Notes on Physical Metallurgy

Ju Li, MIT, December 10, 2015

Chapter 1

Structures

The connections between structures - properties - processing, as illustrated on the cover of *Acta Materialia*, is the basis of materials science and engineering. Because of the relatively simple atomic structure of metals and the wide applications of metals (Bronze Age, Iron Age, copper interconnects in microprocessors), these connections were first discovered and distilled by metallurgists, and such methods and outlook are then extended to other kinds of materials: ceramics, polymers, semiconductors, biomaterials, etc. Although details vary a lot across different materials classes, especially in synthesis, the spirit and outlook are preserved a great deal in most areas of materials science and engineering - including the reliance on thermodynamic (Josiah Gibbs) and kinetic (Lars Onsager, John Cahn) theories, the realization of the importance of intervening scales ("**microstructure**" controls properties), the respect for processing details (physicists don't appreciate the detailed recipe- and history-dependencies of processing as much as materials scientists).

The narrowest definition of Physical Metallurgy was the control of materials properties by thermo-mechanical processing ("heat and beat"), as distinguished from Chemical Metallurgy (changing the chemical composition), and Mechanical Metallurgy ("just beat, no heat"). Such definition is quite arbitrary, however, fundamentally. By "heat", people mean raising temperature T significantly above $T_{\text{room}} = 300K$, as most applications (let's say $\sim 90\%$ of where metals are applied in tonnage) are at T_{room}. But 300K for steel is quite different from 300K for Sn. And even though $k_{\text{B}}T_{\text{room}} = 1/40$ eV is quite "small" compared to the primary bond energies in steel (vacancy formation energy in α-Fe is 1.5 eV, which is about 4 bonds' worth), it turns out thermal fluctuations still cannot be ignored for most

processes of interest. [1] Also, physical metallurgists rely a lot on chemical thermodynamics and phase diagram in treating diffusion and phase transformations, so the boundary between Physical and Chemical Metallurgy is not sharp. For this reason, I would like to regard Physical Metallurgy as **Physical**, **Chemical** and **Mechanical** Metallurgy combined, but with emphasis on **thermo-mechanical processing**.

The students in this class are expected to have taken 3.022 *Microstructural Evolution in Materials* and *3.032 Mechanical Behavior of Materials* or their equivalents already. Some topics of our class may overlap with ealier courses. But because we are 3.14 (upper undergrads) / 3.40J / 22.71J (grad student), this course is expected to be offered at a somewhat higher level, and also will have an integrative flavor. If all goes well, at end of this course you may agree that materials science is a very subtle science: one may think one understands something, until one looks at it once again from another angle, or at a different time/lengthscale.

We begin with the word "structure". By now you have heard about (a) **atomic structure**, (b) **molecular structure**, as in *double-stranded helical* DNA, (c) **microstructure**, as in the (perhaps somewhat chauvinistic) old-school metallurgical matra "microstructure controls properties", (d) **nanostructure**, as in "One nanometer (one billionth of a meter) is a *magical* point on the dimensional scale. Nanostructures are at the confluence of the smallest of human-made devices and the largest molecules of living systems..." - yes, they did use the word "magical" - from a call for proposal from the US National Science Foundation [2]. You may have also heard the word **electronic structure** from researchers doing so-called *ab initio* or first-principles calculations, in solving Schrodinger equation for many electrons, to obtain atomic interactions and interatomic forces numerically. So, which of these "structures" is the most important?

The modern view is that none of the above structures, electronic structure - atomic structure (molecular structure) - nanostructure - microstructure, and even macro-structures (such as a trussed roof), could be ignored. In other words, all these structures potentially could be "equally" important. According to this "multiscale materials" view, there is no need to be particularly chauvinistic about any particular lengthscale, from the Å to the m. There are interesting physics and theory about these physics at all these scales, and they cascade through each other ("handshake"). Like a chain, no link can be missed on this chain of logic.

According to [1], p.1, microstructure is what can be observed under an optical microscope

[1] If in doubt, some properties of steel at 300K, such as ductility, can be really different from at 0K.
[2] Nanoscale Science & Engineering Center program, U.S. National Science Foundation, 2001

(such as grains in a polycrystal), with 100× to 1000× magnification power. However, when the grain size D shrinks to say, 80nm, optical microscope can't see the grains (one has to use electron microscopes), but the dependence of properties such as yield strength σ_Y on D (so called **Hall-Petch relation**: $\sigma(D) = \sigma_0 + kD^{-1/2}$) is no less sensitive than when D is 10 μm when optical microscope can see them [2]. Thus, putting a hard lengthscale bound on a concept based on the resolution limit of some observation instrument is convenient but not very enlightening. In this course, we intentionally smear the concept of microstructure to make the concept inclusive. We would **not** separate nanostructure from microstructure. **Dislocations, cracks, grain boundaries, surfaces, phase boundaries, etc.** are all components of the microstructure. In our course, the term microstructure just denotes some structural order or descriptor, **beyond** the Angstrom-level **atomic structure**.

The term "Microstructural Evolution" is one of the most frequently used word in metallurgy. The choice of the word "Evolution" is intriguing. What is common between Darwin's Evolution in Earth's biosphere, and "Microstructural Evolution" inside a piece of metal?

A modern view, mostly coming from physicists, is that biological evolution and microstructural evolution all belong to so-called emergent phenomena (also called self-organization, spontaneous order) or emergence in *dissipative systems*. A dissipative system is a system **out of equilibrium**, where free energy is *being spent* ($\Delta S_{\text{universe}} > 0$) instead of conserved ($\Delta S_{\text{universe}} = 0$). In thermodynamics, we know that an isolated system at equilibrium, such as a canister of gas with no energy or mass in or out, will reach the state of maximum disorder (homogeneous in density, temperature, no flow), or maximum entropy, for the given constraints (the adiabatic, mechanically strong and impermeable box that contains the gas molecules). However, Earth is not an isolated system: there is high-quality sunlight ($T_{\text{blackbody}} = 5800$K) coming in, and lower-quality light emission going out (think $T_{\text{blackbody}} = 2.7$K for cosmic background radiation). So Earth is in effect a giant heat engine, enjoying the great benefits of a huge free-energy influx, even though the raw energy flux is nearly balanced (energy in = energy out). This free-energy **dissipation** is what supports biosphere, from the smallest plankton to whales, and Darwin's Evolution, where order emerges - the *species* - with some highly conserved traits, despite of rich and intricate interactions between the species.

The same principle is true for a piece of metal. When one "heats and beats" a metal, or react oxygen or hydrogen or lithium with it, or irradiate it with high-energy neutrons, one is injecting free energy into the system, or "dissipating good energy". In exchange for such wasteful behavior, one gets to see beautiful patterns that emerge spontaneously inside the

metals, the "microstructures".

Like "following the money" in *All the President's Men*, it is crucial to account for the flows of free energy

$$G \;=\; E + PV - TS \tag{1.1}$$

in a metal, because the flow/dissipation of G generates the microstructures. Indeed, without a large flux of outside free energy, most "microstructures" cannot form. From the Boltzmann formula

$$c_0 \;\propto\; \exp(-E_{\rm f}/k_{\rm B}T) \tag{1.2}$$

where c_0 is the equilibrium concentration of some defect, and $E_{\rm f}$ is its formation energy, one cannot explain the presence of extended defects like dislocations or grain boudaries since their $E_{\rm f} \to \infty$, if based on just equilibrium thermal fluctuations. In other words, most defects except the point defects (like vacancies) are **manifestations** of out-of-equilibrium-ness. And oftentimes even the vacancy concentrations themselves are out-of-equilibrium (such as quenched-in vacancies in aluminum alloys, or materials under irradiation). One needs to beat the metal, quench the metal, irradiate the metal with 1MeV neutrons, do something more "dramatic" like the above, to create the extended defects. Just like plants tend to cover everywhere there is sunlight, dislocations tend to multiply (there is a word "dislocation breeding" in the field as if dislocations were animals), and cracks tend to grow, to take advantage of elastic strain energy density stored in the material. Dislocations or cracks "feed" on elastic strain energy density: $e_{\rm strain}(\mathbf{x}) = \sigma(\mathbf{x})^2/2C$, where C is the elastic constant, $\sigma(\mathbf{x})$ is the stres at location $\mathbf{x}$, and $e_{\rm strain}(\mathbf{x})$ is the elastic strain energy density. Free energy to a metal is like money to a society, or food/ATP to biosphere.

When one stops heating/quenching, beating on or irradiating the metals, one stops injecting more "good energy" into the system. But there is often still some free energy stored inside the system to drive further evolution, for instance in the form of residual stress (due to - surprise - microstructures!), which induces residual elastic strain energy. Then, the microstructures would start a game of scavenging and cannibalization. They would eat each other, annihilate, polygonize, and coarsen. This is called **relaxation**, or recovery, or repair. In relaxation, one fixes all the external constraints, so the spigot of external fresh free energy is shut off, and the system goes from a high-free-energy state to a low-free-energy state for that **fixed set of external constraints**. Eventually, after infinitely long time of relaxation, the system may reach total internal equilibrium, where nothing will change any more (if the external constraints are kept fixed).

Also, just like a whale is a beautiful multi-cellular organism with many levels of organization (the exterior shape, the organs, the tissues, the cells, the cell organelles, etc.), the entire microstructure of a piece of metals manefest at many lengthscales, with organization at multiple levels. One does not ignore a vacancy because it's small ($E_{\rm f} = 1.27$eV in FCC Cu [3]), just like large mammals like us cannot ignore bacteria or virus. But one also needs to appreciate that a dislocation tangle or dislocation cell could manifest much richer behavior than the average property of single dislocations, as there could be new physics at every level of organization that is worth a scientist's attention, which is the central tenet of the Complexity Theory and Emergence.

NOW INTRODUCE the instructor's own research and relation to Microstructural Evolution.

With the above perspective and background, let us dive into the atomic structure of metals, which is an important rung on the ladder of multiscale structures. You are expected to be familiar with Miller indices for crystallographic directions and crystallographic planes already, and I just want to take the case of hexagonal crystals (Mg, Zn, Ti, Zr,... and graphite) to do some practice. For hexagonal system with three-fold basal symmetry, people tend to use a redundant coordinate frame $\mathbf{a}_0, \mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$, where the bold font denotes a vector, and translation by each of $\mathbf{a}_i$ would take one atomic site to an equivalent site in the Bravais primitive cell (e.g. a Bravais translational vector). People also tend to call $\mathbf{a}_3$ vector $\mathbf{c}$. So $[u_0u_1u_2u_3]$ denotes a direction $u_0\mathbf{a}_0 + u_1\mathbf{a}_2 + u_2\mathbf{a}_2 + u_3\mathbf{a}_3$, where u_i may be rational numbers to denote crystallographic directions. However, since

$$\mathbf{a}_0 + \mathbf{a}_1 + \mathbf{a}_2 \ = \ 0 \tag{1.3}$$

$[u_0 + \lambda, u_1 + \lambda, u_2 + \lambda, u_3]$ would denote the same direction as $[u_0u_1u_2u_3]$. This creates one floating degree of freedom in the notation. One can get rid of the float by choosing a "gauge", which is an arbitrary but necessary convention to make the numerical representation unique, like the reference states people use in chemical thermodynamics. The convention people choose is:

$$u_0 + u_1 + u_2 \ = \ 0. \tag{1.4}$$

Thus, $[11\bar{2}0]/3$ would be connecting to a nearest neighbor on the closed-packed plane, and $[10\bar{1}0]$ would be connecting to a 2nd nearest neighbor.

Note that FCC has period-3, ...-◯+-◯+-◯+... stacking, where ◯ denotes a close-packed layer at $z = 0$, + denotes a close-packed layer on top of the ◯ layer (and shifted in plane), and - denotes a closed-packed layer below the the ◯ layer (shifted in plane also, differently).

In contrast, HCP has period-2 stacking: ...◯+◯+◯+... If one looks down onto the HCP basal plane, there are infinite tunnels at the - sites.

For labeling of crystallographic planes in hexagonal system, we note that an atomic plane is defined as

$$\mathbf{n} \cdot \mathbf{x} \;=\; d \tag{1.5}$$

if $\mathbf{x}$ is the separation between an atom on that plane and an observer atom at the origin, d is a constant for the plane, and $\mathbf{n}$ is the surface normal. For the same $\mathbf{n}$, there is a set of equivalent-site atoms (one essentially considers only 1 representative atom in a primitive Bravais cell) forming a plane with closest distance that is yet nonzero:

$$\mathbf{n} \cdot \mathbf{x} \;=\; d_{\rm min} \;>\; 0 \tag{1.6}$$

which we call the minimum-distance plane. Since $\mathbf{a}_0, \mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ are lattice vectors (combination of Bravais lattice vectors) that translates to equivalent positions, the atom at $\mathbf{x} = \mathbf{a}_0$ would be sitting on a plane with $d_0 \equiv \mathbf{n} \cdot \mathbf{a}_0$, the atom at $\mathbf{x} = \mathbf{a}_1$ would be sitting on a plane with $d_1 \equiv \mathbf{n} \cdot \mathbf{a}_1$, the atom at $\mathbf{x} = \mathbf{a}_2$ would be sitting on a plane with $d_2 \equiv \mathbf{n} \cdot \mathbf{a}_2$, with constraint:

$$d_0 + d_1 + d_2 \;=\; 0 \tag{1.7}$$

due to (1.3). According to the "intercept rule" in labeling planes, one takes the minimum plane for a particular $\mathbf{n}$, and looks where it intercepts along the the $\mathbf{a}_0$ direction, with

$$\frac{d_0}{d_{\rm min}} \;\equiv\; m_0 \quad\leftrightarrow\quad d_{\rm min} \;=\; \frac{d_0}{m_0} \tag{1.8}$$

and m_0 being an integer (could be zero), since atom at $\mathbf{x} = \mathbf{a}_0$ should be equivalent to atoms on the minimum plane. Similarly,

$$\frac{d_1}{d_{\rm min}} \;\equiv\; m_1, \quad \frac{d_2}{d_{\rm min}} \;\equiv\; m_2, \tag{1.9}$$

with

$$m_0 + m_1 + m_2 \;=\; 0. \tag{1.10}$$

So in HCP crystals, one uses $(m_0 m_1 m_2 m_3)$ to label planes, with the understanding that there is one slaved variable, $m_0 = -m_1 - m_2$.

Suppose someone tells you the four integers $(m_0 m_1 m_2 m_3)$, how to find $\mathbf{n}$ in Cartesian coor-

dinates? One can first set up the linear algebra equation:

$$\begin{pmatrix} a_{0x} & a_{0y} & a_{0z} \\ a_{1x} & a_{1y} & a_{1z} \\ a_{2x} & a_{2y} & a_{2z} \\ a_{3x} & a_{3y} & a_{3z} \end{pmatrix} \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = d_{\min} \begin{bmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \end{bmatrix} \tag{1.11}$$

But since the first row vector of the 4×3 matrix is a linear combination of the second and third row vectors, the first equation is redundant, so we only need to solve

$$\begin{pmatrix} a_{1x} & a_{1y} & a_{1z} \\ a_{2x} & a_{2y} & a_{2z} \\ a_{3x} & a_{3y} & a_{3z} \end{pmatrix} \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = d_{\min} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix} \tag{1.12}$$

One does not know $d_{\min}$ at the beginning, but it does not matter, since one can first solve for an unnormalized $\tilde{\mathbf{n}}$ first,

$$\tilde{\mathbf{n}} = \begin{pmatrix} a_{1x} & a_{1y} & a_{1z} \\ a_{2x} & a_{2y} & a_{2z} \\ a_{3x} & a_{3y} & a_{3z} \end{pmatrix}^{-1} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix} \equiv \mathbf{G} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \end{bmatrix} \tag{1.13}$$

and then normalize it later (in fact, this is how one finds $d_{\min}$):

$$\mathbf{n} = \frac{\tilde{\mathbf{n}}}{|\tilde{\mathbf{n}}|}. \tag{1.14}$$

BTW, the column vectors of the 3×3 matrix $\mathbf{G}$ is called the reciprocal vector (sometimes people multiply 2π on it),

$$\mathbf{G} = (\mathbf{g}_1|\mathbf{g}_2|\mathbf{g}_3), \tag{1.15}$$

since

$$\begin{pmatrix} a_{1x} & a_{1y} & a_{1z} \\ a_{2x} & a_{2y} & a_{2z} \\ a_{3x} & a_{3y} & a_{3z} \end{pmatrix} (\mathbf{g}_1|\mathbf{g}_2|\mathbf{g}_3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.16}$$

and we see from (1.13) that

$$\tilde{\mathbf{n}} = m_1\mathbf{g}_1 + m_2\mathbf{g}_2 + m_3\mathbf{g}_3, \quad \mathbf{n} = \frac{m_1\mathbf{g}_1 + m_2\mathbf{g}_2 + m_3\mathbf{g}_3}{|m_1\mathbf{g}_1 + m_2\mathbf{g}_2 + m_3\mathbf{g}_3|}. \tag{1.17}$$

The above does not presume $\mathbf{a}_3 = \mathbf{c}$ is perpendicular to $\mathbf{a}_0, \mathbf{a}_1, \mathbf{a}_2$, or $\mathbf{a}_1$ and $\mathbf{a}_2$ extends 120°

angle. It only presumes $\mathbf{a}_0 + \mathbf{a}_1 + \mathbf{a}_2 = 0$. This is important because we would like to talk about crystallography (and do diffraction and imaging) on elastically strained lattice. Note that according to the so-called Cauchy-Born rule, if a macroscopic deformation gradient $\mathbf{J}$ is applied:

$$\Delta\tilde{\mathbf{x}} \;=\; \mathbf{J}\Delta\mathbf{x} \tag{1.18}$$

where $\mathbf{x}$ is an original macro-position in the body, and if there is no plasticity or phase transformation in the body, then any Bravais lattice vector would transform as

$$\tilde{\mathbf{a}}_i \;=\; \mathbf{J}\mathbf{a}_i. \tag{1.19}$$

Note that (1.19) only applies to the Bravais lattice vector, and not necessarily to the internal coordinates within a complex unit cell (if the unit cell contains ≥ 2 atom). Thus, in an elastically strained lattice,

$$\tilde{\mathbf{a}}_0 + \tilde{\mathbf{a}}_1 + \tilde{\mathbf{a}}_2 \;=\; \mathbf{J}(\mathbf{a}_0 + \mathbf{a}_1 + \mathbf{a}_2) \;=\; 0 \tag{1.20}$$

Even though the above was motivated by HCP metal, the above formulas are completely general for any Bravais crystals. Indeed, the vector form (1.16) is just

$$\mathbf{g}_i \cdot \mathbf{a}_j \;=\; \delta_{ij}, \quad i,j \in 1,2,3 \tag{1.21}$$

which is the fundamental property of reciprocal vectors $\mathbf{g}_1$,$\mathbf{g}_2$,$\mathbf{g}_3$. Also, it is easy to show from the definition $\mathbf{n} \cdot \mathbf{a}_i = m_i d_{\min}$ that

$$d_{\min} \;=\; \frac{1}{|m_1\mathbf{g}_1 + m_2\mathbf{g}_2 + m_3\mathbf{g}_3|}. \tag{1.22}$$

For arbitrarily strained hexagonal crystal, if we want to know whether a direction $[u_0u_1u_2v]$ lies in the plane $(m_0m_1m_2l)$ or not, we can compute

$$\begin{aligned}
& (u_0\mathbf{a}_0 + u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + v\mathbf{c}) \cdot (m_1\mathbf{g}_1 + m_2\mathbf{g}_2 + l\mathbf{g}_3) && (1.23)\\
= \;& ((u_1 - u_0)\mathbf{a}_1 + (u_2 - u_0)\mathbf{a}_2 + v\mathbf{c}) \cdot (m_1\mathbf{g}_1 + m_2\mathbf{g}_2 + l\mathbf{g}_3) && (1.24)\\
= \;& (u_1 - u_0)m_1 + (u_2 - u_0)m_2 + vl && (1.25)\\
= \;& u_1m_1 + u_2m_2 - u_0(m_1 + m_2) + vl && (1.26)\\
= \;& u_1m_1 + u_2m_2 + u_0m_0 + vl && (1.27)
\end{aligned}$$

is zero or not. So even though the hexagonal system uses a redundant coordinate system that is also non-orthogonal, we can perform **naive inner product on the two 4-vectors**. For example, if we want to know whether $[20\bar{2}2]$ is an in-plane direction of $(\bar{1}012)$ or not, we just simply compute $2\times\bar{1}+0\times 0+\bar{2}\times 1+2\times 2$ vanishes or not. ("**3** $\to$ **4D**")

People use $\langle\rangle$ to represent family of [] vectors, including all crystallographic symmetry operations. In a cubic crystal, for instance, the crystallographic symmetry operations involve permutation and mirror operations, thus

$$\langle 110\rangle \ : \ [110],[1\bar{1}0],[\bar{1}10],[\bar{1}\bar{1}0],[101],[10\bar{1}],[\bar{1}01],[\bar{1}0\bar{1}],[011],[01\bar{1}],[0\bar{1}1],[0\bar{1}\bar{1}] \tag{1.28}$$

which has $3\times 2\times 2=12$ family members (6 if inversion doesn't count as distinct direction).

$$\langle 123\rangle \ : \ [123],[132],[312],[213],[231],[321],... \tag{1.29}$$

which has $3!\times 2\times 2\times 2=48$ family members (24 if inversion doesn't count). Similarly, people use $\{\}$ to represent family of () plane inclinations. The family members couting of planes has the same rule as $[]\to\langle\rangle$.

"**3** $\to$ **2D**": Human retina is 2D and we are inherently good at manipulating 2D graphs. The stereographic projection is a technique to project 3D directions $\mathbf{n}$ onto a 2D chart:

$$\mathbf{s}(\mathbf{n}) \ \equiv \ \frac{\hat{\mathbf{n}}-(\hat{\mathbf{n}}\cdot\mathbf{m})\mathbf{m}}{1+|\mathbf{m}\cdot\hat{\mathbf{n}}|}, \quad \hat{\mathbf{n}}\equiv\frac{\mathbf{n}}{|\mathbf{n}|} \tag{1.30}$$

where $\mathbf{m}$ is the projection (or "line of sight") direction, which should be a normalized vector.

A simpler projection is

$$\mathbf{p}(\mathbf{n}) \ \equiv \ \hat{\mathbf{n}}-(\hat{\mathbf{n}}\cdot\mathbf{m})\mathbf{m} \tag{1.31}$$

without the denominator in (1.30). It is easily seen that both $\mathbf{p}(\mathbf{n})$ and $\mathbf{s}(\mathbf{n})$ lie in the unit circle for any $\mathbf{n}$. $\mathbf{p}(\mathbf{n})$ is just Parallel projection with a point light source at $\mathbf{x}_{\text{light}}=-\infty\mathbf{m}$, and projection screen at $\mathbf{x}_{\text{screen}}=\mathbf{m}$. $\mathbf{p}(\mathbf{n})$ converts any circle on the unit sphere to an ellipse on the screen with no cusp.

In contrast, $\mathbf{s}(\mathbf{n})$ also converts great circles on the unit sphere to an ellipse-like curve, but with two cusps, due to the discontinuity from the absolute value operator in the denominator of (1.30). For $\mathbf{n}\cdot\mathbf{m}\geq 0$, we have the light source at $\mathbf{x}_{\text{light}}=-\mathbf{m}$, and screen at origin; and for $\mathbf{n}\cdot\mathbf{m}<0$, we have the light source at $\mathbf{x}_{\text{light}}=+\mathbf{m}$, and screen at origin. (The light

actually hits the screen before it hits the sphere, so it's not a true projection screen).

Also, the mapping from $\mathbf{p} \to \hat{\mathbf{n}}$ or $\mathbf{s} \to \hat{\mathbf{n}}$ is almost unique (has a degeneracy of 2). The spherical surface $\mathbf{n}/|\mathbf{n}|$ has two hemispheres. We can label the forward hemispherical half ($\mathbf{n} \cdot \mathbf{m} \geq 0$) by filled symbol, and rear hemispherical half ($\mathbf{n} \cdot \mathbf{m} < 0$) by open symbol.

A crystallographic plane can either be represented by the set of $\{\mathbf{v}_i\}$ that satisfies $\mathbf{n} \cdot \mathbf{v}_i = 0$, where $\mathbf{v}_i$ is an in-plane separation vector, or by the plane normal vector $\mathbf{n}$ itself. In the former representation on stereogram, we get a solid-line half curve plus a dash-line half curve, with representative $\mathbf{s}(\mathbf{v}_i)$s labelled by "$[u_1u_2u_3]$" of the $\mathbf{v}_i$'s. In the latter representation, we just plot $\mathbf{n}$ itself, with a label "$(m_1m_2m_3)$ pole". The two representations are equivalent. A pole with two-fold rotational symmetry is called diad, represented by an ellipse. A pole with three-fold symmetry is called triad, represented by a triangle. A pole with four-fold rotational symmtry is called tetrad, represented by a square.

In microscopy, a **zone axis z** is a direction, that defines a bunch of planes $\{\mathbf{n}_i\}$ by $\mathbf{n}_i \cdot \mathbf{z} = 0$, sort of turning the normal definition around. One can label a zone axis by "$[u_1u_2u_3]$ ZA direction", or by the set of planes $\mathbf{n}_i$, each using the first or second representation of planes. Take a $\mathbf{z} = [111]$ zone axis (one direction of the $\langle 111 \rangle$ family, that consists of [111], $[\bar{1}11]$, $[1\bar{1}1]$, $[11\bar{1}]$, $[\bar{1}\bar{1}\bar{1}]$, $[1\bar{1}\bar{1}]$, $[\bar{1}1\bar{1}]$, $[\bar{1}\bar{1}1]$). There are three from the {110} planes ($(1\bar{1}0)$, $(10\bar{1})$, $(01\bar{1})$), three from the {112} planes ($(1\bar{2}1)$, $(11\bar{2})$, $(2\bar{1}\bar{1})$), six from the {123} planes ($(12\bar{3})$, $(21\bar{3})$, $(1\bar{3}2)$, $(2\bar{3}1)$, $(3\bar{1}\bar{2})$, $(3\bar{2}\bar{1})$), and their inversions which belong to the rear hemisphere, that falls into the zone.

The significance of the above is that when an incoming X-ray or e-beam has $\mathbf{k}_{\text{in}} \parallel \mathbf{e}_z$, the diffraction condition is

$$(\mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}}) \cdot (\mathbf{x}_l^i - \mathbf{x}_m^j) \;=\; (l - m)2\pi n \tag{1.32}$$

where $\mathbf{k}_{\text{out}}$ is the scattered wave vector, $\{\mathbf{x}_l^i\}$ are atoms that belong to plane number l, and $\{\mathbf{x}_m^j\}$ are atoms that belong to plane number m (both having the same normal $\mathbf{n}$). Eqn. 1.32 is the condition where constructive interference of waves occur. From Eqn. 1.32, we see that

$$(\mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}}) \cdot (\mathbf{x}_l^i - \mathbf{x}_l^{i'}) \;=\; 0 \tag{1.33}$$

or

$$\mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}} \;\parallel\; \mathbf{n} \tag{1.34}$$

For elastic scattering,

$$|\mathbf{k}_{\text{out}}| \;=\; |\mathbf{k}_{\text{in}}| \tag{1.35}$$

(the Ewald's sphere), and for small-angle scattering (in TEM for example, a 200 keV electron has wavelength 0.025079 Å, which is much smaller than interplanar spacing), $\mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}}$ is almost perpendicular to $\mathbf{k}_{\text{in}}$, so the set of $\mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}}$ that gives strong diffraction on the screen are very close to the normal directions of the crystal planes that belong to the zone axis.

BTW people usually denote the angle between $\mathbf{k}_{\text{out}}$ from $\mathbf{k}_{\text{in}}$ as 2θ, where θ is called the Bragg angle. (If we fix $\mathbf{k}_{\text{in}} \parallel \mathbf{e}_z$, the above would have factor of 2 difference in labeling $\hat{\mathbf{k}}_{\text{out}}$ from the typical spherical coordinate label (r, θ, ϕ) convention: $\theta \to 2\theta$). We see that in order for constructive interference to between scattered waves of equivalent atoms:

$$\sum_j \exp(i(\mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}}) \cdot \Delta\mathbf{x}_j) \quad \to \quad (\mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}}) \cdot \Delta\mathbf{x}_j = 2\pi N \tag{1.36}$$

there needs to be

$$\mathbf{q} \equiv \mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}} \;=\; 2\pi(m_1\mathbf{g}_1 + m_2\mathbf{g}_2 + m_3\mathbf{g}_3)n \tag{1.37}$$

or

$$|\mathbf{q}| \;=\; \frac{2\pi n}{d_{\text{min}}} \tag{1.38}$$

but since $|\mathbf{q}| = 2\sin(\theta)|\mathbf{k}_{\text{in}}|$, we get

$$2\sin(\theta)|\mathbf{k}_{\text{in}}| \;=\; \frac{2\pi n}{d_{\text{min}}} \;\to\; 2\sin(\theta)|\hbar\mathbf{k}_{\text{in}}| \;=\; \frac{nh}{d_{\text{min}}} \;\to\; 2\sin(\theta)d_{\text{min}} \;=\; n\frac{h}{|\mathbf{p}_{\text{in}}|} \;=\; n\lambda \tag{1.39}$$

where we used $p = \hbar k$ and the de Broglie relation $\lambda = h/p$ that holds even for relativistic electrons, to get the famous Bragg's law:

$$d_{\text{min}} \;=\; \frac{n\lambda}{2\sin(\theta)} \tag{1.40}$$

where λ is the incoming and outgoing electron wavelength.

Just like we need grids on a Cartesian graph paper, we also needs grids on a stereogram to read out the approximate angular values for any $\hat{\mathbf{n}}$ plotted on the stereogram. This is called a Wulff net, typically with $\Delta\theta = \Delta\phi = 2^\circ$. θ is the polar angle (latitude, constant-latitude lines are called parallels), ϕ is the azimuthal angle (longitude, constant-longitude lines are

called meridians). With θ, ϕ read out, we can reconstruct

$$\hat{\mathbf{n}} \;=\; (\sin\theta)(\sin\phi)\mathbf{e}_x + (\cos\theta)\mathbf{e}_y + (\sin\theta)(\cos\phi)\mathbf{m} \tag{1.41}$$

where $\theta = 0$ is the "north pole" and $\phi = 0$ is the central vertical line. Also, we would have

$$\mathbf{s}(\mathbf{n}) \;=\; \frac{(\sin\theta)(\sin\phi)\mathbf{e}_x + (\cos\theta)\mathbf{e}_y}{1 + |\sin\theta\cos\phi|} \tag{1.42}$$

The beauty of stereographic projection $\mathbf{s}(\mathbf{n})$ is that angle is preserved from spherical surface ("3D") to 2D, the so called "conformal" property. Consider a move on the forward hemisphere:

$$\begin{aligned} d\mathbf{s} \;&=\; \frac{d\hat{\mathbf{n}} - (d\hat{\mathbf{n}}\cdot\mathbf{m})\mathbf{m}}{1+\mathbf{m}\cdot\hat{\mathbf{n}}} - \frac{(\hat{\mathbf{n}} - (\hat{\mathbf{n}}\cdot\mathbf{m})\mathbf{m})(\mathbf{m}\cdot d\hat{\mathbf{n}})}{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^2} \\ &=\; \frac{d\hat{\mathbf{n}} - (d\hat{\mathbf{n}}\cdot\mathbf{m})\mathbf{m} + (\mathbf{m}\cdot\hat{\mathbf{n}})d\hat{\mathbf{n}} - (\mathbf{m}\cdot\hat{\mathbf{n}})(d\hat{\mathbf{n}}\cdot\mathbf{m})\mathbf{m} - (\hat{\mathbf{n}} - (\hat{\mathbf{n}}\cdot\mathbf{m})\mathbf{m})(\mathbf{m}\cdot d\hat{\mathbf{n}})}{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^2} \\ &=\; \frac{d\hat{\mathbf{n}} - (d\hat{\mathbf{n}}\cdot\mathbf{m})\mathbf{m} + (\mathbf{m}\cdot\hat{\mathbf{n}})d\hat{\mathbf{n}} - \hat{\mathbf{n}}(\mathbf{m}\cdot d\hat{\mathbf{n}})}{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^2} \\ &=\; \frac{(1+(\mathbf{m}\cdot\hat{\mathbf{n}}))d\hat{\mathbf{n}} - (d\hat{\mathbf{n}}\cdot\mathbf{m})\mathbf{m} - \hat{\mathbf{n}}(\mathbf{m}\cdot d\hat{\mathbf{n}})}{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^2} \end{aligned} \tag{1.43}$$

Now consider two moves $\hat{\mathbf{n}} \to \hat{\mathbf{n}} + d\hat{\mathbf{n}}_1$ and $\hat{\mathbf{n}} \to \hat{\mathbf{n}} + d\hat{\mathbf{n}}_2$, thus

$$\begin{aligned} d\mathbf{s}_1 \cdot d\mathbf{s}_1 \;&=\; \frac{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^2 d\hat{\mathbf{n}}_1\cdot d\hat{\mathbf{n}}_1 + 2(\mathbf{m}\cdot d\hat{\mathbf{n}}_1)^2 + 2(d\hat{\mathbf{n}}_1\cdot\mathbf{m})^2(\mathbf{m}\cdot\hat{\mathbf{n}}) - 2(1+\mathbf{m}\cdot\hat{\mathbf{n}})(d\hat{\mathbf{n}}_1\cdot\mathbf{m})^2}{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^4} \\ &=\; \frac{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^2 d\hat{\mathbf{n}}_1\cdot d\hat{\mathbf{n}}_1}{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^4} \end{aligned} \tag{1.44}$$

as $d\hat{\mathbf{n}}_1 \perp \hat{\mathbf{n}}$, and also $d\hat{\mathbf{n}}_2 \perp \hat{\mathbf{n}}$

$$d\mathbf{s}_2 \cdot d\mathbf{s}_2 = \frac{(1+(\mathbf{m}\cdot\hat{\mathbf{n}}))^2 d\hat{\mathbf{n}}_2\cdot d\hat{\mathbf{n}}_2}{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^4} \tag{1.45}$$

$$\begin{aligned} d\mathbf{s}_1 \cdot d\mathbf{s}_2 \;=\; &\frac{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^2 d\hat{\mathbf{n}}_1\cdot d\hat{\mathbf{n}}_2 + (\mathbf{m}\cdot d\hat{\mathbf{n}}_1)(\mathbf{m}\cdot d\hat{\mathbf{n}}_2) + (\mathbf{m}\cdot d\hat{\mathbf{n}}_1)(\mathbf{m}\cdot d\hat{\mathbf{n}}_2)}{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^4} + \\ &\frac{2(d\hat{\mathbf{n}}_1\cdot\mathbf{m})(d\hat{\mathbf{n}}_2\cdot\mathbf{m})(\mathbf{m}\cdot\hat{\mathbf{n}}) - 2(1+\mathbf{m}\cdot\hat{\mathbf{n}})(d\hat{\mathbf{n}}_1\cdot\mathbf{m})(d\hat{\mathbf{n}}_2\cdot\mathbf{m})}{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^4} \end{aligned}$$

$$= \quad \frac{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^2 d\hat{\mathbf{n}}_1 \cdot d\hat{\mathbf{n}}_1}{(1+\mathbf{m}\cdot\hat{\mathbf{n}})^4}. \tag{1.46}$$

Thus,

$$\cos(d\hat{\mathbf{n}}_1, d\hat{\mathbf{n}}_2) \equiv \frac{d\hat{\mathbf{n}}_1 \cdot d\hat{\mathbf{n}}_2}{|d\hat{\mathbf{n}}_1||d\hat{\mathbf{n}}_2|} \;=\; \frac{d\mathbf{s}_1 \cdot d\mathbf{s}_2}{|d\mathbf{s}_1||d\hat{\mathbf{s}}_2|} \equiv \cos(d\mathbf{s}_1, d\mathbf{s}_2) \tag{1.47}$$

QED. The above proof is not so transparent, the steps below are more "geometric":

$$d\mathbf{s}(\mathbf{n}) \;=\; \frac{\mathbf{p}(d\mathbf{n}) - (\mathbf{m}\cdot d\mathbf{n})\mathbf{s}(\mathbf{n})}{1+\mathbf{m}\cdot\mathbf{n}} \;=\; \frac{d\mathbf{n} - (\mathbf{m}\cdot d\mathbf{n})(\mathbf{m}+\mathbf{s}(\mathbf{n}))}{1+\mathbf{m}\cdot\mathbf{n}} \tag{1.48}$$

we then note that $\cos(d\hat{\mathbf{n}}_1, d\hat{\mathbf{n}}_2)$ does not depend on the amplitude of $d\hat{\mathbf{n}}_1$, $d\hat{\mathbf{n}}_2$, so we can drop the denominator below. Furthermore, by definition

$$\mathbf{m}+\mathbf{s}(\mathbf{n}) \;=\; \frac{\mathbf{m}+\mathbf{n}}{1+\mathbf{m}\cdot\mathbf{n}} \tag{1.49}$$

So the Wulff net which maps parallels and meridians all have 90° angle at crossings, just like on the spherical surface.

Take a cubic crystal, if we plot all the $\langle 100\rangle$ (degeneracy=6), $\langle 111\rangle$ (degeneracy=8), and $\langle 110\rangle$ (degeneracy=12) on the stereogram, we find we divide the 4π solid angle into 48 equivalent patches. Each patch is a "triangle", with one vertex from each of the $\langle 100\rangle$, $\langle 111\rangle$, $\langle 110\rangle$ families.[3] Thus, when we would like to tabulate the response of a single crystal of cubic symmetry to an external input with direction $\mathbf{n}$, for example uniaxial tensile stress of the form $\sigma\mathbf{n}\mathbf{n}^T$, or the dielectric polarization due to electric field $E\mathbf{n}$, there is no need to collect data for the entire 4π solid angle. We only need to collect and parametrize data in the $4\pi/48$ solid angle, which is called the standard stereographic triangle that is [100]-[110]-[111] connected by geodesic lines (part of the great circle or "flight path"). The response, when $\mathbf{n}$ falls into any of the other 47 triangular patches, can be simply related to what happens in the standard triangle by symmetry permutations.

Anisotropy means some material response function, for example electrical resistivity $r(\mathbf{n})$, depend on $\mathbf{n}$, i.e. directional nature of the applied probe [4]. For a cubic crystal, this means angular non-uniformity within the standard stereographic triangle. For example, BCC iron is the most easy to magnetize along $\langle 100\rangle$ (the magnetic field or induction $\mathbf{B}$ rises the fastest

[3] The so-called rule of addition works for these vertices, that is, [111] is on the line connecting [001] and [110] on the stereogram, so [111] = [001] + [110]

[4] There is a definition using tensors[4], but we use scalar response for pedagogical simplicity.

with the magnetizing field $\mathbf{H}$ in that direction), so when making transformer cores, one should aim to align the $\langle 100 \rangle$ direction with the magnetic flux lines, which reduces magnetic hysteresis and improves energy efficiency of transformers. In FCC nickel, however, the easiest-to-magnetize direction is $\langle 111 \rangle$ family. So, anisotropy is generally crystal-structure and materials dependent.

In real applications, it is often difficult to come up with large single crystal. It is much easier to produce polycrystals, which is aggregate of many single crystals. The lattice orientation of a single crystal consists of three angular degrees of freedom, since a rotational matrix in 3D:

$$\tilde{\mathbf{x}} \;=\; \mathbf{R}\mathbf{x} \tag{1.50}$$

has three degrees of freedom, due to the constraints imposed by $\mathbf{R}^T\mathbf{R} = \mathbf{R}^T\mathbf{R} = \mathbf{I}$. $\mathbf{R}$ may be mapped onto three Euler angles (α, β, γ), in a decomposition:

$$\mathbf{R} \;=\; \mathbf{r}(\mathbf{e}_{z'}, \gamma)\mathbf{r}(\mathbf{e}_{x'}, \beta)\mathbf{r}(\mathbf{e}_z, \alpha), \tag{1.51}$$

where $\mathbf{r}(\mathbf{e}_z, \alpha)$ means rotating about $\mathbf{e}_z$ axis by angle α, $\mathbf{e}_{x'}$ is the new position of $\mathbf{e}_x$ after the first operation, and $\mathbf{e}_{z'}$ is the new position of $\mathbf{e}_z$ after the second operation. So we have

$$(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z) \;\to\; (\mathbf{e}_{x'}, \mathbf{e}_{y'}, \mathbf{e}_z) \;\to\; (\mathbf{e}_{x'}, \mathbf{e}_{y''}, \mathbf{e}_{z'}) \;\to\; (\mathbf{e}_{x''}, \mathbf{e}_{y'''}, \mathbf{e}_{z'}) \tag{1.52}$$

Note that the order of α, β, γ is important. Also, there is a theorem that any $\mathbf{R}$ can be written as $\mathbf{r}(\mathbf{n}, \theta)$, where $\mathbf{n}$ has two degrees of freedom, and θ has one.[5]

Thus, any crystalline grain's lattice orientation embedded in a polycrystal can be specified by these three numbers (with respect to a reference crystal). A random rotation is specified by

$$dP = f_0(\alpha, \beta, \gamma)d\alpha d\beta d\gamma \;=\; \frac{d\alpha}{2\pi} \cdot \frac{d\cos\beta}{2} \cdot \frac{d\gamma}{2\pi} \;=\; \frac{|\sin\beta|}{8\pi^2} d\alpha d\beta d\gamma \tag{1.53}$$

where α is uniformly distributed in $[0, 2\pi)$, $\cos\beta$ is uniformly distributed in $[-1, 1]$, and γ is uniformly distributed in $[0, 2\pi)$. The reason for the "special treatment" of β is that we notice that $\mathbf{e}_z$ undergoes only one transformation in (1.52), and we need to make sure $\mathbf{e}_{z'}$ covers the 4π spherical angle evenly, since although $(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z)$ are treated procedurally differently in (1.52), they must end up physically equivalent (for example the $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ could be [100],

[5] The real polynomial $\det|\mathbf{R} - \lambda\mathbf{I}| = 0$ has three roots λ_1, λ_2, λ_3, with $|\lambda_i|^2 = 1$. So one of $\{\lambda_i\}$ must be ± 1, while the other two could be complex conjugates. The -1 root possibly is due to inversion, which flips the right-handedness of the three axis, and therefore can be separated out from the operation. If $\mathbf{R}$ is differentially produced that preserves the right-handedness, the -1 root cannot be there.

[010] and [001] of a cubic lattice), and thus "equally random", which means $\mathbf{e}_z = [001]$ should cover the sphere evenly.

Texture means the distribution of crystal orientation differs from the random orientation:

$$f(\alpha, \beta, \gamma) \neq f_0(\alpha, \beta, \gamma) \tag{1.54}$$

and there are preferred lattice orientations in the polycrystal. In wire drawing, where one pulls the metal through a die, a $\langle 110 \rangle$ wire texture could develop in BCC iron, which means the grains tend to have one of their $\langle 110 \rangle \parallel$ the wire drawing direction. This happens through a complex, multi-step process, where severe plastic deformation and dislocation storage first occurs, followed by so-called dynamic recovery / recrystallization. We will address how texture forms by plastic deformation later in the course. Note here that even if $\langle 110 \rangle$ is fixed (2 degrees of freedom), the lattice orientation still has a random degree of freedom, in that the grain can rotate around its $\langle 110 \rangle$ axis.

If one rolls a BCC polycrystal into a sheet (rolling direction (RD) + transverse direction (TD) in the sheet plane, plus rolling plane normal (ND)), the favoable lattice orientation is either (A) $\langle 110 \rangle \parallel$ RD, or (B) $\langle 100 \rangle \parallel$ RD. In both (A) and (B), $\langle 001 \rangle \parallel$ ND. Clearly, for making transformer core by repeatedly folding the sheets, (B) texture is better. For BCC Fe-4wt%Si soft magnet, a special process was developed to produce texture (B). Note that if one has "perfect" rolling texture (A) or (B), then one just get a single crystal. In reality, though, even in a strongly textured polycrystal, there is likely $\pm 10^\circ$ variance in adjacent grains.

In recent years, so-called grain orientation imaging using automated electron backscatter diffraction (EBSD) [5, 6, 7] was developed and widely used, which can identify $\alpha_i, \beta_i, \gamma_i$ as well as grain size D_i and shape, for individual grains $\{i\}$ on the surface. Thousands of contiguous grains can be characterized automatically, which can then be analyzed into single-grain statistics, or even multi-grain correlations (for example the grain boundary misorientation is a 2-grain correlation, and grain boundary path/network is a multi-grain correlation[8]). To characterize texture, which is a single-grain statistics, so-called pole figure is developed, which are stereograms with the line of sight $\mathbf{m} \parallel$ ND, RD is $\mathbf{e}_y$, and $\mathbf{e}_x$ is TD. We note this is different from the standard stereographic triangle where the reference directions are the crystallographic directions, and external probing $\mathbf{n}$ are plotted in reference to them ("**properties**" anisotropy of single crystal). Here, the reference directions are the processing geometry, while what are plotted are the crystallographic directions of an assmebly of

polycrystals, since the crystal orientation themselves are changing. This contrast is because the texture pole figures are for characterizing "**processing→structures**" anisotropy. Thus, from the stereogram tools, which is used to characterize anisotropy in 3D, we can already see "properties", and "processing→structures", which are the fundamental ingredients of materials science.

In this chapter, we had a flavor of atomic structure and polycrystalline structure. This is only the beginning - there are many fascinating "structures" in metallurgy, that await us in this course.

Chapter 2

Metallic Bonding, Ideal Strength and the Dislocations Machinery

People's impression of metals are shiny (lights cannot transmit through, much is reflected), malleable objects, and cool when touched (relatively high thermal conductivity κ [W/m/K]). If they measure the electrical conductivity σ [1/Ohm/m], they will also find it to be high. These behaviors are fundamentally connected. A high electrical conductivity means low- and medium-frequency electromagnetic fields will be strongly screened by the free, mobile electrons inside the metal, and thus cannot penetrate far. From an energy band point of view, this also makes sense since zero band gap means arbitrarily low-frequency photons can be absorbed (and maybe re-emitted later). The higher thermal conductivity is related to electrical conductivity through the Wiedemann-Franz law

$$\kappa \approx T\sigma L, \quad L = \frac{\pi^2}{3}\left(\frac{k_{\mathrm{B}}}{e}\right)^2 = 2.44 \times 10^{-8}\mathrm{W/Ohm/K}^{-2} \tag{2.1}$$

which holds reasonably well for simple metals, when the main charge carriers are also the main heat carriers. The last property, that they are malleable (relatively speaking, and varying between metals), originates from metallic bonding (vs. covalent and ionic bonding), which also has to do with the aforementioned delocalized, free electrons.

Before we delve into the nature of metallic bonding, and the nature of malleability or ductility, I just want to mention that the mere existence of shiny solids on Earth is kind of a miracle! This is because metals themselves are quite reactive with oxygen gas molecules,

which surround the Earth. In fact, the standard Gibbs free energy of formation of Fe_2O_3:

$$2\text{Fe(bcc)} + \frac{3}{2}\text{O}_2\text{(g)} \;=\; \text{Fe}_2\text{O}_3(\text{rhombohedral }\; ''\text{hematite}'') \tag{2.2}$$

is $\Delta G^0 = -740$ kJ/mol at $T = 298.15$K and $P(O_2) = 1$atm, or -3.8 eV per Fe atom! To appreciate the magnitude of this, most liquid→solid and solid→solid phase transformations, such as solidification and martensitic transformations, have driving force less than 0.05eV per atom. The -3.8 eV per Fe atom is a huge thermodynamic driving force, Damocles sword up the neck of metallic Fe so to speak, that is not going away any time soon! Our stone-age ancestors saw trees, earth, rocks, clouds and animals, but in their daily lives they did not see much **shiny, opaque, ductile** solids around. That's perhaps why they were fascianted by gold, which is the most stable of metallic elements chemically, as reflected in the electromotive force series, which characterize the willingness of metals to give up their free electrons and turn into a hydrated cation[1]. As metallurgy and human civilization advances, more and more shiny opaque strong and tough objects are produced, so now we are surrounded by them. However, if not tended to, the shiny facade of the Stata Center (made of titanium), the metal trim of your iphone, WILL oxidize away after many many years. So, the whole concept of metallurgy, the usage of metal in human civilization is in fact **exploiting the metastable state of matter**! Corrosion, which is the chemo-mechanical degradations of matter including oxidation reactions, is forever the enemy of most metals. Fortunately, metals developed a defense mechanism called "parabolic kinetics", that slows down (but not stops) the advance of oxidation with time. This miracle of the **long metastability** of metal, which has a kinetic origin, is called passivation.[9] A thin passivation layer develops on the surfaces of metals, that are barriers against oxygen transporting inward, or metal transporting outward to meet oxygen. So it is this thin passivation layer of tens of nanometers (the oxidation layer can be microns or even mm scale, but the region that offers true resistance, the region that is truely atomically compact, only needs to be tens of nanometers) that saves gigantic metal structures of our world that runs cm to meter or even kilometer lengthscale!

If all works out according to the plan, metals should have nothing to worry about within a few millennia. But "parabolic kinetics" does not always work, due to for instance localized pitting corrsion [10] and "friable" oxide, which has a lot of cracks and whose thickness grows linearly with time. And also this skin could be quite fragile mechanically, so when there is

[1]The electronegativity, as characterized by the standard potential, tends to increases from left to right on the periodic table for the same period.

a crack, and stress is applied and amplified at the crack-tip, the skin could be broken, and then the material could fail by **combined** chemical and mechanical attack, which manifest as **stress-corrosion cracking** for instance.

It is not always O_2(g) that is the problem. The "environment" that the metal is embedded in could contain all kinds of agents that is bad for the metal. In Fukushima Daiichi nuclear disaster, the following reaction

$$\mathrm{Zr(hcp)} + 2\mathrm{H_2O(l)} \;=\; \mathrm{ZrO_2(s)} + 2\mathrm{H_2(g)} \tag{2.3}$$

happened, as Zr is more active than H_2 in the electromotive series[2]. The reaction also happens under normal reactor operating temperatures, but is greatly accelerated at high temperatures. So after Fukushima, nuclear metallurgists are developing accident-tolerant nuclear fuel claddings, that does not react with the environment rapidly at high temperatures.

Now coming to the main theme of mechanical response. Metals have valence electrons (incomplete shells) that are loosely bound to the nuclei. When we assemble metal atoms into condensed matter by bringing them into proximity, these valence electrons start to meander between ions, even without thermal agitation (electronic temperature T_e=0K), since there is enough benefit in quantum kinetic energy $\langle\psi|-\nabla^2|\psi\rangle$ not to be constrained around one ion and being able to propagate around.[3] These "itinerant" electrons are shared between many ions, and do not "belong" to a particular ion. (Metallic vs ionic/covalent bonding is like Communism vs GOPism for bachelors and married couples). Under an external electric field, these itinerant electrons can freely flow, giving high electrical and thermal conductivities. Like a glue, this sea of itinerant electrons provide **bonding**, which is to say bringing the atoms together reduced the total energy compared to the individual isolated atomic states:

$$E_\mathrm{b}(N) \;\equiv\; E_\mathrm{together}(N) - N e_\mathrm{isolated} \;<\; 0 \tag{2.4}$$

where N is the number of atoms. We can also define binding energy or cohesive enery per particle:

$$e_\mathrm{b} \;\equiv\; \lim_{N\to\infty} \frac{E_\mathrm{b}(N)}{N} \tag{2.5}$$

[2]The standard potential is $U_0 = -1.45$V, meaning 1.45eV per electron more acive than hydrogen in http://en.wikipedia.org/wiki/Standard_electrode_potential_(data_page). Minor alloying and/or high-temperature structural phase transformations will not change the ball-park value from HCP Zr.

[3]No thermal fluctuation is needed, since quantum fluctuation and Pauli exclusion is already sufficient to delocalize.

where the surface contribution is filtered out by the **large-number limit**. e_b of course is **crystal structure** and **lattice constant (elastic strain)** dependent. In FCC Cu, $e_\mathrm{b} = -3.54$ eV/atom. -3.54 eV/atom can be considered to the energy gain of embedding the Cu core ion in the electron glue, after donating its valence electrons to the glue also.

The usual way people described metallic bonding is the embedded-atom model [11]:

$$U(\{\mathbf{x}_i\}) \;=\; \sum_i \frac{\sum_{j\neq i} u(r_{ij})}{2} + F_i(\sum_{j\neq i} \rho(r_{ij})), \tag{2.6}$$

where $\rho(r_{ij})$ is the "glue projection" function, and F_i is the "ion embedding" function. $U(\{\mathbf{x}_i\})$ is called the interatomic potential or the atomistic **potential energy landscape (PEL)**. The $u(r_{ij})$ is the pair and additive contributions like the simplest Lennard-Jones potentials, but the second term makes the many-body nature of bonding manifest. The embedding function, $F_i(\cdot)$, is often chosen to be $\sqrt{\cdot}$ in the so-called Finnis-Sinclair forms.[12] This provides a bonding energy benefit that scales as $-\sqrt{Z}$, where Z is the coordination number. This $-\sqrt{Z}$ form has a coordinate-saturation effect that stablizes lower-coordination crystal lattices such as BCC, relative to the FCC and HCP close-packed lattices. With a many-body potential form like (2.6), we can sum over lattice sites to obtain e_b for a given lattice structure geometry at $T = 0$. The plot of $e_\mathrm{b}(\Omega)$, where Ω is the atomic volume, is called the **cohesive energy curve**. This would allow us to compare the stability of different crystal structures at zero pressure, as well as at finite pressures (after adding the $+P\Omega$ term).

With $U(\{\mathbf{x}_i\})$, we can also calculate the total potential for an assmebly of **non-perfectly arranged** atoms (imagine thermal fluctuations of ion positions, aka phonons, and/or defects - a defect is defined by a set of atoms having atomic-neighbor relations or **bond topologies significantly different** from those in the perfect reference lattice - defects tend to have higher energy and sit in PEL's **metastable** energy basins), and run molecular dynamics (MD) simulations with it

$$m_i \frac{d^2\mathbf{x}_i}{dt^2} \;=\; -\frac{\partial U(\{\mathbf{x}_i\})}{\partial \mathbf{x}_i} \tag{2.7}$$

From a pure theorist point of view, (2.7) creates a complete "world", in the sense that all crystal and defect structures, their time evolutions and therefore thermomechanical properties can in principle be obtained by integrating (2.7) forward in time.[4]

[4]The practical computability is another matter. In this course, even though we do not teach how to implement (2.7) in the computer, we do want to ask people to think from the "atomistic world" perspective, which is one of the multiscale perspectives of thinking about materials.

It should also be cautioned that the "free electron glue" picture is only an idealization that works best for s- and p- valence electrons, which are the most delocalized. d-electrons are more spatially localized and also have a narrower energy distribution than s- and p-electrons, which makes transition metals harder to deal with from ab initio calculations. f-electron metals, such as rare-earth metals (such as La, Ce) and actinide metals (such as Pu), are even harder to deal with, due to even stronger electron localization and electron-electron correlation. Also, there is in fact some angular dependence in metallic bonding, even in sp-bonded simple metal like Aluminum [13]. This is the reason for the development of Modified Embedded-Atom Model (MEAM) [14] family of empirical interatomic potentials for metals. The form of MEAM is general enough that it may even be used to describe covalently bonded semiconductors, which is important for metallurgy if we want to consider impurities in metals, or compounds such as metal silicides.

The lattice at mechanical equilibrium at $T = 0$ is the result of

$$a_0(c_0, ...) \equiv \arg \min_{\text{structure}} e_{\text{b}}. \tag{2.8}$$

Elastic deformation is defined as "**small**", **reversible**, but **diffuse/delocalized** change to the Bravais lattice vectors $\{\mathbf{a}_i(\mathbf{x})\}$ of a perfect crystal, where $\mathbf{x}$ is a coarse-grained position inside the material. By definition, elastic deformation **excludes highly localized changes** in atomic geometry, which samples the **nonlinear nonconvex** part of the atomistic potential energy landscape (PEL). The elastic response can be probed by applying an external stress $\boldsymbol{\sigma}_{\text{ext}}$:

$$\min_{\boldsymbol{\epsilon}} e_{\text{b}}(\boldsymbol{\epsilon}) - \Omega \text{Tr}(\boldsymbol{\sigma}_{\text{ext}}\boldsymbol{\epsilon}) \quad \rightarrow \quad \boldsymbol{\sigma}_{\text{int}} \equiv \frac{1}{\Omega}\frac{\partial e_{\text{b}}(\boldsymbol{\epsilon})}{\partial \boldsymbol{\epsilon}} = \boldsymbol{\sigma}_{\text{ext}}. \tag{2.9}$$

We could define shear modulus G by

$$G \equiv \left.\frac{\partial \sigma_{\text{int}}^{\text{shear}}}{\partial \epsilon_{\text{shear}}}\right|_{\boldsymbol{\epsilon}=0} = \frac{1}{\Omega}\frac{\partial^2 e_{\text{b}}}{\partial \epsilon_{\text{shear}}^2} \tag{2.10}$$

and the bulk modulus B by

$$B \equiv \left.\frac{\partial \sigma_{\text{int}}^{\text{hydro}}}{\partial \epsilon_{\text{hydro}}}\right|_{\boldsymbol{\epsilon}=0} = \frac{1}{\Omega}\frac{\partial^2 e_{\text{b}}}{\partial \epsilon_{\text{hydro}}^2} \tag{2.11}$$

where ϵ_{shear} is the shear elastic strain (generally, we will use the engineering shear strain, for example $\epsilon_{\text{shear}} = \gamma_{xz} = \partial_x u_z + \partial_z u_x$), and ϵ_{hydro} is the hydrostatic invariant. For pedagogical

simplicity, we could envision a representative elastic deformation that looks like

$$\boldsymbol{\epsilon} \;=\; \begin{pmatrix} \frac{\epsilon_{\rm hydro}}{3} & 0 & \frac{\epsilon_{\rm shear}}{2} \\ 0 & \frac{\epsilon_{\rm hydro}}{3} & 0 \\ \frac{\epsilon_{\rm shear}}{2} & 0 & \frac{\epsilon_{\rm hydro}}{3} \end{pmatrix}. \tag{2.12}$$

In a crude sense, the elastic constants G and B (generally, c_{ijkl} tensor) characterize the **curvature** of the energy landscape with respect to **small, diffuse** changes to $\{\mathbf{a}_i(\mathbf{x})\}$ (the *elastic* strains):

$$e_{\rm b}(\epsilon_{\rm hydro}, \epsilon_{\rm shear}) \;=\; e_{\rm b}(0,0) + \frac{\Omega}{2}(G\epsilon_{\rm shear}^2 + B\epsilon_{\rm hydro}^2) + O(\epsilon^3) \tag{2.13}$$

At finite T, we just need to add $-Ts$ term to $e_{\rm b}$, and use $f_{\rm b} = e_{\rm b} - Ts = -N^{-1}k_{\rm B}T\ln Z$, the Helmholtz free energy of binding per particle, instead of $e_{\rm b}$. This is so-called thermoelasticity formalism. Z is the partition function in statistical mechanics.

In simple lattices, $\epsilon_{\rm hydro}$ stretches the bonds (changes the atomic distance $r_{ij} \equiv |\mathbf{x}_j - \mathbf{x}_i|$) without changing the bond angles θ_{ijk}, a 3-atom quantity. But $\epsilon_{\rm shear}$ changes the bond angles. As the names "free electron gas" or "Fermi liquid" (even at $T_{\rm e}$=0K) imply, people consider the aforementioned electron glue to be somewhat isotropic. The result is that metallic bonding tends to be more isotropic than covalent bonding, as reflected in lower shear modulus to bulk modulus ratio of most metals compared to semiconductors [15, 16]:

$$\frac{G}{B}(\text{metals}) \;<\; \frac{G}{B}(\text{Si, Ge, SiC, etc.}). \tag{2.14}$$

Metallic bonding is thus "**shear-weak**" or "**shear-soft**" compared to ceramics, all the way till spontaneous bond switching driven by shear, when the so-called ideal shear strength $\boldsymbol{\sigma}_{\rm ideal}$ is reached.

The ideal strength $\boldsymbol{\sigma}_{\rm ideal}$ is defined by the following thought experiment (Gedankenexperiment). Imagine a perfect crystal without any defects and at $T = 0$. Now we gradually elastically strain up the lattice, according to a path $\boldsymbol{\epsilon}(\lambda)$, which could be a simple straight line in the 6D strain space

$$\boldsymbol{\epsilon}(\lambda) \;=\; \lambda\boldsymbol{\epsilon}_0, \quad \lambda = [0, \lambda_{\rm C}), \tag{2.15}$$

at what point would a critical $\lambda_{\rm C}$ be reached, that the homogeneity of the lattice can no longer be maintained, and the deformation loses reversibility?

We could imagine that along the ϵ_{shear} axis, we can shear the bonds more and more, until at some point, the original set of nearest-neighbor bonds snap, or break spontaneously. Then we reach the ideal shear strength $\sigma_{\text{ideal}}^{\text{shear}}$ and ideal shear strain $\epsilon_{\text{ideal}}^{\text{shear}}$. We could also imagine that along the ϵ_{hydro} axis, we stretch the bonds more and more, until the nearest-neighbor bonds snap; then we reach the ideal hydrostatic tensile strength $\sigma_{\text{ideal}}^{\text{hydro}}$ and ideal hydrostatic tensile strain $\epsilon_{\text{ideal}}^{\text{hydro}}$. Generally speaking, $\boldsymbol{\epsilon}_{\text{ideal}}$ is a 5-dimensional surface in the 6-dimensional strain space. Moving the strain path $\boldsymbol{\epsilon}(\lambda)$ anyway inside the $\boldsymbol{\epsilon}_{\text{ideal}}$ surface is completely reversible - one can fully recover the perfect crystal upon unloading (all at 0K). But if the path ever touches the surface, BOOM!

Ab initio calculations can be used to calculate $\boldsymbol{\epsilon}_{\text{ideal}}$ and $\boldsymbol{\sigma}_{\text{ideal}}$. The results tend to be huge values [16]. For instance, BCC Fe has $\epsilon_{\text{ideal}}^{\text{shear}} = 0.178$ and $\sigma_{\text{ideal}}^{\text{shear}} = 8$ GPa. (Have you ever seen a piece of bulk Fe that can elastically shear 17% and sustain critical resolved shear stress (CRSS) of 8 GPa reversibly? The key, however, is the qualifier **bulk** Fe and what defects may be contained in your typical polycrystalline bulk Fe: dislocations, GBs, inclusions, surface damages, voids, outright microcracks....)

If you don't believe the numerical ab initio calculations, the large ideal strength can be still be justified on theoretical grounds. The renowned physicist Yakov Frenkel proposed the famous "Frenkel sinusoid"[17] in 1926. Imagine a material whose electron glue is local, i.e., its energy response only cares about the atomic plane immediately above, and the atomic plane immediately below. We can then perform the so-called generalized stacking fault (GSF) energy calculation, which characterize a sharp slip between two rigidly upright blocks of crystals. Let us define the slip displacement as $\mathbf{x}$ (note $\mathbf{x}$ does not mean position here!), and we can calculate the energy increase as the top plane rides above the bottom plane, $\Delta E_1(\mathbf{x})$, the subscript 1 denotes there is just one glue layer that is being sheared (between just two planes). Clearly, $\Delta E_1(\mathbf{x})$ is extensive quantity and needs to be normalized by the slip plane area A_0, and we can define an intensive quantity called one-layer GSF:

$$\gamma_1(\mathbf{x}) \;=\; \frac{\Delta E_1(\mathbf{x})}{A_0} \tag{2.16}$$

We note that $\gamma_1(\mathbf{x})$ resembles the most localized deformation possible in the vertical direction, **very distinct** from the elastic deformation before. We will address this difference later. Right now, however, focus on $\gamma_1(\mathbf{x})$, which has the unit of energy per area, same as the surface or interfacial energies (it is a kind of stacking fault). We note that $\gamma_1(\mathbf{x})$ must be a

periodic function:

$$\gamma_1(\mathbf{x}+\mathbf{b}) \;=\; \gamma_1(\mathbf{x}) \tag{2.17}$$

where $\mathbf{b}$ is a Bravais translational vector. And

$$\gamma_1(0) \;=\; \gamma_1(n\mathbf{b}) \;=\; 0 \tag{2.18}$$

where for a simple cubic solid, one likely have a very high energy for $\mathbf{x} \sim \mathbf{b}/2$, since we will have an energy saddle point. Thus, a most crude fitting form for the slip-shear response would be

$$\gamma_1(x) \;=\; \frac{\gamma^*}{2}[1-\cos\left(\frac{2\pi x}{b}\right)] \tag{2.19}$$

where γ^* is the unstable stacking energy. We can also define

$$\frac{d\gamma_1(x)}{dx} \;=\; \frac{\pi\gamma^*}{b}\sin\left(\frac{2\pi x}{b}\right), \tag{2.20}$$

which can be regarded as the **traction-displacement** response of the **local** electron glue. (The "metallic bonding" really comes from the electron glue, as we have seen before). $\frac{d\gamma_1(x)}{dx}$ has the unit of stress.

Now consider a series of constrained deformation, $E_2(\mathbf{x})$, $E_3(\mathbf{x})$, $E_4(\mathbf{x})$, ..., $E_n(\mathbf{x})$, where the deformation is more and more delocalized (diffuse) in the z-direction. But we can normalize the energy by n, the number of glue layers being sheared:

$$\gamma_n(\mathbf{x}) \;=\; \frac{\Delta E_n(\mathbf{x})}{nA_0}. \tag{2.21}$$

There is clearly also:

$$\gamma_n(\mathbf{x}+\mathbf{b}) \;=\; \gamma_n(\mathbf{x}) \tag{2.22}$$

and we can now directly compare intensive quantity $\gamma_n(\mathbf{x})$ with intensive quantity $\gamma_1(\mathbf{x})$. As it turns out, in FCC Cu,

$$\gamma_n(\mathbf{x}) \;\approx\; \gamma_1(\mathbf{x}) \tag{2.23}$$

indicating the electron glue in Cu is indeed quite local [18]. In FCC Al, $\gamma_n(\mathbf{x})$ and $\gamma_1(\mathbf{x})$ differ somewhat - the difference thus indicates the glue is not entirely local, there is some bond angle dependence in the energy which generate triple-layer interactions. Nontheless, $\gamma_n(\mathbf{x})$ (up to $\gamma_\infty(\mathbf{x})$, which characterizes elastic deformation) are of similar magnitude with

$\gamma_1(\mathbf{x})$. For pedagogical simplicity, let us pretend

$$\gamma_\infty(\mathbf{x}) \;=\; \gamma_1(\mathbf{x}) \tag{2.24}$$

and the electron glue is very local in this course.

From (2.9), we see that

$$\sigma_{\text{shear}} \;=\; \lim_{n\to\infty} \frac{1}{V_n}\frac{\partial E_n}{\partial \epsilon_{\text{shear}}} = \frac{1}{nA_0 d_0}\frac{\partial nA_0\gamma_n(x)}{\partial(x/d_0)} \;=\; \frac{d\gamma_\infty(x)}{dx} \tag{2.25}$$

From (2.24) and (2.20), we then get

$$\sigma_{\text{shear}} \;=\; \frac{d\gamma_1(x)}{dx} = \frac{\pi\gamma^*}{b}\sin\left(\frac{2\pi x}{b}\right). \tag{2.26}$$

From the very simple physical reasoning above, two conclusions can be drawn:

1. For small deformation, $x \ll b$,

$$\sigma_{\text{shear}} \;\approx\; \frac{\pi\gamma^*}{b}\frac{2\pi x}{b} = \frac{\pi\gamma^*}{b}\frac{2\pi\epsilon_{\text{shear}} d_0}{b} \tag{2.27}$$

   so we get

$$G \;=\; \frac{2\pi^2\gamma^* d_0}{b^2} \tag{2.28}$$

   or

$$\gamma^* \;=\; \frac{Gb^2}{2\pi^2 d_0}. \tag{2.29}$$

   as an estimate of the energy barrier (actually energy/area) for localized shear, or **slip**.

2. The peak shear stress is obtained at:

$$\sigma^{\text{ideal}}_{\text{shear}} \;=\; \frac{\pi\gamma^*}{b} \;=\; \frac{Gb}{2\pi d_0} \tag{2.30}$$

   when $x = b/4$ and

$$\epsilon^{\text{ideal}}_{\text{shear}} \;=\; \frac{b}{4d_0}. \tag{2.31}$$

   At this point,

$$\frac{\partial^2 E}{\partial \epsilon^2_{\text{shear}}} \;=\; 0, \tag{2.32}$$

and one enters into the non-convex region of the PEL. The local elastic stability is lost, and homogeneity of the lattice can no longer be maintained.

(2.30) is called the Frenkel ideal shear strength estimate. Generally speaking, in a simple metal, $b = |\mathbf{b}|$ is the nearest-neighbor distance. With respect to the reference atom on one plane, the adjacent plane below should also have one of its nearest neighbors, but the separation is not perfectly parallel to the plane normal $\mathbf{n}$, so there tends to be

$$b \ > \ d_0 \tag{2.33}$$

Thus, a reasonable estimate for the ideal shear strength is

$$\sigma_{\text{shear}}^{\text{ideal}} \ \approx \ \frac{G}{5} \tag{2.34}$$

from the Frenkel sinusoid model. However, as we have mentioned before, metals are "shear-soft", and the sinusiod is actually tilted [19] and peaks earlier than $b/4$, so a better approximation for metals might be

$$\sigma_{\text{shear}}^{\text{ideal}} \ \approx \ \frac{G}{10}. \tag{2.35}$$

Thus, if we take the $\{0001\}\langle 11\bar{2}0\rangle$ shear system of HCP Mg, $G = 19.2$ GPa, the ideal shear strength should be around 2 GPa, which is close to the density functional theory (DFT) calculated value of 1.84 GPa[15].

In an actual experiment on a bulk metal, say HCP Mg, what one gets is a plastically flowing metal at much lower stresses than the ideal strength:

$$\sigma \ = \ \sigma(\epsilon), \quad \epsilon = \dot{\epsilon} t \tag{2.36}$$

where a typical applied strain rate is $\dot{\epsilon} = 10^{-4}$/s. A notable bend occurs in the curve at $\sigma = \sigma_{\text{y}}$. People usually define σ_{y} by the "0.2% offset strain" rule. The rationale for this is that the unloading modulus is often a good (sometimes even better) estimate of the elastic modulus as the loading modulus, so if one imagines unloading, the amount of residual plastic strain at zero load would be 0.2%, which is small but measurable amount of sample-scale plasticity. Thus, the point of σ_{y} can be considered to have initiated measurable sample-scale plasticity, on top of whatever elasticity that have occurred. Hollomon's equation is

$$\sigma \ = \ K\epsilon_{\text{p}}^{n}, \tag{2.37}$$

where n is the (plastic) **strain hardening exponent** (between 0.1 and 0.5 for most metals), and $\epsilon_{\rm p}$ is the plastic strain component of the total applied strain

$$\epsilon \;=\; \epsilon_{\rm e} + \epsilon_{\rm p} \tag{2.38}$$

and $\epsilon_{\rm e}$ is the elastic component of the total applied strain. Also, for traditional macroscopic experiments, it is a very good approximation to have

$$\sigma \;=\; E\epsilon_{\rm e} \tag{2.39}$$

where E is the Young's modulus. Thus, combining the equations, we have

$$\sigma \;=\; K\left(\epsilon - \frac{\sigma}{E}\right)^n, \tag{2.40}$$

which gives the total stress-strain curve.

$\sigma_{\rm y}$ is very small for pure bulk Mg, if we align $(0001)_{\rm Mg}$ 45° to the uniaxial pulling direction. The contrast between $\sigma_{\rm y} \sim 0.7$ MPa and $\sigma_{\rm ideal}^{\rm shear} = 1.8$ GPa is really stark, off by a factor of more than 2000! Has Frenkel gone mad?

In 1934, G. I. Taylor [20], Egon Orowan [21] [5] and Michael Polanyi [22] simultaneously introduced the concept of dislocations, which resolve the paradox or discord between ideal strength and practically observed strength of bulk metals. If we regard Frenkel's estimate as pure physicists' answer to strength of crystals, the answer by Taylor, Orowan and Polanyi has more pessimistic realism in it, which is the typical view of material scientists. The 2000-fold difference is attributed to initial condition in the material, ie. *microstructures* or defects, namely dislocations. These dislocations are line defects that move inside the crystal, like crawling caterpillars or rolling carpet creases [23]. Dislocations are giant atomic-bond harvesting machines: as a dislocation core move in the crystal, it cuts some old bonds but also simultaneously stitches some new bonds together, promoting so-called **bond-switching** (not permanent **bond-loss** as in crack propagation), which is the essence of inelastic or plastic shear. Dislocations are not thermal-equilibrium defects: they must be generated by "beating".

Let us back up a little. Scientists in the 1800s have envisioned elastic distortions on aether [23] that contain localized defects. Anton Timpe [24] and Vito Volterra [25] indeed solved the elastic stress fields of these defects. Volterra further classified these line defects into six types

[5] Orowan was a professor of metallurgy at the MIT from 1950.

of *distorsioni*, three turns out to be dislocations, and three turns out to be disclinations. The dislocations are the 1D edges of a 2D translational fault ($\Delta\mathbf{x} = \mathbf{b}$ to $\Delta\mathbf{x} = 0$), or slip fault. The disclinations are the 1D edges of a 2D rotational fault ($\Delta\theta = 10°$, a grain boundary, to $\Delta\theta = 0$, no grain boundary). Disclinations are prohibitively expensive in 3D crystals, but they can exist in 2D crystals embedded in 3D [26, 27] and liquid crystals [28, 29].

Dislocations were first directly observed by transmission electron microscopy (TEM) by the team led by Sir Peter B. Hirsch at Oxford in 1956. [30] Thus, in this case, materials theory was ahead of experimentation by more than 20 years!

Dislocations are created to relax (gradually reduce) elastic strain energy. As previously mentioned, elastic strain energy is small (small amplitude) but diffuse (long wavelength) "pain" inside the crystal. The most common treatment of such small-amplitude, long-wavelength pain is so-called linear elasticity theory, where stress-strain relation is linear but energy is approximately by quadratic fitting of the bottom. Basically one attempts to fit the PEL by a quadratic expansion near the local minimum. But even if the strain amplitude is somewhat larger and needs to go beyond the quadratic fitting (so-called nonlinear elasticity), the main toplogical features of crystal bonding remains as at the bottom of the energy basin, and reversibility is ensured upon unloading. In contrast, dislocations represent extremely localized, large-amplitude, highly nonlinear (convex→concave→convex) and metastable deformation. The key to plasticity is the lock-in effect, which can be seen from (2.17) and (2.22) already. Namely, if one abuses a crystal by shearing, initially the crystal will cry out for pain, but if one keeps up the abuse, and push it through the nonlinear regime, then the crystal will start to feel less pain, and in the end would see no difference from its comfort zone. Until the next round of abuse starts. This "locks in" the large-amplitude, highly localized slip displacement. **Nonlinearity and non-convexity in the PEL is the essence of plasticity.** (as versus elasticity, which focusses on and is limited by the quadratic fit).

Why then, is dislocation slip preferred over, say, shearing 3 layers together? (From here on, slip means most localized, large shearing between two atomic planes.) We notice even that the generalized stacking fault calculation of $\gamma_1(\mathbf{x})$ looks kind of "unnatural", in that one must rigidly constrains the top and bottom blocks, and only allow relative displacement between the two rigid block. Why would one artificially apply such constraint?

The reason turns out to have more to do with the nonlinear response, than with the linear response of the crystal. If one fixes the external displacement Δ that spreads over n layers, one should plot and compare $\gamma_1(\Delta)$ with $n\gamma_n(\Delta/n)$ (From now on, we use Δ to denote shear

displacement instead of $\mathbf{x}$, since we will talk about spatially dependent displacement $\Delta(\mathbf{x})$). It turns out that, if we assume the local electron glue, (2.23), then for small Δ:

$$\gamma_1(\Delta) \ll n\gamma_n(\Delta/n) \tag{2.41}$$

Indeed, the curvature of the former is n times larger than that of the latter. So, for small Δ, diffuse deformation is preferred, the more diffuse, the better. However, once we requires large shear offset Δ, the situation is seen to be reversed. The saddle-point energy to overcome a diffuse barrier is n times larger than that of $\gamma_1(\Delta)$! Thus, for the most localized slip deformation, the pain comes quickly, but peaks earlier; whereas for the delocalized deformation, the pain comes later, but is ultimately greater. This basically says that, if one must cuts bonds to achieve large traction relaxation, then doing the bond cutting on one atomic plane is the best choice.

The above is the argument for the strongest possible localization in the z-direction, which is localizing down to a single slip plane between two adjacent atomic planes. There is also an argument for localization in the xy-plane, the so-called Peierls-Nabarro theory of the dislocation core. [31, 32] Basically, Peierls argues that if only pain on the slip plane ("localized pain") is counted:

$$E_{\text{slip}} \;=\; \sum_{\text{atom } i \text{ in core}} \gamma_1(\Delta_i) \;\approx\; \int dx \gamma_1(\Delta(x)) \tag{2.42}$$

this energy would prefer a core as narrow as possible. However, since a dislocation must make the transition from $\Delta = 0$ (outside of slipped plate) to $\Delta = \mathbf{b}$ (inside slipped plate), the slip offset Δ changes with position $\mathbf{x}$, and therefore elastic energy in other places ("diffuse pain") must also be involved. One could show it is of the form

$$E_{\text{elastic}} \;=\; \int dx \int dx' \frac{d\Delta(x)}{dx} K \ln|x - x'| \frac{d\Delta(x')}{dx'}, \tag{2.43}$$

where K depends on the elastic constants only. The above is a quadratic form: it is easy to show that if the dislocation core is wider by 2:

$$\Delta(x) \;\to\; \Delta\left(\frac{x}{2}\right) \tag{2.44}$$

E_{elastic} would drop by a factor of 4. Therefore E_{elastic} prefers as wide core as possible. Peierls

solved the variational problem:

$$E_{\text{dislocation}} \;=\; \int dx\gamma_1(\Delta(x)) + \int dx \int dx' \frac{d\Delta(x)}{dx} K \ln|x-x'| \frac{d\Delta(x')}{dx'}, \tag{2.45}$$

and obtained the in-plane size of the dislocation core.[31] The problem with (2.45) is that there is no barrier against the translation

$$\Delta(x) \;\rightarrow\; \Delta(x-s) \tag{2.46}$$

for arbitrary shift s of the dislocation core, forming so-called Goldstone mode, due to the continuum formulation. This is not true in reality, because so-called lattice friction does exist on all dislocations, for example screw dislocation in BCC metal, and dislocations in semiconductors, are known to have very significant lattice frictions.

Nabarro solved the zero-friction problem by resorting back to the atomistic sum:

$$E_{\text{slip}} = \int dx\gamma_1(\Delta(x)) \;\;\rightarrow\;\; E_{\text{slip}} = \sum_{\text{atom } i \text{ in core}} \gamma_1(\Delta_i) \tag{2.47}$$

using the Peierls core solution. The key results from Nabarro's work [32] are: (a) Nabarro obtained an energy barrier for dislocation translation, paradoxially called the Peierls energy barrier (in terms of stress needed to overcome this barrier, the Peierls stress), and (b) the Peierls barrier has strong (exponential) dependence on the core size. The wider the dislocation core, the lower the Peierls barrier. So, dislocations in FCC metals have wider cores (due to Shockley partials splitting), and the lattice friction is small. But screw dislocation in BCC crystals have narrow cores, and therefore the lattice friction can be very large, so large that it can dominate the overall plastic flow strength.

So dislocation is basically a machine to cut bonds on one plane, and then re-stitch them together. It should not be surprising that dislocation is the fundamental agent of plastic deformation, which is basically irreversible shape change, because dislocation slip gives the most localized (in z and in x) way to cut the bonds.

A dislocation is characterized by its line direction $\boldsymbol{\xi}$, $|\boldsymbol{\xi}| = 1$, and the Burgers vector $\mathbf{b}$, with

$$\mathbf{b} \;=\; \oint_{\overline{\mathrm{C}}} \frac{\partial \mathbf{u}}{\partial l} dl \;=\; \oint_{\mathrm{C}} \left(\frac{\partial \mathbf{u}_{\text{elastic}}}{\partial l} \right) dl \tag{2.48}$$

where the line integral is taken in a right-handed sense relative to $\boldsymbol{\xi}$. C is a closed loop in an

original perfect crystal far from the dislocation core (the Lagrangian frame of reference), and $\mathbf{u}$ is the total displacement after the dislocation has sheared into inside the loop, creating a branch cut. $\overline{\mathrm{C}}$ is the same loop as C, except it is open and avoiding the branch cut. $\frac{\partial \mathbf{u}}{\partial l}$ is a strain-like quantity, so we have

$$\frac{\partial \mathbf{u}}{\partial l} = \frac{\partial \mathbf{u}_{\text{elastic}}}{\partial l} + \frac{\partial \mathbf{u}_{\text{inelastic}}}{\partial l} \tag{2.49}$$

where $\frac{\partial \mathbf{u}_{\text{elastic}}}{\partial l}$ is small-amplitude but diffuse (away from the core), and $\frac{\partial \mathbf{u}_{\text{inelastic}}}{\partial l}$ is a delta-function like quantity in space, tracking the 2D branch cut. (The 2D branch cut ends at the 1D dislocation core.) Since stress $\boldsymbol{\sigma} \propto \frac{\partial \mathbf{u}_{\text{elastic}}}{\partial l}$, and the material at the branch cut is perfectly repaired and has the same load-bearing ability as the uncut material, $\boldsymbol{\sigma}$ is continuous across the branch cut and in fact is not even aware of its existence. (We will later see this from the stress solution of screw and edge dislocations). So $\frac{\partial \mathbf{u}_{\text{inelastic}}}{\partial l}$ is also continuous across the branch cut and not aware of the branch cut's existence. The second equality in (2.48) holds because in the continuum representation of $\mathbf{u}(\mathbf{x})$, the 2D branch cut is infinitely thin, and since $\frac{\partial \mathbf{u}_{\text{elastic}}}{\partial l}$ is finite, integrating $\frac{\partial \mathbf{u}_{\text{elastic}}}{\partial l}$ across the zero-thickness branch cut gives zero anyway. In the literature, one often sees

$$\mathbf{b} = \oint_{\mathrm{C}} \frac{\partial \mathbf{u}}{\partial l} dl \tag{2.50}$$

But one must understand this is an abbreviated notation due to "notational laziness". The branch cut unaware second equality in (2.48) is my favorite version because of its subtlety, and to make it even more subtle we can even use the (2.50) form, but keeping in mind that $\mathbf{u}$ there is the elastic component, i.e. modulo $\mathbf{b}$ at the branch cut plane to make $\frac{\partial \mathbf{u}}{\partial l}$ not divergent.

From (2.48) we see that $\boldsymbol{\xi}$ definition and $\mathbf{b}$ definition is related. $(-\boldsymbol{\xi}, -\mathbf{b})$ actually describes the same dislocation defect configuration as $(\boldsymbol{\xi}, \mathbf{b})$.

If $\mathbf{b} \parallel \boldsymbol{\xi}$, it is called screw dislocation. If $\mathbf{b} \perp \boldsymbol{\xi}$, it is called edge dislocation. Otherwise it is called mixed dislocation.

Because a loop integral of purely elastic displacements $\oint_{\mathrm{C}} \frac{\partial \mathbf{u}_{\text{elastic}}}{\partial l} dl$ should always give 0 (imagine we apply a diffuse but single-valued elastic distortion field in which C is embedded), (2.48) gives the *purely inelastic* excess displacement, which is the slip displacement $\mathbf{b}$ between two adjacent atomic planes (in continuum mechanics, this is idealized as infinitely sharp slip

fault). Because of this, there should be Burgers vector conservation law:

$$\mathbf{b}_1 \;=\; \mathbf{b}_2 + \mathbf{b}_3. \tag{2.51}$$

as one could distort C purely elastically from one location to another in Fig. 1-24 of [33].

For an infinite straight dislocation in isotropic elastic medium, the stress field is

$$\sigma_{xz} = -\frac{\mu b}{2\pi}\frac{y}{x^2+y^2}, \quad \sigma_{yz} = \frac{\mu b}{2\pi}\frac{x}{x^2+y^2}, \quad \sigma_{xy} = \sigma_{xx} = \sigma_{yy} = \sigma_{zz} = 0 \tag{2.52}$$

for "positive" screw dislocation:

$$\boldsymbol{\xi} \;=\; \frac{\mathbf{b}}{|\mathbf{b}|} \;=\; \mathbf{e}_z. \tag{2.53}$$

where μ is the shear modulus (we use G for crystallographic shear modulus). In cylindrical coordinate, this is

$$\sigma_{\theta z} \;=\; \frac{\mu b}{2\pi r}, \quad \sigma_{rz} = \sigma_{r\theta} = \sigma_{rr} = \sigma_{\theta\theta} = \sigma_{zz} = 0. \tag{2.54}$$

For edge dislocation,

$$\boldsymbol{\xi} \;=\; \mathbf{e}_z, \;\; \mathbf{b} \;=\; b\mathbf{e}_x \tag{2.55}$$

the formula is a little bit more complicated:

$$\sigma_{xx} \;=\; -\frac{\mu b}{2\pi(1-\nu)}\frac{y(3x^2+y^2)}{(x^2+y^2)^2}, \;\; \sigma_{yy} \;=\; \frac{\mu b}{2\pi(1-\nu)}\frac{y(x^2-y^2)}{(x^2+y^2)^2}, \;\; \sigma_{xz} = \sigma_{yz} = 0 \tag{2.56}$$

$$\sigma_{xy} \;=\; \frac{\mu b}{2\pi(1-\nu)}\frac{x(x^2-y^2)}{(x^2+y^2)^2}, \;\; \sigma_{zz} \;=\; \nu(\sigma_{xx}+\sigma_{yy}) \;=\; -\frac{\mu b\nu}{\pi(1-\nu)}\frac{y}{x^2+y^2}, \tag{2.57}$$

In cylindrical coordinates:

$$\sigma_{rr} = \sigma_{\theta\theta} \;=\; -\frac{\mu b\sin\theta}{2\pi(1-\nu)r}, \quad \sigma_{r\theta} \;=\; \frac{\mu b\cos\theta}{2\pi(1-\nu)r}, \tag{2.58}$$

$$\sigma_{zz} \;=\; \nu(\sigma_{rr}+\sigma_{\theta\theta}) \;=\; -\frac{\mu b\nu\sin\theta}{\pi(1-\nu)r}, \quad \sigma_{rz} = \sigma_{\theta z} = 0. \tag{2.59}$$

Taking the screw dislocation as example (the edge dislocation has the same scaling, but is algebraically more complex). We note in above that the dislocation stress field decays as r^{-1}. This means the elastic strain field decays also as r^{-1}, and the elastic strain energy density

behaves like

$$e_{\text{elastic}}(\mathbf{x}) \;=\; \frac{\sigma_{\theta z}^2}{2\mu} \;=\; \frac{\mu^2 b^2}{8\pi^2 \mu} r^{-2} \tag{2.60}$$

Thus, with a standalone dislocation, the total elastic energy per length scales as

$$\frac{E_{\text{elastic}}}{L} \;=\; \int_{R_0}^{R_1} dr 2\pi r \frac{\mu b^2}{8\pi^2} r^{-2} \;=\; \int_{R_0}^{R_1} dr \frac{\mu b^2}{4\pi} r^{-1} = \frac{\mu b^2}{4\pi} \ln \frac{R_1}{R_0}, \tag{2.61}$$

which is the diffuse "pain" in a ring of materials between R_0 and R_1. Obviously there is a problem with convergence in both the inner cutoff R_0 and the outer cutoff R_1. The inner cutoff can be handled by recognizing that elastic strain has a limit of $\sim 10\%$. Once that limit is reached, we get into the inelastic region of the core, and the pure elasticity theory no longer applies, and one has to use the Peierls-Nabarro theory of the dislocation core that has some handle on the nonlinear non-convex part of PEL, the E_{slip} term in (2.42). [31, 32]. When that nonlinear energy inside R_0 is included, the self energy can be written as

$$\frac{E_{\text{self}}}{L} \;=\; \frac{\mu b^2}{4\pi} \ln \frac{R_1}{R_0} + e_{\text{inelastic}} \;\equiv\; \frac{\mu b^2}{4\pi} \ln \frac{R_1}{\tilde{R}_0}. \tag{2.62}$$

Quite often people find $\tilde{R}_0$ to be around the order of b from exact atomistic calculations. [34].

There is also a problem with the outer cutoff R_1. This in fact means the dislocation cares about its environment. If a single screw dislocation exists in the center of a nanowire [35], then one can expect R_1 to be of the order the cylinder radius R. Generally speaking, in a bulk metal, if there are other dislocations which screen the field of the dislocation in question, and those nearest-neighbor screening dislocations are of the order R_{screen}, we would have the dislocation self energy as

$$\frac{E_{\text{self}}}{L} \;=\; \frac{\mu b^2}{4\pi} \ln \frac{R_{\text{screen}}}{\tilde{R}_0}. \tag{2.63}$$

A rule of thumb in the literature is to take

$$\eta \;\equiv\; \frac{E_{\text{self}}}{L} \;\sim\; \alpha \mu b^2 \tag{2.64}$$

with $\alpha \sim 0.5 - 1$. From (2.63), we see this implies the screening distance is of the order

$$\alpha = 0.5 : R_{\text{screen}} \;\sim\; e^{2\pi} \tilde{R}_0 = 535 \tilde{R}_0, \quad \alpha = 1 : R_{\text{screen}} \;\sim\; e^{4\pi} \tilde{R}_0 = 286751 \tilde{R}_0 \tag{2.65}$$

Assuming $\tilde{R}_0 = b = a_0/\sqrt{2} = 2.556$Å in Cu, this converts to $R_{\text{screen}} = 137$nm for $\alpha = 0.5$, to

$R_{\text{screen}} = 73\mu\text{m}$ for $\alpha = 1$, which covers most of the physically sensible ranges, from heavily work-hardened metal (a mediumly cold-worked Cu has dislocation density $\rho = 4 \times 10^{14}/\text{m}^2$ [36], which implies a characteristic spacing of 50 nm), to highly annealed metal.

From (2.64), we see the cost of creating a dislocation scales with b^2. Thus, whenever possible, the dislocation tends to split into the smallest crystallographic unit. Also, if the interplanar spacing d_0 is large, one tends to have smaller shear moduli. So to minimize the cost of dislocation μb^2, the preferred slip system tend to have (a) the smallest Burgers vector, and (b) the widest planar spacing. (a) and (b) are in fact not unrelated, because the smallest Burgers vector tend to occur on in the closest packing plane. But since the atomic density (a scalar) is the same no matter which planes and corresponding normal direction we count, the *closest packing plane* also tends to be the *loosest stacking plane*. All these point to choice of slip plane with the largest d_0 and smallest b.

Thus, in HCP metals, when the c/a-ratio is significantly smaller than the ideal value of $\sqrt{8/3} = 1.633$, like in Ti and Zr, the prismatic slip $\{10\bar{1}0\}\langle 1\bar{2}10\rangle$ is triggered, instead of basal slip $\{0001\}\langle 1\bar{2}10\rangle$.

The above logic naturally leads to Shockley partials. Assuming G is isotropic in plane (it is if the plane has 3-fold symmetry), the **Frank's rule** says that whenever

$$\mathbf{b}_1 \;=\; \mathbf{b}_2 + \mathbf{b}_3, \quad |\mathbf{b}_1|^2 \;>\; |\mathbf{b}_2|^2 + |\mathbf{b}_3|^2 \tag{2.66}$$

the $\mathbf{b}_1$ dislocation can reduce its energy by splitting into a $\mathbf{b}_2$ dislocation separated some distance from the $\mathbf{b}_3$ dislocation.

Consider (111) plane. The normal of this plane is $\mathbf{n} = [111]/(\sqrt{3}a_0)$. To orient ourselves (see Fig.2.1), we can take

$$\mathbf{e}_{x'} \;=\; \frac{[11\bar{2}]}{\sqrt{6}a_0}, \quad \mathbf{e}_{y'} \;=\; \frac{[\bar{1}10]}{\sqrt{2}a_0}, \quad \mathbf{e}_{z'} \;=\; \frac{[111]}{\sqrt{3}a_0} \tag{2.67}$$

we can check that $\mathbf{e}_{x'} \times \mathbf{e}_{y'} = \mathbf{e}_{z'}$. On this plane, there are six full Burgers vectors:

$$\mathbf{b}_1 \;\equiv\; \frac{[01\bar{1}]}{2}, \quad \mathbf{b}_2 \;\equiv\; \frac{[\bar{1}01]}{2}, \quad \mathbf{b}_3 \;\equiv\; \frac{[1\bar{1}0]}{2}, \tag{2.68}$$

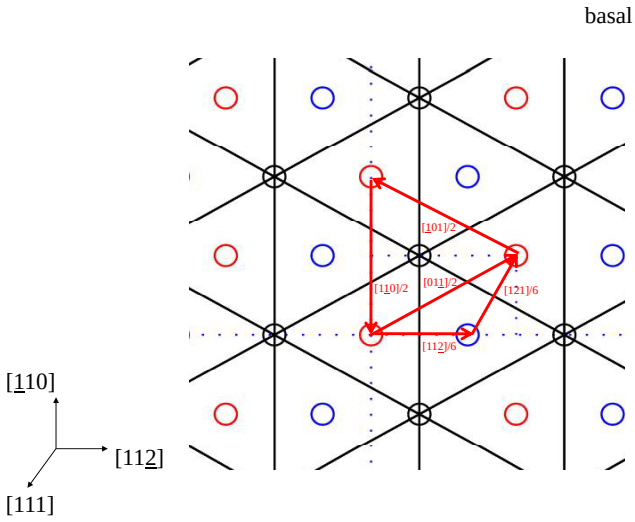


Figure 2.1: Looking down onto the (111) plane.

and $-\mathbf{b}_1$, $-\mathbf{b}_2$, $-\mathbf{b}_3$. Generally,

$$\boldsymbol{\epsilon}_{\text{inelastic}}^{\text{unsymmetrized}} = \frac{A_{\text{slip}}\mathbf{n}\mathbf{b}^T}{V} \tag{2.69}$$

where V is the total same volume, A_{slip} is how much area has slip occurred on this slip plane, and the superscript "unsymmetrized" means we have not carried out the symmetrization process in computing strain:

$$\boldsymbol{\epsilon}_{\text{inelastic}} = \frac{\boldsymbol{\epsilon}_{\text{inelastic}}^{\text{unsymmetrized}} + (\boldsymbol{\epsilon}_{\text{inelastic}}^{\text{unsymmetrized}})^T}{2}. \tag{2.70}$$

So $(\mathbf{n}, -\mathbf{b})$ are often considered to be a different slip system from $(\mathbf{n}, \mathbf{b})$.

Bruce Lee: *Now you put water in a cup, it becomes the cup; You put water into a bottle it becomes the bottle; You put it in a teapot it becomes the teapot. Now water can flow or it can crash. Be water, my friend.* One needs 5 indepedent slip systems to be able to deform arbitrarily.

The point here is that

$$\mathbf{b}_1 = \mathbf{b}_{\text{p1}} + \mathbf{b}_{\text{p2}}, \tag{2.71}$$

where the partial dislocations

$$\mathbf{b}_{\mathrm{p}1} \;=\; \frac{[11\bar{2}]}{6}, \quad \mathbf{b}_{\mathrm{p}2} \;=\; \frac{[\bar{1}2\bar{1}]}{6} \tag{2.72}$$

Since

$$|\mathbf{b}_{\mathrm{p}1}|^2 \;=\; |\mathbf{b}_{\mathrm{p}2}|^2 \;=\; \frac{a_0^2}{6}, \tag{2.73}$$

we have

$$|\mathbf{b}_{\mathrm{p}1}|^2 + |\mathbf{b}_{\mathrm{p}2}|^2 \;=\; \frac{a_0^2}{3} \tag{2.74}$$

which is smaller than

$$|\mathbf{b}_1|^2 \;=\; \frac{a_0^2}{2}. \tag{2.75}$$

Thus, two partials, separated far away, would have smaller energy than a full dislocation. In reality, they will not separate infinitely far apart because of the stacking fault ribbon they generated. Roughly speaking, the reduction in elastic energy is proportional to

$$\propto \;\; \frac{G\Delta(b^2)}{4\pi} \ln \frac{s}{\tilde{R}_0} \tag{2.76}$$

where s is the splitting separation between the two partials, so the total energy is like

$$E \;=\; -\frac{G\Delta(b^2)}{4\pi} \ln \frac{s}{\tilde{R}_0} + s\gamma_{\mathrm{ISF}}, \tag{2.77}$$

where γ_{ISF} is the intrinsic stacking fault energy. So the equilibrium splitting distance scales as

$$s_{\mathrm{eq}} \;=\; \frac{G\Delta(b^2)}{4\pi\gamma_{\mathrm{ISF}}}. \tag{2.78}$$

For low-stacking fault FCC crystal like pure Cu, $\gamma_{\mathrm{ISF}} = 40$ mJ/m^2, the splitting distance is large, like $s = 2$nm. For high-stacking fault FCC crystal like pure Al, $\gamma_{\mathrm{ISF}} = 160$ mJ/m^2, the splitting distance is small, like $s = 4$Å. This has severe consequences on the dislocation dynamics. For example, it is much more difficult for screw dislocations in Cu to cross-slip, because in order to do so, it must first constrict. But a widely separated ribbon would make the energy barrier for constriction larger.

The so-called Thompson tetrahedron describes the arrangement of full and partial Burgers vectors on slip planes. There are four faces (ABCδ, BCDα, CDAβ, DABγ, the last Greek letter is the center of each equilateral triangle), representing the $\{111\}$ planes. Clearly, if we

want to have $\overrightarrow{\mathrm{DA}}$ slip on DABγ slip plane, we can go:

$$\overrightarrow{\mathrm{DA}} \;=\; \overrightarrow{\mathrm{D}\gamma} + \overrightarrow{\gamma\mathrm{A}} \tag{2.79}$$

or

$$\overrightarrow{\mathrm{DA}} \;=\; \overrightarrow{\gamma\mathrm{A}} + \overrightarrow{\mathrm{D}\gamma} \tag{2.80}$$

where $\overrightarrow{\mathrm{DA}} \equiv \mathrm{A} - \mathrm{D}$ denotes the translation direction of the top block versus the bottom block across the slip plane (γ_1). The order of the decomposition matters, as one moves from $\Delta = 0$ region across the dislocation core, to the $\Delta = \overrightarrow{\mathrm{DA}}$ region. Only one choice among (2.79), (2.80) would be allowed. For (2.79), the atom at D site in the top block would be translated to

$$\mathrm{D} + \overrightarrow{\mathrm{D}\gamma} \;=\; \gamma \tag{2.81}$$

at the intermediate state. For (2.80), the atom at D site in the top block would be translated to

$$\mathrm{D} + \overrightarrow{\gamma\mathrm{A}} \tag{2.82}$$

which is also a crystallographic site. The key question here is whether γ or $\mathrm{D} + \overrightarrow{\gamma\mathrm{A}}$ is on top of a - site, or on top of an $\bigcirc$ site. The former (intrinsic stacking fault) is the much lower in energy than the latter on-top configuration. Since γ is on top of C when we look down on the DABγ plane of a Thompson's tetrahedron, we can determine that (2.80) is always right, when we perceive DABγ to be the top (+) plane. There is no $\mathbf{b}_\mathrm{p} \leftrightarrow -\mathbf{b}_\mathrm{p}$ symmetry in FCC or HCP crystals.

The Lomer-Cottrell (LC) lock is formed by the following reaction:

$$\frac{[11\bar{2}]_{(111)}}{6} + \frac{[\bar{1}\bar{2}1]_{(\bar{1}11)}}{6} \;=\; \frac{[0\bar{1}\bar{1}]}{6} \equiv \mathbf{b}_\mathrm{LC}, \tag{2.83}$$

since

$$\frac{a_0^2}{6} + \frac{a_0^2}{6} \;>\; \frac{a_0^2}{18}. \tag{2.84}$$

However, note that $\mathbf{b}_\mathrm{LC} = \frac{[0\bar{1}\bar{1}]}{6}$ is not our usual Burgers vector. Slip by $\frac{[0\bar{1}\bar{1}]}{6}$ on *any* atomic plane is likely to creates a very high energy stacking fault. Furthermore, there is another fundamental conflict if the Lomer-Cottrell dislocation is to move by glide. Note that by the way LC is formed, its line direction $\boldsymbol{\xi}_\mathrm{LC}$ must be a common direction on both (111) and ($\bar{1}11$) planes, namely $\boldsymbol{\xi}_\mathrm{LC} \parallel [111] \times [\bar{1}11] \parallel [01\bar{1}]$. However, $\mathbf{b}_\mathrm{LC} = \frac{[0\bar{1}\bar{1}]}{6}$ does not belong to either (111) or ($\bar{1}11$) "old" planes. It does belong to the ($11\bar{1}$) and ($\bar{1}1\bar{1}$) "new" planes, but

$\boldsymbol{\xi}_{\rm LC}$ does not belong to these "new" planes. Thus, there is no {111} plane where the LC dislocation could move as edge dislocation. That plane should be $\boldsymbol{b}_{\rm LC} \times \boldsymbol{\xi}_{\rm LC} = (100)$, but this cube plane is unusual for slip. For this reason, the Lomer-Cottrell dislocation is called "sessile", or "lock" or "junction", meaning it is a low-energy trap state, but once formed, it would be difficult to move. The LC dislocations are important for dislocation storage and forest dislocation hardening in FCC metals. [6]

The so-called Peach-Koehler force on a dislocation can be derived by virtual work:

$$\delta W \;=\; V\mathrm{Tr}(\boldsymbol{\sigma}\delta\boldsymbol{\epsilon}_{\rm inelastic}) = \mathbf{b}^T\boldsymbol{\sigma}(\boldsymbol{\xi}dl\times\delta\mathbf{x}) \;=\; dl(\mathbf{b}\cdot\boldsymbol{\sigma})\cdot(\boldsymbol{\xi}\times\delta\mathbf{x}) \;=\; dl\delta\mathbf{x}\cdot((\mathbf{b}\cdot\boldsymbol{\sigma})\times\boldsymbol{\xi}) \quad (2.85)$$

since $\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = \mathbf{c}\cdot(\mathbf{a}\times\mathbf{b})$. So the force per unit length of dislocation is

$$\frac{d\mathbf{F}}{dl} \;=\; (\mathbf{b}\cdot\boldsymbol{\sigma})\times\boldsymbol{\xi}. \quad (2.86)$$

In index form this would be

$$\frac{dF_i}{dl} \;=\; \epsilon_{ijk}b_l\sigma_{lj}\xi_k. \quad (2.87)$$

where repeated indices are summed over, and ϵ_{ijk} is the Levi-Civita permutation symbol:

$$\epsilon_{123} = \epsilon_{231} = \epsilon_{312} = 1, \quad \epsilon_{213} = \epsilon_{132} = \epsilon_{321} = -1, \quad \text{all others} = 0. \quad (2.88)$$

This force is always perpendicular to $\boldsymbol{\xi}$. For a non-screw dislocation, the slip plane would have normal

$$\mathbf{m} \;=\; \frac{\mathbf{b}\times\boldsymbol{\xi}}{|\mathbf{b}\times\boldsymbol{\xi}|} \quad (2.89)$$

with $\mathbf{m}\perp\boldsymbol{\xi}$, and gliding direction

$$\mathbf{g} \;=\; \boldsymbol{\xi}\times\mathbf{m}. \quad (2.90)$$

So the total force can be written as

$$\frac{d\mathbf{F}}{dl} \;=\; \frac{d\mathbf{F}_{\rm glide}}{dl} + \frac{d\mathbf{F}_{\rm climb}}{dl} \quad (2.91)$$

with

$$\frac{d\mathbf{F}_{\rm glide}}{dl} \;=\; \mathbf{g}(\mathbf{g}\cdot((\mathbf{b}\cdot\boldsymbol{\sigma})\times\boldsymbol{\xi})), \quad \frac{d\mathbf{F}_{\rm climb}}{dl} \;=\; \mathbf{m}(\mathbf{m}\cdot((\mathbf{b}\cdot\boldsymbol{\sigma})\times\boldsymbol{\xi})). \quad (2.92)$$

Generally speaking, dislocation climb is called "non-conservative" process, because a net

[6] It is not impossible to move, or at least remove LC dislocations, however, if we consider dislocation reactions under stress, or dislocation climb.

flux of atoms toward the core by diffusion is needed in order to drive climb. Thus at lower temperatures when long-range diffusion is impossible, even with finite driving force $\frac{d\mathbf{F}_{\text{climb}}}{dl}$, dislocation won't climb.

Dislocation glide, however, is called "conservative" or displacive process, where all that is needed is for the atoms that are already there to shift their positions by a small and semi-deterministic amount. Dislocation glide is much more ready process when $\frac{d\mathbf{F}_{\text{glide}}}{dl}$ exceeds some threshold. Below we look at a famous case. Consider a pure applied shear stress $\sigma_{xy} = \tau$ for a curved dislocation on y-plane with $\mathbf{b} = b\mathbf{e}_x$, and

$$\boldsymbol{\xi}(l) \;=\; \xi_x \mathbf{e}_x + \sqrt{1-\xi_x^2}\mathbf{e}_z. \tag{2.93}$$

$$\frac{d\mathbf{F}_{\text{climb}}}{dl} \;=\; 0, \quad \frac{d\mathbf{F}_{\text{glide}}}{dl} \;=\; b\tau \mathbf{e}_y \times \boldsymbol{\xi}, \tag{2.94}$$

We can call η in (2.64) the *line tension* of a dislocation. If we pretend

1. η to be independent of $\boldsymbol{\xi}$. (In reality η depends on $\boldsymbol{\xi}$.)
2. Besides the self energy, the dislocations do not interact with each other elastically. (In reality, they do).

we come to the so-called line tension model of a dislocations. This is an extremely simple model because it is local.

Consider the line direction $\boldsymbol{\xi}(l)$ as a function of the dislocation length l. If the dislocation is a straight line, then locally we have force equilibrium from the line tension. But, if $\boldsymbol{\xi}(l)$ has curvature, this would generate

$$d\mathbf{F} \;=\; \eta\boldsymbol{\xi}(l+dl) - \eta\boldsymbol{\xi}(l) \;=\; \eta\frac{d\boldsymbol{\xi}}{dl}dl \;=\; \eta\frac{\mathbf{e}_R(l)}{R(l)}dl \tag{2.95}$$

where $R(l)$ is the radius of curvature, and $\mathbf{e}_R$ points towards the center of the local tangent circle.

If a dislocation is pinned between two fixed ends with distance $2a$, then we would have a circular segment, with line tension balancing the PK force:

$$b\tau \;=\; \frac{\eta}{R(l)} \quad \rightarrow \quad R(l) = \frac{\eta}{b\tau} \tag{2.96}$$

From the derivation above, we see R is actually independent of l when the dislocation reaches equilibrium. This means at equilibrium, the dislocation is always arc of a *perfect circle* in the isotropic line tension model.

The critical configuration is actually when $R = a$ (R first decreases with $\tau \uparrow$, but after reaching the minimum value of a, would start to increase again, so $R = a$ is the "saddle" configuration), so the critical external stress for bow-out is

$$\tau_{\mathrm{C}} \;=\; \frac{\eta}{ba} = \frac{\alpha G b^2}{ba} = \frac{\alpha G b}{a}. \tag{2.97}$$

In reality, $a = 10^{-6}$m, but $b \sim 2 \times 10^{-10}$m, so we get

$$\tau_{\mathrm{C}} \;\sim\; 10^{-4} G. \tag{2.98}$$

The above immediately explains the $1000\times$ difference with Frenkel estimate of ideal shear strength.

The dislocation density ρ [unit $1/\mathrm{m}^2$] is defined as the total length of all dislocations in a unit volume of material. ρ in mediumly work-hardened Cu is typically on the order of $4 \times 10^{14}/\mathrm{m}^2$ (number of etch pits per unit area) $= 4 \times 10^{14}\mathrm{m}/\mathrm{m}^3$ (dislocation line length per m^3 of material - in reference, circumference of earth is 4×10^7m, circumference of sun is 4×10^9m - it would take light 15 days to traverse the dislocation line in $1\mathrm{m}^3$ of copper! so to simulate plasticity by tracking dislocations is quite a challenge). We can estimate the mean spacing between between dislocations to be

$$2a \;=\; \rho^{-1/2} \tag{2.99}$$

Plugging into (2.97), we get

$$\tau_{\mathrm{C}} \;=\; 2\alpha G b \rho^{1/2} \tag{2.100}$$

The above $\rho^{1/2}$ dependence is called the Taylor hardening law. It comes from forest dislocation resistance. There can be other sources of plastic flow resistance, for example lattice friction, solute hardening, precipitate/dispersion hardening, grain boundary hardening etc. The typical way of modeling them is to add all together:

$$\tau_{\mathrm{C}} \;=\; 2\alpha G b \rho^{1/2} + \tau_{\mathrm{Peierls}} + \tau_{\mathrm{solute}} + \tau_{\mathrm{precipitate/dispersion}} + \tau_{\mathrm{GB}} \tag{2.101}$$

τ_{C} is called the critical resolved shear stress (CRSS). The resolved shear stress τ on a slip

system is generally computed as

$$\tau \equiv \frac{(\mathbf{b}\cdot\boldsymbol{\sigma})\cdot\mathbf{n}}{|\mathbf{b}|} = \frac{b_i\sigma_{ij}n_j}{b} \tag{2.102}$$

If we put uniaxial tension/compression along a direction $\mathbf{u}$, we have

$$\boldsymbol{\sigma} = \sigma\mathbf{u}\mathbf{u}^T \tag{2.103}$$

we have

$$\tau = \frac{\sigma(\mathbf{b}\cdot\mathbf{u})(\mathbf{n}\cdot\mathbf{u})}{b} = \sigma\cos(\theta_b)\cos(\theta_n) \tag{2.104}$$

where θ_b is angle between $\mathbf{u}$ and $\mathbf{b}$, and θ_n is angle between $\mathbf{u}$ and $\mathbf{n}$. $\cos(\theta_b)\cos(\theta_n)$ is called the Schmid factor. Since $\mathbf{n}\perp\mathbf{b}$, the maximum Schmid factor is $\frac{1}{2}$, when $\mathbf{u}$ is 45° between $\mathbf{n}$ and $\mathbf{b}$.

The so-called Schmid's Law means all that matter is scalar CRSS τ_{C}, no matter what is the tensor $\boldsymbol{\sigma}$ that generates this scalar.

In above we have been talking about shear, i.e. bond switching, where there is transient loss of coordination for the atoms involved, but over long timescale no net loss of total coordination (or very little). This is fundamentally different from the cleavage process, where there is often irreversible loss of metal-metal coordination [7] Shear and cleavage are the two fundamental categories of inelastic events inside the solid. For small elastic deformation, they are roughly characterized by G and B, respectively. Then for ideal strength calculation, there is no formal distinction, but practially the tensile and shear ideal strength and strains can be used to characterize the intrisic brittleness of materials [15, 16]. But for large nonlinear inelasticity, the inelastic shear and inelastic cleavage are very different. The metal-metal bond switching is a reversible source of dissipation: an arrays of bond switched this way can be reswitched later, converting mechanical energy to heat many many times. But if there is a loss of coordination, for example by voiding and surface creation, then this can only be used one time. For this reason, metals which are shear soft have a larger fracture toughness, because the soft shear entices the shear relaxation again and again. Bonding shearing is a *sustainable* way of dissipating energy, whereas cleavage is basically a one-off thing.

Just like the Frenkel relation for shear, there is a popular form for fitting decohesion called Universal Binding Energy Relation (UBER)[38]. The details are not that important, the

[7] Imagine, that, once two metal surfaces are opened by the Griffith process [37], the metal surfaces are passivated by oxygen, and one cannot recover the metal-metal coordination even if the crack is closed later.

key is that $e_{\rm b}(\epsilon_{\rm hydro})$ is not a periodic function, but is a function with a minimum, followed by a turning point where the 2nd derivative vaishes. So $\sigma_{\rm hydro}(\epsilon_{\rm hydro})$ has a maximum, then decays to zero as $\epsilon_{\rm hydro} \to 0$. Also, it can be shown that to separate a material, the best way is to localize the bond cutting on one plane. In other words, consider a crystal with 10^{24} atoms, thus 10^8 planes on each side. It takes only cutting the bonds on one plane out of the 10^8 to achieve separation. Brittle ceramics basically do this. It turns out that metals are wily, and do not fall for this generally. It takes a whole lot of bond shearing in metals before one coordination loss is achieve in metals, by for instance dislocation emission in front of the crack tip.

Having reconciled the $\sigma_{\rm shear}^{\rm ideal} = 2$GPa for Mg vesus the measured CRSS = 0.35 MPa for Mg (Basically dislocation is like a lever, that breaks bond in its core, and then restitches them back together), I would like to mention an interesting possibility of elastic strain engineering [39, 2]. All physical properties are function of the elastic strain. Because "smaller is stronger", nanostructured materials such as nanowires, nanotubes, nanoparticles, thin films, atomic sheets etc. can dynamically withstand non-hydrostatic (e.g. tensile and shear) stresses up to a significant fraction of its ideal strength without inelastic relaxation by plasticity or fracture. For example, large elastic strains can be generated by epitaxy in thin films, or by static or dynamical external loading on small-volume materials, and can be spatially homogeneous or inhomogeneous. This leads to new possibilities for tuning the physical and chemical (e.g. electronic, optical, magnetic, phononic, catalytic, etc.) properties of a material, by varying the 6-dimensional elastic strain as continuous variables. By controlling the elastic strain field statically or dynamically, one opens up a much larger parameter space (probably on par with chemical alloying) for optimizing functional properties of materials, imparting a new meaning to Feynman's statement "there's plenty of room at the bottom".

Chapter 3

Linear Response Theory and Long-Range Diffusion

The chemical potential

$$\mu_i \equiv \left.\frac{\partial G}{\partial N_i}\right|_{N_{j\neq i},T,P} = g(T,P,\mathbf{X}) + (\mathbf{p}_i - \mathbf{X})\cdot\nabla_{\mathbf{X}} g(T,P,\mathbf{X}), \tag{3.1}$$

where

$$g = \frac{G}{N}, \quad N \equiv N_1 + N_2 + N_3 + ... + N_c, \tag{3.2}$$

$$\mathbf{X} = [X_2, X_3, ..., X_c], \quad X_i \equiv \frac{N_i}{N}, \quad (\mathbf{p}_i)_j = \delta_{i,j+1} \tag{3.3}$$

is the thermodynamic "price" of a type-i atom. If the price varies spatially $\mu_i = \mu_i(\mathbf{x})$, there will be an incentive for the atom to move from locations of higher price to locations of lower price (social analogy: migrant worker, currency arbitrage). Such a system is not in global thermodynamic equilibrium, but the non-equilibrium is of global nature ("type A" non-equilibrium in chapter A) so *local* $T(\mathbf{x})$, $P(\mathbf{x})$, $\mu_i(\mathbf{x})$ can be defined. In other words, each RVE (representative volume element), were it isolated, would be infinitesimally close to an equilibrium state (atoms in this RVE are "happy" with nearby atoms in the same RVE; they just try to "keep up with the Joneses" - the other RVEs).

Define atom flux $\mathbf{J}_i$ to be the number of type-i atoms jumping though a unit area per unit time:

$$dN_i(\text{in} \to \text{out}) - dN_i(\text{out} \to \text{in}) = (\mathbf{J}_i \cdot \mathbf{n}) dA dt \tag{3.4}$$

Note that in d-dimensional space, flux is a d-component vector $\mathbf{J}_i = [J_{ix}, J_{iy}, J_{iz}]$, of unit atoms/m^2/s. Conjugate to the flux vector, define the *concentration* of type-i atoms to be $c_i \equiv N_i(\text{RVE})/\text{volume(RVE)}$, which has unit atoms/m^3. The total concentration (also unit atoms/m^3) is

$$c \equiv \sum_i c_i = \frac{N(\text{RVE})}{V(\text{RVE})} = \frac{1}{v} \tag{3.5}$$

the concentration and the composition are related simply by:

$$c_i = cX_i \tag{3.6}$$

since the common volume factors cancel out. Because the total number of atoms is conserved, one RVE's gain must be other RVEs' loss, there will be an atom conservation equation (Fick's second law):

$$\partial_t c_i = -\nabla \cdot \mathbf{J}_i = -\partial_x J_{ix} - \partial_y J_{iy} - \partial_z J_{iz}. \tag{3.7}$$

(3.7) can be simply appreciated in the case of 1D transport: $J_{iy} = J_{iz} = 0$, $J_{ix} = J_{ix}(x,t)$. If we take interval $[x - \Delta x/2, x + \Delta x/2]$ and unit area in yz plane, and count the number of "red Ferraris" N_i in the interval, as well as observing how many red Ferraris have passed observation posts (police patrol cars) $x - \Delta x/2$ and $x + \Delta x/2$, there must be conservation of red Ferraris:

$$J_{ix}(x - \Delta x/2, t) - J_{ix}(x + \Delta x/2, t) = \dot{N}_i \quad \to \quad -\partial_x J_{ix}(x) \approx \partial_t c_i. \tag{3.8}$$

The *thermodynamic driving force* for diffusion is the negative *gradient* of chemical potential: $\mathbf{F}_i \equiv -\nabla \mu_i(\mathbf{x}) = -[\partial_x \mu_i, \partial_y \mu_i, \partial_z \mu_i]$. This is because if there were no gradients: $\mu_i(\mathbf{x}) = \mu_i^{\text{ref}}$ for all i, then no matter how high or low are the uniform μ_i^{ref}s' the absolute magnitude, there will be no *diffusion*. We know that $\mathbf{J}_i$ somehow depends on the driving forces $\mathbf{J}_i = \mathbf{J}_i(\mathbf{F}_1, \mathbf{F}_2, .., \mathbf{F}_c)$, so we can do a Taylor expansion on the driving forces around $\mathbf{F}_1 = \mathbf{F}_2 = .. = \mathbf{F}_c = 0$, and when the driving forces are small, only the leading-order terms are important, which are:

$$\mathbf{J}_i = \sum_j \mathbf{L}_{ij} \mathbf{F}_j = -\sum_j \mathbf{L}_{ij} \nabla \mu_j \tag{3.9}$$

where $\mathbf{L}_{ij}$'s are called Onsager linear-response coefficients. For example, in a 3-component system ($C = 3$), we would have

$$\mathbf{J}_1 = \quad \mathbf{L}_{11}\mathbf{F}_1 + \mathbf{L}_{12}\mathbf{F}_2 + \mathbf{L}_{13}\mathbf{F}_3 = -\mathbf{L}_{11}\nabla\mu_1 - \mathbf{L}_{12}\nabla\mu_2 - \mathbf{L}_{13}\nabla\mu_3$$

$$\begin{aligned}\mathbf{J}_2 &= \mathbf{L}_{21}\mathbf{F}_1 + \mathbf{L}_{22}\mathbf{F}_2 + \mathbf{L}_{23}\mathbf{F}_3 = -\mathbf{L}_{21}\nabla\mu_1 - \mathbf{L}_{22}\nabla\mu_2 - \mathbf{L}_{23}\nabla\mu_3 \\ \mathbf{J}_3 &= \mathbf{L}_{31}\mathbf{F}_1 + \mathbf{L}_{32}\mathbf{F}_2 + \mathbf{L}_{33}\mathbf{F}_3 = -\mathbf{L}_{31}\nabla\mu_1 - \mathbf{L}_{32}\nabla\mu_2 - \mathbf{L}_{33}\nabla\mu_3 \end{aligned} \tag{3.10}$$

Regarding (3.9), several important observations can be made:

1. Since the Taylor expansion is about $\mathbf{F}_1 = \mathbf{F}_2 = .. = \mathbf{F}_c = 0$, the linear coefficients $\mathbf{L}_{ij}$ themselves are independent of the driving forces ("linear response"), but they could depend on the local composition, as well as temperature and pressure: $\mathbf{L}_{ij} = \mathbf{L}_{ij}(\mathbf{X}) = \mathbf{L}_{ij}(\mathbf{X}(\mathbf{x}))$, as well as the position through $\mathbf{X}(\mathbf{x})$.

2. Each $\mathbf{L}_{ij}$ is a $d \times d$ matrix. If the material is isotropic or if we are in a quasi-1D situation, however, we can consider $\mathbf{L}_{ij}$ as a scalar: $\mathbf{L}_{ij} = L_{ij}\mathbf{I}_{d\times d}$.

3. $\mathbf{L}_{ii}$ ($\mathbf{L}_{11}$, $\mathbf{L}_{22}$, $\mathbf{L}_{33}$) are so called diagonal or *direct coefficients*. $\mathbf{L}_{ij}$ with $j \neq i$ are so called off-diagonal or *cross-coupling* coefficients. Cross-coupling effect can be appreciated, in for instance gas diffusion [40]. Say species 3 gas atoms have uniform chemical potential in the chamber: $\mu_3(\mathbf{x}) = \mu_3^{\text{ref}}$, so there is no *incentive* for species 3 atoms to move left or right overall, macroscopically. However, imagine there is a finite diffusional driving force $\nabla\mu_1(\mathbf{x})$ for species 1 atoms, which causes them to move macroscopically from left to right. There are unavoidable collisions between type-3 and type-1 atoms: since on average more type-1 atoms are moving to the right when they collide, type-3 atoms may be *entrained* to move to the right as well, even if it is not in their own "interest" to move to the right. This is similar to eating more than you typically do in a convivial feast with giants. A remark should be made here that even if type-3 atoms are driven to diffuse *up* its own $\nabla\mu_3$ gradient, the overall $dS_{\text{universe}} > 0$ (Chap.2 of [41]), since type-1 atoms will gain more by diffusing *down* the $\nabla\mu_1$ gradient.

The cross-coupling effect is very general. Not only the atom (ion) fluxes, but heat flux $\mathbf{J}_Q$ (unit W/m^2) and electron flux $\mathbf{J}_e$ (unit #electrons/m^2/s, the electrical current in a metallic wire $I = -e\mathbf{J}_e A$ where A is cross-section of the wire) can be related to gradients in the thermal potential ($\ln T$) and electron's chemical potential μ_e (see Chap 2.1 of [41]). The cross-coupling effects manifest in for instance in electro-migration:

$$\mathbf{J}_i = L_{ii}(-\nabla\mu_i) + L_{ie}(-\nabla\mu_e), \quad \mathbf{J}_e = L_{ei}(-\nabla\mu_i) + L_{ee}(-\nabla\mu_e) \tag{3.11}$$

where $\mathbf{J}_i$ stands for some ion/atom flux. μ_e is the electron chemical potential, and is related to the electrode potential U simply as $\mu_e(\mathbf{x}) = -eU(\mathbf{x})$. In [42], one sees indium atom

transfer from larger indium nanoparticle to smaller indium nanparticle, which is reverse of typical coarsening process. This is because there is an electrical current and "electron wind force" that are blowing indium atoms downstream.

The cross-coupling effect is also the basis for the thermoelectric effects (Fig. 2.2, 2.3 in [41]), due to:

$$\mathbf{J}_Q \;=\; L_{QQ}(-\nabla \ln T) + L_{Qe}(-\nabla \mu_e), \quad \mathbf{J}_e \;=\; L_{eQ}(-\nabla \ln T) + L_{ee}(-\nabla \mu_e) \tag{3.12}$$

A well-known special case is the Seebeck effect, where temperature gradient can lead to electromotive force (voltage). To show this, take two metallic wires A and B made of different materials. Take a large-impedance voltmeter at constant temperature T, connect the two wire to two leads of the voltmeter at T, then *join* the other two ends at another constant-temperature reservoir at $T+\Delta T$. Due to the large electrical impedance of the voltmeter, there is nearly no current in the circuit, $\mathbf{J}_e = 0$ in either wire. This means $L_{eQ}(\nabla \ln T) = eL_{ee}(\nabla \phi)$, so $\Delta\phi^{\mathrm{A}} = \frac{L^{\mathrm{A}}_{eQ}}{eL^{\mathrm{A}}_{ee}} \ln \frac{T+\Delta T}{T}$, $\Delta\phi^{\mathrm{B}} = \frac{L^{\mathrm{B}}_{eQ}}{eL^{\mathrm{A}}_{ee}} \ln \frac{T+\Delta T}{T}$. Since $\phi^{\mathrm{B}} = \phi^{\mathrm{A}}$ at the place they join, the voltmeter must measure a potential difference. This potential difference (electromotive force) can be exploited (thermal energy $\rightarrow$ electrical energy), although the calculation becomes more involved when the current is finite. Because there are no moving parts, this can be used for low-grade thermal energy scavenging. Also, this is the mechanism behind some thermocouples (temperature sensors). For example, a Cu - $Cu_{55}Ni_{45}$ (constantan, eureka) thermocouple has a *Seebeck coefficient* (aka thermopower) of 41 microvolts per Kelvin at room temperature.

One can also use electrical energy to transport heat, which is so-called Peltier effect. Consider replacing the voltmeter above by a battery, and let us start by having $\Delta T = 0$, i.e. the whole apparatus starts isothermally. Then, there is $\mathbf{J}_Q/\mathbf{J}_e = L_{Qe}/L_{ee}$. Consider two wires having equal cross section, then there is electron current in the circuit driven by the battery, with electron current conservation $\mathbf{J}^{\mathrm{A}}_e = \mathbf{J}^{\mathrm{B}}_e$. But since $L^{\mathrm{A}}_{Qe}/L^{\mathrm{A}}_{ee} \neq L^{\mathrm{B}}_{Qe}/L^{\mathrm{B}}_{ee}$ in general, we will have $\mathbf{J}^{\mathrm{A}}_Q \neq \mathbf{J}^{\mathrm{B}}_Q$. Unlike electrons, heat can easily accumulate (C_P) and can also irradiate, and in fact we can inject heat $\mathbf{J}^{\mathrm{A}}_Q - \mathbf{J}^{\mathrm{B}}_Q$ on one end, and take out heat $\mathbf{J}^{\mathrm{B}}_Q - \mathbf{J}^{\mathrm{A}}_Q$ on the other, and maintain steady-state operation. This is then a *heat pump*, and is in fact how refrigerator works. There are now thermoelectric refrigerators on market: they have no moving parts and is therefore dead quiet. Because electrical dissipation (finite electrical impedance) is involved in this setup, the efficiency of the thermoelectric refrigerator is less than the ideal Carnot efficiency ($\eta^{\mathrm{ideal}} = \infty$ in fact for isothermal heat pump).

Note that the Onsager (3.12) is a master equation ("Grandaddy of all them equations") that can describe thermal conduction or electrical conduction individually, as well as their cross-coupling effect. For example, the well-known Ohm law $\Delta\phi = IR$ for a resistor under voltage drop $\Delta\phi$ is just a special case of (3.12). Consider a wire of length l and cross-sectional area A. If the wire is under isothermal condition, then $\nabla \ln T = 0$, and then the only driving force to drive electron flux in (3.12) would be $\nabla\mu_e = -e\nabla\phi$: $\mathbf{J}_e = eL_{ee}\nabla\phi = eL_{ee}\Delta\phi/l$, and the electrical current in the wire is then $I \equiv (-e)\mathbf{J}_e A = -e^2 L_{ee} A/l\Delta\phi = \Delta\phi/R_{\text{wire}}$. We can thus identify

$$R_{\text{wire}} = \frac{l}{e^2 L_{ee} A} \tag{3.13}$$

the minus sign is because electrical current always flow from high ϕ to low ϕ. We see then R_{wire} is proportional to the wire length, and inversely proportional to its cross-sectional area, which agrees with intuition. Similarly, when there is no electrostatic potential gradient, $\nabla\phi = 0$, we have $\mathbf{J}_Q = L_{QQ}(-\nabla \ln T) = -(L_{QQ}/T)\nabla T$ and a simple thermal conduction problem, and we can identify L_{QQ}/T to be the so-called *thermal conductivity* with unit J/m/s/K.

4. (3.9) is correct *only if* we measure the flux in an observation frame co-moving with the material. If we talk about diffusion in crystalline materials, this frame of observation is called the local lattice or crystal frame (C-frame) . All fluxes in (3.9) and (3.10) then have $^{\text{C}}$ superscript on them: $\mathbf{J}_i = \mathbf{J}_i^{\text{C}}$, which denote *diffusional* contribution to the flux. It is quite obvious that if the *whole material* is translating with respect to an observer, with velocity $\mathbf{v}_{\text{C}}^{\text{L}}$, where $^{\text{L}}$ denotes measurements performed by the observer in the lab, there should be $\mathbf{J}_i^{\text{L}} = \mathbf{J}_i^{\text{C}} + c_i\mathbf{v}_{\text{C}}^{\text{L}}$, where $c_i\mathbf{v}_{\text{C}}^{\text{L}}$ is the *convective* contribution to the flux. Another way to see this is to define $\mathbf{v}_i^{\text{C}} \equiv \mathbf{J}_i^{\text{C}}/c_i$, the average diffusional velocity seen in the crystal frame. $\mathbf{v}_i^{\text{C}}$ is simply the average atomic velocities of all type-i atoms in the RVE, measured by a crystal-frame observer who is attached to the lattice and who thinks the RVE is not moving. Since velocities are additive, there must be $\mathbf{v}_i^{\text{L}} = \mathbf{v}_i^{\text{C}} + \mathbf{v}_{\text{C}}^{\text{L}}$, and $\mathbf{J}_i^{\text{L}} = c_i(\mathbf{v}_i^{\text{C}} + \mathbf{v}_{\text{C}}^{\text{L}}) = \mathbf{J}_i^{\text{C}} + c_i\mathbf{v}_{\text{C}}^{\text{L}}$. This distinction between crystal-frame and lab-frame observations will be important when we later discuss about the Kirkendall effect, because it turns out that the crystal planes can actually move with respect to the lab (creep) in diffusion-couple experiments, due to finite divergence of the vacancy flux. The creeping velocities $\mathbf{v}_{\text{C}}^{\text{L}}$ can depend on position, in which case $\mathbf{J}_i^{\text{L}} = \mathbf{J}_i^{\text{C}} + c_i\mathbf{v}_{\text{C}}^{\text{L}}(\mathbf{x})$. The physical basis of decomposing flux into diffusional and convective contributions is because thermodynamics is fundamentally local, and decoupled from macroscopic motion. An atom in a glass of water in a spaceship sees only atoms nearby and has no idea how fast the spaceship is moving, or even how fast the water is twirling.

The chemical potential of atom which drives the Onsager flux is thus decoupled from the macroscopic velocity of the spaceship or even the twirling of water in the glass. The task of solving for the convective velocity $\mathbf{v}_{\mathrm{C}}^{\mathrm{L}}(\mathbf{x})$ in a whole continuum body falls in the realm of mechanics (elasto-plasto dynamics and fluid dynamics) and outside of the realm of materials thermodynamics and kinetics, although in extremely simple geometries like Kirkendall effect in quasi-1D diffusion couple we can just make statements about $\mathbf{v}_{\mathrm{C}}^{\mathrm{L}}(\mathbf{x})$ by inspection without doing mathematical mechanics. This is true for ion/atom flux as well as electron flux.

When one is talking about gas or liquid, there is no longer the concept of a site lattice or "crystal frame". In that case, the correct procedure is to define $\mathbf{v}_{\mathrm{C}}^{\mathrm{L}} \equiv \sum_{i\in\mathrm{RVE}} m_i\mathbf{v}_i / \sum_{i\in\mathrm{RVE}} m_i$ for atoms within an RVE, which is the convective velocity of that RVE. The diffusive fluxes are defined when $\mathbf{v}_{\mathrm{C}}^{\mathrm{L}}$ is subtracted off from the atom velocities. The reason for this is that $\bar{\mathbf{v}}$ in the Maxwellian distribution (A.1) is $\mathbf{v}_{\mathrm{C}}^{\mathrm{L}}$.

Based on statistical mechanics, Lars Onsager (1968 Nobel prize in chemistry) proposed a fundamental reciprocal relation $\mathbf{L}_{ij} = \mathbf{L}_{ji}^{T}$. In the case of isotropy (for diffusion, this means cubic symmetry or higher [4]) or quasi-1D situation, this means

$$L_{ij} \;=\; L_{ji}. \tag{3.14}$$

For example, in the case of $c = 3$,

$$\begin{aligned}
\mathbf{J}_1 \;&=\; -L_{11}\nabla\mu_1 - L_{12}\nabla\mu_2 - L_{13}\nabla\mu_3 \\
\mathbf{J}_2 \;&=\; -L_{21}\nabla\mu_1 - L_{22}\nabla\mu_2 - L_{23}\nabla\mu_3 \\
\mathbf{J}_3 \;&=\; -L_{31}\nabla\mu_1 - L_{32}\nabla\mu_2 - L_{33}\nabla\mu_3
\end{aligned} \tag{3.15}$$

and we have $L_{12} = L_{21}$, $L_{13} = L_{31}$, $L_{23} = L_{32}$.

The general formalism above is true for gas, liquid or solid-state diffusion, as long as convective contribution to flux is taken out. In the gas phase, according to kinetic theory of gases the diffusion rate is controlled by the rate of atomic collisions, where atoms can be thought of as hard spheres that move in space, that occupy much less volume than the empty spaces (free volume), so collisions are infrequent, mean free path is long and diffusion rate is high. In the liquid phase, the free volume on average is *less* than the occupied volume, so atomic collisions are frequent and diffusion is more difficult than in the gas phase. In liquids, free volume is typically spread out between clusters of atoms (this cluster of thirty atoms have higher average volume per atom than that cluster of forty atoms, see Fig. 3.1(b)), and is

not sharply localized at a certain *site*. Diffusion is faster in a local *region* with higher free volume. Envision the crowd at departmental Christmas party: the room is jam-packed, and for one person to move ("excuse me"), several persons have to adjust collectively. The zones in the room where there are bit more space on average would allow diffusion to happen faster *there*.

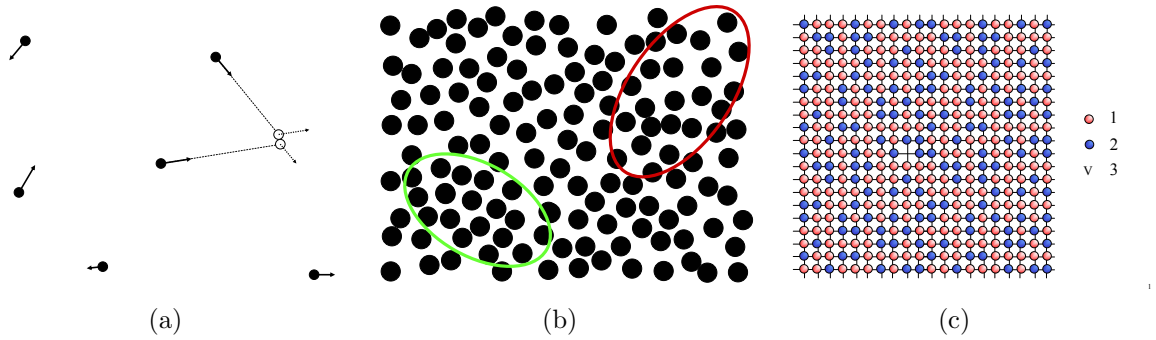


Figure 3.1: The concept of free volume in (a) gas (b) liquid (c) crystal.

Inside a crystal lattice, free volume is *sharply localized* as lattice site vacancies. Excess free volume in crystals also exists inside dislocation cores, grain boundaries and near surfaces, where the free volumes tend to be more delocalized, percolate in space and larger in magnitude compared to other parts of the crystal (i.e. vacancy nanoporosity). The trend of larger free volume $\rightarrow$ higher diffusion rate generally holds true in crystals as well. Thus, the relative ease of diffusion is ranked as surface diffusion $>$ grain boundary diffusion $\sim$ dislocation core (pipe) diffusion $>$ lattice diffusion. We will quantify this ranking later.

Inside crystalline lattice away from line and planar defects, by far the more common mechanism of diffusion is the exchange of atoms with vacancy, shown in Fig. 3.1(c). A careful analysis of the thermodynamics of vacancies is therefore critical for understanding solid-state diffusion. Before we proceed, we must make a distinction between atomic *sites* and *atoms* in a crystal. This distinction is similar to the difference between US government structure (white house, senate, supreme court etc.) with who are occupying the offices now. The government structure (site lattice) tends to be more permanent than the office holder, in crystalline solids; although sites can be destroyed as well, such as during climb of an edge dislocation. The sites can certainly be moved, which is the essence of plastic deformation (when we touch some object and feel it is deformed, we are not registering which labelled atom goes where, only the shifting of atomic site which are occupied by *some* atom - in other word our hand canot tell tracer or self diffusion).

A vacancy can be regarded as the occupation of a lattice site by a "Vacadium" species, denoted by V. In Fig. 3.1(c), a site is *always* occupied by either a red atom (1), a blue atom (2), or V (3). Thus:

$$X_1 + X_2 + X_V \;=\; 1. \tag{3.16}$$

Solution thermodynamics typically ignores the existence of X_V because X_V is small, often at ppm level and below, although near the melting temperature it can reach $\sim$0.1% [43]. But vacancies are more critical to the kinetics than to the thermodynamics. In the crystal observation frame, due to conservation of lattice sites there must be:

$$\mathbf{J}_1 + \mathbf{J}_2 + \mathbf{J}_3 \;=\; 0, \tag{3.17}$$

which means

$$(L_{11} + L_{21} + L_{31})\nabla\mu_1 + (L_{12} + L_{22} + L_{32})\nabla\mu_2 + (L_{13} + L_{23} + L_{33})\nabla\mu_3 \;=\; 0 \tag{3.18}$$

The above will be true in all situations if

$$L_{11} + L_{21} + L_{31} \;=\; 0, \quad L_{12} + L_{22} + L_{32} \;=\; 0, \quad L_{13} + L_{23} + L_{33} \;=\; 0, \tag{3.19}$$

or $\sum_i L_{ij} = 0$ for all j. And since $L_{ij} = L_{ji}$, we will also have:

$$L_{11} + L_{12} + L_{13} \;=\; 0, \quad L_{21} + L_{22} + L_{23} \;=\; 0, \quad L_{31} + L_{32} + L_{33} \;=\; 0, \tag{3.20}$$

or $\sum_j L_{ij} = 0$ for all i. Then we can simplify (3.15) as

$$\begin{aligned}
\mathbf{J}_1 \;&=\; -L_{11}\nabla(\mu_1 - \mu_3) - L_{12}\nabla(\mu_2 - \mu_3) \\
\mathbf{J}_2 \;&=\; -L_{21}\nabla(\mu_1 - \mu_3) - L_{22}\nabla(\mu_2 - \mu_3) \\
\mathbf{J}_3 \;&=\; -L_{31}\nabla(\mu_1 - \mu_3) - L_{32}\nabla(\mu_2 - \mu_3)
\end{aligned} \tag{3.21}$$

The above is the consequence of *network constraint* (see chap 2.2.2 of [41]), where the true compositional degrees of freedom are $N_c - 1$ instead of N_c, and thus there are only $N_c - 1$ driving forces. If we have 1-2(V) only (monatomic solid with vacancy), the equation would be simplified to be:

$$\begin{aligned}
\mathbf{J}_1 \;&=\; -L_{11}\nabla(\mu_1 - \mu_2) \\
\mathbf{J}_2 \;&=\; -L_{21}\nabla(\mu_1 - \mu_2) = L_{11}\nabla(\mu_1 - \mu_2)
\end{aligned} \tag{3.22}$$

where $L_{11} = -L_{12} = -L_{21} = L_{22} > 0$. Rewriting 2 as V, we would have

$$\mathbf{J}_1 \;=\; -L_{VV}\nabla(\mu_1 - \mu_V), \quad \mathbf{J}_V \;=\; L_{VV}\nabla(\mu_1 - \mu_V). \tag{3.23}$$

Now we can discuss about what controls μ_V. Consider a Kossel crystal with nearest-neighbor springs $u(r) = -\epsilon + k(r-a_0)^2/2$ and Z nearest neighbors ($Z = 4$ in 2D and 6 in 3D). At 0K, if there is no vacancy, each atom would have $e_1 = -Z\epsilon/2$ *cohesive energy* since each atom is connected to Z springs, shared with another atom. By creating vacancy, the total energy would have risen by $e_V = Z\epsilon/2$ per vacancy created, since when plucking out an atom from Kossel crystal Z springs are broken, but when we re-attach this atom to a surface ledge, $Z/2$ springs are formed anew. The total energy thus can be written as $E = N_1 e_1 + N_V e_V$ at 0K, so long as $N_V \ll N_1$ so the probability of two vacancies sitting side by side is small. At finite temperature, this vacancy formation energy e_V would be modified by vibrational contribution, so $e_V \to f_V^f$, the vacancy formation *free energy* (no configurational entropy contribution, only vibrational entropy contribution). Similarly, the cohesive energy e_1 will be modified by vibrational energy contribution, $e_1 \to f_1^\circ$. The total Helmholtz free energy of the system would thus look like:

$$F \;=\; N_1 f_1^\circ + N_V f_V^f + (N_1 + N_V) k_\mathrm{B} T (X_1 \ln X_1 + X_V \ln X_V) + ... \tag{3.24}$$

where $F_0 = N_1 f_1^\circ$ is a fully dense reference state with $N_V = 0$. At zero stress ($P = 0$), $G = F$, and $F = F_0 + N_V f_V^f + k_\mathrm{B} T (N_1 \ln X_1 + N_V \ln X_V)$ will be minimized at

$$f_V^f + k_\mathrm{B} T \left(\frac{N_1}{X_1} \cdot \frac{dX_1}{dN_V} + \frac{N_V}{X_V} \cdot \frac{dX_V}{dN_V} + \ln X_V \right) \;=\; 0. \tag{3.25}$$

or simply

$$f_V \equiv f_V^f + k_\mathrm{B} T \ln X_V \;=\; 0. \tag{3.26}$$

Note that $\mu_V \equiv f_V$ at zero stress, thus $\mu_V^{\text{boundary}} = 0$ if the RVE has reached equilibrium with the adjacent surface vacancy source/sink. Many textbooks call the vacancy formation free energy G_V^f, but the word Gibbs free energy is sometimes overused. In some occasions, when people say Gibbs free energy, they actually mean the Helmholtz free energy. At $P = 0$ the two are equivalent, but to keep the discussion clean we will stick to the $f_V \equiv f_V^f + k_\mathrm{B} T \ln X_V$ notation even at finite stress.

$\mu_V^{\text{boundary}} = 0$ because unlike in a typical A-B solution, where A and B have to come from

some mass sources, here the solid chooses its own optimal degree of porosity or atomic-scale free volume. To have more nanoporosity all the solid needs to do is to encroach on adjacent vacuum, which is in infinite supply at $P = 0$. So to reach equilibrium with the surface, the source of this vacuum, the boundary condition is just $f_V = \mu_V^{\text{boundary}} = 0$, where $f_V \equiv f_V^f + k_\text{B}T \ln X_V$. In this case then, $X_V = \exp(-f_V^f/k_\text{B}T)$, and a plot of $\ln X_V$ versus $1/T$ would give h_V^f/k_B, the vacancy formation enthalpy. There is then $f_V^f = h_V^f - Ts_V^f(\text{vib})$. s_V^f contains only the vibrational entropy contribution. In copper, h_V^f is about 1.27 eV, s_V^f is about $2.35k_\text{B}$. [3]

In this course the vacancy formation volume Ω_V^f, a concept parallel to the vacancy formation energy f_V^f, is assumed to be simply $\Omega_V^f = \Omega$, where Ω is the atomic volume. That is to say, for simplicity we will assume "Vacadium" is exactly as large as the solvent atom. Or, there is zero *vacancy relaxation volume* $\Omega - \Omega_V^f$ after we pluck out an atom, which is true in the Kossel crystal. Then we have $c = N/V = 1/\Omega$, $X_V = c_V/c = c_V\Omega$. And so the equilibrium vacancy concentration at zero stress is $c_V^0 = \Omega^{-1} \exp(-f_V^f/k_\text{B}T)$.

If there is normal traction $t_{nn} = \mathbf{n}\boldsymbol{\sigma}\mathbf{n} = \sigma_{nn}$ on the surface[1], then vacuum no longer comes at zero price. The thermodynamic balance then requires $f_V \equiv f_V^f + k_\text{B}T \ln X_V = t_{nn}\Omega$ to linear order in stress, which gives $X_V = \exp(-(f_V^f - t_{nn}\Omega)/k_\text{B}T)$, whereby tension favors more vacancy (since work can be done on the boundary when nanoporosity is created inside), and compression favors less vacancy. The boundary traction will need to be equilibrated mechanically with the internal stress. If we identify $t_{nn}\Omega$ as work, then still $\mu_V^{\text{boundary}} = f_V - t_{nn}\Omega = f_V^f + k_\text{B}T \ln X_V - t_{nn}\Omega = 0$. This may be understood by the following argument: $\mu_V = \partial G/\partial N_V$, where G is the total thermodynamic potential including surface work. Initially, when $X_V = 0$, μ_V is very negative, so by taking in additional porosity, $N_V \to N_V + 1$, G decreases. The solid can take as much nanoporosity as it wants, and this only stops when μ_V approaches 0. Taking in more porosity then would make the total thermodynamic potential G go back up again.

The effects of internal stress on f_1° and f_V^f are 2nd order in stress (strain energy), which in most cases may be ignored, whereas stress come into the *boundary condition* as $f_V^f = t_{nn}\Omega$, which is linear order in stress if $t_{nn} \neq 0$. In the case of uniform hydrostatic pressure, $t_{nn} = -P$ for whichever exposed surface, so the solid body can be in global thermodynamic equilibrium if the vacancy density is uniform $X_V = \exp(-(f_V^f + P\Omega)/k_\text{B}T)$. On the other hand, if the solid is in uniaxial tension or shear, the solid body can *never be* in global thermodynamic

[1]Traction $\mathbf{t}$ is a boundary quantity, $\boldsymbol{\sigma}$ is a bulk quantity: even though the two are related by $\mathbf{t} = \boldsymbol{\sigma}\mathbf{n}$ on the surface, their physical meaning are distinct.

equilibrium. This is because the local equilibrium value of X_V would depend on which surface the RVE is adjacent to (which "market" the RVE is "trading with", and like people, the closest market is the most important one). There will be more vacancies near surface under tensile normal traction, and less vacancies near surface under compressive normal traction. The vacancy flux will move to surface under compression, which will drive deformation of the solid by diffusional creep.

In (3.22), if we take 2 to be V, then $\mathbf{J}_1 = -L_{VV}\nabla(\mu_1 - \mu_V)$, $\mathbf{J}_V = -L_{VV}\nabla(\mu_V - \mu_1)$. What drives diffusion inside the body is always $\mu_{V1} \equiv \mu_V - \mu_1$ instead of μ_1 or μ_V alone. Note that μ_{V1} corresponds to a magical operation of directly extracting atom from inside the material: $(N_1, N_V) \to (N_1 - 1, N_V + 1)$ without the necessity of putting the atom on a surface ledge [2]. As such μ_{V1} is entirely *local* and depend only on $\mathbf{x}$, as it should be for any quantity used in the PDE. If $\Omega_1 = \Omega_V = \Omega$ (no relaxation volume), then in terms of strain, the μ_{V1} operation is silent, thus the exchange potential μ_{V1} would have no dependence on $\boldsymbol{\sigma}(\mathbf{x})$ in the linear order, and it would depend only on concentration:

$$\mu_{V1} \;=\; f_V^\circ - f_1^f + k_\mathrm{B}T \ln \frac{X_V}{X_1} \;=\; f_V^\circ - f_1^f + k_\mathrm{B}T \ln \frac{X_V}{1 - X_V} \tag{3.27}$$

Note that $\ln(1 - X_V)$ is not a sensitive function of X_V when X_V is very small. In contrast, $\ln X_V$ diverges violently as $X_V \to 0$. Thus $\mu_{1V} \approx f_1^\circ - f_V^f - k_\mathrm{B}T \ln X_V = f_1^\circ - f_V^f - k_\mathrm{B}T \ln \Omega c_V$, and $\nabla\mu_{1V} = -k_\mathrm{B}T/c_V \nabla c_V$. So,

$$-\mathbf{J}_1 = \mathbf{J}_V = -L_{VV} k_\mathrm{B}T/c_V \nabla c_V. \tag{3.28}$$

Define $D_V \equiv L_{VV} k_\mathrm{B}T/c_V$, we get $\mathbf{J}_V = -D_V \nabla c_V$ (Fick's 1st law), and so

$$\partial_t c_V \;=\; \nabla \cdot (D_V \nabla c_V). \tag{3.29}$$

The above assumes vacancies are only created/annihilated on the surface or grain boundary, and once they move inside the lattice can only be transported but not created/annihilated. If there *are* internal vacancy sources/sinks inside the crystal besides planar surfaces or grain boundaries, for instance around the half-planes of edge dislocations (another "market" with which the RVE can trade atomic-scale porosity), then (3.29) needs to be modified appropri-

[2] Note the canonical surface vacancy creation process is $(N_1, N_V) \to (N_1, N_V + 1)$ where we do worry about the ledge. To make $(N_1, N_V) \to (N_1 - 1, N_V + 1)$ less magical, what we actually do is to take the extracted atom to an isloated atom dump (infinitely dilute gas).

ately:

$$\partial_t c_V \;=\; \nabla \cdot (D_V \nabla c_V) + (\partial_t c_V)_{\text{source}}. \tag{3.30}$$

where $(\partial_t c_V)_{\text{source}}$ reflects the rates of internal creation/annihilation of atomic-scale porosity/free volume. Typically, $(\partial_t c_V)_{\text{source}}$ would induce motion of the lattice planes, as would the surface / grain boundary sources (imagine playing a game of Tetris).

Consider a square block of solid of size d under uniaxial tension $\sigma_{11} > 0$, $\sigma_{22} = 0$. RVEs near the vertical surfaces have $c_V = c_V^0$. RVEs near the horizontal surfaces have $c_V = c_V^0 \exp(t_{nn}\Omega/k_{\text{B}}T) \approx c_V^0(1 + t_{nn}\Omega/k_{\text{B}}T)$, when t_{nn} is very small. Thus, ∇c_V will be of the order $c_V^0 t_{nn}\Omega/k_{\text{B}}Td$, and the vacancy flux will be of the order $D_V c_V^0 t_{nn}\Omega/k_{\text{B}}Td$. Because each vacancy arriving at vertical surface destination would cause local sink-in of volume Ω, the displacement rate would be of the order $D_V c_V^0 t_{nn}\Omega^2/k_{\text{B}}Td$ and strain rate would be of the order $\dot{\epsilon} \sim D_V c_V^0 t_{nn}\Omega^2/k_{\text{B}}Td^2$. If we recognize $D_V c_V^0 \Omega = D_V X_V^0 \sim D^*$, the self diffusivity of type-1 atoms (to be shown later), then the creep strain rate of this block of solid is simply $\dot{\epsilon} \sim (D^*\Omega/k_{\text{B}}Td^2)t_{nn}$. This is called Nabarro-Herring creep [44]. The creep rate is supported by lattice vacancy transport, and is proportional to the inverse 2nd power of the sample size or grain size, where vacancy flows from its source (high X_V^{boundary}) to its sink (low X_V^{boundary}).

In addition to surfaces and grain boundaries, edge dislocations (half-planes in crystals) can also be source and sink of vacancies. When edge dislocation acts as vacancy source, its half plane extends, and simultaneously a vacancy appears in the adjoining crystal. When it acts as sink of vacancies, its half plane shrinks (see Fig. 3.3(b) of [41]). Thus, if there are plenty of edge dislocations of different Burgers vectors inside the crystal, the crystal can also creep with strain rate $(D^*\Omega/k_{\text{B}}Td^2)t_{nn}$, where d is now the average spacing from dislocation vacancy source to dislocation vacancy sink. Having plenty of dislocations inside the crystal (half-planes for vacancy to be generated near and annihilated with) would ensure $\mu_V \approx 0$ everywhere inside the crystal, if the atomic processes of emitting and absorbing vacancies near the dislocation core is not too difficult (that is, if it is not so-called *reaction limited* kinetics). Also note that just like vacancies annihilating on surface would induce macroscopic motion ($\dot{\epsilon} \sim t_{nn}D^*\Omega/k_{\text{B}}Td^2$), vacancies annihilating on half-planes would also induce macroscopic motion, $\mathbf{v}_{\text{C}}^{\text{L}}$.

One also have the situation where local σ_{nn} is generated by the Young-Laplace pressure, then $f_V \equiv f_V^f + k_{\text{B}}T \ln X_V = -\gamma\kappa\Omega$, where $\kappa = R_1^{-1} + R_2^{-1}$, and $X_V = \exp(-(f_V^f + \gamma\kappa\Omega)/k_{\text{B}}T)$. Thus hill top would have less vacancy, valley would have more vacancy, which will drive vacancy flux uphill (= atom flux downhill), and smooth out the surface. Note that if there

is no vacancy relaxation volume: $\Omega - \Omega_V^f = 0$, then μ_{1V} and internal diffusion is independent of pressure. A detailed mathematical treatment of surface smoothing is given in Chap. 14.1 of [41].

Since $\mathbf{J}_V = c_V \mathbf{v}_V$ (this is still C-frame), we have $\mathbf{v}_V = -L_{VV} k_\mathrm{B} T / c_V^2 \nabla c_V = L_{VV}/c_V(-\nabla \mu_V)$. Since $-\nabla \mu_V$ can be identified as the thermodynamic driving force for vacancy motion, we can define $M_V \equiv L_{VV}/c_V$ to be the *vacancy mobility*. We have previously defined $D_V \equiv L_{VV} k_\mathrm{B} T / c_V$. Thus we have derived the Einstein relation $M_V = D_V / k_\mathrm{B} T$.

Among all three inter-related transport quantities, $L_{VV} = -L_{1V} = -L_{V1} = L_{VV}$, D_V and M_V, M_V has the most direct physical interpretation. If the vacancy has charge, as in ionic crystals, M_V may be measured by applying an external electric field to the crystal and checking how much faster the vacancies move on average. In the limit of dilute c_V, M_V should be nearly a constant of c_V, since the driven motion of one vacancy should be independent from the driven motion of another vacancy, as they seldom cross each other's path when vacancy concentration is so dilute:

$$M_V(c_V) \;=\; M_V^0 + \mathcal{O}(c_V) \tag{3.31}$$

Therefore

$$D_V(c_V) \;=\; k_\mathrm{B} T M_V(c_V) \;=\; D_V^0 + \mathcal{O}(c_V). \tag{3.32}$$

This means $L_{VV} = -L_{1V} = -L_{V1} = L_{VV} = c_V M_V$ must scale as c_V for small c_V. Within the small c_V approximation then, D_V may be taken out of $\partial_t c_V = \nabla \cdot (D_V \nabla c_V) = D_V^0 \nabla^2 c_V$ as the leading order term. This is still in the C-frame.

How to physically interpret D_V^0? Let us perform a thought experiment, and see what the field equation $\partial_t c_V = D_V^0 \nabla^2 c_V$ implies. Imagine we put a single vacancy at the origin at $t = 0$. This like a delta-function in the initial vacancy concentration. We know the solution to $\partial_t c_V = D_V^0 \nabla^2 c_V$ in the case of 1D is $c_V(x,t) = \frac{1}{\sqrt{4\pi D_V^0 t}} \exp(-\frac{x^2}{4 D_V^0 t})$. Since $c_V(x,t)$ is normalized, $\int_{-\infty}^{\infty} dx c_V(x,t) = 1$, it can be understood as a probability density of finding the vacancy migrating to x at time t, given that it was at 0 at time 0. The mean squared displacement is thus

$$\int_{-\infty}^{\infty} dx x^2 c_V(x,t) \;=\; 2 D_V^0 t. \tag{3.33}$$

In the case of 3D, the normalized concentration / probability density is just

$$c_V(x,y,z,t) \;=\; \frac{1}{(4\pi D_V^0 t)^{3/2}} \exp\left(-\frac{x^2+y^2+z^2}{4D_V^0 t}\right) \tag{3.34}$$

with mean squared displacement

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdydz(x^2+y^2+z^2)c_V(x,y,z,t) \;=\; 6D_V^0 t. \tag{3.35}$$

This is what the macroscopic field equation tells us. But what is the *microscopic basis* for this vacancy's mean squared displacement?

Atoms are always rattling around the vacancy, and once in a while a nearest-neighbor atom jumps across to fill this vacancy, but also leaving an empty spot behind. We then say "the vacancy has hopped" or "the vacancy has interchanged with an nearest-neighbor atom". In 1D, the vacancy can hop left or right. In 2D, the vacancy can hop up, down, left, right ($Z=4$), etc. The rate of successful hop to one specific nearest neighbor (say right) can be modeled as $\Gamma'_V = \nu\exp(-\frac{g_V^m}{k_{\rm B}T})$, where ν is a physical attempt frequency (typically $\sim 10^{12}$/s) and g_V^m is the vacancy migration free energy barrier: $g_V^m \equiv G^* - G$, where G^* is the system's total free energy at saddle point (when vacancy has moved *halfway* from one lattice site to an adjacent lattice site). When there is no external bias (zero driving force), Γ'_V should be the same in all Z channels, so the total hop rate of the vacancy is $\Gamma_V = Z\Gamma'_V$. The vacancy will essentially be performing unbiased random walk on the site lattice, each hop labeled by an integer $k = 1..K$, where $K = \Gamma_V t \gg 1$. Let us define $\mathbf{r}_k$ to be the vectorial hopping distance at kth-hop. In 1D, $\mathbf{r}_k$ would be a_0 or $-a_0$ with equal probability. In 2D, $\mathbf{r}_k$ would be $(a_0,0)$, $(0,a_0)$, $(-a_0,0)$, $(0,-a_0)$ with equal probability, etc. If we let $\mathbf{x}_V(t)$ be the position of the vacancy at time t, then

$$\mathbf{x}_V(t) \;=\; \mathbf{r}_1 + \mathbf{r}_2 + ... + \mathbf{r}_K. \tag{3.36}$$

In the theory of probability, if A and B are two random variables, then $\mathrm{E}[A+B] = \mathrm{E}[A] + \mathrm{E}[B]$, where $\mathrm{E}[X]$ means the expectation (average) value of X. Thus $\mathrm{E}[\mathbf{x}_V(t)] = \mathrm{E}[\mathbf{r}_1] + \mathrm{E}[\mathbf{r}_1] + ... + \mathrm{E}[\mathbf{r}_K] = 0$, which means the centroid of $\mathbf{x}_V(t)$ distribution is still 0. The variance of X is defined as $\mathrm{Var}[X] \equiv \mathrm{E}[(X - \mathrm{E}[X])^2]$. If A and B are *independent* random variables, there is further $\mathrm{Var}[A+B] = \mathrm{Var}[A] + \mathrm{Var}[B]$. So $\mathrm{Var}[\mathbf{x}_V(t)] = \mathrm{Var}[\mathbf{r}_1] + \mathrm{Var}[\mathbf{r}_1] + ... + \mathrm{Var}[\mathbf{r}_K] = K\mathrm{Var}[\mathbf{r}_k] = Ka_0^2$, where a_0 is the Kossel crystal lattice constant. This is because $\mathbf{r}_k \cdot \mathbf{r}_k = a_0^2$ with probability 1 when the vacancy is hopping on the simple cubic Kossel lattice (in BCC, $Z=8$, $\mathrm{E}[\mathbf{r}_k] = 0$ and $\mathrm{Var}[\mathbf{r}_k] = 3a_0^2/4$; in FCC, $Z=12$, $\mathrm{E}[\mathbf{r}_k] = 0$ and $\mathrm{Var}[\mathbf{r}_k] = a_0^2/2$).

The final piece of the puzzle is to apply the so-called central limit theorem from the theory of probability, which states that if many random variables are added together, the probability distribution function of that *sum* will approach Gaussian, no matter how the individual random variables are distributed:

$$dP(\mathbf{x}_V(t) \text{ in } dxdydz) \;=\; \frac{dxdydz}{(2\pi\sigma^2)^{3/2}} \exp\left(-\frac{x^2+y^2+z^2}{2\sigma^2}\right), \tag{3.37}$$

with $3\sigma^2 = \text{Var}[\mathbf{x}_V(t)] = Ka_0^2 = \Gamma_V t a_0^2$. Matching the field equation solution (3.35) with that from discrete random walk, we may identify $6D_V^0 t$ in (3.35) as $\Gamma_V t a_0^2$, so $D_V^0 = \Gamma_V a_0^2/6$.

A more mundane interpretation of $D_V^0 = \Gamma_V a_0^2/6$ is also possible, but this case we imagine a *driven* system, where there is concentration gradient ∇c_V, and thus *finite* thermodynamic driving force. Specifically, imagine $\nabla c_V = (\partial_x c_V, 0, 0)$, and consider two adjacent atomic planes at $x = 0$ and $x = a_0$, respectively. The plane at $x = 0$ will have on average $c_V(x = 0)a_0$ vacancies per unit area, whereas the plane at $x = a_0$ will have on average $c_V(x = a_0)a_0$ vacancies per unit area, and the two are generally not equal. The number of vacancies hopping from left plane $\rightarrow$ right plane is $c_V(x = 0)a_0\Gamma'_V$, the number of vacancies hopping from the right plane $\rightarrow$ left plane is $c_V(x = a_0)a_0\Gamma'_V$, so the net vacancy flux (if measurement is done at $x = 0.5a_0$) will be $(c_V(x = 0) - c_V(x = a_0))a_0\Gamma'_V \approx -\partial_x c_V a_0^2 \Gamma'_V$. On the other hand, the field equation says the flux should be $-D_V^0 \partial_x c_V$, so we identify $D_V^0 = a_0^2\Gamma'_V$.

In the random walk model of diffusion, the tagged vacancy just wanders around without any driving force or preferred direction of motion. In the alternative derivation, there is a macroscopic concentration gradient and chemical potential driving force. We see the two models give identical result for D_V^0, which is a manifestation of so-called *fluctuation-dissipation* theorem, which states that equilibrium fluctuations (equilibrium position fluctuations in the case of random walk) are governed by the same laws as an externally driven system (macroscopic $\partial_x c_V$ and $\nabla\mu_V$) if the driving force is small. The Einstein relation $M = D/k_\text{B}T$ is a manifestation of this idea as well, where M characterizes the velocity of a driven object, and D characterizes the random-walk response of that object when there is no driving force. Aside from this insight, the random walk model also gives us a microscopic physics expression $D_V^0 = a_0^2\Gamma'_V$. Thus, the slope of $\ln D_V^0$ versus $1/T$ would give us h_V^m/k_B, the enthalpy of vacancy migration, with $g_V^m = h_V^m - Ts_V^m$, where s_V^m is the entropy of vacancy migration.

Now we can discuss about self-diffusion in solids. Self-diffusion means that instead of identifying a vacancy and tracking its motion $\mathbf{x}_V(t)$, we "tag" an atom and tracks its motion. The way self-diffusion is measured experimentally is to use radioactive isotope $2 = 1^*$, which

is chemically identical to 1. The system 1-2(1*)-3(V) satisfies:

$$J_1 \;=\; -L_{11}\nabla\mu_{1V} - L_{12}\nabla\mu_{2V},\;\; J_2 \;=\; -L_{21}\nabla\mu_{1V} - L_{22}\nabla\mu_{2V}. \tag{3.38}$$

For example, in a diffusion couple experiment, one weld two bar together, one bar has $c_1^{-\infty}$, $c_{1^*}^{-\infty}$, the other has $c_1^{\infty} = c_1^{-\infty} + c_{1^*}^{-\infty}$ (Fig. 3.1 of [41]), and then heat it up for some hours for 1^* to diffuse inward. If the system is under $P = 0$, the vacancy density should be uniform everywhere $c_V^0 = \Omega^{-1}\exp(-f_V^f/k_{\mathrm{B}}T)$, since the vacancy can't "tell" the difference between 1^* and 1, and would have no motivation to change its statistical and dynamical behavior before/after c_1 is replaced by c_{1^*}.

Generally we also have the Gibbs-Duhem relation:

$$X_1 d\mu_1 + X_2 d\mu_2 + X_V d\mu_V \;=\; 0 \tag{3.39}$$

for an isothermal isobaric system, with only $\mathbf{X}$ is changing. Therefore

$$X_1\nabla\mu_1 + X_2\nabla\mu_2 + X_V\nabla\mu_V = 0 \;\;\rightarrow\;\; X_1\nabla(\mu_1-\mu_V) + X_2\nabla(\mu_2-\mu_V) = -\nabla\mu_V. \tag{3.40}$$

If $\mu_V = 0 = \frac{\partial G}{\partial N_V}|_{N_1,N_2}$ everywhere (vacancies thermodynamically equilibrated with surface/GB/dislocation sources), we will have

$$X_1\nabla(\mu_1-\mu_V) + X_2\nabla(\mu_2-\mu_V) = 0 \;\;\rightarrow\;\; \nabla\mu_{2V} \;=\; -\frac{c_1}{c_2}\nabla\mu_{1V}. \tag{3.41}$$

Plugging this back into (3.38), we get

$$J_1 \;=\; -(L_{11} - \frac{L_{12}c_1}{c_2})\nabla\mu_1, \quad J_2 \;=\; -(L_{22} - \frac{L_{21}c_2}{c_1})\nabla\mu_2, \tag{3.42}$$

or applying this to 1-1*-V system

$$J_1 \;=\; -(L_{11} - \frac{L_{11^*}c_1}{c_{1^*}})\nabla\mu_1, \quad J_{1^*} \;=\; -(L_{1^*1^*} - \frac{L_{1^*1}c_{1^*}}{c_1})\nabla\mu_{1^*}. \tag{3.43}$$

Because 1-1* always mix ideally no matter what is c_{1^*}, there is

$$\nabla\mu_1 \;=\; \frac{k_{\mathrm{B}}T}{c_1}\nabla c_1, \quad \nabla\mu_{1^*} \;=\; \frac{k_{\mathrm{B}}T}{c_{1^*}}\nabla c_{1^*}, \tag{3.44}$$

and so

$$J_1 \;=\; -k_{\mathrm{B}}T(\frac{L_{11}}{c_1} - \frac{L_{11^*}}{c_{1^*}})\nabla c_1, \quad J_{1^*} \;=\; -k_{\mathrm{B}}T(\frac{L_{1^*1^*}}{c_{1^*}} - \frac{L_{1^*1}}{c_1})\nabla c_{1^*}. \tag{3.45}$$

Define self diffusivity as

$$D^* \;\equiv\; k_{\mathrm{B}}T(\frac{L_{1^*1^*}}{c_{1^*}} - \frac{L_{1^*1}}{c_1}), \tag{3.46}$$

we get Fick's 1st law for radioactive tracers:

$$J_{1^*} \;=\; -D^*\nabla c_{1^*}. \tag{3.47}$$

Even though vacancies seem to disappear from the derivations above, we will argue below based on physical grounds that in fact $D^* = fX_VD_V$, where f is a correlation correction factor of order 1 (Chap. 8.2 of [41]). The argument is the following. If we track a tagged *atom*: its rate of changing site is actually much lower than the average hopping rate of a vacancy Γ'_V, because unless there is a vacancy just next to this tagged atom, there is no chance for the atom to hop. The unconditional probability of finding a vacancy right next to our tagged atom is ZX_V, so the total rate of atom hopping is approximately $ZX_V\Gamma'_V$. The method of random walk then gives us $D^* \approx ZX_V\Gamma'_Va_0^2/6 = X_VD_V$. From this we see that unlike the vacancy diffusivity itself $D_V \approx D_V^0$, the self-diffusivity D^* is actually a very sensitive function of X_V. Ultimately this is because self-diffusion in substitutional alloys is physically a *side effect* of vacancy diffusion. This also illustrates the point that although vacancy is almost always negligible thermodynamically (on μ of other species; μ_V does care about X_V), it is almost always non-negligible kinetically.

There is a tricky point in the derivation above, which is that even if the "vacancy hops" are truly uncorrelated, the atom jumps would still be somewhat correlated. The reason can be seen from the following: if we do not know where the vacancy is coming from (left, right, up, down), the atom would jump left, right, up, down with equal probability. However, once we know what $\mathbf{r}_1$ is for the atom jump, for example $\mathbf{r}_1 = (a_0, 0)$, we now know that immediately after the jump, the vacancy must be immediately to the left of the atom. This makes $\mathbf{r}_2$ more likely to be $(-a_0, 0)$, the so-called "back flow". Thus, $\mathbf{r}_1$ and $\mathbf{r}_2$ for the atom jumps would be correlated somewhat negatively. This has an effect of reducing the self diffusivity $D^* = fX_VD_V$, where $0 < f \leq 1$ is a correlation correction factor. This is treated more extensively in Chap. 8.2 of [41], and f is found to be 0.78 in FCC crystals and 0.73 in BCC crystals.

The self-diffusivity D^* is widely used because it can be experimentally measured. When we

plot $\ln D^*$ versus $1/T$, the slope is $(h_V^f + h_V^m)/k_\mathrm{B}$, the sum of the enthalpy of formation and the enthalpy of migration of the vacancy. In copper, h_V^m is about 0.71 eV, which means the effective activation enthalpy $h_V^f + h_V^m$ is about 2 eV since h_V^f is about 1.27 eV. [3]

In self-diffusion, there is no vacancy flux: $J_V = 0$, since whether atoms surrounding the vacancy are radioactive or not have no bearing on the vacancy trajectory $\mathbf{x}_V(t)$. In such case, there is no macroscopic lattice plane motion inside the crystal, $\mathbf{v}_\mathrm{C}^\mathrm{L} = 0$, and C-frame and L-frame are identical. Now let us discuss an example where this is not the case, which leads to the famous Kirkendall effect [45, 46]. Imagine 1-2-V system where 1 is chemically different from 2, in particular imagine 1-V exchange is much easier than 2-V exchange. For simplicity let us assume this difference in rate of exchange comes entirely from $h_V^m(1) \neq h_V^m(2)$, and not in the vacancy formation energy. That is to say, unless the vacancy attempts to hop, it cannot tell the difference between 1 and 2. Then even though there is a concentration gradient in 1 vs 2, say the left side is richer in 1 and the right side is richer in 2, we have $c_1(x) + c_2(x) \approx 1$ and c_V should be approximately constant throughout the sample. We also assume there is no volume difference between 1 and 2: $\Omega_1 = \Omega_2 = \Omega_V^f = \Omega$.

Consider again two atomic planes at $x = 0$ and $x = a_0$, and there is a concentration gradient $\partial_x c_2$, so the $x = a_0$ plane has more type-2 atoms than the $x = 0$ plane. This will be *compensated mainly* by $-\partial_x c_1$, since $X_1 + X_2 + X_V = 1$, and X_V cannot exceed a certain limit, say 0.1% before voids start to appear, while change in X_2 in the sample can be tens of percent. Thus, $|\partial_x c_2| \approx |\partial_x c_1| \gg |\partial_x c_V|$. In fact, if we assume there are plenty of dislocation sources inside the crystal, then $\mu_V \approx 0$, and X_V is uniform (the non-uniform X_V case of Nabarro-Herring creep is because there is macroscopic stress bias on different surfaces; when there is no appreciable global stress bias, as in typical diffusion couple experiment, X_V should be nearly uniform). We thus expect there will more type-1 atoms on $x = 0$ plane than on $x = a_0$ plane. Vacancies on $x = a_0$ plane will therefore see more type-1 atoms and less type-2 atoms on the $x = 0$ plane, whereas vacancies on $x = 0$ plane will see more type-2 atoms and less type-1 atoms on the $x = a_0$ plane. Since 1-V exchange is facile while 2-V exchange is sluggish, there will be *more* vacancies hopping from right plane to left plane than vice versa. Thus, there will be a net *vacancy* flux $J_V \neq 0$ observed in the crystal frame. The nano porosities are being *pumped* backward, to the 1-rich side. We have seen in our previous discussion of Nabarro-Herring creep that $J_V \neq 0$ has the ability of modifying the site lattice by "tetris"-like action and inducing a macroscopic motion of the lattice planes $\mathbf{v}_\mathrm{C}^\mathrm{L}(x) \approx \dot{\epsilon} x$. This is the same story in the Kirkendall effect.

Formally, we have

$$J_V^{\mathrm{C}} \;=\; -J_1^{\mathrm{C}} - J_2^{\mathrm{C}}, \tag{3.48}$$

where we add the $^{\mathrm{C}}$ superscript back on. (3.42) is still applicable, but the thermodynamic interactions between 1-2 may be non-ideal, so

$$\mu_1 \;=\; \mu_1^{\circ} + k_{\mathrm{B}}T\ln\gamma_1 X_1, \quad \mu_2 \;=\; \mu_2^{\circ} + k_{\mathrm{B}}T\ln\gamma_2 X_2 \tag{3.49}$$

where γ_1, γ_2 are the activity coefficients. The effect of uniform and small X_V is ignored in μ_1 and μ_2 expression. (Generally speaking, X_V is not important *thermodynamically* to any species' chemical potential *except for* μ_V, although X_V is diffusion kinetically critical to *all* species). Therefore

$$d\mu_1 \;=\; k_{\mathrm{B}}T(1+\frac{d\ln\gamma_1}{d\ln c_1})\frac{dc_1}{c_1}, \quad d\mu_2 \;=\; k_{\mathrm{B}}T(1+\frac{d\ln\gamma_2}{d\ln c_2})\frac{dc_2}{c_2}. \tag{3.50}$$

Plugging it back into (3.42), we get

$$J_1^{\mathrm{C}} = -k_{\mathrm{B}}T(\frac{L_{11}}{c_1}-\frac{L_{12}}{c_2})(1+\frac{d\ln\gamma_1}{d\ln c_1})\nabla c_1, \quad J_2^{\mathrm{C}} = -k_{\mathrm{B}}T(\frac{L_{22}}{c_2}-\frac{L_{21}}{c_1})(1+\frac{d\ln\gamma_2}{d\ln c_2})\nabla c_2. \tag{3.51}$$

Using the Gibbs-Duhem relation it can be shown that the thermodynamic factors $\frac{d\ln\gamma_1}{d\ln c_1} = \frac{d\ln\gamma_2}{d\ln c_2}$.

Define **intrinsic diffusivities**

$$D_1 \equiv k_{\mathrm{B}}T(\frac{L_{11}}{c_1}-\frac{L_{12}}{c_2})(1+\frac{d\ln\gamma_1}{d\ln c_1}), \quad D_2 \equiv k_{\mathrm{B}}T(\frac{L_{22}}{c_2}-\frac{L_{21}}{c_1})(1+\frac{d\ln\gamma_2}{d\ln c_2}), \tag{3.52}$$

we have

$$J_1^{\mathrm{C}} \;=\; -D_1\nabla c_1, \quad J_2^{\mathrm{C}} \;=\; -D_2\nabla c_2. \tag{3.53}$$

As we previously discussed, $\nabla c_1 \approx -\nabla c_2$, but D_1 can be larger than D_2, therefore $J_1^{\mathrm{C}} \neq J_2^{\mathrm{C}}$. So

$$J_V^{\mathrm{C}} \;=\; D_1\nabla c_1 + D_2\nabla c_2. \tag{3.54}$$

What is the physical effect of J_V^{C}? If we imagine all these nano porosities are dumped directly on the free surface on the left end of the diffusion couple, this would induce a sink-in/pop-out velocity of $-J_V^{\mathrm{C}}\Omega$ of that free surface. If that free surface is held fixed, the whole sample would then shift with velocity $\mathbf{v}_{\mathrm{C}}^{\mathrm{L}} = J_V^{\mathrm{C}}\Omega$. In reality, the vacancy flux $J_V^{\mathrm{C}}(x)$ does not need

to go all the way to the free surface to annihilate, they can annihilate on edge dislocations as shown in Fig. 3.3 of [41]. But it can be seen by direct inspection, that whether annihilating inside or the end surface would give the same $\mathbf{v}_{\rm C}^{\rm L} = J_V^{\rm C}(x)\Omega$ (like in the game of tetris), if the end surface is held fixed in the laboratory frame.

Thus,

$$J_1^{\rm L} \;=\; J_1^{\rm C} + c_1\mathbf{v}_{\rm C}^{\rm L} = -D_1\nabla c_1 + X_1(D_1\nabla c_1 + D_2\nabla c_2) \approx -(X_2D_1 + X_1D_2)\nabla c_1 \tag{3.55}$$

$$J_2^{\rm L} \;=\; J_2^{\rm C} + c_2\mathbf{v}_{\rm C}^{\rm L} = -D_2\nabla c_2 + X_2(D_1\nabla c_1 + D_2\nabla c_2) \approx -(X_2D_1 + X_1D_2)\nabla c_2 \tag{3.56}$$

where we ignored the small X_V. The vacancy flux is $J_V^{\rm L} = J_V^{\rm C} + c_V\mathbf{v}_{\rm C}^{\rm L} = J_V^{\rm C} + X_V J_V^{\rm C} \approx J_V^{\rm C} = D_1\nabla c_1 + D_2\nabla c_2 = (D_2 - D_1)\nabla c_2$. If we define **interdiffusivity** $\tilde{D} \equiv X_2D_1 + X_1D_2$, a lab-frame quantity, then

$$J_1^{\rm L} = -\tilde{D}\nabla c_1, \quad J_2^{\rm L} = -\tilde{D}\nabla c_2, \tag{3.57}$$

which is what we really need in solving most diffusion problems. Fick's 2nd law says that, in lab frame (the conservation of "red Ferraris" works in lab frame, or in any uniformly translating frame - that is to say if the two police patrol cars move with identical speed - but not if with different speeds):

$$\partial_t c_1 \;=\; \nabla\cdot(\tilde{D}\nabla c_1), \quad \partial_t c_2 \;=\; \nabla\cdot(\tilde{D}\nabla c_2). \tag{3.58}$$

$\tilde{D}$ should be a function of composition $\tilde{D}(X_2)$, where we assumed X_V (degree of porosity) take the equilibrium value for given X_2. From the Onsager coefficient representation, we also know that $\tilde{D}$ should have similar order of magnitude as the self-diffusivity D^* (1 or 2), which means that $\tilde{D} \propto X_V$. (3.58) interdiffusion in the lab frame is the starting point for solving most diffusion problems.

How to experimentally measure the interdiffusivity $\tilde{D}(X_2)$ as a function of composition? Matano devised a graphical method, where $\tilde{D}(X_2)$ can be determined from a single diffusion couple experiment:

$$\partial_t c_2 = \partial_x(\tilde{D}\partial_x c_2), \quad x \in (-\infty, \infty), \tag{3.59}$$

$$c_2(x, t=0) \;=\; \begin{cases} c_2^{-\infty}, & x < 0 \\ c_2^{\infty}, & x > 0 \end{cases}. \tag{3.60}$$

The first step in this analysis is the so-called Boltzmann transform. Define $\eta \equiv x/2\sqrt{t}$.

There is reversible mapping $(\eta, t) \leftrightarrow (x, t)$. For any function $y(\eta, t) = y(x/2\sqrt{t}, t)$, there is

$$\partial_t y \;=\; \partial_\eta y \times \frac{-x}{4t^{3/2}} + D_t y, \tag{3.61}$$

where $D_t y \equiv \partial y/\partial t|_\eta$. Similarly

$$\partial_x y \;=\; \partial_\eta y \times \frac{1}{2\sqrt{t}}. \tag{3.62}$$

Thus (3.59) can be rewritten as

$$D_t c_2 - \frac{\eta}{2t}\partial_\eta c_2 \;=\; \frac{1}{4t}\partial_\eta(\tilde{D}\partial_\eta c_2). \tag{3.63}$$

Because the initial profile (3.60) has no intrinsic lengthscale (the interface of $c_2^{-\infty} \to c_2^{\infty}$ is infinitely sharp), Boltzmann argued that any $c_2(x, t)$ we see at finite time must already have fallen into a *self-similar attractor profile* $c_2(x, t) = c_2(\eta)$, and $D_t c_2 = 0$. This self-similar attractor profile would satisfy:

$$-2\eta\partial_\eta c_2 \;=\; \partial_\eta(\tilde{D}\partial_\eta c_2). \tag{3.64}$$

Without the D_t dependence, ∂_η is really $d/d\eta$, so

$$-2\eta\frac{dc_2}{d\eta} \;=\; \frac{d}{d\eta}\left(\tilde{D}\frac{dc_2}{d\eta}\right). \tag{3.65}$$

Thus,

$$-2\int_{-\infty}^{\eta} d\eta'\eta'\frac{dc_2}{d\eta'} \;=\; \tilde{D}(X_2)\frac{dc_2}{d\eta}, \tag{3.66}$$

since we know $\tilde{D}\partial_\eta c_2$ is zero at minus infinity. Thus

$$\tilde{D}(X_2) \;=\; -2\frac{d\eta}{dc_2}\int_{-\infty}^{\eta}\eta' dc_2 \;=\; -\frac{1}{2t}\frac{dx}{dc_2}\int_{-\infty}^{x} x' dc_2(x'). \tag{3.67}$$

On the right-hand side since we know dc_2/dx approaches zero at plus infinity as well, and $\tilde{D}$ must be finite at plus infinity, there must be:

$$0 \;=\; \int_{-\infty}^{\infty} x' dc_2. \tag{3.68}$$

The $x = 0$ plane is also called the Matano plane or the Matano interface. It is a **stationary plane in the lab frame**, and does not correspond to any lattice plane (since all lattice planes are moving). The Matano plane is the *lab-frame* location where the two sides of the

diffusion couple are first joined. Formula (3.67) is called the Boltzmann-Matano analysis.

To experimentally measure the interdiffusivity using (3.67) requires knowing where the origin $x = 0$ is, where the two sides first met. If you are given a sample that is the outcome of a diffusion-couple experiment, but the person who did the experiment has died, so no one remembers where $x = 0$ is, what are you going to do? (think in forensics or archaeological context). Well, you can measure the concentration profile $c_2(\tilde{x})$ using EDS, where $\tilde{x}$ is with respect to your current, arbitrarily chosen, origin. Suppose $\tilde{x} = \Delta$ is where the Matano plane is, then $\tilde{x} = x + \Delta$ (coordinate transformation between current frame and original experimenter's frame). Then, obviously

$$\int_{-\infty}^{\infty} \tilde{x}' dc_2(\tilde{x}') \;=\; \int_{-\infty}^{\infty} (x + \Delta) dc_2 = 0 + \Delta(c_2^{\infty} - c_2^{-\infty}), \tag{3.69}$$

$$\Delta \;=\; \frac{\int_{-\infty}^{\infty} \tilde{x}' dc_2(\tilde{x}')}{c_2^{\infty} - c_2^{-\infty}}. \tag{3.70}$$

Equation (3.70) tells you where is the Matano plane in your current, arbitrarily chosen, frame, even if you did not do the experiment yourself. Thus, we see that the terminology "Matano plane" is not a completely trivial thing, even though initially it may sound trivial.

In above we have considered the microscopic physics of lattice diffusion coefficient D_{lattice}, where free volumes are in the form of localized lattice vacancies, the concentration and mobility of which control the effective mobility of atoms in lattice. It is typical to write

$$D_{\text{lattice}}(T) \;=\; D^0_{\text{lattice}} \exp\left(-\frac{h^*_{\text{lattice}}}{k_{\text{B}}T}\right) \tag{3.71}$$

where both D^0_{lattice} and h^*_{lattice} are *temperature-independent* to leading order. Note that the effective activation energy $g^*(T) \equiv g^f_{\text{V}} + g^m_{\text{V}}$ has both enthalpic and entropic contributions: $g^*(T) = h^* - Ts^*$. The s^* term is due to vibrational entropy, including vibrational entropy of the saddle-point configuration when vacancy migrates. Thus, the slope of $\exp(-\frac{g^*(T)}{k_{\text{B}}T})$ in log scale versus $1/T$ is not $g^*(T)$, because $g^*(T)$ is itself a function of T, due to finite s^*. On the other hand, we may assume h^* and s^* to be independent of T, effectively truncating a Taylor expansion of $g^*(T)$ in T to first order. Then, $\exp(-\frac{g^*(T)}{k_{\text{B}}T}) = \exp(\frac{s^*}{k_{\text{B}}})\exp(-\frac{h^*}{k_{\text{B}}T})$, and the $\exp(\frac{s^*}{k_{\text{B}}})$ can be absorbed (together with the physical trial frequency ν) into the temperature-independent prefactor D^0_{lattice}.

In above we have dealt with lattice diffusion in substitutional alloys, specifically vacancy-exchange dominated lattice diffusion. Prior to Smigelskas and Kirkendall's famous experi-

ment [45], people thought that interdiffusion in alloy occur by the direct-exchange mechanism, that is, a Cu atom and an adjacent Ni atom directly swap position without the aid of vacancy. Smigelskas and Kirkendall's experiment proved that vacancies play a huge role is material kinetics: as a side effect, it also causes motion of the lattice - a surprising effect like continental drift or glacier motion.

Lattice diffusion may also happen by the motions of interstitials. If one follows the life of an interstitial (after it is created, before it is annihilated, at surface/GB/climbing dislocations), it also performs random walk, with an apparent interstitial diffusivity $D_I = M_I k_\mathrm{B} T$. The contribution of interstitials to lattice diffusivity can also be estimated as $\Delta D_\mathrm{lattice} \propto X_I D_I$ (the atom at the interstitial can exchange/kick out atom on the lattice, causing mixing of lattice atoms). The total lattice diffusivity can thus be estimated as $D_\mathrm{lattice} \propto X_I D_I + X_V D_V$. The equilibrium interstitial concentration is estimated as $X_I^0 = \exp(-g_I^f/k_\mathrm{B}T)$, the interstitial diffusivity may be estimated as $D_I = \nu \exp(-g_I^m/k_\mathrm{B}T)a_0^2$, so the total temperature sensitivity of the $X_I D_I$ is governed by $h_I^f + h_I^m$, the formation energy plus the migration energy of interstitial. In typical metals, the migration energy of interstitial is much lower than that of vacancy, for example, in Cu $h_I^m = 0.1\mathrm{eV}$ [3]. However, the formation energy of interstitial is much higher, $h_I^f = 3\mathrm{eV}$ [3]. Thus the total activation parameter for interstitial-exchange diffusion is usually higher than that for vacancy-exchange diffusion, for substitutional alloys (in interstitial alloys like Fe-C system it's another story), inside crystals near point-defect equilibrium. However, in irradiated materials where the point-defect distribution could be far from equilibrium, interstitial-exchange could cause abnormally high diffusivity. Also note that, for interstitials, there is no network constraint, and the Onsager equations would look somewhat different. We will not delve into the details here.

Since there are also free volumes near surface (D_surface), in grain boundaries (GBs, D_GB) and dislocation cores (D_core), which often form percolating paths, we expect

$$D_\mathrm{surface}(T) \gg D_\mathrm{GB}(T) \sim D_\mathrm{core}(T) \gg D_\mathrm{lattice}(T) \tag{3.72}$$

This is because although the trial frequencies of atom hops are comparable or even a bit lower near surface, GB and dislocation cores, the activation enthalpies for atom hops should

$$h^*_\mathrm{surface} < h^*_\mathrm{GB} \sim h^*_\mathrm{core} < h^*_\mathrm{lattice} \tag{3.73}$$

such that the rate of successful hops is much higher and (3.72) holds true at all $0 < T < T_\mathrm{melt}$. In FCC metals, $h^*_\mathrm{GB} \sim 0.5 h^*_\mathrm{lattice}$. So mass transport is much accelerated near these extended

defects, which are capable of forming percolating networks.

On the other hand, there are less number of atoms near these extended defects (surface, GB, dislocation) than atoms in the lattice. In the case of GB enhanced diffusion, envision the idealized geometry shown in Fig. 2.26 of [47], where the grain boundary thickness is δ, and the rest is lattice crystal of thickness $d - \delta$. Atoms inside δ layer are supposed to diffuse faster down a concentration gradient $\partial_x c$. Typically, we can take the grain boundary thickness to be something like 3Å, and d is the grain size (something like 1μm). The total flux averaged over the d cross-section would be

$$J_{\text{app}} \;=\; -\frac{D_{\text{GB}}\delta + D_{\text{lattice}}(d-\delta)}{d}\partial_x c \;\equiv\; -D_{\text{app}}\partial_x c \tag{3.74}$$

$$D_{\text{app}} \;=\; D_{\text{lattice}} + (D_{\text{GB}} - D_{\text{lattice}})\frac{\delta}{d} \;\approx\; D_{\text{lattice}} + D_{\text{GB}}\frac{\delta}{d} \tag{3.75}$$

the approximation in the end is because the absolute magnitude of D_{GB} is always much larger than D_{lattice}. From Fig. 2.27 of [47] we see that there is then a transition temperature T_{trans} for diffusion rate in polycrystal, above which $D_{\text{app}} \approx D_{\text{lattice}}$, below which $D_{\text{app}} \approx D_{\text{GB}}\delta/d$, with different temperature slopes. $T_{\text{trans}} \sim 0.75 - 0.8T_{\text{melt}}$ in many materials. The same behavior holds true if we have a single-crystal nanowire of diameter d and *surface thickness* δ, in which atoms hop faster. We see that interfacial or surface mass transport paths will always dominate over lattice crystal mass transport at low enough temperatures (or small enough d's) because of lower activation enthalpies.

In the case of dislocation core diffusion (aka pipe diffusion), the correspondent quantity to δ/d would be $\rho\delta^2$, which is the ratio of atoms in the dislocation core to total number of atoms, where ρ is dislocation density and δ is dislocation core thickness:

$$D_{\text{app}} \;=\; D_{\text{lattice}} + (D_{\text{core}} - D_{\text{lattice}})\rho\delta^2 \;\approx\; D_{\text{lattice}} + D_{\text{core}}\rho\delta^2 \tag{3.76}$$

ρ in mediumly work-hardened Cu is typically on the order of $10^{14}/\text{m}^2$ (number of etch pits per unit area) = $10^{14}\text{m}/\text{m}^3$ (dislocation line length per m^3 of material - in reference, circumference of earth is 4×10^7m, circumference of sun is 4×10^9m). Taking δ =3Å, we see that the dimensionless quantity $\rho\delta^2 \sim 10^{-5}$, i.e. per hundred thousand atoms in the lattice, there is one atom in the dislocation core. Although this weighting factor looks small, at low enough temperatures $D_{\text{core}}(T)$ can become so much larger relative to $D_{\text{lattice}}(T)$ that mass transport (for whatever rate of mass transport that can occur at such low temperatures) will be governed by dislocation core diffusion instead of lattice diffusion.

In MSE we emphasize solids, but fluids are also worth a brief mention. Even under optical microscope, one can see small particles embedded in a fluid (pollen in water, fat droplet in milk) executing agitated random motion, with no perceivable macroscopic forcing. This provides direct visualization of random walk, aka Brownian motion, and thermal fluctuation forces. Consider a fat droplet of size r (around μm): the British fluid dynamicist George Gabriel Stokes derived a relation between drag force F and steady state velocity of a sphere embedded in a continuum Newtonian fluid ($\sigma = \eta\dot{\epsilon}$) of viscosity η as $\mathbf{F} = 6\pi r\eta\mathbf{v}$. Einstein wrote it as $\mathbf{v} = \frac{1}{6\pi r\eta}\mathbf{F}$ and identify $\frac{1}{6\pi r\eta}$ as the mobility M (not mass) of the fat droplet in the fluid medium. Then, using Einstein formula $D = Mk_{\mathrm{B}}T$, he predicted

$$D \;=\; \frac{k_{\mathrm{B}}T}{6\pi r\eta} \tag{3.77}$$

which agrees perfectly with the measured random-walk characteristics of the fat droplets. (3.77) is called the Stokes-Einstein formula. Although (3.77) was intended for larger particles like pollens or fat droplets, people attempt to correlate macroscopic hydrodynamic quantity η with self-diffusivity of molecules. For example, liquid water has $\eta = 10^{-3}$ Pa·s (centipoise) at 20°C, which one can measure macroscopically with a viscometer, and water molecule size can be taken to 2Å. Plugging into (3.77) gives one 10^{-9} m^2/s. The actual self-diffusivity of water molecule in liquid water is about 2×10^{-9} m^2/s [48, 49].

One expects diffusion in solids, even surface diffusion $D_{\mathrm{surface}}(T)$, to be smaller than diffusion in liquids. Thus, 10^{-9} m^2/s or 10^{-5} cm^2/s can be taken as the upper bound on solid-state diffusivity. For example, yttria stabilised zirconia (YSZ) with yttria doping and around 10% oxygen vacancies on the anion sub-lattice, is known to be a fast oxygen ion conductor. The magnitude of oxygen diffusivity is about 3×10^{-6} cm^2/s in YSZ at the highest temperatures of 2000K [50]. An alternative way to think about this number is that $D \sim X_V \nu a_0^2 e^{-g/k_{\mathrm{B}}T}$: with trial frequency $\nu = 10^{12}$/s (atomic vibration frequency), $a_0 = 3 \times 10^{-10}$m, we see that even when the success rate of barrier hopping per trial reaches 1/10, that is, one succeeds for every 10 trials of mounting the barrier, D would be 10^{-9} m^2/s. When the success rate is 1/10, the diffusing species can really be thought as more "fluid" than "solid". The typical success rate in solids per trial is much lower, though, especially at lower temperatures.

Chapter 4

Capillary Energy Effects

All previous discussions ignored the role of surfaces (area and shape), which are asymptotically correct for large bodies with infinite volume-to-area ratio. Actual bodies are not infinite. Define surface free energy as the *excess* Helmholtz free energy:

$$F(\mathbf{N}, T, V, \mathbf{A}) \equiv F_{\text{bulk}}(\mathbf{N}, T, V) + \int_{\mathbf{A}} dA\gamma(\mathbf{n}, \mathbf{N}, T, V, \mathbf{A}) \tag{4.1}$$

where $\mathbf{A}$ denotes the area and shape of V's border, and

$$F_{\text{bulk}}(\mathbf{N}, T, V) \equiv \lim_{\lambda\to\infty} \frac{F(\lambda^3\mathbf{N}, T, \lambda^3 V, \lambda^2\mathbf{A})}{\lambda^3}, \tag{4.2}$$

is $\mathbf{A}$-independent. The above are well-defined recipes for $F_{\text{bulk}}(\mathbf{N}, T, V)$ and γ, given state function $F(\mathbf{N}, T, V, \mathbf{A})$, and arbitrary but consistent choices of the dividing surface $\mathbf{A}$ and V. The dependent variables of γ is a bit involved. γ definitely depends on $\mathbf{n}$, as well as T, chemistry and density of the substrate.

In the case of a liquid, $\gamma(\mathbf{n}) = \gamma$,

$$F = F_{\text{bulk}} + \gamma A. \tag{4.3}$$

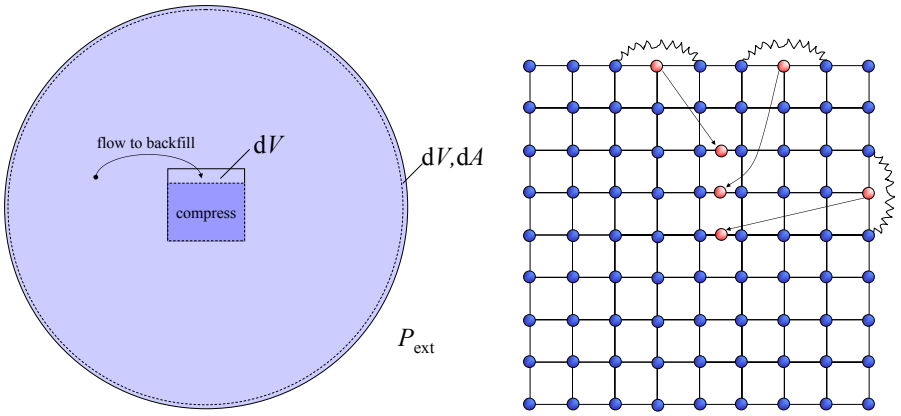
If $\gamma > 0$, the *shape* that minimizes G is clearly a sphere: $V = 4\pi R^3/3$, $A = 4\pi R^2$.

What determines R? This may sounds like a trivial question since you know the number of

atoms, and can write down something reasonable like

$$N\bar{v} \;=\; \frac{4\pi R^3}{3} \tag{4.4}$$

the question is then what is $\bar{v}$. The inside of the liquid can choose to shrink against its own internal pressure, to reduce the outside surface area (like you would wrap up in a fetal position when a bunch of thugs attack you - you *generate* internal pressure to minimize surface pain - if surface pain is extreme, you may shrink smaller).



(a) (b)

Figure 4.1: The origin of Young-Laplace pressure: (a) in liquid droplet (b) in Kossel crystal with free surface

Consider the thought experiment in Fig. 4.1(a). The square RVE squeeze by dV, its Helmholtz free energy increases by $P_{\text{int}}dV$. The normal liquid backflow to occupy the vacated volume, simultaneously reducing the exposed surface area by dA. The external pressure does work $P_{\text{ext}}dV$. To reach equilibrium:

$$P_{\text{ext}}dV + \gamma dA \;=\; P_{\text{int}}dV, \tag{4.5}$$

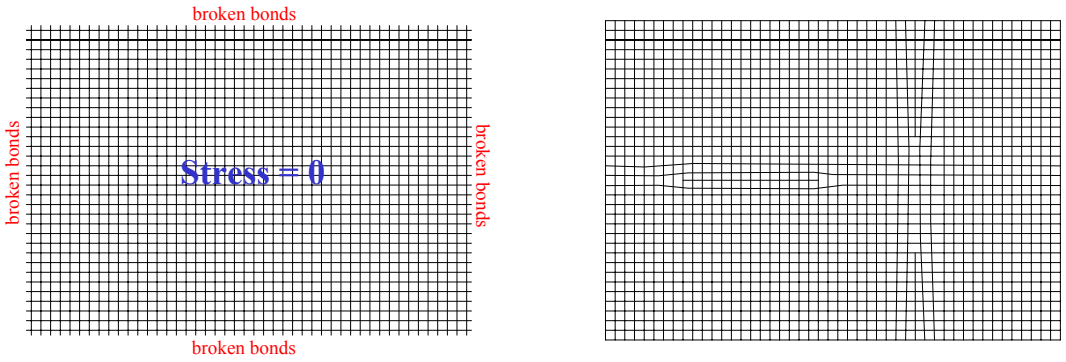
so

$$P_{\text{int}} \;=\; P_{\text{ext}} + \gamma\frac{dA}{dV} \;=\; P_{\text{ext}} + \frac{2\gamma}{R}, \tag{4.6}$$

The above pressure difference is called Young-Laplace pressure, which can be derived purely mechanically.

For the Young-Laplace relation to work, the ability of the material to *flow* is important.

Consider the following paradox. There is a simple cubic crystal with nearest neighbor springs $u(r) = -\epsilon + k(r - a_0)^2/2$ (Kossel crystal). The equilibrium lattice constant is just a_0. If you cut out the surfaces, you break surface springs and there is finite surface energy ($\epsilon/2a_0^2$). But there is no internal pressure generated, at least not immediately - if the crystal is to remain elastic (no change in bonding topology) - because elastic shrinkage in the center causing elastic displacement on the outside do not reduce the number of broken bonds (the *surface stress* of solid turns out to be zero in this case, despite of finite *surface energy*). If however there are plenty of dislocations inside the crystal, which can help to reconfigure and eliminate the number of surface sites with broken bonds by absorbing and emitting vacancies, then over a long time, an internal pressure wil be generated, and that *despite of finite interal elastic strain energy thus created* will still have lower total energy than the original stress-free configuration because of less number of surface sites (imagine.



(a) (b)

Figure 4.2: (a) A Kossel crystal with broken bonds at surface will still have interal lattice constant a_0 and zero interal stress, despite of finite surface energy due to the surface dangling bonds. (b) A solid is able to shear as well as eliminate surface sites by dislocation climb.

In the literature, you see three terms frequently: *surface energy*, *surface stress* and *surface tension*. For liquid surfaces all three terms mean the same thing. Surface tension basically means isotropic surface stress, like hydrostatic pressure is a special case of general stress state in 3D. For solid surfaces, one must be very careful: surface stress (force per length that is attributed to a geometric surface) is not necessarily surface energy, and does not have to be isotropic. Also, whether the Young-Laplace pressure exists depends on the timescale of observation. If the timescale of observation is long enough that diffusion is allowed to happen and surface sites reconfigured (see Fig. 4.1(b)), then the solid behaves more like a

liquid (for even "the mountains flowed before the Lord"[51]), and surface stress may start to be related to the surface energy. On the other hand, in short timescale and with no bond reconfiguration, the surface stress does not have to be the surface energy and Young-Laplace pressure may not be established inside the solid, as the Kossel crystal paradox showed.

Consider now the chemical potential change of atoms in the finite sized body. Since

$$\left.\frac{\partial \mu_i}{\partial P}\right|_{\mathbf{X},T} \;=\; v_i \tag{4.7}$$

For an RVE of composition $\mathbf{X}$, compared to the same RVE embedded in an infinite particle, the chemical potential would have risen by

$$\Delta\mu_i \;=\; v_i \frac{2\gamma}{R}. \tag{4.8}$$

The above is called the Gibbs-Thomson effect (Young and Laplace were mechanicians in 1805, but chemical potential concept was only invented after Gibbs work in 1876 at Yale). It will play a role in driving bulk and surface diffusion.

Near a general surface

$$\Delta\mu_i \;=\; \gamma v_i \left(\frac{1}{R_1}+\frac{1}{R_2}\right) \;=\; \gamma v_i(\kappa_1+\kappa_2), \tag{4.9}$$

where κ_1 and κ_2 are the two *principal curvatures*, since there is the general geometric relation

$$\frac{dA}{dV} \;=\; \kappa_1(\mathbf{x})+\kappa_2(\mathbf{x}). \tag{4.10}$$

Note that (4.9) is a *local* condition, and $R_1, R_2, \kappa_1, \kappa_2$ all depend on the position $\mathbf{x}$ of the surface: $R_1(\mathbf{x}) = 1/\kappa_1(\mathbf{x})$, $R_2(\mathbf{x}) = 1/\kappa_2(\mathbf{x})$.

Consider a 1D height profile $h(x)$ of a surface. We will show in below that the local curvature $K(x) = 1/R(x) \approx -\partial_x^2 h$, if $h(x)$ is a gently varying curve, with small $|h|$. The reason that curvature is related to the *2nd-order* derivative is because a straight line profile $h(x) = a+bx$, which has finite 0th and 1st-order derivatives, has *no* curvature. Curvature is defined by a local fit to $h(x)$ by a perfect circular arc. Consider $h(0) = h'(0) = 0$ (since 0th and 1st-order derivatives are unimportant to curvature value), and we would like to fit to $h(x)$ by a circle that is tangent to the horizontal axis at $x = 0$. With such a circle of radius R, we would have $(R-h)^2 + x^2 = R^2$, or $-2Rh + h^2 + x^2 = 0$. For $|h| \ll |x| \ll R$, we can ignore the h^2

term, and have $h \approx x^2/2R$, so $h''(0) = 1/R$, Q.E.D.

Consider now a thin film with no internal stress attached to a bottom substrate, with an initially undulating top surface height profile $h(x)$. If the surface energy is isotropic, the total capillary energy $\gamma \int dx\sqrt{1 + (h'(x))^2}$ can be reduced by taking a flat $h(x) = h_0$, and choosing h_0 such that the total number of atoms is conserved. This is the "macroeconomic" view of why surface undulations should disappear for a stress-free thin film. The question is what are the "microeconomic" mechanisms for undulations decay by diffusion. That is, what are in it for the individual atoms at different locations to smooth out the profile undulations. Note that according to the sign convention in (4.6), there should be an extra negative sign $K(x) = 1/R(x) \approx -\partial_x^2 h$ if the material is *below* the surface. Then, in this quasi-1D problem, $\kappa_1(x) = -\partial_x^2 h$, $\kappa_2(x) = 0$, and the Gibbs-Thomson effect says that $\Delta\mu = -\gamma\partial_x^2 h\Omega$, in reference to same system with flat surface. Thus, RVEs immediately beneath the profile crest (negative $\partial_x^2 h$) will have positive $\Delta\mu$, whereas RVEs beneath the profile valley (positive $\partial_x^2 h$) will have negative $\Delta\mu$. Thus, there will be a driving force for atoms near the crest to migrate to the bottom. The migration can occur by bulk diffusion, e.g. vacancy mechanism, but it can also occur by surface diffusion.

Let us consider the case of mass transport dominated by surface diffusion, due to for example low temperature or small-wavelength undulations. In this case we consider there is a surface skin channel of width δ where atom mobilities are high ($D_{\text{surface}}/k_{\text{B}}T \gg D_{\text{bulk}}/k_{\text{B}}T$). Atoms in the skin channel must be equilibrated with RVEs immediately beneath the skin, and therefore see the same thermodynamic driving force $-\partial_x\Delta\mu = \gamma\Omega\partial_x^3 h$ to move. We have flux in the skin channel to be $J_{\text{surface}} = c_{\text{surface}}v_{\text{surface}} = c_{\text{surface}}(D_{\text{surface}}/k_{\text{B}}T)(-\partial_x\Delta\mu) = (\gamma c_{\text{surface}}\Omega D_{\text{surface}}/k_{\text{B}}T)\partial_x^3 h = (\gamma D_{\text{surface}}/k_{\text{B}}T)\partial_x^3 h$. The last equality is because $c_{\text{surface}}\Omega \approx 1$ (we assumed the surface channel has comparable atom density as bulk: a small difference of tens of per cent can anyhow be absorbed into the definition of D_{surface}). So the total rate of "red Ferrari" transport across the surface skin channel is $J_{\text{surface}}(x)\delta$. If $J_{\text{surface}}(x)\delta$ is a constant across a certain range of x, then we just have uniform transport of matter with no accumulation. On the other hand, if $J_{\text{surface}}(x + \Delta x)\delta \neq J_{\text{surface}}(x)\delta$, then we will have accumulation of "red Ferraris". The rate of accumulation is $-\partial_x(J_{\text{surface}}(x)\delta)$ per unit distance of x. Since each "red Ferrari" carries volume Ω, the rate of surface height increase would be

$$\partial_t h \;=\; -\Omega\partial_x(J_{\text{surface}}\delta) \;=\; -(\gamma\delta D_{\text{surface}}\Omega/k_{\text{B}}T)\partial_x^4 h \tag{4.11}$$

It is easy to check that the unit works out.

Define $B \equiv \gamma\delta D_{\text{surface}}\Omega/k_{\text{B}}T$ (unit m^4/s, Eq. 14.10 of [41]), the equation $\partial_t h = B_{\text{surface}}\partial_x^4 h$ describes so-called *curvature-driven surface diffusion*. In contrast to the typical diffusion equation derived from Fick's 1st and 2nd law, $\partial_t c = D\partial_x^2 c$, which is a parabolic PDE (parabolic spatial derivative), curvature-driven flow is a quartic PDE. That is, the highest-order spatial derivative is 4th order. This difference is fundamentally because $\partial_t c^* = D^*\partial_x^2 c^*$ is driven by *thermochemistry*, whereas $\partial_t h = -B\partial_x^4 h$ is driven by *geometry* and the Young-Laplace pressure *derived from geometry*. Introduction of capillary energy now adds an $\mathbf{A}$-dependent term in the chemical potential, which is now *geometry-sensitive* for finite-sized objects. The power of the Onsager formalism which depends on $-\nabla\mu$ and mobility M / Onsager coefficient L, rather than the concentration gradient $-\nabla c$ and diffusivity D, should now be more apparent. We see that two very different equations, $\partial_t h = -B_{\text{surface}}\partial_x^4 h$ and $\partial_t c = D\partial_x^2 c$, can both be derived from the same grandaddy Onsager equation. Therefore, physically, $-\nabla\mu$ and M are definitely the more fundamental quantities to remember than $-\nabla c$ and D and Fick's 1st law, though in practice $-\nabla c$ and D are probably more convenient to use and measure. So the connection from $-\nabla\mu$ Onsager language to $-\nabla c$ Fick language is important to demonstrate, which we have already done in Sec. 3.

Suppose we are given initial height profile $h(x, t = 0) = h_0 + a_0 \sin(kx)$, we may guess the solution to $\partial_t h = -B\partial_x^4 h$ to be $h(x, t) = h_0 + a(t)\sin(kx)$. This guess is called *separation of variables*, and it happens to work out:

$$\frac{da(t)}{dt}\sin(kx) \;=\; -Ba(t)(-k^2)^2\sin(kx) \tag{4.12}$$

$$\frac{da(t)}{dt} = -Bk^4 a(t) \;\;\rightarrow\;\; \frac{d\ln a(t)}{dt} = -Bk^4 \;\;\rightarrow\;\; a(t) = a(t=0)e^{-Bk^4 t} \tag{4.13}$$

Thus $h(x, t) = h_0 + a_0 e^{-Bk^4 t}\sin(kx)$. Because $\partial_t h = -B\partial_x^4 h$ is linear PDE (it is nonlinear in ∂_x operator but linear in h), the solutions are additive. According to Fourier, almost any initial profile can be decomposed into sine and cosine of different wavevectors:

$$h(x, t = 0) \;=\; \int_0^\infty dk(a_0(k)\sin(kx) + b_0(k)\cos(kx)) \tag{4.14}$$

including initial profiles with sharp steps. So the finite-time solution is simply:

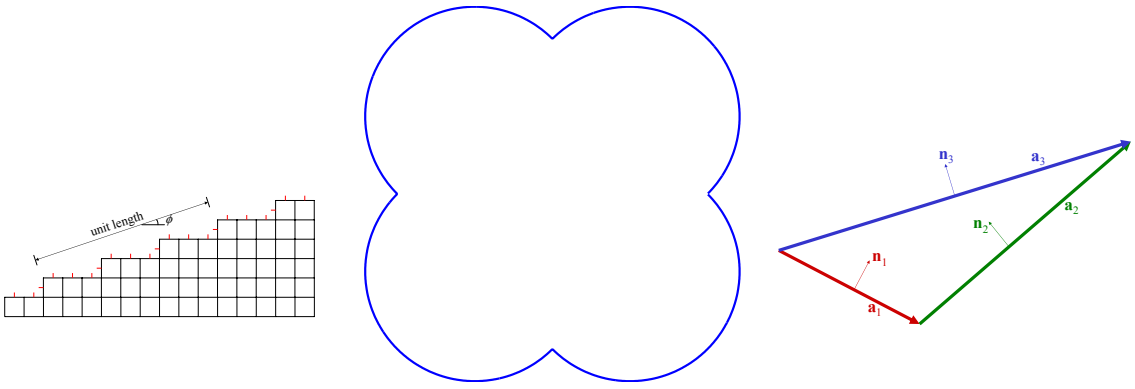
$$h(x, t = 0) \;=\; \int_0^\infty dk\left(a_0(k)e^{-Bk^4 t}\sin(kx) + b_0(k)e^{-Bk^4 t}\cos(kx)\right) \tag{4.15}$$

We see that the smaller-wavelength component (sharp features) dies out much faster than longer-wavelength component. By increasing the wavelength by a factor of 2, the amplitude

decay halflife increases by a factor of 16. (The same is true for $\partial_t c^* = D^* \partial_x^2 c^*$ for composition modulations: here increasing the wavelength by a factor of 2 increases the decay halflife by a factor of 4.) Indeed, for $k = 0$ component (the average height profile h_0), the amplitude does not decay at all. Note that $\{\sin(kx), \cos(kx)\}$ or $\{e^{ikx}\}$ is a good basis for infinite domain $x \in (-\infty, \infty)$ problems, where there is no boundary condition, or periodic domain problems, where the boundary condition is trivial. For problems with nontrivial boundary conditions (finite spatial support), $\{e^{ikx}\}$ is no longer good spatial basis and other *eigenfunction* basis suitable for this particular boundary condition would be needed. Separation of variables approach still works in those situations though, as long as the PDE is linear.

Wulff plot: $\gamma(\mathbf{n})\mathbf{n}$, and inverse Wulff plot: $\gamma^{-1}(\mathbf{n})\mathbf{n}$.

Kossel crystal show that surface energy naturally have $\sin|\phi|$ type singularities, with cusps (locally minimal surface energy) occurring at certain special ϕ's that have especially well packed surface structure ($\{111\}$, $\{110\}$, $\{100\}$ surfaces in FCC crystals). When ϕ deviates just a little bit (either $+$ or $-$) from these special angles, there will be crystallographic *ledges* whose density is $\propto \sin|\Delta\phi|$, causing a singular cusp in the energy vs ϕ plot. Such singularity is due to crystallography, and ultimately, the discreteness of atoms.



(a) (b) (c)

Figure 4.3:

Stability of a certain thin film surface (constrained on substrate) against decomposition. Consider $\mathbf{a}_1 + \mathbf{a}_2 = \mathbf{a}_3$. First we would like to show

$$a_1\mathbf{n}_1 + a_2\mathbf{n}_2 \;=\; a_3\mathbf{n}_3 \tag{4.16}$$

where $|\mathbf{a}_1| = a_1$, $|\mathbf{a}_2| = a_2$, $|\mathbf{a}_3| = a_3$. Since $a_1\hat{\mathbf{a}}_1 + a_2\hat{\mathbf{a}}_2 = a_3\hat{\mathbf{a}}_3$, we only need to apply 90°

rotation matrix $\mathbf{R}$ to both left and right-hand side to prove (4.16). There is a more general proof (applicable to tetrahedron in 3D) using Gauss theorem. Define all $\mathbf{n}_i$ of a polyhedra to be pointing outward. The claim is that

$$\sum_i A_i \mathbf{n}_i \;=\; 0. \tag{4.17}$$

The proof is to consider

$$\mathbf{b}\cdot\sum_i A_i\mathbf{n}_i \;=\; \int_{\text{surface}} dA\mathbf{b}\cdot\mathbf{n} = \int_{\text{body}} d^3\mathbf{x}(\nabla\cdot\mathbf{b}) \;=\; 0. \tag{4.18}$$

for arbitrary $\mathbf{b}$. So (4.17) must be true, and (4.16) is a 2D special case, with normal of 1,2 defined inward as shown in Fig. 4.3(c).

Now the energy of 1+2 combination is $\gamma_1 a_1 + \gamma_2 a_2$. Define

$$\gamma_3^* \;\equiv\; \frac{\gamma_1 a_1 + \gamma_2 a_2}{a_3} \tag{4.19}$$

If the actual $\gamma_3 > \gamma_3^*$, the $\mathbf{n}_3$ facet would be unstable against decomposition into 1+2. However, the geometric equality (4.16) could be rewritten as

$$a_1\gamma_1\gamma_1^{-1}\mathbf{n}_1 + a_2\gamma_2\gamma_2^{-1}\mathbf{n}_2 \;=\; a_3\gamma_3^*\gamma_3^{*-1}\mathbf{n}_3 = \gamma_1 a_1\gamma_3^{*-1}\mathbf{n}_3 + \gamma_2 a_2\gamma_3^{*-1}\mathbf{n}_3 \tag{4.20}$$

So:

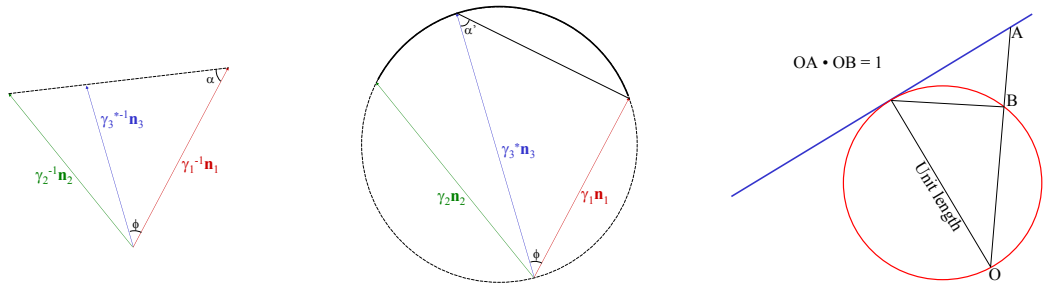
$$a_1\gamma_1(\gamma_1^{-1}\mathbf{n}_1 - \gamma_3^{*-1}\mathbf{n}_3) \;=\; a_2\gamma_2(\gamma_3^{*-1}\mathbf{n}_3 - \gamma_2^{-1}\mathbf{n}_2) \tag{4.21}$$

which means $\gamma_3^{*-1}\mathbf{n}_3$ must be on the straightline connecting $\gamma_1^{-1}\mathbf{n}_1$ and $\gamma_2^{-1}\mathbf{n}_2$. If the actual γ_3^{-1} lies *inside* of this γ_3^{*-1} line segment, then γ_3 will be unstable against decomposition.

So when we plot the inverse Wulff plot, $\gamma^{-1}(\mathbf{n})\mathbf{n}$. Any facet that is inside the common tangent construction of $\gamma^{-1}(\mathbf{n})\mathbf{n}$ will be unstable against decomposition (read p. 346-349, 608-615 of [41], ignore the discussion about the capillary vector $\xi(\mathbf{n})$). Note that it is possible to adjust the relative position of 1+2 to 3, such that beneath 3 contains exactly the same number of atoms.

Define the angle between $\mathbf{n}_3$ and $\mathbf{n}_1$ to be ϕ. From the law of sine in inverse Wulff plot, we get

$$\frac{\sin(\pi-\alpha-\phi)}{\gamma_1^{-1}} \;=\; \frac{\sin\alpha}{\gamma_3^{*-1}}. \tag{4.22}$$



(a) (b) (c)

Figure 4.4: .

In above ϕ is variable as $\mathbf{n}_3$ scans between $\mathbf{n}_1$ and $\mathbf{n}_2$, but α is constant, set by $\gamma_1^{-1}\mathbf{n}_1$ and $\gamma_2^{-1}\mathbf{n}_2$. We may rewrite the equation then as

$$\gamma_3^*(\phi) \;=\; \gamma_1 \frac{\sin(\pi-\alpha-\phi)}{\sin\alpha}. \tag{4.23}$$

It turns out that $\gamma_3^*(\phi)$ must be part of a circle which goes through three points: the origin, $\gamma_1\mathbf{n}_1$ and $\gamma_2\mathbf{n}_2$. This can proven by the following, consider Fig. 4.4(b). Let us call the angle shown in Fig. 4.4(b) as α'. By the law of sine, we have

$$\frac{\sin(\pi-\alpha'-\phi)}{\gamma_3^*(\phi)} \;=\; \frac{\sin\alpha'}{\gamma_1} \quad\rightarrow\quad \gamma_3^*(\phi) = \gamma_1 \frac{\sin(\pi-\alpha'-\phi)}{\sin\alpha'}. \tag{4.24}$$

Comparing with (4.23), the only way this can be true is $\alpha' = \alpha$, which is constant. The set of points with such property forms a perfect circle (inscribed angle inside a circle facing a constant chord is constant). An alternative and simpler proof is that a straight line with unity distance to the origin maps to a circle after r^{-1} transformation.

Define $\gamma^*(\mathbf{n})\mathbf{n}$ as the *stable* Wulff plot. Given $\gamma(\mathbf{n})\mathbf{n}$ (from say, a first-principles total energy calculation), one plots $\gamma^{-1}(\mathbf{n})\mathbf{n}$ and eliminate segments of $\gamma^{-1}(\mathbf{n})\mathbf{n}$ that lies *inside* the common tangent construction. The montage of straight-line common tangent segments plus uneliminated $\gamma^{-1}(\mathbf{n})\mathbf{n}$ segments form $\gamma^{*-1}(\mathbf{n})\mathbf{n}$. We then invert $\gamma^{*-1}(\mathbf{n})$ to get $\gamma^*(\mathbf{n})\mathbf{n}$.

Alternatively, the above can be formulated in Wulff space directly. **Tangent circle theorem**: Given $\gamma(\mathbf{n})\mathbf{n}$, both the necessary and sufficient condition that $\gamma^*(\mathbf{n}') = \gamma(\mathbf{n}')$ for a particular $\mathbf{n}'$ is that if one draws a circle through the origin and tangent to $\gamma(\mathbf{n})\mathbf{n}$ at $\mathbf{n}'$, such tangent circle lies completely *within* $\gamma(\mathbf{n})\mathbf{n}$ and do not hit any other points on $\gamma(\mathbf{n})\mathbf{n}$. This

is because a tangent line of $\gamma^{-1}(\mathbf{n})\mathbf{n}$ that does not hit $\gamma^{-1}(\mathbf{n})\mathbf{n}$ at any other point maps to a tangent circle inside $\gamma(\mathbf{n})\mathbf{n}$.

The tangent circle theorem and decomposition test is useful for thin-film surface on substrate. For free-standing crystallite such as formed in deposition, where surface energy dominates the shape, we need **Wulff construction**: consider a crystal with f possible surface orientations $\mathbf{n}_i$. Denote their distance to the center as h_i. Then the exposed length is a_i. Clearly,

$$a_i \;=\; a_i(h_{i-1}, h_i, h_{i+1}). \tag{4.25}$$

We also have the following reciprocal relation:

$$\frac{\partial a_i}{\partial h_{i-1}} \;=\; \frac{\partial a_{i-1}}{\partial h_i} \;=\; \frac{1}{\sin\theta_{i,i-1}}, \tag{4.26}$$

which can be proven from inspecting the geometry, where $\theta_{i,i-1}$ is the angle between $\mathbf{n}_i$ and $\mathbf{n}_{i-1}$.

Now consider a free-standing particle of fixed volume V. We seek the shape that minimizes its surface energy:

$$F_{\text{surface}} \;=\; \sum_i \gamma_i a_i, \tag{4.27}$$

with the shape completely determined by the $\{h_i\}$. Change in volume must be constrained to zero:

$$0 \;=\; \sum_i a_i dh_i, \tag{4.28}$$

and

$$dF_{\text{surface}} \;=\; \sum_i \left(\gamma_{i-1}\frac{\partial a_{i-1}}{\partial h_i} + \gamma_i \frac{\partial a_i}{\partial h_i} + \gamma_{i+1}\frac{\partial a_{i+1}}{\partial h_i}\right) dh_i, \tag{4.29}$$

so there must be

$$\gamma_{i-1}\frac{\partial a_{i-1}}{\partial h_i} + \gamma_i \frac{\partial a_i}{\partial h_i} + \gamma_{i+1}\frac{\partial a_{i+1}}{\partial h_i} \;=\; \beta a_i, \tag{4.30}$$

where a_i is the Lagrange multiplier. Using the reciprocal relation:

$$\gamma_{i-1}\frac{\partial a_i}{\partial h_{i-1}} + \gamma_i \frac{\partial a_i}{\partial h_i} + \gamma_{i+1}\frac{\partial a_i}{\partial h_{i+1}} \;=\; \beta a_i. \tag{4.31}$$

On the other hand, $a_i(h_{i-1}, h_i, h_{i+1})$ is a homogeneous function of degree 1 (in 2D):

$$a_i(lh_{i-1}, lh_i, lh_{i+1}) \;=\; la_i(h_{i-1}, h_i, h_{i+1}) \tag{4.32}$$

So by taking derivative against l on both sides, and then setting $l = 1$, there is

$$h_{i-1}\frac{\partial a_i}{\partial h_{i-1}} + h_i\frac{\partial a_i}{\partial h_i} + h_{i+1}\frac{\partial a_i}{\partial h_{i+1}} \;=\; a_i. \tag{4.33}$$

In 3D, there is $a_i(lh_{i-1}, lh_i, lh_{i+1}) = l^2 a_i(h_{i-1}, h_i, h_{i+1})$ and $h_{i-1}\frac{\partial a_i}{\partial h_{i-1}} + h_i\frac{\partial a_i}{\partial h_i} + h_{i+1}\frac{\partial a_i}{\partial h_{i+1}} = 2a_i$.

Comparing the two equations, we see that

$$... = \frac{\gamma_{i-1}}{h_{i-1}} = \frac{\gamma_i}{h_i} = \frac{\gamma_{i+1}}{h_{i+1}} = ... = \beta \tag{4.34}$$

for all i, will be a variational extremum. In fact, $dF_{\text{surface}} = dF_{\text{bulk}} = (P_{\text{int}} - P_{\text{ext}})dV$ is the original Young-Laplace pressure argument (Fig. 4.1(a)), and the facet-independent Lagrange multiplier β can be identified to be simply the Young-Laplace pressure difference $\Delta P = P_{\text{int}} - P_{\text{ext}}$. So in 2D, we have $\Delta P = \frac{\gamma_i}{h_i}$.

The above means that the inner envelope formed by all Wulff planes (a Wulff plane lies perpendicular to $\gamma(\mathbf{n})\mathbf{n}$ at $\gamma(\mathbf{n})\mathbf{n}$) gives the equilibrium shape of a free-standing nanocrystal. This is called **Wulff construction**, which minimizes the total surface energy of a free-standing nanoparticle. Note that the Wulff construction serves a different purpose from the tangent circle theorem. The tangent circle theorem deals with the stability of *one* surface constrained to have overall inclination $\mathbf{n}'$ because it must conform to the substrate, whereas the Wulff construction needs to optimize all facets of the nanocrystal simultaneously.

In 3D, there is an extra factor of $\frac{1}{2}$ on RHS, and we get

$$... = \frac{\gamma_{i-1}}{h_{i-1}} = \frac{\gamma_i}{h_i} = \frac{\gamma_{i+1}}{h_{i+1}} = ... = \frac{\beta}{2} = \frac{\Delta P}{2} \tag{4.35}$$

or $\Delta P = \frac{2\gamma_i}{h_i}$ to be the pressure increase inside the solid particle. We see that for isotropic surface energy and spherical particle, this reduces to the familiar expression $\Delta P = \frac{2\gamma}{R}$.

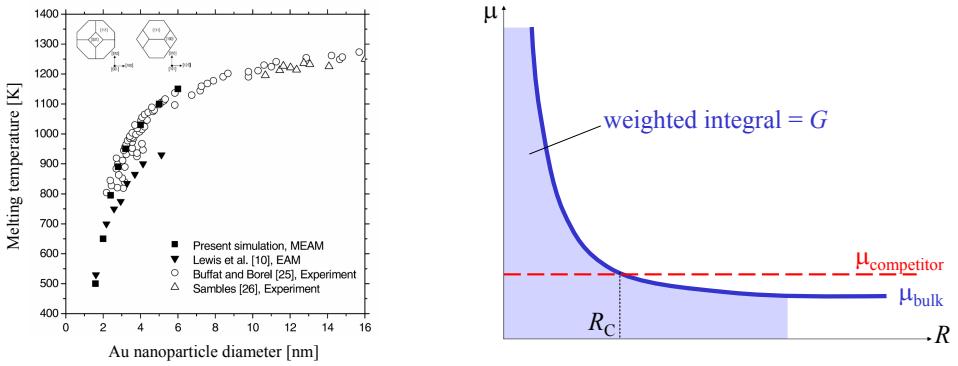
Here comes a paradox. Ignoring the more subtle effect of surface pre-melting, the $\Delta P = \frac{2\gamma_i}{h_i}$ pressure increase should change the melting point of the solid nanoparticle [52, 53, 54, 55],

like a pressure cooker. According to the Clausius-Clapeyron equation:

$$\Delta T_{\rm melt} \;=\; \frac{\Delta s}{\Delta v}\Delta P \;=\; \frac{\Delta s}{\Delta v}\frac{2\gamma_i}{h_i} \;\propto\; \frac{\Delta s}{\Delta v}\frac{2\gamma}{R} \tag{4.36}$$

Thus, the change in melting point should scale as $1/R$, which seems to agrees with Fig. 4.5(a) (Fig. 2 of [54]). Indeed, Fig. 4.5 shows significant *change* (by more than 50%!) of the melting point of pure Au solid nano-particles when the diameter reaches 2nm. However, a more careful inspection indicates the sign of $T_{\rm bulkmelt} - T_{\rm nanomelt}$ is wrong! While naive application of the Clausius-Clapeyron relation indicates the melting point should increase due to positive Young-Laplace pressure, the actual experimental and simulation data indicate the melting point is *suppressed*. What is going on?

Here are some data: $v_{\rm Au}^{\rm fcc} \approx 17.8$ Å^3, $v_{\rm Au}^{\rm liquid} \approx 18.9$ Å^3, and suppose Richard's rule holds (entropy of melting is about $1.1k_{\rm B}$), formulate the problem, define and estimate the physical quantity you could extract from Fig. 4.5. Even though the actual solid particle is faceted as shown in the inset, here in the "spherical cow" approximation take them to be spherical.



(a) (b)

Figure 4.5: Nanoparticle stability. (a) Melting point of Au nanoparticles as a function of diameter, observed experimentally as well as in molecular dynamics simulations [54]. (EAM: Embedded Atom Method; MEAM: Modified Embedded Atom Method, both are interatomic potentials for performing computer simulations). (b) Chemical potential of (energy of attaching) an atom in nanoparticle as a function of nanoparticle radius.

The resolution of the paradox is the following. In deriving the Wulff construction for free-standing solid particle, we mentioned $\gamma_i/h_i = \Delta P/2$, so there is pressure increase inside the solid particle. If one blindly feeds this Young-Laplace pressure increase into the Clausius-Clapeyron equation for bulk material $dP/dT = \Delta s/\Delta v$, one gets melting point *increase*,

since Au melts with positive volume of melting, $\Delta v = v_{\rm Au}^{\rm liquid} - v_{\rm Au}^{\rm fcc} = 1.1$ Å^3, and positive entropy of melting, $\Delta s = s_{\rm Au}^{\rm liquid} - s_{\rm Au}^{\rm fcc} = 1.1 k_{\rm B}$. Using any formula without remembering how it was derived can be dangerous. Recall that P in the Clausius-Clapeyron relation is the *external pressure* $P_{\rm ext}$, to be shared equally between the solid and liquid phases upon $\Delta P_{\rm ext}$. Apparently, the Young-Laplace pressure difference does not work like a common $P_{\rm ext}$.

The right mental setup here is to envision a competing Au *liquid droplet*. There will also be a Young-Laplace pressure difference, but the pressure difference in the liquid droplet will be proportional to $\gamma_{\rm L}$, the liquid surface energy, rather than $\gamma_{\rm S}$, the isotropic solid surface energy in the "spherical cow" approximation. Thus, capillary pressure will be of *different magnitudes* inside the solid particle and liquid particle. If $\gamma_{\rm S}$ is quite a bit larger than $\gamma_{\rm L}$, then the increase of chemical potential inside the solid particle will exceed the increase of chemical potential inside the liquid particle, destabilizing the solid particle relative to the liquid particle and suppressing the melting point.

The discussion is further complicated by the phenomenon of surface pre-melting, where a thin layer of liquid covers the solid particle before the core of the solid particle melts [55]. To take this into account requires more extensive modeling (Model II in [53, 52]). For pedagogical reason let us pretend pre-melting does not happen. The physical effect and parameters extracted from Model II and Model I are not fundamentally different.

Recall that $G^{\rm solid}(N,T,P_{\rm ext},A^{\rm solid}) = G^{\rm solid}_{\rm bulk}(N,T,P_{\rm ext}) + 4\pi R_{\rm S}^2\gamma_{\rm S}$, $G^{\rm liquid}(N,T,P_{\rm ext},A^{\rm liquid}) = G^{\rm liquid}_{\rm bulk}(N,T,P_{\rm ext}) + 4\pi R_{\rm L}^2\gamma_{\rm L}$, where $R_{\rm S}$ and $R_{\rm L}$ are the radius of the solid and liquid nanoparticles, respectively. Setting them equal, we have

$$4\pi(R_{\rm S}^2\gamma_{\rm S} - R_{\rm L}^2\gamma_{\rm L}) \;=\; G^{\rm liquid}_{\rm bulk}(N,T,P_{\rm ext}) - G^{\rm solid}_{\rm bulk}(N,T,P_{\rm ext}). \tag{4.37}$$

Also recall that at bulk melting point $T_{\rm bulkmelt}$, we have

$$G^{\rm liquid}_{\rm bulk}(N,T_{\rm bulkmelt},P_{\rm ext}) - G^{\rm solid}_{\rm bulk}(N,T_{\rm bulkmelt},P_{\rm ext}) \;=\; 0, \tag{4.38}$$

so we can perform Taylor expansion of the right-hand side around $T = T_{\rm bulkmelt}$ and keep only the leading-order term in the spirit of spherical-cow approximation, and get:

$$4\pi(R_{\rm S}^2\gamma_{\rm S} - R_{\rm L}^2\gamma_{\rm L}) \;=\; N\Delta s(T_{\rm bulkmelt} - T). \tag{4.39}$$

We also have the following relations: $Nv_{\rm Au}^{\rm fcc} = 4\pi R_{\rm S}^3/3$, $Nv_{\rm Au}^{\rm liquid} = 4\pi R_{\rm L}^3/3$, $N = 4\pi R_{\rm S}^3/3v_{\rm Au}^{\rm fcc} = 4\pi R_{\rm L}^3/3v_{\rm Au}^{\rm liquid}$, and so $R_{\rm L}^2 = R_{\rm S}^2(v_{\rm Au}^{\rm liquid}/v_{\rm Au}^{\rm fcc})^{2/3} = 1.04R_{\rm S}^2$ (the liquid particle of equal mass is

slightly larger in size). Thus we have:

$$T_{\text{bulkmelt}} - T \;=\; \frac{4\pi R_{\text{S}}^2(\gamma_{\text{S}} - 1.04\gamma_{\text{L}})}{\Delta s \cdot 4\pi R_{\text{S}}^3/3v_{\text{Au}}^{\text{fcc}}} \;=\; \frac{3v_{\text{Au}}^{\text{fcc}}(\gamma_{\text{S}} - 1.04\gamma_{\text{L}})}{\Delta s \cdot R_{\text{S}}}. \tag{4.40}$$

We get a melting point suppression that goes like R_{S}^{-1}, which is good news because that is how the figure looks like. If we take $T_{\text{bulkmelt}} - T$ to be 700 K at $R_{\text{S}} = 1$nm (diameter 2nm), then we could estimate $\gamma_{\text{S}} - 1.04\gamma_{\text{L}}$ to be 0.2 J/m^2. This is quite reasonable number. In Table 3.1 of [47], γ_{S} is listed to be 1.39 J/m^2. So from (4.40) we may deduce that $\gamma_{\text{L}} = 1.14$ J/m^2.

In Table 3.4 of [47], γ_{SL} is listed to be 0.132 J/m^2. We know that Au liquid, as most metallic liquids, wets its own solid: $\gamma_{\text{S}} - \gamma_{\text{L}} - \gamma_{\text{SL}} > 0$, and that certainly does not conflict with our result. In reality, bulk measurement gives $\gamma_{\text{L}} = 1.135$ J/m^2 [53], which is embarrassingly close to the prediction of (4.40).

Instead of comparing the total free energy, there is an alternative derivation based on comparing the chemical potentials. Since

$$\mu^{\text{solid}} \;=\; \mu_{\text{bulk}}^{\text{solid}} + \frac{2\gamma_{\text{S}} v_{\text{Au}}^{\text{fcc}}}{R_S}, \quad \mu^{\text{liquid}} \;=\; \mu_{\text{bulk}}^{\text{liquid}} + \frac{2\gamma_{\text{L}} v_{\text{Au}}^{\text{liquid}}}{R_L}, \tag{4.41}$$

if we equate μ^{solid} with μ^{liquid}, and use the expression $\mu_{\text{bulk}}^{\text{liquid}} - \mu_{\text{bulk}}^{\text{solid}} = \Delta s(T_{\text{bulkmelt}} - T)$, we will get:

$$T_{\text{bulkmelt}} - T \;=\; \frac{1}{\Delta s}\left(\frac{2\gamma_{\text{S}} v_{\text{Au}}^{\text{fcc}}}{R_S} - \frac{2\gamma_{\text{L}} v_{\text{Au}}^{\text{liquid}}}{R_L}\right) \;=\; \frac{2v_{\text{Au}}^{\text{fcc}}(\gamma_{\text{S}} - 1.04\gamma_{\text{L}})}{\Delta s \cdot R_{\text{S}}}, \tag{4.42}$$

since $v_{\text{Au}}^{\text{liquid}}/v_{\text{Au}}^{\text{fcc}} = R_L^3/R_S^3$. The rationale for equating the chemical potentials is that, *if* there is an existing liquid particle of *equal mass* to the solid particle, the liquid particle would grow and the solid particle would shrink, via vapor phase transport. This assumption was clearly stated on p.2289 of [53], even though the scenario is not very realistic (where does this liquid particle of just the right size come from...) On the other hand, equating the total energy, (4.40), assumes one solid particle is *totally transformed* into a liquid particle, and an accounting of the total thermodynamic profit is done assuming the kinetic barrier can be overcome.

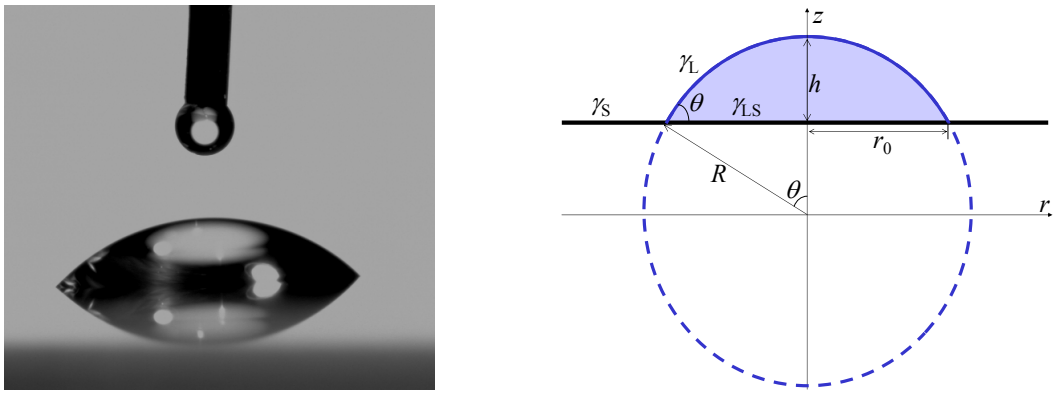
For the sake of a rough estimate in the present problem, (4.40) and (4.42) are equally fine, differing just by a numeric factor of 3/2. It is worth noting, however, that for a sphere

the relative capillary contribution to the *integral energy*, the total energy, always has this factor of 3/2 over the relative capillary contribution to the *differential energy*, the chemical potential. For a cylinder (nanowire), this ratio becomes 2/1, so this distinction between total and differential energy contribution is no longer small and is worth some discussion. First, we note that for a finite-sized object ($G = G_{\rm bulk} + \gamma A$), the physical meaning of μ as always is the change in G if we add one more atom/molecule to the object: $\mu \equiv \frac{\partial G}{\partial N} = \mu_{\rm bulk} + \gamma \frac{dA}{dV}\frac{\partial V}{\partial N}$ ($= \mu_{\rm bulk} + 2\gamma v/R$ in the case of a sphere, and $\mu_{\rm bulk} + \gamma v/R$ in the case of a cylinder), which agrees with the result of using $\partial\mu/\partial P = v$ and the Young-Laplace pressure, as they must. We see that the chemical potential is no longer a size-independent quantity, but depends on R, and in fact *always* diverges to $+\infty$ as $R \to 0$ even if $\mu_{\rm bulk}$ is very favorable, as illustrated in Fig. 4.5(b). A reverse statement is that $G(R)$ is just the integral of $\mu(R)$, as we build up the object from $R' = 0$ to the present size $R' = R$. Initially, this integral may look hazardous because of the $1/R'$ singularity. Fortunately, it is a weighted integral, the weighting factor is $4\pi R'^2 dR'$ in the case of sphere and $2\pi R' dR'$ in the case of cylinder. For sphere, we get $\int_0^R 4\pi R'^2 dR' (2\gamma v/R') / \int_0^R 4\pi R'^2 dR' = 3\gamma v/R$, that is to say the averaged price from 0 to R is 3/2 higher than the going price at R. This can be directly seen already in Fig. 4.5(b), since the previous stuff bought at lower R' is more expensive than the stuff bought at the present R. Similarly, for a nanowire, $\int_0^R 2\pi R' dR' (\gamma v/R') / \int_0^R 2\pi R' dR' = 2\gamma v/R$, so the averaged price from 0 to R is twice as expensive as the going price. The philosophical point in judging which is more reasonable, (4.40) or (4.42), in governing the melting point, is that you cannot buy cheap stuff without buying the expensive stuff first. (where does this liquid particle of just the right size in (4.42) come from...)

Despite of the error in neglecting surface pre-melting [55], our "spherical cow" approach is still a resounding success because it gives the correct behavior of melting point suppression versus particle size ($\propto 1/R$), as well as the correct order of magnitude for $\gamma_{\rm S} - 1.04\gamma_{\rm L}$, which is of the same order of magnitude as $\gamma_{\rm SL}$.

For many metals, $\gamma_{\rm S}$ is close to but greater than $\gamma_{\rm L} + \gamma_{\rm LS}$.

With the above introduction of $\gamma_{\rm LS}$, now we generalize surface energy to interfacial energy (rigorously speaking, even surface is solid-vapor or liquid-vapor interface). Consider liquid-solid energy $\gamma_{\rm LS}$. For the moment assume diffusion is slow in solid and the solid surface is flat. The liquid has surface energy (liquid-vapor interface) $\gamma_{\rm L}$, and the solid has surface energy (solid-vapor interface) $\gamma_{\rm S}$. The angle formed between $\gamma_{\rm L}$ and $\gamma_{\rm LS}$ is called the contact



(a) (b)

Figure 4.6: (a) Water drop on glass. Taken from Wikipedia. (b) Cylindrical coordinate frame.

angle, θ. From energy or force balance, we should have:

$$\gamma_S = \gamma_{LS} + \gamma_L \cos\theta. \tag{4.43}$$

The above is called Young's equation. θ expresses a relationship between γ_L and $\gamma_S - \gamma_{LS}$. From one contact angle measurement, one could get the difference $\gamma_S - \gamma_{LS}$, but not the absolute values.

Two remarks on Young's equation: (a) it is derived as a force balance on a *point*, the contact point. As such, body force such as gravity which is proportional to volume does not enter into it directly. Thus, the same Young's equation works if the solid surface is inclined or even vertical, with the contact angle (defined as the angle of liquid meniscus to solid surface at contact point) remaining the same - θ is a material constant. θ itself will not change, although the shape of the overall droplet could change, due to gravity. (b) We only considered balance of force components parallel to the solid surface. The component perpendicular to surface, $\gamma_L \sin\theta$, is nonzero and not at equilibrium, because we assume solid diffusion is slow. If significant time is allowed for solid-state diffusion, a cusp will in fact develop on the solid surface. This balance of vertical capillary force is demonstrated in so-called *grain boundary grooving* (Fig. 14.9 of [41]), when a GB meets perpendicular to a surface, and surface diffusion has occurred for long enough to enable equilibrium of this vertical capillary force. Let the grooving angle be ϕ, there is

$$2\gamma_S \cos(\phi/2) = \gamma_{GB} \tag{4.44}$$

when equilibrium is finally reached. In fact the above equation is how grain boundary energy γ_{GB} is measured. (γ_{S} needs to be measured in previous experiments).

If we ignore the effect of gravity, the rest shape must be a truncated sphere. This is because of the Young-Laplace relation $\Delta P = 2\gamma/R$, where R is the local radius of curvature. Since the fluid is quiescent (not flowing), if we ignore the hydrostatic pressure difference, the pressure must be the same everywhere inside the fluid, which means the local curvature R^{-1} is the same everywhere. Mathematically this can be seen as $P_{\mathrm{L}}(h) - P_{\mathrm{ext}} = \frac{\gamma}{R_1(h)} + \frac{\gamma}{R_2(h)}$, where $P_{\mathrm{L}}(h) = P_{\mathrm{L}}(0) - \rho g h$ is hydrostatic pressure inside the quiescent liquid drop, and the atmospheric pressure is assumed to be independent of h.

If $\gamma_{\mathrm{S}} - \gamma_{\mathrm{LS}} = -\gamma_{\mathrm{L}}$, $\theta = 180°$. This means the liquid would not wet the solid, and would rather be a stand-alone spherical droplet detached from the solid surface. You want your raincoat surface to have this property.

If $\gamma_{\mathrm{S}} - \gamma_{\mathrm{LS}} = \gamma_{\mathrm{L}}$, $\theta = 0°$. If $\gamma_{\mathrm{S}} - \gamma_{\mathrm{LS}} \geq \gamma_{\mathrm{L}}$, then on a horizontal surface, the liquid would spread out and totally cover the horizontal surface with equal thickness because this is energetically favorable. This is called complete wetting. It turns out that for most metals: $\gamma_{\mathrm{S}} - \gamma_{\mathrm{LS}} > \gamma_{\mathrm{L}}$, which means the solid metal likes to be covered completely by its own melt, instead of as liquid domes or droplets on top. Later in discussing nucleation, we will see that this means the nucleation barrier of solid→liquid phase transformation will be zero upon heating. Liquid metals could sustain significant under-cooling before transforming to solid due to nucleation barrier posed by $\gamma_{\mathrm{LS}} > 0$, but solid metals seldom manifest significant superheating, because the interfacial energy difference actually *helps* the nucleation of liquids on solid surface.

What if the surface is vertical, and $\gamma_{\mathrm{S}} - \gamma_{\mathrm{LS}} > \gamma_{\mathrm{L}}$? The meniscus (Greek for crescent) is the curved liquid surface in response to presence of the container, such as a vertical test tube. In the case of complete wetting ($\gamma_{\mathrm{S}} - \gamma_{\mathrm{LS}} > \gamma_{\mathrm{L}}$), there are two considerations. First, uniform spread of liquid on the vertical surface is against gravity, unlike on a horizontal surface. Secondly, one gains the same amount of interfacial energy reduction whether it is 1mm of liquid covering the solid or 1μm liquid covering the solid surface. Because of gravity, you want the surface covered with as *thin* layer of liquid as possible. Thus, what will actually happen is that a layer of liquid *molecular dimension thin* will creep up and cover the vertical solid surface in its entirety, and modify the effective "solid" surface energy from γ_{S} to γ_{S}^*, such that $\gamma_{\mathrm{S}}^* = \gamma_{\mathrm{LS}} + \gamma_{\mathrm{L}}$ and the contact angle is exactly $\theta = 0°$. In the case of capillary tube under complete wetting condition, this will cause the meniscus to have a negative curvature of $R = -D/2$, which induces a negative pressure inside the liquid, which will pump up a

liquid jet. From the height of the liquid jet, one can infer the liquid surface energy $\gamma_{\rm L}$.

A rough estimation of the ability of a liquid *monolayer* (only a single molecule thick) to climb up a vertical wall. Suppose $\gamma_{\rm S}-\gamma_{\rm LS}-\gamma_{\rm L}=0.1$ J/m^2: this is the thermodynamic driving force of a completely dry surface to be wetted, and 0.1 J/m^2 is reasonable magnitude for such. Previously in homework we have computed the molecular volume of one H_2O molecule to be 30 Å^3. So the area covered by one molecule is around 10 Å^2. Then the adsorption energy of one H_2O molecule on such dry hydrophilic surface should be about 10^{-20} J or 0.0624 eV. Sounds small (since a primary bond enegy is already $\epsilon\sim$ 1eV), but how high can it climb on the dry container wall? Since $m=$ 18amu $\approx 2.9890\times10^{-26}$ kg, it will climb up until $mgh=10^{-20}$J, or $h=3.3456\times10^4$ m, the height of approximately 4 Mount Everest. This in fact is not surprising, if we recognize that air pressure on Everest is about a third of sea level pressure, and air molecule has kinetic energy $k_{\rm B}T_{\rm room}/2$ in the z direction, or 0.0125 eV. $k_{\rm B}T_{\rm room}$ is really already a lot for gravity, and 10^{-20} J $\approx 5k_{\rm B}T_{\rm room}$.

Having said that, the *2nd* monolayer will have a lot less adsorption energy than the 1st monolayer because of longer distance to the wall, and will only climb up to say, half Mount Everest. The 3rd monolayer may only climb up 500 m, and so on. This "solid surface" with a few adsorbed liquid monolayers will have an effective "surface" energy $\gamma_{\rm S}^*=\gamma_{\rm LS}+\gamma_{\rm L}$ as far as the remainder of the bulk liquid is concerned.

When a cylindrical hollow tube of diameter D is inserted into liquid, the contact-angle equation gives

$$R\cos\theta \;=\; \frac{D}{2} \tag{4.45}$$

where we assumed the meniscus is part of a perfect sphere of radius R. The Young-Laplace pressure in the fluid is then

$$\Delta P \;=\; P(h)-P_{\rm ext} \;=\; \rho gh \;=\; \frac{2\gamma_{\rm L}}{R} \;=\; \frac{4\gamma_{\rm L}\cos\theta}{D} \tag{4.46}$$

Depending on whether $0\le\theta<90^\circ$ or $90^\circ<\theta\le180^\circ$, a negative / positive Young-Laplace pressure is generated, which either pumps up or pushes down a liquid jet of height h. Note that $\gamma_{\rm S}$, $\gamma_{\rm LS}$ also come into (4.46) via

$$\cos\theta \;=\; \frac{\gamma_{\rm S}-\gamma_{\rm LS}}{\gamma_{\rm L}} \tag{4.47}$$

Even though it appears that (4.47) may not always have legitimate real θ solution for certain

($\gamma_{\rm S}$, $\gamma_{\rm LS}$, $\gamma_{\rm L}$) combination, in reality when this happens from the discussion above on "invisible monolayer" either $\gamma_{\rm S}$ is renormalized to $\gamma_{\rm S}^*$, or $\gamma_{\rm LS}$ is renormalized to $\gamma_{\rm LS}^*$, such that $\theta = 0$ or $\theta = 180^\circ$ *becomes* legitimate solutions.

The molecular-scale monolayer comes up quite often in science. One is in the context of Ben Franklin's experiment at Clapham pond [56], and Langmuir-Blodgett film. The other is so-called surface pre-melting phenomena, defined as the loss of long-range order in the top few layers of atoms on a crystal, at $T < T_{\rm bulkmelt}$: the most of the crystal still maintains long-range order, but if one does electron diffraction only on the top few layers of atoms and look at the structure factor, one sees a liquid-like arrangement.

Suppose β precipitate is in contact with phase α. Suppose in the case of a planar interface ($K = 0$), the two have reached equilibrium:

$$\mu_i^\alpha(\mathbf{X}_0^\alpha, T, P^\alpha) \;=\; \mu_i^\beta(\mathbf{X}_0^\beta, T, P^\beta = P^\alpha), \tag{4.48}$$

where P^α is the "external" pressure. If α is vapor phase, there is then equilibrium vapor pressure $P_i^{\rm eq}(\infty)$ of species i in α, in contact with β. Now imagine a curved interface across which two phases try to reach equilibrium, and their interfacial energy γ is isotropic. Everything is the same in β, but now with a curved interface, the Young-Laplace pressure causes the chemical potential of i inside the precipitate to increase, which must be matched by an equal raise of chemical potential of i outside:

$$\frac{2v_i^\beta\gamma}{R} \;=\; k_{\rm B}T\ln\frac{P_i^{\rm eq}(R)}{P_i^{\rm eq}(\infty)} \tag{4.49}$$

Similarly, if α is liquid or solid, there is equilibrium solubility $X_i^\alpha(\infty)$ of species i in α, and if the solution can be approximated as ideal for i, the solubility of i in α will be enhanced by

$$\frac{2v_i^\beta\gamma}{R} \;=\; k_{\rm B}T\ln\frac{X_i^\alpha(R)}{X_i^\alpha(\infty)}. \tag{4.50}$$

The above can in turn drive diffusion in α, which makes larger β particles bigger, and smaller β particles smaller.

In Wulff plot we focused on the inclination degrees of freedom, ϕ, since the vapor phase and liquid phase (fluid) has random atomic arrangement with no intrinsic orientation (zero dof). Crystal surface has two dofs, the inclination ϕ. Crystal-crystal interfaces are generally more complicated, possessing 5 dofs. A crystal-crystal interface can be a grain boundary,

if the two crystals are of the same structure, just rotated; or a *phase boundary*, if different lattice structures and/or significantly different compositions. In addition to $\mathbf{n}$ (called two inclination degrees of freedom), there are also three misorientation degrees of freedom in how one crystal is rotated ($\mathbf{R}^T\mathbf{R} = \mathbf{I}$) with respect to the other. Let us represent this misorientation by an abstract and generic angle θ. In the case of GB, if $\theta = 0$, $\gamma = 0$ for all inclinations. But this is not so for phase boundary: imagine Kossel on Kossel *epitaxy* with $\theta = 0$ but different lattice constants. There will be finite γ.

Epitaxy is a particular kind of thin-film deposition where the deposited material takes on the same structure and orientation as the substrate ($\theta = 0$). Thin-film deposition is less constrained than some bulk solid-solid phase transformation such as alloy decomposition / precipitation, because the added material comes from a fluid phase (vapor, liquid). Thus in depositing a thin film, vertically there is a stress-free surface, so if the added material wants to dilate vertically, it can do so without incurring elastic energy penalty. Laterally, there is less constraint as well (a semi-coherent interface would be able to relax all long-ranged elastic energy to zero).

Let us talk about GBs first. There are low-angle ($\theta < 10 - 15°$) GBs, special high-angle GBs such as twin boundaries, and high-angle "random" GBs (see Fig. 3 of [57]). Near $\theta = 0$ as well as the special high-angle GBs, the grain boundary energy varies with $\Delta\theta$ as:

$$\Delta\gamma \;=\; A|\Delta\theta|(B - \ln|\Delta\theta|) \tag{4.51}$$

which represents the cusps (vicinal boundaries are those that are few degrees off from special high-angle GBs). This Read-Shockley formula is explained by so-called dislocation representation of crystal-crystal interfaces. Because dislocations have $1/r$ like stress field, the strain energy density is $\propto 1/r^2$, and so the energy stored near one such dislocation is $\int_{R_0}^{l} 2\pi r dr/r^2 \propto \ln(l/R_0)$, where R_0 is some cutoff distance. The dislocation density on the interface (unit 1/m) can be shown to be $\rho_{\text{int}} \equiv 1/l \propto |\Delta\theta| > 0$, thus the energy goes like $-|\Delta\theta|\ln(|\Delta\theta|R_0)$. Similar kind of argument can be made for ϕ-dependence: it is "cuspy", because crystallographically the vicinal boundaries must exist as long stretches of coherent GBs, plus misfit steps.

Coherency means atoms on two sides of the interface match well geometrically, at the atomic scale, an admittedly somewhat fuzzy concept. There are coherent interfaces, semi-coherent interfaces, and incoherent interfaces. The above classification works for both phase boundaries and grain boundaries. In GBs, the special high-angle GBs are coherent (all atoms

along the interface are "good" material), the low-angle GBs and vicinal boundaries are semi-coherent (long stretches of "good" material $l - 2R_0$, interspersed by "bad" material $2R_0$), and the random high-angle GBs are incoherent ($l \sim 2R_0$, the dislocation cores overlap and basically all materials along the interface are "bad", with grotesquely misaligned bonds). Example of coherent and incoherent twin boundaries shown in Fig. 3.12 of [47].

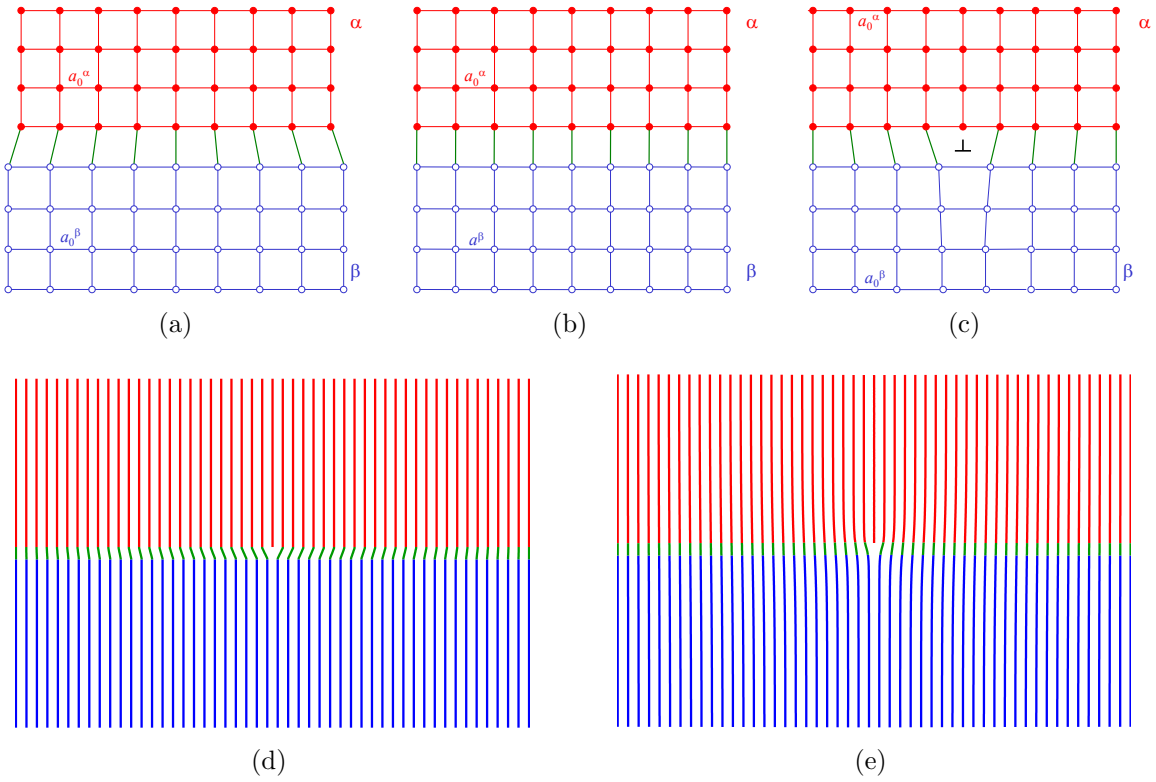


Figure 4.7: (a) 9 planes matching 9 planes, with zero elastic energy but huge average glue energy ($\propto \gamma_{\text{incoherent}} A$), (b) 9 planes matching 9 planes, with small glue energy ($\propto \gamma_{\text{coherent}} A$ where $\gamma_{\text{coherent}} \propto \epsilon_\alpha - \epsilon_{\alpha\beta}$, but finite elastic energy ($\propto \delta^2 V^\beta$), (c) 9 planes matching 8 planes, with smaller elastic energy than (b) and smaller glue energy than (a). It turns out that by *choosing appropriate matching*, the total energy of (c) can be $\propto \gamma_{\text{semicoherent}} A$. (d) 51 planes matching 50 planes, equivalent to (a), where dis-registry function $d(x)$ is piecewise linear. (e) 51 planes matching 50 planes, equivalent to (c), where $d(x)$ is sigmoidal.

Consider a phase boundary between α and β, with same crystal structure and orientation ($\theta = 0$), but different equilibrium lattice constants (a_0^α and a_0^β). How might a β particle be embedded in α, or epitaxially grow on α? Define misfit strain $\delta = a_0^\beta / a_0^\alpha - 1$. Let us first suppose the misfit strain is small, say 1%. There are a few limiting possibilities, illustrated in Fig. 4.7. Consider a model of Kossel crystal on Kossel crystal, and let us suppose that

the energy per bond $\epsilon_\alpha = \epsilon_\beta > \epsilon_{\alpha\beta}$, so α and β *wants* to phase separate. Let us define elastic energy to be energy stored in the blue and red springs, and "glue" energy to be energy stored in the interfacial green springs. In (a), the elastic energy is zero, but the glue energy is only zero at the center and gets worse and worse further out. If one defines dis-registry function $d(x)$ to be the offset between red and blue atoms at the interface $(x_{\text{red}} - x_{\text{nearestblue}})/a_0^\alpha$, then $d(x)$ is linear / piecewise linear function in x, which incurs huge interfacial energy penalty on average. In fact, (a)'s interface will not be very different from general incoherent interfaces in terms of energy, since around half of the green bonds are grotesquely dis-registered, defined by $d(x) > 1/4$. In such situation expect (a)'s interfacial energy to be $\gamma_{\text{incoherent}} \sim 1$ J/m^2, or ~ 0.1 eV per interfacial atom.

In (b), there will be finite elastic energy ($\propto \delta^2 V^\beta$) to compress β, but the interfacial energy becomes much smaller than $\gamma_{\text{incoherent}}$ ($\propto \gamma_{\text{coherent}} A \propto (\epsilon_\alpha - \epsilon_{\alpha\beta})A$). Typical γ_{coherent} is about 0.1 J/m^2 (ranges from $1 - 200$ mJ/m^2). For fully coherent precipitates ($\theta = 0$) like Guinier-Preston zones in Al-4% Ag (Fig. 3.39 of [47]), and tertiary γ' particles in Ni-based superalloys, since V^β's are small (tens of nanometers), the elastic energy is not overly expensive, and the comfort of γ_{coherent} is well worth the effort of compressing or dilating the precipitate volumetrically. The shape of these fully coherent particles, if not spherical, tends to be "blocky", with aspect ratio not *too* far from 1, determined mainly by the Wulff plot.

As the fully coherent precipitate particle gets larger, however, the $\propto \delta^2 V^\beta$ energy becomes more and more expensive, and at one point there will be *coherency loss*, when the body is no longer willing to keep up the elastic strain for the sake of the glue. Misfit dislocations will be injected into the interface, which are either nucleated afresh or gathered from the surrounding of the originally fully coherent particle. By tuning the density of misfit dislocations ρ_{misfit} appropriately, one can eliminate *long-range* elastic pain that permeates through the precipitate volume, although *short-range* elastic pain on the lengthscale of $l = 1/\rho_{\text{misfit}} \propto 1/\delta$, will still persist. The dis-registry function $d(x)$, instead of a linear function in $[-l/2, l/2]$ as in (a), will be a sigmoidal function, whose width $2R_0$ (R_0 is the dislocation core radius) is fixed and do not change with l. Alternatively, if one uses (a) as the reference state (stress-free in the body, huge pain in the glue), the (a) $\rightarrow$ (c) transformation can be represented by injecting an array of misfit dislocations of distance l plus compensating infinitesimal "coherency dislocations" of the opposite sign (Fig. B.8(b) of [41]).

Fig. 4.7(d) is the same situation as Fig. 4.7(a), and Fig. 4.7(e) is the same as Fig. 4.7(c), except we now have 51 α planes matching 50 β planes, and we use real misfit dislocation solution instead of cartoon. Notice that most of the green bonds in Fig. 4.7(e) are "happy"

- maybe 40 out of the 50 green bonds have $d(x) < 0.1$, and only a few green bonds in Fig. 4.7(e) are grotesquely dis-registered and in extreme pain. It should be clear that the glue pain in Fig. 4.7(e) is much smaller than in Fig. 4.7(d). To achieve the relaxation from Fig. 4.7(d)→(e), local elastic energy is necessary, but not long-range compression of V^β. The lumped sum of glue energy near the core (smaller than (a)) and elastic energy (smaller than (b)) will be defined as $\gamma_{\text{semicoherent}} A$, where

$$\gamma_{\text{semicoherent}} \;=\; \gamma_{\text{coherent}} + U|\delta|(W - \ln|\delta|), \tag{4.52}$$

same as in the Read-Shockley formula for GBs [58]. From this we can thus infer that coherency loss should occur at

$$\gamma_{\text{semicoherent}} A \;=\; \gamma_{\text{coherent}} A + \delta^2 V^\beta \;\;\rightarrow\;\; U|\delta|(W - \ln|\delta|)A \;=\; \delta^2 V^\beta \tag{4.53}$$

$$h_c \;\equiv\; \frac{V^\beta}{A} \;\propto\; \frac{W - \ln|\delta|}{|\delta|}. \tag{4.54}$$

The formula above also governs coherency loss in hetero epitaxy. For instance Ge (free-standing Germanium lattice constant $a_0^{\text{Ge}} = 5.64613$Å) on Si (free-standing Silicon lattice constant $a_0^{\text{Si}} = 5.43095$Å), whose $\delta = 0.04$: we expect the first tens of layers of Ge atoms on Si substrate will be fully coherent, but when that critical deposition thickness h_c is reached, there will be a transition from (b) → (c).

We see that $\gamma_{\text{semicoherent}} - \gamma_{\text{coherent}}$ saturates at large δ, when the misfit dislocation cores start to overlap and one can no longer make out individual misfit dislocations, but a jumbled mess of grotesquely dis-registered green bonds ($d(x) > 1/4$). Typically, when $\delta > 0.2$, there is no longer much difference between semi-coherent and incoherent interfaces. $\gamma_{\text{semicoherent}}$ is around 200-500 mJ/m^2.

In above we have considered β having the same structure as α, $\theta = 0$, and there is only lattice constant mismatch. In such case, forming fully coherent precipitate (on all sides) is possible at small sizes. But if β has different crystal structure from α, forming fully coherent precipitate is generally impossible at any sizes. It is however still possible to form some coherent or semi-coherent interfaces between α and β, if the orientation of β is carefully chosen, in so-called special *orientation relationship*. The Wulff plot for such a given orientation relationship then may look like Fig. 3.40 of [47], where a particular interfacial inclination is greatly favored over all others, which then gives a plate-like morphology (Fig. 3.42 of [47]), with broad faces being this special coherent/semicoherent interface, bounded

by narrow strips of high-energy incoherent interfaces.

Generally speaking, an *orientation relationship* is denoted like

$$(001)_{\text{tetragonal}} \parallel (001)_{\text{FCC}}, \quad [100]_{\text{tetragonal}} \parallel [100]_{\text{FCC}}, \tag{4.55}$$

if the precipitate is a tetragonal ordered compound, with $[001]_{\text{tetragonal}}$ being its c-axis (Fig. 3.41 of [47]). Since it is embedded in a cubic matrix, there are two other different but equivalent orientation *variants*:

$$(001)_{\text{tetragonal}} \parallel (010)_{\text{FCC}}, \quad [100]_{\text{tetragonal}} \parallel [100]_{\text{FCC}}, \tag{4.56}$$

$$(001)_{\text{tetragonal}} \parallel (100)_{\text{FCC}}, \quad [100]_{\text{tetragonal}} \parallel [010]_{\text{FCC}}. \tag{4.57}$$

There are three degrees of freedom in orientation relationship, 2 in plane orientation matching, and 1 in axis matching (in (4.55), $[100]_{\text{tetragonal}}$ must belong to $(001)_{\text{tetragonal}}$ plane, and $[100]_{\text{FCC}}$ must belong to $(001)_{\text{FCC}}$ plane, and one performs in-plane rotation to match them). Orientation relationship can be determined by selected area electron diffraction and contains the same amount of information. (4.56),(4.57) would give different set of diffraction peaks from (4.55), and therefore are distinct orientation variants. On the hand, changing $[100]_{\text{FCC}}$ in (4.55) to $[010]_{\text{FCC}}$ or $[\bar{1}00]_{\text{FCC}}$ or $[0\bar{1}0]_{\text{FCC}}$ (4 ways to perform 90° rotation in-plane after plane matching has been achieved) would not change the diffraction peaks, and therefore do not represent distinct orientation variants. Orientation relationship like (4.55) does not tell you the exact *inclination* of the interface (although strongly hinting it is near the two planes) - to determine that one must go to the imaging mode.

When HCP precipitate comes out of FCC matrix, it is possible to have a habit plane of coherent or semi-coherent interfaces, plus short incoherent interfaces on the side as dictated by the Wulff plot (Fig. 3.40 of [47]), forming plate-like (Widmanstätten) precipitate morphology. In this case, we would be matching the basal plane of hcp with close-packed plane of FCC, as in

$$(0001)_{\text{HCP}} \parallel (111)_{\text{FCC}}, \quad [2\bar{1}\bar{1}0]_{\text{HCP}} \parallel [1\bar{1}0]_{\text{FCC}} \tag{4.58}$$

As another example, when BCC interfaces with FCC, say in a Nb-Cu multilayer, there are two well-known possible orientation relationships:

$$\text{Kurdjumov} - \text{Sachs}: \quad (110)_{\text{BCC}} \parallel (111)_{\text{FCC}}, \quad [1\bar{1}1]_{\text{BCC}} \parallel [0\bar{1}1]_{\text{FCC}}, \tag{4.59}$$

$$\text{Nishiyama} - \text{Wasserman}: \quad (110)_{\text{BCC}} \parallel (111)_{\text{FCC}}, \quad [001]_{\text{BCC}} \parallel [\bar{1}01]_{\text{FCC}}, \tag{4.60}$$

Chapter 5

Elastic Energy Effects

Bulk solution thermodynamics is all about $\mathbf{X}, T, P$. There is nothing in bulk solution thermo that defines *shape/orientation* (or size, for that matter, as long as it is "large enough"). Three factors control shapes and sizes of *microstructures / nanostructures*: (a) capillary energy, (b) elastic energy, (c) kinetics. In this chapter we briefly discuss about (b). Because both (a) and (b) are energies and they always add together in driving kinetics, we can already assert now that for "larger" objects, minimizing elastic energy is more important than minimizing capillary energy, whereas for "smaller" objects, minimizing capillary energy is more important than minimizing elastic energy. Since capillary energy $\propto A$ is usually positive (interface as energetic "overhead"), capillary energy will tend to drive things to "coarsen", that is, small things *merge* together to form larger things to reduce the interfacial energy overhead. This can happen by for instance solute diffusion in the matrix. On the other hand, we will see that (b) could sometimes drive microstructures to "refine" or "split", as well as to "organize" into **patterns**. The basic reason is that $\Delta G_{\text{elastic}} \propto V^\beta \delta^2$: by refining the microstructure and mix-and-matching different strain variants, one could reduce the effective δ averaged over a transformed volume.

For a fully coherent precipitate, the total elastic energy goes as $\bar{C}\delta^2 V^\beta$. $\bar{C}$ depends on the elastic constants of both α and β. If the elastic constants are the same between the matrix and the precipitate, and furthermore are isotropic, and if the transformation strain is isotropic as well, then it can be shown that $\bar{C}$ is *independent* of the shape of the precipitate. But if the elastic constants are different/anisotropic, or if the transformation strain is anisotropic ($\delta_{xx} \neq \delta_{yy} \neq \delta_{zz}$), $\bar{C}$ will also depend on the shape of the precipitate. It turns out that the disk shape confers some advantage over the spherical shape. See Fig. 3.48

of [47]. Imagine there is a disk-shape hole in a matrix with normal in z, convince yourself that δ_{zz}, δ_{xz}, δ_{yz} transformation strains in the hole are easier to accommodate (softer or more compliant, take less elastic energy) than δ_{xx}, δ_{yy}, δ_{xy} transformation strains, because δ_{zz}, δ_{xz}, δ_{yz} take advantage of "lever-like" action of the hole, while δ_{xx}, δ_{yy}, δ_{xy} cannot (jacking up the hole is geometrically softer than stretching the hole). One naively would expect $\Delta G_{\text{elastic}} \propto V^\beta(E_x\delta_{xx}^2 + E_y\delta_{yy}^2 + E_z\delta_{zz}^2)$ by symmetry, if $\delta_{xz} = \delta_{yz} = \delta_{xy} = 0$, and E_x,E_y,E_z are like the Young's modulus in three directions. It turns out that by taking the disk shape, the $E_z\delta_{zz}^2$ term can be significantly reduced (geometrical softening), and then the elastic energy becomes dominated by the $V^\beta(E_x\delta_{xx}^2 + E_y\delta_{yy}^2)$ part. $V^\beta(E_x\delta_{xx}^2 + E_y\delta_{yy}^2)$ can then be further reduced by choosing the orientation of the precipitate ("orientational softening") or equivalently the z-plane of the disk. Most cubic metals are the softest in $\langle 100\rangle$ and hardest in $\langle 111\rangle$, so there will be three variants of fully coherent disks if elastic energy is dominant (capillary energy of fully coherent precipitate tends to favor blocky shape). The [100], [010], [001] oriented disks can capitalize on the soft direction ($E_x = E_y = E_z = E_{\langle 100\rangle}$) to accommodate δ_{xx}, δ_{yy} more cheaply.

For larger precipitates or different crystal structures, full coherency can no longer be maintained. An Eshelby operation of inserting an object into the hole, where the object is larger/smaller than the hole as shown in Fig. 3.49 of [47], is necessary, if no long-range diffusion is possible. (If long-range mass transport by diffusion is allowed, the elastic stress can be plastically relaxed to zero, by for instance plating atoms to an outer free surface in diffusional creep, so the object exact fits in the hole, and there is no elastic energy and only interfacial energy). If the interface is semi-coherent or incoherent, it will be "shear-weak", so it can be considered "greased", which means the interface binds α,β vertically but horizontally they can easily slide with respect to each other. One could also say, at least ideally, that "friction" is zero on incoherent or semi-coherent interfaces, and infinite on coherent interfaces. This would significantly change $\Delta G_{\text{elastic}}$ versus the fully coherent "shear-strong" interface case. In the case of such zero-friction Eshelby interfaces with hydrostatic dilation $\delta_{xx} = \delta_{yy} = \delta_{zz} = \delta$ and isotropic elasticity, Nabarro has solved the total elastic energy (section 19.1.3 of [41]) to be

$$\Delta G_{\text{elastic}} \;=\; V^\beta \cdot 6\mu\delta^2 E\left(\frac{c}{a}\right) \tag{5.1}$$

for a stiff ellipsoid $(x/a)^2 + (y/a)^2 + (z/c)^2 = 1$ embedded in a compliant isotropic matrix of shear modulus μ, and the ellipsoid is larger than the hole by δ. The dimensionless function $E(0) = 0$, $E(1) = 1$ and $E(\infty) = 3/4$ [59]. For small x, $E(x) \approx 3\pi x/4$. Thus, spheres

($c/a = 1$) and needles ($c/a \gg 1$) are not that much different in elastic energy per volume, but disks ($c/a \ll 1$) are extremely cheap in elastic energy, per volume. Since $V^\beta = 4\pi a^2 c/3$, $\Delta G_{\text{elastic}} \propto ac^2$ for a thin disk.

The results for thin disk can be rationalized by the following. According to elementary elasticity solution, there is a stress amplification factor a/c right in front of a disk-shape (penny-shaped) hole, like in a lever. Thus, if the stiff ellipsoid expands by strain δ, the matrix material right in front would sustain strain $\delta a/c$, with associated strain energy density $\propto \delta^2 a^2/c^2$. However, the spatial extent of such highly stress region is very small. In the radial direction, the extent is only the local radius of curvature $R = (d^2x/dz^2)^{-1} = c^2/a$. The circumferential length is $2\pi a$, so the total volume of such highly stressed region is only $\propto c^4/a$. So the total elastic energy stored in the matrix is just $\propto ac^2$. Since the interface is incoherent, even ϵ_{xx}, ϵ_{yy} at the center of the disk can be relaxed, as the interface is "greased".

The remarkably simple (5.1) allows us to do some analysis. The total thermodynamic driving force for phase transformation:

$$\Delta G \;=\; \Delta G_{\text{soln}} + \Delta G_{\text{capillary}} + \Delta G_{\text{elastic}} \tag{5.2}$$

We see $\Delta G_{\text{elastic}}$ is positive, as usually is $\Delta G_{\text{capillary}}$ (except in the case of surface or GB premelting), so the latter two terms usually impede phase transformation. In a temperature-driven transformation:

$$\Delta G_{\text{soln}}(T) \;\approx\; \Delta S_{\text{soln}}(T_e)(T_e - T) \tag{5.3}$$

where ΔS_{soln} is the entropy of transformation from homogeneous solution thermo, and T_e is the temperature at which homogeneous solution thermo predicts the transformation would happen, if without the contraints of interfaces (such as very large objects) and elastic interactions (free-standing solid, or liquid,gas). We also note that $\Delta G_{\text{soln}} \propto V^\beta$. So the cheaper is $\Delta G_{\text{elastic}}$ per volume, the better. A blob of β would thus like to split into smaller disks with higher aspect ratios, to reduce the elastic cost per volume. Such spatial refinement (at least in one direction) will however be counterbalanced by the increase in interfacial energy.

In above we have only considered hydrostatic transformation strain, which involves only one strain variant for β. Many phase transformations could involve shear transformations of

several equivalent variants, for instance in cubic→tetragonal transformation:

$$\epsilon_{\text{variant1}} = \delta \begin{pmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}, \quad \epsilon_{\text{variant2}} = \delta \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}, \quad \epsilon_{\text{variant3}} = \delta \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & -1 \end{pmatrix} \tag{5.4}$$

corresponds to tetragonal precipitates of different orientation relationships with respect to the surrounding matrix. These precipitates act like electric dipoles in electromagnetism: they interact with each other via long-ranged stress fields, and so there are favorable mutual arrangements. (Previously we have only talked about a single precipitate in an infinite matrix; now we are talking about precipitate-precipitate interactions).

An extreme example of elasticity-stabilized fine microstructure and microstructural self-organization is the twinned martensite plate in steels. At high temperature steel is FCC (austenite, γ), whose solubility of carbon interstitials is large. At low temperature, the iron-rich phase in steel is BCC (ferrite, α) if the carbon atoms manage to leave to form cementite Fe_3C. If however the cooling rate $-dT/dt$ is large, the carbon may not get the chance to leave, in which case body-centered-tetragonal (bct, martensite, α') phase may form, which is supersaturated in carbon, with local tetragonal strains (Fig. 24.3 of [41]). Because cooling is so fast and diffusion is sluggish (that even carbon did not manage to leave), α' is formed purely by highly collective displacements (mainly shearing) of atoms, without diffusion. Such diffusionless shear-dominant phase transformation is generally called *martensitic transformation*. Before and after martensitic transformation, the atomic registry sustains a *deterministic* change (1st atomic neighbor may become 2nd atomic neighbor, but it will be the same way for *all atoms* within a single variant, unlike in a diffusive transformation). The particular diffusionless shear-dominant transformation in steel is called the Bain transformation, proposed by American metallurgist Edgar Bain who worked at US Steel in Pittsburgh. It is pretty closely described by (5.4), with $\delta \approx 0.2$, although the actual transformation ($\delta_{zz} = -0.2$, followed by lateral expansion $\delta_{xx} = \delta_{yy} = 0.13$, and $\sqrt{2} \times 0.8 \approx 1.1314$) is not exactly volume-preserving.

The reason for the mind-boggling nanostructure in a twin-accommodated martensite plate is this: a single variant growing as a disk along its shear directions is fine, it incurs very little elastic energy, as long as its aspect ratio c/a is small. However, unlike in ordinary alloy decomposition/precipitation reactions ($\gamma \to \alpha$+cementite), where just a finite volume fraction $f_\alpha < 1$ will become precipitate and then the system is happy, here as temperature cools *all* the austenites want to transform: $\gamma \to \alpha'$, $f_{\alpha'} = 1$, without the help of carbon diffusion.

So the transformed region must thicken to convert the adjacent unhappy austenites. But it cannot thicken indefinitely as just a single variant. In (5.1), if you double the thickness you double the elastic energy price, not for just the newly added region, but the old one as well. (the total elastic energy pricetag actually quadruples since $\Delta G_{\text{elastic}} \propto ac^2$). What can the martensite do? The surrounding austenites want to convert, but the elastic energy of adding onto this existing martensite variant has simply become too high.

This martensite variant then comes up with an ingenious idea of collaborating with his other two alter-egos (in fact one at a time - inside one martensite plate: $\gamma \rightarrow \alpha'_{\text{variant1}} + \alpha'_{\text{variant2}}$). As far as the austenite is concerned, it does not really matter which of the three variants it is converted into: there is the same $\Delta S_{\text{soln}}(T_e)(T_e - T)$ solution thermodynamic benefit to be gained. Co-transforming into two finely-spaced variants together does involve the additional overhead of interfacial energy: so both bct variants choose to grow in the twin plane that they share, across which they are twin related, so the interfacial energy overhead can be minimized (coherent twin boundary is the cheapest of all interfaces that can separate the two variants). They also choose to appear in different volume fractions within one martensite plate:

$$\bar{\boldsymbol{\epsilon}}_{\text{plate}} \;=\; w\boldsymbol{\epsilon}_{\text{variant1}} + (1-w)\boldsymbol{\epsilon}_{\text{variant2}}. \tag{5.5}$$

And here is a crucial trick: it turns out that by choosing the weight w judiciously (turns out to be 2/3), it is possible to for the strain fields of the two variants to largely cancel each other. In fact, by periodically stacking the two variants c_1/a, c_2/a, $c_1/c_2 = (2/3)/(1/3) = 2$ into a composite plate (the martensite plate), it is possible to create a martensite plate of length $c' \gg a$ and aspect ratio c'/a, where there is basically *zero* elastic stress over the lengthscale of c', despite the fact that $\delta \approx 20\%$, a huge transformation strain. This is something a single shear variant can never do. He can grow laterally as a single $c/a \gg 1$, but he can never thicken. To thicken without incurring exorbitant elastic energy, he must enlist the help of his alter-ego.

The broad side (length c') of this composite plate is called the habit plane, with normal z'. Mathematically, it is called an invariant plane, because there is no macroscopic distortion in the plane or rotation of the plane before and after the co-transformation: $\epsilon_{x'x'} = \epsilon_{x'y'} = \epsilon_{y'y'} = 0$. This can be achieved by tuning w and z' normal (the martensite plate "chooses its own hatching ground"). In the case of steel, the habit plane is irrational - it is not a crystallographic plane of the γ or α' phase. A great triumph in the development of the phase transformation theory was Wechsler, Lieberman and Read were able to predict in 1953 the habit plane inclination of twin-accommodated martensite plate in steels to within

1° from experimental observation [60] based on elastic energy argument alone, with *no fitting parameter*. The composite martensite plate does have finite macroscopic strains $\epsilon_{x'z'}$, $\epsilon_{y'z'}$, as well as dilation $\epsilon_{z'z'}$ (which are geometrically softened because $c' \gg a$): those in turn stimulate secondary organizations of the martensite plates.

To rigorously work through the above assertions requires quite elaborate matrix formalism, but there is a simple analogy to packing a suitcase. Imagine you have a suitcase of toys called Transformer Blocks$^{\text{TM}}$. These blocks are initially cubic, and pack well in your suitcase. But when you bring the suitcase on a train, suddenly they decide to transform, to either of two shear variants: $\epsilon_{xy}^{\text{variant1}} = -\epsilon_{xy}^{\text{variant2}} = 1$, with equal pleasure. Clearly, if they all decide to transform to one variant, you suitcase will burst due to excessive elastic energy. It is however, possible to have all Transformer Blocks$^{\text{TM}}$ transformed, half to variant 1 and half to variant 2. If you pack them in a careful way (variant 1 in one row, variant 2 in other row, and repeat), you can still fit them all inside your suitcase, incurring very little elastic energy on average. There are jagged edges in this composite, but there is only localized elastic energies near the suitcase, and there is very little pain inside the body. You may also decide to stack 2 variant1 rows + 2 variant2 rows + 2 variant1 rows + ... This arrangement has the merit of reducing the number of internal twin boundaries, by half. However, now the jagged edges are thicker, and the elastic pain penetrates a bit deeper into the body. So a finer microstructural organization of a multi-variant system reduces elastic energy but increases interfacial energy, while coarser microstructural organization reduces interfacial energy but increases elastic energy. The optimal microstructural lengthscale is then a compromise of the two effects. In steels, the twin laminar thickness is governed by the competition between twin boundary energy and elastic energy, and it turns out that tens of nanometers bi-layer spacing achieves the best compromise.

We thus see that elastic energy can promote microstructures to self-organize over a very long lengthscale, and sometimes even forming hierarchical structures. The root cause is that in most solid-solid phase transformations, there is shear transformation strain. Unlike hydrostatic transformation strain, shear transformation strain by nature tends to have multiple variants (shear transformation this way is as pleasurable as shear transformation some other way). A single shear variant may never grow very big under zero external load, constrained in a solid. But two shear strains can partly cancel each other's long-range elastic field, and grow *bigger together* in a particular arrangement. Thus elastic interactions promote collaboration between shear variants and partial cancellation of shear transformation strain over a fine lengthscale. In addition to twin-accommodated martensite, there are dislocation ac-

commodated martensite, where dislocation plasticity produces the so-called *lattice-invariant* shear, that is superimposed onto the transformation shear strain, to enable fitting "into the suitcase": $\gamma \to \alpha'_{\text{variant1}} + \alpha'_{\text{variant1plastic}}$. Because of the many possible plastic strains that can be generated by dislocation slip, dislocation accommodated martensite is more "flexible" and "variable shaped" than twin-accommodated martensite, but is less mathematically elegant. Kinetically, martensitic transformation happens very fast: the transformation front moves locally with nearly the speed of sound. Unlike in diffusional transformations, where one may need to wait minutes to hours for a certain transformation to go from 10% transformed to 90% transformed, martensitic transformation is almost "instantaneous", that is, holding at a certain temperature and waiting very long does not get a lot more transformed.

The supersaturated carbon in α' and the finely-spaced twin boundaries hardens the steel greatly (Fig. 19.20 of [1]). Martensitic transformation is technologically important, because one can harden steels a lot without expensive alloying elements (both iron and carbon are abundant), by simply quenching (i.e. fast-cooling) below M_{s}, the *martensite start temperature* (typically $\sim$300°C). After martensitic transformation is finished (below M_{f}, the *martensite finish temperature*), the material is so hard and still has considerable residual stress (remember this is low temperature and plastic flow is not as expedient as at higher temperatures when diffusion can happen), that it is prone to cracking. To improve the ductility, a *tempering* processing step usually ensues, where the temperature is raised to an intermediate level, to allow small-scale plastic flow to happen to relieve the internal stress, as well as allowing some supersaturated carbon to precipitate out as ϵ-carbide (hexagonal, $Fe_{2-3}C$), η-carbide (orthorhombic, Fe_2C) or cementite (orthorhombic, Fe_3C). Tempering improves the ductility without sacrificing the strength too much.

Martensitic transformation is also the basis for shape memory alloys (SMA) such as Ni-Ti, used in medical equipment. A stent can be inserted into blood vessel, and upon heating, returns to a pre-set shape to open up the blood vessel. An associated *super-elasticity* effect (same as *pseudo-elasticity* effect) can be used to make cell phone antennas, actuators, graspers, endoscope etc. Macroscopically, pseudo-elasticity entails energy dissipation [61], but unlike plasticity does not lock in another stress-free shape. Microscopically, pseudo-elasticity is based on deformation twinning, where one variant of transformation shear is converted into another variant of transformation shear, and shape change is accomplished (at finite stress) by varying the volume fraction of one variant. The shape change is reversible (but with dissipation) because there is no secondary process (like two slip systems intersect to form Lomer-Cottrell lock in general dislocation plasticity) or high-order processes that

strongly locks in the inelastic strain and make the reversed path almost impossible.

Chapter 6

Interfacial Mobility

As an introduction to *kinetic* factors, let us consider what can be accomplished by *short-range* diffusion, essentially one atom deciding to jump across an interface. Consider the cartoon Fig. 3.23(a) of [47] showing the GB of a pure metal. Suppose there are n_1 atoms per unit area (unit atoms/m^2) in grain 1 ready to make a jump to grain 2, with success rate given by $\nu_1 \exp(-Q_{1\to 2}/k_{\mathrm{B}}T)$ as in transition state theory. Similarly, there are n_2 atoms per unit area in grain 2 ready to make a jump to grain 1, with success rate given by $\nu_2 \exp(-Q_{2\to 1}/k_{\mathrm{B}}T)$. The total speed of GB motion would then be

$$v \;=\; \Omega \left(n_2 \nu_2 \exp(-Q_{2\to 1}/k_{\mathrm{B}}T) - n_1 \nu_1 \exp(-Q_{1\to 2}/k_{\mathrm{B}}T) \right). \tag{6.1}$$

(From now on we will use $\mathbf{v}$ to denote vector velocity, v to denote magnitude of velocity, and Ω to denote volume). $Q_{2\to 1}$ and $Q_{1\to 2}$ are related:

$$Q_{1\to 2} \;=\; G^* - G_1, \;\; Q_{2\to 1} \;=\; G^* - G_2 \;=\; Q_{1\to 2} - \Delta G, \quad \Delta G \;\equiv\; G_2 - G_1. \tag{6.2}$$

If $G_2 > G_1$, then the boundary should move to the right and v is positive. The above assumes there is only one microscopic pathway to go from 1 to 2, and going from 2 to 1 uses the same pathway in reverse. In reality there can be multiple microscopic pathways, but so-called *principle of detailed balance* will make the result qualitatively the same.

Suppose the GB is flat, then $G_2 = G_1$, and we must have $v = 0$. That is to say, the rule of kinetics must be consistent with the laws of thermodynamics. So there must be $n_1\nu_1 = n_2\nu_2$,

and we can define the equilibrium *exchange* flux:

$$J_{\text{exchange}} \;=\; n_1\nu_1 \exp(-Q_{\text{exchange}}/k_{\text{B}}T), \quad Q_{\text{exchange}} = G^* - G_1. \tag{6.3}$$

Now consider a small elevation in G_2, due to for instance a curved GB with radius of curvature towards grain 2. The same elevation in free energy can achieved also by an elastic stress - as far as the jumping atoms are concerned it does not matter *how* $\Delta G = G_2 - G_1$ is generated, "money is money". Then, we have:

$$\begin{aligned} v = \quad & \Omega\left(n_2\nu_2 \exp(-(Q_{\text{exchange}} - \Delta G)/k_{\text{B}}T) - n_1\nu_1 \exp(-Q_{\text{exchange}}/k_{\text{B}}T)\right) \\ = \quad & \Omega J_{\text{exchange}}(\exp(\Delta G/k_{\text{B}}T) - 1) \\ \approx \quad & \frac{\Omega J_{\text{exchange}}}{k_{\text{B}}T}\Delta G, \end{aligned} \tag{6.4}$$

if ΔG is finite but $\ll k_{\text{B}}T$. Recall that in the case of Young-Laplace pressure, $\Delta G = \Omega\Delta P$, so we get:

$$v \;=\; \frac{\Omega^2 J_{\text{exchange}}}{k_{\text{B}}T}\Delta P \;=\; M\Delta P, \tag{6.5}$$

where M is called the *boundary mobility*.

Mobility is always the ratio between a velocity and a force (unit J/m) for discrete moving objects, or a force density (J/m^3) in the case of continuous moving boundaries. For the mobility concept to be applicable, the driving force must be sufficiently small, so it is in the so-called *linear-response* regime, where the response (velocity, flux, current etc.) is linearly proportional to the driving force. Probably the best known example of mobility is that of a spherical microbead embedded in viscous liquid:

$$\mathbf{v} \;=\; M\mathbf{F} = \frac{\mathbf{F}}{6\pi\eta R}, \tag{6.6}$$

where $\mathbf{F}$ is a persistent dragging force, η is the liquid viscosity and R is the bead's radius, because Einstein encountered this formula, initially derived by Stokes, in his study of Brownian motion in 1905. A previously unexpected connection between mobility

$$M \;=\; \frac{D}{k_{\text{B}}T} \tag{6.7}$$

and D, diffusivity of microbeads was revealed in Einstein's 1905 study. In fact, $\Omega^2 J_{\text{exchange}}$ in our (6.5) can already be identified as the "diffusivity" D^* of the GB location if we consider

fluctuations in the exchange flux will cause the location of a unit-area GB to perform random walk with mean squared displacement $2D^*t$ [62].

Atom hops are thermally activated. To extract the effective activation energy, one could plot $\ln(Mk_{\mathrm{B}}T)$ with respect to $1/T$, and the slope would be

$$-\frac{\partial(Q_{\mathrm{exchange}}/k_{\mathrm{B}}T)}{\partial(1/T)} = -\frac{H_{\mathrm{exchange}}}{k_{\mathrm{B}}}, \quad Q_{\mathrm{exchange}} = H_{\mathrm{exchange}} - TS_{\mathrm{exchange}}, \tag{6.8}$$

where Q_{exchange} is the activation free energy, H_{exchange} is the activation enthalpy, and S_{exchange} is the activation entropy, between the saddle state and 1 (or 2) at equilibrium.

It turns out that solutes can have a huge effect on interfacial mobility. GBs in pure metals can move many orders of magnitude faster than GBs in alloys. This is because impurities like to be trapped inside GBs, where there is larger "free volume". Reciprocally, they exert a "frictional" force on GB motion, like dusts for cogs in the wheel. Since random GBs have larger free volume, their mobilities are more susceptible to alloying (see Fig. 3.27 of [47], just 0.006 wt% of tin is able to reduce the mobility of random GBs in high-purity lead by factor of 10^4!). Thus, one important reason for alloying is to stablize the microstructure, i.e. grain size. Grain size is really key for many properties, for instance grain size contributes to the overall strength by Hall-Petch relation $\sigma(D) = \sigma_0 + kD^{-1/2}$. Without control of grain size, one has no control over microstructure.

Let us consider the problem of *grain growth kinetics*. One should realize that a polycrystal is at best a metastable system thermodynamically. It is favorable to evolve the microstructure to have coarser and coarser grains, to reduce internal boundaries and capillary energy. But exactly how does this occur and at what rate?

One starts by considering the GB triple junctions (triple lines in 3D). If γ is isotropic, the equilibrium dihedral angle should be 120°. Let us assume the triple junction mobility is much more facile than GB mobility, that is to say let us assume the triple junction moves *as if instantaneously* so the total force on the triple junction is always zero and the dihedral angle is always 120°, and only the GB mobility effectively impedes grain growth. Let us compare a grain with N sides and N triple junctions to a polygon with N vertices. The sum of the interior angles of any N-sided polygon is $(N-2)\pi$. Compare this to the sum of dihedral angles $2N\pi/3$, we see that $(N-2)\pi < 2N\pi/3$ if $N < 6$. Therefore if $N < 6$, there must be some dihedral angles that will include the polygon interior angle at the same vertex, associated with which are concave inward GB segments that will shrink the grain,

due to positive Young-Laplace pressure. On the other hand, if $N > 6$, then there must exist polygon angles which includes the dihedral angle, so the corresponding GB segments are convex outward, and will tend to grow the grain. We also expect the smaller grains to have less number of sides, while the larger grains have more sides. Thus, smaller grains with small number of sides will tend to shrink, and the larger grains with larger number of sides will tend to expand.

In fact, consider columnar grains (so-called 2D grain growth), we have the von Neumann-Mullins equation (Chap. 15 of [41])

$$\begin{aligned}\frac{dA}{dt} &= \sum_{\text{sides}} \int dlv = -\sum_{\text{sides}} \int dlM(\gamma\kappa) = -M\gamma \sum_{\text{sides}} \int dl\kappa = -M\gamma \sum_{\text{sides}} \int \frac{dl}{R} \\ &= -M\gamma \sum_{\text{sides}} \int d\theta = -M\gamma(2\pi - N \times \pi/3) = \frac{M\gamma\pi}{3}(N-6), \end{aligned} \tag{6.9}$$

which was recently generalized to 3D by MacPherson and Srolovitz [63]. Therefore, the average grain size of surviving grains would go as $\frac{d\bar{A}}{dt} = \frac{M\Omega\pi}{3}(\bar{N}-6)$. [1] In a self-similar grain growth, we expect the lengthscale of microstructure to change with time, but not topological characteristics, so we expect $\bar{N}$ is constant, then $\bar{A} = \bar{A}_0 + kt$, and then we expect the grain size to grow as $\bar{D} = \sqrt{\bar{D}_0^2 + kt}$.

The above is an example of *parabolic kinetics*: the grain size growth rate is fast at beginning when the lengthscale is small, but slows down as the lengthscale coarsens. If we sacrifice a bit of rigor, there is an alternative derivation that is perhaps physically more illuminating. We say that

$$\frac{d\bar{D}}{dt} \propto v \propto M\frac{\gamma}{R} \propto M\frac{\gamma}{\bar{D}} \rightarrow \frac{d(\bar{D}^2)}{dt} \propto M\gamma \rightarrow \bar{D}^2 = \bar{D}_0^2 + kt, \tag{6.10}$$

with $k \propto M\gamma$.

Even though as time goes on the growth rate slows down, such microstructural evolution may still not desirable. How can we shut down grain growth *completely* at some desired grain size $\bar{D}_{\text{desired}}$? It turns out that with small second-phase particles such as oxide, sulfide and silicate inclusions, we can pin down GB motion, with so-called Zener pinning mechanism. Consider

[1] There are several ways to prove/interpret this. See for example p. 378 of [41]. A simpler way is to define area-weighted average: $\bar{Y} \equiv \sum_i Y_i A_i / \sum_i A_i$, where the vanished grains have $A_i = 0$. Thus $\bar{A} \equiv \sum_i A_i^2 / \sum_i A_i$, $\dot{\bar{A}} = 2\sum_i A_i \dot{A}_i / \sum_i A_i = 2\sum_i A_i \frac{M\Omega\pi}{3}(6-N_i)/\sum_i A_i = 2\frac{M\Omega\pi}{3}(\bar{N}-6)$. We get an extra factor of 2, but still linear growth in $\bar{A}(t)$.

a situation where the interface between inclusion and matrix is completely incoherent, and therefore the inclusion does not care about the orientation of matrix. When GB intersects the inclusion, the grain boundary area is reduced by πr^2, and thus the total energy is reduced by $\gamma\pi r^2$ compared to when the GB is detached from the inclusion. The detachment occurs over a lengthscale r, so the maximum pinning force is estimated to be $F_{\max} \propto \gamma r$. A more precise calculation is to say that $F = 2\pi r \sin\theta \cdot \gamma\cos\theta$, where $2\pi r\sin\theta$ is the circumferential length of intersection, and $\gamma\cos\theta$ is because the GB must be perpendicular to the inclusion surface. The total drag force is maximized at $\theta = 45°$, and $F_{\max} = \pi\gamma r$.

Consider volume fraction $f \ll 1$ of these oxide inclusions, in random dispersion (often f is between 10^{-2} and 10^{-3}). The mean distance between two nearest-neighbor inclusions is $l \propto rf^{-1/3}$. So one grain boundary of area $A \propto D^2$ can make contact with at most $N_{\text{inc}} \propto D^2/l^2 = f^{2/3}D^2/r^2$ inclusions. The maximum force exerted by these particles against motion of one GB is thus $\propto f^{2/3}D^2/r^2 \cdot \gamma r$, which is $\gamma f^{2/3}/r$ per GB area (call it the "pinning pressure", which can be $\pm$ that is always against the direction of motion). Contrast this with the original driving force:

$$\frac{d\bar{D}}{dt} \;\propto\; M\left(\frac{\gamma}{\bar{D}} - \frac{\gamma f^{2/3}}{r}\right), \tag{6.11}$$

we see that the growth can be arrested when $\bar{D}$ reaches $\bar{D}_{\text{stall}} = rf^{-2/3}$. N_{inc} used above is an upper bound, a more conservative estimate would be $N'_{\text{inc}} \propto fD^2/r^2$, if we assume the GB is a random *flat* plane, that is not trying to *bend* to "touch" the particles. Then we would derive the growth can be arrested at $\bar{D} = \bar{D}_{\text{stall}} \propto rf^{-1}$. Either $\bar{D}_{\text{stall}} = rf^{-1}$ or $rf^{-2/3}$, we see that for a fixed volume fraction f, the finer the particles (and thus more numerous), the more potent is Zener pinning, since smaller grain size can be stabilized. In oxide dispersion strengthened (ODS) steels, the (Y,Ti,Al,Mg)-O particles is often <10nm in diameter and are very potent GB pinners. The flip side to this is that if the material is exposed to high temperature, and diffusion of oxygen and solutes is allowed in the matrix, then these particles would coarsen: $r \uparrow$ while the volume fraction f is constant. Then as the oxide particles coarsen, grain growth would also happen, and very soon your material becomes garbage. The point made here is that the stabilities of microstructures are interrelated: if one kind of microstructure is destabilized, it is likely to impact the stability of other microstructures.

In above, the GBs move in the direction of positive (concave) curvature. Is it possible for the GBs to move in the direction of negative (convex) curvature? (see Fig. 3.26 of [47]). Yes, if

the thermodynamic driving force contains more terms than just the Young-Laplace pressure. In the case of recrystallization, a heavily worked material leaves a great amount of dislocation debris inside the grain. Most of these dislocation content is statistically stored, that is to say having no net Burgers vector, such as dense arrays of dislocation dipoles. These content can be eliminated when a GB sweeps across, working like a trash collector and incinerator. The thermodynamic driving force in this case is the elimination of dislocation core energy and short-range elastic energy associated with the dislocation dipoles. The boundary is convex because the γ is actually trying to keep up:

$$\frac{d\bar{D}}{dt} \propto M\left(\frac{G_{\text{dislocation}}}{V} - \frac{\gamma}{\bar{D}}\right). \tag{6.12}$$

Chapter 7

Nucleation, Growth and Coarsening

Imagine a homogeneous matrix phase α. Nucleation means the appearance of something quite different from α in a localized region. If we use order parameter $\eta(\mathbf{x})$ (could be local density, for example, in the case of liquid-solid transition) to describe the whole system, we have $\eta(\mathbf{x}) = \eta^\alpha$ uniformly before the nucleation, and $\eta(\mathbf{x}) = \eta^\beta$ in small regions of $\mathbf{x}$ after the nucleation, where η^β differs from η^α by a *finite amount*. In other words, nucleation are disturbances to $\eta(\mathbf{x})$ which are large in *amplitude* and small in *spatial extent*. In a multi-component system, the concentration field $\mathbf{c}(\mathbf{x})$ is often a natural choice for the order parameter field $\eta(\mathbf{x})$; however associated with chemical changes there can be structural, electrical and magnetic changes as well.

Growth means enlargement of the $\eta(\mathbf{x}) = \eta^\beta$ spatial domain. Nucleation and growth is Nature's strategy to accomplish all *first-order* phase transitions. In contrast, in second-order phase transitions such as spinodal decomposition and order-disorder transformation, the system does not necessarily need to go through a nucleation stage. In that case, disturbances to $\eta(\mathbf{x})$ which are infinitesimal in *amplitude* and large in *spatial extent* can increase in amplitude.

Nucleation and growth can be regarded as involving two players: the matrix phase α, and a single contiguous region of β. In the coarsening stage, there will be three or more players, where *multiple* β regions interact, sometimes mediated by the matrix.

Referring back to (5.2), let us first discuss about ΔG_{soln}, the bulk/volumetric/solution free energy. ΔG_{soln} has two characteristics: (a) $\Delta G_{\text{soln}} \propto V^\beta$, the volume of transformed region. (b) In temperature-driven first-order transitions, $\Delta G_{\text{soln}} \propto \Delta T \equiv T_e - T$, where T_e is the bulk equilibrium temperature between α and β, if only solution thermodynamics is taken

into account.

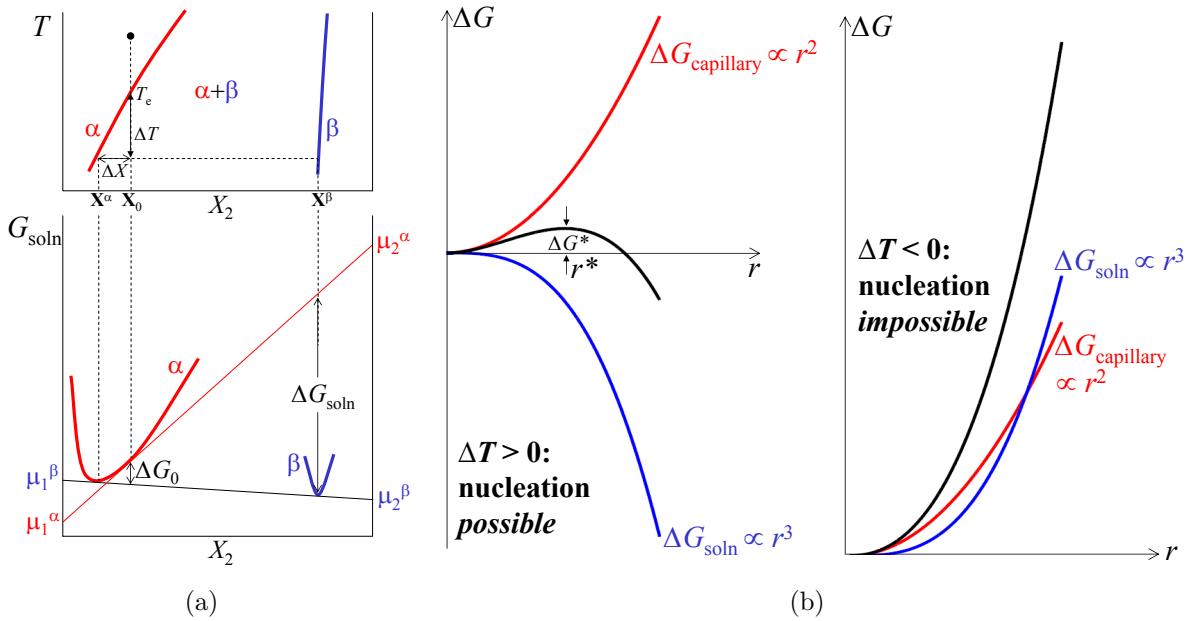


Figure 7.1: (a) Solution thermodynamics driving force for nucleation of β precipitate in α matrix. (b) Total thermodynamics driving force for nucleation, $\Delta T > 0$ versus $\Delta T < 0$.

To illustrate this in a multi-component system, consider precipitation reaction $\alpha(\mathbf{X}_0) \rightarrow \alpha(\mathbf{X}^\alpha) + \beta(\mathbf{X}^\beta)$ in a binary alloy, driven by ΔT as shown in Fig. 7.1(a). $\Delta\mathbf{X} \equiv \mathbf{X}_0 - \mathbf{X}^\alpha$ is called solute *supersaturation*. The supersaturation is linearly proportional to ΔT for small ΔT, as shown on the phase diagram. Let us consider how a small β domain of volume V^β and N^β atoms can be nucleated inside an infinite α matrix. $N^\beta X_1^\beta$ type-1 atoms and $N^\beta X_2^\beta$ type-2 atoms are needed to constitute the nuclei. Let us assume $\Omega_1^\alpha = \Omega_1^\beta = \Omega_2^\alpha = \Omega_2^\beta = \Omega$ for simplicity, so we do not have to consider elastic energy for the moment (the Eshelby "hole" is exactly the same size as the precipitate). We need to take $N^\beta X_1^\beta$ type-1 atoms and $N^\beta X_2^\beta$ type-2 atoms from the matrix (now in supersaturated composition $\mathbf{X}_0$), which requires $N^\beta X_1^\beta \mu_1^\alpha(\mathbf{X}_0) + N^\beta X_2^\beta \mu_2^\alpha(\mathbf{X}_0)$ free energy. When these atoms are remixed and form the β phase, the energy becomes $N^\beta X_1^\beta \mu_1^\beta + N^\beta X_2^\beta \mu_2^\beta$, so the solution driving force in this initial stage of phase transformation is

$$\Delta G_{\text{soln}} \;=\; N^\beta X_1^\beta(\mu_1^\beta - \mu_1^\alpha(\mathbf{X}_0)) + N^\beta X_2^\beta(\mu_2^\beta - \mu_2^\alpha(\mathbf{X}_0)) \tag{7.1}$$

Note that although $\mu_1^\beta(\mathbf{X}^\beta) = \mu_1^\alpha(\mathbf{X}^\alpha)$ and $\mu_2^\beta(\mathbf{X}^\beta) = \mu_2^\alpha(\mathbf{X}^\alpha)$ by definition, the solute-depleted matrix composition $\mathbf{X}^\alpha$ is nowhere to be had at the *beginning stage* of precipitation.

Compared to the supersaturated matrix composition which is what we have now, there is $\mu_1^\beta \neq \mu_1^\alpha(\mathbf{X}_0)$ and $\mu_2^\beta \neq \mu_2^\alpha(\mathbf{X}_0)$. We see that $\Delta G_{\text{soln}} \propto N^\beta \propto V^\beta$. From the graphical construction in Fig. 7.1(a), we also see that $\Delta G_{\text{soln}} \propto \Delta\mathbf{X} \propto \Delta T$. Thus, we can define *solution driving force per transformed volume* $g_s \equiv -\Delta G_{\text{soln}}/V^\beta$. And we know that $g_s \propto \Delta T$.

Ignoring elastic energy for the moment (it should be zero for purely diffusional transformation with $\Omega_1^\alpha = \Omega_1^\beta = \Omega_2^\alpha = \Omega_2^\beta = \Omega$), we have

$$\Delta G \;=\; -g_s V^\beta + \int \gamma dA \tag{7.2}$$

If γ is isotropic, the best shape for ΔG for a given V^β is a sphere. Thus,

$$\Delta G \;=\; -g_s \frac{4\pi r^3}{3} + \gamma 4\pi r^2. \tag{7.3}$$

The first term is proportional to r^3, the second term is proportional to r^2. For a normal system with positive surface/interfacial energy γ, the second term is positive. Thus, for small nuclei sizes, the total energy always *increases* with increasing r, due to Young-Laplace pressure and the prevalence of surface effects (large surface-to-volume ratio) at small nuclei sizes. This means when an atom attaches to a small nuclei, it always find the small nuclei "unattractive", and would preferentially detach. The only reason we see small nuclei at all is because according to the Boltzmann distribution, the probability of seeing N^β-cluster $\propto \exp(-\Delta G(N^\beta)/k_{\text{B}}T)$, that even if $\Delta G(N^\beta)$ is very unattractive there is still finite chance (though very small) of finding it.

If $g_s > 0$ ($\Delta T > 0$), however, there exists a critical radius as shown in Fig. 7.1(b) where $\Delta G(N^\beta)$ finally starts to go down with additional atom attachment: $N^\beta \to N^\beta + 1$. This *saddle-point* configuration or critical nuclei occurs at size

$$0 \;=\; -g_s 4\pi r^{*2} + 8\pi r^* \gamma \quad\to\quad r^* \;=\; \frac{2\gamma}{g_s}. \tag{7.4}$$

Thus, $r^* \propto (\Delta T)^{-1}$. The smaller the undercooling, the larger the critical nuclei needs to be. Plugging $r^* = 2\gamma/g_s$ back into (7.3), we get

$$\begin{aligned} \Delta G^* \;&=\; -g_s \frac{4\pi(2\gamma/g_s)^3}{3} + \gamma 4\pi (2\gamma/g_s)^2 \\ &=\; -\frac{32\pi\gamma^3}{3g_s^2} + \frac{16\pi\gamma^3}{g_s^2} \end{aligned}$$

$$= \frac{16\pi\gamma^3}{3g_s^2}. \tag{7.5}$$

Two observations can be made: (1) $\Delta G^* \propto (\Delta T)^{-2}$, which is a strong dependence: when ΔT is small, ΔG^* diverges and there is no chance. (2) The volumetric ΔG_{soln} contribution is negative, the capillary $\Delta G_{\text{capillary}}$ is positive, and right at r^* the former is always 2/3 of the latter. This is generically true for any $\max_r -ar^3 + br^2$. N^β is related to r as $N^\beta = 4\pi r^3/3\Omega$, if we assume 1 and 2 have the same volume.

If $g_s < 0$ ($\Delta T < 0$), then ΔG is a monotonically increasing function of β size. Equilibrium distribution (fluctuation) of β cluster sizes can be achieved in this case, whose concentration should be $C(N^\beta) = \Omega^{-1}\exp(-\Delta G(T, N^\beta)/k_{\text{B}}T)$ where $c = \Omega^{-1}$ is the atom concentration. Typically in metals $c \sim 10^{29}/\text{m}^3$. $C(N^\beta)$ here is monotonically smaller for larger cluster sizes. This means β-like cluster shows up occasionally, but quickly decompose by $N^\beta \to N^\beta - 1$, which is energetically more favorable than $N^\beta \to N^\beta + 1$.

Now imagine ΔT is suddenly switched from negative to positive. The energy landscape changes at every r, which is significant especially for $r > r^*$. This system is like a leaky kettle: it cannot reach true thermodynamic equilibrium unless all the water is leaked (all α transformed into β). But we are interested in the leaking process. Consider $N^\beta \to N^\beta + 1$: it is still energetically punishing at small sizes, but less so than originally, so one should see *more* larger clusters by and by. But this cannot happen immediately. Some time is needed to see changes in the cluster size distribution $\tilde{C}(N^\beta, t)$. This is because *atom attachment* takes time. It takes some some time after the $-\Delta T \to \Delta T$ switch for the original $C(N^\beta)$ to develop into $\tilde{C}(N^\beta, t)$ distribution that has significant quasi-steady state value at r^*. This waiting time for the $\tilde{C}(N^\beta, t)$ distribution to transform significantly is called the incubation time t_{inc}. After t_{inc} the nucleation rate will approach a quasi-steady state value, if the supersaturation and α volume are held constant.

A crude estimate for the nucleation rate $\mathcal{N}$ (unit $1/\text{m}^3/\text{s}$) is the following. The quasi-steady state $\tilde{C}(N^{\beta *}, t)$ value should look something like $\Omega^{-1}\exp(-\Delta G^*/k_{\text{B}}T)$, which is the average number of critical sized nuclei per volume. We multiply this by the frequency scale of atom attachment, Γ. This would give the number of newly generated *super-critical* nuclei per volume per time. These super-critical nuclei are likely to grow larger and larger, since $N^\beta \to N^\beta + 1$ by then would start to reduce energy rather than increase energy. Of course, the critical sized nuclei needs to be resupplied to maintain quasi-steady state value. To solve this whole problem consistently requires the so-called Master Equation (Chap.

19 of [41]) approach, which will give some dimensionless prefactor on the nucleation rate (Zeldovich factor) that is on the order of 10^{-1}. Brushing over such niceties, $\mathcal{N}$ may be modeled phenomenologically as:

$$\mathcal{N} \;=\; \Gamma\Omega^{-1}\exp(-\Delta G^*/k_{\mathrm{B}}T). \tag{7.6}$$

Γ for liquid→solid transitions is approximately a constant, 10^{11}/s. This means that in order to reach a "significant" nucleation rate $\mathcal{N}_{\text{significant}} = 1/\text{cm}^3/\text{s}$ with homogeneous nucleation, ΔG^* needs to be about $78k_{\mathrm{B}}T$. Taking the melting point of Cu, $T_M = 1357.77$K as the temperature scale, this means $\Delta G^* \approx 9$ eV.

Γ in solid→solid transitions should be proportional to the inter-diffusivity $\tilde{D}$, since solute partition (more type-2 atoms flowing into the nuclei while more type-1 atoms flowing into the matrix) is needed. Thus, Γ is also a strong function of T, roughly with $\ln\Gamma$ versus $1/T$ giving $(h_V^f + h_V^m)/k_{\mathrm{B}}$, the vacancy formation enthalpy plus vacancy migration enthalpy. Since $\exp(-\Delta G^*/k_{\mathrm{B}}T)$ is a growing function of ΔT (with sharp thresholding behavior), while Γ is a decreasing function of ΔT, $\mathcal{N}(\Delta T)$ should have a maximum at some $(\Delta T)^{\text{best}}_{\text{nucleation}}$. If $\Delta T < (\Delta T)^{\text{best}}_{\text{nucleation}}$ the saddle point is too high, if $\Delta T > (\Delta T)^{\text{best}}_{\text{nucleation}}$ the rate of atom attachment is too low.

(7.5) shows that the interfacial energy γ appears in cubic power in ΔG^*. In solid→solid transformation, since coherent interfaces have $\gamma = 1 - 200$ mJ/m^2 while semicoherent interfaces has $\gamma = 200 - 500$ mJ/m^2, the precipitate will try everything it can (adjusting shape and orientation relationship) to minimize the interfacial energy. Sometimes it even means precipitating out metastable phases *first*, such as GP zones, which have lower g_s but lower γ as well due to coherent interfaces. When the GP zones grow large enough, it will then transform to the more stable precipitate β later with associated coherency loss transition: $\alpha \to \text{GP} \to \beta$, in a two-step process.

From homework, we have seen that if there is a *pre-existing* rigid boundary (for example mold wall in casting liquid→solid), the total volume and interfacial energy of the spherical cap solid is:

$$V^\beta \;=\; \frac{4\pi r^3}{3}\frac{(2+\cos\theta)(1-\cos\theta)^2}{4}, \quad \Delta G_{\text{capillary}} \;=\; 4\gamma\pi r^2\frac{(2+\cos\theta)(1-\cos\theta)^2}{4}, \tag{7.7}$$

where r is the radius of curvature of the spherical cap, θ is the contact angle: $\gamma_{\text{ML}} = \gamma\cos\theta + \gamma_{\text{MS}}$, M stands for mold wall in the case of solidification (solid spherical cap grows

on mold wall), and $\gamma \equiv \gamma_{\rm SL}$ here. Since both the volume and capillary energy are scaled by the same factor $(2+\cos\theta)(1-\cos\theta)^2/4$, the activation energy of such *heterogeneous nucleation* on the mold wall is simply:

$$\Delta G_{\rm het} \;=\; S(\theta)\Delta G_{\rm hom}, \quad S(\theta) = \frac{(2+\cos\theta)(1-\cos\theta)^2}{4}. \tag{7.8}$$

When $\theta = 30^\circ$, $\Delta G_{\rm het}$ is only 1.286% of $\Delta G_{\rm hom}$. When $\theta = 10^\circ$, $\Delta G_{\rm het}$ is only 0.017% of $\Delta G_{\rm hom}$! (and $\Delta G_{\rm het}$ then goes into the exponential). Thus, heterogeneous nucleation is usually much more likely than homogeneous nucleation. The only way homogeneous nucleation rate has been measured was by breaking up the liquid into tiny droplets supported on amorphous substrate [64]. In such cases undercooling as large as 200K has been observed.

Most metals has $\gamma_{\rm S} > \gamma_{\rm LS} + \gamma_{\rm L}$, which means the solid metal likes to be covered completely by its own melt, instead of as liquid domes or droplets on top. The contact angle $\theta = 0$, and $S(\theta) = 0$. So no super-heating is typically seen in melting. The solid surface disorders and the surface liquid layer thickens as soon as the thermodynamic melting point T_M is reached.

Next we discuss about growth, which is defined as enlargement of the $\eta(\mathbf{x}) = \eta^\beta$ domain after nucleation. This enlargement can either be accomplished by *glissile* motion of the $\alpha\beta$ interface that involves shear (think gliding motion of $[11\bar{2}]a_0/6$ partial dislocation on a FCC twin interface), in which case it is called a *military* transformation; or it could be accomplished by randomized atom jumps across the interface, in which case it is called a *civilian* transformation. In a military transformation, the nearest-neighbor relation between atoms is either unchanged (first-nearest-neighbor on the left becomes first-nearest-neighbor on the right), or if changed (first-nearest-neighbor on the left becomes second-nearest-neighbor on the left), change in a *deterministic* and *uniform* way for *all* atoms of the same precipitate. The atoms move as a *group* with no individuality, and thus the whole variant will tend to carry large shape strain. Martensitic transformations (chapter 5) and deformation twinning are examples of military transformations. Because of the collectivity of the transformation mechanism (large activation volume [65]), the speed of military transformation is either essentially zero if the driving force is insufficient, or exceedingly fast when the driving force reaches the threshold (the moving speed of $\alpha\beta$ interface approaches speed of sound in some martensitic transformations). Such kinetics is sometimes called "athermal", which means the growth speed - when the interface moves - does not depend sensitively on temperature.

In contrast, the atomic nearest-neighbor relations are disrupted in a *randomized* and *nonuniform* way before and after a civilian transformation. Instead of "thinking and acting" as a

group, each atom "thinks and acts on its own". There can be shape strain as well (in the Eshelby sense, chapter 5), but they tend to be smaller in magnitude and of a more hydrostatic character. This is because without the *constraint* of keeping nearest-neighbor relation and moving as group, the randomized atomic jumps may relax away some shear strain. As we have seen, both the rate of short-range diffusion (interfacial migration) and long-range diffusion (interdiffusion for solute partitioning) are very sensitive to temperature, but are less sensitive to the driving force (linear response means flux/rate is proportional to driving force, such exponent of 1 is much less than the more collective military shear transformations, due to the small activation volume of diffusion [65]). It does not take a great driving force beyond the critical nucleus to grow in a civilian fashion (in linear response, if the driving force is halved, the flux/growth speed is halved - there is no sharp threshold), but the speed of growth when it is growing depends quite sensitively on temperature. Such growth kinetics is called "thermally activated" growth. The words "athermal" and "thermally activated" growth can be confusing: in reality every process is thermally activated, it is just that there are quantitative differences in the driving force and temperature dependencies due to differing degree of collectivity [65].

Because both short-range and long-range diffusions alter nearest-neighbor relations in a random fashion, a 100% pure military transformation cannot involve diffusion, and must be completely *displacive*. Military transformation is therefore sometime also called *displacive transformation*, and civilian transformation called *diffusional transformation*. Many phase transformations are of mixed military/civilian character: there may be large collective shear displacements, but diffusion is also necessary. For example, austenite→bainite transformation in steel at intermediate temperatures involve large collective shear displacements indicated by the appearance of surface reliefs [66], but diffusion must also have occured due to solute partitioning detected experimentally. This is in contrast to austenite→ferrite or pearlite transformation at higher temperatures which is completely civilian, or austenite→martensite transformation at lower temperatures which is completely military. As John Wyrill Christian remarked, "the main categories of transformation are called military and civilian, but rigid classifications is not required since soldiers may sometimes be out of step and civilians may sometimes form paramilitary organizations!" (p.6 of [67]).

Growth in civilian transformation can be *interface controlled* or *diffusion controlled*, depending on *where* the free energy is mostly spent on: driving interfacial migration or long-range diffusion. (Here the term "diffusion controlled" means *long-range diffusion*). Consider the quasi-1D binary system shown in Figure 3.67 of [47] as an example. Let us assume the

precipitate β is essentially pure type-2 atoms, so the motions of type-2 atoms drives the kinetics. Let us define the composition profile in α to be $\mathbf{X}^\alpha(x)$. At $x = \infty$, the matrix is in supersaturation: $\mathbf{X}^\alpha(x = \infty) = \mathbf{X}_0$. Right at the interface $x = x_i$, it takes a limiting value $\mathbf{X}^\alpha(x_i) = \mathbf{X}_i^\alpha$, which is smaller than $\mathbf{X}_0$ since the supersaturation is being drained to grow β. On the other hand, $\mathbf{X}_i^\alpha$ must be larger than $\mathbf{X}_e^\alpha$, of $\alpha(\mathbf{X}_0) \to \alpha(\mathbf{X}_e^\alpha) + \beta(\mathbf{X}_e^\beta)$ in the equilibrium phase diagram. This is because the jumping across of type-2 atoms over the $\alpha\beta$ interface requires some chemical potential driving force as well. Based on our assumption, $\mathbf{X}_e^\beta = \mathbf{p}_2$ (pure 2), the composition profile in β is $\mathbf{X}^\beta(x) = \mathbf{p}_2$ uniformly. Mimicking what we have done for grain boundary migration (chapter 6), we can write down:

$$v \;=\; M\frac{\mu_2^\alpha(\mathbf{X}_i^\alpha) - \mu_2^\beta(\mathbf{p}_2)}{\Omega} \tag{7.9}$$

where $\frac{\mu_2^\alpha(\mathbf{X}_i^\alpha) - \mu_2^\beta(\mathbf{X}_e^\beta)}{\Omega}$ has the unit of pressure, and M is the mobility of this continuous boundary (unit $\mathrm{m^2s/kg}$).

We note that $\mu_2^\beta(\mathbf{p}_2) = \mu_2^\alpha(\mathbf{X}_e)$, so the total driving force $\mu_2^\alpha(\mathbf{X}_0) - \mu_2^\beta(\mathbf{p}_2) = \mu_2^\alpha(\mathbf{X}_0) - \mu_2^\alpha(\mathbf{X}_e)$ can be decomposed into $\mu_2^\alpha(\mathbf{X}_0) - \mu_2^\alpha(\mathbf{X}_i^\alpha)$ plus $\mu_2^\alpha(\mathbf{X}_i^\alpha) - \mu_2^\alpha(\mathbf{X}_e)$, the former used to drive long-range diffusion, the latter used for jumping across the interface. In fact, if α is an ideal solution, then

$$\mu_2^\alpha(X_2) \;=\; \mu_2^{\alpha\circ} + k_\mathrm{B}T\ln X_2, \quad \mu_2^\alpha(X_2 + \Delta X) - \mu_2^\alpha(X_2) \;\approx\; \frac{k_\mathrm{B}T}{X_2}\Delta X. \tag{7.10}$$

So the driving forces are approximately

$$\Delta_\mathrm{interface}\mu_2 \;\approx\; \frac{k_\mathrm{B}T}{X_{e2}^\alpha}(X_{i2}^\alpha - X_{e2}^\alpha), \quad \Delta_\mathrm{diffusion}\mu_2 \;\approx\; \frac{k_\mathrm{B}T}{X_{e2}^\alpha}(X_{02} - X_{i2}^\alpha). \tag{7.11}$$

If during quasi-steady state growth, $X_{02} - X_{i2}^\alpha \ll X_{i2}^\alpha - X_{e2}^\alpha$, then the growth will be called *interface controlled*, since most the free energy driving force is spent on driving interface migration. On the other hand, if $X_{02} - X_{i2}^\alpha \gg X_{i2}^\alpha - X_{e2}^\alpha$, then it is called *diffusion controlled* since most the free energy driving force is spent on driving long-range diffusion. If $X_{02} - X_{i2}^\alpha$ is comparable to $X_{i2}^\alpha - X_{e2}^\alpha$, the growth will be called under *mixed* control.

Mass conservation of type-2 atoms *inside* the α phase is expressed as:

$$\partial_t c_2 \;=\; \partial_x(\tilde{D}\partial_x c_2) \tag{7.12}$$

in the lab frame. The flux at the interface coming from α is $J_2^{\mathrm{L}} = -\tilde{D}\partial_x c_2(x_i)$, so $\tilde{D}\partial_x c_2(x_i)dt$ type-2 atoms are arriving from the right per unit area in dt. With a moving interface of velocity v, on the other hand, $\Omega^{-1}(1 - X_{i2}^{\alpha})(vdt)$ type-2 atoms will be needed to build up the β phase from α. So one must have:

$$(v/\Omega)(1 - X_{i2}^{\alpha}) \;=\; \tilde{D}\partial_x c_2(x_i). \tag{7.13}$$

Such boundary condition relating flux at boundary to moving boundary velocity is called the *Stefan problem*. Type-1 atom also need to escape (since phase β is "intolerant" of type-1 atoms) via bulk flux in α. So generally we will have the Stefan boundary condition

$$v(\mathbf{c}_e^{\beta} - \mathbf{c}_i^{\alpha}) \;=\; \tilde{D}\partial_x \mathbf{c}^{\alpha}. \tag{7.14}$$

Let us first consider diffusion-controlled growth: $\mathbf{c}_i^{\alpha} \approx \mathbf{c}_e^{\alpha}$, in which case

$$v \;=\; \frac{\tilde{D}\partial_x \mathbf{c}^{\alpha}}{\mathbf{c}_e^{\beta} - \mathbf{c}_e^{\alpha}} \tag{7.15}$$

the denominator is given by bulk phase diagram. Imagine a planar front of β has just been nucleated inside α at $t = 0$. To estimate v from (7.15), we need an order-of-magnitude estimate for $\partial_x \mathbf{c}^{\alpha}$. Since it takes some time for the information "here is a precipitate plate" to propagate outward by diffusion, we may infer that

$$\partial_x \mathbf{c}^{\alpha} \;\propto\; \frac{\mathbf{c}_0 - \mathbf{c}_e^{\alpha}}{l_D} = \frac{\Delta\mathbf{c}}{l_D}, \quad l_D \equiv \sqrt{6\tilde{D}t}, \tag{7.16}$$

where the numerator is supersaturation, and denominator is the diffusion length. Plugging the above into (7.15), we see that

$$v \;\propto\; \sqrt{\frac{\tilde{D}}{t}}\frac{\Delta\mathbf{c}}{\mathbf{c}_e^{\beta} - \mathbf{c}_e^{\alpha}}. \tag{7.17}$$

Several comments can be made regarding this moving planar front: (1) the growth speed is not constant, it slows down as $t^{-1/2}$. (2) the thickness of the precipitate $l_P = \int v dt \propto \sqrt{\tilde{D}t} \propto l_D$, so the size of the precipitate (type-2 atoms rich) scales with the size of the diffusion-affected *depleted zone* in the matrix (type-2 atoms poor), which makes sense from a mass-conservation point of view. (3) We know the supersaturation $\Delta\mathbf{c}$ is proportional to ΔT, thus the larger ΔT the larger the thermodynamic driving force for diffusion; however,

the kinetics of diffusion is slowing down with larger ΔT, so the growth speed $v(\Delta T)$ actually peaks at some intermediate ΔT. This is similar to the nucleation rate $\mathcal{N}(\Delta T)$, except $\mathcal{N}(\Delta T)$ has a sharp threshold, while $v(\Delta T)$ is initially linear in ΔT.

The above discussion would hold true to $t = \infty$ if there is only one precipitate nucleated in an infinite matrix. In reality, multiple precipitates can be nucleated at different places, and they would grow pretty much independently *until* their l_D's, their "spheres of influence" the depleted zones, start to overlap (Fig. 5.17 of [47]). Thus commences coarsening, a global war between different precipitate for the possession of type-2 atoms. It is usually a fight to the death, unless $\tilde{D}$ is shut off by external control of the temperature.

The above may be applied to marching of the broad face of a precipitate plate. Let us now analyze marching of the edge of a precipitate plate. The edge may be approximated by a semi-circle with radius of curvature r (Fig. 5.19 of [47]). This growth then differs thermodynamically from growth of the the broad face which has $r = \infty$, in that the growth will be working *against* the Young-Laplace pressure $\Delta P = \gamma\Omega/r$. There is less *thermodynamic* incentive for type-2 atom to join the edge compared to the broad face. On the other hand, this is compensated by the larger "view angle": the broad face can only pick up solutes from the front, while a promontory-like edge can pick up solutes from any angle within π. So edge growth is *diffusion kinetically facilitated*. The edge also tends to be incoherent interfaces, which facilitates atom attachment. Such thermodynamic/kinetic balance will eventually determine the optimal r that gives the maximum edge growth speed.

Assuming the growth is still diffusion controlled (plenty of atom attachments/detachments at the $\alpha\beta$ interface), the equilibrium solubility of type-2 atom in α just outside the edge is enhanced by:

$$\tilde{X}^{\alpha}_{e2} \;=\; X^{\alpha}_{e2}\exp(\frac{\gamma\Omega}{rk_{\rm B}T}) \;\approx\; X^{\alpha}_{e2}(1+\frac{\gamma\Omega}{rk_{\rm B}T}) \tag{7.18}$$

according to the Gibbs-Thomson effect. Thus, the *effective* supersaturation that drives diffusion in α is reduced:

$$\tilde{\Delta}X_2 \;\equiv\; X_{02}-\tilde{X}^{\alpha}_{e2} = X_{02}-X^{\alpha}_{e2}(1+\frac{\gamma\Omega}{rk_{\rm B}T}) = (X_{02}-X^{\alpha}_{e2})(1-\frac{X^{\alpha}_{e2}}{X_{02}-X^{\alpha}_{e2}}\frac{\gamma\Omega}{rk_{\rm B}T}). \tag{7.19}$$

If we define

$$r^* \;\equiv\; \frac{X^{\alpha}_{e2}}{X_{02}-X^{\alpha}_{e2}}\frac{\gamma\Omega}{k_{\rm B}T} \;=\; \frac{\gamma\Omega}{\mu^{\alpha}_2(\mathbf{X}_0)-\mu^{\alpha}_2(\mathbf{X}^{\alpha}_e)} \tag{7.20}$$

the above can be simplified to

$$\tilde{\Delta}X_2 \;=\; \Delta X_2(1-\frac{r^*}{r}). \tag{7.21}$$

In fact, r^* is exactly the *critical nucleus size* of nucleating a cylinder in supersaturated α, when in $N^\beta \to N^\beta+1$ the reduction in bulk free energy is exactly balanced by the increase in chemical potential due to Young-Laplace pressure. Diffusion would stop if $r=r^*$, $\tilde{\Delta}X_2=0$. Indeed, if $r<r^*$, the edge would not grow but *retreat*, and *emit* type-2 atoms rather than absorbing type-2 atoms.

To estimate the steady-state edge growth speed, we convert (7.15) to 2D radial coordinate,

$$v \;=\; \frac{\tilde{D}\partial_r \mathbf{c}^\alpha}{\mathbf{c}_e^\beta - \mathbf{c}_e^\alpha} \tag{7.22}$$

and assume

$$\partial_r \mathbf{c}^\alpha \;\propto\; \frac{\tilde{\Delta}c}{r} \;=\; \frac{\tilde{\Delta}X_2}{r\Omega} \;=\; \frac{\Delta X_2(1-\frac{r^*}{r})}{r\Omega} \tag{7.23}$$

since in radial diffusion, there is now an intrinsic lengthscale r. $\mathbf{c}_e^\alpha$ in the denominator should really be $\tilde{\mathbf{c}}_e^\alpha$, but since the difference between $\mathbf{c}_e^\beta$ and $\mathbf{c}_e^\alpha$ is much larger than the difference between $\mathbf{c}_0$ and $\mathbf{c}_e^\alpha$, this distinction is less important than in the numerator and can be ignored. So we see that:

$$v \;\propto\; \frac{\tilde{D}\Delta c_2}{c_{e2}^\beta - c_{e2}^\alpha} \times \frac{1}{r}(1-\frac{r^*}{r}) \tag{7.24}$$

which is maximized when $r=2r^*$. We note that $v_{\max} \propto \frac{\tilde{D}\Delta c_2}{4(c_{e2}^\beta - c_{e2}^\alpha)r^*}$, which is proportional to $(\Delta T)^2$ for small ΔT. At large ΔT, the temperature dependence in $\tilde{D}$ kicks in, which means $v_{\max}$ will have a maximum at some intermediate ΔT.

We note that $v_{\max}$ is a constant of time. So the edge would cut into α with constant speed. The broad face would grow after that, with slower and slower speed. Like in evolution, the fastest moving edge has the advantage in survival because it can imbibe the full supersaturation in α. The slower moving edges with $r<2r^*$ or $r>2r^*$ would gradually be "snuffed out" due to the broad-face growth of the faster moving plates, which drain away the supersaturation and make the slower moving precipitate edges move even slower. So Nature selects an optimal lengthscale, an optimal tip/edge curvature in this case. If r is too small, there is not enough thermodynamic driving force for diffusion; if r is too big, the solutes have to diffuse too long from the matrix, so kinetically the growth becomes too slow. Such

thermodynamic/kinetic balance is quite generic. The same argument not only selects the edge curvature of precipitate plates in solid-solid transformation, but also the tip curvature of dendrite arms in solidification, and the lamellae thickness in eutectic reaction.

Let us now consider the opposite situation of interface-controlled growth. During such growth, the supersaturation holds all the way up to the interface (spatially uniform in α but temporally decreasing): $X_2^\alpha(x) = X_{i2}^\alpha$, and the growth speed is simply:

$$v \;=\; M\frac{k_\mathrm{B}T}{\Omega X_{e2}^\alpha}(X_{i2}^\alpha - X_{e2}^\alpha). \tag{7.25}$$

One might wonder how could long-range diffusion be easy, when type-2 atoms trekked *many* lattice spacings to get to the interface, but short-range diffusion (interfacial migration) is hard, where the type-2 atoms just need to make *one or several* atomic jumps to get to the other side. The answer is that there might be structural difficulties like the one shown at FCC/HCP interface in Fig. 3.68 of [47]. Also, there might be *interfacial chemistry* at play, different from bulk solution chemistry. Rare solute elements (may not be type-2) of ppm-level bulk concentration could segregate to the interface, and *significantly* decrease interfacial mobility. For example, just 0.006 wt% of tin is able to reduce the mobility of random GBs by factor of 10^4! (chapter 6). These segregated elements may exert strong barrier/trapping forces on type-2 atoms, that the type-2 atoms just find it very difficult to go across the interface, despite being so close to the target phase. (In human history there is no shortage of soldiers who trekked thousands of miles just to die in front of a wall.) The last reason for interface-controlled growth is that some phase transformations *do not require* long-range diffusion. For example, in the so-called *massive transformation*, α and β have the same composition but different structures. No solute partitioning and long-range diffusion is needed. All that is needed in massive transformation is for atoms to jump across the interface, i.e. short-range diffusion. In that case, the growth speed v would be approximately a constant.

We are now ready to study the well-known Johnson-Mehl-Avrami-Kolmogorov equation [68, 69, 70, 71], which combines nucleation and growth rates to give the total volume fraction of transformed β as a function of time. There are two simplifying assumptions in the Avrami equation: (a) the nucleation rate $\mathcal{N}$ is constant in untransformed α, and (b) the growth speed v is constant. As we have seen from previous models of $\mathcal{N}$ and v, in binary systems $\mathcal{N}$,v depend on the local supersaturation and the geometry, and may or may not be constant temporally or spatially. (a) and (b) are probably more appropriate for mas-

sive transformations such as liquid↔solid, solid→solid transitions of elemental materials, recrystallization, etc. Because the assumptions (a) and (b) are *simple* and lead to *analytical* solutions, the Avrami equation is a good starting model.

Even with the simplifying assumptions (a) and (b), there are many confusing or mathematically incorrect derivations of the Avrami equation in textbooks. A correct derivation is given by John Cahn, in the so-called time-cone approach. Consider as starting point a 1D system: $d = 1$. A long nanowire that undergoes phase change would be a good example [72]. Initially, the wire is completely α. Assumption (a) states that if $(x, x + dx)$ is α at time t, there is probability $dP = \mathcal{N} dx dt$ that $(x, x + dx)$ will contain one super-critical β nucleus at time $t + dt$, which can grow from that point on. Implicit in (a) is the assumption that the critical nucleus size r^* is so small compared to the wire length, that it can be practically regarded as 0. We just declare some previously-α *point* on x to be β at certain time, the probability of this declaration is proportional to the *space-time* volume $dxdt$. Assumption (b) is about growth: it says that once a point has been declared β, it will encroach on surrounding points and convert them into β, with spreading velocity v. Such conversion will stop if and only if the surrounding point is already β, when the time cones of two β nuclei meet. Lastly, assumption (a) states that once a point has been declared β, it will stay β forever, with no new nucleation probability associated with it. These assumptions are illustrated graphically as overlapping time cones in Fig. 21.1 of [41].

Aided by the graph, it is easy to prove that both the necessary and sufficient condition for a point (x, t) to stay α is that all points in the *reverse time cone* has *refrained* from nucleation. The probability of this refrainment can be calculated by sequential interrogation of different time slices: $t = (0, \Delta t)$, $(\Delta t, 2\Delta t)$, ..., $(t - \Delta t, t)$. In the first time slice $t = (0, \Delta t)$, the probability of nucleation at different spatial points are uncorrelated. The probability that a particular space-time volume element $\Delta t \times \Delta x$ of the first time slice refrains from nucleation is $1 - \mathcal{N}\Delta t \Delta x \approx \exp(-\mathcal{N}\Delta t \Delta x)$, so the total probability that no nucleation occurs within the first time slice of the reverse time cone is

$$\exp(-\mathcal{N}\Delta t \Delta x) \exp(-\mathcal{N}\Delta t \Delta x) ... \exp(-\mathcal{N}\Delta t \Delta x) \;=\; \exp(-\mathcal{N} w(t = 0) \Delta t) \qquad (7.26)$$

where $w(t = 0)$ is width of the reverse time cone at $t = 0$. If there is no nucleation in the first time slice[1], we may ask what is the probability that there is also no nucleation in the second time slice. The answer is $\exp(-\mathcal{N} w(t = \Delta t)\Delta t)$, so the probability of no nucleation

[1] If there is nucleation in the first time slice, the proper followup question becomes more complicated because one cannot nucleate again in the transformed region; fortunately we are not asking that question.

in the first two time slices is just $\exp(-\mathcal{N}w(t=0)\Delta t)\exp(-\mathcal{N}w(t=\Delta t)\Delta t)$. We keep asking these "no nucleation" questions sequentially: if any slice says "yes nucleation", it is game over for (x,t) to remain as α. The probability that the entire reverse time cone has refrained from nucleation is therefore

$$P_\alpha(x,t) \;=\; \exp(-\mathcal{N}(w(t=0)+w(t=\Delta t)+...+w(t-\Delta t))\Delta t) = \exp(-\mathcal{N}\mathcal{V}), \qquad (7.27)$$

where $\mathcal{V}$ is the total space-time volume of the reverse time cone. In 1D, $\mathcal{V}=\int_0^t (2vt')dt' = vt^2$. In 2D, we have circles whose radius grows as vt, so $\mathcal{V}=\int_0^t \pi(vt')^2dt' = \pi v^2t^3/3$. In 3D, we have spheres whose radius grows as vt, so $\mathcal{V}=\int_0^t 4\pi(vt')^3/3dt' = \pi v^3t^4/3$. Thus in 3D, we expect the *volume fraction* of α to decay with time as

$$f^\alpha \;=\; P_\alpha(x,t) \;=\; \exp(-\pi\mathcal{N}v^3t^4/3), \qquad (7.28)$$

and the volume fraction of β would increase with time as

$$f^\beta \;=\; 1-\exp(-\pi\mathcal{N}v^3t^4/3). \qquad (7.29)$$

In short times, f^β grows as $\pi\mathcal{N}v^3t^4/3$. In long times, f^β approaches 1. The shape of $f^\beta(t)$ looks like a sigmoidal function.

Note that the time origin of (7.29) should be right after the incubation time $t_{\rm inc}$, when $C(N^\beta)$ cluster size distribution has developed into $\tilde{C}(N^\beta,t)$ distribution that has significant quasi-steady state value at r^*, and the nucleation rate is abruptly reaching quasi-steady-state value. If one uses $-\Delta T\to\Delta T$ quench as the time origin, there should be a time shift

$$f^\beta \;=\; \left[1-\exp(-\pi\mathcal{N}v^3(t-t_{\rm inc})^4/3)\right]H(t-t_{\rm inc}). \qquad (7.30)$$

where $H()$ is the Heaviside step function.

A time-temperature-transformation (TTT) diagram is a contour plot of f^β in time-temperature space. Typically, two contour lines are drawn, $f^\beta=0.01$ and $f^\beta=0.99$. Using the Avrami equation, the TTT contours can be drawn:

$$t_{0.01} \;=\; \left(\frac{3\ln 0.99}{-\pi\mathcal{N}v^3}\right)^{1/4}, \;\; t_{0.99} \;=\; \left(\frac{3\ln 0.01}{-\pi\mathcal{N}v^3}\right)^{1/4} \qquad (7.31)$$

If one draws the contour plot in $\ln(t)$-T space, the two contour lines would shift by a constant

horizontal amount at different temperatures.

The shape of the TTT contours typically looks like a left-pointing nose. This is because both $\mathcal{N}$ and v peak at intermediate ΔT's, so the product $\mathcal{N}v^3$ also peaks at some intermediate ΔT. At small ΔT, ΔG^* is so large that nucleation takes a long time. At large ΔT, diffusion become sluggish.

The TTT diagram aids design of heat treatment. If one wants to form β as a strengthening phase, one could take hold at ΔT near the nose, to minimize the time of treatment (long holding times tie down capital equipment, and increase energy cost). Vice versa, if one want to rapidly quench a liquid to form a glass, one would want to avoid the crystallization - one could then design the minimal quench rate dT/dt on the TTT diagram, so one could avoid the nose.

There are actually two kinds of TTT diagrams. One is for isothermal heat treatment, where the temperature is held fixed during transformation. The other is for continuous cooling, where $T(t)$ is a continuous curve (typically a straight line). The Avrami TTT curves (7.31) are for isothermal treatment. However, the isothermal TTT diagram and continuous cooling TTT diagram share some common features. For order-of-magnitude estimates, the two may be used interchangeably.

In the Johnson-Mehl-Avrami model, one did not specify what happens *after* two β time-cones meet, except saying that "once β, always β". However, as we have seen before, there can be multiple β *variants*. For example, if ordered intermetallic compound β precipitates out from disordered random A-B solid solution α, the ordering may be ABABAB (β_1) or BABABA (β_2) on the site lattice. β_1 and β_2 are clearly degenerate in energy. The interface between β_1 and β_2: ABABAB|BABABA, is called the anti phase boundary (APB). Just like grain coarsening, where *some* orientation variant grains grow in size whereas others disappear, different β phase variants may also compete with, and later literally "eat" each other, even though different variants have the same bulk free energy G_{soln} *a priori*.

The Johnson-Mehl-Avrami model does not have enough physical ingredients to resolve realistic coarsening processes in multi-component multi-phase alloys. This is because the β nuclei in Johnson-Mehl-Avrami does not carry depletion zone with it in the matrix, which is a very important feature that also give rise to non-constant velocity v. Indeed, in a two-phase precipitation reaction $\alpha \to \beta + \alpha'$, v will approach zero and f^β will approach f_e^β when the supersaturation is spent, and there will be finite volume fraction of α' left in the end. The Avrami equation (7.29) really is only physically appropriate for $\alpha \to \beta$, where α can

be transformed to β 100% with constant v, such as massive transformations and recrystallization. Nonetheless, a vertically rescaled Avrami equation $f^\beta = f_e^\beta(1-\exp(-kt^n))$ may be still a reasonable *fitting form* to fit the degree of completion of many transformations, including technological transformations [73] like the replacement of vinyl records by CDs in the marketplace, or the spreading of an infectious disease. To extract the Avrami exponent n, one may plot $\ln(-\ln(1-f^\beta(t)/f^\beta(\infty)))$ versus $\ln t$.

Let us consider a two-phase alloy, where the total β volume fraction has already approached a constant, f_e^β. That is to say, there is on average not much more supersaturation to be had in the matrix, and type-2 atoms must come mainly from other β-precipitates. If we label the particles by i, there is

$$\sum_i \frac{4\pi r_i^3}{3} \;=\; \text{const}, \quad \sum_i r_i^2 \dot{r}_i = 0. \tag{7.32}$$

The mean-field approximation assumes each particle of radius r_i sees the same environment, which means $\dot{r}_i$ will be a deterministic function of r_i only. If we assume the background has uniform concentration $\bar{c}_2$ (note $\bar{c}_2$ is different from the initial supersaturation c_{02}, since now we are at a late coarsening stage), and we insert a spherical particle inside, we can use the previous result of (7.24) for diffusion-controlled growth. The only difference is the critical radius is now

$$r^* \;\equiv\; \frac{c_{e2}^\alpha}{\bar{c}_2 - c_{e2}^\alpha}\frac{2\gamma\Omega}{k_{\rm B}T} \;=\; \frac{2\gamma\Omega}{\mu_2^\alpha(\bar{c}_2) - \mu_2^\alpha(c_{e2}^\alpha)} \tag{7.33}$$

instead of (7.20), since there are two principal radii of curvature for a spherical particle, instead of one for a cylindrical edge. So from (7.24) we have

$$\dot{r}_i \;=\; \frac{a\tilde{D}(\bar{c}_2 - c_{e2}^\alpha)}{(c_{e2}^\beta - c_{e2}^\alpha)} \times \frac{1}{r}(1-\frac{r^*}{r}) = \frac{a\tilde{D}2\gamma\Omega c_{e2}^\alpha}{(c_{e2}^\beta - c_{e2}^\alpha)k_{\rm B}Tr^*} \times \frac{1}{r}(1-\frac{r^*}{r}) \tag{7.34}$$

where a is a dimensionless constant. Thus $\sum_i r_i^2\dot{r}_i = 0$ requires

$$\sum_i (r_i - r^*) \;=\; 0 \quad \to \quad r^* \;=\; \frac{\sum_i r_i}{\sum_i 1} \tag{7.35}$$

so r^* can be interpreted as the average particle radius. For $r_i < r^*$, the particle will shrink. For $r_i > r^*$, the particle will grow. We may rewrite it as:

$$\dot{r}_i \;=\; \frac{b\tilde{D}}{r_i}(\frac{1}{r^*} - \frac{1}{r_i}) \tag{7.36}$$

where $b \equiv \frac{2a\gamma\Omega c_{e2}^{\alpha}}{(c_{e2}^{\beta}-c_{e2}^{\alpha})k_{\rm B}T}$ is a constant lengthscale.

We can model the particle size distribution:

$$dC \;\equiv\; f(r,t)dr \tag{7.37}$$

to be the concentration (#particles/m^3) of β particles with radius between r and $r+dr$. Using the same argument as counting "red Ferraris", but now in r-distribution space instead of x-real space, one gets

$$\partial_t f \;=\; -\partial_r(f(r,t)\dot{r}) \;=\; -\partial_r\left(\frac{b\tilde{D}}{r}(\frac{1}{r^*}-\frac{1}{r})f(r,t)\right) \tag{7.38}$$

While we can initialize the particle size distribution $f(r,t=0)$ any way we like, over long time the distribution will approach a self-similar attractor distribution (Fig. 15.5 and 15.6 of [41]) of the form:

$$f(r,t) \;\;\rightarrow\;\; \frac{1}{r^{*4}(t)}g\left(\frac{r}{r^*(t)}\right), \tag{7.39}$$

like in the Boltzmann transform and self-similar solutions of the diffusion equation. The reason for the prefactor is because of the normalization condition

$$f_e^{\beta} \;=\; \int_0^{\infty} dr f(r,t)\frac{4\pi r^3}{3} \;=\; \text{const}, \tag{7.40}$$

assumed for the coarsening stage.

If we define $\tilde{r} \equiv r/r^*(t)$, then the $(\tilde{r},t) \leftrightarrow (r,t)$ mapping goes as $F(\tilde{r},t) = F(r/r^*(t),t)$, so:

$$\partial_r F \;=\; r^{*-1}\partial_{\tilde{r}}F, \quad \partial_t F \;=\; D_t F + (\partial_{\tilde{r}}F)(-\tilde{r}\dot{r}^*/r^*), \tag{7.41}$$

where

$$\partial_t F \;\equiv\; \left.\frac{\partial F}{\partial t}\right|_r, \;\; D_t F \;\equiv\; \left.\frac{\partial F}{\partial t}\right|_{\tilde{r}}. \tag{7.42}$$

So (7.38) becomes

$$-4r^{*-5}g(\tilde{r})\dot{r}^* - r^{*-5}\dot{r}^*\tilde{r}\partial_{\tilde{r}}g \;=\; -b\tilde{D}r^{*-1}\partial_{\tilde{r}}(r^{*-6}b\tilde{D}\tilde{r}^{-1}(1-\tilde{r}^{-1})g(\tilde{r})), \tag{7.43}$$

so

$$r^{*2}\dot{r}^* \;=\; b\tilde{D}\frac{\partial_{\tilde{r}}(\tilde{r}^{-1}(1-\tilde{r}^{-1})g(\tilde{r}))}{4g+\tilde{r}\partial_{\tilde{r}}g} \tag{7.44}$$

The left-hand side depends only on t, the right-hand side depends only on $\tilde{r}$, so both must be constant. We thus have:

$$r^{*3}(t) - r^{*3}(0) \;=\; kb\tilde{D}t. \tag{7.45}$$

This is called Lifshitz-Slyozov-Wagner (LSW) model [74, 75]. The shape of the self-similar particle size distribution is shown in Fig. 15.5 of [41]. It has an abrupt cutoff at $r = 1.5r^*$.

We see that the LSW coarsening rate is $\propto \gamma c_{e2}^{\alpha}\tilde{D}$. Thus three strategies can be used to reduce the rate of coarsening (p. 316 of [47]). (a) Low interfacial energy: Ni-based super-alloys γ (fcc random solid solution) / γ' (ordered Ni_3Al) are used in aircraft engines and power turbines, and need to be high-T creep-resistant. The γ/γ' interfaces are coherent with exceptionally low interfacial energies, 10-30 mJ/m^2. Furthermore, the misfit strain can be fine tuned to essentially zero by changing alloy composition, so the coherency loss during deformation - which increases γ - may be reduced. The creep-rupture life of such zero-misfit alloy can be increased by a factor of fifty compared to 0.2%-misfit alloy. (b) low solubility c_{e2}^{α} in the matrix: oxygen has low solubility in most metal phases, and would like to precipitate out as stoichiometric oxide. Oxide dispersion strengthened (ODS) steels take advantage of this to become creep-resistant. (c) low diffusivity: pure cementite (Fe_3C) in steel coarsen very quickly because C diffuse interstitially, which is much faster than substitutional diffusion that relies on vacancies. Adding additional alloying elements M that form more stable carbides of the form $Fe_xM_yC_z$ than cementite would have two effects: (i) it ties down C so c_{eC}^{α} is lower, (ii) the diffusion of these metallic alloying elements is substitutional, which is much slower than C diffusion, which slows down coarsening.

If the kinetics of coarsening is interface controlled, then

$$\dot{r}_i \;=\; M\frac{\mu_2^{\alpha}(\bar{c}_2) - \mu_2^{\alpha}(c_{e2}^{\alpha}) - \frac{2\gamma\Omega}{r_i}}{\Omega} \tag{7.46}$$

we can also define a critical radius

$$r^* \;=\; \frac{2\gamma\Omega}{\mu_2^{\alpha}(\bar{c}_2) - \mu_2^{\alpha}(c_{e2}^{\alpha})} \tag{7.47}$$

and

$$\dot{r}_i \;=\; 2\gamma M\left(\frac{1}{r^*} - \frac{1}{r_i}\right). \tag{7.48}$$

Thus $\sum_i r_i^2\dot{r}_i = 0$ requires

$$r^* \;=\; \frac{\sum_i r_i^2}{\sum_i r_i}. \tag{7.49}$$

Following similar procedure as LSW model, it can be shown that

$$r^{*2}(t) - r^{*2}(0) \;=\; k\gamma M t, \tag{7.50}$$

after the particle size distribution has fallen into a self-similar attractor. This parabolic kinetics of coarsening is no different from grain growth, which is also interface controlled. Thus, whether it is long-range diffusion controlled or short-range diffusion controlled will lead to difference in the coarsening exponent by 1.

Chapter 8

Solidification

Solidification means liquid→solid transformation. Much has already been discussed about solidification in the previous chapters, such as nucleation theory, so only new aspects are discussed here. First of all, solidification involves much larger heat of transformation ΔH than most solid→solid transformations, so heat conduction may need to be taken into account in solidification kinetics. This intuitively makes sense, since how long a bottle of water freezes fully into ice when you put it outside of your window in the winter obviously may depend on how fast the bottle, the ice and the water conducts heat, and how much latent heat is in the water→ice transformation. From Onsager relation, one can derive the heat transport equation:

$$\partial_t T \;=\; \alpha \nabla^2 T, \tag{8.1}$$

which looks the same as the mass transport equations, where α is the thermal diffusivity (unit m^2/s). Typically, $\alpha^{\mathrm{solid}} > \alpha^{\mathrm{liquid}}$, $D^{\mathrm{liquid}} \sim 10^{-4}\alpha^{\mathrm{liquid}}$, $D^{\mathrm{solid}} \sim 10^{-8}\alpha^{\mathrm{liquid}}$ at the melting point (p.285 of [76]). Note that α^{liquid} is somewhat lower than the corresponding solid, whereas liquid's mass diffusivity is *much higher* than the corresponding solid. For this reason, long-range and short-range mass diffusion are relatively facilitated in liquids compared to heat diffusion, so we may *not* regard solidification of liquids as an isothermal process in a system. In contrast, many solid→solid transformations (interface controlled or long-range mass diffusion controlled) can be practically considered to be isothermal due to the low heat of transformation and high thermal diffusivity in solids ($D^{\mathrm{solid}} \sim 10^{-8}\alpha^{\mathrm{liquid}}$, and interfacial mobility which is short-range mass diffusion may follow the same general trend as D^{solid}, so mass transport is even more of an issue in solid→solid transformations, such as interface-controlled grain growth).

It is instructive to consider how a planar solid-liquid interface may maintain its stability in a non-isothermal situation. Let us assume (a) it is pure liquid $\rightarrow$ pure solid, so no solute partitioning and long-range mass transfer are needed, and (b) atom attachment/detachment at the interface is facile, so the solidification is not interface controlled (imagine a market on the interface where there is no trading fee/tax, there is a lot of trading, and there is no cheating and no trading fraud - then prices would have reached their "fair" values given the constraints of the far-field supply and demand). Then the velocity of the interface is controlled by how quickly the latent heat can be conducted away by long-range heat diffusion, a Stefan problem very much like the precipitate growth problem, except α, T are used instead of $\tilde{D}$, c_2. Also, $T_i \equiv T(\mathbf{x}_i) = T_e$ at the interface, i.e. the interface follows an isotherm contour (if Young-Laplace pressure can be ignored for initial coarse lengthscale - see later when lengthscale is refined) since (b) means no under-cooling is needed to drive interfacial motion.

If the solid is connected to external heat sink, for example if the solid was originally nucleated heterogeneously on the mold wall and maintains connection to the mold, then heat will be conducted away from the interface via the solid. In that case, it can be shown that the planar interface should be stable. This is because if a bulge develops on the solid, the T contour line would move *further* from the sink at the bulge and closer to the hotter zone, which means less heat is conducted away into the solid, so the bulge will move slower than the rest of the interface until the bulge vanishes.

If on the other hand if heat is conducted away from the interface via the *liquid*, the planar interface will not be able to maintain morphological stability. This may happen if the solid is nucleated homogeneously or heterogeneously (around a floating oxide particle, for instance) away from the mold wall, or if a nuclei originally nucleated on the mold wall is swept into the liquid by convective current. In that case, if the interface develops a bulge, the T contour line would move closer to the heat sink, develops a larger heat current locally that conducts away the latent heat, which further accelerates the growth of this bulge. Eventually, a thermal dendrite will develop. After the primary dendrite arm grows long enough, the lateral surface of the dendrite arm may become unstable again, and offshoots secondary and tertiary dendrite arms. These dendrite arms typically follow crystallographic directions ($\langle 100 \rangle$ in cubic metals and $\langle 1\bar{1}00 \rangle$ in hcp metals) which impart it some additional interfacial mobility advantage.

Just like precipitate edge growth, there is an optimal radius of curvature r of the dendrite tip for maximizing the dendrite velocity. The rate heat is conducted away is $\propto r^{-1}$, since the

isotherm contours have this lengthscale. On the other hand, too narrow r gives large Young-Laplace pressure to grow against: the thermodynamic driving force thus scales as $1 - r^*/r$, where r^* is the critical nucleus for homogeneous nucleation. This thermodynamics-kinetics $v \propto r^{-1}(1 - r^*/r)$ trade-off leads to $r_{\text{optimum}} = 2r^*$ for the winning dendrite. The slower moving dendrites will be stopped by the secondary and tertiary arms of the faster moving dendrite.

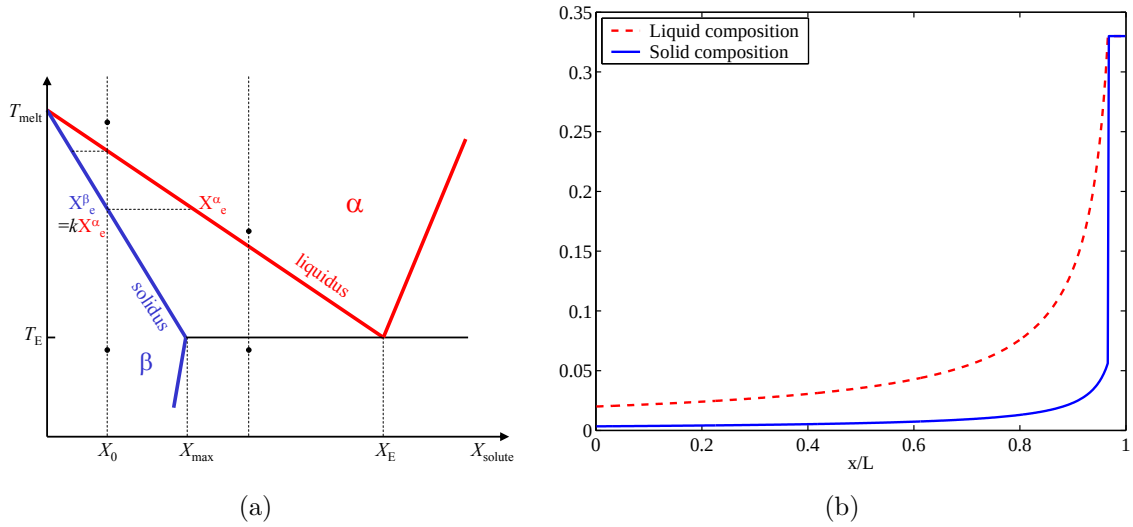


Figure 8.1: (a) A typical eutectic phase diagram. (b) Scheil-Gulliver solution of the concentration profile, when there is no diffusional mixing in solid, but complete mixing by convection in liquid.

Since $D^{\text{liquid}} \sim 10^{-4}\alpha^{\text{liquid}}$, when the liquid is not perfectly pure and contains just a small amount of solutes, the nature of the solidification kinetics may change from heat diffusion controlled to mass diffusion controlled. Consider a typical binary phase diagram with an eutectic point at T_{E}. The melting temperature of pure 1 is T_{melt}; the maximum solubility of 2 in the solid phase is X_{max} (from now on the scalar X symbol stands for X_2), and the eutectic liquid composition is X_{E}. For simplicity we assume that both the liquidus and the solidus are straight lines: $dX^{\alpha}_e/dT = X_{\text{E}}/(T_{\text{E}} - T_{\text{melt}})$, $dX^{\beta}_e/dT = X_{\text{max}}/(T_{\text{E}} - T_{\text{melt}})$, where α stands for liquid and β stands for solid.

The partition coefficient k is defined as $k \equiv X^{\beta}_e/X^{\alpha}_e = X_{\text{max}}/X_{\text{E}}$. If $k < 1$, then the solid phase "hates" solutes and wants to eject them into the liquid. An enriched liquid on the other hand can sustain a bit more cooling before decomposing again. In the discussions that follow we are going to assume that atom attachment/detachment at the solid/liquid interface

is easy, so interfacial mobility is not the controlling factor of interface speed. When this is the case, the solid and the liquid should reach thermodynamic equilibrium right at the interface, which is to say $X^\beta(x = x_i) = X_e^\beta(T_i)$, $X^\alpha(x = x_i) = X_e^\alpha(T_i)$, $\mu_2^\beta(x_i) = \mu_2^\alpha(x_i)$, namely the compositions at the interface follow some T_i cut of the equilibrium phase diagram where T_i is the temperature right at the interface (if the interface is planar or if the Gibbs-Thomson effect can be ignored).

We are going to study unidirectional solidification (directional solidification), which is an industrially important process to make single-crystal turbine blades, high-purity semiconductors etc. Consider a bar of liquid $x \in (0, L)$. A heat sink is placed at $x = 0$, so the solid phase will grow from the left $x < x_i$, with composition profile $X^\beta(x < x_i, t)$. The liquid phase is retreating on the right $x > x_i$ with composition profile $X^\alpha(x > x_i, t)$. $x_i(t = 0) = 0$, and the interfacial velocity is $v \equiv \dot{x}_i$. For simplicity, we will assume $\Omega_1^\alpha = \Omega_2^\alpha = \Omega_1^\beta = \Omega_2^\beta = \Omega$. The initial liquid composition (average composition of the alloy) is $X^\alpha(0 < x < L, t = 0) = X_0$. We also note that while the concentration profile sustains a tie-line jump at the interface, the temperature profile $T(x)$ should be continuous in value across the interface and takes value T_i at the interface.

Three limiting scenarios will be considered: (a) plenty of diffusional mixing in the solid (thus plenty of diffusional mixing in the liquid as well since $D^\alpha \sim 10^4 D^\beta$), (b) no diffusional mixing in the solid, but complete mixing *by convection* in the liquid, and (c) no diffusional mixing in the solid, but partial mixing *by diffusion* in the liquid.

In scenario (a), full diffusional equilibrium is achieved in both α and β due to the slow-moving interface and quasi-static cooling. Heat is removed so slowly from the system that the bar may be considered isothermal. In this case, both the solid phase and the liquid phase have plenty of time to diffuse and therefore take a uniform composition: $X^\beta(x < x_i) = X^\beta$, $X^\alpha(x > x_i) = X^\alpha$, and X^β and X^α must also be equilibrated across the interface, thus $X^\beta(x < x_i) = X_e^\beta(T_i)$, $X^\alpha(x > x_i) = X_e^\alpha(T_i)$, i.e. the uniform compositions are just those indicated by the phase diagram at mutual equilibrium. Since heat diffusivities are larger than the mass diffusivities, the heat diffusion lengths are longer than the mass diffusion lengths, therefore the temperature must be uniform as well, $T(x) = T_i$. In Fig. 8.1(a), the first solid that forms has composition kX_0, coming out at $T_i = T_{\text{melt}} + X_0(T_{\text{E}} - T_{\text{melt}})/X_{\text{E}}$. The solutes in the region now occupied by the solid get ejected into the liquid, making it richer. The richer liquid can cool down a bit further, before a new solid, also a bit richer, comes out and attaches to the interface. However, since we assume diffusion is "fast" in the solid (or at least given enough time to happen to completion) compared to motion of the interface,

this richer solid gets homogenized with the original kX_0 solid, and the entire uniform solid gets a bit richer. This process repeats, and the uniform liquid and solid compositions just "slide down" the liquidus and solidus lines in Fig. 8.1(a), until two possible eventualities. If $X_0 < X_{\max}$, the maximum solid solubility in β, then the last drop of liquid should disappear at $T_i = T_{\rm melt} + X_0(T_{\rm E} - T_{\rm melt})/X_{\max} > T_{\rm E}$, with uniform solid composition $X^\alpha = X_0$ just before the disappearance. If $X_0 > X_{\max}$, however, then the last drop of liquid disappears at $T_{\rm E}$. Just before $T_{\rm E}$, the uniform α region has composition $X_{\max}$, which occupies volume fraction $f^\beta = (X_{\rm E} - X_0)/(X_{\rm E} - X_{\max})$, and the uniform liquid region has composition $X_{\rm E}$ which occupies volume fraction $f^\alpha = (X_0 - X_{\max})/(X_{\rm E} - X_{\max})$. This f^α liquid region will turn into an eutectic solid (some γ/β precipitate-in-matrix microstructure) below $T_{\rm E}$ at the end of the bar.

Scenario (a) is the easiest to understand but rarely happens in practice because solid-state diffusion is slow, $D^\beta \sim 10^{-4} D^\alpha$, therefore it might take extremely long time to reach uniform composition in the solid part. In many cases, it is more appropriate to assume there is *no diffusional mixing* in the solid during solidification: the solid *keeps the composition that it first came out with*. Mixing in the liquid by diffusion is much easier, and can be further aided by convective mixing, such as vigorous stirring (what you do if you want to dissolve sugar in water). So one may consider the (b) scenario, which is no mixing in the solid, but full mixing in the liquid. Then we have: $X^\beta(x < x_i)$ and $X^\alpha(x > x_i) = X^\alpha$. Again we assume interfacial mobility is not an issue, so two sides of the flat interface reach thermodynamic equilibrium at the interfacial temperature T_i: $X^\beta(x = x_i) = X_e^\beta(T_i) = kX_e^\alpha(T_i)$, $X^\alpha = X_e^\alpha(T_i)$. Now imagine the interface moves by $dx_i = vdt$. The original $X^\alpha dx_i$ in the region will be replaced by the new $X_e^\beta(T_i)dx_i = kX^\alpha dx_i$, so solutes will be ejected into the liquid. Due to the fast mixing, the ejected solutes "instantaneously" gets dispersed everywhere in the entire fluid. Define solid volume fraction $f^\beta \equiv x_i/L$ and liquid volume fraction $f^\alpha = 1 - f^\beta$, we have $df^\beta = -df^\alpha = dx_i/L$. Mass conservation requires that

$$(kX^\alpha - X^\alpha)dx_i + (L - x_i)dX^\alpha \;=\; 0 \quad \leftrightarrow \quad (1-k)X^\alpha df^\alpha + f^\alpha dX^\alpha \;=\; 0 \tag{8.2}$$

The above is just a special limit of the Stefan problem. Then we have $(1-k)d\ln f^\alpha + d\ln X^\alpha = d\ln((f^\alpha)^{1-k}X^\alpha) = 0$, or $(f^\alpha)^{1-k}X^\alpha = \text{const}$. At time 0, the entire region is fluid, $f^\alpha(t=0) = 1$, $X^\alpha(t=0) = X_0$, so we get const $= X_0$ and

$$X^\alpha \;=\; X_0(f^\alpha)^{k-1}, \quad X^\beta(x_i) \;=\; kX_0(f^\alpha)^{k-1} \;=\; kX_0\left(1 - \frac{x_i}{L}\right)^{k-1}. \tag{8.3}$$

The above is called the Scheil-Gulliver equation. The solution is plotted in Fig. 8.1(b).

With $k < 1$, we see that both X^α and $X^\beta(x)$ are monotonically increasing function of x. Because there is no diffusional mixing in the solid, the solid composition at the interface is *richer* than in scenario (a) for the corresponding x_i. As a result, the liquid composition is also pushed higher. Indeed, the Scheil-Gulliver equation shows that the liquid composition would diverge as $f^\alpha \to 0$. In reality this does not happen because as soon X^α hits X_E at $T_i = T_\mathrm{E}$, the liquid cannot enrich further and must undergo eutectic decomposition $\alpha \to \beta + \gamma$. From the Scheil-Gulliver equation we see that $f^\alpha = (X_\mathrm{E}/X_0)^{1/(k-1)}$ portion of the bar at the end will have $\beta + \gamma$ eutectic microstructure. No matter what is $X_0 \in (0, X_\mathrm{E})$, the last drop of liquid in scenario (b) always solidifies at T_E with composition X_E.

From Fig. 8.1(b), we see that unidirectional solidification (heat sink at $x = 0$) drives large amount of solutes to the end of the bar. This is ultimately because the solid "hates" solutes ($k = 1$) and keep pushing them into the liquid. This leads to a general idea for *purification*. We could for instance cut out the last 10% of the bar in Fig. 8.1(b) after it solidifies. The remaining 90% of the bar will have an average composition $\bar{X}$ considerably less than X_0, say $\bar{X} = rX_0$, where $r < 1$ is some reduction factor. This shorter bar can then be *remelted* into a homogeneous liquid, and unidirectional solidified again. Since (8.3) is linear in X_0, all that is going to happen is that we replace X_0 by rX_0, and the process repeats itself self-similarly in the shorter bar. Then after n passes, the average composition will be just $r^n X_0$ (the bar is also shorter, $0.9^n L$). If say $r = 0.25$, just 5 passes will make the bar (still 59% of the original length) a thousand times purer on average than the original bar!

In the semiconductor industry very high-purity Si is needed as a base material before intentional doping by ion implantation or diffusion. The amount of undesired solutes may need to be restricted below ppm level or lower in the base material. *Zone refinement* is a process invented by W. G. Pfann [77] whereby a hot zone is repeatedly passed though a bar of materials, locally melting the bar and giving solutes in the solid an opportunity to go into the liquid, and pass them on and "sweeping" them towards the end of the bar. Even though different in details and mathematically more complicated [78], the general idea is the same as the Scheil-Gulliver equation. Zone refinement also gives a nearly exponential dependence of the average purity on the number of passes.

Finally, let us consider scenario (c), where diffusional mixing in the liquid is allowed, but there is no convective mixing. As we mentioned before, diffusion is a very effective means of mass transport at small lengthscales, but it gets progressively more sluggish at longer

lengthscales (in contrast, convective mixing does not suffer from this "longer-gets-lazier" shortcoming). Imagine that at the beginning, a small sliver of kX_0 solid come out, and there is excess solute $X^\alpha(x > x_i) = X_0 + \Delta X^\alpha(x - x_i, t)$ built up in the liquid in front of the interface. $\Delta X^\alpha(x - x_i, t)$, which is a decreasing function of x, may be called the "bow wave", if we think of x_i as the tip of a moving boat on a lake. When x_i is small, the spatial extent of the bow wave is small, so it is relatively easy for the solutes to get away from the solidification front by diffusion. But as time goes on, the spatial extent of the bow wave gets longer and longer, and diffusion gets progressively more sluggish. As a result, solutes build up in the bow wave, so the bow wave not only gets fatter in spatial extent but also larger in amplitude. Starting out from kX_0, $X^\beta(x_i)$ then gets larger and larger progressively as well, but it can't exceed X_0 which is the liquid feedstock composition.

In fact, because of the longer-gets-lazier property of diffusion inside the bow wave, it is possible to reach *steady state* in scenario (c), meaning the solid that comes out no longer change composition with x_i and time. Steady-state propagation is not feasible in scenarios (a) and (b), which have some kind of global mixing and therefore are always sample-size aware. Because diffusion is "short-sighted", it is possible to establish a steady-state propagation in (c) where the interfacial velocity $\dot{x}_i = v$ is a constant, and the local condition of the steady-state-propagating bow wave is L-independent.

In any steady-state propagation there must be $X^\beta(x_i) = X_0$, since what feeds into the bow wave must be what is left behind, composition wise. So there must be $X^\alpha(x_i) = k^{-1}X_0$ due to the interfacial equilibrium assumption. Thus during steady-state propagation, the interfacial temperature must be pinned at $T_i = T_{\text{melt}} + X_0(T_{\text{E}} - T_{\text{melt}})/X_{\max}$. The diffusion equation that governs the bow wave in liquid phase is

$$\partial_t X^\alpha \;=\; D^\alpha \partial_x^2 X^\alpha \tag{8.4}$$

where D^α is conceptually similar to interdiffusivity in solids. The Stefan boundary condition is

$$(k^{-1} - 1)X_0 v \;=\; -D^\alpha \partial_x X^\alpha(x_i). \tag{8.5}$$

The following exponential-decay form

$$X^\alpha(x) \;=\; (k^{-1} - 1)X_0 \exp\left(\frac{x_i - x}{w}\right) + X_0 \tag{8.6}$$

would satisfy $X^\alpha(x_i) = k^{-1}X_0$ and $X^\alpha(\infty) = X_0$ boundary conditions. Plugging it into the

PDE, we need

$$(k^{-1}-1)X_0\exp(\frac{x_i-x}{w})\frac{v}{w} \;=\; D^\alpha(k^{-1}-1)X_0\exp(\frac{x_i-x}{w})\frac{1}{w^2}, \tag{8.7}$$

or the characteristic bow wave width

$$w \;=\; \frac{D^\alpha}{v}. \tag{8.8}$$

Fortunately, this choice of w also satisfies the Stefan boundary condition:

$$(k^{-1}-1)X_0 v \;=\; -D^\alpha\cdot(k^{-1}-1)X_0\exp(\frac{x_i-x}{w})\frac{-1}{w}\Big|_{x=x_i}. \tag{8.9}$$

Thus indeed (8.6) is a valid steady-state solution.

In reality, v is controlled by the rate of heat removal. When the solidification is complete and one does a chemical analysis of the solid bar, in the initial $x\in(0,x_{\text{start}})$ section of the bar, where $x_{\text{start}}\propto w$, the solid composition is below X_0. In the final $x\in(x_{\text{finish}},L)$ section of the bar where $L-x_{\text{finish}}\propto w$, the solid composition will be higher than X_0 as the richer-composition bow wave "crashes" to the end of the liquid container. For $x\in(x_{\text{start}},x_{\text{finish}})$ the solid composition is approximately X_0.

If $X_0>X_{\max}\equiv kX_{\text{E}}$, the maximum solubility in the solid, then obviously a steady state cannot be established, since α cannot accept so much solutes. In this case, we could envision the bow wave amplitude build up, until the apex hits X_{E}, at which point an eutectic zone develops.

Up to now in (a),(b),(c) we have assumed the planar interface can be maintained in unidirectional solidification. Recall that in the solidification of pure liquids, as long as heat is conducted *away* from the liquid, $dT/dx>0$, the planar interface would be stable, because a bulge in the solid would get closer to hotter liquid zones and will soon be molten away. With the compositional degree of freedom added in and with composition-dependent melting temperatures, the problem can be more complicated. For (a) and (b), since the liquid composition is uniform, $dT/dx>0$ in the liquid would still guarantee stability of the planar interface. For (c), it turns out that dT/dx not only needs to be positive, it need to be greater than a critical value, dT_e/dx. Otherwise fingers can form: the solid finger/dendrite will attempt to break away from the rich liquid layer and chases after the cleaner X_0 liquid beyond thickness w.

The basic idea is that as one falls off the bow wave in the liquid, the solute concentration becomes less, and so the equilibrium temperature T_e gets higher (liquidus line in Fig. 8.1(a)). One can compute the rate of equilibrium temperature increase as

$$\begin{aligned} \frac{dT_e}{dx} &= \frac{dT_e}{dX^\alpha} \cdot \frac{dX^\alpha}{dx} = \frac{T_{\mathrm{E}} - T_{\mathrm{melt}}}{X_{\mathrm{E}}} \cdot (k^{-1} - 1) \exp(\frac{x_i - x}{w}) \frac{-X_0}{w}\Big|_{x=x_i} \\ &= \frac{X_0(T_{\mathrm{melt}} - T_{\mathrm{E}})(k^{-1} - 1)}{X_{\mathrm{E}} w}. \end{aligned} \tag{8.10}$$

Even if $dT/dx > 0$ and the real temperature is *increasing* as one goes into the liquid, if the real temperature does not increase *as fast* as T_e does, the real temperature in the liquid could still fall *below* the freezing temperature for *that local composition*. This is called *constitutional supercooling*. The term means that even though the temperature is apparently increasing as one goes deeper into the liquid, $dT/dx > 0$, still the liquid is being supercooled due to the negative *compositional gradient*. Constitutional supercooling in effect says that in a multi-component liquid, what is important is not the absolute magnitude of dT/dx, but the difference $d(T - T_e)/dx$, as one leaves the solid/liquid interface. It is like in a society with no inflation (pure system), a raising salary year by year means improved standard of living; but in a society with inflation (alloy, where the equilibrium temperature changes with composition and therefore position), one has to get a raise year by year that beats the inflation, in order to have real improved standard of living. When $d(T - T_e)/dx < 0$, one is "beat by inflation" and fingering instability and breakup of the planar interface may happen. One way to think about it is that new solids may now be able to grow independently in the liquid's diffusion layer, breaking up the planar interface. These dendrites may sharpen their tips to get a kinetic advantage, but eventually the Young-Laplace pressure becomes too large to grow against. The classic thermodynamics-kinetics $v \propto r^{-1}(1 - r^*/r)$ trade-off applies here as well, and the optimal dendrite tip radius would be $r_{\mathrm{optimum}} = 2r^*$.

The same argument may apply to diffusion-limited growth of planar solid-solid interfaces. Without a large enough dT/dx to stabilize the planar interface (very difficult to achieve in solids due to the large thermal diffusivity), the planar precipitate-matrix interface will spontaneously break up into fingers. There is also an optimal wavelength selection for such fingering, balancing capillary energy with growth kinetic advantage gained by the fingers. Such diffusional instability and fingering kinetics are generally called the Mullins-Sekerka instability [79, 80]. The basic reason for such instability is that the β phase needs to eject solutes, and the quicker the accumulated solutes can diffuse away the faster β can grow. Therefore the dendrites form and chase after low-concentration but high-supersaturation

(by low tempertaure) matrix regions, to dispense away the solutes accumulated at the tip.

Chapter 9

Point defects: Climb, Anelasticity, Strain aging

Point defects such as vacancies and interstitials affect macroscopic behavior of materials. This is very clear from radiation damage of materials[81], where high-energy radiation knocks atoms off their lattice, and create out-of-equilibrium concentrations of vacancies and other point defect and defect clusters, which in turn causes swelling and embrittlement. Below, we will outline some other prominent effects caused by the point defects.

Consider a single edge dislocation of Burgers vector $\mathbf{b} \perp \boldsymbol{\xi}$, which is embedded in a material cylinder of radius R. Imagine the half plane of the dislocation extend by Δh, which requires $\Delta N = b\Delta h/\Omega$ atoms to be attached to the dislocation core, translating the local fields (local dilation, shear, stres, energy density) associated with the dislocation by Δh. We are going to assume these atoms attached to the core to be plucked out the perfect lattice, thus creating ΔN vacancies in the lattice. For pedagogical purpose we will take the rigid-framework assumption for the crystalline site lattice:

$$V(N_{\mathrm{A}}, N_{\mathrm{V}}) \;=\; \frac{N_{\mathrm{A}} + N_{\mathrm{V}}}{N_{\mathrm{A}}} V(N_{\mathrm{A}}, 0), \tag{9.1}$$

which means the formation volume of vacancy is the same as the formation volume of an atom:

$$\Omega_{\mathrm{A}} \;=\; \Omega_{\mathrm{V}} \;=\; \Omega. \tag{9.2}$$

In this case, the cylinder would uniaxially expand with averaged strain

$$\boldsymbol{\epsilon} \;=\; \frac{\Delta N \Omega \hat{\mathbf{b}}\hat{\mathbf{b}}^T}{V_{\text{cylinder}}} \tag{9.3}$$

over the entire cylinder, due to the climb (if this is hard to visualize, imagine a rectanglar block instead of a cylinder). If there is far-field stress on the cylinder, it would do work

$$\Delta N \Omega \text{Tr}(\hat{\mathbf{b}}\hat{\mathbf{b}}^T\boldsymbol{\sigma}) \;=\; b\Delta h \text{Tr}(\hat{\mathbf{b}}\hat{\mathbf{b}}^T\boldsymbol{\sigma}) \tag{9.4}$$

which agrees with the climb component of the Peach-Koehler force (2.86). This mass action would also increasing the free energy by $\Delta N g_V^f$ (only vibrational entropy included) and the configurational entropy by $-(\Delta N)k_{\text{B}} \ln X_V$, assuming the ΔN vacancies are sufficiently far away from the dislocation that the vacancies do not feel the dislocation's own stress field (these "thermally emitted vacancies" have "fully escaped" from the dislocation). So when this process reaches equilibrium, we would have

$$\Omega \text{Tr}(\hat{\mathbf{b}}\hat{\mathbf{b}}^T\boldsymbol{\sigma}) \;=\; g_V^f + k_{\text{B}}T \ln X_V \tag{9.5}$$

and we get

$$X_V \;=\; X_V^0 \exp(\Omega \text{Tr}(\hat{\mathbf{b}}\hat{\mathbf{b}}^T\boldsymbol{\sigma})/k_{\text{B}}T) \tag{9.6}$$

where X_V^0 is our usual thermal equilibrium vacancy concentration (say, from the canonical surface vacancy source) without the dislocation and without the stress.

The above derivation describes a mass action, which can have a reciprocal effect, on the vacancy concentration, and on the dislocation. (Just like Newton's 3rd law: for each action, there is an equal and opposite reaction). The effect on the vacancy concentration is that near an edge dislocation that is biased by stress, the equilibrium vacancy concentration will be either elevated (if $\text{Tr}(\hat{\mathbf{b}}\hat{\mathbf{b}}^T\boldsymbol{\sigma})$ is tensile), or suppressed (if $\text{Tr}(\hat{\mathbf{b}}\hat{\mathbf{b}}^T\boldsymbol{\sigma})$ is compressive). The dislocation core offers a market to create/annihilate vacancies, just like the surface does. We note this elevation/suppression effect is independent of the sign of $\mathbf{b}$ (whether the half plane is pointing "up" or "down"), but not the sign of $\boldsymbol{\sigma}$, which makes sense. A tensile stress should create more "atomic-scale porosity", which are the vacancies.

The effect of the vacancy concentration on the dislocation is that there will be an additional osmotic force, in addition to the well-known Peach-Koehler force (2.86), if the vacancy concentration near the dislocation core (but still far enough that the vacancy/dislocation has

negligble elastic interactions - if truly $\Omega_{\rm V} = \Omega$ then there will not be any elastic interaction, that they can be considered separate entities) is different from the (9.6) value. So the force per unit length of dislocation is

$$\frac{d\mathbf{F}}{dl} \;=\; (\mathbf{b}\cdot\boldsymbol{\sigma})\times\boldsymbol{\xi} + \frac{k_{\rm B}T(\boldsymbol{\xi}\times\mathbf{b})}{\Omega}\ln\frac{X_V}{X_V^0}. \tag{9.7}$$

To see the above vector form, we note that strain×volume created by dislocation translation $\delta\mathbf{x}$ is $(dl\boldsymbol{\xi}\times\delta\mathbf{x})\mathbf{b}^T$ (a canonical edge dislocation has $\hat{\mathbf{b}} = \mathbf{e}_x$, the extra half plane in $+\mathbf{e}_y$ and $\boldsymbol{\xi} = \mathbf{e}_z$), and so the extra volume created is $\mathrm{Tr}((dl\boldsymbol{\xi}\times\delta\mathbf{x})\mathbf{b}^T) = \mathbf{b}\cdot(dl\boldsymbol{\xi}\times\delta\mathbf{x}) = \delta\mathbf{x}\cdot(\mathbf{b}\times dl\boldsymbol{\xi})$, and the number of vacancies emitted by the moving dislocation core is

$$\Delta N \;=\; \frac{\delta\mathbf{x}\cdot(\mathbf{b}\times dl\boldsymbol{\xi})}{\Omega}. \tag{9.8}$$

To appreciate how large the osmotic force is, we know that

$$\frac{k_{\rm B}T_{\rm room}}{\mathring{\rm A}^3} \;=\; 4.14{\rm GPa} \tag{9.9}$$

so with $\Omega = 11.8\mathring{\rm A}^3$ in Cu, we get

$$\frac{k_{\rm B}T_{\rm room}}{\Omega} \;=\; 350{\rm MPa} \tag{9.10}$$

so with $\frac{X_V}{X_V^0} = 2$, we will need the equivalence of 243 MPa of Peach-Koehler force to balance the osmotic force that would otherwise drive the edge dislocation "up" in y. So this is not a small effect.

Friction between two bodies is well known effect, but **internal friction** within a single solid body is a little bit less well known, which characterizes how much the solid deviates from perfect elastic body. Internal friction can be characterized by a torsion balance [82]. This instrumentation, and the more general machinery of Dynamic Mechanical Analysis (DMA) spectroscopy, studies small-stress dynamical behavior of materials. Unlike plasticity which is large-stress nonlinear behavior, the DMA spectroscopy is linear-response but focusing on frequency space characteristics. Consider the following partition:

$$\epsilon \;=\; \epsilon_{\rm e} + \epsilon_{\rm i} \tag{9.11}$$

where $\epsilon_{\rm i}$ is the inelastic strain, also called stress-free strain, transformation strain. To appre-

ciate this concept of transformation strain, consider the state of affairs of carbon interstitials in α-iron. Without carbon,α-iron would be perfectly cubic. With carbon, it is not necessarily so. The carbon can sit at edge centers or face centers, which are actually equivalent (octahedral site), so we only need consider edge center. Clearly, if the carbon is on [100]-bond, there will be uniaxial dilation in x. The only reason that ferrite (with carbon) is still cubic is because the three populations of carbon interstials have equal concentration:

$$c_{\mathrm{C}}^x \;=\; c_{\mathrm{C}}^y \;=\; c_{\mathrm{C}}^z \tag{9.12}$$

However, if somehow we can accomplish a population bias, then that configuration will no longer be cubic, and will have a transformation strain with respect to the cubic state. We can model this transformation strain as

$$\boldsymbol{\epsilon}_{\mathrm{i}} \;=\; a \begin{pmatrix} c_{\mathrm{C}}^x - \frac{c_{\mathrm{C}}^x + c_{\mathrm{C}}^y + c_{\mathrm{C}}^z}{3} & 0 & 0 \\ 0 & c_{\mathrm{C}}^y - \frac{c_{\mathrm{C}}^x + c_{\mathrm{C}}^y + c_{\mathrm{C}}^z}{3} & 0 \\ 0 & 0 & c_{\mathrm{C}}^z - \frac{c_{\mathrm{C}}^x + c_{\mathrm{C}}^y + c_{\mathrm{C}}^z}{3} \end{pmatrix} + b(c_{\mathrm{C}}^x + c_{\mathrm{C}}^y + c_{\mathrm{C}}^z)\mathbf{I} \tag{9.13}$$

the second term above is irrelevant in the present discussion, because we presume the total carbon concentration in DMA experiment is unchanged.

To develop a model for torsion balance, we note that the amount of torsion $\theta \propto \epsilon$, and to achieve apparent acceleration $\ddot{\theta}$ requires force $\propto m\ddot{\theta} \propto \sigma = G\epsilon_{\mathrm{e}}$, where G is the shear modulus, so the basic kinematic equation is

$$\ddot{\epsilon} \;=\; kG\epsilon_{\mathrm{e}} \;=\; \ddot{\epsilon}_{\mathrm{e}} + \ddot{\epsilon}_{\mathrm{i}} \tag{9.14}$$

where k depends on the geometry. In above, the only question is how $\ddot{\epsilon}_{\mathrm{i}}$ depends on $G\epsilon_{\mathrm{e}}$. If we assume that

$$\dot{\epsilon}_{\mathrm{i}} \;=\; \frac{G}{\nu}\epsilon_{\mathrm{e}} \tag{9.15}$$

where ν is an apparent "viscosity" that relates the inelastic strain rate with

Appendix A

Review of Bulk Thermodynamics

Equilibrium: given the constraints, the condition of the system that will eventually be approached if one waits long enough.

Example: gas-in-box. Box is the constraint (**volume**, heat: isothermal/**adiabatic**, permeable/**non-permeable**). One initialize the atoms any way one likes, for example all to the left half side, and suddenly remove the partition: BANG! one gets a non-equilibrium state. But after a while, everything settles down.

Atoms in solids, liquids or gases at equilibrium satisfy **Maxwellian velocity distribution**:

$$dP \;\propto\; \exp\left(-\frac{m(v_x-\bar{v}_x)^2}{2k_{\rm B}T}\right)dv_x, \quad \langle v_x^2\rangle \;=\; \frac{k_{\rm B}T}{m}. \tag{A.1}$$

$k_{\rm B} = 1.38\times10^{-23}$ J/K is the Boltzmann constant, it is the gas constant divided by 6.022×10^{23}. If I give you a material at equilibrium without telling you the temperature, you could use the above relation to measure the temperature.

But in high-energy Tokamak plasma, or dilute interstellar gas, the velocity distribution could be non-Gaussian, bimodal for example. Then T is ill-defined. Since entropy is conjugate variable to T, entropy is also ill-defined for such far-from-equilibrium states.

Equilibrium is however yet a bit more subtle: it is possible to reach equilibrium among a subset of the degrees of freedom (all atoms in a shot) or subsystem, while this subsystem is not in equilibrium with the rest of the system.

This is why engineering and material thermodynamics is useful for cars and airplanes. Imagine a car going 80 mph on highway: the car is not in equilibrium with the road, the axel is not in equilibrium with the body, the piston is not in equilibrium with the engine block. Yet, most often, we can define temperature (local temperature) for rubber in the tire, steel in the piston, hydrogen in the fuel tank, and apply equilibrium materials thermodynamics to analyze these components individually.

This is because of **separation of timescales**. The atoms in condensed phases collide much more frequently (10^{12}/second) than car components collide with each other. Thus, it is possible for atoms to reach equilibrium with adjacent atoms, before components reach equilibrium with each other.

Define "Type A non-equilibrium", or "local equilibrium": atoms reach equilibrium with each other within each **representative volume element (RVE)**; the RVE may not be in equilibrium with other RVEs.

For "Type A non-equilibrium", we can define local temperature: $T(\mathbf{x})$, and local entropy.

In this course, we will be mainly investigating "Type A non-equilibrium", and study how the RVEs reach equilibrium with each other across large distances compared to RVE size. Type B non-equilibrium, such as in Tokamak plasma, or radiation knockout in radiation damage, can be of interest, but is not the main focus of this course.

Consider a binary solid solution composed of two types of atoms, N_1, N_2 in absolute numbers (we prefer to use absolute number of atoms instead of moles in this class). **Helmholtz free energy** $F \equiv E - TS = F(T, V, N_1, N_2)$: $dF = dE - TdS - SdT$ is a complete differential. For closed system $dN_1 = dN_2 = 0$, the first law says $dE = \delta Q - PdV$, where PdV is work (coherent energy transfer) and δQ is heat (incoherent energy transfer via random noise).

For open system, $dE = \delta Q - PdV$ needs to be modified as

$$dE \;=\; \delta Q - PdV + \mu_1 dN_1 + \mu_2 dN_2 \tag{A.2}$$

μ_1, μ_2 are the **chemical potentials** of type-1 and type-2 atoms, respectively. To motivate the additional terms $\mu_1 dN_1 + \mu_2 dN_2$ for open systems, consider a process of atom attachment at $P = 0$, $T = 0$. And for simplicity assume for a moment $N_2 = 0$ (just type-1 atoms). In this case, before and after attaching an additional atom, kinetic energies K are zero. $E = U + K = U(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{3N_1})$. $U(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{3N_1})$ is called the **interatomic potential**

function, a function of $3N_1$ arguments. For some materials, such as rare-gas solids, it is a good approximation to expand $U(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{3N_1}) \approx \sum_{i<j} u_{ij}(|\mathbf{x}_j - \mathbf{x}_i|)$, where i, j label the atoms and run from $1..N_1$, and $u_{ij}(r)$ is called the pair potential (energy=0 reference state is an isolated atom infinitely far away). Clearly then, E will change, since there is one more atom in the sum, within interaction range from the previous set of atoms. Since $P = 0$, $PdV = 0$. In order to maintain $T = 0$, $\delta Q = 0$. To do this there must be an "intelligent magic hand" to drag on the atom to have a "soft landing". The energy input by the "intelligent magic hand" is coherent energy transfer, $\delta Q = 0$ (if not convinced, consider a layer of atoms adding on top of solid by a "forklift" - the added layer will move like a piston - no heat is needed). Also, the "intelligent magic hand" or "forklift" accomplishes so-called "mass action" (addition or removal of atoms), and is different from traditional PdV work, which describes a process of changing volume *without* changing the number of atoms. And thus μ_1 is motivated. In fact, from this microscopic idea experiment we have derived $\mu_1(T = 0, P = 0) = \sum_j u_{ij}(|\mathbf{x}_j - \mathbf{x}_i|)/2$ when $\mathbf{x}_j$ runs over lattice sites.

A well-known pair potential is the Lennard-Jones potential:

$$u_{ij}(r) \;=\; 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right], \tag{A.3}$$

which achieves minimum potential energy $-\epsilon_{ij}$ when $r = 2^{1/6}\sigma_{ij} = 1.122\sigma_{ij}$. For an atom inside a perfect crystal lattice, its number of nearest neighbors (aka coordination number) is denoted by Z. For instance, in BCC lattice $Z = 8$, in FCC lattice $Z = 12$. To further simplify the discussion, we can assume the pair interaction occurs only between nearest-neighbor atoms, and the Lennard-Jones potential is approximated by expansion $u_{ij}(r) = -\epsilon_{ij} + k_{ij}(r - 2^{1/6}\sigma_{ij})^2/2$ (perform a Taylor expansion on Lennard-Jones potential and truncate at $u = 0$).

The simplest model for a crystal is a simple cubic crystal with nearest neighbor springs $u_{ij}(r) = -\epsilon_{ij} + k_{ij}(r - a_0)^2/2$ (Kossel crystal), where a_0 is the lattice constant of this simple cubic crystal. With Z nearest neighbors ($Z = 4$ in 2D and 6 in 3D), $\mu(T = 0, P = 0) = -Z\epsilon/2$.

From dimensional argument, we see μ is some kind of energy per atom, thus on the order of minus a few eV (eV=1.602×10^{-19}J), in reference to isolated atom. To compare, at room temperature, thermal fluctuation on average gives $k_{\mathrm{B}}T_{\mathrm{room}} = 4.14 \times 10^{-21}\mathrm{J} \approx 0.0259$ eV $=$ eV/40 per degree of freedom.

Second law says $TdS = \delta Q$ when comparing two adjacent equilibrium states (integral form is $S_2 - S_1 = \int_{\text{any quasi-static path connecting 1-2}} \delta Q/T$). Thus

$$dF(T, V, N_1, N_2) \;=\; -PdV - SdT + \mu_1 dN_1 + \mu_2 dN_2 \tag{A.4}$$

We thus have:

$$P \;=\; -\left.\frac{\partial F}{\partial V}\right|_{T,N_1,N_2}, \;\; S \;=\; -\left.\frac{\partial F}{\partial T}\right|_{V,N_1,N_2}, \;\; \mu_1 \;=\; \left.\frac{\partial F}{\partial N_1}\right|_{T,V,N_2}, \;\; \mu_2 \;=\; \left.\frac{\partial F}{\partial N_2}\right|_{T,V,N_1}. \tag{A.5}$$

(T, V, N_1, N_2) describes the outer characteristics of (or outer constraints on) the system, and (A.4) describes how F would change when these outer constraints are changed, and could go up or down. But there are also inner degrees of freedom inside the system (for example, precipitate/matrix microstructure, which you cannot see or fix from the outside, and can only observe when you open up the material and take to a TEM). When the inner degrees of freedom change under fixed (T, V, N_1, N_2), the 2nd law states that F must decrease with time.

From theory of statistical mechanics it is convenient to start from F, since there is a direct microscopic expression for F, $F = -k_{\text{B}}T \ln Z$, where Z is so-called **partition function** [83, 84]. Plugging into (A.5), one then obtains direct microscopic expressions for P, the so-called internal pressure (or its generalization in 6-dimensional strain space, the stress tensor $\boldsymbol{\sigma}$, in so-called Virial formula), as well as S, μ_1, μ_2. This then allows atomistic simulation people to calculate so-called equation-of-state $P(T, V, N_1, N_2)$ and thermochemistry $\mu_i(T, V, N_1, N_2)$, if only the correct interatomic potential $U(\mathbf{x}^{3(N_1+N_2)})$ is provided. The so-called first-principles CALPHAD (CALculation of PHAse Diagrams) [85] is based on this approach, and is now a major source of phase diagram and thermochemistry information for alloy designers (metal hydrides for hydrogen storage, battery electrodes where you need to put in and pull out lithium ions, and catalysts). Since atomistic simulation can access metastable states and even saddle-points, there is also first-principles calculations of *mobilities*, such as diffusivities, interfacial mobilities, chemical reaction activation energies, etc. So F is important quantity computationally.

For experimentalist, however, most experiments are done under constant external pressure instead of constant volume (imagine melting of ice cube on the table, there is a natural tendency for volume change, illustrating the concept of *transformation volume*). For discussing phase change under constant external pressure, we define Gibbs free energy

$G \equiv F + PV = E - TS + PV$. The full differential of G is

$$dG \;=\; VdP - SdT + \mu_1 dN_1 + \mu_2 dN_2 \tag{A.6}$$

so

$$V \;=\; \left.\frac{\partial G}{\partial P}\right|_{T,N_1,N_2}, \;\; S \;=\; -\left.\frac{\partial G}{\partial T}\right|_{P,N_1,N_2}, \;\; \mu_1 \;=\; \left.\frac{\partial G}{\partial N_1}\right|_{T,P,N_2}, \;\; \mu_2 \;=\; \left.\frac{\partial G}{\partial N_2}\right|_{T,P,N_1}. \tag{A.7}$$

The above describes how a homogeneous material's G would change when its T, P, N_1, N_2 are changed, which could go up or down. If the system has internal inhomogeneities that are evolving under constant T, P, N_1, N_2, however, then G must decrease with time. Internal microstructural changes under constant T, P, N_1, N_2 that *increase* G are *forbidden*.

Also,

$$d(E + PV) \;=\; \delta Q + VdP + \mu_1 dN_1 + \mu_2 dN_1 \tag{A.8}$$

so if a closed system is under constant pressure, the heat it absorbs is the change in the enthalpy $H \equiv E + PV = G + TS$. H is also related to G through the so-called **Gibbs-Helmholtz relation**:

$$H \;=\; \left.\frac{\partial (G/T)}{\partial (1/T)}\right|_{N_1,N_2,P}. \tag{A.9}$$

Putting Δ before both sides of (A.9), the *heat of transformation* ΔH is related to the free-energy *driving force of transformation* as

$$\Delta H \;=\; \left.\frac{\partial (\Delta G/T)}{\partial (1/T)}\right|_{N_1,N_2,P}. \tag{A.10}$$

Now we formally introduce the concept of *thermodynamic driving force* for phase transformation. Consider two possible phases $\phi = \alpha, \beta$ that the system could be in. Both phases have the same numbers of atoms N_1, N_2, the same T and P. Consider pressure-driven phase transformation, $dG^\alpha = V^\alpha dP$, $dG^\beta = V^\beta dP$. Suppose $V^\alpha > V^\beta$, when we plot G^α and G^β graphically on the same plot, we see that at low pressure, the high-volume phase α may win; but at high pressure, the low-volume (denser phase) β will win. As a general rule, when P is increased keeping T fixed, the denser phase will win. So liquid phase will win over gas, and typically solid phase will win over liquid. Consider for example Fig. A.1(a). Density ranking: $\epsilon > \gamma > \alpha$. For fixed T, N_1, N_2, there exists an equilibrium pressure P_{eq} where the

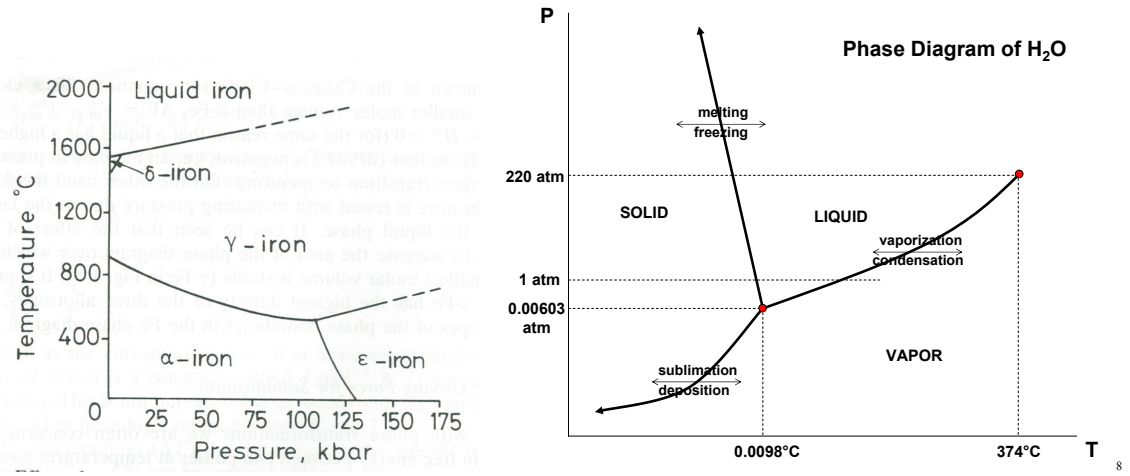


Figure A.1: (a) Figure 1.5 of Porter & Easterling [47]. (b) Phase diagram of pure H_2O: the solid-liquid boundary has negative dP/dT, which is an anomaly, because ice has larger volume than liquid water.

Gibbs free energy curves cross, at which

$$G^\alpha(P_{\rm eq}, T, N_1, N_2) \;=\; G^\beta(P_{\rm eq}, T, N_1, N_2). \tag{A.11}$$

At $P > P_{\rm eq}$, the driving force for $\alpha \to \beta$ is $\Delta G \approx (V^\alpha - V^\beta)(P - P_{\rm eq})$. Vice versa, at $P < P_{\rm eq}$, the driving force for $\beta \to \alpha$ is $\Delta G \approx (V^\alpha - V^\beta)(P_{\rm eq} - P)$ (by convention, we make the driving force positive). $P - P_{\rm eq}$ ($P_{\rm eq} - P$) may be called the overpressure (underpressure), respectively.

We could also have temperature-driven transformation, keeping pressure fixed: $dG^\alpha = -S^\alpha dT$, $dG^\beta = -S^\beta dT$. So G vs T is a downward curve. The question is which phase is going down faster, G^α or G^β. The answer is that the state that is more disordered (larger S) will go down faster with $T \uparrow$. So at some high enough T there will be a crossing. Liquid is going down faster than solid, gas is going down faster than liquid, with $T \uparrow$ holding P constant. For a fixed pressure, there exists an equilibrium temperature $T_{\rm eq}$ where the Gibbs free energy curves cross, at which

$$G^\alpha(P, T_{\rm eq}, N_1, N_2) \;=\; G^\beta(P, T_{\rm eq}, N_1, N_2). \tag{A.12}$$

Consider for example solid↔liquid transformation. In this case, $T_{\rm eq} = T_{\rm M}(P)$, the equilibrium bulk melting point. α=liquid, β=solid, $S^\alpha > S^\beta$. At $T > T_{\rm eq}$, the more disordered

phase is favored, and the driving force for $\beta \to \alpha$ transformation, which is melting, is $\Delta G \approx (S^\alpha - S^\beta)(T - T_{\rm M})$. Vice versa at $T < T_{\rm eq}$, the more ordered phase is favored, which is solidification, and the driving force for $\alpha \to \beta$ is $\Delta G \approx (S^\alpha - S^\beta)(T_{\rm M} - T)$. Because we are doing first-order expansion, it is OK to take $S^\alpha - S^\beta$ to be the value at $T_{\rm M}$. However, at $T_{\rm M}$ we have $E^\alpha + PV^\alpha - T_{\rm M}S^\alpha = H^\alpha - T_{\rm M}S^\alpha = H^\beta - T_{\rm M}S^\beta = E^\beta + PV^\beta - T_{\rm M}S^\beta$, we have $S^\alpha - S^\beta = (H^\alpha - H^\beta)/T_{\rm M}$. $H^\alpha - H^\beta$ is in fact the heat released during phase change under constant pressure, and is called the **latent heat** L. So we have

$$\Delta G \;\approx\; \frac{L}{T_{\rm M}}|T_{\rm M} - T|. \tag{A.13}$$

$|T_{\rm M} - T|$ is called undercooling / superheating for solidification / melting. We see that the thermodynamic driving force for phase change is proportional to the amount of undercooling / superheating (in Kelvin), with proportionality factor $\frac{L}{T_{\rm M}} = \Delta S$. Later we will see later why a finite thermodynamic driving force is needed, in order to observe phase change within a *finite amount of time*. (If you are extremely leisurely and have infinite amount of time, you can observe phase change right at $T_{\rm eq}$).

solid/liquid: melting, freezing or solidification. liquid/vapor: vaporization, condensation. solid/vapor: sublimation, deposition. At low enough pressure, the gas phase is going to come down in free energy significantly, that the solid goes directly to gas, without going through the liquid phase.

Thus, typically, high pressure / low temperature stabilizes solid phase, low pressure / high temperature stabilizes gas phase. The tradeoff relation can be described by the **Clausius-Clapeyron relation** for polymorphic phase transformation (single-component) in $T - P$ plane. The question we ask is that suppose you are already sitting on a particular (T, P) point that reaches perfect equilibrium between α, β,

$$G^\alpha(N_1, N_2, T, P) \;=\; G^\beta(N_1, N_2, T, P) \tag{A.14}$$

in which direction on the (T, P) plane should one go, $(T, P) \to (T+dT, P+dP)$, to *maintain* that equilibrium, i.e.:

$$G^\alpha(N_1, N_2, T + dT, P + dP) \;=\; G^\beta(N_1, N_2, T + dT, P + dP) \tag{A.15}$$

$$G^\alpha(N_1, N_2, T, P) - S^\alpha dT + V^\alpha dP \;=\; G^\beta(N_1, N_2, T, P) - S^\beta dT + V^\beta dP. \tag{A.16}$$

So:

$$-S^\alpha dT + V^\alpha dP \;=\; -S^\beta dT + V^\beta dP. \tag{A.17}$$

and the direction is given by

$$\frac{dP}{dT} \;=\; \frac{S^\alpha - S^\beta}{V^\alpha - V^\beta} \;=\; \frac{L}{T(V^\alpha - V^\beta)}. \tag{A.18}$$

The above equation keeps one "on track" on the $T-P$ phase diagram. It's like in pitch darkness, if you happen to stumble upon a rail, you can *follow* the rail to map out the whole US railroad system. The Clausius-Clapeyron relation tells you how to follow that rail. L is called "latent heat". $V^\alpha - V^\beta$ is the volume of melting/vaporization/sublimation, you may call it the "latent volume".

In above we have only considered the scenario of so-called congruent transformation $\alpha \leftrightarrow \beta$, where α and β are single phases with the same composition. We have not considered the possibility of for example $\alpha \leftrightarrow \beta + \gamma$, where γ has different composition or even structure from β. To understand the driving force for such transformations which are indeed possible in binary solutions, we need to further develop the **language of chemical potential**.

The total number of particles is $N \equiv N_1 + N_2$. Define mole fractions $X_1 \equiv N_1/N$, $X_2 \equiv N_2/N$. Since there is always $X_1 + X_2 = 1$, we cannot regard X_1 and X_2 as independent variables. Usually by convention one takes X_2 to be the independent variable, so-called *composition*. Composition is dimensionless, but it could be a multi-dimensional vector if the number of species $C > 2$. For instance, in a ternary solution, $C = 3$, and composition is a 2-dimensional vector $\mathbf{X} \equiv [X_2, X_3]$. Composition can spatially vary in inhomogeneous systems, for instance in an inhomogeneous binary solution, $X_2 = X_2(\mathbf{x}, t)$. In order for $\alpha \leftrightarrow \beta + \gamma$ to happen kinetically, for instance changing from $X_2(\mathbf{x}) = 0.3$ uniformly (initially α phase) to some region with $X_2(\mathbf{x}) = 0.5$ (in β phase, "solute sink") and some region with $X_2(\mathbf{x}) = 0.1$ (in γ phase, 'solute source"). This requires would require long-range diffusion of type-2 solutes over distances on the order of the sizescale of the inhomogeneities, which is called **solute partitioning**.

We can define the *particle average Gibbs free energy* to be $g \equiv G/N = G(T, P, N_1, N_2)/(N_1 + N_2)$. Like the chemical potentials, g will be minus a few eV in reference to isolated atoms ensemble. It can be rigorously proven, but is indeed quite intuitively obvious, that $g = g(X_2, T, P)$, which is to say the particle average Gibbs free energy depends on chemistry but not quantity (think of $(N_1, N_2) \leftrightarrow (N, X_2)$ as a variable transform that decomposes

dependent variables into *quantity* and *chemistry*). It is customary to plot g versus X_2 at constant T, P. It can be mathematically proven that μ_1, μ_2 are the *tangent extrapolations* of $g(X_2)$ to $X_2 = 0$ and $X_2 = 1$, respectively. Algebraically this means

$$\begin{aligned} \mu_1(X_2, T, P) &= g(X_2, T, P) + \left.\frac{\partial g}{\partial X_2}\right|_{T,P} (0 - X_2) \\ \mu_2(X_2, T, P) &= g(X_2, T, P) + \left.\frac{\partial g}{\partial X_2}\right|_{T,P} (1 - X_2). \end{aligned} \tag{A.19}$$

It is also clear from the above that $g(X_2, T, P) = X_1\mu_1 + X_2\mu_2$, so

$$G(T, P, N_1, N_2) \;=\; N_1\mu_1 + N_1\mu_2 \;=\; N_1 \left.\frac{\partial G}{\partial N_1}\right|_{T,P,N_2} + N_2 \left.\frac{\partial G}{\partial N_2}\right|_{T,P,N_1} \tag{A.20}$$

On first look, the above seems to imply that particle 1 and particle 2 do not interact. But this is very far from true! In fact, $\mu_1 = \mu_1(X_2, T, P)$, $\mu_2 = \mu_2(X_2, T, P)$.

For pure systems: $X_2 = 0$, $g(X_2 = 0, T, P) = \mu_1(X_2 = 0, T, P) \equiv \tilde{\mu}_1(T, P)$; or $X_2 = 1$, $g(X_2 = 1, T, P) = \mu_2(X_2 = 1, T, P) \equiv \tilde{\mu}_2(T, P)$. $\tilde{\mu}_1(T, P)$, $\tilde{\mu}_2(T, P)$ are called Raoultian reference-state chemical potentials (they are not the isolated-atoms-in-vaccuum reference states, but already as interacting-atoms). In this class we take the $\tilde{\mu}_1$, $\tilde{\mu}_2$ reference states to the same structure as the solution, but in pure compositions (so-called Raoultian reference states).

When plotted graphically, it is seen that $g(X_2)$ is typically convex up with $\mu_1(X_2, T, P) < \tilde{\mu}_1(T, P)$ and $\mu_2(X_2, T, P) < \tilde{\mu}_2(T, P)$ (if not, what would happen?) This negative difference is defined as the *mixing chemical potential*

$$\mu_i^{\text{mix}} \;\equiv\; \mu_i(X_2, T, P) - \tilde{\mu}_i(T, P), \;\; i = 1, 2 \tag{A.21}$$

and *mixing free energy*

$$g^{\text{mix}} \;\equiv\; X_1\mu_1^{\text{mix}} + X_2\mu_2^{\text{mix}} = g - X_1\tilde{\mu}_1(T, P) - X_2\tilde{\mu}_2(T, P), \quad G^{\text{mix}} = Ng^{\text{mix}} \tag{A.22}$$

respectively. Clearly, by definition, $G^{\text{mix}} = 0$ at pure competitions. $g^{\text{mix}}(X_2, T, P)$ can be interpreted as the driving force to react pure 1 and pure 2 of the same structure as the solution to obtain a solution of non-pure composition, per particle in the mixed solution. $\Delta G = -Ng^{\text{mix}}(X_2, T, P)$ is in fact the *chemical driving force* to make a solution by mixing

pure constituents.

It turns out there exists "partial" version of the full differential (A.6):

$$dg(X_2, T, P) \;=\; vdP - sdT + \left.\frac{\partial g}{\partial X_2}\right|_{T,P} dX_2 \tag{A.23}$$

$$d\mu_i(X_2, T, P) \;=\; v_i dP - s_i dT + \left.\frac{\partial \mu_i}{\partial X_2}\right|_{T,P} dX_2 \tag{A.24}$$

where

$$\begin{aligned} v_1 \;&\equiv\; \left.\frac{\partial V}{\partial N_1}\right|_{T,P,N_2}, \quad v_2 \;\equiv\; \left.\frac{\partial V}{\partial N_2}\right|_{T,P,N_1}, \quad s_1 \;\equiv\; \left.\frac{\partial S}{\partial N_1}\right|_{T,P,N_2}, \quad s_2 \;\equiv\; \left.\frac{\partial S}{\partial N_2}\right|_{T,P,N_1}, \\ e_1 \;&\equiv\; \left.\frac{\partial E}{\partial N_1}\right|_{T,P,N_2}, \quad e_2 \;\equiv\; \left.\frac{\partial E}{\partial N_2}\right|_{T,P,N_1}, \quad h_1 \;\equiv\; \left.\frac{\partial H}{\partial N_1}\right|_{T,P,N_2}, \quad h_2 \;\equiv\; \left.\frac{\partial H}{\partial N_2}\right|_{T,P,N_1}, \quad ... \end{aligned} \tag{A.25}$$

Generally speaking, for arbitrary extensive quantity A (volume, energy, entropy, enthalpy, Helmholtz free energy, Gibbs free energy), "particle partial A" is defined as:

$$a_i \;\equiv\; \left.\frac{\partial A}{\partial N_i}\right|_{N_{j\neq i},T,P}. \tag{A.26}$$

The meaning of a_i is the increase in energy, enthalpy, volume, entropy, etc. when an additional type-i atom is added into the system, keeping the temperature and pressure fixed. The *particle-average* a is simply

$$a \;\equiv\; \frac{A}{N} \;=\; \sum_{i=1}^{C} X_i a_i. \tag{A.27}$$

For instance, the particle average volume and particle average entropy

$$v \;\equiv\; \frac{V}{N} = X_1 v_1 + X_2 v_2, \quad s \;\equiv\; \frac{S}{N} = X_1 s_1 + X_2 s_2, \tag{A.28}$$

is simply the composition-weighted sum of particle partial volumes and partial entropies of different-species atoms, respectively. While (A.27) relates all $a_i(X_2, ..., X_C, T, P)$s to $a(X_2, ..., X_C, T, P)$, it is also possible to obtain individual $a_i(X_2, ..., X_C, T, P)$ from $a(X_2, ..., X_C, T, P)$

by the tangent extrapolation formula:

$$a_i(X_2, ..., X_C, T, P) \;=\; a(X_2, ..., X_C, T, P) + \sum_{k=2}^{C} (\delta_{ik} - X_k) \frac{\partial a(X_2, ..., X_C, T, P)}{\partial X_k}, \qquad \text{(A.29)}$$

where δ_{ik} is the Kronecker delta: $\delta_{ik} = 1$ if $i = k$, and $\delta_{ik} = 0$ if $i \neq k$. Note in (A.29), although the k-sum runs from 2 to C, i can take values 1 to C. (A.19) is a special case of (A.29): for historical reason the particle partial Gibbs free energy is denoted by μ_i instead of g_i.

The so-called Gibbs-Duhem relation imposes constraint on the partial quantities when composition is varied while holding T, P fixed:

$$0 \;=\; \sum_{i=1}^{C} X_i da_i|_{T,P}, \qquad \text{(A.30)}$$

For binary solution, this means

$$0 = X_1 d\mu_1|_{T,P} + X_2 d\mu_2|_{T,P} = X_1 dv_1|_{T,P} + X_2 dv_2|_{T,P} = X_1 ds_1|_{T,P} + X_2 ds_2|_{T,P} = ... \qquad \text{(A.31)}$$

The above can be proven, but we will not do it here.

The above is the general solution thermodynamics framework. To proceed further, we need some detailed models of how g depends on X_2. In so-called ideal solution:

$$\mu_1^{\text{ideal}-\text{mix}}(X_2, T, P) \;=\; k_{\text{B}} T \ln X_1, \;\; \mu_2^{\text{ideal}-\text{mix}}(X_2, T, P) \;=\; k_{\text{B}} T \ln X_2. \qquad \text{(A.32)}$$

And so

$$g^{\text{ideal}-\text{mix}}(X_2, T, P) \;\equiv\; k_{\text{B}} T (X_1 \ln X_1 + X_2 \ln X_2), \qquad \text{(A.33)}$$

which is a symmetric function that is always negative (that is to say it always prefer mixing), with $-\infty$ slope on both sides. Ideal solution is realized nearly exactly in isotopic solutions such as ^{235}U - ^{238}U. In such case, there is no chemical difference between the two species ($\epsilon_{\text{AA}} = \epsilon_{\text{BB}} = \epsilon_{\text{AB}}$), so the enthalpy of mixing is zero. The driving force for mixing is entirely entropic in origin, because there would be many ways to arrange ^{235}U and ^{238}U atoms on a lattice, whereas there is just one in pure ^{235}U or pure ^{238}U crystal (^{235}U atoms are indistinguishable among themselves, and so are ^{238}U atoms). This can be verified from the formula $s^{\text{mix}} = -\partial g^{\text{mix}}/\partial T$, $h^{\text{mix}} = \partial(g^{\text{mix}}/T)/\partial(1/T)$.

We define excess as difference between the actual mix and the ideal-mix functions:

$$g^{\text{excess}} \equiv g^{\text{mix}}(X_2, T, P) - g^{\text{ideal-mix}}(X_2, T, P), \quad \mu_i^{\text{excess}} \equiv \mu_i^{\text{mix}} - k_{\text{B}} T \ln X_i. \tag{A.34}$$

Clearly, excess quantities for ideal solution is zero.

In so-called regular solution model,

$$g^{\text{excess}}(X_2, T, P) = \omega X_1 X_2, \tag{A.35}$$

where ω is X_2,T,P independent constant. Using (A.19), we get

$$\mu_1^{\text{excess}} = \omega X_2^2, \quad \mu_2^{\text{excess}} = \omega X_1^2. \tag{A.36}$$

And so

$$\mu_1(X_2) = \tilde{\mu}_1 + k_{\text{B}} T \ln X_1 + \omega X_2^2, \quad \mu_2(X_2) = \tilde{\mu}_2 + k_{\text{B}} T \ln X_2 + \omega X_1^2. \tag{A.37}$$

It is also customary to define *activity coefficient* γ_i, so that

$$\mu_i(X_2, T) \equiv \tilde{\mu}_i(T) + k_{\text{B}} T \ln \gamma_i X_i. \tag{A.38}$$

Contrasting with (A.37), we see that in the regular solution model, the activity coefficients are $\gamma_2(X_2, T) = e^{\omega X_1^2 / k_{\text{B}} T}$, $\gamma_1(X_2, T) = e^{\omega X_2^2 / k_{\text{B}} T}$.

When $\omega < 0$, the driving force for mixing is *greater* than in ideal solution. When one uses the formula $s = -\partial g / \partial T$, $h = \partial (g/T) / \partial (1/T)$, we can see that the ideal-mixing contribution is entirely *entropic*, whereas the excess contribution is entirely *enthalpic* if ω is independent of temperature. In fact, it can be shown from statistical mechanics that

$$\omega = Z \left((\epsilon_{\text{AA}} + \epsilon_{\text{BB}})/2 - \epsilon_{\text{AB}} \right), \tag{A.39}$$

where ϵ_{AB} is the Kossel spring binding energy between A-B ("heteropolar bond"), and ϵ_{AA} and ϵ_{BB} are the Kossel spring binding energy between A-A and B-B (homopolar bonds).

Derivation of the **regular solution model** (this has been shown in MSE530 Thermodynamics of Materials): arrange $X_{\text{A}} N$ A atoms and $X_{\text{B}} N$ B atoms on a lattice. The number

of choices:

$$\Omega \;=\; \frac{N!}{(X_{\rm A}N)!(X_{\rm B}N!)} \tag{A.40}$$

Assume all these choices (microstates) have the same enthalpy:

$$\begin{aligned} H &= -Z(X_{\rm A}N(X_{\rm B}\epsilon_{\rm AB}+X_{\rm A}\epsilon_{\rm AA})+X_{\rm B}N(X_{\rm B}\epsilon_{\rm BB}+X_{\rm A}\epsilon_{\rm AB}))/2 \\ &= -NZ(2X_{\rm A}X_{\rm B}\epsilon_{\rm AB}+X_{\rm A}^2\epsilon_{\rm AA}+X_{\rm B}^2\epsilon_{\rm BB})/2 \end{aligned} \tag{A.41}$$

in contrast to reference state of pure A and pure B

$$H^{\rm ref} \;=\; -NZ(X_{\rm A}\epsilon_{\rm AA}+X_{\rm B}\epsilon_{\rm BB})/2 \tag{A.42}$$

so the excess is:

$$\begin{aligned} H^{\rm excess} &= -NZ(2X_{\rm A}X_{\rm B}\epsilon_{\rm AB}+X_{\rm A}^2\epsilon_{\rm AA}-X_{\rm A}\epsilon_{\rm AA}+X_{\rm B}^2\epsilon_{\rm BB}-X_{\rm B}\epsilon_{\rm BB})/2 \\ &= -NZ(2X_{\rm A}X_{\rm B}\epsilon_{\rm AB}-X_{\rm A}X_{\rm B}\epsilon_{\rm AA}-X_{\rm B}X_{\rm A}\epsilon_{\rm BB})/2 \\ &= NZX_{\rm A}X_{\rm B}\left((\epsilon_{\rm AA}+\epsilon_{\rm BB})/2-\epsilon_{\rm AB}\right) \;=\; N\omega X_{\rm A}X_{\rm B}. \end{aligned} \tag{A.43}$$

According to the Boltzmann formula $S = k_{\rm B}\ln\Omega$, the entropy is

$$\begin{aligned} S &= k_{\rm B}\ln\frac{N!}{(X_{\rm A}N)!(X_{\rm B}N!)} \approx k_{\rm B}(N\ln N - X_{\rm A}N\ln X_{\rm A}N - X_{\rm B}N\ln X_{\rm B}N) \\ &= -Nk_{\rm B}(X_{\rm A}\ln X_{\rm A}+X_{\rm B}\ln X_{\rm B}), \end{aligned} \tag{A.44}$$

using the Stirling formula: $\ln N! \approx N\ln N - N$ for large N. S is the same as that in ideal solution, because the regular solution model takes the "mean-field" view that all possible configurations are iso-energetic. The regular solution model in the form of (A.35) is a well-posed *model* with algebraic simplicity, but it may not reflect reality very well.

For positive ω, spinodal decomposition will happen below a critical temperature $T_{\rm C}$: a random 50%-50% A-B solution α would separate into A-rich solution α_1 and B-rich solution α_2 - see plots of $g(X_2, T)$ at different T. We have studied this model in detail in MSE530.

For negative ω, although nothing will happen as seen from the regular solution model, in reality **order-disorder transition** will happen below a critical temperature $T_{\rm C}$, where the A-B solution starts to posses **chemical long-range order** (CLRO). A good example is β-brass, a Cu-Zn alloy in BCC structure ($Z = 8$). See Chap. 17 of [41]. Cu and Zn atoms like each other energetically, more than Cu-Cu, and Zn-Zn. Suppose $X_{\rm Zn} = 0.5$, at $T = 0$, what

would be the optimal microscopic configuration? Since $F = E - TS$, at $T = 0$ minimization of F is the same as minimization of $E = U$, the system will try to maximize the number of Cu-Zn bonds. Indeed, so-called long-range chemical order, that is, Cu occupying one sub-lattice (') and Zn occupying another sub-lattice ("), or Cu occupying sub-lattice " and Zn occupying sub-lattice ' would give the maximum number of Cu-Zn bonds. The regular solution model did not distinguish between the two sub-lattices, statistically speaking. In order to be able to distinguish, let us define sub-lattice compositions $X'_{\rm A} + X'_{\rm B} = 1$, $X''_{\rm A} + X''_{\rm B} = 1$. Clearly the overall composition

$$X_{\rm A} \;=\; \frac{1}{2}(X'_{\rm A} + X''_{\rm A}), \;\; X_{\rm B} \;=\; \frac{1}{2}(X'_{\rm B} + X''_{\rm B}). \tag{A.45}$$

By defining sub-lattice compositions, we have effectively added one more "coarse" degree of freedom to describe our alloy, the so-called η order parameter:

$$\eta \;\equiv\; \frac{1}{2}(X''_{\rm B} - X'_{\rm B}). \tag{A.46}$$

$Cu_{50}Zn_{50}$ taking the CsCl structure at $T = 0$ would have $\eta = 0.5$ or $\eta = -0.5$. Previously, the regular solution model constrains $\eta = 0$ (because it does not entertain an η order parameter). Now, with η, we would have

$$X''_{\rm B} = X_{\rm B} + \eta, \;\; X'_{\rm B} = X_{\rm B} - \eta, \;\; X''_{\rm A} = 1 - X_{\rm B} - \eta, \;\; X'_{\rm A} = 1 - X_{\rm B} + \eta. \tag{A.47}$$

Still under the mean-field approximation (so called **Bragg-Williams** approach [86, 87] in alloy thermochemistry), as the regular solution model, we can estimate the proportion of A(')-A(") bonds:

$$p_{\rm AA} \;=\; X''_{\rm A}X'_{\rm A} = (1 - X_{\rm B} - \eta)(1 - X_{\rm B} + \eta), \tag{A.48}$$

the proportion of B(')-B(") bonds:

$$p_{\rm BB} \;=\; X''_{\rm B}X'_{\rm B} = (X_{\rm B} + \eta)(X_{\rm B} - \eta), \tag{A.49}$$

the proportion of A(')-B(") bonds:

$$p_{\rm AB} \;=\; X'_{\rm A}X''_{\rm B} = (1 - X_{\rm B} + \eta)(X_{\rm B} + \eta), \tag{A.50}$$

the proportion of A(")-B(') bonds:

$$p_{\mathrm{BA}} \;=\; X'_{\mathrm{B}} X''_{\mathrm{A}} = (X_{\mathrm{B}} - \eta)(1 - X_{\mathrm{B}} - \eta) \tag{A.51}$$

among all the nearest-neighbor bonds in the alloy. Clearly, the above Bragg-Williams estimation satisfies the sum rule constraint:

$$p_{\mathrm{AA}} + p_{\mathrm{BB}} + p_{\mathrm{AB}} + p_{\mathrm{BA}} \;=\; 1. \tag{A.52}$$

The particle-average energy is thus just

$$h \;=\; -\frac{Z}{2}(p_{\mathrm{AA}}\epsilon_{\mathrm{AA}} + p_{\mathrm{BB}}\epsilon_{\mathrm{BB}} + (p_{\mathrm{AB}} + p_{\mathrm{BA}})\epsilon_{\mathrm{AB}}) \tag{A.53}$$

From derivations of the regular solution model and discussions in the last semester, we see that if we chose our reference state appropriately, then we can say $\epsilon_{\mathrm{AA}} = 0$, $\epsilon_{\mathrm{BB}} = 0$, $\epsilon_{\mathrm{AB}} = -\omega/Z$, to simplify the algebra:

$$h(X_{\mathrm{B}}, \eta) \;=\; \omega(X_{\mathrm{A}} X_{\mathrm{B}} + \eta^2). \tag{A.54}$$

which we see is the same as the regular solution model if $\eta = 0$. The physics of the above expression is that, if with CLRO and solute partitioning onto the two sub-lattices, one can increase the number of A-B bonds from $X_{\mathrm{A}} X_{\mathrm{B}}$ to $X_{\mathrm{A}} X_{\mathrm{B}} + \eta^2$.

The entropy is just the sum of the entropies of the two sub-lattices (in other words, the total number of possible microstates is the product of the numbers of microstates on ' sublattice and that on " sublattice). Therefore:

$$s(X_{\mathrm{B}}, \eta) \;=\; -\frac{k_{\mathrm{B}}}{2}(X'_{\mathrm{A}} \ln X'_{\mathrm{A}} + X'_{\mathrm{B}} \ln X'_{\mathrm{B}} + X''_{\mathrm{A}} \ln X''_{\mathrm{A}} + X''_{\mathrm{B}} \ln X''_{\mathrm{B}}). \tag{A.55}$$

The free energy (of mixing) per particle is thus

$$g(X_{\mathrm{B}}, \eta) \;=\; \omega(X_{\mathrm{A}} X_{\mathrm{B}} + \eta^2) + \frac{k_{\mathrm{B}} T}{2}(X'_{\mathrm{A}} \ln X'_{\mathrm{A}} + X'_{\mathrm{B}} \ln X'_{\mathrm{B}} + X''_{\mathrm{A}} \ln X''_{\mathrm{A}} + X''_{\mathrm{B}} \ln X''_{\mathrm{B}}) \tag{A.56}$$

with

$$\frac{\partial g}{\partial \eta} \;=\; 2\omega\eta + \frac{k_{\mathrm{B}} T}{2}\left(-\ln \frac{X'_{\mathrm{B}}}{X'_{\mathrm{A}}} + \ln \frac{X''_{\mathrm{B}}}{X''_{\mathrm{A}}}\right), \tag{A.57}$$

$$\frac{\partial^2 g}{\partial \eta^2} \;=\; 2\omega + \frac{k_{\mathrm{B}}T}{2}\left(\frac{1}{X'_{\mathrm{B}}X'_{\mathrm{A}}} + \frac{1}{X''_{\mathrm{B}}X''_{\mathrm{A}}}\right). \tag{A.58}$$

In a real material, both X_{B} and η are fields: $g(X_{\mathrm{B}}(\mathbf{x},t),\eta(\mathbf{x},t))$. However, we note there is a fundamental difference between X_{B} and η. $X_{\mathrm{B}}(\mathbf{x},t)$ is conserved:

$$\int d\mathbf{x} X_{\mathrm{B}}(\mathbf{x},t) \;=\; \text{const} \tag{A.59}$$

if integration is carried out in the entire space. Thus, when optimizing

$$G \;=\; \frac{1}{\Omega}\int d\mathbf{x} g(X_{\mathrm{B}}(\mathbf{x}),\eta(\mathbf{x})) \tag{A.60}$$

we can not do an unconstrained optimization on $g(X_{\mathrm{B}})$: there has to be a Lagrange multiplier (the chemical potential) on the total free energy minimization. On the other hand, there is no such constraint on η: we can do an unconstrained optimization with respect to η (and indeed that is what Nature does). More involved discussions [41] show that X_{B} is so-called **conserved order parameter**, and evolve according to the so-called **Cahn-Hilliard** evolution equation [88] (basically diffusion equation), whereas **non-conserved order parameter** like the CLRO evolve according to the so-called **Allen-Cahn** equation [89], in the linear response regime.

For a given T, X_{B}, we thus have

$$g(X_{\mathrm{B}}) \;=\; \min_{\eta} g(X_{\mathrm{B}},\eta) \tag{A.61}$$

at thermodynamic equilibrium. So:

$$\ln\frac{(X_{\mathrm{B}}-\eta)(1-X_{\mathrm{B}}-\eta)}{(X_{\mathrm{B}}+\eta)(1-X_{\mathrm{B}}+\eta)} \;=\; \frac{4\omega\eta}{k_{\mathrm{B}}T} \tag{A.62}$$

We note that $\eta = 0$ is always a solution to above, i.e. it is always a stationary point in the variational problem. But is $\eta = 0$ a minimum or a maximum? From (A.58) we note that at high enough T, $\eta = 0$ would always be a free energy minimum. But as T cools down, at

$$T_{\mathrm{C}}(X_{\mathrm{B}}) \;=\; \frac{-2\omega X_{\mathrm{B}}(1-X_{\mathrm{B}})}{k_{\mathrm{B}}} \tag{A.63}$$

$g(X_{\mathrm{B}},\eta)$ would lose stability with respect to η at $\eta = 0$, in a manner of 2nd order phase transformation (for example, magnetization at Curie temperature). This is called order-disorder

transformation, where chemical long-range order emerges at a low enough temperature. In particular, the highest temperature where chemical order may emerge is at $X_\mathrm{B} = 0.5$, where the enthalpic driving force for two sub-lattice partition is especially strong:

$$T_\mathrm{C}^* \;=\; -\frac{\omega}{2k_\mathrm{B}}. \tag{A.64}$$

We also note that T_C^* exists only for $\omega < 0$. If $\omega > 0$, $\frac{\partial^2 g}{\partial \eta^2} > 0$ always and $\eta = 0$ stays stable global minimum. Thus the Bragg-Williams model is the same as the regular solution model for $\omega > 0$. The Bragg-Williams model gives only different results from the regular solution model for $\omega < 0$, and in that case for

$$T \;<\; T_\mathrm{C}(X_\mathrm{B}) = 4T_\mathrm{C}^* X_\mathrm{B}(1 - X_\mathrm{B}) \tag{A.65}$$

only. At $T < T_\mathrm{C}(X_\mathrm{B})$, we have the CLRO at equilibrium:

$$\ln \frac{(X_\mathrm{B} + \eta)(1 - X_\mathrm{B} + \eta)}{(X_\mathrm{B} - \eta)(1 - X_\mathrm{B} - \eta)} \;=\; \frac{8\eta T_\mathrm{C}^*}{T}, \tag{A.66}$$

from which we can solve for η.

The above is called the Bragg-Williams approach, which is at the same level of theory (mean-field approximation) as the regular solution model, and only gives different results ($\eta \neq 0$) if $\omega < 0$ and $T < T_\mathrm{C}$. There are certain solid-state chemistries where ω is very negative, in which case CLRO is close to the maximum possible value for a large temperature range. These are so-called *line compounds* (because off-stoichiometry solubility range is so low, these phases appear as lines in $T - X_2$ phase diagrams) or *ordered phases*, with formulas like $\mathrm{A}_m\mathrm{B}_n$ where m and n are integers. Many crystalline ceramics (oxides, nitrides, carbides etc.) are line *compounds*, as the solubility range is typically very narrow besides the ideal stoichiometry. In metallic alloys, these would be called *intermetallics compound* phases. These phases are typically very strong mechanically (stability due to very negative ω), and are used as strengthening phases (precipitates) to impede dislocation motion. There are special symbols to denote these phases with long-range chemical order, such as $\mathrm{L2}_0$ (bcc based), $\mathrm{L1}_2$ (fcc based), $\mathrm{L1}_0$ (fcc based), $\mathrm{D0}_3$, $\mathrm{D0}_{19}$, Laves phases, etc.

There is still a higher-level of theory called the **quasi-chemical approximation** [90, 91], originating from a series of approximations by Edward A. Guggenheim [83]. It proposes the concept of **chemical short-range order** (CSRO): even in so-called random solid solution ($\omega > 0$, or $\omega < 0$ but $T > T_\mathrm{C}$) which has no long-range chemical order, $\eta = 0$, the atomic

arrangements may not be random as in the mean-field sense, and manifest "correlations". For example, a pair "correlation" means the probability of finding a particular kind of A-B bond is larger than the product of average probabilities of finding A in a particular sublattice and B in another sublattice. Beyond pair correlations, there are also triplet correlations, quartet correlations, ..., in a so-called **cluster expansion** approach [85], each addressing an excess probability beyond the last level of theory. Specifically, in the quasi-chemical approximation one uses the pair probabilities p_{AA}, p_{BB}, p_{AB}, p_{BA} as coarse degrees of freedom. These are valid order parameters, because at least in principle one could count the fraction of A(')-A("), B(')-B("), A(')-B("), A(")-B(') bonds in a given RVE. These coarse-grained statistical descriptors will take certain values, and one can formulate a variational problem based on them.

p_{AA}, p_{BB}, p_{AB}, p_{BA} must satisfy sum rule (A.52). Therefore, in addition to X_{B}, η, the quasi-chemical approximation introduces three more degrees of freedom. In systems where CLRO vanish, there is no statistical distinction between the two sub-lattices, so $p_{\mathrm{AB}} = p_{\mathrm{BA}}$, in which case only two additional degrees of freedom from the quasi-chemical approach. The quasi-chemical free energy reads:

$$\begin{aligned} g(X_{\mathrm{B}}, \eta, p_{\mathrm{AB}}, p_{\mathrm{BA}}, p_{\mathrm{BB}}) \;=\; & \frac{\omega(p_{\mathrm{AB}} + p_{\mathrm{BA}})}{2} + \\ & \frac{k_{\mathrm{B}}T}{2}(X'_{\mathrm{A}} \ln X'_{\mathrm{A}} + X'_{\mathrm{B}} \ln X'_{\mathrm{B}} + X''_{\mathrm{A}} \ln X''_{\mathrm{A}} + X''_{\mathrm{B}} \ln X''_{\mathrm{B}}) + \\ & \frac{Zk_{\mathrm{B}}T}{2}(p_{\mathrm{BB}} \ln \frac{p_{\mathrm{BB}}}{X'_{\mathrm{B}}X''_{\mathrm{B}}} + p_{\mathrm{AB}} \ln \frac{p_{\mathrm{AB}}}{X'_{\mathrm{A}}X''_{\mathrm{B}}} + p_{\mathrm{BA}} \ln \frac{p_{\mathrm{BA}}}{X'_{\mathrm{B}}X''_{\mathrm{A}}} + \\ & (1 - p_{\mathrm{BB}} - p_{\mathrm{AB}} - p_{\mathrm{BA}}) \ln \frac{1 - p_{\mathrm{BB}} - p_{\mathrm{AB}} - p_{\mathrm{BA}}}{X'_{\mathrm{A}}X''_{\mathrm{A}}}) \end{aligned} \tag{A.67}$$

with sub-lattice compositions X'_{A}, X''_{A}, X'_{B}, X''_{B} taken from (A.47) The actual chemical free energy at local equilibrium is

$$g(X_{\mathrm{B}}) \;=\; \min_{\eta, p_{\mathrm{AB}}, p_{\mathrm{BA}}, p_{\mathrm{BB}}} g(X_{\mathrm{B}}, \eta, p_{\mathrm{AB}}, p_{\mathrm{BA}}, p_{\mathrm{BB}}) \tag{A.68}$$

As a general remark, a compound phase would tend to manifest as sharp "needle" in $g(X_{\mathrm{B}})$, which means small deviation from the ideal stoichiometry $\mathrm{A}_m\mathrm{B}_n$ would cause large "pain" or increase in $g(X_{\mathrm{B}})$, since A-A and B-B bonds must be formed (due to the host lattice structure) which are much more energetically costly than A-B bonds.

Both spinodal decomposition and order-disorder transformation are *2nd-order phase trans-*

formations, defined by a vanishingly small jump in the *order parameter*, as one crosses the transition temperature $T_{\rm C}$. In contrast, 1st-order phase transition are characterized by a *finite jump* in order parameter. For instance, in melting, we can use the local density as order parameter to distinguish between liquid and solid, or some feature of the selected area electron diffraction (SAED) pattern. In either case, before and after melting, there is a finite jump in this order parameter field ($\rho(\mathbf{x}, T^-_{\rm melt}) = \rho^s$ but $\rho(\mathbf{x}, T^+_{\rm melt}) = \rho^l$ for some $\mathbf{x}$). Thus, melting is a 1st-order phase transitions. Also, consider an eutectic decomposition reaction: $l \rightarrow \alpha + \beta$, defined by $(T^{\rm E}, X_2^{l{\rm E}}, X_2^{\alpha{\rm E}}, X_2^{\beta{\rm E}})$. If one uses the local composition as the order parameter: then there is also a finite change ($X_2(\mathbf{x}, T^{{\rm E}+}) = X_2^{l{\rm E}}$ but $X_2(\mathbf{x}, T^{{\rm E}-}) = X_2^{\alpha{\rm E}}$ or $X_2^{\beta{\rm E}}$, for some $\mathbf{x}$). In contrast, in the case of $\omega > 0$ and spinodal decomposition $\alpha \rightarrow \alpha_1 + \alpha_2$ which is 2nd-order phase transformations, $X_2^{\alpha_2} - X_2^{\alpha_1} \propto \sqrt{T_{\rm C} - T}$. Whereas $X_2(\mathbf{x}, T^-_{\rm C}) = X_2^{\alpha}$ uniformly $T^+_{\rm C}$, one sees only *infinitesimal* compositional modulations at $T^-_{\rm C}$: $X_2(\mathbf{x}, T^-_{\rm C}) = X_2^{\alpha_1}$ or $X_2^{\alpha_2}$. The *amplitude* of the concentration wave (concentration is our order parameter here) is infinitesimal.

Common tangent construction: $\mu_2^{\alpha}(X_2^{\alpha}, T) = \mu_2^{\beta}(X_2^{\beta}, T)$, $\mu_1^{\alpha}(X_2^{\alpha}, T) = \mu_1^{\beta}(X_2^{\beta}, T)$ manifest as common tangent between $g^{\alpha}(X_2)$ and $g^{\beta}(X_2)$ curves. This equation has two unknowns, X_2^{α} and X_2^{β}, and we need to solve two joint equations which are generally nonlinear (thus numerical solution by computer may be needed). Show graphically how this may be established for two phases α, β, rich in A and B, respectively, by diffusion. Since

$$dG \;=\; VdP - SdT + \sum_{i=1}^{C} \mu_i dN_i, \tag{A.69}$$

atoms/molecules will always migrate from high chemical potential phase/condition to low chemical potential phase/condition.

Let us now investigate situations where a large-solubility phase (α) is in contact with a line compound phase (β). The common tangent construction can be simplified in these situations. Let us consider two limiting cases (**a**) and (**b**), where the $g^{\beta}(X_2, T)$ needle is "around" (**a**) $X_2 \approx 0$ and (**b**) $X_2 \approx 1$, respectively. (**a**) corresponds to an example of adding antifreeze to water, where the liquid solution delays freezing due to addition of solutes. (**b**) corresponds to an unknown solubility problem, which is to say how much can be dissolved in α for a given temperature when it is interfaced with a precipitate β phase that is nearly pure 2.

(**a**): people add antifreeze to say liquid water, to suppress the freezing temperature. How

does that work?

In this case, $g^{\beta}(X_2, T)$ is a needle "around" $X_2 \approx 0$ (the ice phase), whereas α is the liquid phase. The first thing to realize is the solubility of B is typically lower in solids than in liquids. Energetic interaction between atoms is more important in solids than liquids, since atoms in solids are bit closer in distance, and also put a premium on periodic packing. "Misfit" molecules B would feel much more comfortable living in a chaotic environment like liquid, than in a crystal (think about societal analogies). To first approximation, we can assume the ice crystals that first precipitates out as temperature is cooled is pure ice: $\mu^{\rm ice}_{\rm H_2O}(X^{\rm ice}_{\rm B}, T, P) \approx \tilde{\mu}^{\rm ice}_{\rm H_2O}(T, P)$.

The second thing to realize is that

$$\mu^{\rm liquid}_{\rm H_2O} \;\approx\; \tilde{\mu}^{\rm liquid}_{\rm H_2O}(T, P) + k_{\rm B}T \ln X^{\rm liquid}_{\rm H_2O} \tag{A.70}$$

If the $\approx$ in above is $=$, then it is an ideal solution. Raoult's law says that no matter what kind of solution (solid,liquid,gas), as long as the solutes become *dilute enough*, the *solvent* molecule's chemical potential approaches that in an ideal solution. This is in fact also true for the ice crystals, but $X^{\rm ice}_{\rm B}$ is so small that it's not going to have any effect on H_2O in ice. For the liquid, we have

$$\ln X^{\rm liquid}_{\rm H_2O} \;=\; \ln(1 - X^{\rm liquid}_{\rm B}) \;\approx\; -X^{\rm liquid}_{\rm B}. \tag{A.71}$$

So the chemical potential of water in liquid solution is lowered by $X^{\rm liquid}_{\rm B} k_{\rm B}T$ due to the presence of B in liquid. How much does that lower the melting point? (compared to what?)

$$\tilde{\mu}^{\rm liquid}_{\rm H_2O}(T, P) - k_{\rm B}T X^{\rm liquid}_{\rm B} \;=\; \tilde{\mu}^{\rm ice}_{\rm H_2O}(T, P) \tag{A.72}$$

Remember that $T^{\rm pure}_{\rm melt}$ is *defined* by

$$\tilde{\mu}^{\rm liquid}_{\rm H_2O}(T^{\rm pure}_{\rm melt}, P) \;=\; \tilde{\mu}^{\rm ice}_{\rm H_2O}(T^{\rm pure}_{\rm melt}, P). \tag{A.73}$$

Perform Taylor expansion with respect to T:

$$-\Delta s^{\rm pure}_{\rm melt}(T - T^{\rm pure}_{\rm melt}) \;=\; k_{\rm B}T X^{\rm liquid}_{\rm B}, \tag{A.74}$$

we get

$$T^{\rm pure}_{\rm melt} - T \;\approx\; \frac{k_{\rm B}T^{\rm pure}_{\rm melt}}{\Delta s^{\rm pure}_{\rm melt}} X^{\rm liquid}_{\rm B}. \tag{A.75}$$

The pure liquid with larger entropy of melting will have less relative melting point suppression (essentially steeper $\mu_i(T)$ will be less sensitive). What is interesting about (A.75) is that the potency of an antifreeze is independent of the chemical type of the antifreeze, at least when only a tiny amount of antifreeze is added. When the solution is very dilute, the stabilization of the *solvent* is entirely entropic.

Richard's rule: simple metals have $\Delta s^{\text{pure}}_{\text{melt}} \approx 1 - 2k_{\text{B}}$. Water has $\Delta s^{\text{pure}}_{\text{melt}} \approx 2.65k_{\text{B}}$.

Trouton's rule: $\Delta s^{\text{pure}}_{\text{evap}} \approx 10.5k_{\text{B}}$, for various kinds of liquids. Water has $\Delta s^{\text{pure}}_{\text{evap}} \approx 13.1k_{\text{B}}$.

Now consider the opposite limit (**b**): in this case, $g^\beta(X_2, T)$ is a needle around $X_2 \approx 1$. Then for a given T, $g^\beta(X_2^\beta, T) \approx \mu_2^\beta(X_2^\beta, T) \approx \tilde{\mu}_2^\beta(T)$, and we just need to solve

$$\mu_2^\alpha(X_2^\alpha, T) \;=\; \tilde{\mu}_2^\beta(T) \tag{A.76}$$

It can be shown mathematically, but is quite obvious visually, that the second equation $\mu_1^\alpha(X_2^\alpha, T) = \mu_1^\beta(X_2^\beta, T)$ for the solvent atoms becomes "unimportant" (still rigorously true, just that whether we solve it or not has little bearing on what we care about - one can draw a bunch of tangent extrapolations on $g^\beta(X_2^\beta)$ with slight differences in X_2^β, we can see huge changes in μ_1^β but little changes in μ_2^β, due to the vast difference in extrapolation distances - such equations are called "stiff" - stiff equations can make analytical approaches easier, but general numerical approaches more difficult). So we have effectively reduced to 1 unknown and 1 equation (or rather, we have decoupled a previously 2-unknowns-and-2-equations into two nearly indepedent 1-unknown-and-1-equations).

Suppose α=simple cubic, β=BCC. Suppose α phase can be described by regular solution with $\omega > 0$ (see Fig. 1.36 of [47], there is an eutectic phase diagram and $g^\alpha(X_2^\beta)$ bulges out in the middle):

$$\tilde{\mu}_2^\alpha(T) + k_{\text{B}}T \ln X_2^\alpha + \omega(1 - X_2^\alpha)^2 \;=\; \tilde{\mu}_2^\beta(T) \tag{A.77}$$

Rearranging the terms we get

$$X_2^\alpha \;=\; \exp\left(-\frac{\tilde{\mu}_2^\alpha(T) - \tilde{\mu}_2^\beta(T) + \omega(1 - X_2^\alpha)^2}{k_{\text{B}}T}\right) \tag{A.78}$$

The above can be solved iteratively. We first plug in $X_2^\alpha = 0$ on RHS, get a finite X_2^α on the LHS, then plug this new X_2^α to RHS and iterate. From the very first iteration, however, we

get

$$X_2^\alpha \;=\; \exp\left(-\frac{\tilde{\mu}_2^\alpha(T) - \tilde{\mu}_2^\beta(T) + \omega}{k_\mathrm{B}T}\right) \tag{A.79}$$

and if $Q(T) \equiv \tilde{\mu}_2^\alpha(T) - \tilde{\mu}_2^\beta(T) + \omega \gg k_\mathrm{B}T$, X_2^α would be small and then the first iteration would be close enough to convergence. $\tilde{\mu}_2^\alpha(T) - \tilde{\mu}_2^\beta(T)$ is how much more uncomfortable it is for a type-2 atom to be living in pure-2 α structure compared to pure-2 β structure. ω is still how much more uncomfortable it is for type-2 atom to be living among a vast sea of type-1 atoms rather than among its own kind (at 0K, $\tilde{\mu}_2^\alpha = -Z^\alpha\epsilon_{22}/2$, $\omega = Z^\alpha(-\epsilon_{12} + (\epsilon_{11} + \epsilon_{22})/2)$, so $\tilde{\mu}_2^\alpha + \omega = Z^\alpha(-\epsilon_{12}) - (-Z^\alpha\epsilon_{11}/2)$, which corresponds to the process of squeezing out a type-1 atom and placing it on a ridge, then inserting a type-2 atom into this sea of 1). Thus $Q(T)$ is an energy that can be interpreted as how much more uncomfortable it is to transfer a B atom from pure β phase to dilute α phase, excluding the configurational entropy of B in α phase. Exponential forms of the kind $e^{-Q/k_\mathrm{B}T}$ are called Boltzmann distribution in thermodynamics, and Arrhenius expression when one talks about rates in kinetics. It says that even though some places are (very) uncomfortable to be at or somethings are (very) difficult to do, there will always be some fraction of the population who will do those, because thermal fluctuations reward disorder and risk-taking. A prominent feature of these Boltzmann/Arrhenius forms, especially at low temperatures, is that $k_\mathrm{B}T$ in the denominator is a very violent term. A change in T by 100°C can conceivably cause many orders of magnitude change in the solubility.

The above train of thought can be extended to vacancies. A monatomic crystal made of type-A atoms, but with the possibility of "porosity" inside (non-occupancy of lattice sites), can be regarded as a fully dense A-B crystal with B identified as "Vacadium". In this case, $\epsilon_{\mathrm{BB}} = \epsilon_{\mathrm{AB}} = 0$, so $\omega = Z\epsilon_{\mathrm{AA}}/2$, i.e. it is enthalpically costly to mix Vacadium with A, and they would prefer to segregate if based entirely from enthalpy standpoint or at $T = 0$ K. However, entropically A and Vacadium would prefer to mix. When you mix a block of pure Vacadium (in β phase) with pure A in α (fully dense), the solubility of Vacadium in α would be $X_\mathrm{V} = e^{-Q/k_\mathrm{B}T}$. Also, when you are 100% Vacadium it does not matter what structure the Vacadium atoms are arranged, so $\tilde{\mu}_2^\alpha(T) - \tilde{\mu}_2^\beta(T) = 0$ thus $Q = \omega = Z\epsilon_{\mathrm{AA}}/2$. Q is called the *vacancy formation energy* in this context. Physically, Q is identified as the energy cost to extract an atom from lattice (break Z bonds) and attach it to an ledge on surface (form $Z/2$ bonds), in a Kossel crystal. In this class the above process is called the *canonical vacancy creation process*. The canonical vacancy creation process creates porosity inside the solid, making the solid appear *larger* than the fully dense state (social analogy would be "hype"). Note that the canonical vacancy creation process is not an *atomization* process, where one

extracts an atom and put it away to infinity.

An abstract view of phase transformation. Define order parameter η, which could be density, structure factor, magnetic moment, electric polarization, etc. η is a scalar of your choice that best reflects the nature of the problem (phase transition). The Gibbs free energy is defined as $G(N_1, N_2, ..., N_C, T, P; \eta)$. There are global minimum, metastable minima, and saddle point. For example, at low temperature, for pure iron, both $G(\eta_{\text{FCC}})$ and $G(\eta_{\text{BCC}})$ are local minima of $G(\eta)$, but $G(\eta_{\text{FCC}}) > G(\eta_{\text{BCC}})$. To go from $\eta_1 = \eta_{\text{FCC}}$ to $\eta_2 = \eta_{\text{BCC}}$, $G(\eta)$ must first go even higher than $G(\eta_1)$. This energy penalty is called the activation energy, and $\eta \in (\eta_1, \eta_2)$ is called the reaction coordinate. Define η^* to be the position of saddle point, we have

$$Q_{1\to 2} \;=\; G(\eta^*) - G(\eta_1), \quad Q_{2\to 1} \;=\; G(\eta^*) - G(\eta_2). \tag{A.80}$$

According to statistical mechanics, all possible states of η can exist, just with different probability. The rate of transition, if one is at η_1, to η_2, is given by:

$$R_{1\to 2} \;=\; \nu_0 \exp(-\frac{Q_{1\to 2}}{k_{\text{B}}T}), \tag{A.81}$$

where ν_0 is some attempt frequency (unit 1/s), corresponding to the oscillation frequency around η_1 (imagine a harmonic oscillator coupled to heat bath). The rate of transition, if one is already at η_2, to η_1, is given by:

$$R_{2\to 1} \;=\; \nu_0 \exp\left(-\frac{Q_{2\to 1}}{k_{\text{B}}T}\right). \tag{A.82}$$

If $G(\eta_1) > G(\eta_2)$, then $Q_{1\to 2} < Q_{2\to 1}$, and $R_{1\to 2} \gg R_{2\to 1}$ since Q's are in the exponential, and $Q_{2\to 1} - Q_{1\to 2} = G(\eta_1) - G(\eta_2)$ is proportional to the sample size.

One can also express η as function of position, $\eta(\mathbf{x})$, to represent an interface. Consider the condition when FCC is in equilibrium with BCC: $G(\eta_{\text{FCC}}) = G(\eta_{\text{BCC}})$, and there is an interface that separates them. $\eta(x)$ is then a sigmoid-like curve, with characteristic width defined as interfacial width. The interfacial energy arises because atoms in the interface are neither FCC or BCC, and have energy density higher than either of them. This would lead to a positive interfacial energy (Chap. 3)

The common tangent construction gives unique solution in composition when T, P is fixed. If T, P come into play, however, then the game is richer. The single-component Clausius-Clapeyron relation (A.18) can be generalized to C-component solutions. If we consider i in

α of composition $\mathbf{X}^\alpha \equiv [X_2^\alpha, ..., X_C^\alpha]$, or in β of composition $\mathbf{X}^\beta \equiv [X_2^\beta, ..., X_C^\beta]$, there needs to be

$$\mu_i^\alpha(\mathbf{X}^\alpha, T, P) \;=\; \mu_i^\beta(\mathbf{X}^\beta, T, P) \tag{A.83}$$

to maintain mass action equilibrium (chemical equilibrium), to make sure atom i is "equally happy" in α as in β. Let us investigate what dP/dT needs to be in order to maintain that way, if $\mathbf{X}^\alpha$ and $\mathbf{X}^\beta$ are fixed (for instance two "compound" phases, or one compound phase in contact with a large constant-composition reservoir): because we have

$$d\mu_i^\alpha \;=\; v_i^\alpha dP - s_i^\alpha dT, \quad d\mu_i^\beta \;=\; v_i^\beta dP - s_i^\beta dT. \tag{A.84}$$

To maintain (A.83), we need

$$\frac{dP}{dT} \;=\; \frac{s_i^\alpha - s_i^\beta}{v_i^\alpha - v_i^\beta} \;=\; \frac{h_i^\alpha - h_i^\beta}{T(v_i^\alpha - v_i^\beta)}, \tag{A.85}$$

the latter equality is because if α, β are already at chemical equilibrium for i at a certain (T, P), there is:

$$\mu_i^\alpha = h_i^\alpha - Ts_i^\alpha \;=\; \mu_i^\beta = h_i^\beta - Ts_i^\beta. \tag{A.86}$$

Consider for example, the equilibria between pure liquid water (β) and air (α): air is a solution. Then one has:

$$\frac{dP}{dT} \;\approx\; \frac{h_i^\alpha - h_i^\beta}{T(v_i^\alpha)} \tag{A.87}$$

since v_i^α is larger than v_i^β by a factor of 10^3. For the air solution $\mathbf{N} = (N_1, N_2, N_3, ...N_c)$, we have

$$V \;\approx\; \frac{Nk_\mathrm{B}T}{P} \quad \rightarrow \quad v_i \equiv \left.\frac{\partial V}{\partial N_i}\right|_{N_{j\neq i}, T, P} = \frac{k_\mathrm{B}T}{P}. \tag{A.88}$$

Thus

$$\frac{dP}{dT} \;\approx\; \frac{h_i^\alpha - h_i^\beta}{T(k_\mathrm{B}T/P)}, \quad \frac{d\ln P}{d(1/T^2)} \;\approx\; -\frac{\Delta h_i}{k_\mathrm{B}}. \tag{A.89}$$

So:

$$\ln\frac{P^\mathrm{eq}}{P_\mathrm{ref}^\mathrm{eq}} \;\approx\; \frac{\Delta h_i}{k_\mathrm{B}}\left(\frac{1}{T_\mathrm{ref}} - \frac{1}{T}\right), \tag{A.90}$$

when temperature is raised, the equilibrium vapor pressure goes up.

Notice that the gas phase always beats *all* condensed phases at low enough (but still positive) pressure. One can thus draw a $\ln P$-T diagram, and down under it is always the gas phase.

This is because chemical potential in the gas phase goes as

$$\mu_i^{\rm gas}(\mathbf{X}^{\rm gas}, T, P) \;\approx\; k_{\rm B}T \ln X_i P + \tilde{\mu}_i^{\rm gas}(T, 1{\rm atm}), \tag{A.91}$$

which goes to $-\infty$ as $P \to 0$, whereas chemical potentials in condensed phases are bounded. (The physical reason for going to $-\infty$ as $P \to 0$ is that the entropy of gas blows up as $k_{\rm B} \ln v$). Thus, all condensed phases (liquid,solid) become metastable at low enough pressure (see water phase diagram, Fig. A.1 (b)). Another way of saying it is that there always exists an equilibrium vapor pressure for any temperature and composition, which may be small but always positive, below which components in the liquid or solid solution would rather prefer to come out into the gas phase (volatility).

However, they are two manners by which vapor can come out. When you heat up a pot of water, at say 80°C, you can already feel vapor coming out if you stand over the pot, and maybe see some steam, but it's very peaceful *evaporation* process. However, when the temperature reaches 100°C, there is a very sharp transition. Suddenly there is a lot of commotion, and there is *boiling*. What *defines* the boiling transition?

The commotion is caused by the presence of *gas bubbles*, not present before T reaches $T_{\rm boil}$. The boiling transition is defined by $P^{\rm eq} = 1$ atm, the atmospheric pressure. Before $T < T_{\rm boil}$, there may be $P^{\rm eq} > P_{\rm H_2O}^{\rm ambient}$, so the water molecules would like to come out. But they can only come out from the gas-liquid interface, not inside the liquid, so the evaporation action is limited only to the water molecules in the narrow interfacial region $<$ 1nm. This is because any pure H_2O gas bubbles formed inside would be *crushed* by the hydrostatic pressure AND surface tension. But when $P^{\rm eq} > 1$ atm, pure H_2O gas bubbles can now nucleate **inside** the liquid. These bubbles nucleate, grow, and eventually rise up and break. At $T > T_{\rm boil}$ the whole body of liquid can join the action of phase transformation, not just the lucky few near the gas-liquid interface. Thermodynamically, there is nothing very special about the boiling transition, but if you look at the rate of water vapor coming out, there is a drastic upturn at $T = T_{\rm boil}$. So the boiling transition is a *transition in kinetics*. The availability of nucleation sites is important for such kinetic transitions. In the case of boiling, the nucleation sites are likely to be the container wall (watch a bottle of coke). Without the heterogeneous nucleation sites, it is possible to significantly superheat the liquid past its boiling point, without seeing the bubbles.

One can have superheating/supercooling because of the barriers to transformation. The

amount of thermodynamic driving force in a temperature-driven phase transformation is:

$$\Delta G \equiv \mu_i^{\alpha} - \mu_i^{\beta} \equiv \Delta\mu_i \approx \Delta s_i^{\mathrm{eq}} \Delta T = \frac{\Delta h_i^{\mathrm{eq}}}{T^{\mathrm{eq}}} \Delta T \tag{A.92}$$

if the reaction coordinate is identified as mass transfer from one phase to another (η_1 state: $N_i^{\alpha} + 1$ in α, N_i^{β} in β; η_2 state: N_i^{α} in α, $N_i^{\beta} + 1$ in β). To drive kinetics at a finite speed, the driving force (thermodynamic potential loss or dissipation) must be finite. (Chap. 2)

Appendix B

Spinodal Decomposition and Gradient Thermodynamics Description of the Interface

First-order phase transition is characterized by *finite jump* in the order parameter $\eta^\alpha \to \eta^\beta$ as soon as $T = T_e^\pm$ (the nucleation rate may be very small, but theoretically suppose one waits long enough one can witness this finite jump at $T_e^\pm$). For example, melting of ice at $P = 1$atm is a first-order transition because as soon as T rises up to 0.0001°C and melting can occur, there is a finite density change from ice to liquid water, and there is an *obvious* change in the viscosity as well. Also spatially, the transition from $\eta(\mathbf{x}) = \eta^\alpha$ to $\eta(\mathbf{x}') = \eta^\beta$ typically occurs over a *very narrow region*: the shortest distance between $\mathbf{x}$ and $\mathbf{x}'$ (interfacial thickness w) is *typically* less than 1nm. Previously, we assigned a capillary energy γ to this interfacial region without discussing this region's detailed *structure*. Such "sharp interface" view, where one ignores the detailed interfacial structure and represent it as a geometric dividing surface, is sufficient for most first-order phase transition problems. If one is really interested in the physical thickness of this interfacial region however, one must use so-called gradient thermodynamics formulation [88] to be introduced below, where the capillary energy $\int \gamma dA$ in the sharp-interface representation is replaced by a 3D integral involving a gradient squared term $\int K|\nabla\eta(\mathbf{x})|^2 d^3\mathbf{x}$ with $K > 0$. The above replacement makes sense intuitively, since the interfacial region is characterized by large gradients in $\eta(\mathbf{x})$, absent in the homogeneous bulk regions of α or β. Nucleation and growth is a must for all first-order phase transitions, where *large change* ($\eta^\alpha \to \eta^\beta$) occurs in a *narrow region*

(the interface) even during nucleation.

In contrast, second-order phase transition is characterized by *initially infinitesimal changes* over a *wide region*. These initially infinitesimal changes appear spontaneously in the system and grow with time, without going through a nucleation (large change in a small region) stage. For example, in the paramagnetic (α)→ferromagnetic ($\alpha1$,$\alpha2$) transition of pure iron as T is cooled below $T_c = 1043$K (the Curie temperature, also called the critical point), both the spin-down $\alpha1$ and the spin-up $\alpha2$ phase have very small magnetic moments: $\eta^{\alpha1} = -m$, $\eta^{\alpha2} = m$, with $m \propto (T_c - T)^{1/2}$. Microscopically, going from $\alpha1$ to $\alpha2$ near T_c would involve the flipping of a very small number of spins. So the high-temperature paramagnetic phase, and the two low-temperature ferromagnetic phases are very similar to each other near T_c: $|\eta^{\alpha} - \eta^{\alpha1}|, |\eta^{\alpha} - \eta^{\alpha2}| \propto (T_c - T)^{1/2}$, where η is the magnetic moment. The breakup of a uniform paramagnetic domain into multiple ferromagnetic domains upon a drop in temperature below T_c is spontaneous and instantaneous and *does not require* a nucleation stage: it is growth, off the bat. In other words, no under-cooling is required for observing *the start* of second-order phase transition within a given observation period. The growth happens essentially instantaneously at $T = T_c^{\pm}$. Although, to see the growth and coarsening to a certain amplitude would require time.

The way a system can accomplish second-order transition vis-à-vis first-order transition is best illustrated using the binary solution example: $g_{\text{soln}}(X_2, T) \equiv G_{\text{soln}}(N_1, N_2, T)/(N_1+N_2)$. Suppose $\Omega_1 = \Omega_2 = \Omega$, we may define specific volume free energy as

$$g_v(c_2) \;\equiv\; \Omega^{-1} g_{\text{soln}}(X_2 = c_2\Omega) \tag{B.1}$$

so the bulk solution free energy for a homogeneous system is just

$$G_{\text{soln}} \;=\; \left(\int d^3\mathbf{x}\right) g_v(c_2). \tag{B.2}$$

$g_v(c_2)$ is the same function as $g_{\text{soln}}(X_2)$ after horizontal and vertical scaling. So the tangent extrapolation of $g_v(c_2)$ to $c_2 = 0$ (corresponding to $\mathbf{X} = \mathbf{p}_1$) would give $\Omega^{-1}\mu_1$, and tangent extrapolation of $g_v(c_2)$ to $c_2 = \Omega^{-1}$ (corresponding to $\mathbf{X} = \mathbf{p}_2$) would give $\Omega^{-1}\mu_2$. $c_2(\mathbf{x})$ is our order parameter field $\eta(\mathbf{x})$ here. For an inhomogeneous system, the solution free energy should intuitively be written as

$$G_{\text{soln}} \;=\; \int d^3\mathbf{x} g_v(c_2(\mathbf{x})). \tag{B.3}$$

Using the above as reference, the total free energy then looks like:

$$G \;=\; \int d^3\mathbf{x}(g_v(c_2(\mathbf{x})) + K|\nabla c_2(\mathbf{x})|^2) + G_{\text{elastic}} \tag{B.4}$$

where the gradient squared term replaces the capillary energy $\int \gamma dA$. $G_{\text{elastic}} = 0$ if $\Omega_1 = \Omega_2 = \Omega$. (B.4) is a unified model that can be used to investigate both finite interfacial thickness in first-order transitions [88], as well as second-order transitions [92]. Since $K > 0$, the model (B.4) punishes sharp spatial gradients, the origin of interfacial energy. On the other hand if all changes occur smoothly over a large wavelength with small spatial gradients, then G approaches G_{soln}. Since G_{soln} is the driver of phase transformation (gradient/capillary and elastic energies are typically positive), let us consider what G_{soln} wants to do first.

For a closed system, c_2 is *conserved*:

$$\int d^3\mathbf{x}c_2(\mathbf{x}) \;=\; \text{const} \tag{B.5}$$

which means it is possible to *partition* the solutes, but it is not possible to change the total amount of solutes in the entire system. For instance, if one starts out with a uniform concentration $c_2(\mathbf{x}) = c_2^\alpha$, a partition may roughly speaking occur as:

$$c_2^\alpha \;=\; f^{\alpha 1}c_2^{\alpha 1} + f^{\alpha 2}c_2^{\alpha 2}, \tag{B.6}$$

where volume fraction

$$f^{\alpha 1} \;=\; \frac{c_2^{\alpha 2} - c_2^\alpha}{c_2^{\alpha 2} - c_2^{\alpha 1}}, \quad f^{\alpha 2} \;=\; 1 - f^{\alpha 1} = \frac{c_2^\alpha - c_2^{\alpha 1}}{c_2^{\alpha 2} - c_2^{\alpha 1}} \tag{B.7}$$

of the region has $c_2(\mathbf{x}) = c_2^{\alpha 1}$ and $c_2(\mathbf{x}) = c_2^{\alpha 2}$, respectively, separated by sharp interfaces. The solution free energy of the *partitioned* system is then

$$G_{\text{soln}} \;=\; \left(\int d^3\mathbf{x}\right)(f^{\alpha 1}g_v(c_2^{\alpha 1}) + f^{\alpha 2}g_v(c_2^{\alpha 2})) \tag{B.8}$$

compared to the unpartitioned and uniform original system $(\int d^3\mathbf{x})\, g_v(c_2^\alpha)$.

Local stability means G_{soln} is stable against small perturbations in $c_2(\mathbf{x})$. The necessary and sufficient condition for local stability is that

$$\frac{\partial^2 g_v}{\partial c_2^2} \;>\; 0. \tag{B.9}$$

If $\frac{\partial^2 g_v}{\partial c_2^2} < 0$, a small partition with $c_2^{\alpha 1} \approx c_2^\alpha \approx c_2^{\alpha 2}$ would be able to decrease G_{soln}. For example, with $c_2^{\alpha 2} = c_2^\alpha + \Delta c$, $c_2^{\alpha 1} = c_2^\alpha - \Delta c$, $f^{\alpha 1} = f^{\alpha 2} = 1/2$, one has

$$\frac{G_{\text{soln}}}{\int d^3\mathbf{x}} \;=\; \frac{1}{2} g_v(c_2^\alpha - \Delta c) + \frac{1}{2} g_v(c_2^\alpha + \Delta c) \;=\; g_v(c_2^\alpha) + \frac{1}{2}\frac{\partial^2 g_v}{\partial c_2^2}(c_2^\alpha)(\Delta c)^2 + ... \tag{B.10}$$

which would be lower than uniform $g_v(c_2^\alpha)$ if $\frac{\partial^2 g_v}{\partial c_2^2} < 0$. A sinusoidal perturbation

$$c_2(\mathbf{x}) \;=\; c_2^\alpha + a(t)\sin(\mathbf{k}\cdot\mathbf{x}) \tag{B.11}$$

would also have equal amount of "ups and downs", and would thus also reduce G_{soln}. The reason sinusoidal perturbation is preferred (at least initially) compared to the step function between $c_2^\alpha - \Delta c$ and $c_2^\alpha + \Delta c$ is that it minimizes the gradient energy by spreading the gradients around. Therefore if $\frac{\partial^2 g_v}{\partial c_2^2} < 0$, its amplitude $a(t)$ will increase with time. This is the trick behind spinodal decomposition, or more generally second-order phase transitions, which can reduce the system free energy without nucleation. *Nucleation is not needed here because the system's initial state does not have local stability.* The loss of local stability is induced by temperature, i.e.

$$\frac{\partial^2 g_v}{\partial c_2^2}(c_2^\alpha, T_{\text{C}}^+) \;>\; 0, \quad \frac{\partial^2 g_v}{\partial c_2^2}(c_2^\alpha, T_{\text{C}}^-) \;<\; 0 \tag{B.12}$$

thus

$$\frac{\partial^2 g_v}{\partial c_2^2}(c_2^\alpha, T_{\text{C}}) \;=\; 0. \tag{B.13}$$

During initial growth of the sinusoidal profile in the unstable composition range, the solutes appears to diffuse *up* the concentration gradient (Fig. 5.39 of [47]). According to the phenomenological Fick's 1st law $\mathbf{J}_2 = -\tilde{D}\nabla c_2$, this would mean a negative interdiffusivity $\tilde{D}(c_2) < 0$. This is in fact not surprising, because $\tilde{D}$ (from D_1, D_2) contains thermodynamic factor $1 + \frac{d\ln\gamma_2}{d\ln c_2}$, which can be shown to be $\frac{X_2(1-X_2)}{k_{\text{B}}T}\frac{\partial^2 g}{\partial X_2^2}$ and thus have the same sign as $\frac{\partial^2 g}{\partial X_2^2}$. When $\frac{\partial^2 g}{\partial X_2^2}$ is negative, $\tilde{D}$ is negative. This means that at the most fundamental level, diffusion is driven by the desire to reduce free energy or chemical potential, and *not* by the desire to smear out the concentration gradient.

Mathematically, while a positive diffusivity tends to smear out the profile (the shorter the wavelength, the faster the decay of the Fourier component amplitude), a negative diffusivity would tend to increase the roughness of the profile. The growth of very-small wavelength fluctuations in spinodal decomposition will be punished by the gradient energy, though.

Thus an optimal wavelength will be selected initially, which can be tens of nms. Later, after the compositions have deviated largely from c_2^α, the microstructural lengthscale may further coarsen, although the interfacial lengthscale will sharpen. Because $\alpha 1$ and $\alpha 2$ do not come out of a nucleation and growth process, but amplification of sinusoidal waves of certain optimal wavelength, they lead to unique-looking interpenetrating microstructures.

In contrast to spinodal instability, *in a first-order phase transition the system's initial state has never lost its local stability.* At $T = T_e^+$, one is in a *globally stable* uniform composition, which means

$$g_v(c_2^\alpha, T_e^+) \; < \; f^{\alpha 1} g_v(c_2^{\alpha 1}, T_e^+) + f^{\alpha 2} g_v(c_2^{\alpha 2}, T_e^+) \tag{B.14}$$

for small and large deviations $|c_2^{\alpha 2} - c_2^\alpha|$ alike (thus a globally stable system must be locally stable, but not vice versa). Then at $T = T_e^-$, $c_2(\mathbf{x}) = c_2^\alpha$ becomes locally stable *only*, which means small deviations would still induce the system energy to go up, but large deviations may induce the system energy to go down. *Thus, small perturbations like (B.11) would decay and die, but large enough perturbations may survive.* The chance survival of large enough perturbations/fluctuations in the order-parameter field is just nucleation.

(B.4) can be used to estimate interfacial thickness w in the following manner. Since $\nabla c_2 \propto (c_2^\beta - c_2^\alpha)/w$ inside the interface, the gradient energy integral scales as $K(c_2^\beta - c_2^\alpha)^2/w$, so the wider the interface the better for the gradient energy. On the other hand, right at $T = T_e$, $g_v(c_2)$ of the first term connects two energy-degenerate states $g_v(c_2 = c_2^\beta) = g_v(c_2 = c_2^\alpha)$, with a bump $g_v^* - g_v(c_2^\alpha)$ in between. The solution free energy first term thus gives an excess $\propto (g_v^* - g_v(c_2^\alpha))w$, that punishes wide interfaces. The best compromised is thus reached at $w \propto K^{1/2}|c_2^\beta - c_2^\alpha|(g_v^* - g_v(c_2^\alpha))^{-1/2}$, with interfacial energy $\gamma \propto K^{1/2}|c_2^\beta - c_2^\alpha|(g_v^* - g_v(c_2^\alpha))^{1/2}$. It turns out that for T_e near T_c, $|c_2^\beta - c_2^\alpha| \propto (\Delta T)^{1/2}$, where $\Delta T = T_c - T_e$, and $g_v^* - g_v(c_2^\alpha) \propto (\Delta T)^2$, so the interfacial width near the critical temperature would diverge as $(\Delta T)^{-1/2}$, and the interfacial energy would vanish as $(\Delta T)^{3/2}$ [88].

Science advances greatly when two seemingly different concepts are connected, for instance the Einstein relation $M = D/k_\mathrm{B}T$. Cahn and Hilliard made a similar contribution when they connected interfacial energy to critical temperature and second-order phase transformation. Based on the insight that gradient term should be added to thermodynamic field theories (fundamentally this is because of atomic discreteness), they developed gradient thermodynamics formalism for chemical solution systems that predict finite interfacial width, interfacial energy, as well as wavelength selection in spinodal decomposition [92], under one unified framework. The development can in fact be traced back to the work of van der Waals for

single-component systems, using density as order parameter[93]. Another offshoot of this approach was provided by Ginzburg and Landau in the theory of superconductivity.

Finally, if $\Omega_1 \neq \Omega_2$ the 1-rich $\alpha 1$ phase and 2-rich $\alpha 2$ will have different stress-free volumes, and to accommodate this mismatch coherently would involve finite elastic energy $G_{\text{elastic}} > 0$. Growth of the sinusoidal concentration wave would require growth of the associated transformation strain wave. This would delay the onset of the spinodal instability.

Bibliography

[1] Robert E. Reed-Hill and Reza Abbaschian. *Physical Metallurgy Principles*. PWS, Boston, third edition, 1992.

[2] T. Zhu and J. Li. Ultra-strength materials. *Prog. Mater. Sci.*, 55:710–757, 2010.

[3] Y. Mishin, M. J. Mehl, D. A. Papaconstantopoulos, A. F. Voter, and J. D. Kress. Structural stability and lattice defects in copper: Ab initio, tight-binding, and embedded-atom calculations. *Phys. Rev. B*, 6322:224106, 2001.

[4] John F. Nye. *Physical properties of crystals: their representation by tensors and matrices*. Clarendon Press, Oxford, 1957.

[5] B. L. Adams, S. I. Wright, and K. Kunze. Orientation imaging - the emergence of a new microscopy. *Metall Trans A*, 24:819–831, 1993.

[6] R. D. Doherty, D. A. Hughes, F. J. Humphreys, J. J. Jonas, D. J. Jensen, M. E. Kassner, W. E. King, T. R. McNelley, H. J. McQueen, and A. D. Rollett. Current issues in recrystallization: a review. *Mater. Sci. Eng. A-Struct. Mater. Prop. Microstruct. Process.*, 238:219–274, 1997.

[7] F. J. Humphreys. Review - grain and subgrain characterisation by electron backscatter diffraction. *J. Mater. Sci.*, 36:3833–3854, 2001.

[8] C. A. Schuh, M. Kumar, and W. E. King. Analysis of grain boundary networks and their evolution during grain boundary engineering. *Acta Mater.*, 51:687–700, 2003.

[9] D. D. Macdonald. Passivity - the key to our metals-based civilization. *Pure Appl. Chem.*, 71:951–978, 1999.

[10] G. S. Frankel. Pitting corrosion of metals - a review of the critical factors. *J. Electrochem. Soc.*, 145:2186–2198, 1998.

[11] M. S. Daw and M. I. Baskes. Embedded-atom method - derivation and application to impurities, surfaces, and other defects in metals. *Phys. Rev. B*, 29:6443–6453, 1984.

[12] M. W. Finnis and J. E. Sinclair. A simple empirical n-body potential for transition-metals. *Philos. Mag. A-Phys. Condens. Matter Struct. Defect Mech. Prop.*, 50:45–55, 1984.

[13] S. Ogata, J. Li, and S. Yip. Energy landscape of deformation twinning in bcc and fcc metals. *Phys. Rev. B*, 71:224102, 2005.

[14] M. I. Baskes. Modified embedded-atom potentials for cubic materials and impurities. *Phys. Rev. B*, 46:2727–2742, 1992.

[15] S. Ogata, J. Li, N. Hirosaki, Y. Shibutani, and S. Yip. Ideal shear strain of metals and ceramics. *Phys. Rev. B*, 70:104104, 2004.

[16] S. Ogata and J. Li. Toughness scale from first principles. *J. Appl. Phys.*, 106:113534, 2009.

[17] J. Frenkel. The theory of the elastic limit and the solidity of crystal bodies. *Z. Phys.*, 37:572–609, 1926.

[18] S. Ogata, J. Li, and S. Yip. Ideal pure shear strength of aluminum and copper. *Science*, 298:807–811, 2002.

[19] A. J. Foreman, M. A. Jaswon, and J. K. Wood. Factors controlling dislocation widths. *Proc Phys Soc Lond A*, 64:156–163, 1951.

[20] G. I. Taylor. The mechanism of plastic deformation of crystals. part i.-theoretical; part ii.-comparison with observations. *Proc. R. Soc. Lond. A*, 145:362–404, 1934.

[21] E. Orowan. Zur kristallplastizitat. i. tieftemperaturplastizitt und beckersche formel; ii. die dynamische auffassung der kristallplastizitat; iii. uber den mechanismus des gleitvorganges. *Z. Phys.*, 89:605–659, 1934.

[22] M. Polanyi. Uber eine art gitterstorung, die einen kristall plastisch machen konnte. *Z. Phys.*, 89:660–664, 1934.

[23] J. P. Hirth. A brief history of dislocation theory. *Met Trans A-Phys Met Mater Sc*, 16:2085–2090, 1985.

[24] A. Timpe. *Diss. Gottingen (Leipzig, 1905); Z. Math. Phys.*, 52:348, 1905.

[25] Vito Volterra. Sur l'quilibre des corps lastiques multiplement connexes. *Annales scientifiques de l'cole Normale Suprieure*, 24:401–517, 1907.

[26] A. R. Bausch, M. J. Bowick, A. Cacciuto, A. D. Dinsmore, M. F. Hsu, D. R. Nelson, M. G. Nikolaides, A. Travesset, and D. A. Weitz. Grain boundary scars and spherical crystallography. *Science*, 299:1716–1718, 2003.

[27] W. T. M. Irvine, V. Vitelli, and P. M. Chaikin. Pleats in crystals on curved surfaces. *Nature*, 468:947–951, 2010.

[28] P. E. Cladis and M. Kleman. Non-singular disclinations of strength s = + 1 in nematics. *J Phys-Paris*, 33:591–&, 1972.

[29] P.M. Chaikin and T.C. Lubensky. *Principles of condensed matter physics*. Cambridge University Press, Cambridge, 1994.

[30] P. B. Hirsch, R. W. Horne, and M. J. Whelan. Direct observations of the arrangement and motion of dislocations in aluminum. *Philos Mag*, 1:677–&, 1956.

[31] R. Peierls. The size of a dislocation. *Proc. Phys. Soc*, 52:34–37, 1940.

[32] F. R. N. Nabarro. Dislocations in a simple cubic lattice. *Proc Phys Soc Lond*, 59:256–272, 1947.

[33] J.P. Hirth and J. Lothe. *Theory of dislocations*. Wiley, New York, second edition, 1982.

[34] J. Li, C. Z. Wang, J. P. Chang, W. Cai, V. V. Bulatov, K. M. Ho, and S. Yip. Core energy and peierls stress of a screw dislocation in bcc molybdenum: A periodic-cell tight-binding study. *Phys. Rev. B*, 70:104113, 2004.

[35] M. J. Bierman, Y. K. A. Lau, A. V. Kvit, A. L. Schmitt, and S. Jin. Dislocation-driven nanowire growth and eshelby twist. *Science*, 320:1060–1063, 2008.

[36] H. Mughrabi. Deformation-induced long-range internal stresses and lattice plane misorientations and the role of geometrically necessary dislocations. *Philos. Mag.*, 86:4037–4054, 2006.

[37] A.A. Griffith. The phenomena of rupture and flow in solids. *Philos. Trans. R. Soc. London A*, 221:163–198, 1920.

[38] J. H. Rose, J. Ferrante, and J. R. Smith. Universal binding-energy curves for metals and bimetallic interfaces. *Phys. Rev. Lett.*, 47:675–678, 1981.

[39] J. Li, Z-W. Shan, and E. Ma. Elastic strain engineering for unprecedented materials properties. *MRS Bulletin*, 39:108–114, 2014.

[40] H. Verweij, M. C. Schillo, and J. Li. Fast mass transport through carbon nanotube membranes. *Small*, 3:1996–2004, 2007.

[41] Robert W. Balluffi, Samuel M. Allen, and W. Craig Carter. *Kinetics of Materials*. Wiley, New York, 2005.

[42] B. C. Regan, S. Aloni, R. O. Ritchie, U. Dahmen, and A. Zettl. Carbon nanotubes as nanoscale mass conveyors. *Nature*, 428:924–927, 2004.

[43] R. O. Simmons and R. W. Balluffi. Measurements of equilibrium vacancy concentrations in aluminum. *Phys. Rev.*, 117:52–61, 1960.

[44] C. Herring. Diffusional viscosity of a polycrystalline solid. *J. Appl. Phys.*, 21:437–445, 1950.

[45] A. D. Smigelskas and E. O. Kirkendall. Zinc diffusion in alpha-brass. *Tran. Amer. Inst. Min. Met. Eng.*, 171:130–142, 1947.

[46] Y. D. Yin, R. M. Rioux, C. K. Erdonmez, S. Hughes, G. A. Somorjai, and A. P. Alivisatos. Formation of hollow nanocrystals through the nanoscale kirkendall effect. *Science*, 304:711–714, 2004.

[47] David A. Porter and Kenneth E. Easterling. *Phase transformations in metals and alloys*. Chapman & Hall, London, second edition, 1992.

[48] R. Mills. Self-diffusion in normal and heavy-water in range 1-45 degrees. *Journal Of Physical Chemistry*, 77:685–688, 1973.

[49] D. Topgaard and A. Pines. Self-diffusion measurements with chemical shift resolution in inhomogeneous magnetic fields. *J. Magn. Reson.*, 168:31–35, 2004.

[50] M. Kilo, C. Argirusis, G. Borchardt, and R. A. Jackson. Oxygen diffusion in yttria stabilised zirconia - experimental results and molecular dynamics calculations. *Phys. Chem. Chem. Phys.*, 5:2219–2224, 2003.

[51] M. Reiner. The deborah number. *Phys. Today*, 17:62–62, 1964.

[52] J. R. Sambles. Electron microscope study of evaporating gold particles - kelvin equation for liquid gold and lowering of melting point of solid gold particles. *Proc. R. Soc. London A*, 324:339–351, 1971.

[53] P. Buffat and J. P. Borel. Size effect on melting temperature of gold particles. *Phys. Rev. A*, 13:2287–2298, 1976.

[54] J. H. Shim, B. J. Lee, and Y. W. Cho. Thermal stability of unsupported gold nanoparticle: a molecular dynamics study. *Surf. Sci.*, 512:262–268, 2002.

[55] Z. L. Wang, J. M. Petroski, T. C. Green, and M. A. El-Sayed. Shape transformation and surface melting of cubic and tetrahedral platinum nanocrystals. *J. Phys. Chem. B*, 102:6145–6151, 1998.

[56] B. Franklin. Of the stilling of waves by means of oil. extracted from sundry letters between benjamin franklin, ll. d. f. r. s. william brownrigg, m. d. f. r. s. and the reverend mr. farish. *Philosophical Transactions*, 64:445–460, 1774.

[57] W. T. Read and W. Shockley. Dislocation models of crystal grain boundaries. *Phys. Rev.*, 78:275–289, 1950.

[58] J. H. Vandermerwe. On the stresses and energies associated with inter-crystalline boundaries. *Proc. Phys. Soc. London A*, 63:616–637, 1950.

[59] F. R. N. Nabarro. The influence of elastic strain on the shape of particles segregating in an alloy. *Proc. Phys. Soc*, 52:90–93, 1940.

[60] M. S. Wechsler, D. S. Lieberman, and T. A. Read. On the theory of the formation of martensite. *Tran. Amer. Inst. Min. Met. Eng.*, 197:1503–1515, 1953.

[61] S. J. Hao, L. S. Cui, D. Q. Jiang, X. D. Han, Y. Ren, J. Jiang, Y. N. Liu, Z. Y. Liu, S. C. Mao, Y. D. Wang, Y. Li, X. B. Ren, X. D. Ding, S. Wang, C. Yu, X. B. Shi, M. S. Du, F. Yang, Y. J. Zheng, Z. Zhang, X. D. Li, D. E. Brown, and J. Li. A transforming metal nanocomposite with large elastic strain, low modulus, and high strength. *Science*, 339:1191–1194, 2013.

[62] Z. T. Trautt, M. Upmanyu, and A. Karma. Interface mobility from interface random walk. *Science*, 314:632–635, 2006.

[63] R. D. MacPherson and D. J. Srolovitz. The von neumann relation generalized to coarsening of three-dimensional microstructures. *Nature*, 446:1053–1055, 2007.

[64] D. Turnbull and R. E. Cech. Microscopic observation of the solidification of small metal droplets. *J. Appl. Phys.*, 21:804–810, 1950.

[65] J. Li. The mechanics and physics of defect nucleation. *MRS Bull.*, 32:151–159, 2007.

[66] American Society for Metals. *Phase Transformations*. American Society for Metals, Metals Park, OH, 1970.

[67] John Wyrill Christian. *The Theory of Transformations in Metals and Alloys*. Elsevier, Amsterdam, third edition, 2002.

[68] W. A. Johnson and R. F. Mehl. Reaction kinetics in processes of nucleation and growth. *Trans. Am. Inst. Min. Metall. Eng.*, 135:416–442, 1939.

[69] M. Avrami. Kinetics of phase change i - general theory. *J. Chem. Phys.*, 7:1103–1112, 1939.

[70] M. Avrami. Granulation, phase change, and microstructure - kinetics of phase change. iii. *J. Chem. Phys.*, 9:177–184, 1941.

[71] A. N. Kolmogorov. On statistical theory of metal crystallization. *Izv. Akad. Nauk SSSR (Ser: Matem.)*, 3:355–359, 1937.

[72] S. H. Lee, Y. Jung, and R. Agarwal. Highly scalable non-volatile and ultra-lowpower phase-change nanowire memory. *Nature Nanotechnology*, 2:626–630, 2007.

[73] Louis A. Girifalco. *Dynamics of Technological Change*. Van Nostrand Reinhold, New York, 1991.

[74] I. M. Lifshitz and V. V. Slyozov. The kinetics of precipitation from supersaturated solid solutions. *J. Phys. Chem. Solids*, 19:35–50, 1961.

[75] C. Wagner. Theorie der alterung von niederschlagen durch umlosen (ostwald-reifung). *Z. Elektrochem.*, 65:581–591, 1961.

[76] Jonathan A. Dantzig and Charles L. Tucker. *Modeling in materials processing*. Cambridge University Press, Cambridge, 2001.

[77] W. G. Pfann. Principles of zone-melting. *Tran. Amer. Inst. Min. Met. Eng.*, 194:747–753, 1952.

[78] Bruce Chalmers. *Principles of solidification*. Wiley, New York, 1964.

[79] W. W. Mullins and R. F. Sekerka. Morphological stability of a particle growing by diffusion or heat flow. *J. Appl. Phys.*, 34:323–329, 1963.

[80] W. W. Mullins and R. F. Sekerka. Stability of planar interface during solidification of dilute binary alloy. *J. Appl. Phys.*, 35:444–451, 1964.

[81] Gary S. Was. *Fundamentals of Radiation Materials Science: Metals and Alloys*. Springer, 2007.

[82] T. S. Ke. Experimental evidence of the viscous behavior of grain boundaries in metals. *Phys Rev*, 71:533–546, 1947.

[83] Terrell L. Hill. *An Introduction to Statistical Thermodynamics*. Dover, New York, 1987.

[84] Donald A. McQuarrie. *Statistical Mechanics*. Harper & Row, New York, 1976.

[85] A. van de Walle and G. Ceder. Automating first-principles phase diagram calculations. *J. Phase Equilib.*, 23:348–359, 2002.

[86] W. L. Bragg and E. J. Williams. The effect of thermal agitation on atomic arrangement in alloys. *Proc. R. Soc. Lond. A*, 145:699–730, 1934.

[87] W. L. Bragg and E. J. Williams. The effect of thermal agitation on atomic arrangement in alloys - ii. *Proc. R. Soc. Lond. A*, 151:0540–0566, 1935.

[88] J. W. Cahn and J. E. Hilliard. Free energy of a nonuniform system .1. interfacial free energy. *J. Chem. Phys.*, 28:258–267, 1958.

[89] S. M. Allen and J. W. Cahn. Microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. *Acta Met*, 27:1085–1095, 1979.

[90] Claude H.P. Lupis. *Chemical thermodynamics of materials*. North-Holland, New York, 1983.

[91] Mats Hillert. *Phase Equilibria, Phase Diagrams and Phase Transformations: Their Thermodynamic Basis*. Cambridge University Press, New York, 1998.

[92] J. W. Cahn. On spinodal decomposition. *Acta. Met.*, 9:795–801, 1961.

[93] Johannes Diderik van der Waals. The thermodynamik theory of capillarity under the hypothesis of a continuous variation of density. *Konink. Akad. Weten. Amsterdam*, 1:56, 1893.