regimes

The main way lubricants function is by introducing a layer between sliding surfaces that lowers the shear strength. The different types of lubricating layers can be categorized as follows:

- *Hydrostatic lubrication* – The sliding surfaces are separated by a fluid film pressurized by an external supply. The thickness of the fluid film is great enough for the asperities on the opposing surfaces not to make contact.
- *Hydrodynamic lubrication* – In this case, the fluid film separating the sliding surfaces is pressurized by viscous dragging of the lubricant into a narrowing gap. Again, the fluid film thickness is great enough for the asperities on the opposing surfaces not to make contact.
- *Elastohydrodynamic lubrication* – For very thin liquid lubricant films, highly localized pressures within a squeezed lubricant film lead to an enhanced lubricant viscosity and to an elastic distortion of the surfaces. Both effects help prevent asperity–asperity contact from occurring between the sliding surfaces.
• **Boundary lubrication** – Asperity–asperity contact occurs between the sliding surfaces, but the asperity surfaces are covered with a monolayer of adsorbed lubricant molecules that lowers the shear strength.

• **Solid lubrication** – A solid film with low shear strength separates the sliding surfaces.

Within the same lubricated system, the mechanism of lubrication can change as sliding conditions change, as illustrated in Fig. 9.1. This plot, known as the Stribeck curve (Stribeck 1902), illustrates how the friction coefficient of a lubricated bearing changes as the sliding parameters change. Figure 9.2 illustrates the different lubrication regimes of the Stribeck curve for the section of the bearing closest to contact. Moving along the Stribeck curve right to left:

• **Figure 9.2(a): Hydrodynamic lubrication** – Under conditions of light loads and/or high sliding speeds, shearing of the liquid lubricant film provides enough hydrodynamic lift such that the surfaces separate sufficiently for none of the asperities of the opposing surfaces make contact. This is the hydrodynamic lubrication regime.

• **Figure 9.2(b): Elastohydrodynamic lubrication** – With increasing load or decreasing sliding speed or viscosity, the separation distance $D$ decreases. At small separation distances, the hydrodynamic pressure in the lubricant film becomes high enough to deform elastically the surface asperities, delaying the onset of solid–solid contact. This is the elastohydrodynamic lubrication regime and occurs around the minimum in friction coefficient of the Stribeck curve.

• **Figure 9.2(c): Boundary lubrication** – For even higher loads or lower speeds, high pressures squeeze out the liquid lubricant at the contacting asperities, resulting in solid–solid contact and a substantial increase in friction. The friction still remains lower than for sliding on clean surfaces, if a layer of molecules

---

**Fig. 9.1.** Stribeck curve for an oil lubricated journal bearing.
In the lubrication regimes of hydrostatic, hydrodynamic, and elastohydrodynamic, where no solid–solid contact occurs and only a fluid film is being sheared, viscosity is the principal fluid characteristic that determines how well it functions as a lubricant.

### 9.2.1 Definition and units

As discussed in Chapter 3, when a shearing force is applied to a solid, as illustrated in Fig. 9.3(a), it deforms elastically so that the top surface moves sideways a distance $x$, which is proportional to the shear stress $\tau_s$, then stops:

\[
\gamma = \frac{x}{d} = \frac{\tau_s}{G} \quad \text{or} \quad x = \frac{d \tau_s}{G}
\]

where $G$ is the shear modulus. In other words, the solid materials display the property of rigidity: They can withstand a moderate shear stress for indefinite period of time. If the shear stress exceeds the threshold for plasticity, however, the solid breaks.
Fig. 9.3. (a) Shearing a solid against a plate. (b) Shearing a fluid between two plates.

A fluid behaves completely differently in that it has no rigidity at all: if you apply a shearing force, a fluid gives way because it flows, and it continues to flow as long you continue to push on it. Viscosity describes the shear forces that exist in the fluid and, consequently, quantifies the fluid’s resistance to shear.

Suppose we have the situation illustrated in Fig. 9.3(b) where a fluid is being sheared between two parallel plates. The bottom plate is held stationary, and a force $F_f$ is applied to the top plate so that it moves at a velocity $v_0$. For most experimental situations where $v_0$ is not too great, $F_f$ is found to be proportional to the area of the plate $A_f$ and to the shear rate $v_0/d$, where $d$ is the distance between the plates:

$$\frac{F_f}{A_f} = \tau_f = \eta \frac{v_0}{d}. \quad (9.2)$$

The *dynamic* (or *absolute*) viscosity $\eta$ is the proportionality constant between the shear stress $\tau_f$ and the shear rate $v_0/d$. We can rewrite eq. (9.2) to obtain a definition of the dynamic viscosity:

$$\eta = \frac{\tau_f}{v_0/d} \quad (9.3)$$

Within Systeme Internationale units, the shear stress has units N/m² (or *pascal*, Pa) and the shear rate has units (m/s)/m, or s⁻¹, so the dynamic viscosity $\eta$ has units of N·s/m² or Pa·s. In the c.g.s. unit system, the units of dynamic viscosity are dyn·s/cm²; this c.g.s. unit of viscosity has been given the name *poise* (P) after the French doctor Jean Poiseuille (1797–1869) who studied the movement of water through glass tubing of very small diameter in order to understand
how blood flows through capillaries in the human body. Since water has a viscosity close to 0.01 poise, the centipoise (cP) is the unit most commonly used for dynamic viscosity. These units for dynamic viscosity are related to each other by

$$1 \text{ cP} = 10^{-2} \text{ poise} \equiv 10^{-3} \text{ Pa} \cdot \text{s} = 1 \text{ mPa} \cdot \text{s}. \quad (9.4)$$

Typical viscosities of lubricating oils range from 50 to 1000 cP, while air (the most common gaseous lubricant) has a viscosity of 0.019 cP or 55 times less than water.

For eq. (9.2) to be valid, the “no-slip” boundary condition needs to hold: When fluid molecules come into contact with one of the plate surfaces in Fig. 9.3(b), they stick long enough for the average velocity of the molecules next to the plate to be treated as being equal to the plate velocity. To some extent, molecules always slip a small amount while in contact, so the “no-slip” condition is not strictly true, but for most situations the slip is negligible. (A later section covers slip in more detail.) With the no-slip boundary condition, the fluid velocity $v(y)$ between the plates in Fig. 9.3 increases linearly from the bottom plate [$v(0) = 0$] to the top plate, [$v(d) = v_0$] with the velocity profile illustrated in Fig. 9.3(b).

While the dynamic viscosity is the most relevant parameter for lubrication problems, *kinematic viscosity* is often the one tabulated and is defined by the relation

$$\nu = \frac{\eta}{\rho} \quad (9.5)$$

In situations where flow is driven by gravity, the kinematic viscosity turns out to be the more relevant expression for the viscous behavior of the fluid. Since viscosity is usually measured using gravity-driven flow, the kinematic rather than the dynamic viscosity is measured more directly and ends up being the viscosity reported.

In SI units, kinematic viscosity has units of m$^2$/s. The c.g.s. unit is cm$^2$/s, which is called a *stoke* (St), after George Gabriel Stokes (1819–1903), a professor at Cambridge University, who developed much of the mathematical analysis for fluid flow. The most common unit for kinematic viscosity is centistokes (cSt); for example, water has a kinematic viscosity of 1.00 cSt at room temperature. The units for kinematic viscosity are related to each other by

$$1 \text{ cSt} = 10^{-2} \text{ stoke} = 1 \text{ mm}^2/\text{s}. \quad (9.6)$$

9.2.2 *Non-Newtonian behavior and shear degradation*

For low and moderate shear rates ($<10 \text{s}^{-1}$), most fluids follow eq. (9.3), and the shear stress is a linear function of shear rate; these fluids are classified as *Newtonian fluids*. The solid lines in Fig. 9.4 illustrate this Newtonian behavior for several liquids. When the shear stress is not linear with shear rate the fluid is called *non-Newtonian*. The dashed lines in Fig. 9.4 illustrate two types of non-Newtonian behavior: shear thickening and shear thinning.
Latex paint, blood (shear thinning)

Shear rate ($\frac{dv}{dz}$ or $\frac{dg}{dt}$)

Slope = $\eta$

Honey (-200 cP)
Glycerine (1410 cP)
Mineral oil (70 cP)
Water (1 cP)
Wet sand, cornstarch (shear thickening)

Fig. 9.4. Shear stress is linear with shear rate for Newtonian fluids (solid lines) and non-linear for non-Newtonian fluids (dashed lines).

For shear thickening fluids, it’s fairly easily to move them slowly but this becomes harder at higher shear rates. An everyday example of this is a mixture of cornstarch with a modest amount of water. If you tilt the mixture, it flows with the consistency of whipping cream, but if you stir it with a spoon, the mixture becomes remarkably stiff in the region next to the stirring spoon. What is happening? A cornstarch mixture is a concentrated suspension of spherical particles that repel one another. When left unperturbed, the particles arrange in an ordered structure that keeps them reasonably well separated from each other when sheared at a low shear rate. At higher shear rates, the order becomes disrupted, bringing some of the particles into contact resulting in higher friction forces. The repulsive interaction pushes them back so, on average, the spacing between the particles becomes larger. This thickening of the material is called dilatancy. Due to this thickening, the material becomes noticeably drier (as the fraction of water falls) and becomes stiffer (due to the higher meniscus forces between the particles as well as the contact between some of the particles).

Another example of dilatancy is that when you stand on a damp beach, the sand around your feet becomes noticeably drier due to shear thickening.

An exciting new technology based on shear thickening is liquid armor for protective vests (Brown 2006). Liquid armor consists of a mixture of polyethylene glycol liquid and very fine particles of silica, which shear thickens instantly when struck by a hard object, to form a stiff shield. The liquid armor is painted on Kevlar sheets, which are then made into a protective garment. Since the compound fills in the gaps between the Kevlar sheet fibers, it does a better job of spreading the energy of the impact out than the fibers alone. This results in a lighter, less expensive protective vest, more resistant to punctures from stabbing and high speed projectiles.

For shear thinning fluids, it becomes easier to shear the fluid at higher shear rates. Shear thinning is common in suspensions, such as blood and latex paint,
and in liquids composed of long, randomly oriented polymers. For paint, shear thinning is a desirable feature, as the paint spreads easily when sheared by a brush but resists flowing down the vertical wall being painted when the brush is removed. For blood, the application of shear breaks up the aggregates of blood cells, platelets, etc. suspended in the blood plasma while aligning these constituents so that they flow over one another more easily. With the shear thinning of polymer, the application of shear helps to align the polymer chains, making it easier for them to slide past each other, reducing their viscosity. When the viscosity declines at higher shear rates, a lubricated bearing becomes less able to support a load, and the lubricant film thins: hence the term shear thinning.

Obviously, for a liquid, the shear stress cannot increase to infinity with shear rate, as eventually breakdown occurs at the molecular level when the strength of the material is exceeded. Experiments with polymers indicate that, at shear stresses of about one thirtieth of the shear modulus, molecular flow is no longer a thermally activated diffusion process but instead occurs in distinct shear bands. Consequently, at high shear rates, a polymer lubricant becomes plastic and cannot support a shear stress above its limiting shear strength $\tau_{\text{lim}}$. Many possible mechanisms have been proposed for how various materials transition from liquid-to solid-like behavior at high shear rates, but experimental verification of these models has often been difficult to achieve (Bair and Winer 1992; Luengo et al. 1996). Figure 9.5 illustrates a simple rheological model where the shear stress plateaus to a limiting shear strength $\tau_{\text{lim}}$ due to the shear thinning by following the relationship

$$\frac{\tau}{\tau_{\text{lim}}} = 1 - \exp \left( -\frac{\eta_0 \dot{\gamma}}{\tau_{\text{lim}}} \right).$$

(9.7)

![Fig. 9.5. Schematic of how the shear stress plateaus to a limiting shear stress at high shear rates.](image-url)
Shear thinning due to the slippage of lubricant molecules against each other is reversible in that the viscosity returns to its original value when the shear rate is reduced back to the linear region. Some lubricants, however, undergo irreversible degradation at high shear rates. For these lubricants, the molecules are so entangled with one another that high shear rates cannot disentangle them to the extent needed to relieve the stress through slippage alone. Instead, the stress breaks the bonds within the lubricant molecule; the lubricant becomes permanently degraded, and the viscosity no longer returns to its original value at low shear rates. This shear degradation can be thermally promoted by frictional heating and catalyzed by certain impurities in the lubricant or on the sliding surfaces.

Typically, the onset of non-Newtonian behavior occurs when the shear rate exceeds about twice the characteristic molecular frequency $\tau_0^{-1}$ of the liquid. Approximating the interaction between molecules in a liquid by means of a Lennard–Jones potential, a characteristic molecular time $\tau_0$ can be defined:

$$\tau_0 = \left( \frac{ma^2}{\varepsilon} \right)^{1/2}$$

(9.8)

where $m$ is the molecular mass, $a$ the characteristic molecular dimension, and $\varepsilon$ the characteristic interaction energy. For water, which has small mass and size, the time scale is very short, resulting in a very high shear rate value for the onset of non-Newtonian behavior. For polymers, which have much larger mass and length, the onset of non-Newtonian behavior occurs at much more moderate shear rates.

9.2.3 Temperature dependence

Liquid viscosities depend strongly on temperature. For example, the viscosity of water near freezing is 1.8 times higher than at $20^\circ C$. (So, it is not just because it so cold that it is more difficult to swim in freezing water, it indeed takes more physical effort.) For mineral oils—a common type of lubricant—the dependence of viscosity on temperature has empirically been found to follow the relationship:

$$\log \log(\eta/\rho + A) = \text{constant} - c \log T.$$  

(9.9)

Eq. (9.9) indicates that a lubricant’s viscosity drops precipitously with increasing temperature. As the lift generated by a hydrodynamic bearing is proportional to viscosity, it is greatly reduced as the temperature increases. So frictional heating within the bearing can lower the lubricant’s viscosity, leading to a bearing that was initially operating in the hydrodynamic regime to move to the left along the Strubeck curve in Fig. 9.1, into the boundary lubrication regime, resulting in higher friction and wear and early failure of the bearing.

9.3 Fluid film flow in confined geometries

Let’s now look more closely at the fluid confined between the two parallel plates in Fig. 9.3(b), but this time considering the more general case where, in addition to
Fig. 9.6. A fluid between two parallel plates that extend much greater distances in the lateral $x$ and $y$ directions than their separation distance $d$. (a) The pressures and shear forces acting on a fluid element between the plates. (b) Poiseuille flow driven by a pressure gradient $dp/dx$. (c) Couette flow driven by the shear forces generated by the relative motion of the top and bottom plates.

The shear stress applied to the fluid by the plate surface, a pressure gradient exists in the $x$ direction. Figure 9.6 shows schematic side views of the two parallel plates. Sketched within the fluid in Fig. 9.6(a) are the pressures and stresses acting on an infinitesimal-size fluid element that force this element to flow between the plates. For steady-state flow these forces sum up to zero, leading to the relationship

$$\frac{\partial p}{\partial x} = \frac{\partial \tau}{\partial z}. \quad (9.10)$$

Extending eqs. (9.2) and (9.3) to this element:

$$\tau = \eta \frac{\partial v_x}{\partial z}. \quad (9.11)$$
This relationship defines the dynamic viscosity at the microscopic level as the ratio of the local shear stress $\tau$ to the local strain rate $\partial v_x / \partial z$.

Combining eqs. (9.10) and (9.11), we obtain

$$\frac{\partial p}{\partial x} = \eta \frac{\partial^2 v_x}{\partial z^2},$$  

(9.12)

which is the simplified Navier–Stokes equation describing the motion of the fluid element. For the geometries in Fig. 9.6, $\partial p / \partial x$ does not depend on $z$, so integrating eq. (9.12) twice over the $z$ variable obtains the expression for the fluid element’s velocity in the $x$ direction as a function of $z$:

$$v_x(z) = \frac{1}{2\eta} \left( -\frac{dp}{dx} \right) z (d - z) + v_0 \frac{z}{d}$$  

(9.13)

where the no-slip boundary conditions are used of $v_x = 0$ at $z = 0$ and $v_x = v_0$ at $z = d$.

The first term on the right-hand side of eq. (9.13) represents the flow velocity driven by the pressure gradient $dp/dx$, with a negative pressure gradient driving the flow to the right. This pressure-induced flow is called Poiseuille flow, as this is the type of flow that was found in Dr. Poiseuille’s original studies for water through the glass capillaries. The Poiseuille flow velocity has a parabolic dependence on $z$, as illustrated in Fig. 9.6(b).

The second term in eq. (9.13) represents the flow velocity driven by shear due to the relative motion of the two plates. This shear-induced flow is called Couette flow after Maurice Couette who described this type of flow in his 1890 PhD thesis, for the first successful viscometer based on shearing a liquid between coaxial cylinders (Couette 1890). The Couette flow depends linearly on $z$, as illustrated in Fig. 9.6(c) and Fig. 9.3(b).

Eq. (9.13) can be integrated again over $z$ to obtain the volumetric flow rate per unit width of the plate:

$$q = \int_0^d v_x(z) dz = \frac{d^3}{12\eta} \left( -\frac{dp}{dx} \right) + \frac{v_0 d}{2}.$$  

(9.14)

Again, the first term on the left is the pressure-driven (Poiseuille) volumetric flow rate and the second term the shear-driven (Couette) volumetric flow rate.

### 9.4 Slippage at liquid–solid interfaces

If one keeps increasing the shear rate, say for example by factors of ten, it is unreasonable to expect that the resistance to flow, as expressed by eqs. (9.2) and (9.14), keeps rising forever by factors of ten; instead, something eventually gives, either at the walls or within the fluid. An example of where the breakdown occurs within the liquid is the shear thinning at high shear rates; then, slippage occurs between the molecules lowering the liquid’s viscosity. An obvious example for slippage at a liquid–solid interface is when one pushes a piece of glass (which
is a liquid with a super high viscosity) across a solid surface: If enough force is applied to overcome the static friction, it slides with perfect slip rather than flow.

The equations in the previous section for fluid flow assume that the fluid adjacent to a solid surface moves at the same velocity as the solid, the no-slip boundary condition. For most ordinary fluids in most ordinary situations, this assumption works quite well. However, if we look closely at how fluid molecules interact with solid surfaces, we realize that the contacting fluid molecules are going to slip to some degree across the surface. For ordinary liquids flowing in macro-sized channels and gaps, slippage has an imperceptible influence on flow, but, as dimensions become smaller, slip effects become more important.

In the following few sections, we explore the mechanisms for slip at a liquid–solid interface. This is valuable for the several reasons:

- **Scientific importance** – Once we understand how liquid molecules slide against solid surfaces, we will have a good starting point for discussing the nanoscale mechanisms that underlie the dynamic friction of one material sliding against another, such as when we discuss how molecular slippage impacts boundary lubrication in the next chapter.

- **Current technology** – When working with high viscosity liquids—such as suspensions, foodstuffs, and emulsions—taking advantage of slip is often essential for achieving reasonable flow rates (Barnes 1995). Probably, the most prosaic example of this is when you squeeze toothpaste out of its tube: Due to slippage between the paste and the walls of the tube’s orifice, the paste comes out easily onto your toothbrush with a rod-like shape.

- **Future technology** – In the emerging technology of microfluidics, the goal is to manipulate minute amounts of liquid through channels no more than a few microns across. A major problem is getting large enough flow rates through these narrow channels when a no-slip boundary condition exists at the channel walls. Consequently, many researchers are developing new surface treatments that produce substantial slippage in order to improve the throughput in microfluidic devices.

### 9.4.1 Definition of slip length

While the no-slip boundary condition has been successfully used since the beginning of the 19th century when scientists began analyzing fluid flows, doubt has always existed as to how precise this approximation was, and the possibility of slip has been continuously debated. The standard way of characterizing slip was introduced early on by Navier (Navier 1823): The slip velocity \( v_s \) of the fluid contacting the surface is proportional to the shear rate at the surface or

\[
v_s = b \frac{dv(z)}{dz} \bigg|_{z=0}.
\]

(9.15)

Since the shear rate is also proportional to the shear stress [eq. (9.11)], this approach is equivalent to the slip velocity \( v_s \) being proportional to the shear
stress $\tau_s$ at the surface

$$v_s = \frac{b}{\eta \tau_s}.$$  \hspace{1cm} (9.16)

From eqs. (9.15) and (9.16), we see that the parameter $b$ characterizes the slippage of fluid at a surface; $b$ is called the *slip length* as it has the unit of length. Generally, $b$ is assumed to be constant, making this a linear boundary condition. If $b$ is a function of $v_s$ or $\tau_s$, the boundary condition is non-linear.

Figure 9.7 illustrates the velocity profiles of a fluid flowing next to a surface with various degrees of slip: no slip, partial slip, and perfect slip. From Fig. 9.7 we see that the slip length $b$ corresponds to the distance below the surface where flow velocity extrapolates to zero.

### 9.4.2 Measuring slip at liquid–solid interfaces

#### 9.4.2.1 Pressure drop versus flow rate method

A common way of determining slip lengths is to measure the flow rate driven by a known pressure gradient through a narrow channel or pipe. The measured flow rate is then fitted with the calculated flow rate using the slip length as the adjustable. Considering a channel with length $L$ and width $w$, which are much greater than its height $d$, we integrate eq. (9.12) assuming a slip velocity $v_s$ exists at the channel walls; then the flow rate $q$ for an applied pressure drop $\Delta p$ becomes

$$q = \frac{wd^3 \Delta p}{12\eta L} + wd \cdot v_s \hspace{1cm} (9.17)$$

where the first term on the right is the Poiseuille flow with no-slippage [same as eq. (9.14)] and the second term is the additional flow due to slip. With the

![Fig. 9.7. Profiles of the fluid flow velocity next to a surface as a function of the distance $z$ away from the surface, where the extrapolation length $b$ characterizes the amount of slippage. Three types of slip are illustrated: no slip ($b = 0$), partial slip ($0 < b < \infty$), and perfect slip ($b = \infty$).](image-url)
Navier condition eq. (9.15) relating slip velocity to slip length, the slip length $b$ can be determined by from the flow rate through this channel:

$$b = \frac{2\eta q L}{\Delta pw d^2} - \frac{d}{6}.$$  \hspace{1cm} (9.18)

So, we see that the flow rate with slip $q(b)$ is the following factor larger than the flow rate with no-slip $q_{\text{no-slip}}$:

$$
\frac{q(b)}{q_{\text{no-slip}}} = 1 + \frac{6b}{d}.
$$  \hspace{1cm} (9.19)

Eq. (9.19) indicates that the extra flow rate due to slip only becomes significant when the slip length $b$ becomes comparable to the channel height $d$. As $b$ is typically sub-micron for ordinary liquids like water flowing against smooth surfaces, the channel needs to be fairly narrow—no more than a few microns in height—for slippage to have a noticeable effect on flow. Since such small values for $b$ are often smaller than the precision with which $d$ can be measured, relative experiments are often done instead, where the flow rate is measured for the same channel with different surface treatments while keeping the channel dimension constant. A typical surface treatment is to passivate an otherwise hydrophilic channel surface with a low energy surface film in order to promote slip. From these experiments, lower energy surfaces are found to induce slip lengths of a few tens of nanometers for liquid flowing through capillaries (Lauga et al. 2007; Neto et al. 2005).

9.4.2.2 Drainage versus viscous force

When a curved surface is pushed toward a flat or another curved surface while immersed in a liquid, the liquid drains via Poiseuille flow, generating a viscous force that opposes the change in separation distance $D$. For a sphere of radius $R$ near a flat surface or two crossed cylinders with mean radius $R$, this viscous force $F$ can be expressed as

$$F = -f^* \frac{6\pi \eta R^2 V}{D}$$  \hspace{1cm} (9.20)

where $V = \dot{D}$ is the velocity of the change in separation and $f^*$ is the slip factor that quantifies how the slippage affects the viscous force: $f^* = 1$ if the no-slip condition occurs, and $f^* < 1$ if slip occurs. For the case where the slip length $b$ is the same on the two opposing surfaces, the slip factor $f^*$ can be expressed (Hocking 1973; Vinogradova 1996) as

$$f^* = \frac{D}{3b} \left[ \left(1 + \frac{D}{6b}\right) \ln \left(1 + \frac{6b}{D}\right) - 1 \right].$$  \hspace{1cm} (9.21)

In the limit of small separations where $D \ll b$, then $f^*$ goes to zero as $f^* \approx D \ln (6b/D) / 3b$; therefore, as the separation distance goes to zero, the viscous force in eq. (9.20) varies logarithmically with $D$ due to slippage.
Numerous studies have exploited this drainage technique, using either the surface force apparatus (SFA) or atomic force microscopy (AFM) for measuring slip lengths. SFA is particularly effective at these measurements as not only can $D$ be made as small as a single layer of liquid molecules, but also the absolute value of $D$ can be measured with the precision of a few ångstoms. With $R$ in an SFA typically being a few centimeters, the viscous force is easily measurable, either as a function of separation velocity $V$ or separation distance $D$. Zhu and Granick have conducted the most extensive series of slip lengths measurements with the SFA technique and have found that the no-slip condition generally holds for hydrophilic surfaces and for surfaces with any significant amount of roughness, while slip lengths a few tens of nanometers to a few microns occur for molecularly smooth surfaces with low surface energies (Zhu and Granick 2001; Zhu and Granick 2002; Zhu and Granick 2002; Zhu and Granick 2002; Granick et al. 2003). Zhu and Granick have also observed that a critical shear stress needs to be exceeded for a significant slip length to be observed and the slip length increases with shear rate, indicating that $b$ is not a constant as typically assumed for a linear boundary condition. Channel experiments have also observed slip lengths increasing with shear stress (Choi et al. 2003).

9.4.3 Mechanisms for slip at liquid–solid interfaces

Eqs (9.15) and (9.16) characterize the slippage within the continuum description for fluid flow, but do not tell us how the slip occurs or what determines the slip length $b$. In our discussion, we will divide possible mechanisms for slippage into two categories:

1. **Molecular slip** where the liquid molecules remain in intimate contact with the solid surface
2. **Apparent slip** where the liquid is not in intimate contact with the solid, but instead a thin medium with much lower viscosity separates the liquid from the solid.

9.4.3.1 Molecular slip

When a liquid flows against a solid surface, molecular slip occurs when the hydrodynamic forces are sufficient to drive the molecules adjacent to a solid across its surface. The primary question to be addressed is: What determines the resistance that the liquid molecules experience as they move along the interface between the liquid and solid?

9.4.3.1.1 Molecular slip at low energy surfaces

Intuitively, one expects that slip would become greater as the interaction energy becomes weaker between the liquid and the solid. This concept has led to many experimental and theoretical investigations focusing on how slip depends on surface energies.

Tolstoi developed one of the first theories quantifying the relationship between surface energies and slip (Tolstoi 1952; Blake 1990). In his analysis, Tolstoi
extended concepts for explaining the mobility of molecules in a liquid to explaining their mobility at the liquid–solid interface. In a liquid, an individual molecule's mobility is determined by the probability of a void being created within the liquid large enough for the molecule to diffuse into this hole. The probability of thermally exciting the creation of this molecule is related to the energy required to form the void: If \(a\) is the characteristic molecular dimension and \(\gamma\) the liquid surface tension, then the energy to create a molecular size hole is roughly \(\gamma a^2\). When a liquid wets the solid, the work of adhesion of the liquid with the solid is greater than the liquid’s internal cohesion (Section 5.4.2.1.3), and it takes more energy to create a hole within the liquid monolayer adjacent to the solid; this suppresses the mobility within that layer and results in a slip length less than the size of the molecule. When the liquid has a finite contact angle with the solid (partial wetting and non-wetting), the liquid has a lower affinity for the solid than for itself. So, more holes are thermally excited in the molecular layer adjacent to a non-wetting surface, increasing the diffusivity within this layer and resulting in a finite slip length. One shortcoming of the Tolstoi analysis is that it predicts that the slip length increases exponentially with increasing contact angles. This exponential dependence leads to predicted slip lengths of millimeters, many orders of magnitude higher than observed in experiments. To get around this dilemma, it has been proposed that surface heterogeneity or surface roughness may limit the magnitude of achievable slip length to much less than the Tolstoi prediction.

Molecular dynamic simulations also indicate that slip lengths increase as the liquid–solid interaction energy is reduced, but these simulations generally predict slip lengths much smaller than what is measured experimentally. (Barrat and Bocquet 1999a, b).

In experiments, a number of research groups have measured slip lengths on both wetting and non-wetting surfaces and have systematically found larger slip lengths for the non-wetting surfaces. As pointed out by Lauga et al., however, when results from different groups are compared with each other, slip lengths correlate poorly with contact angles (Lauga et al. 2007). The poor correlation between slip and contact angles among results from research groups (as well as the occasional observation of no-slip on non-wetting surfaces) may result from other parameters that influence slip, such as roughness, that are varying from one experimental setup to the next and masking the underlying relationship between surface energy and slip.

A variation on this concept of voids being thermally excited at liquid–solid surfaces is the notion that a small gap, a few ångströms wide, is thermally excited between the liquid and solid due to short range repulsive forces acting between the low energy surface and the adjacent liquid layer. As discussed in Section 7.4.2.2, hydrophobic forces can repel water molecules away from hydrophobic surfaces [x-ray reflectivity measurements have confirmed the existence of a drop in water density at water–hydrophobic surface interfaces (Jensen et al. 2003)]. This slight separation of a few ångströms greatly reduces the overlap of the
atomic electron clouds on the opposing liquid and solid surfaces, greatly reducing
their ability to generate lateral repulsive forces that resist slip. If gases dissolved
in the water segregate out to a water–hydrophobic surface interface, they can
swell this low density gap to form “nano-bubbles” tens of nanometers high,
further promoting slip at these interfaces; in this case, slip occurs via the shearing
of the gas in the nano-bubbles, and the slippage mechanism is more appropriately
classified as apparent slip rather than molecular slip.

9.4.3.1.2 Slippage of polymers melts In addition to occurring at interfaces
with low surface energies, molecular slip can also happen at interfaces where
the liquid wets the solid, if the hydrodynamic forces are great enough to drag
the liquid molecules across the solid. One example of this was mentioned above,
where a piece of glass slides across a surface with perfect slip since the viscous
forces are so high that the atoms that make up the glass are forced to move in
unison. The flow of polymer melts is another example where high viscous forces
lead to slippage. The slippage of polymer melts has been well studied in the
field of polymer processing, as slippage often dominates the flow properties of
polymers through the pipes used in their processing, and processing additives
that promote slip are often added to plastics to facilitate their flow through
moulds (Achilleos et al. 2002).

Since these polymers are entangled, they strongly oppose being sheared as
this would tend to disentangle them; the entanglement results in viscosities that
are enormously enhanced compared with normal liquids and non-entangled poly-
mers. With such high viscosities, it becomes easier to concentrate the shear at the
interface with the solid rather than spread the shear rate out over the polymer
melt. The basic model for slippage at a polymer–solid interface was proposed by
de Gennes (de Gennes 1979) and predicts that the slip length $b$ is related to the
bulk viscosity $\eta$ of the polymer and to the molecular size $a$ by

$$b = a \frac{\eta}{\eta_0}$$  \hspace{1cm} (9.22)

where $\eta_0$ is the viscosity of a liquid consisting of monomer units of the polymer
(i.e., a liquid with the same interactions, but no entanglements). Since entan-
glement greatly enhances the viscosity of the polymer melt, $\eta$ can be many
orders of magnitude greater than $\eta_0$, resulting in slip lengths of many microns
to millimeters.

9.4.3.2 Apparent slip

When a lower viscosity component segregates out between a liquid and a solid
surface, it shears more easily leading to apparent slip at the liquid–solid interface.
Figure 9.8(a) illustrates the velocity profile for flow in this situation. Notice that
the velocity profile extrapolates through the low viscosity boundary film and
below the solid surface to give an apparent slip length $b$, even though a no-slip
boundary actually exists where the low viscosity fluid meets the solid surface.
Since the shear stress is continuous across the boundary between the high and
Fig. 9.8. Profiles of the liquid flow velocity next to a solid surface. (a) Apparent slip occurs due to a lower viscosity layer forming next to the surface. (b) A higher viscosity boundary layer forms next to the surface shifting the no-slip boundary condition a distance $a$ away from the surface.

Low viscosity fluids, the velocity gradient in the boundary layer must be higher than the rest of the liquid by the ratio of the two viscosities [eq. (9.11)]. This leads to an apparent slip length

$$b = c(\eta_l/\eta_b - 1)$$  \hspace{1cm} (9.23)

where $c$ is the boundary layer thickness, $\eta_l$ is the liquid viscosity, and $\eta_b$ is the boundary layer viscosity.

Below are some of the mechanisms for forming low viscosity layers at liquid–solid interfaces, that lead to apparent slip.

- **Lotus effect of trapped air around surface roughness** – This mechanism occurs for water flowing over hydrophobic surfaces where the slopes of the surface roughness are very high, as illustrated in Fig. 9.9(a). The high water contact angle on the sides of these surface features with a high aspect ratio prevents

Fig. 9.9. For water-hydrophobic interfaces, a low viscosity gaseous layer causes apparent slip. (a) Pockets of air trapped between the peaks of the surface roughness. (b) Nano-bubbles formed by gas dissolved in the water segregating to the surface.
the water from penetrating below the roughness summits, so only air or vapor exists around the roughness features. When water flows over this surface, slippage occurs over the trapped air pockets, which dominates over the much smaller fraction at the roughness summits where the no-slip boundary condition applies. Since the ratio of water viscosity to air viscosity equals 55, then eq. (9.23) indicates that the slip length $b$ should be 54 times the thickness of the trapped air layer. This mechanism is often referred to as the lotus effect, since it is responsible for water droplets quickly running off lotus leaves, which are covered with tall, thin hairs of a hydrophobic, wax-like material, providing the leaves with the right roughness for this effect (Barthlott and Neinhuis 1997). These surfaces are referred to as super-hydrophobic. Several research groups have demonstrated that large slip lengths—tens of microns—can be engineered onto surfaces by nano-fabricating them to have micron high surface features made or covered with low energy films (Martines et al. 2005; Choi and Kim 2006).

• **Nano-bubbles** – For smoother hydrophobic surfaces, the air pockets prevalent on the super-hydrophobic surfaces become thermodynamically unstable and collapse. Even though air bubbles are not supposed to be stable on these surfaces, numerous recent studies have found evidence that bubbles, only a few nanometers in height (*nano-bubbles*) form at these water–hydrophobic surface interfaces. While the mechanism for nano-bubble formation is not well understood, they are suspected to form from gases previously dissolved in water segregating to the hydrophobic surface [Fig. 9.9(b)]. [The evidence for nano-bubbles and possible explanations have been recently reviewed (Lauga et al. 2005; Neto et al. 2005).]

• **Polymer solution and particle suspensions** – If a polymer is dissolved in a solution where the solvent molecules with lower viscosity preferentially adsorb on the surface, a low viscosity layer forms to provide for apparent slip. Similarly, if the particles in a suspension are repelled by a nearby surface, apparent slippage occurs due to the reduced concentration of particles at the suspension–solid interface (Barnes 1995).

Figure 9.8(b) illustrates the opposite effect to apparent slip, where a layer of higher viscosity material segregates to the liquid–solid interface, moving the plane for the no-slip boundary condition into the liquid flow a distance $a$, where $a$ is often on the order of the size of a molecule. For lubricated bearings, this high viscosity layer is often formed by the additives blended into the base lubricant, which segregates out to serve as a boundary lubricant layer. These are often polymers with end-groups that have an affinity for the solid surface, driving them to self-assemble out of solution and anchoring them to the surface with a high effective viscosity. Another example of a boundary layer formed by molecular deposition, occurs when lipid plaques deposit on the walls of arteries; the thicker these lipid deposits are, the more restricted the flow through blood vessels and the higher the risk for heart attacks and strokes.
9.4.4 Example: Shear stress in the presence of slip

For a liquid film between two parallel surfaces with no slippage at the surfaces, eq. (9.2) indicates how much shear stress $\tau_f$ needs to be applied to move one surface at a velocity $v_0$ relative to the other surface. If slippage occurs with a slip length $b$ at both plate surfaces, eq. (9.2) becomes

$$
\tau_f = \frac{F_f}{A_f} = \frac{\eta v_0}{d + 2b} = \eta \frac{v_0}{d} \left(1 + \frac{2b}{d}\right)^{-1} \tag{9.24}
$$

So, slippage reduces by a factor $(1 + 2b/d)^{-1}$ the shear stress needed to generate a shear rate $v_0/d$. Figure 9.10 shows how the shear stress is reduced by various amount of slip for a 100 cP liquid being sheared between two parallel surfaces separated by $d = 1 \mu$m. Only when $b$ becomes comparable to the separation distance $d$ does the shear stress become significantly less than for the no-slip case.

9.4.5 Why does the no-slip boundary condition work so well?

Before leaving this discussion of slip, it is useful to return the question: Why does the no-slip boundary condition works so well for most ordinary liquid–solid interfaces?

One reason why has already been touched on: When the liquid molecules make contact with the walls, the intermolecular forces cause them to stick and resist their movement to the adjacent surface sites. Only when these interactions are
reduced, for example, by lowering the interface energy, are slip lengths increased above a few molecular diameters.

Another reason is that the roughness prevalent on most surfaces suppresses the potential for slip. In essence, the viscous dissipation of flowing over surface roughness brings the liquid to rest, regardless of the strength of the molecular interactions with the surface (Granick et al. 2003). Zhu and Granick have shown an rms roughness of 6 nm is sufficient to suppress slippage at a hydrophobic surface (Zhu and Granick 2002). In general, any surface heterogeneity, of which roughness is just one type, can pin the flowing liquid, suppressing slip. The situation is physically similar to the pinning of liquid droplets on surfaces, due to chemical heterogeneity and roughness of the surface, that leads to contact angle hysteresis (Section 5.4.2.1.5). Increasing the shear rate of the liquid flowing over these surface irregularities eventually exerts enough shear stress to overcome this pinning and induces slip. For most practical surfaces, where the roughness is typically more than the few nanometers needed to suppress slip, the critical shear stress is never reached for ordinary flow conditions, and the no-slip boundary condition holds. Super-hydrophobic surfaces represent a special case where a certain kind of surface roughness is combined with a hydrophobic surface to induce slip.

9.5 Fluid film lubrication

Over the past century, perhaps the biggest achievement of tribology engineering has been the ability to design bearings with lubricating fluid films. For a well designed fluid film bearing, the sliding surfaces almost never come into solid–solid contact as shearing takes place entirely across the fluid film. Such a bearing can operate within its designed operating range practically forever with low friction and minimal wear, so long as the fluid is not lost or degraded. Virtually every machine, where components move relative to each other for long periods of time, now relies on fluid bearings; some examples are automobile engines, electric motors, turbines, pumps, gearboxes, and computer hard drives.

While lubricated bearings have been around for centuries, scientific understanding of what it takes to achieve effective fluid film lubrication only started in the late 1880s, with the work of Beauchamp Tower and Osborne Reynolds.

First came Tower’s experimental finding of hydrodynamic lubrication. Tower was employed by a British railway company to investigate friction in railroad journal bearings. During the course his work, he had drilled a hole in the top of a bearing, which was lubricated from the bottom by an oil bath, with the intention of adding more lubricant from the top (Fig. 9.11). To his amazement, he found that, when the shaft rotated against the bearing, not only was the oil pulled into the bearing, but also rose up the hole with enough pressure to pop out the wooden plug he had used to stop up the hole. His subsequent investigations showed that this high oil pressure was correlated with very low friction coefficients for the bearing, from $\mu = 0.001$ to 0.01.
Fig. 9.11. Schematic of the partial journal bearing used by Beauchamp Tower for simulating the bearing tribology of railway axles. In the later stages of his experiments, he added the pressure gauge shown at the top of the bearing, for determining the oil pressure.

After hearing of Tower’s intriguing result, Osborne Reynolds analyzed how a journal could drag oil into a bearing and how a sufficient pressure builds up to support the loading force. In 1886, he published a detailed analysis of fluid film bearings—including a differential equation now called the Reynolds equation—that still forms the theoretical basis for fluid film lubrication. This equation is used to determine the pressure distribution within a fluid film bearing, from the geometry of the surfaces, the relative sliding velocities, and the fluid’s density. The Reynolds equation can be derived from the general equations of motion for viscous fluids, the Navier–Stokes equations, by assuming that the fluid flows in a laminar manner (not turbulent), that the fluid film thickness is small compared to the lateral dimensions, and that mainly viscous forces act within the fluid. The reader is referred to specialized texts on fluid flow and lubrication theory for the derivation of Reynolds equation and more detailed discussion of fluid film lubrication (Hamrock 1994; Williams 1994; Szeri 1998).

Only in a few simple cases (some of which are discussed below) can the Reynolds equation be solved analytically. These days, most analyses of fluid film lubrication problems are done numerically using computer software specifically written for this task. The coupling over the past few decades of computer aided design (CAD) of bearing geometries with new precision engineering methods for fabricating surface topographies and new specialty lubricants has led to a renaissance of high performance bearing engineering. The greatly improved lubrication
performance of modern bearing designs has led to key advances for numerous technologies such as disk drives, gyroscopes, ultracentrifuges, MRI machines, and dental drills.

9.5.1 Hydrodynamic lubrication

In hydrodynamic lubrication, fluid is dragged into a bearing by the relative motion of the two sliding surfaces. For hydrodynamic lubrication to work, a portion of the bearing needs to converge so that the outlet for the entrained fluid is smaller than the inlet. This convergence causes the fluid to be squeezed as the relative motion of the two surfaces pulls lubricant into the gap; this pressurizes the fluid film, generating a lifting force that pushes the surfaces apart. Figure 9.12 shows three such convergent geometries often encountered in hydrodynamic bearings: (a) an inclined plane or tilted pad bearing, (b) a Rayleigh step bearing, and (c) a non-concentric journal bearing.

It should be kept in mind that the wedge angles in Fig. 9.12 are greatly exaggerated; the actual convergence angles in hydrodynamic bearings are quite small, typically only about a quarter of a degree. Also, the lift force generated does not depend so much on the precise shape of the converging geometry, but more strongly, as we will see, on the ratio of inlet to outlet height $h_i/h_o$.

In order for hydrodynamic lubrication to occur without the sliding surfaces making contact, the fluid film thickness throughout the bearing needs to be much

![Fig. 9.12. Upper: Convergent bearing geometries frequently encountered in hydrodynamic lubrication. The bearings are sufficiently wide in the direction out of the page so that the fluid flow can be treated as being entirely parallel to the plane of the page. Lower: The distributions of pressure $p$ above ambient along the bearings that support the load $L$ per unit width, where the width is the length of the bearing perpendicular to the page. Note that the pressure distributions are not symmetrical about the bearing midpoint.](image-url)
larger than the surface roughness, so that the asperities on the opposing surfaces do not rub against each other. A typical rule of thumb is that the minimum separation distance between the surface means $[h_o$ or $h_{\text{min}}$ Fig. 9.12] should be at least a factor of five times the standard deviation of roughness $\sigma$ to ensure that the two surfaces do not contact. The film thickness in oil bearings typically ranges from 1 $\mu$m to 1 mm.

9.5.1.1 Inclined plane bearing
First let’s discuss the pressure distribution that develops for one of the simplest and most common types of fluid bearing geometries: the inclined plane geometry shown in Fig. 9.12(a). To simplify the problem, we consider an incompressible fluid (most liquids can be treated as incompressible) and assume that viscosity is constant throughout the bearing. For this situation, the Reynolds equation for determining the variation in pressure along the bearing can be expressed as

$$\frac{dp}{dx} = -6\eta U (h - h^*)/h^3$$

where $h^*$ is the separation distance at the $x$ location where the pressure is maximum ($dp/dx = 0$). If we integrate eq. (9.25) with the boundary conditions $p = 0$ at $x = 0$ and $x = B$, we obtain the pressure distribution shown in the lower part of Fig. 9.12(a). Integrating again, we obtain an expression for how much load $L$ the bearing supports per unit width (i.e., the lift generated by the bearing):

$$L = 6\eta K U \left( \frac{B}{h_o} \right)^2$$

where

$$K = \frac{\ln(1 + n)}{n^2} - \frac{2}{n(2 + n)}$$

and $n = h_i/h_o - 1$. Therefore, the value of $K$ depends on the ratio of the inlet and outlet heights, but for this convergent geometry turns out to be fairly insensitive to this ratio. An inclined plane bearing generates maximum lift at $h_i/h_o = 2.2$ where $K = 0.027$.

9.5.1.2 Rayleigh step bearing
In 1918 Lord Rayleigh was able to show, using variational calculus, that the bearing shape that provides maximum lift consists of two parallel sections, such as that shown in Fig. 9.12(b). (An optimal Rayleigh step bearing generates a factor of 1.29 more lift than an optimal inclined plane bearing.)

The parallel sections of a Rayleigh step bearing make the analysis more straightforward than for other bearing types. First, the Reynolds equation predicts a constant pressure gradient in the parallel sections of the step bearing:

$$\left\{ \frac{dp}{dx} \right\}_i = \frac{p_s}{B_i} \quad \text{and} \quad \left\{ \frac{dp}{dx} \right\}_o = -\frac{p_s}{B_o}$$

(9.28)
where \( p_s \) is the peak pressure, which occurs at the step, and \( B_o \) and \( B_i \) are the lengths of the two parallel sections. This leads to the pressure profile shown in the bottom of Fig. 9.12(b).

As before, we assume that the fluid is incompressible (i.e., a liquid) and that the viscosity is constant throughout the bearing. Since the volume flow rate is the same in both sections, from eq. (9.14) we can write

\[
q = -\frac{h_i^3}{12\eta} \left\{ \frac{dp}{dx} \right\}_i + \frac{Uh_i}{2} = -\frac{h_o^3}{12\eta} \left\{ \frac{dp}{dx} \right\}_o + \frac{Uh_o}{2}.
\]

Combining this equation with eq. (9.28) and rearranging to solve for \( p_s \) obtains

\[
p_s = \frac{P(H - 1)}{H^3 + P} \frac{6U\eta B_o}{h_o^2} \left( H^3 + P \right)
\]

where \( H = h_i/h_o \) and \( P = B_i/B_o \).

The load \( L \) per unit width that can be carried by a Rayleigh step bearing is given by

\[
L = \int_0^\frac{B}{2} p\ dx = \frac{U\eta B^2}{h_o^2 (1 + P)(H^3 + P)} 3P(H - 1) \]

For a given viscosity \( \eta \) and sliding speed \( U \), maximum lift occurs when \( H = 1.87 \) and \( P = 2.59 \) (Williams 1994 p. 248).

### 9.5.1.3 Journal bearings

Figure 9.12(c) illustrates a journal bearing where a shaft (journal) rotates within a sleeve (bearing or bushing), dragging lubricant into the clearance gap. Since the loading force shifts the shaft axis off center from the bearing sleeve, a convergent wedge of lubricant forms, which generates the hydrodynamic pressure that counterbalances the load during steady state hydrodynamic lubrication.

At first glance, a journal bearing looks like an inclined plane bearing that has been wrapped round on itself. The solution for the Reynolds equation, however, is complicated by the presence of the divergent portion of the lubricant film in the journal bearing. If the Reynolds equation is simply integrated around the bearing, one obtains the plot labeled *full Sommerfeld solution* in Fig. 9.13 for the pressure as function of angle. The problem with this solution is that the areas underneath the positive and negative portions of the pressure curve cancel, so this solution provides no net load support. That journal bearings do provide load supports suggests that liquid lubricants can only sustain a limited
amount of negative pressure. When a liquid is subjected to a negative pressure, it is under a tensile stress which causes the liquid to rupture when a threshold is exceeded. This rupture process is known as cavitation as it creates voids in the liquid. The physics of cavitation is still poorly understood (Furukawa and Tanaka 2006; Grando et al. 2006). Cavitation can be promoted by dissolved gases in the lubricant, which segregate out to form the voids, or the cavitation process can pull in air from outside the bearing, displacing lubricant within the bearing. Often, cavitation is initiated at the walls of the bearing, where the roughness lowers the activation energy, and this process can be further promoted by the presence of nano-bubbles on these surfaces.

In the Reynolds equation analysis, the phenomenon of cavitation is typically handled by requiring that the pressure in the liquid be greater than or equal to zero, which provides solutions close to what is measured in experiments. In Reynolds’ original 1886 paper, he used the boundary condition that the pressure and its first derivative with respect to position around the bearing go to zero simultaneously as the liquid exits the convergent portion of the journal bearing; this Reynolds solution is plotted in Fig. 9.13.

The expression for the load bearing capacity of a journal bearing, even though the analysis is more complicated, ends up being similar in form to that for an inclined plane bearing [eq. (9.26)]; the load $L$ per unit width that a journal bearing supports is given by:

$$L = \eta SU \left( \frac{R}{h} \right)^2$$  \hspace{1cm} (9.32)
where $h$ is the mean thickness of the lubricant film, $R$ the shaft radius, and $U = \omega R$ the peripheral speed of the shaft. Here $h$ is also the maximum distance that the shaft axis can be displaced from the bearing center before contacting the bearing surface, so it is also the maximum eccentricity. For typical journal bearings, $R/h$ is usually in the range of 1000 to 10 000.

In eq. (9.32), $S$ is a dimensionless number, called the Sommerfeld number, determined by the eccentricity of the shaft in its journal housing and by the ratio of the bearing width to its diameter, and so it expresses how this ratio and the ratio of the maximum to the minimum lubricant film thickness ($h_{\text{max}}/h_{\text{min}}$) impacts the load bearing capacity. $S$ can be obtained by numerically solving the Reynolds equation for any eccentricity and width to diameter ratio, and is tabulated in many standard texts on bearing lubrication. An obvious case is when the shaft is concentric with journal housing: $S = 0$, as no convergent wedge forms to generate a load bearing capacity. For a more typical case when $h_{\text{max}}/h_{\text{min}} = 4$, $S$ is about 2 when bearing width is the same as the diameter, and $S \approx 7$ for an infinitely long bearing. As the load force $L$ increases, the bearing becomes more eccentric, increasing $h_{\text{max}}/h_{\text{min}}$ and the pressures in the convergent film. The maximum load bearing capacity is determined by the smallest $h_{\text{min}}$ that can be achieved before the shaft roughness starts to contact the roughness of the journal housing.

A good estimate for the friction or tangential force $F$ per unit width acting on a typical bearing can be estimated from the viscous drag acting on an infinitely wide bearing:

$$F = 2\pi \eta U R/h.$$  \hspace{1cm} (9.33)

Therefore the friction coefficient $\mu$ is estimated by

$$\mu = F/L = \left( \frac{2\pi}{S} \right) \left( \frac{h}{R} \right).$$  \hspace{1cm} (9.34)

From eq. (9.34), we can see that very low values of $\mu$ can be obtained with hydrodynamic lubrication; for example, if $S = 7$ and $h/R = 1/1000$, then $\mu = 0.0009$.

9.5.2 Gas bearings

While liquid lubricants are used within most fluid film bearings, for some bearings gas is a more desirable lubrication fluid. Air is the most common gas used due to its inherent advantages of ease of supply, non-degradability, and little environmental risk when exhausted. This is particularly true for most hydrostatic gas bearings (or more properly aerostatic bearings) where they are continually exhausted into the environment. Aerostatic bearings are frequently used on assembly lines to guide parts through with low friction without contaminating the surfaces of the moving parts. An air hockey game is another example where an aerostatic bearing provides the puck with ultra-low friction motion.
Since gas viscosities are typically 0.1% to 0.01% that of oil lubricants, sliding speeds for an aerodynamic bearing need to be many times that of a comparable liquid hydrodynamic bearing to generate a similar load bearing capacity. For high speed bearings, however, gas lubrication provides some distinct advantages:

1. Gases are not susceptible to the degradation that plagues oil lubricants at high shear rates.
2. The low gas viscosity provides for low friction.
3. Gases can lubricate over a wide temperature range without the risk of freezing at low temperatures or boiling at high temperatures.

Therefore, gas journal bearings are commonly used in high rotation speed devices, such as turbomachinery, machine tool spindles, gyroscopes, and dental drills.

Since the lower viscosity provides less lift, gas bearings typically have a smaller separation distances than liquid bearings to compensate. The smaller gap sizes mean that cleanliness is a more critical issue for gas bearings, since hard contaminant particles with dimensions smaller than the clearance can cause serious damage if ingested into the bearing. The smaller gap size also means that the manufacturing tolerances and surface roughness of gas bearings have to be much more tightly controlled than for oil bearings. An extreme example of tight tolerances for an aerodynamic bearing occurs in a disk drive, where the recording head slider is designed to fly with a gap less than ten nanometers at its trailing edge. To ensure that a few nanometers of clearance is achieved between the slider and disk, these surfaces are now manufactured with rms roughnesses less than a nanometer.

For gas bearings, the compressibility needs to be taken into account. One way of doing this is to define a dimensionless compressibility or bearing number when analyzing the bearing. For the inclined plane geometry in Fig. 9.14, the bearing number is

$$\Lambda = \frac{6U \eta_a B}{p_a h_o^2}$$  \hspace{1cm} (9.35)

where $p_a$ is the ambient pressure and $\eta_a$ is the gas viscosity at ambient pressure. For low values of the ratio $U/h_o$, the bearing number is small, the variation in pressure along the bearing is small relative to the ambient pressure, and the gas can be treated as incompressible. At high values of $U/h_o$ corresponding to large bearing numbers, changes in pressure along the bearing become comparable to the ambient pressure, and compressibility effects need to be taken into account when analyzing the Reynolds equation for the bearing. Figure 9.14 illustrates how compressibility affects the resulting pressure distribution for an inclined plane bearing. From Fig. 9.14, we see that, with a compressible gas, the center of pressure shifts toward the bearing’s trailing edge relative to that with an incompressible liquid.
234 LUBRICATION

9.5.2.1 Slip flow in gas bearings
Just as for liquids, molecular slippage also occurs for gases flowing over a solid surface, due to the effect of the slip velocity of the gas molecules impinging on the surfaces. When the thickness $h$ of the gas film is much greater than the mean free path $\lambda$ of the gas molecules, this slip has a negligible impact on the gas flow, and the no-slip boundary condition applies. When $h$ is comparable to $\lambda$, gas flow through the bearing increases significantly above what would be expected with the no-slip boundary condition. When flows are analyzed using the Reynolds equation in continuum theory, the effect of slip is often treated as an effective rarefaction of the gas, that decreases the effective viscosity of the gas as the film becomes thinner, rather than as a slip length.
The earliest approach, for estimating the effective viscosity of a gas in a narrow gap was that of Burgdorfer, who applied the kinetic theory of gases to gas film lubrication:

$$\eta = \frac{\eta_0}{1 + (6a\lambda/h)}$$

(9.36)

where $\eta_0$ is the unrarefied gas viscosity and $a$ is the surface correction coefficient for $\lambda$ (Burgdorfer 1959). This approach is often referred to as the first-order slip model and has been extended over the years to include higher order correction terms in a second-order slip model (Hsia and Domoto 1983) and a 1.5-order slip model (Mitsuya 1993). A more recent approached, introduced by Fukui and Kaneko, is to use a generalized lubrication equation based on the Boltzmann equation, which describes the behavior of the gas molecules statistically (Fukui and Kaneko 1988). In this method, the corrections to the Poiseuille flow rates are calculated in advance and a lookup table generated for future numerical simulations (Fukui and Kaneko 1990).

### 9.5.3 Elastohydrodynamic lubrication

Previously we noted that, to achieve effective hydrodynamic lubrication, a fluid film bearing should generate enough lift to ensure that the asperities on the opposing sliding surfaces do not contact at the bearing location with minimum lubricant film thickness. For many years, one of the puzzles of hydrodynamic lubrication was that many of the most commonly used lubricants provided effective lubrication even in situations where the standard hydrodynamic lubrication theory predicted insufficient lift to prevent asperity–asperity contact. Some common examples of where standard hydrodynamic theory predicts that the local pressures during sliding should be sufficient to squeeze the lubricant out of the bearing thereby inducing contact include:

- Gear teeth rubbing against each other
- Cam rubbing against a follower
- Balls in a ball bearing rolling against their race.

Beginning in the 1940s, the theory of elastohydrodynamic lubrication (EHL) was developed to explain why effective fluid film lubrication still occurs in these situations. This incorporates two additional effects into the standard hydrodynamic theory, that manifest themselves at higher contact pressures and help maintain a continuous liquid hydrodynamic film between lubricated sliding surfaces:

1. The increase in a lubricant’s viscosity with increasing pressure
2. The elastic deformation of surfaces due to the high local pressures within a lubricant film.

#### 9.5.3.1 Pressure dependence of viscosity

In addition to their viscosity decreasing with temperature [as described by eq. (9.9)], lubricants also have the characteristic that their viscosity increases
steeply with pressure. This is particularly true for oil based lubricants and helps account for their effectiveness in many situations. The viscosity of an oil can usually be modeled as having exponential dependence on pressure, described by the Barus equation:

$$\eta = \eta_0 \exp (\alpha p)$$  \hspace{1cm} (9.37)

where $\eta_0$ is the viscosity at one atmosphere, $p$ is the hydrostatic pressure, and $\alpha$ is a constant for the particular lubricant. For example, for a typical mineral oil where $\alpha \sim 2 \times 10^{-8} \text{Pa}^{-1}$, a peak pressure of 230 MPa results in the peak local viscosity $\eta$ being 100 times higher than $\eta_0$. So, while the high pressure is trying to push the lubricant out from between the sliding surfaces, the exponential rise in viscosity works to resist the flow of lubricant out of the bearing, countering the effect of the high pressure and thus maintaining effective hydrodynamic lubrication. From eqs. (9.32), (9.31), and (9.26), the lift generated by a bearing scales with the lubricant viscosity, so the higher viscosity due to pressure works to increase the separation distance between the sliding surfaces. It can be shown (Williams 1994 pp. 326–9) that the additional lift from this pressure-enhanced viscosity increases the minimum separation by as much as 2.3 times.

9.5.3.2 Pressure-induced elastic deformation

Usually, a bigger factor in keeping sliding surfaces out of solid–solid contact is the elastic deformation away from contact due to the high pressures within lubricant films. Even for light loading conditions, small elastic deformations can have a major impact, maintaining a continuous film of lubricant within a contact zone.

Analyzing these elastohydrodynamic lubrication situations is necessarily complex as the Reynolds equation must be solved while taking into account the elastic distortion of surfaces due to the hydrodynamically generated pressure distributions and also how the viscosity varies with pressure.

Figure 9.15 illustrates the essential features of a line contact distorted by elastohydrodynamic lubrication. For the most part, the elastic deformation and the pressure distribution follow closely to that of an unlubricated Hertzian contact. At the fluid inlet, the surfaces becomes flatter so that the sliding surfaces are nearly parallel with a fluid film thickness $D$. Near the outlet, the elastic deformation deviates substantially from a Hertzian contact as a constriction or ‘nip’ develops, reducing the film thickness by about 25%. This constriction causes a sharp spike in pressure before it returns back to ambient pressure as the fluid exits.

It may initially seem odd that the pressure can vary as much as shown in Fig. 9.15 where the surfaces are nearly parallel. What is happening: When $\exp(\alpha p)$ is much larger than 1, the local viscosity becomes very pressure sensitive, which leads to minute changes in film thickness resulting in large variations in pressure. A consequence of this is that the mean separation between the surfaces has a fairly weak dependence on load, since the pressure of the liquid in the gap rises rapidly to counter an increasing load as the separation is slightly reduced.
For analyzing an elastohydrodynamic lubrication sliding geometry, the typical approach is to solve the Reynolds equation numerically and then fit power laws to the numerical solutions to determine the dependence on various parameters. One such example for a line contact (Fig. 9.15) is the power law expression for the minimum film thickness \( h_{\text{min}} \) (Dowson and Higginson 1959; Dowson and Higginson 1966):

\[
\frac{h_{\text{min}}}{R} = 2.65 \left( 2\alpha E_c \right)^{0.54} \left( \frac{\bar{U} \eta_o}{2E_c R} \right)^{0.7} \left( \frac{L}{2E_c R} \right)^{-0.13}
\]  

(9.38)

where \( L \) is the load per unit length, \( R \) is the reduced radius, \( \eta_o \) and \( \alpha \) are defined by eq. (9.37), \( E_c \) is the composite elastic modulus, and \( \bar{U} \) is the entraining velocity, which is the mean velocity of the two surfaces. The final term \( \left( L/2E_c R \right)^{-0.13} \) is a non-dimensional parameter that describes how the minimum film thickness depends on the loading force. From this term we can see that the minimum film thickness has a very weak dependence on the load force \( L^{-0.13} \); this is a critical factor in successful elastohydrodynamic lubrication, as the film thickness can only be reduced to the point of solid–solid contact by applying an exceptionally high load.

\[ \text{Fig. 9.15. (a) How elastohydrodynamic lubrication causes a cylindrical roller bearing to deform as it slides over a flat surface (z scale exaggerated for clarity). The lubricant film over most of the Hertzian contact zone has nearly constant thickness } D, \text{ except near the outlet where a constriction reduces the separation to } h_{\text{min}}. \text{ This constriction leads to a spike in the pressure distribution (b). The dashed semi-ellipse indicates what the pressure would be for a Hertzian contact.} \]
9.5.3.3 Example: Minimum film thickness between sliding gear teeth

To get a better feel for what the film thickness is in a typical elastohydrodynamic situation, let’s use eq. (9.38) to compute the minimum lubricant thickness between two sliding gear teeth, as illustrated in Fig. 9.16. For this example, we assume:

- The lubricant is a mineral oil with \( \eta = 70 \text{ mPa} \cdot \text{s} \) and \( \alpha = 2 \times 10^{-8} \text{ Pa}^{-1} \).
- The gears are made of steel with \( E_c = 115 \text{ GPa} \).
- The loading force per unit tooth width is \( L = 10^6 \text{ N/m} \).
- The relative sliding velocity at the line of action is \( U = 10 \text{ m/s} \), so \( \bar{U} = U/2 = 5 \text{ m/s} \).
- The gear tooth radius where contact occurs is \( R_t = 4 \text{ cm} \), so the reduced radius is \( R = R_t^2/(R_t + R_t) = R_t/2 = 2 \text{ cm} \).

\[
h_{\text{min}} = 2.65 \left\{ \frac{2\alpha E_c}{2E_c R} \right\}^{0.7} \left\{ \frac{L}{2E_c R} \right\}^{-0.13} R
\]

\[
= 2.65 \left\{ 2 \left(2 \times 10^{-8} \text{ Pa}^{-1}\right) \left(115 \text{ GPa}\right) \right\}^{0.54} \left\{ \frac{(5 \text{ m/s}) (70 \text{ mPa} \cdot \text{s})}{4 \left(115 \text{ GPa}\right) (0.02 \text{ m})} \right\}^{0.7} \left\{ \frac{10^6 \text{ N/m}}{2(115 \text{ GPa})(0.02 \text{ m})} \right\}^{-0.13} 0.02 \text{ m} = 0.77 \mu\text{m} \quad (9.39)
\]

Fig. 9.16. (a) One gear driving another. (b) Magnified view of the meshing gearing teeth, which are typically coated with or immersed in an oil based lubricant. Elastohydrodynamic lubrication occurs where the teeth slide against each other at the line of action.
So, a film of lubricant is maintained between sliding surfaces, if the composite roughness of the gear teeth is less than three quarters of a micron, enabling the gears to rotate with low friction and wear.

This elastohydrodynamic analysis should be contrasted with the minimum film predicted for rigid surfaces lubricated with an isoviscous ($\alpha = 0$) lubricant under the same sliding conditions (Williams 1994 eq. 7.79):

$$h_{\text{min}} = 4.9 \frac{U}{2L} \eta R$$

$$= 4.9 \frac{(10 \text{ m/s}) (70 \text{ mPa} \cdot \text{s}) 0.02 \text{ m}}{2 \times 10^6 \text{ N/m}}$$

$$= 0.034 \mu\text{m}$$ (9.40)

Therefore, for this situation, elastohydrodynamic effects lead to a lubricant film’s needing to be a factor of 36 thicker than that predicted by a standard hydrodynamic theory.

9.5.3.4 Experimental measurements of elastohydrodynamic lubrication

Developing experimental techniques for characterizing elastohydrodynamic (EHD) lubricant has long been challenging since these EHD films are, by nature, localized to small volumes that are difficult to access. As a consequence, the initial development of elastohydrodynamic lubrication theory was done without the aid of experimental input. Fortunately, over the past few decades advances in scientific techniques for characterizing nanoscale phenomena have led to a number of valuable approaches being developed specifically for characterizing thin EHD films. With these new techniques, it is now possible to determine, almost down to molecular dimensions, the physical and chemical properties of EHD films, as well as the complex rheology of films at high shear rates and contact pressures (Spikes 1999).

The thickness of the lubricant film under EHD conditions has long been considered to be the most basic parameter to be characterized by experiment. This thickness can now be measured using a variety of techniques, such as optical interferometry, capacitance, x-ray transmittance, and laser fluorescence. One of the principal uses of these thickness measurement techniques has been to confirm how accurately elastohydrodynamic lubrication theory predicts the lubricant film thickness profile for a wide variety of EHD conditions.

Figure 9.17 shows an example from Glovnea et al. (2003), where optical interferometry is used to measure the thickness of a hexadecane lubricant film as a function of entraining speed, for a steel ball rolling on a glass disk. The remarkable aspect of the data in Fig. 9.17 is that the film thickness continues to follow the power law dependence expected from EHD theory for a spherical rolling contact down to a thickness of less than 1 nm, which is comparable to the size of the hexadecane molecule. For this sphere-on-flat geometry, theory (Hamrock
Fig. 9.17. Average lubricant film thickness for a steel ball rolling on a glass disk lubricated with hexadecane with an applied load of 20 N. Film thickness is measured at the center of the contact zone using optical interferometry. The bars correspond to the standard deviation. For this experiment, the ball had an rms roughness of 9.5 nm and the disk 8.2 nm. From Glovnea et al. (2003) with permission of Springer Science and Business Media.

and Dowson 1981) predicts that the EHD film thickness $D$ follows

$$D = k(\bar{U}\eta)^x$$

(9.41)

where $x$ is a constant with a value between 0.6 and 0.75; the slope 0.72 of the log–log plot in Fig. 9.17 falls as expected in this range. This is even more remarkable when you realize that the composite surface roughness is 12.5 nm. So, even though the contact pressures are high enough to flatten any asperity roughness (in this situation, the glass surface elastically conforms to the steel surface), a molecularly thin film of lubricant is maintained between the rolling surfaces with the predicted thickness, instead of being squeezed out. For thicknesses less than 1 nm, the film thickness in Fig. 9.17 deviates slightly from the expected linearity, which is attributed by Glovnea et al. to a thin layer—1 to 2 Å thick—of hexadecane becoming attached to one of the surfaces.

One of the principal reasons why the data in Fig. 9.17 agrees well with theory is that the lubricant is entrained into the EHD interface by the rolling action of the ball on the disk. Since the surfaces roll past each other without sliding, there is no shearing of the lubricant film in the EHD zone. Reducing the degree of rolling and increasing the relative sliding speed increases the shear rate
applied to the lubricant film, but, as mentioned before, the shear stress cannot keep increasing forever with shear rate: eventually something has to give, either through slippage at the lubricant–solid interfaces or through the lubricant’s limiting shear strength. In these situations, elastohydrodynamic theory overestimates the shear stress expected for purely Newtonian behavior with no-slip boundary conditions; indeed, EHD experiments with sliding interfaces commonly observe a weaker than expected shear stress. Currently, a major area of elastohydrodynamic research is to determine the appropriate rheological models for incorporating non-Newtonian and slippage effects into EHD theory (Bair and Winer 1992; Ehret et al. 1998).

If one of the EHD surfaces is transparent, optical spectroscopies, such as infrared and Raman, can be used to study the molecular character of the lubricant film under EHD conditions. These spectroscopies can be useful in establishing the film’s molecular composition, the degree of molecular alignment, and whether molecular degradation is occurring.

Mapping of thermal emission from EHD films is another recently developed technique that is showing great promise. With this technique, the temperature rise due to frictional heating can be mapped and compared with theoretical predictions. Knowing the temperature distribution within the EHD zone can be extremely valuable, as excessive temperatures can lead to lubricant degradation and damage of the sliding surfaces. Another application of thermal emission mapping is to deduce the shear stress that causes the increase in temperature; mapping temperature increases can then be used to generate a map of the shear stress over the EHD film (Spikes et al. 2004).

9.6 Important physical and chemical properties of lubricants

As discussed at the beginning of this chapter, viscosity is the most important physical property for determining the lubricating properties, as borne out by the numerous expressions in the chapter containing viscosity $\eta$ as a parameter. Here, we briefly discuss a few other physical and chemical properties of lubricants that are often important for determining their success or failure.

9.6.1 Surface tension

Generally, it is desirable for lubricants to have low surface tensions for two reasons:

1. As discussed in Chapter 5, low surface tension corresponds to low cohesion energy. Generally, the weaker the forces are between the molecules, the easier it is for molecules to slide over each other, and the lower is the lubricant shear strength of the lubricant. So, low surface tension tends to correlate with low shear strengths and low friction in the boundary lubrication regime.

2. In order for a liquid to provide lubrication, it must first cover the solid sliding surfaces. This is not a problem if the sliding surfaces are immersed in
the liquid, such as for a sealed bearing filled with lubricant. In many situations however, either the bearing is only partially immersed or a thin film of lubricant is applied topically to the surfaces to provide the necessary lubricant. In both cases, the liquid lubricant needs to wet the surfaces to ensure that sliding does not occur on an uncovered portion of the surfaces. As most polymer lubricants have surface tensions in the range of 20 to 35 mN/m, wetting is typically not a problem with metal or oxide surfaces, which have surface energies above this range. Lubrication for plastics can be more problematic, however, as their surface energies overlap the high end of the lubricant surface tension range. So, either lubricant with very low surface tension should be used or the plastic parts should be immersed in the lubricant. Even with immersion, slippage as discussed in Section 9.4.3 can occur if the lubricant surface tension is higher than the solid surface energies; this slippage leads to smaller than expected separations between the sliding surfaces.

9.6.2 Thermal properties

One of the principal drawbacks of polymer lubricants is the limited range of temperatures over which they can provide hydrodynamic lubrication. The exponential dependence of viscosity on temperature means that the viscosity varies rapidly with temperature. At temperatures that are too low, the viscosity becomes so high that the lubricant effectively solidifies; at high temperatures, the viscosity can be too low to provide sufficient lift. At temperatures of a few hundred degrees Celsius, many lubricant molecules begin to break down chemically (for most polymer lubricants, this typically occurs well below their boiling temperature). For example, mineral oils are among the least thermally stable with a thermal stability of only 135 °C, while perfluoropolyethers (thermal stability = 370 °C) and polyphenyl ethers (thermal stability = 430 °C) are among the most thermally stable. Consequently, care must be given when choosing a lubricant to ensure that it has the appropriate viscosity characteristics and is thermally stable over the expected operating temperature range.

Even if the operating temperatures are within acceptable limits, thermal degradation of the lubricant can still be a problem if high flash temperatures occur from frictional heating at contacting asperities. This can be alleviated somewhat if lubricants have high thermal conductivity to help dissipate heat away from these flash points.

For hydrocarbon based lubricants, the main way in which degradation takes place is through oxidation due to the presence of air or oxygen. This oxidation is detrimental as the degradation products form smears or coalesce into solid-like deposits on the sliding surfaces, hindering lubrication, or they can form organic acids that corrode the surfaces or promote further lubricant decomposition. Consequently, most commercial hydrocarbon based lubricants are blended with anti-oxidation agents that react preferentially with oxygen dissolved in the lubricant.
9.7 References


At the beginning of the last chapter, we discussed the different lubrication regimes for a bearing immersed in a lubricating oil as they corresponded to different portions of the Stribeck curve (Fig 9.1). In the hydrodynamic lubrication regime, which occurs at high sliding speeds and light loads, the sliding surfaces are separated by a thick oil film, and friction depends the oil’s hydrodynamic properties, particularly viscosity, and not on the nature of the solid surfaces. Increasing the load or decreasing the relative speed thins the lubricating film within the bearing and, eventually, contact occurs between the sliding surfaces. With this contact, friction and wear not only increase dramatically, but also start to depend strongly on the specific nature of the solid surfaces and the molecular films adsorbed on them.

In this chapter, we look more closely at how the properties of the lubricating films change as the space between the sliding surfaces becomes very small—down to molecular dimensions. As we shall see, several interesting phenomena dominate the lubricating properties of this nanoscale regime:

- **Confined liquids** – Enhanced viscosity and solidification of nanoscale liquid films
- **Boundary lubrication** – An adsorbed monolayer of lubricant that resists penetration of contacting asperities
- **Capillary and disjoining pressures** – Mechanisms responsible for meniscus formation and meniscus forces from lubricant films

### 10.1 Confined liquids

If we consider the example of a smooth ball of radius $R$ being pushed with a loading force $L$ against a liquid droplet on a flat surface (Fig. 10.1), the liquid between the two surfaces undergoes viscous flow as it is squeezed out (Granick 1991; Granick 1999). When the thickness of this liquid film is macroscopic, the viscous flow is well described by Reynolds theory of hydrodynamic lubrication, and the rate at which the liquid film thickness decreases is expressed by eq. (9.20). When the ball–flat separation distance $D$, however, becomes less than about 10 molecular diameters, interesting things begin to happen due to the confinement of the liquid molecules in this narrow gap.

One of these things which was discussed in Section 7.4.1, is that the between two solid surfaces liquid molecules tend to order into a layered structure.
Fig. 10.1. A ball being dropped onto a liquid droplet. As the ball falls, it reduces the thickness of the liquid between the surfaces over time before reaching an equilibrium thickness at long times. The finite thickness of liquid that remains in the gap between the ball and flat originates from the tendency of the liquid molecules to order into layers within the gap: it takes a certain amount of loading force to overcome the solvation forces of the ordered structures and squeeze out the remaining layers. This tendency to form layers manifests itself as oscillations in the local density of the liquid, with a period of molecular dimensions that decay away from the solid surfaces. Reproduced with permission from Granick (1991). Copyright 1991, American Association for the Advancement of Science (AAAS).

This tendency to form layers manifests itself as oscillations in the local density of the liquid, as illustrated on the right side of Fig. 10.1; the period of this oscillation corresponds to a molecular dimension, and the oscillation amplitude decays away from the solid surfaces. This ordered structure gives rise to solvation or structural forces that oppose the squeezing out of individual layers and remarkably enable the liquid film to support the weight of the ball. Of course, the larger the loading force is, the smaller the number of molecular layers at the final equilibrium. Quite a substantial pressure, however, is needed to squeeze out the last molecular layer, ranging, for example, from 300 MPa for short alkanes to 800 MPa for long alkanes (Sivebaek et al. 2004).

Another interesting effect when liquids are confined in small gaps is how their viscoelastic properties change. One way of determining the viscoelastic properties of the liquid confined in the gap is to oscillate one surface parallel to the other so as to shear the liquid film. A convenient way of doing this is to use the surface force apparatus (SFA), which has the advantage that the solvation forces elastically deform the mica surfaces to form two parallel surfaces, which sandwich a thin film of liquid of uniform thickness $D$ between them, as illustrated in...
Fig. 10.2. (a) Cross cylinder geometry of an SFA experiment. Bottom row shows enlarged cross-sectional view of the encircled contact zone for (b) undistorted mica surfaces when $D$ is large or (c) elastically distorted mica surfaces when $D$ is small as a few layers of molecules. The repulsive solvation forces from the molecular layers sandwiched between the mica surfaces elastically deform the mica sheets, and the glue attaching them to the cylindrical lenses, to form two parallel surfaces. In the figure, the scale normal to the surface is greatly exaggerated: the parallel region typically extends up to 100 microns across, while the gap between the surfaces is only a few nanometers.

Fig. 10.2(c). This parallel geometry simplifies the expression of the force needed to shear the liquid film to [eq. (9.2)]

$$F = \frac{\eta Av}{D}. \quad (10.1)$$

Another advantage of using an SFA is that the mica surfaces are atomically smooth, allowing the film thickness $D$ be made arbitrarily small if enough pressure can be applied to squeeze out the last molecular layer. With the SFA technique, the area $A$ over which the surfaces are parallel (typically tens of microns in diameter) is measured by optical microscopy; the separation $D$ is measured by optical interference; and the lateral sliding velocity $v$ is measured using a displacement sensor for lateral deflections. Then, by applying a known lateral force $F$ to one of the mica surfaces, the effective viscosity of the liquid in the gap is determined from

$$\eta = \frac{DF}{Av} \frac{F}{A(d\gamma/dt)}. \quad (10.2)$$
where $d\gamma/dt$ is the shear rate.

In a technique pioneered by Prof. Granick’s group (Peachey et al.
1991), a known sinusoidal force is applied in the lateral direction to one mica surface while the amplitude and phase of the resulting induced velocity of that surface are measured. With this sinusoidal excitation, eq. (10.2) determines the real and imaginary parts of $\eta$, which correspond to the viscous loss and elastic modulus of the liquid film at that frequency and corresponding shear rate $d\gamma/dt$.

The general result of these viscoelastic experiments is that the more a liquid is confined, the more sluggish it becomes. This slowing down of the liquid’s molecular motion with decreasing film thickness can be divided into three regimes:

1. **Bulk liquid** – For thick liquid films, the viscosity and other dynamic properties remain close to that of the bulk liquid.
2. **Enhanced viscosity** – When the liquid film thickness is reduced below four to ten molecular diameters (depending on the liquid), the liquid begins to transition to a more solid-like phase as manifested by a rapidly increasing viscosity and elastic modulus.
3. **Solid-like** – When the film thickness becomes less than two to four molecular diameters (again depending on the liquid), a solid-like phase forms, which deforms elastically when sheared until its yield stress is exceeded.

This transition with decreasing film thickness, that is, bulk liquid $\rightarrow$ enhanced viscosity $\rightarrow$ solid-like, has been observed for many kinds of liquids, such as linear and branched alkanes, polymers, glass formers, liquid crystals, and aqueous solutions, and so it seems to be a generic feature of confinement rather than related to the molecular structure or chemical composition.

Figure 10.3 shows how the viscosity of a simple alkane liquid, dodecane, increases rapidly with decreasing film thickness. For dodecane films less than 50 Å thick, oscillatory forces are observed with a period equal to the dodecane chain diameter, indicating that dodecane molecules are forming layers with their chains oriented parallel to the surfaces. At 40 Å film thickness, the effective viscosity is already greatly enhanced over the bulk viscosity and seems to diverge as the film thickness is further reduced, consistent with the liquid becoming solid-like at small film thickness.

Not only does a liquid viscosity become strongly enhanced when confined, but it also becomes highly non-Newtonian. In general, the non-Newtonian behavior of a liquid is described by a power law:

$$\eta_{\text{eff}} \propto \left(\frac{d\gamma}{dt}\right)^n$$

(10.3)

where a positive exponent $n$ indicates shear thickening and a negative exponent shear thinning. Figure 10.4 shows how the shear thinning occurs for 27 Å thick films of dodecane and OMCTS: the effective viscosity decreases back towards the
Fig. 10.3. How the effective viscosity increases with decreasing film thickness for dodecane being sheared between mica surfaces in an SFA at 28°C. Dodecane has the chemical structure CH$_3$(C$_2$H$_2$)$_{10}$CH$_3$, so can be thought as a flexible linear chain of twelve carbon atoms, with 18 Å length and 4 Å chain diameter. At 28°C, dodecane’s bulk viscosity is 0.01 poise and is independent of shear rate. Reprinted with permission from Granick (1991). Copyright 1991, AAAS.

Fig. 10.4. Non-Newtonian shear thinning of liquids confined between mica surfaces in an SFA. Circles correspond to a 27 Å thick film of dodecane. Triangles correspond to a 27 Å thick film of OMCTS, [Si(CH$_3$)$_2$O-]$_4$, a compact (but flexible) ring shaped molecules of ~9 Å diameter. Reprinted with permission from Hu et al. (1991). Copyright 1991 by the American Physical Society.
bulk value with an exponent of $n = -2/3$. That the same power law dependence is observed both for films of the linear dodecane molecule and of OMCTS, a silicone oil with a globular shape, indicates that this shear thinning with the enhanced viscosity is a fairly general phenomenon.

So how does confinement lead to enhanced viscosity and solid-like behavior? The enhanced viscosity indicates that the molecular relaxation times in confined films are much longer than in the bulk liquid. For dodecane confined to a 27 Å film, a relaxation time of 0.05 seconds can be estimated from the onset of shear thinning at a shear rate of $20 \text{ s}^{-1}$ in Fig. 10.4; this relaxation time is about $10^8$ times slower than in the bulk (Hu et al. 1991). The likely explanation for the slower relaxation times during confinement is that when one molecule diffuses through the liquid, its neighbors have to be displaced a little bit to make room for this motion, and these molecules in turn have to displace their neighbors. When the liquid is confined between two stiff surfaces, however, the liquid molecules quickly bump up against these hard walls, severely limiting their displacements in the direction normal to the film. So, the thinner the film, the greater these displacements have to extend out laterally before a molecule can diffuse in the liquid film. This leads to higher activation volumes and activation energies for molecular diffusion through the liquid film as it becomes thinner (Hu et al. 1991). Attractive interactions between the walls and the liquid molecules further slow molecular motion by pinning molecules to the walls (Alba-Simionesco et al. 2006). Another way of looking at this is that confinement lowers the entropy of the liquid film by reducing the number configuration states available. This lowering of entropy shifts the film’s freezing point to higher temperatures increasing the liquid’s viscosity and eventually inducing a liquid-to-solid transition (Thompson et al. 1992; Weinstein and Safran 1998; Braun and Peyrard 2003; Gao et al. 2004). Similar liquid-to-solid transitions have also been observed for liquids in other types of confined spaces such as porous media (Christenson 2001; Alba-Simionesco et al. 2006).

Consistent with the liquid transitioning into a solid-like structure due to confinement, SFA experiments with these thin films also show development of a measurable yield stress when sheared, a clear indicator of solid-like behavior. For example, with the silicone oil OMCTS, this liquid-to-solid transition occurs when the film thickness is reduced from 62 to 54 Å (seven to six molecular layers) (Gee et al. 1990; Klein and Kumacheva 1998; Kumacheva and Klein 1998). This yield stress $\sim 0.1 \text{ MPa}$ is several orders of magnitude smaller than that needed to shear a bulk OMCTS crystal, indicating that this “solid-like” film is a long way from matching the crystalline solid.

Figure 10.5 shows how the yield stress or static friction force increases as the individual liquid OMCTS layers are squeezed out between mica surfaces in an SFA. The friction versus time trace in Fig. 10.5 also exhibits a sawtooth fine structure due to stick–slip motion of the sliding surfaces (Section 4.3.1), indicating that static friction is larger than kinetic friction. For these thin liquid films, the transition from stick to slip is usually interpreted as being due to
Fig. 10.5. How the friction force increases as the individual layers of OMCTS are squeezed out from the sliding mica surfaces. Reprinted with permission from Israelachvili (2005). Copyright 2005, Materials Research Society.

The shear stress inducing a melting transition within the solid film to initiate sliding (Thompson and Robbins 1990). Figure 10.6 illustrates how this might occur for spherically shaped molecules such as OMCTS. In Fig. 10.6, several possible disordered or slip processes are depicted as it is often unclear how the molecules slide past each other in these narrow gaps. When liquid films solidify in SFA experiments, the shear stresses are found to depend only weakly on sliding velocity, but, when the sliding velocity is raised above a threshold value, the stick–slip behavior disappears and smooth sliding occurs.

Recent experiments have cast some doubt, however, on the reliability of the SFA technique for determining the degree of solidification through measurements of yield stresses. In SFA experiments where the surfaces are prepared using a technique that minimizes the number of nanoparticles deposited (Frantz and Salmeron 1998), much lower yield stresses are observed than in earlier experiments, suggesting that some of the yield stress previously attributed to the solidifying of the liquid films may be coming from shearing of the contacts between the nanoparticles and the mica surfaces (Becker and Mugele 2003; Zhu and Granick 2003; Zhu and Granick 2004).

This model, that an order-to-disorder or melting transition occurs before sliding is initiated, also predicts that the separation distance $D$ between the sliding surfaces should increase slightly to provide the extra space needed for the disordered state: In order for the surfaces to be able to move sideways, they must first move slightly apart to unlock the molecules from their ordered structure. This dilatancy is seen both in the simple picture drawn in Fig. 10.6 and in molecular
CONFINED LIQUIDS

fig. 10.6. Idealization of how spherical molecules rearrange in response to a shear that induces a solid-to-liquid transition. Note that there are number of ways that disorder can occur during the slip process: (c) total disorder of the liquid molecules, (c)′ disorder of the internal layers, and (c)″ slippage between ordered layers. Courtesy of J. N. Israelachvili, University of California, Santa Barbara.

dynamic simulations (Thompson et al. 1992). SFA experiments have confirmed that about 0.1 Å dilatancy occurs or about one-tenth the width of the confined molecule (Demirel and Granick 2002).

Another mechanism that accounts for friction continuing to drop after sliding has been initiated is that the shearing process induces alignment of the liquid molecules enabling them to slide more easily past each other. This mechanism is thought to be prevalent for liquids whose molecular structure is more complicated than a simple spherical structure (Israelachvili 2005). Figure 10.7 shows this process schematically: When the shearing starts, the film first dilates to accommodate the molecular movement [Fig. 10.7(b)]. With continued shear, the molecules disentangle and start to orient with the shear direction enabling them to slide more easily past each other [Fig. 10.7(c)–(e)]. One’s first expectation might be that this alignment process takes place on the timescale of an individual molecular relaxation time (pico- to nanoseconds); in SFA experiments,
however, the timescale is found to be minutes or longer, during which time the sliding surfaces move many microns. The reason for this is that, when confined to these small gaps, the molecular movement becomes a highly cooperative affair, with many molecules needing to move to accomplish a molecular realignment.
This long timescale for the liquid structure to adapt the shearing process results in the friction of these molecules exhibiting a “memory” of how it has been previously sheared (Gao et al. 2004).

While these thin liquid films exhibit some solid-like features when confined, i.e. static friction, tendency to order, and high viscosity, they should not be thought of as a crystalline solid film; the ordering is more diffuse than in a solid, and the molecules continue to move about within the film, just at a much slower rate compared with the bulk. Fluorescence correlation spectroscopy experiments confirm that molecular diffusion still occurs within these solidified films. In these experiments, the diffusion of a tracer molecule is monitored within a solid-like film, and the diffusion coefficient was found to decrease exponentially away from the edge of the contact zone towards its center, indicating that the liquid film becomes more sluggish towards the center of the contact zone (Mukhopadhyay et al. 2002).

It should be noted that the SFA geometry provides a highly idealized contact environment for studying the enhanced viscosity and solidification during confinement. To what extent do these phenomena persist in less idealized contact geometries? Molecular dynamic simulations have found that, if the surfaces have sufficient roughness to prevent ordering of the liquid molecules into layers between them, that this prevents the solidification process (Gao et al. 2000; Jabbarzadeh et al. 2006). We should also not be surprised that the enhanced viscosity diminishes as the lateral extent of contact diminishes, as the fluorescence correlation spectroscopy experiments of Mukhopadhyay indicate that the molecules become less sluggish toward the edges of the SFA contact zone. AFM experiments, where the contact zone is orders of magnitude smaller in area than an SFA, have also shown much less enhancement of viscosity (Mate 1992b; Friedenberg and Mate 1996; O’Shea and Welland 1998).

While enhanced viscosity and solidification occur in confined contact zones, this sensitivity to roughness and to the lateral extent of the contact zone makes it difficult to predict how these phenomena will influence the tribological performance of a particular contacting interface.

10.2 Boundary lubrication

As discussed in Chapters 5 and 6, clean surfaces of metals and other high surface energy materials experience high adhesion forces when brought into contact. These high adhesive forces typically result in high friction and high wear rates, which can be reduced by the application of a boundary lubricant: a molecular monolayer deposited onto the contacting surfaces that resists penetration of the contacting asperities while reducing the adhesion and shear strength of the contacting interface.

Figure 10.8 shows an example where the friction between two metal surfaces is measured under ultra-high vacuum conditions to ensure that the surfaces can be kept clean until a molecular layer can be deliberately adsorbed onto them
Fig. 10.8. Static coefficient of friction measured at 120 K in an ultra-high vacuum chamber between two copper single crystals with (111) surfaces. The static friction decreases as the coverage of trifluoroethanol, CF₃CH₂OH, increases; ML, monolayer coverage. Reprinted from Gellman and Ko (2001) with permission from Springer Science and Business Media. Copyright 2001, Plenum Publishing Corporation.

(Gellman and Ko 2001). For the clean Cu(111) surfaces in Fig. 10.8, as well as for other clean metal surfaces under vacuum conditions, the static friction is very high, with \( \mu_s > 4 \). As a small prototype lubricant molecule, trifluoroethanol, is absorbed on the Cu(111) surface, the friction decreases with sudden drops occurring when the first and second monolayers (ML) complete and a limiting value of \( \mu_s \approx 0.3-0.4 \) being reached at high coverages.

10.2.1 Molecular mechanisms of boundary lubrication

Figure 10.9(a) illustrates the adsorbed structure of a boundary lubricant that has a long hydrocarbon chain and a reactive or polar end-group. These long chain
molecules adsorb with their reactive or polar end-group adhering strongly to the solid surface; close packing of the chains causes them to line up perpendicularly to the surface; and the neutral terminating end-group presents a surface with low surface energy and low shear strength. When opposing surfaces with this type of boundary lubricant layer are brought into contact [Fig. 10.9(b)], both the molecular adhesion between the chains from van der Waals forces and the bonding of the end-group to the surface provide substantial resistance to penetrating asperities, forcing most of the shearing to take place at the interface between the boundary lubricant layers except for those asperity summits where high contact pressures displace the lubricant layer.

This concept of boundary lubrication originates from William Hardy (1864–1933) who in the 1920s studied lubrication by long chain hydrocarbons with various reactive end-groups (Hardy and Bircumshaw 1925); some of his results are shown in Fig. 10.10. He found that, on metal surfaces, the hydrocarbons with reactive end-groups had lower friction coefficients than those without them, and that the friction coefficients decreased with increasing molecular weight. Friction coefficients of as low as 0.05 could be obtained for the longest chain molecules, much lower than for unlubricated surfaces. Since a film one molecule thick can provide effective boundary lubrication, a small concentration (0.1% to 1%) of boundary lubricant in the base lubricant oil is usually a sufficient source for this layer. For situations where the contacting surfaces cannot be immersed in a lubricant, a monolayer of boundary lubricant can be applied to the sliding

Fig. 10.9. (a) Self-assembled or Langmuir–Blodgett monolayer with the polar or reactive end-group adhering to a solid surface and the closely packed alkane chains standing upright. (b) Bowden and Tabor’s concept of how a boundary layer lubricant is displaced from an asperity contact (Bowden and Tabor 1950).
surfaces during their fabrication, in order to minimize friction, adhesion, and wear during use.

In recent years, the molecular details of the type of boundary lubrication illustrated in Fig. 10.9 (self-assembled or Langmuir–Blodgett monolayers with upright alkane chains) have become well established (Carpick and Salmeron 1997; Barrena et al. 1999; Tutein et al. 2000; Perry et al. 2001; Salmeron 2001; Chandross et al. 2004; Mikulski et al. 2005).

Low friction and strong resistance to penetration is maximized when the highest packing occurs between the alkane chains, as this packing maximizes the van der Waals cohesion between the chains. Higher friction occurs in low packing density situations, such as for short chain lengths or mixtures of different chain lengths; the higher friction originates from the chain segments having more room to adopt configurations away from linear, providing more channels for energy dissipation.

In contrast to the view of Bowden and Tabor that these molecules remain as rigid rods when squeezed between contacting surfaces (Fig. 10.9), current evidence indicates that the chains first bend at their terminal –CH₃ methyl groups when a contact pressure is applied. This is illustrated in Fig. 10.11, which shows how the terminal methyl groups of an attached C₁₃ alkane chain bend away from the sliding interface at moderate contact pressures, as determined by molecular dynamic simulations (Mikulski et al. 2005).

These molecular dynamic simulation results are consistent with the results of AFM and sum-frequency generation spectroscopy (Carpick and Salmeron 1997;
Fig. 10.11. Molecular dynamic simulation results of the molecular structure of a self-assembled monolayer in a sliding contact (Mikulski et al. 2005). The self-assembled monolayer is composed of an alkane chain with 13 carbon atoms attached at the bottom to a diamond (111) surface in the (2×2) arrangement. The top surface is amorphous carbon, which is sliding over the boundary lubricant at 0.22 Å/ps. At a low contact pressure of 0.9 GPa, the methyl groups are only slightly perturbed away from their original configuration; increasing the contact pressure to 7.2 GPa decreases monolayer thickness by bending the terminal methyl groups away from the sliding interface. Courtesy of Judith A. Harrison and Paul T. Mikulski, United States Naval Academy.

Barrena et al. 1999; Perry et al. 2001; Salmeron 2001). Fig. 10.12 shows a few of the molecular structures adopted at different levels of compression by these linear chain boundary lubricants (Carpick and Salmeron 1997; Perry et al. 2001; Salmeron 2001). For relatively light contact pressures of 10–200 MPa, the terminal –CH₃ groups are bent over in a gauche distortion that slightly reduces the film thickness while maintaining packing density. With increasing pressure under the contacting asperity, the molecular layer compresses by the linear chains’ collectively adopting discrete tilt angles away from normal (not shown
260 LUBRICATION IN TIGHT SPOTS

Terminal gauche defects (–CH$_3$)

Plastic deformation of gold substrate

Displacement

Non-contact

10–200 MPa

~1–2 GPa

~1 GPa

Fig. 10.12. Structure of a boundary lubricant layer subjected to different levels of compression. (Top right) At light pressures of 10–200 MPa, the molecular chains remain upright but the terminal CH$_3$ groups develop gauche defects. (Bottom left) A sharp AFM tip or sharp asperity displaces the lubricant layer and penetrates the film when the contact exceeds about 1 GPa. (Bottom right) For blunt contacts, the molecules remain trapped in the contact zone and the solid surfaces plastically deform around them. Reproduced with permission from Carpick and Salmeron (1997). Copyright 1997 American Chemical Society.

in Fig. 10.12). For sharp asperity contacts with $R < 100$ nm, all the lubricant molecules are displaced from the contact zone, and the asperity breaks through the boundary lubricant layer to make solid–solid contact when the contact pressure exceeds a threshold on the order of 1 GPa. For blunt contacts (typically when $R > 100$ nm), the lubricant molecules remain trapped in the contact zone while the solid surfaces undergo plastic deformation.

10.2.2 Molecularly thin liquid boundary lubricant layers

The strong cohesion between the close packed chains of the boundary lubricants depicted in Figs. 10.9, 10.11, and 10.12 results in these lubricant layers behaving like a soft-solid film. While this strong cohesion is good for resisting penetration of contacting asperities, the soft-solid character means that the films are slow to recover from damage. So, while these boundary layers may protect the surface the first time a sliding contact passes over it, the film may be too damaged to protect the surface on subsequent passes unless a mechanism exists for quickly replenishing the displaced lubricant. Within bearings that are immersed in a lubricating oil, this replenishment is accomplished by having a sufficiently high concentration of the boundary lubricant in the base oil (typically 0.1% to 1%) to ensure rapid adsorption onto the exposed surface.
Asperity

Fig. 10.13. (a) A liquid lubricant film being displaced from an asperity by a contacting slider. (b) Lubricant from the surrounding area then flows or diffuses to replenish the lost lubricant before (c) the slider recontacts the asperity on a subsequent pass.

For surfaces that cannot be immersed in oil or covered with grease containing a boundary layer additive, the closely packed, soft-solid layer depicted in Figs. 10.9, 10.11, and 10.12 may not provide sufficient boundary lubrication for interfaces that experience frequent contacts. In these cases, a more appropriate choice is to have a very thin (often only a few nanometers thick) liquid film of lubricating molecules, as depicted in Fig. 10.13. Then, when contact displaces lubricant from an asperity summit, the liquid lubricant molecules from the surrounding area flow or diffuse to cover the exposed surface. Since the molecules within these films maintain their liquid character, they do not form the ordered, upright structures depicted in Figs. 10.9, 10.11, and 10.12, but instead have the much more disordered liquid molecular structure.

As discussed in the example below, reactive or polar end-groups still help improve the lubricating properties, as the bonding of the end-group helps resist displacement during contact. These reactive end-groups have the disadvantage,
however, of acting like “anchors” when these molecules diffuse across surfaces, greatly slowing down the replenishment process (Min et al. 1995; Ma et al. 1999; Scarpulla et al. 2003). As a consequence, close attention is paid to controlling the degree to which the liquid lubricant molecules are “bonded” and the degree to which the remaining film is “free” to flow or to diffuse over a surface, in technical applications such as the lubricant film deposited on the disk in a disk drive.

For those lubricant films exposed to ambient air or vacuum conditions, the lubricants need to have very low volatility. For tribological systems in a sealed environment, however, more volatile lubricants can be used, which can provide an additional replenishment mechanism of vapor redeposition. While used infrequently, vapor lubrication, where a reservoir of volatile boundary lubricant provides a continual replenishment source, remains a potentially attractive means of providing boundary lubrication in sealed environments (Hanyaloglu and Graham 1994; Henck 1997; Gellman 2004).

10.2.2.1 Example of the importance of end-groups in a liquid lubricant film

The atomic force microscope (AFM) provides a convenient way to study how molecular end-groups alter the surface forces and lubrication properties of a single asperity contact at the molecular level. Figure 10.14 shows the force vs. distance curves as an AFM tip is brought into contact with silicon wafers covered with thin liquid films of perfluoropolyether polymer lubricants, that differ slightly in that one has a reactive alcohol (–OH) end-group and the other an unreactive and relatively neutral –CF₃ end-group.

When the tip first contacts the lubricant films at point M on the curve, a sudden attractive force is observed due to the liquid lubricant’s forming a meniscus around the tip as discussed in Section 6.3.1.1. For the lubricant with unreactive end-groups, the normal force becomes more attractive with decreasing separation distance. This attractive force is the sum of the capillary force from the meniscus and the solid–solid attraction, mostly likely the van der Waals force mediated by the intervening liquid lubricant. The force quickly turns repulsive once the hard wall of solid–solid contact is reached.

When the lubricant has alcohol end-groups, a very different force behavior is observed as the tip penetrates the lubricant film: the normal force becomes increasingly more repulsive with decreasing separation distance, indicating that the attractive van der Waals force is counteracted by a repulsive force associated with the alcohol end-groups. This repulsive force comes first from the compression of the polymer molecules beneath the tip, followed by the force needed to overcome the hydrogen bonding between the alcohol end-groups and the silicon oxide surface. Figure 10.15 illustrates how the AFM tip interacts with these two polymer lubricants with different termination.

A major way in which reactive end-groups on lubricant molecules facilitate boundary lubrication is a dramatic increase in the contact pressure that a thin layer of lubricant molecules can support before the lubricant is squeezed out and
Fig. 10.14. Force versus separation distance as a tungsten AFM tip with \( \sim 100 \text{ nm} \) radius approaches a silicon wafer covered with \( 30 \text{ Å} \) thick perfluoropolyether polymer lubricant films that have either (top) unreactive \(-\text{CF}_3\) end-groups or (bottom) reactive alcohol end-groups. These perfluoropolyether lubricants [with the chemical structure \( R-\text{CF}_2(\text{OC}_2\text{F}_4)_m(\text{OCF}_2)_n\text{O-R} \) where \( R = -\text{CF}_3 \) or \(-\text{CH}_2\text{OH}\)] are linear chain polymers, about \( 100 \text{ Å} \) in length and \( 7 \text{ Å} \) in diameter, and are commonly used as lubricants in disk drive, aerospace, and MEMS applications. The zero separation distance is defined to occur at the onset of hard wall repulsion. A negative load force corresponds to an attractive force acting on the tip. Reprinted with permission from Mate (1992b). Copyright 1992, by the American Physical Society.

Fig. 10.15. Schematic of a polymer lubricant forming a liquid meniscus around an AFM tip or asperity contact. (a) If the lubricant has neutral end-groups, it is easily squeezed out. (b) If the lubricant has reactive or polar end-groups, it resists being squeezed out due to the bonding forces between the end-groups and the tip and sample surfaces.
solid–solid contact occurs. For the two lubricants shown in Fig. 10.14, an extra 100 nN load force has to be exerted on the tip to make hard wall contact for the polymer lubricant with alcohol end-groups compared to the same polymer with neutral end-groups. As the contact zone underneath this AFM tip is roughly about $10^2$ nm$^2$, this extra 100 nN load corresponds to an average contact pressure of 1 GPa for squeezing out the lubricant with alcohol end-groups.

The AFM friction experiments in Fig. 10.16 show that the alcohol-terminated lubricant maintains its load bearing capacity during sliding. Again, the liquid meniscus forms at point M when the tip contacts the lubricant film during approach [Figs. 11.16(a) and (c)], but the friction is negligible as the tip penetrates the lubricant film and only increases when solid–solid contact is reached at zero separation [Figs. 11.16(b) and (d)]. This negligible friction indicates that little, if any, enhancement of viscosity occurs for lubricant confined in the AFM contact geometry, in stark contrast to what happens in an SFA experiment. This difference is attributed to the lubricant molecules being much less confined underneath an AFM tip, as the AFM contact zone is more than $10^3$ times smaller than an SFA contact zone. The smaller volume of liquid being sheared by an AFM tip

![Figure 10.16](image.png)

**Fig. 10.16.** Friction and normal force as a function of the normal sample position for 30 Å thick perfluoropolyether polymer lubricant films on silicon wafers (same as the lubricants in Fig. 10.14). A 100 nm radius tungsten tip slides back and forth over the sample at a velocity of 1 μm/s as the sample is slowly brought into contact and then withdrawn. Each data point is the average over one sliding oscillation. (a) and (b) are for the lubricant with unreactive or neutral end-groups, while (c) and (d) are for the lubricant with alcohol end-groups. Reprinted with permission from Mate (1992b). Copyright 1992, by the American Physical Society.
sets a much lower upper limit on the number of molecules that can be involved in the collective motions believed to be responsible for the enhanced viscosity (Hu et al. 1991). It also means that the liquid molecules only have to diffuse a short distance to escape from underneath the sliding tip.

When the neutral end-group lubricant sample is withdrawn, an adhesive force of almost 300 nN is needed to break the solid–solid sliding contact [Fig. 10.16(a)]. For the alcohol-terminated lubricant, however, no adhesive force is present when the tip breaks the solid–solid sliding contact [Fig. 10.16(c)]. This low adhesion is attributed to the alcohol end-groups forcing their way back into the sliding interface to reestablish the hydrogen bonding with the tip and sample surfaces. The friction coefficient of \( \mu = 0.6 - 0.7 \) [Fig. 10.16(b) and (c)] is similar to the value observed when unlubricated surfaces were used in these experiments. This indicates that once hard-wall contact occurs, the lubricant molecules are sufficiently squeezed out from between the sliding surfaces that they longer influence the shear strength of the asperity contact junctions. This shear strength can be lowered substantially by using end-groups on the lubricant, that form covalent bonds with the sliding surfaces, providing superior resistance to penetration by the tip at high contact pressures.

### 10.3 Capillary and disjoining pressures

As discussed in previous chapters, for contacting surfaces not under vacuum or immersed in a liquid, menisci form around the contact points, either from the capillary condensation of vapors or the migration of adsorbates toward the contact points. Relying on adventitious contaminants for lubrication, however, can lead to unpredictable results. To better control the material that ends up in the menisci around the contact points, either a thin film of mobile lubricant is applied the surfaces prior to contact or a volatile lubricant is added to the vapor surrounding the contacting surfaces.

Next we discuss how a thin film of liquid deposited on surfaces redistributes when contact occurs, as illustrated in Fig. 10.17. The two driving forces for this redistribution are the capillary pressure, which pulls the liquid lubricant into the menisci around the contact points, and the disjoining pressure, which draws the liquid lubricant out of the meniscus and back onto the solid surfaces (Mate 1992a). The physical origins of capillary pressure were discussed in Section 5.2. The disjoining pressure of a liquid film was briefly discussed in Section 7.3.8 in connection with the van der Waals interaction between the film and substrate; in the next section, we present a fuller discussion of this fundamental parameter of liquid films.

#### 10.3.1 Disjoining pressure

The disjoining pressure \( \pi \) represents the interaction energy per unit volume between a liquid film and the adjacent solid substrate, as measured relative to the interaction energy experienced by the same liquid molecules at the bulk liquid’s surface. Formally, the disjoining pressure is defined as the negative derivative of
Fig. 10.17. (a) Two surfaces coated with a thin film of liquid lubricant prior to contact. (b) After contact the lubricant redistributes to form “toe dipping” and “pillbox” menisci around the contact points.

Fig. 10.18. A simplistic view of disjoining pressure. A polymer molecule on the surface of the bulk liquid experiences a force $F_0$ pulling that molecule towards the liquid, while the same molecule on the surface of a liquid film experiences a slightly different force $F_0 + F'$. The disjoining pressure corresponds to $F'/A_{mol}$ where $A_{mol}$ is the molecular area.

The Gibbs energy per unit area with respect to film thickness; but, informally, the disjoining pressure can be thought as the extra force $F'$ experienced by a molecule at the surface of the liquid film due to the presence of the nearby substrate, divided by the area $A_{mol}$ of that molecule

$$\pi(h) = \frac{F'}{A_{mol}}$$

where $h$ is the thickness of the liquid film. This is illustrated in Fig. 10.18.

If we adopt the sign convention of $F'$ being positive when the substrate exerts an attractive interaction on the liquid film higher than that of the bulk liquid, then a positive disjoining pressure $\pi$ corresponds to an attractive interaction that acts to thicken the liquid film. Such an attractive interaction acts to pull liquid out of reservoirs, such as the menisci around contact points, to thicken the film or to pull a non-uniform liquid film flat (Robbins et al. 1991).

If the strength of the interaction between the substrate and the liquid film is less than that of the liquid with itself, the disjoining pressure $\pi$ is negative, and
the interaction is considered repulsive. With such a repulsive interaction, the substrate repels the liquid in the film, causing the film to thin with the excess liquid forming dewetting droplets (Kim et al. 1999).

Following Derjaguin and Churaev, the disjoining pressure can be represented as the sum of three main contributors

$$\pi = \pi_{vdw} + \pi_{bond} + \pi_{struct}$$

where the $\pi_{vdw}$ is the contribution from the van der Waals forces acting between the film and the substrate, $\pi_{bond}$ is the contribution from the other bonding interactions between the liquid molecules and substrate (covalent, ionic, hydrogen bonding, etc.), and $\pi_{struct}$ is the contribution from the liquid molecules having a different structure in a film than in the bulk liquid (Derjaguin and Churaev 1974). For many liquid films, the van der Waals term $\pi_{vdw}$ dominates the disjoining pressure and can be expressed as

$$\pi(h) = \frac{A_{SLV}}{6\pi h^3}$$

where $A_{SLV}$ is the Hamaker constant for this solid–liquid–vapor system (Section 7.3.8). Eq. (10.6) predicts that the disjoining pressure diminishes rapidly with a $h^{-3}$ dependence as the film thickness increases. Also, $\pi \to 0$ for thick films, which is to be expected since the film properties must approach those of the bulk liquid as $h \to \infty$.

10.3.2 Distribution of a liquid film around a pore opening

Figure 10.19(a)–(b) illustrates how liquid redistributes itself around a pore opening on a surface so as to minimize energy. At equilibrium, the differential in energy $dE$ equals zero or

$$dE = (\mu_{surface} - \mu_{pore}) \, dn + P_{cap} \, dV = 0$$

where $\mu_{surface}$ and $\mu_{pore}$ are the chemical potentials of the liquid on the surface and in the pore and $P_{cap}$ is the capillary pressure of the liquid in the pore. The first term in the middle portion of eq. (10.7) represents how the potential energy of the liquid molecules changes as they move from the pore to the film on the surrounding surface; the second term represents the work done by the capillary pressure during this move. As the disjoining pressure $\pi$ represents the interaction energy per unit volume between the liquid film and the surface with respect to the bulk liquid in the pore, we can write

$$-\pi dV = (\mu_{surface} - \mu_{pore}) \, dn.$$  

So, the equilibrium condition eq. (10.7) becomes simply

$$\pi(h) = P_{cap}(r_{m-pore}).$$

Therefore, the liquid redistributes itself until the disjoining pressure of the liquid film around the pore opening equals the capillary pressure in the pore opening.
Fig. 10.19. How a liquid redistributes over a pore opening. (a) If the liquid surface is initially flat over the pore, the attraction of the solid for the liquid (positive disjoining pressure) pulls the liquid out of the pore, thickening the liquid film around the pore opening. (b) Equilibrium is reached when the capillary pressure of the liquid in the pore balances the disjoining pressure of the liquid film. (c) Liquid from a thin film with thickness $h$ wicking up around a spherical particle on a solid surface.

Eq. (10.9) is a fairly generally result, applicable to other situations where a thin liquid film is in equilibrium with a nearby meniscus, such as the meniscus that forms from liquid wicking up around the spherical contact illustrated in Fig. 10.19(c).

For a cylindrically shaped pore where the meniscus over the opening has a spherically shaped surface with a radius of curvature $r_{m\text{-pore}}$, the capillary pressure is given by eq. (5.5): $P_{\text{cap}} = 2\gamma_L / r_{m\text{-pore}}$, where $\gamma_L$ is the liquid surface tension. If the liquid film’s disjoining pressure is from van der Waals attraction [eq. (10.6)], then, when a liquid film with thickness $h$ surrounds the pore opening [Fig. 10.19(b)], the equilibrium meniscus radius $r_{m\text{-pore}}$ is given by

$$\frac{A_{\text{SLV}}}{6\pi h^3} = \frac{2\gamma_L}{r_{m\text{-pore}}} \quad (10.10)$$

$$r_{m\text{-pore}} = \frac{12\pi \gamma_L h^3}{A_{\text{SLV}}}$$
How small can this meniscus radius \( r_{m\text{-pore}} \) be? The smallest meniscus radius occurs for the thinnest liquid film that is still continuous, about one layer of molecules lying flat, or \( h \sim 0.5 \text{nm} \). If we also assume that \( \gamma_L = 20 \text{mN/m} \) and \( A_{\text{SLV}} = 10^{-19} \text{J} \), then the smallest possible radius would be

\[
r_{m\text{-pore}} = \frac{12\pi (20 \text{ mN/m}) (0.5 \text{ nm})^3}{10^{-19} \text{ J}} = 0.94 \text{ nm}.
\]

Eq. (10.10) indicates that for any liquid film thickness greater than one molecular layer, \( r_{m\text{-pore}} > h \).

### Example: Measurement of the disjoining pressure of a perfluoropolyether lubricant

A similar analysis can be done for a liquid covering a groove on a surface, except that in this case the capillary pressure is given by \( P_{\text{cap}} = \gamma_L/r_{m\text{-groove}} \) where \( r_{m\text{-groove}} \) is the radius of the cylindrically shaped meniscus over the groove. Figure 10.20 shows an example where the lubricant distribution over a groove has been used to measure the disjoining pressure of the film on the surrounding surface. The lubricant in this case is the same perfluoropolyether lubricant as the one without reactive end-groups discussed for Fig. 10.14. These measurements were done by using an AFM to measure the meniscus radius of a lubricant filled groove on a silicon wafer, for different thicknesses of the lubricant film on the surrounding wafer. From the meniscus radius, the capillary pressure of the liquid in groove is determined; equating this capillary pressure with the disjoining pressure of the liquid film on the silicon wafer determines the disjoining pressure for that film thickness, which is measured by ellipsometry. These measured values are somewhat less than those obtained previously by the author for similar perfluoropolyether polymer lubricants on silicon wafers; for these earlier measurements, the break-free length of an AFM tip being withdrawn from a lubricated sample was used to determine the disjoining pressure (Mate and Novotny 1991).

Since the silicon wafer typically has a 1 nm thick silicon oxide surface layer, a somewhat different equation than eq. (10.6) needs to be used to account for the van der Waals contribution to the disjoining pressure for this three-interface system (air–film–SiOx–Si):

\[
\pi (h) = \frac{A_{\text{SiOx}}}{6\pi h^3} + \frac{A_{\text{Si}} - A_{\text{SiOx}}}{6\pi (h + h_{\text{SiOx}})^3}.
\]

The solid line in Fig. 10.20 is obtained by fitting eq. (10.12) to the measured values of disjoining pressure. This fitting yields the values of \( A_{\text{SiOx}} = 9.1 \times 10^{-21} \text{ J} \) and \( (A_{\text{Si}} - A_{\text{SiOx}}) = 7.0 \times 10^{-20} \text{ J} \) for the two Hamaker constants and \( h_{\text{SiOx}} = 1.0 \text{ nm} \) for the silicon oxide thickness.

From Fig. 10.20, we see that, for a film thickness greater than 10 nm, the disjoining pressure is fairly small (<0.03 atm). For a thickness less than 2 nm
Fig. 10.20. Measurements of how disjoining pressure varies with film thickness for a perfluoropolyether polymer liquid deposited on a silicon wafer (the same lubricant as in Fig. 10.14 without reactive end-groups). The solid line shows the fit to the data by assuming a van der Waals interaction between lubricant and the silicon wafer. These measurements were done by measuring the meniscal radius of the liquid in a narrow groove on a silicon wafer when different thicknesses of liquid lubricant were applied. From the meniscal radius, the capillary pressure of the liquid in the groove is determined; equating this capillary pressure with the disjoining pressure of the liquid film on the silicon wafer around the groove determines the disjoining pressure for that film thickness, which is itself determined by ellipsometry. Reprinted with permission from Fukuzawa et al. (2004). Copyright 2004, IEEE.

The relevant range when these perfluoropolyether lubricants are used to lubricate surfaces in disk drives, the disjoining pressure is quite large: eq. (10.12) indicates that $\pi = 2\text{ atm}$ at $h = 2\text{ nm}$, rising to $10\text{ atm}$ for $h = 1\text{ nm}$ at which point $r_{m\text{-groove}} = 26\text{ nm}$, smaller than what can be measured by AFM but still such that $r_{m\text{-groove}} \gg h$.

10.3.3 Lubricant distribution between contacting surfaces
Figure 10.17 shows schematically how a thin liquid film wets the gap between rough contacting surfaces. As discussed in Section 6.3.3, the different wetting regimes between contacting surfaces are labeled toe dipping, pillbox, flooded, and immersed (Matthewson and Mamin 1988).

As illustrated in the left-hand inset of Fig. 10.21, when the equilibrium meniscus radius $r_m$ is less than $D/2$ (assuming that $h << D$ and where $D$ is a typical gap separation between the surfaces), the system conforms to the toe dipping
If the interaction between the liquid film and the solid surfaces is predominantly van der Waals attraction, then equating the capillary pressure in the meniscus around a toe dipping asperity contact to the lubricant film’s disjoining pressure yields

\[ P_{\text{cap}} = \frac{\gamma_L r_m}{r_m} = \frac{A_{\text{SLV}}}{6\pi h^3}. \]  

(10.13)

[This expression for \( P_{\text{cap}} \) assumes \( r_m << R \) (section 5.2.1).] The equilibrium meniscus radius around the contacting asperity is then

\[ r_m = \frac{6\pi \gamma_L h^3}{A_{\text{SLV}}}. \]  

(10.14)

For a film thicker than a monolayer, the meniscus radius \( r_m \) is always larger than the lubricant film thickness \( h \) and increases rapidly as \( h \) increases.

If the thickness \( h \) is large enough to generate a meniscus radius \( r_m \) greater than \( D/2 \) (but still assuming that \( h << r_m \) and \( D \)), the meniscus enters the pillbox regime—conditions to the right of the line plotted in Fig. 10.21—and starts to flood the gap between the surfaces. In the pillbox regime, the capillary pressure is set by the separation distance: \( P_{\text{cap}} = 2\gamma/D \). Initially, \( P_{\text{cap}} \) is higher than
Fig. 10.22. How the meniscus force $L_{men}$ adds to the external loading force $L$ to increase static friction $F_s$.

$\pi(h)$ and pulls lubricant into the pillbox shaped menisci. This flooding process continues until the film thickness is reduced to the point where the disjoining pressure balances the capillary pressure or until the entire gap floods where $2r_m > D$.

10.3.4 Meniscus force

While a thin film of liquid lubricant can be effective for lubricating surfaces that cannot be immersed in the lubricant, care must be taken to control the lubricant film thickness and the roughness of the contacting parts, to ensure against excessive meniscus and friction forces. The meniscus forces come from the capillary pressure of the lubricant that wicks up around the contact points while the contacting parts are at rest. In recent years, substantial progress has been made at modeling this process so that a suitable combination of lubricant film thickness and surface roughness can be chosen to ensure that the meniscus and friction forces generated by the equilibrium distribution of the lubricant are in an acceptable range (Mate 1992a; Gui and Marchon 1995).

A principal concern is that the meniscus force from the lubricant results in excessive static friction. As illustrated in Fig. 10.22, the net meniscus force $L_{men}$ adds to the externally applied load $L$. From Amontons’ law, the static friction $F_s$ is given by

$$F_s = \mu (L + L_{men})$$  \hspace{1cm} (10.15)

where $\mu$ is the coefficient of friction.

10.3.4.1 Example: Stiction of a recording head slider

In Figure 10.23, the force needed to initiate sliding of a recording head slider on a disk surface is plotted as function of the amount of time the slider has been sitting at rest on that disk. This static friction force is often referred to as stiction. The disk surface used in these experiments was textured with bumps fabricated by high power laser pulses; these laser-textured bumps have a uniform height of 38 nm and are spaced every 50 $\mu$m. (Laser texturing has been previously discussed in Sections 3.4.1.2.1 and 6.3.3.1.1.) The use of a laser-textured disk provides a well defined separation of $D \sim 38$ nm over most of the apparent contact area, as illustrated in Fig. 10.24(a).
Fig. 10.23. Experimental measurements from the author’s lab of the force needed to initiate sliding (static friction or “stiction”) of a recording head slider after sitting on a laser-textured disk for different periods of time. The average height of the laser-textured bumps is 38 nm, making the separation distance $D$ also 38 nm. The lubricant is the perfluoropolyether polymer with alcohol end-groups.

From the plot in Fig. 10.21, we can estimate that for $D \sim 38$ nm the crossover from the toe dipping and pillbox behavior should occur when the lubricant thickness is increased above 16 Å. The data in Fig. 10.23 agree quite well with this crossover estimate: For lubricant thicknesses of 6 and 13 Å, the static friction remains constant at 18 mN out to rest times as long as 6000 minutes (4.2 days), consistent with toe dipping behavior, where the meniscus force is independent of the meniscus size. For a lubricant thickness greater than 16 Å, however, static friction increases with rest time, consistent with pillbox menisci that slowly grow with time, increasing the meniscus loading force and the resulting static friction force. The thicker the lubricant film in this pillbox regime, the faster lubricant flows into the pillbox shaped menisci, and the faster static friction increases.

10.3.4.2 Calculating meniscus force

To minimize the meniscus force and resulting static friction, the lubricant film thickness needs to be sufficiently thin and the roughness sufficiently large to ensure that the system stays in the toe dipping regime to the left of the line plotted in Fig. 10.21. In the toe dipping regime, the meniscus force $f_{td-i}$ acting on a single asperity contact with radius $R_{td-i}$ is given by eq. (6.41)

$$f_{td-i} = 4\pi\gamma_L R_{td-i}.$$  (10.16)

(Note that this equation for the meniscus force assumes that the lubricant contact angle $\sim 0^\circ$; this is reasonable as the lubricant needs to wet the contacting
FIG. 10.24. (a) Disk drive slider sitting on a disk surface textured with a laser to have bumps of a uniformed height. Expanded views illustrate toe dipping and pillbox lubricant menisci forming around the bump–slider contacts. (b) For stiction-free sliders, small pads, tens of nanometers high and tens of microns in width, are deposited on the slider’s bottom surface away from the recording head so that the slider sits on the disk at a slight angle. Since the disk and sliders are very smooth (rms roughness $\sim1$ nm), the lubricant floods the pad–disk interface and forms a meniscus around each pad.

surfaces in order to spread as a thin film.) The total meniscus force $L_{\text{men}}$ in the Toe Dipping regime is the sum over the individual meniscus asperity forces:

$$L_{\text{men}} = \sum f_{\text{td-i}} = 4\pi\gamma_L \sum R_{\text{td-i}} = 4\pi\gamma_L n_{\text{td}} R_{\text{td-ave}}$$

(10.17)

where $n_{\text{td}}$ is the number of asperities in contact and $R_{\text{td-ave}}$ is their average radius. [This is the same result as eq. (6.42)]. From eq. (10.17), we see that,
when in the toe dipping regime, the total meniscus force depends only on the number of asperities in contact and their average radius, and is independent of lubricant film thickness and separation distance. As pointed out in the example in Section 6.3.3.1.1, \( L_{\text{men}} \) and \( n_{\text{td}} \) are coupled: an increase in the number of contacting asperities \( n_{\text{td}} \) increases the total loading force \( (L + L_{\text{men}}) \), which leads to a further increase in \( n_{\text{td}} \). This increasing loading force also reduces the separation distance \( D \), and, at high enough loading force, portions of the contacting interface are pushed into the pillbox regime.

As the lubricant film thickness is increased or the separation between the contacting surfaces is made smaller, those portions of the contacting interface with the smallest separations eventually cross over to the pillbox side of the line plotted in Fig. 10.21. In the pillbox regime, the meniscus force \( f_{p-i} \) generated by each pillbox shaped meniscus is given by the product of its area \( A_{p-i} \) and its capillary pressure \( P_{\text{cap}-i} = 2\gamma_L/D_i \), where \( D_i \) is the separation distance at the edge of that pillbox meniscus. In the pillbox regime, the net meniscus force \( L_{\text{men}} \) is the sum over the toe dipping and pillbox menisci

\[
L_{\text{men}} = 4\pi \gamma_L \sum R_{\text{td}-i} + \sum \frac{2\gamma_L}{D_i} A_{p-i}.
\]  

(10.18)

Estimating the meniscus forces from pillbox menisci is tricky for several reasons:

- The pillbox meniscus areas \( A_{p-i} \) continue to grow until the disjoining pressure equilibrates with the menisci’s capillary pressure. The meniscus growth rate depends on the flow of a viscous lubricant through nanometer thick films, which can be difficult to estimate. The time scales for these menisci to reach equilibrium can range from fractions of a second to many months. For example in Fig. 10.23, the pillbox menisci that form for lubricant thickness >16 Å do not reached equilibrium even after 6000 minutes.

- As the pillbox menisci grow, the lubricant film thins until its disjoining pressure equals the pillbox menisci capillary pressure. Sometimes the entire contacting interface floods before these pressures equilibrate and the system enters the immersed wetting regime (Section 6.3.3.3). This is actually an easy system for which to calculate the meniscus force, as the meniscus area is the total area immersed and the capillary pressure is equal to the disjoining pressure of the lubricant film surrounding the contacting interface.

- As the pillbox menisci grow, the meniscus force grows, pulling the contacting surfaces closer together. The interface can collapse if the separation reduction results in a meniscus force rising faster than the elastic force from the contacting asperities. This is especially a concern for those surfaces where the loading force is supported by a low density of asperities (Gui and Marchon 1995; Gui et al. 1997).

10.3.4.2.1 Example: Calculation of stiction force of disk drive sliders in the pillbox regime Even with these possible complications, the meniscus force in the pillbox regime and its impact on static friction can be calculated by making
a suitable set of assumptions. In this example, we discuss the method of Gui and Marchon for calculating the static friction or stiction force for a disk drive slider sitting on laser-textured bumps when it is in the pillbox regime (Gui and Marchon 1995; see also Bennett et al. 1997). This model assumes:

- The capillary pressure of the lubricant in the menisci equals the disjoining pressure of the lubricant in the film on the disk surface, i.e., that the pillbox menisci have reached equilibrium.
- The volume of the lubricant between the slider and disk is constant.
- The meniscus force causes the separation distance $D$ to decrease due to the elastic deformation of the laser-textured bumps.

Due to the elastic deformation of the laser-textured bumps, the separation distance $D$ decreases with increasing pillbox menisci size, so the final values of $D$, capillary pressure, and lubricant thickness $h$ are coupled. The final values $h$ and $D$ are determined by the equation

$$h_0 = \left( \frac{D}{a} \right)^{1/3} - \frac{N \pi RD^2}{A_{\text{slider}}} + \frac{D}{2 \gamma_L A_{\text{slider}}} \left[ D - \left( \frac{D}{a} \right)^{1/3} \right]$$

$$\times \left( \frac{4}{3} N E_c R^{1/2} (H - D)^{3/2} - L \right)$$

(10.19)

where $h_0$ is the initial lubricant film thickness, $E_c$ is the composite elastic modulus of the bump–slider contacts, $H$ is the height of the laser-textured bumps, $N$ is the number of bumps, and $L$ is the externally applied load, $A_{\text{slider}}$ is the total slider area, and $a = 12 \pi \gamma_L / A_{\text{SLV}}$. Gui and Marchon’s procedure for calculating the meniscus force from the pillbox menisci is to solve eq. (10.19) numerically for $D$, which is then used to calculate the capillary pressure and meniscus area, which are in turn used to calculate the meniscus force $L_{\text{men}}$. Finally, the static friction is calculated from eq. (10.15) using a value of 0.2 for the friction coefficient $\mu$.

Figure 10.25 shows the values that Gui and Marchon calculated for the static friction or stiction force over the range of lubricant thickness and bump heights in the pillbox regime. The stiction force increases as either the initial lubricant thickness $h_0$ or separation distance $D$, as set by bump height $H$, is decreased. If the bump height $H$ drops below a certain height for a given lubricant thickness, however, the elastic deformation of bumps becomes unstable towards the increasing meniscus force, and the interface collapses; in Fig. 10.25, the left ends of the plotted lines terminate at these points. At high bump heights, the menisci enter the toe dipping regime, and the stiction force asymptotes to the value of 12 mN.

10.3.4.2.2 Padded or stiction-free slider To achieve areal storage densities greater than 100 Gbit/in$^2$ in disk drives, the magnetic spacing between the recording head and the disk’s magnetic media needs to be less than 16 nm in the disk surface’s data zones (see Fig. 1.4). At this small spacing, a design dilemma
occurs regarding whether or not to use laser-texturing in the disk landing zone: If the bump heights are less than about 20 nm, the gaps between the bumps tend to flood with lubricant; but, if the bumps are higher than the slider’s flying height, the resulting interference during the start–stop cycles, when the drive is powered on and off, can lead to unacceptable wear of the recording head.

For the most part, this dilemma has been resolved by switching to padded sliders on lightly textured disk surfaces as illustrated in Fig. 10.24(b). These padded sliders are often referred to as stiction-free sliders (Yamamoto et al. 1998), even though they do not actually eliminate stiction, but only reduce its value to an acceptable level. In this design, pads are deposited on the bottom surface of the slider away from the end with the recording head. When the slider flies over the disk with a slight pitch angle (\(\sim 100 \mu \text{rad}\)), the recording head can only be a few nanometers above a smooth disk surface. When the slider lands on the disk surface, the slider sits either solely on the pads or with a slight tilt angle so that its trailing edge also touches the disk.

The pads are fabricated such that their height \(H > 2r_m\), to ensure that the area between the pads does not flood with lubricant when the slider sits on the disk. Instead, menisci form around the pads’ edges, and the pad–disk interfaces are in the immersed wetting regime. Calculating the meniscus force in this geometry is straightforward; it is just the total pad area \(A_{\text{pad}}\) (including any flooded area
at the trailing edge) times the capillary pressure, which equals the disjoining pressure of the lubricant film:

\[ L_{\text{men}} = \frac{\gamma_L}{r_m} A_{\text{pad}} = \frac{A_{\text{SLV}}}{6\pi h^3} A_{\text{pad}} \quad (10.20) \]

where we have assumed that the disjoining pressure is solely from van der Waals interaction [eq. (10.6)]. For example, if \( A_{\text{pad}} = 0.01 \text{mm}^2 \) (0.8% of the area of a pico-size slider), \( h = 1 \text{nm} \), and \( A_{\text{SLV}} = 10^{-19} \text{J} \),

\[
L_{\text{men}} = \frac{10^{-19} \text{J}}{6\pi (10^{-9} \text{m})^3} (10^{-8} \text{m}^2) = 53 \text{mN}. \quad (10.21)
\]

For \( \gamma_L = 25 \text{mN/m} \) (typical for a disk drive lubricant), \( r_m = 4.7 \text{nm} \) when \( h = 1 \text{nm} \), so the pad height \( H \) should be greater than 9.4 nm. If the coefficient of friction is \( \mu = 0.2 \) and the loading force exerted by the slider suspension is \( L = 15 \text{mN} \), then the static friction or stiction force in this example is estimated to be

\[
F_s = \mu (L + L_{\text{men}}) = 0.2 (15 \text{ mN} + 53 \text{ mN}) = 14 \text{ mN}. \quad (10.22)
\]

This stiction force is similar for unpadded sliders on laser-textured disks in the toe dipping regime (Figs. 10.23 and 10.25).

10.3.5 Liquid menisci at high speeds

While the concepts of disjoining and capillary pressures are readily applied for determining the distribution of a liquid lubricant within a stationary contacting interface, this becomes more challenging for sliding surfaces. At moderate speeds, menisci around the contacting asperities no longer maintain their equilibrium shape and adopt a shape determined more by shear forces and viscous flow.

Recent experiments have provided new insights into the shape that lubricant menisci adopt at high speed sliding interfaces and how these non-equilibrium shapes influence friction (Mate et al. 2006). Figure 10.26 illustrates the experimental setup: a small contact pad tilted at a small pitch angle runs in contact with a rotating disk, and the friction force on this pad is measured as a function of the normal contact force. The friction and adhesion forces are found to depend on the direction of the disk rotation, which is attributed to the different shapes of menisci that form around the pad at high speeds, as shown in Fig. 10.26. These menisci are presumed to form from the mechanical action of the pad rubbing against the lubricant film on the disk surface: As the mobile fraction is pushed ahead of the contacting edge of the pad, the lubricant fills up the pad–disk gap for the positive-pitch pad and piles up at the leading edge of the negative-pitch pad. Since the spacing between the positive-pitch pad and the disk at its open
Fig. 10.26. Schematic of experiment done in the author’s lab where sliders are designed to run in light contact at speeds between 4 to 12 m/s, with a disk that rotates either clockwise or counterclockwise (Mate et al. 2006). Contact is only made at a contact pad (30 × 30 µm² in area and 10 nm high) at left end of the slider ABS, which is pitched at a slight angle relative to the disk (∼100 µrad); the pitch angle is defined as positive when the disk moves to the left (clockwise) or negative when the disk moves to the right (counterclockwise). The light gray illustrates how a non-equilibrium meniscus forms around the contact pad due the mechanical action of the pad’s rubbing at high speed against the 1 nm thick lubricant film on the disk.

end is very small (< 4 nm), the meniscus radius $r_m$ underneath this pad is very small (∼1 nm), resulting in a high capillary pressure that leads to a high meniscus adhesive force on the positive-pitch pad. The larger meniscus area sheared by the positive-pitch pad also results in a higher friction force for this pad.

10.4 References


ATOMISTIC ORIGINS OF FRICTION

In Chapter 4, we introduced the macroscopic laws of friction and began to discuss some of the possible microscopic origins of friction. In this chapter we delve deeper into how friction originates at the atomic scale for “dry” or unlubricated surfaces.

When Amontons originally proposed his laws of friction in 1699, many, including Amontons, thought that friction originated from the interlocking of the roughness on opposing surfaces (see Fig. 4.2). By the middle of the twentieth century, however, this roughness mechanism had been ruled out and replaced by the notion that the two main contributors to friction are:

1. Adhesion – Resistance to sliding that comes from the atomic level interaction forces between the atoms on the opposing contacting surfaces
2. Plowing – Resistance during sliding that comes from the work of the asperities of the harder surface plastically deforming the softer surface

The plowing term for friction was discussed in Section 4.2.2, and we will not discuss it further here. In this chapter we focus on the intricacies of adhesion friction—how the interactions between atoms lead to a lateral force when two smooth surfaces slide over each other without wear.

11.1 Simple models for adhesive friction

The notion that friction should be related to adhesion seems fairly natural: the same bonding mechanism that makes it difficult to pull surfaces apart should also make it difficult to slide them over each other. The experimental evidence that friction and adhesion are related goes as far back as Desaguliers, who in 1734 noticed that friction and adhesion forces between metals increase greatly after their surfaces have been polished to a smooth finish (Desaguliers 1734).

One obvious way that adhesion contributes to friction is by increasing the total loading force that pushes the surfaces into contact. If $L_{\text{ext}}$ is the externally applied loading force and $L_{\text{adh}}$ is the adhesive contribution to the loading force, then Amontons’ law for the friction force $F$ becomes

$$F = \mu(L_{\text{ext}} + L_{\text{adh}}).$$

One consequence of eq. (11.1) is that friction can occur even when no external load is applied.
As discussed in Section 4.2.1, based on experiments they conducted from the 1930s through the 1960s, Bowden and Tabor promoted the concept that adhesion also contributes to the coefficient of friction $\mu$ in eq. (11.1) (Bowden and Tabor 1950). They argued that this adhesive component of friction $F_{\text{adh}}$ is proportional to the real area of contact $A_r$:

$$F_{\text{adh}} = A_r s$$  \hspace{1cm} (11.2)

where $s$ is the stress needed to initiate or to maintain sliding—i.e., $s$ is the shear strength of the contact junctions. That friction is proportional to normal force, as expressed by eq. (11.1), then comes about because the real area of contact is proportional to load, which is true for both plastic or elastic deformation of the contacting asperities between rough surfaces (see Section 3.4).

A fundamental feature of Bowden and Tabor’s concept of adhesive friction is that the shear strength $s$ is determined by the interaction forces between the sliding atoms. That adhesion alone is insufficient to cause friction can be seen from the following simple argument: If we consider two solids with ideally flat surfaces, without any atomic structure, and where an attractive interaction pulls them into contact, it then takes no energy to slide one flat surface over another as no work is done against the intermolecular forces. Once the surfaces have atomic structure, however, we have to consider the work to move the atoms of one surface over the atomic structure of the opposing surface.

We start with a very simple atomic level model for adhesive friction (Tabor 1982; Homola et al. 1989; Homola et al. 1990). This model is sometimes referred to as the cobblestone model (Homola et al. 1990), in analogy to pushing a cart over a cobblestone road where a certain amount of lateral force needs to be applied to raise the wheels over the cobblestones to get the cart to move.

To commence sliding, the contacting surfaces must first separate in the direction normal to the sliding surface by a small distance $\Delta D$ (typically a fraction of an angstrom) so that the atoms on the opposing surfaces have enough room to slide over each other. Once sliding starts, the surfaces move laterally a distance of about one atomic dimension $a$ before the atoms find their next minimum energy bonding sites. If the total adhesion force is $L_{\text{adh}}$, the energy required to separate the two surfaces to initiate sliding is $\Delta D \times L_{\text{adh}}$, which is a small fraction $\varepsilon$ of the total adhesion energy $2\gamma A_r$, where $\gamma$ is the material’s surface energy. If we next assume that the contact pressure from the external loading force is negligible compared to the internal pressure generated by the adhesive interactions between atoms (Section 7.3.5) and that all the energy used to separate the surfaces the distance $\Delta D$ and to move them laterally the distance $a$ is dissipated as heat, then

$$a \times F_{\text{adh}} = \Delta D \times L_{\text{adh}} = \varepsilon (2\gamma A_r).$$  \hspace{1cm} (11.3)

This leads to the expression for the shear strength:

$$s = \frac{F_{\text{adh}}}{A_r} = \frac{2\gamma \varepsilon}{a}.$$  \hspace{1cm} (11.4)
For example, if we assume that $\gamma = 25 \text{ mN/m}$ (typical value for a hydrocarbon surface) and that 10% of the surface energy is lost each time the surfaces slide an atomic dimension $a \sim 1 \text{ Å}$, then $s \sim 5 \times 10^7 \text{ N/m}^2$, which compares well with the value measured for shearing a hydrocarbon film between two mica surfaces in the surface force apparatus (SFA) (Homola et al. 1990).

One aspect of shear strength that should be mentioned here is that, in some experiments, the shear strength $s$ is found to increase linearly with the contact pressure $P$:

$$s = s_0 + \alpha P$$  \hspace{1cm} (11.5)  

(Briscoe and Tabor 1978; Briscoe 1981; Briscoe and Smith 1982; Gee et al. 1990; Berman et al. 1998; Israelachvili 2005). A noticeable linear dependence of shear strength with contact pressure is typically observed when the interactions across the sliding interface are relatively weak, leading to the contact pressure from the externally applied load being comparable to the internal pressure from adhesive forces. In these situations, Amontons’ law that friction is proportional to load appears to remain valid even though the real area of contact is constant (Gao et al. 2004; Israelachvili 2005).

Eq. (11.4) would seem to indicate that friction, as determined by the shear strength $s$, should scale with the adhesion energy of the interface as characterized by the interfacial energy $\gamma$. While most interfacial energies do not vary much more than an order of magnitude, the shear strengths for wear-free sliding of crystalline interfaces are found in experiments to vary over a tremendous range—from $10^{-2}$ to $10^{10} \text{ N/m}^2$ (Krim 1995; Krim 1996). This suggests that much more is going on than indicated by the simple cobblestone model. The rest of this chapter is devoted to further unraveling the atomic level origins of shear strength and their impact on friction.

### 11.2 Atomistic models for static friction

When considering how friction originates at the atomic level, it is convenient to separate the problem into two regimes:

1. **Static friction** – The force needed to overcome the potential energy barriers between atoms in order to initiate sliding.

2. **Kinetic friction** – The mechanisms for dissipating energy as atoms slide over each other.

First we examine the modeling of static friction through the simple Frenkel–Kontorova (FK) and Tomlinson models as well as from more advanced molecular dynamic simulations. Later in the chapter, we discuss the origins of kinetic friction by examining the energy dissipation mechanisms of phonon and electronic excitations as measured in experiments of adsorbed monolayers sliding across the surfaces of a quartz crystal microbalance (QCM). At the end of the chapter, we examine wind induced shear flow of thin liquid films in order to understand better the relationship between molecular friction and lubrication.
11.2.1 Frenkel–Kontorova model

First introduced to describe the physics of dislocations in solids, the Frenkel–Kontorova (FK) model (Frenkel and Kontorova 1938; Kontorova and Frenkel 1939) is now widely used to describe many physical phenomena, including the motion of adsorbed monolayers on surfaces and the friction between two solids (Braun and Kivshar 2003; Hirano 2006). As illustrated in Fig. 11.1, in its simplest form, the FK model consists of a one-dimensional chain of atoms placed in a periodic potential. The atoms on the chain interact with their neighbors via harmonic potentials represented by springs, and a lateral force $F$ pulls the chain along the periodic potential. In an FK model of friction, the atoms on the chain represent the atoms of the slider surface or of an adsorbed layer sliding over a substrate surface; as they slide, these atoms feel a friction force from the gradient of the periodic potential of the atoms of the substrate surface. The spring potentials between atoms on the chain then correspond to the interactions between the atoms within the adsorbed layer or the body of a slider. Here we discuss the one-dimensional case, but the FK model can be straightforwardly extended to a more realistic two-dimensional simulation of actual friction situations (McClelland 1989; Hirano 2006).

Fig. 11.1(a) illustrates the special case where the number of atoms on the chain equals the number of minima of the potential. When no lateral force is applied, the equilibrium configuration is for all the atoms to reside at the bottoms of the potential wells. Now, as the end of the chain is pulled to one side, the atoms move up the side of the potential well; when they come up to the top of potential peak, they all slide in unison down into the next potential well. This can be thought of as a tug-of-war between the atoms on the chain and the external lateral force of the periodic potential. This special case of equal density of atoms and minima corresponds to the maximum lateral force needed to initiate sliding, which equals the force to overcome each individual potential well times the number of

![Diagram](attachment:fig11.1.png)

**Fig. 11.1.** Schematic of the one-dimensional Frenkel–Kontorova (FK) model. Atoms interact with each other via overdamped springs while experiencing a periodic potential from the substrate. The chain is pulled to the right by a lateral force $F$. (a) Commensurate structure with one atom per potential minimum. (b) Structure with four atoms per three minimum.
atoms on the chain. This type of friction is called “atomistic locking” (Hirano 2003).

If we add an extra atom to the chain, this creates a kink on the chain, as the new minimum energy configuration has a local compression somewhere along the chain to accommodate the extra atom. The importance of kinks is that the activation barrier for moving a kink is always much smaller than the amplitude of the substrate potential, so kinks can be moved along the chain much more easily than moving all the atoms in unison. The higher the kink concentration is, the higher the system mobility will be. For finite chain lengths, kinks can be generated at one end, propagate down the length of the chain, and disappear at the other end. By this model, the onset of sliding of a finite contact area can be initiated by local compression at the edges of the contact zone creating interface dislocations that then propagate across the contact zone, in a manner similar to dislocations being responsible the ductility of materials (Section 3.1).

An important concept within the FK approach is incommensurability, which can result in frictionless sliding. For an infinitely long chain, the system is said to be incommensurate if the ratio of the natural period of the chain to the substrate period is an irrational number. An incommensurate chain becomes “free” of the substrate potential if the chain is stiff enough compared to the magnitude of the periodic potential. For these situations, the loss and gain of each atom’s interaction energy cancel each other out, and, for an infinite system, the total sum of the forces acting on all the atoms is strictly zero: any applied lateral force moves the chain. The simple explanation for this is that for every atom going up a side of a potential barrier, another atom somewhere along the chain is going down, and these two processes exactly cancel each other. This frictionless condition applies only to the force needed to initiate motion of the chain or static friction; to achieve totally frictionless motion, the viscous energy dissipation mechanisms need to be negligible.

Even for commensurate systems—such as the system illustrated in Fig. 11.1(b), which has four atoms for every three potential minima—the force needed to initiate sliding is greatly reduced, as the atoms within the repeating structure are out of phase with each other as they move along the potential curve.

While frictionless sliding is expected within the FK model for an incommensurate interface where the interactions across the sliding interface are much weaker than the interactions between atoms within the solids, the frictionless behavior disappears, and the chain becomes locked to the substrate, if the chain’s elastic constant $g$ is less than a threshold value called $g_{\text{Aubry}}$, at which point the interactions across the sliding interface become comparable to those in the bulk. This locking occurs due the creation of local regions of common periodicity along the otherwise incommensurate chain. Similarly, the frictionless behavior also disappears if the chain has finite length, as the ends of the chains are always pinned, though friction is much reduced due to atoms being incommensurate between the ends.
Experimental realizations of ultra-low friction in incommensurate sliding systems

One way of verifying FK models of friction is to compare the values of friction predicted from the model with those measured by experiment. While this has been done for the motion of adsorbed layers (Braun and Naumovets 2006) and the friction of metals (Hirano 2006), good agreement between these models and experiment still remains elusive.

The most exciting aspect of the FK analysis to be confirmed in experiments is the prediction of ultra-low friction for incommensurate sliding systems (Sokoloff 1984; McClelland 1989). This phenomenon of ultra-low friction has been named superlubricity by Hirano et al., who first observed reduced friction as the sliding surfaces become incommensurate (Hirano et al. 1991); in their experiments, contacting crystals of muscovite mica were rotated away from commensurability in a surface force apparatus. The term superlubricity is something of a misnomer, because the resistance to motion is never completely absent in experiments, as it is for other “super” effects in physics, such as superconductivity or superfluidity; the name, however, has now stuck for describing ultra-low friction situations. Since the results of Hirano et al., ultra-low friction has also been observed in other incommensurate sliding systems consisting of atomically clean surfaces with a well defined periodic structure (Luthi et al. 1994; Overney et al. 1994; Hirano et al. 1997; Falvo et al. 1999; Ko and Gellman 2000; Dienwiebel et al. 2004).

A particularly striking example of superlubricity is shown in Fig. 11.2. In these experiments (Dienwiebel et al. 2004), an AFM tip is scanned over the basal plane of a graphite crystal. It is well known that friction is very low for AFM tips rubbing against graphite (Mate et al. 1987), and this is now attributed to a small graphite flake being picked up at the end of the AFM tip as illustrated in Fig. 11.3. The small flake leads to the low friction for two reasons:

1. Weak adhesion occurs between the contacting planes of graphite atoms, while the atom lattice within the graphite planes is fairly stiff.
2. The random angle between the lattice angles of these planes means that generally the surfaces will be incommensurate.

In the experiment by Dienwiebel et al. shown in Fig. 11.2, the graphite sample was rotated underneath the AFM tip and the attached flake. Every 60 degrees the misfit angle between the lattices of the graphite planes of the flake and the sample becomes zero, and the friction rises sharply when these two surface structures come into commensuration, that is where all the graphite atoms on the tip flake lock with the potential minima of the (0001) surface of the graphite substrate. Rotating the sample a few degrees away from a zero misfit angle is all that is needed to make the flake’s atomic structure incommensurate with the graphite substrate, drastically reducing the friction down to near the detection limit of the AFM.
11.2.2 Tomlinson model

A somewhat simpler and earlier model for friction was developed by Prandtl (Prandtl 1928) and Tomlinson (Tomlinson 1929) and is usually referred to as the Tomlinson model or the independent oscillator model. A one-dimensional Tomlinson model is illustrated in Fig. 11.4(a). As with the FK model, the slider atoms feel the periodic potential of the substrate surface atoms as they slide over them, the instantaneous friction force on each atom is the gradient of the periodic potential, and the net friction force is the sum of these individual forces. Unlike the FK model, however, the slider atoms are independently connected by springs to the rigid slider.

Initially, the system is locked in a configuration of minimum potential energy. When the slider starts to move, the system exhibits static friction, until sufficient elastic stress is built up in the springs to overcome the potential barriers, at which point the individual atoms pop rapidly to their next stable state. During these pops or slips, the stored elastic energy is converted into kinetic energy, which is dissipated as heat (this dissipation process is incorporated into the model as the damping constant for the spring-atom oscillators).

If the potential barrier height is lowered to below the threshold for slips to occur, the atoms move smoothly through the potential. If, in this smooth sliding case, these atoms are also incommensurate with the substrate, the lateral forces acting on the individual atoms cancel each other leading to
negligible net friction, analogous to the frictionless sliding that occurs in the FK model.

While the Tomlinson model has been successfully applied to many situations [for a list see the review by Robbins and Müser (Robbins and Müser 2000)], it can often be unclear whether the FK or Tomlinson model is the better choice for describing static friction processes at the atomic scale. For describing the friction between three-dimensional solids, both models are likely to be equally inadequate: the FK model neglects the interactions of the interfacial layer (represented by the chain) with the rest of the slider body; the Tomlinson model includes this interaction in a simplified way, but neglects the interactions between the atoms at the slider surface.

11.2.2.1 Example: An AFM tip sliding across an NaCl crystal at ultra-low loads
A situation where the Tomlinson model can work well for modeling friction is for an AFM tip scanning across a sample surface as illustrated in Fig. 11.4(b). A nice
Figs. 11.4. (a) One-dimensional Tomlinson model where atoms, connected to a slider by overdamped springs, experience a periodic potential as they slide across a substrate surface. (b) An AFM configured to measure the friction force acting on a single atom at the apex of the tip, which is a practical realization of the Tomlinson model.

practical realization of the Tomlinson model in an AFM experiment was done by Socoliuc et al. where an AFM tip was scanned over a NaCl surface in ultra-high vacuum at sufficiently small loads that only a few atoms at the end of the tip were in contact (Socoliuc et al. 2004). In this case, having only a few atoms in contact is important for obtaining a periodic substrate potential, since the tip atoms are incoherent with the sample surface atom and having a large number of atoms in contact would have averaged out the period potentials experienced by the individual atoms.

Figs. 11.5(a)–(c) show the lateral force acting on the tip as it is scanned back and forth across the NaCl surface at different loads. At the highest load, in Fig. 11.5(a) (but still low enough at 4.7 nN that only a few atoms touch), the trace of the lateral force exhibits a sawtooth structure with the periodicity of the NaCl crystal surface, characteristic of the atomic scale stick–slip motion discussed in Section 4.3.1.3. The slip motion occurs when the gradient of the lateral periodic force exceeds the effective lateral spring constant of the AFM; at this point, the tip suddenly slips to the next potential minimum, where it sticks until enough elastic stress is built up for the next slip. When the tip is scanned in the opposite direction, the slips occur in the other direction, and a hysteresis or
Fig. 11.5. (a)–(c) Measurement of the lateral force acting on an AFM tip as it scans back and forth over a (001) surface of a NaCl crystal. The applied loading force is (a) $F_N = 4.7 \, \text{nN}$, (b) $F_N = 3.3 \, \text{nN}$, and (c) $F_N = -0.47 \, \text{nN}$. For (c) the negative value of $F_N$ indicates that a slight adhesive force holds the tip in contact as it slides over the NaCl surface. (d)–(f) Results from the Tomlinson model that correspond well with the experimental results. This model uses an effective lateral spring constant $k = 1 \, \text{N/m}$, a lattice periodicity $a = 0.5 \, \text{nm}$, and $\lambda = 5$ for (d), 3 for (e), and 1 for (f). Reprinted with permission from Socoliuc et al. (2004). Copyright 2004, by the American Physical Society.
friction loop is observed. The area of the friction loop equals the energy dissipated by friction during the back-and-forth-scan; this energy dissipation occurs during the slip portions of the scan where the tip’s kinetic energy is dissipated.

When the load is decreased to 3.3 nN [Fig. 11.5(b)], the amplitude of the periodic lateral force decreases, leading to a decrease in the magnitude of the static friction prior to the atomic scale slip and a decrease the energy dissipated in the friction loop. When the loading force is reduced below a threshold value [Fig. 11.5(c)], the slips and the hysteresis disappear, and the forward and backward scans overlay each other perfectly (energy is no longer dissipated during the scan). The disappearance of stick–slip motion at low loading force indicates that the amplitude of the lateral periodic friction force becomes sufficiently small at low loads that its gradient no longer exceeds the effective lateral spring constant of the AFM. Since no slips occur, no energy is dissipated during a back and forth scan, and the average friction is zero. This is another form of superlubricity, which can occur for atomic scale contacts at ultra-low loads.

This transition from stick–slip motion to smooth modulation for AFM friction can be explained in straightforward manner with the one-dimensional Tomlinson model. We assume that the tip experiences a sinusoidal potential for the lateral force with amplitude $E_0$ and period $a$ [Fig. 11.4(b)] and that $x_{\text{tip}}$ and $x_{\text{base}}$ are the lateral motion of the tip atom and the base of the AFM cantilever, respectively. Then the tip’s minimum energy configuration is described by the minimum of the effective potential

$$V = -\frac{E_0}{2} \cos \left( \frac{2\pi x_{\text{tip}}}{a} \right) + \frac{1}{2} k (x_{\text{tip}} - x_{\text{base}})^2$$  \hspace{1cm} (11.6)

where $k$ is the effective spring constant between the tip atom and the AFM base, and represents the combination of the cantilever spring constant and the lateral stiffness of the contact.

Whether the motion is continuous or a series of stick–slips is described by the parameter

$$\lambda = \frac{2\pi^2 E_0}{ka^2}.$$  \hspace{1cm} (11.7)

Continuous motion occurs when $\lambda < 1$, while stick–slip motion occurs when $\lambda > 1$. Figs. 11.5(d)–(f) show the numerical results for the motion for three different values of $\lambda$ that reproduce well the experimental results of Figs. 11.5(a)–(c).

This Tomlinson model further indicates that the effective lateral spring constant for this experiment is on the order of 1 N/m. This is a factor of fifty less than the lateral spring constant of the AFM cantilever, indicating that the effective stiffness is dominated by the lateral stiffness of the contact, which, in turn, implies that the deformability of the tip apex and the surface of the contact zone has a lateral stiffness on the order 1 N/m.
11.2.3 Molecular dynamic simulations

A reasonable question to ask is: How well do the general conclusions of the one-dimensional FK and Tomlinson models carry over to more realistic three-dimensional models? Recently, molecular dynamic (MD) simulations have systematically studied how friction depends on the atomic structure of the sliding materials and the interaction potentials between the atoms. For bare surfaces sliding against each other, the basic concepts of the FK and Tomlinson models are found to remain valid in MD simulations (Müser and Robbins 2000; Hirano 2006). In particular, commensurate sliding surfaces are always found to be “pinned”, meaning that a threshold lateral force has to be applied to overcome the static friction associated with the pinned atoms. Incommensurate sliding surfaces in these simulations are unpinned and exhibit frictionless sliding, unless one of the surfaces becomes so deformable that the atoms rearrange themselves to accommodate the atoms on the opposing surface; in the FK model, this rearrangement is equivalent to the system being unpinned when the elastic constant \( g > g_{\text{Aubry}} \) and pinned when \( g < g_{\text{Aubry}} \).

11.2.4 Example: Cold welding

A practical example of where atoms at an initially incommensurate interface rearrange themselves during contact to generate high friction and adhesion forces is the process of cold welding that occurs when clean metal surfaces are brought together in high vacuum. Since metal surfaces free of any adsorbed contaminants (which can be achieved in high vacuum) have very high surface energies (100 to 1000 mN/m), very strong forces are generated on any atoms that come into contact with them. When one metal contacts another, these high forces drive the diffusion of atoms at that interface to maximize the metal–metal contact and minimize the area of the metal exposed to the vacuum. With this rearrangement of atoms, patches of the interface become commensurate, resulting in high friction and adhesion forces on the order of the bulk yield shear strength. This results in high friction coefficients (\( \mu > 4 \)) and adhesion for clean metals in ultra-high vacuum (Bowden and Rowe 1956; Pashley et al. 1984; Gellman and Ko 2001). Generally, during sliding of these surfaces, fracture occurs within the bulk and microscopic bits of metals are pulled out of one surface and adhere to the other, a process known as adhesive wear.

11.2.5 Why static friction occurs in real-life situations

In molecular dynamic simulations of bare sliding surfaces, it is fairly easy to come up with situations where frictionless sliding occurs (just as for the FK and Tomlinson models). Practical realization of frictionless behavior, however, has only been observed in experiments carefully constructed for this purpose. One might well wonder why, in practice, static friction is so universal, since, at most sliding interfaces, the sliding atoms are incommensurate with those of the opposing surface and the surface forces are usually not strong enough to pull the atoms on opposing surfaces into commensurability.
MD simulations from Mark Robbins’ group at John Hopkins University indicate that this dilemma can be resolved by considering the presence of “third bodies” between the sliding surfaces in the form of adsorbed molecules or atoms (He et al. 1999; Müser and Robbins 2000; He and Robbins 2001; Müser et al. 2001). As discussed many times in this book, surfaces outside of ultra-high vacuum invariably adsorb some contaminants from the surrounding environment—hydrocarbons, water vapor, etc.—and sometimes a molecular layer is deliberately applied in the form of a lubricant film. MD simulations indicate that sandwiching this layer of molecules between the sliding surfaces provides another mechanism for static friction between incommensurate surfaces: While the interactions between the contacting solid surfaces may be too small to rearrange their own atoms into commensuration, these forces are sufficient to rearrange the atoms or molecules in the intervening layer. So, during sliding, these third bodies diffuse within the sliding interface to find the local energy minima between the opposing surface atoms. This idea is illustrated in Fig. 11.6. Once in these local minima positions, the molecules become locked in and resist lateral motion. In the simulations, this locking of the adsorbates generates static friction and kinetic friction between both incommensurate and commensurate sliding interfaces.

It should also be noted that the rearrangement of atoms and molecules at a contacting interface—either those in the adsorbed layer between the contacting surfaces or those in the deformable bulk—also produce the phenomenon of adhesion hysteresis discussed in Section 5.4.3; where the energy gain when the two surfaces are brought into contact is less than the energy required to separate them. This would suggest that a correlation should exist between friction

![Fig. 11.6. Schematic of two incommensurate surfaces separated by a few mobile atoms. The larger the gap, the smaller the energy penalty for occupation by a mobile atom. The black circles indicate the most energetically favorable positions for atoms to fit between the two surfaces, while the dashed circle indicates a less favorable site. Reprinted with permission from Müser et al. (2001). Copyright 2001, by the American Physical Society.](image-url)
and adhesion hysteresis, which indeed has been found in surface force apparatus experiments (Chen et al. 1991; Yoshizawa et al. 1993).

11.3 Atomic origins of kinetic friction

11.3.1 Sliding isolated molecules and monolayers across surfaces

A good starting point for considering how kinetic friction originates at the atomic level is to discuss what happens when individual atomic or molecular adsorbates slide across a surface.

Fig. 11.7 illustrates an idealized situation where a few isolated polymer molecules are sitting on a solid surface that is smooth in the sense that the activation barrier to diffuse from one adsorption site to another is small compared with $k_B T$. On such a surface, a negligible threshold force is needed to initiate sliding, and we can assume that the force $F$ opposing the motion of these molecules across the surface is purely viscous in nature, and so can be described by the relation

$$F = m \eta_m v$$

where $m$ is the molecular mass, $v$ is the sliding velocity, and $\eta_m$ represents the viscosity of the sliding interface.

For most cases, this viscous damping of the atoms and molecules as they move across a surface primarily comes from sliding-induced atomic vibrations (phonon excitations) that are ultimately transformed into heat. The phononic contribution to $\eta_m$ has been studied analytically and with molecular dynamic simulations (Cieplak et al. 1994; Tomassone et al. 1997; Persson 1998; Liebsch et al. 1999; Torres et al. 2006). In analytical models for adsorbed noble gases

![Fig. 11.7. Schematic of a few polymer molecules moving across a surface due to an applied air stress $\tau$. The top surface of the cylinders around each polymer molecule represent the effective area $A_m$ over which the air shear acts to generate a force $F = A_m \tau$ that drives the molecule across the surface with an average velocity $v = \chi F$.](image-url)
sliding over surfaces, a sinusoidal potential with a corrugation $U_0$ is typically used to model the adatom–substrate interaction. Then $\eta_m$ can be expressed as

$$\eta_m = \eta_{\text{subs}} + cU_0^2.$$  

(11.9)

The first term $\eta_{\text{subs}}$ in eq. (11.9) is the energy dissipation from non-phonon excitations, such as electronic excitations, while the second term is the dissipation from phonon excitations where $c$ is a constant depending on temperature and lattice spacing. Good agreement has been found between numerical simulations and experiments for phonon dissipation of noble gas atoms sliding on metal surfaces (Tomassone et al. 1997; Coffey and Krim 2005).

For sliding over metal surfaces, friction dissipation due to electronic excitation is expected to be a major contributor (Persson and Nitzan 1996; Persson 1998; Liebsch et al. 1999; Bruch 2000), but this has been more difficult to model and to measure in experiments. Recently, Highland and Krim have observed a reduced value of $\eta_m$ when the substrate enters a superconducting state, consistent with an electronic dissipation mechanism related to adsorbate polarizability (Highland and Krim 2006).

At sparse coverages, the movement of molecules across this surface can be described by the Einstein model for diffusion of small particles through a viscous medium, where the average velocity for diffusion is given by

$$v = \chi F$$  

(11.10)

where $F$ is the average force acting on the molecule driving its diffusion across the surface and $\chi$ is the molecule’s mobility, which is related to the diffusion coefficient $D$ by the Einstein equation

$$D = \chi k_B T$$

$$\chi = \frac{k_B T}{D}$$  

(11.11)

By comparing eqs. (11.8) and (11.10), we see that

$$\chi = \frac{1}{m\eta_m}.$$  

(11.12)

When the driving force $F$ comes from an air shear stress $\tau$, the average driving force is

$$F = \tau A_m$$  

(11.13)

where $A_m$ is the cross-sectional area of the molecule parallel to the surface, represented by the top of the cylinders in Fig. 11.7. As an isolated molecule may not present a well defined cross section to the impinging air molecules that are generating the air shear stress, it is probably more appropriate to think of $A_m$ as an effective area defined by $A_m = F/\tau$. 
11.3.2 Quartz crystal microbalance

Since the pioneering experiment of Krim and Widom in 1988, where a krypton monolayer was slid over a gold substrate (Krim and Widom 1988), the quartz crystal microbalance (QCM) has become the main technique for measuring friction dissipation for adsorbed layers sliding over surfaces. As illustrated in Fig. 11.8, a QCM consists of a thin single crystal of quartz with metal electrodes deposited in its top and bottom surfaces. Quartz crystals are used as they have very little internal dissipation or friction, enabling them to oscillate with a very sharp resonance frequency, typically between 5 to 10 MHz, driven by a voltage applied to the electrodes. When a thin film of atoms or molecules is adsorbed on the electrode surfaces, the added mass causes the resonance frequency of the crystal to shift to a lower frequency. Due to the very sharp resonance, very small shifts in the resonance frequency, and hence a very small adsorbed mass, can be readily detected, enabling the coverage of a minute fraction of a monolayer to be measured.

Since the resonance motion is a transverse shear oscillation of the quartz crystal, the electrode surfaces move laterally as illustrated in Fig. 11.8. If the adsorbed layer cannot keep up with this motion, it slides over the surface with a relative velocity $v_{\text{rel}}$ and dissipates its kinetic energy via friction, which broadens the resonance. QCM measurements of interfacial friction are limited to sliding

Fig. 11.8. (Bottom) Schematic of the quartz crystal microbalance (QCM) used for measuring the friction dissipation of adsorbed layers sliding across surfaces. (Top) An enlarged cross-sectional view of the center of the QCM showing the motion of the resonance oscillation of the quartz crystal as well as a submonolayer coverage of adsorbed noble gas atoms on the metal electrode surfaces.
systems with very low friction such as absorbed rare gases or small, physisorbed molecules that can slide easily over the electrode surface. For adsorbed layers with higher friction or chemisorbed layers that are more strongly bonded to the substrate, the slippage tends to be too small to generate a measurable broadening of the QCM resonance. Typically, noble metals are used for the electrode surface, as these have a lower bonding energy of the adsorbate layer to the metal surface. These metals are deposited in such a manner that the exposed surface is predominately a smooth, close packed crystalline surface, which further promotes ease of sliding.

Generally, for those weakly bound atoms and molecules that slide easily on a QCM surface, no threshold force is needed to induce sliding (i.e., no static friction). The adsorbed layers experience only viscous friction during sliding, described by eq. (11.8), which can also be described by a slip time

\[ t_{\text{slip}} = \frac{1}{\eta_m} \]  

(11.14)

that is inversely proportional to the frictional damping. This slip time \( t_{\text{slip}} \) corresponds to the time for the velocity of the adsorbed layer to decay by \( 1/e \) after the driving force \( F \) is removed.

11.3.2.1 Example: Xe on Ag(111)

Fig. 11.9 shows the slip time for xenon on a Ag(111) surface as a function of coverage (Daly and Krim 1996). In these experiments, the xenon films slide a distance of 2 nm, which is about one tenth the peak-to-peak amplitude of the lateral motion of the crystal surface, and have a peak sliding velocity on the order of 1 cm/s. The data in Fig. 11.9 have several interesting features:

- As the xenon coverage increases from zero to the completion of an incommensurate monolayer at 0.9 monolayer, the slip time drops from 2.1 ns to 0.8 ns, indicating increasing interfacial friction as the atoms become more closely packed.
- Maxima occur in the slip times when a sparse two-dimensional gas phase exists on either the silver surface or xenon monolayer surfaces (i.e., the slip time maxima occur at coverages of \(< 0.3, 1.2, \) and \( 2.05 \) monolayers).
- At 1 monolayer coverage, the xenon atoms have the same spacing as in bulk xenon and slide over the Ag(111) surface as a single solid layer with a shear strength given by

\[ s = \frac{\rho v_{\text{rel}}}{t_{\text{slip}}} \]  

(11.15)

where \( \rho \) is the mass per unit area of the xenon film. From the results in Fig. 11.9 Daly and Krim deduce that the shear strength for sliding a one-atom-thick monolayer is 11.9 N/m\(^2\), while for a two-atom-thick bilayer it is 15.1 N/m\(^2\) or about 25% higher.
11.3.3 Movement of a liquid film on a surface with the blow-off technique

Fig. 11.9 showed an example where an adsorbed film behaves as a two-dimensional gas at low coverages and as a solid layer at a close packed monolayer. Another important type of adsorbed layer is a molecularly thin liquid film. As mentioned in Chapter 10, these thin liquid films are often used as lubrication layers on surfaces. Here, we discuss how the blow-off technique can be used to elucidate the molecular origins of friction and lubrication associated with these thin films.

The blow-off was originally developed by Deryaguin in the 1940s (Deryaguin et al. 1944; Deryaguin et al. 1992). Recently, Mate and co-workers have updated the blow-off technique with modern instrumentation, greatly improving its sensitivity to submonolayer coverages (Mate and Marchon 2000; Scarpulla et al. 2003; Mate 2004). The blow-off technique is an alternative technique to the QCM for studying the movement of larger molecules on surfaces that would be incapable of causing a measurable broadening of the quartz crystal resonance.
Fig. 11.10. The blow-off technique as implemented by Mate and co-workers. An air shear stress $\tau$ is generated by blowing $N_2$ through a narrow channel, and the stress is calculated from $\tau = \Delta Pd/2L$. The air shear stress drives the flow of a liquid film on a sample that serves as the bottom surface of the channel. The profiles of the liquid films before and after the application of the air shear stress are measured ex situ by ellipsometry. Reprinted with permission from Scarpulla et al. (2003). Copyright 2003, American Institute of Physics.

Fig. 11.10 shows a schematic of the updated blow-off technique (Mate and Marchon 2000; Scarpulla et al. 2003; Mate 2004). In this technique, an air shear stress $\tau$ is generated on a sample surface by flowing nitrogen through a narrow channel. The sample, which is partially coated with a film of non-volatile liquid, serves as the bottom surface of the channel. The shear response of this liquid film is determined by measuring the movement of the film’s step edge downwind due to the applied air shear stress. Because of the low Knudsen number of the gas flow through the channel, slippage at the gas–surface interface is negligible, and the air shear stress is given by

$$\tau = \frac{\Delta P d}{2L}$$

(11.16)

where $\Delta P$ is the pressure drop along the channel.

11.3.3.1 Example: Wind-driven flow of perfluoropolyether lubricants on silicon wafers

Fig. 11.11 shows an example of the thickness profiles of a liquid film on a silicon wafer being sheared by an applied air stress for different periods of time. This particular liquid is a perfluoropolyether (PFPE) lubricant called Demnum-SA (manufactured by Daikin), which has an alcohol group at one end of the perfluoropolyether chain. The alcohol group hydrogen bonds to the silicon oxide surface of the silicon wafer, providing an anchoring effect to the movement of those Demnum-SA molecules next to the silicon oxide surface. Consequently,
for film thicknesses less than one monolayer (which in this case is defined to be twice the radius of gyration $R_g$), the adsorbed molecules move very slowly when subjected to the air shear stress, while molecules in the second monolayer slide rapidly over this first monolayer. This structure of the first and second monolayers of Demnum-SA on silicon is shown schematically in Fig. 11.12(a).

The thickness profiles from the blow-off technique can be analyzed within a continuum model for liquid flow (shown schematically in Fig. 11.13) to determine the shear response of the film as a function of its thickness. For this analysis, it is convenient to first normalize the profile results with respect both to time $t$ and shear stress $\tau$ by dividing the distance $\Delta x$ between profiles at different film thicknesses $h$ by the stress–time product $\tau \cdot t$.

$$\chi_S(h) = \frac{\Delta x(h)}{\tau t}. \quad (11.17)$$

The normalized result $\chi_S$ is called the shear mobility (Scarpulla et al. 2003).

Fig. 11.14 shows how the shear mobility $\chi_S$ varies with film thickness $h$ for the Demnum-SA film profiles in Fig. 11.11. Also shown in Fig. 11.14 is the shear mobility for the perfluoropolyether liquid Demnum-S20, on a bare silicon wafer and on a silicon wafer with a 7 Å thick grafted layer of Demnum-SA.
Demnum-S20 has the same perfluoropolyether backbone structure and molecular weight as Demnum-SA, but has only neutral –CF$_3$ end-groups, and so moves much more easily across the surface without the anchoring effect of the alcohol end-group present on the Demnum-SA molecule.
The plots of $\chi_S$ versus $h$ in Fig. 11.14 can be interpreted within the continuum analysis of liquid film flow in terms of an effective viscosity that is a function of the distance $z$ away from the sample surface. According to eq. (9.11), when a liquid film is subjected to a shear stress $\tau$, momentum is transferred across the film according to

$$\tau = \eta(z) \frac{dv(z)}{dz} \quad (11.18)$$

where $v(z)$ is the velocity of the liquid in the direction of the applied stress at a distance $z$ from the solid–liquid interface and $\eta(z)$ is the dynamic viscosity at $z$ (see Fig. 11.13). The effective viscosity $\eta_{eff}$ of the liquid near the air–liquid interface is determined by setting $z = h$ in eq. (11.18). Using eq. (11.17) and the fact that the velocity of the film’s air–liquid interface is $v(h) = x(h)/t$, we obtain

$$\eta_{eff} = \left[ \frac{d\chi_S}{dh} \right]^{-1} \quad (11.19)$$

Therefore, the inverse slope of a $\chi_S$ vs. $h$ plot can be interpreted as the effective viscosity, so long as the changes in the thickness profiles occur mainly from shear flow. One of the experimental difficulties of the blow-off technique is that the thickness profiles can also change due to diffusion and evaporation; so these experiments need to be implemented in a way that keeps these non-shear-induced
In Fig. 11.14, the effective viscosities are labeled for the different perfluoropolyether films. For Demnum-SA less than one monolayer (2$R_g$) thick, the effective viscosity is very high, $\sim 1000$ mPa·sec, due to the anchoring effect of the alcohol end-group; this is almost an order of magnitude higher than the bulk viscosity of Demnum-SA (140 mPa·sec). For the fast second layer of Demnum-SA, the effective viscosity is 28 mPa·sec, much lower than the bulk viscosity.

For Demnum-S20 on a bare silicon wafer, the shear mobility increases linearly with film thickness when the film thickness exceeds 6 Å, with a slope corresponding to an effective viscosity of 90 mPa·sec, the same as the bulk viscosity of Demnum-S20. So, when this perfluoropolyether polymer has neutral end-groups, it undergoes a shear flow across this surface with a no-slip boundary condition at the liquid–solid interface and with the bulk viscosity, when the coverage exceeds one layer of molecules lying flat on the surface (the polymer chain diameter equals 6 Å). At first, it may seem somewhat surprising that a liquid film would continue to have the bulk viscosity at a film thickness less than the polymer’s diameter of gyration, since all the molecules are partially in contact with the solid surface. As discussed in Section 10.1, a dramatic enhancement of viscosity occurs in surface force apparatus (SFA) experiments at such a thin film thickness. In an SFA experiment, however, the enhanced viscosity comes about from the movement of the liquid molecules being severely restricted when confined in a narrow gap between two hard-wall surfaces. For a liquid film sitting on just one solid surface, the molecules have a much easier time jostling each other as they undergo shear flow and can momentarily displace each other slightly above the liquid–air interface as they flow across the surface, enabling them to maintain the bulk viscosity.

For Demnum-S20 on the grafted layer of Demnum-SA, the shear mobility shown in Fig. 11.14, when the film thickness is less than the molecule’s diameter of gyration, is actually higher than found for the Demnum-S20 on the bare silicon wafer. This implies that slippage occurs for Demnum-S20 molecules in contact with grafted layer when a shear stress is applied. In essence, the grafted perfluoropolyether polymers provide a low surface energy surface that the unattached perfluoropolyether polymers can slide over more easily [Fig. 11.12(b)].

When the film thickness of these perfluoropolyether polymers is less than the 6 Å diameter of the polymer backbone, the surface coverage is incomplete even if the polymers lie flat on the solid surface as illustrated in Fig. 11.12(c). At these sparse coverages, the film movement can no longer be analyzed in terms of liquid film flow as depicted in Fig. 11.13; a better approach is to analyze the movement in terms of wind driven diffusion.

As illustrated in Fig. 11.7 and expressed by eq. (11.10), the average diffusion viscosity at sparse coverages is given by $\nu = \chi F$, where $\chi$ is the Einstein mobility factor related to the diffusion coefficient by eq. (11.11). Combining eq. (11.17) for the shear mobility $\chi_S$ with this expression for $\nu$ links the shear mobility to
molecular diffusion as expressed by $\chi$:

$$\chi_S = \chi A_m. \quad (11.20)$$

If the adsorbed molecules do not interact with each other, $\chi$ and $A_m$ should be constant with coverage, resulting in $\chi_S$ being independent of coverage at a sub-monolayer film thickness. An example of constant $\chi_S$ occurs in Fig. 11.14 for film thicknesses less than 6 Å of Demnum-S20 on bare silicon where $\chi_S = 4 \text{ nm/Pa} \cdot \text{sec}$. Approximating Demnum-S20’s molecular cross-sectional area $A_m$ to air shear as

$$A_m = \text{(polymer chain length)} \times \text{(chain diameter)}$$

$$= (7.5 \text{ nm}) \times (0.6 \text{ nm})$$

$$= 4.5 \text{ nm}^2 \quad (11.21)$$

we obtain

$$\chi = \frac{\chi_S}{A_m}$$

$$= \frac{(4 \text{ nm/Pa} \cdot \text{sec})}{(4.5 \text{ nm}^2)}$$

$$= 0.9 \times 10^{12} \text{ sec/kg} \quad (11.22)$$

which, using eqs. (11.12) and (11.14), corresponds to a slip time of 4 ps. This slip time should be compared to slip times on the order of nanoseconds typically obtained in QCM experiments for sliding noble gas atoms across surfaces (as in Fig. 11.9). Not surprisingly, the viscous friction, as expressed by the inverse slip time, for sliding a polymer like Demnum-S20 is orders of magnitude higher than for the much smaller noble gas atoms, which are weakly bonded to the QCM electrode surface.

11.3.4 Pinning of an absorbed layer

Previously, when discussing adsorbates sliding across a surface as illustrated in Fig. 11.7, we assumed that the surface was smooth and that the adsorbates experience only viscous friction as they slide. In reality, the atomic nature of surfaces means that they can never be perfectly smooth, as atomic corrugation always exists on surfaces. Fig. 11.15(a) illustrates the variation in an adsorbed atom’s binding energy as it travels across a surface. For physisorbed atoms, the maximum gradient of the energy barrier $E_B$ between adsorption sites typically corresponds to a lateral force on the order of $10^{-3} \text{ eV/Å}$; in a typical QCM experiment, however, the force of inertia $F$ is on the order $10^{-9} \text{ eV/Å}$, many orders of magnitude less than what is needed to move a physisorbed atom directly over these barriers, while the air shear force in a blow-off experiment technique is many orders of magnitude less than the inertia force in a QCM experiment. Instead, the external force $F$ lowers the height of the diffusion barrier slightly, making it more likely for adsorbates to diffuse in the direction of $F$. The sliding velocity $v$ then corresponds to the average drift velocity of the adsorbate layer,
which is proportional to $F$. Since this diffusion is a thermally activated process, the drift velocity displays the temperature dependence

$$v \propto \exp\left(\frac{-E_B}{k_B T}\right).$$

(11.23)

So, for a sufficiently large energy barrier $E_B$ or a sufficiently low temperature $T$, $v \sim 0$, and the adsorbates are effectively pinned to the surface.

Let’s consider the example of a QCM experiment done by Bruschi et al. with krypton on gold (Bruschi et al. 2002). At a sufficiently low coverage, all the adsorbed krypton atoms are pinned at high bonding energy defect sites. When this surface was vibrated in the QCM experiment with a low vibration amplitude, the inertia force was insufficient to overcome the energy barriers around these defect sites. When the vibration amplitude was increased to above 3.4 Å, this generated a sufficient inertia force to eject krypton atoms from a few of these defect sites, and these loose atoms knocked the remaining pinned atoms free. This resulted in the slip time jumping suddenly from 0 to 7 ns when the adsorbed layer became unpinned. Once unpinned, the adsorbed layer displayed hysteresis when the amplitude was decreased the adsorbed layer stayed in the unpinned state to a smaller vibration amplitude than when the amplitude was increased. In essence, a pinned adsorbed layer exhibits static friction with a need to apply a threshold shear stress to initiate sliding, and, since the static friction is higher than the kinetic friction, hysteresis is observed.

11.4 References

REFERENCES 309


REFERENCES


WEAR

The most elementary definition of wear is the loss of material from one or both of the contacting surfaces when subjected to relative motion, while a broader definition includes any form of surface damage caused by rubbing one surface against another. Wear is a complex process, making it one of the most difficult aspects of tribology to study.

Wear is the basis of the oldest manufacturing processes, dating back to when humans fashioned their first tools by cutting, grinding, and chipping various sticks and stones they found lying around into more useful shapes. Now, sophisticated grinding, cutting, and polishing processes can fabricate parts with highly controlled shapes and surface finishes. In some industries, parts are machined with nanometer dimensional tolerances with only a few ångstroms of roughness. In semiconductor fabrication, for example, it is often necessary to remove the surface irregularities from the previous processing step to flatten the wafer surface before adding additional circuit elements. This accomplished through a process called chemical-mechanical polishing (CMP) or planarization, where a polishing pad rubs an abrasive and corrosive chemical slurry against the wafer to remove the uneven material. With advanced CMP processes, heterogeneous surfaces can be polished to a roughness of a few ångstroms.

In other situations, however, wear is poorly controlled and detrimental to contacting surfaces. The “wearing out” of contacting parts has widespread economic consequences, for example from the cost of replacement parts, machine downtime, and lost business.

Sometimes the consequences of wear are catastrophic. On January 31, 2000, Alaska Airlines flight 261 crashed into the Pacific Ocean killing all eighty eight people on board, when improper lubrication of a jackscrew had led to excessive wear and eventually to the loss of control of the plane’s horizontal stabilizer.

Despite the technological importance of wear, no simple and universal model has been developed to describe it. As with many other tribological phenomena, the multitude of physical mechanisms contributing to wear makes it difficult to develop a general understanding of how wear occurs at the nanoscale.

However for many specific situations, the underlying wear mechanisms have been determined in a fair amount of detail. In this chapter, we will not attempt a comprehensive review of these previous studies of wear (such a review would require several volumes). Instead, we provide only a brief outline of the most
common mechanisms contributing to wear and of how the modern tribologist considers wear at the nanoscale.

12.1 Simple model for sliding wear

We start with a very simple model for sliding wear, due to Archard and usually referred to as the Archard wear equation (Archard 1953). Sliding wear is the wear that occurs from other mechanisms besides abrasive wear. Abrasive wear occurs when hard particles or hard asperities rub against a surface to cause damage or material removal; it is discussed later in this chapter. The distinction between sliding and abrasive wear is not clear-cut, and often sliding wear processes generate debris that leads to abrasive wear.

This model derives an expression for the wear rate $W$, which is defined to be the volume $V$ of material removed per unit sliding distance $s$:

$$W = \frac{V}{s}. \quad (12.1)$$

So, a basic assumption of this wear model is that the wear volume is proportional to the sliding distance. While this is generally borne out in experiments, it should be kept in mind that it is not universally true, as wear can change the nature of sliding surfaces to such an extent that the wear mechanism changes. For example, when parts first rub against each other, a running-in or breaking-in period often occurs where the initial wear rate is higher or lower while the surfaces are evolving to their “steady state” of wear.

Another main assumption of this model is that, due to surface roughness, contact only occurs where the asperities touch, so the true area of contact is the sum of all the individual asperity contact areas. Figure 12.1 shows schematically the formation of a wear particle at an individual asperity contact during a sliding contact. For this individual contact zone, the maximum area is $\Delta A$, and the characteristic radius is $a$; leading to

$$\Delta A \sim \pi a^2. \quad (12.2)$$

Wear occurs when a fragment of material detaches from an asperity during contact. In this wear model, the volume of this fragment is presumed to be proportional to the cube of the contact dimension $a$. If we approximate the shape of this wear fragment as a hemisphere with radius $a$, the volume of this wear fragment can be approximated as

$$\Delta V \sim \frac{2}{3} \pi a^3. \quad (12.3)$$

While in contact, this asperity slides the short distance

$$\Delta s \sim 2a. \quad (12.4)$$
Most asperity contact events, however, do not generate a wear fragment; instead, detachment only occurs for a small fraction $\kappa$ of these events in a sliding distance $s$. Then, for individual asperity contacts, the average wear rate for generating a wear particle is

$$\frac{\Delta V}{s} = \kappa \frac{\Delta A}{3}. \quad (12.5)$$

The overall wear rate is the sum over the individual asperity contacts:

$$W = \frac{\kappa}{3} \sum \Delta A$$

$$= \frac{\kappa}{3} A_r \quad (12.6)$$

where $A_r$ is the real area of contact between the two surfaces.

In Archard’s original model, he considered the wear at metal surfaces where it is safe to assume that deformation of the asperity contacts is predominately plastic. As discussed in Section 3.4.1.2, this leads to the real area of contact $A_r$ being proportional to the loading force $L$ and described by eq. (3.34): $A_r = L/H$ where $H$ is the hardness of the surface region. To simplify the final equation, we will combine the factor of $1/3$ in eq. (12.6) with the constant of proportionality.
by setting $K = \kappa/3$; this leads to the *Archard wear equation* for wear rate:

$$W = \frac{KL}{H}. \quad (12.7)$$

In this equation the dimensionless quantity $K$ is referred to as the *coefficient of wear* or the *wear coefficient*. It is always less than unity and typically much less: for unlubricated materials, $K$ typically lies in the range from $10^{-2}$ to $10^{-7}$.

One of the primary purposes of lubrication, of course, is to lower wear rates, with the lowest wear rates being obtained for full-film hydrodynamic lubrication, where the wear coefficients are usually insignificant ($<10^{-13}$). For boundary lubrication systems, the wear coefficients are generally range from $10^{-6}$ to $10^{-10}$, which, as expected, is lower than most unlubricated systems.

For those situations where the deformation of the contacting asperities is primarily elastic, the real area of contact is given by eq. (3.35):

$$A_r \simeq 3 \left( \frac{R}{\sigma} \right)^{1/2} \frac{L}{E_c} \quad (12.8)$$

where $E_c$ is the composite elastic modulus of the two surfaces, $R$ the mean radius of curvature of the surface asperities, and $\sigma$ the standard deviation of the asperity heights. In this elastic situation, our simple model indicates that the wear rate should be given by

$$W = \frac{K_eL}{E_c} \quad (12.9)$$

where $K_e$ is the coefficient of wear given by

$$K_e = \kappa \left( \frac{R}{\sigma} \right)^{1/2}. \quad (12.10)$$

As it is often difficult to know what is the hardness or the elastic modulus of the near surface region where wear occurs (or even whether the asperity deformations are predominantly plastic or elastic), the Archard wear equation is often expressed as

$$W = kL \quad (12.11)$$

where $k$ is called the *dimensional wear coefficient*. It is usually quoted in units of mm$^3$ (Nm)$^{-1}$.

For dry sliding systems without lubrication, the wear coefficient varies by more than five orders of magnitude, while coefficients of friction for dry sliding systems typically lie in the fairly narrow range between 0.18 and 0.8. This suggests that the wear rate is much more sensitive to the underlying wear mechanism than friction. While the wear coefficients provide valuable means for comparing the severity of wear, they tell us nothing about the underlying mechanisms that are responsible.
12.2 Major influences on wear rates

The Archard wear equation is deceptively simple—it basically says that the wear rate is proportional to the real area of contact, which, for rough contacting surfaces, is proportional to the loading force. In principle, for a good predictive model of wear rate, we need only to determine the coefficient of wear for a particular sliding system. In practice, however, wear coefficients can vary dramatically as the sliding conditions change. We have already mentioned how transient wear rates occur during an initial running-in period while the sliding surfaces evolve toward their steady-state sliding conditions. Another serious problem is that the wear coefficients tend to depend strongly on the magnitude of the loading force. As a consequence, wear rates are not generally proportional to the loading force as indicated by the Archard wear equation, except over limited ranges.

For most sliding systems, no single wear mechanism dominates at all operating conditions; rather, different wear mechanisms operate, with their relative importance changing as the sliding conditions change, and with abrupt changes in wear rates occurring when one dominant wear mechanism replaces another.

Next, we discuss how the different variables of loading force, sliding speed, sliding time, and environment control the main factors of mechanical stresses, interface temperature, and oxidation that influence the wear processes for metals; much of the complexity of wear comes from the interrelation of these variables and factors:

- Increasing the loading force generates higher contact pressures and shear stresses; this causes more plastic deformation in the metal, which leads to more mechanical damage.
- The temperature rise at a contacting junction depends on the rate of frictional heating and on how fast the heat is conducted away from the contact point. At slow sliding speeds, the temperature rise is small since the low power from frictional heating (sliding speed times friction force) is quickly conducted away; in the limit of slow sliding speeds, the sliding can be considered isothermal. At high sliding speeds, heat is generated much faster than it is conducted away, resulting high interface temperatures; in the limit of high sliding speeds, the sliding is considered adiabatic. Increasing the loading force increases the frictional force, resulting in more frictional heating and higher interface temperatures.
- For metals, these high temperatures can lead to mechanical softening of the asperities and, in the extreme, melting of the near surface region. For ice, it is well known that much of its slipperiness comes from frictional heating melting the ice near the surface to form a thin film of water that lubricates the sliding surfaces (Rosenberg 2005).
- High interface temperatures from frictional heating promote chemical reactions at the sliding surfaces, such as oxidation of metal surfaces in air or degradation of lubricant films.
12.2.1 Wear maps

A useful way of diagramming how the wear of a material responds to changes in conditions is to construct a map of wear behavior based on experimental observations and physical modeling (Lim et al. 1987; Hsu and Shen 1996; Lim 1998). Figure 12.2 shows an example of such a map for unlubricated steel sliding in air. A variety of names have been given to these types of maps: wear-mechanism maps, wear-mode maps, wear-transition maps, and wear-regime maps. Once established for a particular set of contacting materials, these maps can be valuable guides for predicting the conditions where a particular wear mechanism dominates. While Fig. 12.2 illustrates the wear response for steel, wear maps for most metals sliding in air have a similar form.

In a typical wear map, dimensionless variables are used: a normalized contact pressure, defined as the loading force divided by the nominal contact area times
the hardness of the softer material, and a normalized velocity, usually the sliding velocity divided by the velocity of heat flow.

From Fig. 12.2, we see that, at low sliding speeds, the wear is dominated by mechanical processes, which are determined more by mechanical stresses than by sliding speed. When the sliding speed is increased beyond the point where the system ceases to be considered isothermal, wear mechanisms that are more dependent on temperature become dominant. For steel sliding in air, the wear mechanism transitions over a narrow range of velocities in Fig. 12.2 to a regime where oxidational wear dominates because the rise in temperature from frictional heating promotes surface oxidation. In this regime, somewhat counterintuitively, the wear rates decrease with increasing velocity due to the formation of oxide films, which provide a protective layer against further wear.

12.3 Mechanisms of wear

12.3.1 Wear from plastic deformation

For metals sliding at low speeds, mechanical wear processes involving plastic deformation dominate. The general requirement for plastic deformation is that the mechanical stresses generated by the adhesive, loading, and frictional forces exceed the yield stress of one or both of the sliding materials.

During sliding, the shear stresses generated by friction add to those from the loading and adhesive forces, changing the position and magnitude of the maximum stress. For the stationary case \((\mu = 0)\), the stress is concentrated a distance \(\sim 0.5a\) below the surface (Section 3.3.1). As long as \(\mu\) stays below 0.3 during sliding, the maximum shear stress and the associated plastic flow remain beneath the surface. This constraining of the plastic flow region beneath the surface limits the possible plastic strain that can accumulate with each sliding pass. Friction coefficients less than 0.3 are characteristic of lubricated surfaces and metals covered with oxide layers.

As the traction force from friction increases, the location of the maximum stress moves closer to the surface and more towards the trailing edge of the contact zone. When \(\mu\) is above 0.3, the maximum stress is concentrated at the surface at the trailing edge of the contact zone. For those situations where the peak shear stress occurs at the surface, the induced plastic flow is much less constrained than for subsurface yield, resulting in larger shear strains and more surface damage.

Plastic deformation generated by sliding contacts leads to a number of distinct wear mechanisms:

- **Delamination** – Plastic flow nucleates and promotes of the growth of subsurface cracks that propagate parallel to the surface, before extending out to the free surface to form platelet-like wear particles.
- **Surface cracks** – For brittle materials, the high tensile stresses generated by the traction within the contact region lead to surface fractures and cracks as illustrated in Fig. 12.3.
Fig. 12.3. Surface cracks forming in a brittle material behind the trailing edge of a slider, due to the shear stresses generated by the friction $F$ as the slider moves from left to right.

- **Fatigue** – Repetitive variations of stresses from repeated sliding or rolling contact results in fatigue failure in the near-surface region.

- **Mixing** – With the contact sliding of unlubricated ductile materials, the friction induced plastic flow leads to high plastic shear strains (in the range 10 to 1000) at the sliding surfaces (Rigney 2000). One of the consequences of these high strains is that the grains near contacting surfaces are often reshaped down to the nanometer size. Another consequence of this is mixing of the chemical constituents of the sliding surfaces, resulting in the near surface regions and wear debris being a mixture of the material from the two sliding surfaces (Rigney 2000; Karthikeyan et al. 2005).

- **Seizure** – During the sliding of metals at extremely high loads, the high contact pressures and shear stresses acting on the contacting asperities cause them to undergo extensive plastic flow and junction growth (Section 4.2.4). Under these circumstances, the real area of contact approaches the nominal area, leading to such high adhesion and shear forces over these enlarged contact zones that momentary seizure of the moving parts takes place followed by severe wear when they move again (upper portion of the wear map for steel in Fig. 12.2).

- **Adhesion** – Adhesive forces combined with plastic flow pull out wear particles from the tips of the asperities. As this is a fairly common wear mechanism, the next section elaborates further on adhesive wear.

### 12.3.2 Adhesive wear

The concept of adhesive wear is based on the notion that adhesion occurs between asperities when they touch, followed by plastic shearing that plucks off the ends of the softer asperities; these bits then adhere to the opposing surface before eventually becoming loose wear debris.

Adhesive wear can be analyzed using Archard’s simple wear model described in Section 12.1 and Fig. 12.1. According to this model, higher adhesion forces in the contact zones are more likely to pull out a wear fragment. Consequently, higher surface energy should result in higher wear rates, since adhesive forces scale with surface energy (Chapter 6).
Since the surface energy depends strongly on the chemical composition of the surface, wear rates from this adhesive mechanism should be very sensitive to the presence of contamination layers or lubricant films, since these strongly influence adhesion at the solid–solid contacts. For example, clean metals sliding in vacuum experience very high adhesive and friction forces, as the metals’ adsorbed contamination layer, which is present in air, is absent in the vacuum environment. (High adhesive forces in vacuum lead to the phenomenon of cold welding, described in Section 11.2.3.)

For metals sliding in air, the probability of adhesive wear can be dramatically reduced by replacing the adventitious contamination layer with a monolayer of boundary lubricant, as described in Section 10.2. Besides lowering the overall surface energy, a good boundary lubricant layer also resists displacement by the contacting asperities, which prevents the underlying metals from coming into direct contact and forming strong metallic bonds.

12.3.2.1 Example: An atomic level simulation of adhesive wear
Sorensen et al. have used atomic level simulations of copper tips sliding against copper surfaces to investigate how friction and wear occur for clean metal surfaces (Sorensen et al. 1996). In these simulations, they found that, in fcc metals, slippage (and hence plastic flow during wear) occurs preferentially along the close packed (111) crystal planes. So, if the end of the tip with a flat (111) crystal surface is slid over another (111) facing crystal surface, slippage occurs between these two (111) crystal faces, and the tip slides without wear. The sliding motion in this case occurs by an atomic scale stick-slip motion (Sections 4.3.1.3 and 11.2.2.1).

When the tip end and substrate both have (100) rather than (111) crystal faces, a very different sliding behavior occurs. Figure 12.4 shows the case where the lattices of the two (100) are aligned so that atoms on the opposing (100) surfaces are commensurate. (As discussed in Section 11.2, friction and adhesion forces between contacting surfaces are highest when the opposing surfaces are commensurate.) So, when the tip moves across the surface, slippage occurs along the (111) crystal planes within the tip rather than at the (100) crystal plane of the original interface as the (100) planes have a higher critical shear yield stress than the (111) planes. Eventually, a pyramid shaped pile of atoms detaches from the tip to form a potential wear particle.

12.3.3 Abrasive wear
In abrasive wear, the rubbing of hard particles or hard asperities against a surface removes or displaces the material from that surface. For abrasive wear to occur, these particles or one of the contacting surfaces needs to be considerably harder than the surface being abraded (at least 1.3 times harder). The abrasive wear process leads to a characteristic surface topography of long grooves running in
the same direction as the sliding motion. Figure 12.5 illustrates several types of abrasive wear; two general types occur:

1. *Three-body abrasion* – Wear is caused by hard particles that are free to roll or slide between two sliding surfaces. Some examples of particles that cause abrasive wear: polishing slurries, wear debris generated by other wear mechanisms, and grit contaminants in a lubricated bearing.

2. *Two-body abrasion* – Wear is caused by the hard protrusions on one surface scapping and gouging the other surface. Cutting tools and sand paper are examples of two-body abrasive wear.

A simple way to model the abrasive wear of an asperity in a two-body process or the corner of a grit particle in a three-body process is to consider a cone of a hard material gouging a groove into a softer material as shown in Fig. 12.6.
Fig. 12.5. Types of abrasive wear: (a) Three body abrasive wear where free hard particles rotate or slide between the sliding surfaces. (b) Two body abrasive wear where abrasive grit is embedded in the top surface and abrades the lower surface. (c) Two body abrasive wear where asperities of a rough, hard surface abrade a softer surface.

Fig. 12.6. A cone shaped asperity plowing through a softer surface to form an abrasive wear groove.
This model is closely related to the model for plowing friction discussed in Section 4.2.2. If the depth of the groove is $h$ and the slope of the side of the cone is $\tan \theta$, the volume $V$ of material displaced when this asperity slides a distance $s$ is given by

$$V = sh^2 \tan \varphi = sh^2 \cot \theta. \quad (12.12)$$

Since only the front half the asperity supports the loading force $L$ over an area $\pi a^2/2$, we have

$$L = \frac{1}{2} H \pi a^2 = \frac{1}{2} H \pi h^2 \cot^2 \theta \quad (12.13)$$

where $H$ is the hardness of the soft surface. Combining eqs. (12.12) and (12.13) to eliminate $h$, we obtain the following expression for the wear rate:

$$W = \frac{V}{s} = \frac{2 \tan \theta}{\pi} \frac{L}{H}. \quad (12.14)$$

Though derived from different initial assumptions, eq. (12.14) has the same form as the Archard wear equation [eq. (12.7)], with a wear coefficient for abrasive wear of $K_a = 2 \tan \theta/\pi$, where $\tan \theta$ is the average slope of the asperities on the hard surface. From eq. (12.14), we see that the wear coefficient is proportional to the average slope of the asperities doing the abrading.

One problem with eq. (12.14) is that it predicts wear rates much higher than actually observed. An obvious source for this overestimation is the assumption that all the material displaced is lost from the surface. In reality, with each sliding pass, only a fraction of the displaced material also detaches, while the remainder piles up along the edges of the groove.

The amount of material removed with each pass depends on the attack angle and the degree of penetration of the cutting edge into the softer surface. Figure 12.7 is a wear-mechanism map showing the ranges of penetration and interfacial shear strength over which the different abrasive mechanisms of cutting, plowing, and wedge formation occur, for a hard spherical indenter sliding against unlubricated metals. At small penetration depths and shear strengths, mainly plowing occurs, with displaced material piling up along the groove edges. These ridges along the groove undergo fatigue with subsequent passes of the asperity contact before detaching as wear particles; for this plowing/fatigue type wear, the wear coefficient is relatively low ($<10^{-3}$) compared to other abrasive wear mechanisms. When the shear strength is relatively large ($f > 0.5$), the friction forces in the contact zone push the softer material out in front of the indenter in a wedge shaped formation, and material is eventually removed by the propagation of a crack. For degrees of penetration greater than 20%, the cutting mechanism dominates where the displaced material forms micro-chips out in front of the indenter that are easily detached as wear particles. In this cutting regime, wear coefficients over 0.1 can be achieved.
Fig. 12.7. The wear mechanism map for a spherical indenter abrading an unlubricated metal. The degree of penetration $D_p$ is defined as the penetration depth $h$ divided by the radius of contact $a$. The dimensionless shear strength $f$ is the ratio of the interfacial shear strength and the shear yield stress of the material being abraded. Reproduced with permission from Hokkirigawa and Kato (1988). Copyright 1988, Elsevier.

12.3.4 Oxidative wear
12.3.4.1 Metals

Many materials, such as metals, react with oxygen in air to form oxide layers on their surfaces. Due to its universality, the role of oxidation in the wear of metals has been studied extensively.

A fairly typical case is the oxidative wear of steel, discussed already in connection with the wear-mechanism map in Fig. 12.2. For unlubricated steel surfaces, sliding at speeds below about 1 m/s, wear occurs predominately through plastic deformation of the near surface region, to generate mainly metallic debris.

A sliding speed of 1 m/s works out to be just sufficient to generate the flash temperatures of about 700°C needed to cause oxidation at contacting asperities on the steel surfaces. Rapid oxidation of these contacts is promoted by the mechanical action of sliding, which generates numerous defects (voids, vacancies, dislocations, etc.) in the near surface region that provide channels for easy diffusion of oxygen atoms into the metal. Oxide films reduce the wear rate of metals by decreasing the shear strength of the sliding interface, which reduces the large subsurface shear strains that are necessary for the plastic deformation wear mechanisms.
At sliding speeds just above the oxidation threshold, the oxide is thin, patchy, and brittle, and mild oxidational wear occurs as these patches are scraped off and the exposed metal is reoxidized.

With increased frictional heating at higher sliding speeds, oxidation becomes more widespread, resulting in a thicker, more continuous oxide film. The higher interface temperature at these higher speeds also promotes plastic flow of the oxide film, leading to more severe oxidational wear.

12.3.4.2 Carbon overcoats
In the past few decades, amorphous carbon films (also called diamond-like carbon or DLC) have come into common use as protective films that are deposited on many types of surfaces (disks in disk drives, razor blades, etc.) to reduce corrosion and wear. In unlubricated sliding, these carbon overcoats suffer from oxidative wear when the high temperatures generated by frictional contacts cause the carbon to burn off as carbon dioxide (Marchon et al. 1990; Yen 1996).

Similarly, oxidative wear is the main form of wear for the carbon–carbon composite material used in high performance brakes such as those in the wheels of jet airplanes.

12.3.4.3 Ceramics
Non-oxide ceramics (such as silicon nitride, silicon carbide, titanium nitride, and titanium carbide) form oxide layers when exposed to air and water vapor, and these layers can have a strong influence on the wear of the surfaces. The friction and wear of these materials are therefore sensitive to the presence of water and the relative humidity in the sliding environment.

For example (Fischer and Tomizawa 1985), silicon nitride forms an oxide layer in the presence of water via the reaction

$$\text{Si}_3\text{N}_4 + 6\text{H}_2\text{O} \rightarrow 3\text{SiO}_2 + 4\text{NH}_3.$$  

Subsequently, the surface of this oxide film is hydrated via the reaction

$$\text{SiO}_2 + 2\text{H}_2\text{O} \rightarrow \text{Si(OH)}_4.$$  

The presence of water on ceramic surfaces is also known to promote crack propagation and plasticity of ceramics (the Rebinder effect). For silicon nitride, the presence of a hydrated surface layer reduces its shear strength, resulting in lower wear rates at high humidity. As the sliding speed is increased, the higher temperatures from frictional heating decrease the hydration layer and its protective influence, increasing wear rates. For some ceramics such as silicon carbide, alumina, and zirconia, the presence of water increases wear rates, rather than decreasing them, due to the increased plasticity of the surface from the Rebinder effect.
12.4 Plasticity at the nanoscale

Many of the wear processes described above involve plastic deformation of material in small localized regions where asperities of the opposing surfaces or hard particles make contact. It turns out that the onset of plasticity for such small regions can be dramatically different than the plastic yield stress determined by macroscopic measurements. As described in Section 3.1, macroscale plastic deformation within crystalline materials occurs mainly through the motion of flaws and defects in the crystal structure, called dislocations. When the size of the region undergoing plastic flow becomes much smaller than the typical distance between dislocations, however, the contribution of dislocations to the yield stress becomes insignificant; instead, the yield stress is governed by the force needed to slide one plane of atoms over another. For a perfect crystal with no dislocations, the theoretical shear stress needed to slide a plane of atoms over another is given by

\[
\tau_c \sim \frac{G}{2\pi} \quad (12.15)
\]

where \(G\) is the shear modulus (Section 3.2.1); \(\tau_c\) is typically a factor of 100 higher than the measured value of yield stress in actual crystals.

As discussed Section 3.3.2, when a hard spherical indenter is pushed against a surface in elastic contact so that the contact zone has a radius \(a\), the peak pressure and shear stress are localized a distance \(\sim 0.5a\) below surface. As the loading force on the indenter is increased, these peak stresses increase until the yield stress of the material in this region is exceeded. Since the pioneering work of Oliver and Pharr (Oliver and Pharr 1992), great strides have been made in developing nanoindentation techniques for measuring the elasticity and yield stresses in very localized regions a few tens of nanometers across. In these nanoindentation instruments, typically a Berkovich diamond indenter is used that has a triangular pyramid shape. Even for the sharpest Berkovich indenters, the apex is always a little bit rounded; while this tip radius can be as small as ten nanometers, more typical radii are a few hundred nanometers, which results in contact zones tens of nanometers across, at a few millinewtons of loading force.

Figure 12.8 shows the results from a typical nanoindentation experiment on a single crystal, in this case, a Cr$_3$Si crystal (Bei et al. 2005). In these standard indentation experiments, the loading force \(P\) on the indenter is first increased (loading) then decreased (unloading), while the depth the indenter penetrates into the sample (or its displacement \(h\) into the sample) is measured; the resulting data is plotted as \(P\) vs. \(h\).

In Fig. 12.8, the left-hand plot shows the case where the maximum load \(P\) is below the threshold for nucleating dislocations; since few or no dislocations occur in the contact zone, the deformation is entirely elastic, and the unloading curve retraces the loading curve.
The right-hand plot shows the case where the loading force is increased above the threshold for nucleating dislocations. At this threshold of $P_c = 3.9 \text{ mN}$, the material underneath the indenter becomes elastically unstable and abruptly undergoes plastic flow, causing a “pop-in” that is characteristic of this flow. As further evidence that plastic deformation has occurred, the unloading curve for the right-hand plot does not retrace the loading curve, but instead unloads to a finite $h$, which is the depth of the residual indentation hole. From the geometry of the indenter and the value of $P_c$, Bei et al. estimate that the maximum shear stress underneath the indenter at the onset of the pop-in is close to the theoretical maximum in eq. (12.15). Similar pop-ins have been observed in other nanoindentation experiments with other crystalline material, and the peak shear stresses in these cases are also estimated to be close that predicted by eq. (12.15). Theory and atomic simulations indicate that these pop-ins occur when the maximum shear stress reaches the value need to slide atoms over the crystal planes with the lowest shear strength (Kelchner et al. 1998; Rodriguez de la Fuente et al. 2002; Li 2007). Once this slippage starts, it creates dislocations that enable atomic motion along the other crystal planes, and the process cascades with rapid plastic flow during a pop-in event until the stress is reduced to the yield stress determined by the motion of dislocations.
So, while a sliding surface may start off with a low density of flaws and defects, during the initial stages of wear, plastic deformation by the contacting asperities quickly builds up dislocation density in the near-surface layer so that it develops a fine grain or almost amorphous structure.

12.5 References


INDEX

adhesion hysteresis 104–110
adhesive force 6, 114–135, 154–160, 195–204
Auger electron spectroscopy (AES) 12–13
bearing area curve 29
bearings
air bearing surface (ABS) 8–9
gas 232–235
history 11
incline plane 229
journal 230–232
Rayleigh step 229–230
blow-off technique 301–307
capillary condensation 6, 90–92, 131–132, 197–199
capillary pressure 85–90, 267–271, 275–279
carbon overcoat 8–10, 326
Casimir force 146
chemical force microscopy 194
cold welding 295
contact area
elastic sphere on flat 45–48
multi asperity 51–59
real 50–59
contact angles 96–105
Fowkes method 98
hysteresis 103–104
measurement 101–102
Young–Dupre equation 97
Young’s equation 96–97, 120
Zisman method 98–100
contact electrification 171–180
contact potential 173–177
contact recording 10
Couette flow 216
cows 24, 45
Debye force see induction force
Derjaguin approximation 113–114
digital micro-mirror devices (DMD) 6–7
disjoining pressure 90, 133, 160–162, 265–272, 275–278
disk drives 7–10, 33, 55, 61, 99, 175, 272–273, 275–279
dislocations 42–43, 94, 327–328
dispersion force 140
Dupre equation 93
elastic constant 43, 45
elastic deformation 41, 43–48
elastic modulus see elastic constant
electron spectroscopy for chemical analysis (ESCA) 12–13
electrostatic double-layer 167–169, 202–204
gear oil 4–5
Fourier transform infrared spectroscopy (FTIR) 12–13
Frenkel–Kontorova model 287–288
friction
adhesive 11, 68–69, 284–286
Amontons’ law 10, 63–66
automotive 4–5
coefficient 64, 68–69, 71
Coulomb’s law 64
electronic contribution 298
kinetic 72–74, 286
phonon contribution 297–298
plowing 68–69
static 72–74, 286
gecko adhesion 158–160
Greenwood and Williamson model 50–59, 155
hardness 49–50
hydrophobic attraction 170–171
height distribution 29
history of tribology 10–12
hydration repulsion 169–170
induction force 140
inelastic impact 59–61
interfacial energy 92
Johnson–Kendall–Roberts (JKR) theory 117–121
Jost Report 3
junction growth 70–72
Keeson force see orientation force
Kelvin equation 91
Laplace pressure see capillary pressure
laser texture 55, 130, 272, 276–277
Lennard–Jones potential 138
London force see dispersion force
lubrication
boundary 11, 208–209, 255–265
elastohydrodynamic 207–208, 235–241
hydrodynamic 11, 207–208, 228–235
hydrostatic 207
lubricant 9–11, 241–242
regimes 207–209
meniscus force 122–135, 272–276
micro-electromechanical systems (MEMS) 5–7, 133–135
micro-motor 6
millipede 18–19
molecular dynamic simulations 14
boundary lubrication 258–259
friction 295–296
nanoimprinting 16–18
nanotechnology 19–20
nanotribology 11, 19
orientation force 139–140
oscillatory force see solvation force
pillbox menisci 131–133, 270–276
plasticity index 54
plastic deformation 41–42, 48–49, 319–320, 327–329
Poiseuille flow 216
probability distribution 29
quartz crystal microbalance (QCM) 299–300, 308
Raman spectroscopy 12–13
repulsive atomic potentials 138–139
retardation of dispersion force 142
roughness 24–37
average 26–27
parameters 26–29
rms 26–27
sandcastles 128
water menisci in sand 126–128
scanning electron microscopy (SEM) 13, 30
scanning tunneling microscopy (STM) 13, 30
secondary ion mass spectroscopy (SIMS) 12–13
shear strength 66, 68, 71
limiting 213
pressure dependence 286
simple model 285–286
shear thickening 212
shear thinning 212–213, 251
slider-disk interfaces 7–10, 130, 175–177, 272–273, 275–279
slippage 216–226
apparent slip 222–224
gas bearings 234–235
lotus effect 223–224
molecular slip 220–222
nano-bubbles 224
polymer melts 222
slip length 217–218
super-hydrophobic 224
solidification of a liquid 248–252
solvation force 162–167, 247
stick-slip 74–81
atomic scale 79–81, 292–294
Stribeck curve 208
sum frequency generation 12–13
superlubricity 20, 289
surface energy 82–108
critical surface tension 99–100
lubricants 241–242
van der Waals 152–153
surface force apparatus (SFA) 14, 114, 121, 188–192, 247–252
surface tension see surface energy
toe dipping menisci 128–130, 270–276
Tomlinson model 290–294
triboelectricity see contact electrification
tribology name 2–3
triboluminescence 179
van der Waals interactions 6, 20, 32, 139–162, 195–197
Hamaker constant 145–151
 Lifshitz’s theory 146–150
rough surface 154–157
surface energy 152–153
viscosity
dynamic 209
enhanced 248–255
kinematic 210
non-Newtonian 211–214
pressure dependence 235–236
temperature dependence 214
waviness 25
wear
abrasive 321–325
adhesive 11, 320–321
Archard equation 314–316
coefficient 316
INDEX

maps 318–319
oxidative 325–326
wetting 101
at contacting interfaces 128–133
work hardening 70

x-ray diffraction 13
x-ray reflectivity 13
Young and Laplace equation 86