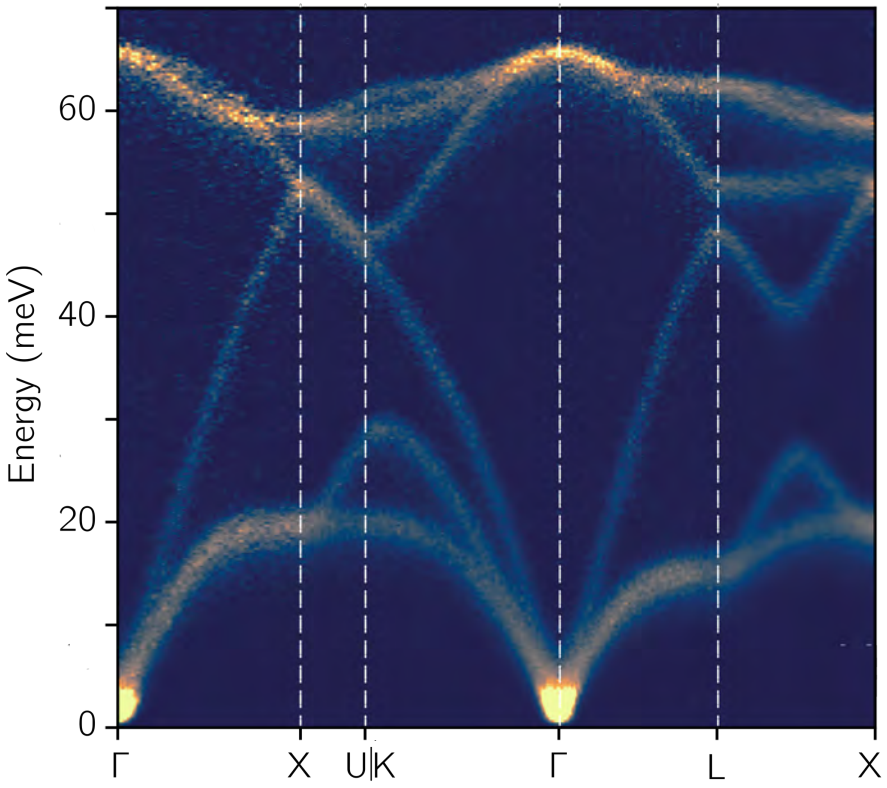
Inelastic Scattering



Revision 1.0
5 Mar. 2020

Inelastic Scattering

Brent Fultz
Barbara and Stanley Rawn, Jr., Professor of
Materials Science and Applied Physics
with *Tim Kelley, Mike McKerns, Jiao Lin, JaeDong Lee,*
Hillary Smith, Olivier Delaire

Dept. Applied Physics and Materials Science, $\langle \mathrm{AP}|\hat{h}|\mathrm{MS}\rangle$
California Institute of Technology
Pasadena, California 91125 USA

Typesetting and production by the author.
Cover design by the author.

This book is dedicated to the new user – may you benefit from our mistakes.

Preface

The bright future has arrived for inelastic scattering studies of materials, molecules, and condensed matter. The international science community and national science agencies have invested heavily in new instruments and new sources of neutrons and x-rays that are overcoming many of the historical limitations of inelastic scattering research. The newest generation of transmission electron microscopes offers highly monochromatic electrons that enable measurements of thermal excitations that would otherwise be buried beneath the elastic line. A new family of inelastic laser scattering experiments are now possible, such as those that couple vibrations of the cavity itself to photons in from laser.

Higher intensities and instrument sensitivities are major and obvious improvements. Experimental *inelastic* scattering has been particularly constrained by low countrates, forcing experimental compromises in energy resolution and intensity. The ARCS inelastic neutron spectrometer with its location at the high-power target station of the Spallation Neutron Source and with its high detection efficiency, for example, provides unprecedented experimental productivity, overcoming many of the restrictions caused by the low countrates that have accompanied inelastic neutron scattering experiments to date.

As features in the data from inelastic scattering experiments become more visible, and averaging processes are less necessary, more information can be extracted from the experimental measurements. The goal of this book is to describe the underlying scattering physics and dynamic processes in materials, and show opportunities to elevate the level of science done with inelastic spectrometers. A large body of specialized knowledge is required to design modern experiments for inelastic instruments such as time-of-flight chopper spectrometers. This body of knowledge will only grow as new capabilities become available. Unfortunately, the underlying concepts are scattered over many disciplines. For example, the books on the theory of thermal neutron scattering by S. W. Lovesey and G. L. Squires are superb. Similarly, excellent solid-state physics texts by J. M. Ziman, C. Kittel, U. Rössler, and N. W. Ashcroft and N. D. Mermin are available for understanding the principles of excitations in condensed matter. Today the concepts from neutron scattering and condensed matter physics are not connected well by existing written texts. It is a challenge to organize a coherent presentation of this wide body of knowledge, connected by the needs of experimental inelastic scattering.

This document did not begin as a textbook. The original draft was a manual

of specifications for the ARCS data analysis software. Defining specifications is a major step in planning a software project and setting its scope. Writing a manual of specifications forces a high degree of detailed planning of classes and modules. As we struggled with these details on software structure, it became obvious that they should parallel as closely as possible the science and practice of experimental inelastic neutron scattering research. Besides the challenge of organizing the higher-level intellectual concepts, practical problems with notation became apparent almost immediately. (Should the scattering vector be $\vec{Q}$, $\vec{\kappa}$ or $\Delta\vec{k}$? What about its sign?)

The present book is intended for a spectrum of readers spanning from graduate students beginning their doctoral research in inelastic neutron scattering, researchers who need to learn how to use chopper spectrometers and their data analysis, and ourselves, the authors, who need a reference manual. Our heartfelt concern, however, is for the graduate student who enters the field of inelastic neutron scattering with no experience with instruments and only a sketchy understanding of the scientific principles. This text was designed to help the reader understand the principles and methods of inelastic scattering, and do so efficiently.

Explanations are presented at the minimum level of detail required to understand physical concepts. There is some sacrifice of the care in development found in *The Theory of Neutron Scattering from Condensed Matter* by S. W. Lovesey and *Introduction to the Theory of Inelastic Neutron Scattering* by G. L. Squires, but our goal was to provide explanations that are "best buys," producing the most physical insight for the amount of intellectual effort required to understand them. Another goal was to present the field of inelastic scattering as a codified intellectual discipline, showing inter-relationships between different topics. To do so, the notation from other books has been altered in places, for example $\vec{Q}$, defined as $\vec{k}_\mathrm{i} - \vec{k}_\mathrm{f}$ was selected for the scattering vector so that κ could be consistent with its usage in *The Theory of Lattice Dynamics in the Harmonic Approximation* by A. A. Maradudin, et al. (It is an unfortunate, but well-established convention that the scattering vector in the diffraction literature is of opposite sign, $\Delta\vec{k} \equiv \vec{k}_\mathrm{f} - \vec{k}_\mathrm{i}$.)

Graduate students learning the concepts presented in this book are assumed to have some understanding of scattering experiments – a good understanding of x-ray diffraction would be suitable preparation. The student should have some competence with the manipulation of Patterson functions in Fourier space to the level developed, for example, in *Transmission Electron Microscopy and Diffractometry of Materials* by B. Fultz and J. W. Howe, and should have some understanding of solid-state physics at the level of *Principles of the Theory of Solids* by J. M. Ziman or *Solid-State Physics* by H. Ibach and H. Lüth.

Brent Fultz
Pasadena
April, 2020

Contents

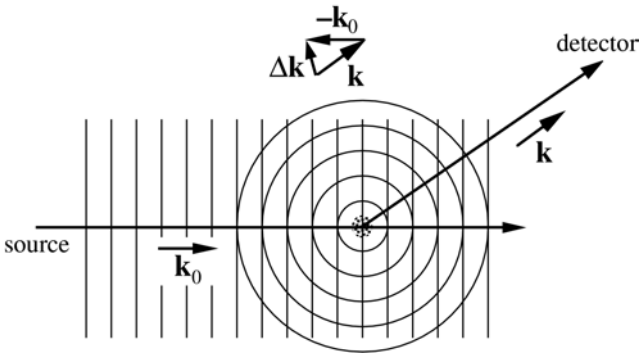
Chapter 1

Scattering

This chapter on scattering assumes some background knowledge about elastic scattering as used for diffraction studies of atomic structure. (After all, x-ray diffraction is common currency for chemists, materials scientists and condensed matter physicists.) Nevertheless, some of the important concepts in diffraction physics are developed here, and recast it into a formulation based on Patterson functions.

We begin with a discussion of coherence and energy, in part because these concepts are sometimes taken for granted by persons having extensive experience with diffraction experiments. In Sect. 1.4, elastic scattering from static potentials is developed in the Born approximation, giving the basic Fourier transform relationship between the wave and the scattering factor distribution: $\psi(\vec{Q}) = Ff(\vec{r})$. This is followed by a brief explanation that shows how time-varying potentials cause inelastic scattering by modulating the frequency of the scattered wave.

More progress in understanding scattering experiments is possible by analyzing the correlation functions that are derived from the measured intensity, rather than the neutron wavefunction (which is not measured directly). Section



1.5 returns to elastic scattering to develop the concept of the Patterson function, $P(\vec{r})$. The Patterson function is the spatial correlation function that includes all information about the diffracted intensity, as opposed to the diffracted wave. The corresponding correlation function for inelastic scattering is the Van Hove space-time correlation function. It is developed in Sect. 2.1 in Chapter 3 in a path that parallels the Patterson function, but is more general. The double Fourier transform of the Van Hove function, $G(\vec{r}, t)$, provides all information about the scattered intensity (elastic and inelastic). The Patterson function is a special case of the Van Hove function for static scattering potentials. Graphical examples are used to demonstrate fundamental features of the scattered intensity for the elastic case with the Patterson function: $I(\vec{Q}) = F_{\vec{r}} P(\vec{r})$, and for the inelastic case with the Van Hove function: $I(\vec{Q}, E) = F_{\vec{r}} F_t G(\vec{r}, t)$. The two Fourier relationships between the scattered intensity and the correlation functions are discussed in detail in this chapter and the next:

- $I(\vec{Q}) = F_{\vec{r}} P(\vec{r})$, which relates the diffracted intensity, $I(\vec{Q})$, to the Fourier transform (from space to momentum) of the Patterson function, $P(\vec{r})$.
- $I(\vec{Q}, E) = F_{\vec{r}} F_t G(\vec{r}, t)$, which relates the scattered intensity, $I(\vec{Q}, E)$, to two Fourier transforms (from space to momentum, and time to energy) of the Van Hove function, $G(\vec{r}, t)$.

For purposes of orientation, we begin with an example.

1.1 A Sampler of Scattering Mechanisms

Neutrons are scattered from materials by several mechanisms. Each scattering mechanism can be taken as independent of the others, since in quantum mechanics the scattering either occurs with full effect, or it does not occur at all. You can usually assume that the scattering mechanisms do not mix, although the average over many scatterings gives an average behavior predicted by classical mechanics. Some categorizations of the different mechanisms of scattering are

- Nuclear or Magnetic. Is the neutron scattered by interaction with the nucleus, or with unpaired (magnetic) electron spins at the atom?
- Elastic or Inelastic. Is the neutron scattered without change of its kinetic energy, or is there an energy transfer between the neutron and the sample?
- Coherent or Incoherent. Is the phase of the neutron wavefunction predictable after scattering, or is there no relationship between the phase of the incident and scattered wavefunction?
- Multiphonon or single phonon. Does the scattering of the neutron cause only one phonon to be excited, or are two excited simultaneously?

- Multiple or single scattering. Is the neutron scattered one time on its way through the sample, or is it scattered more than once?

Combinations across these categories are possible. There is a mechanism of coherent inelastic magnetic single scattering, for example. It is, however, an oxymoron to speak of incoherent coherent scattering.[1] Obviously many mechanisms for neutron scattering are possible. The essentials of these processes will be covered in the next two chapters. For an overview, however, we illustrate several of these possibilities with Fig. 1.1.

Neutron scattering from vanadium is almost entirely incoherent, meaning that the scattering works with the intensities of scattering, but not the phases. This causes a loss of information along the Q direction in Fig. 1.1Va. For example, Bragg diffractions are not seen in the data of Fig. 1.1Va for vanadium. On the other hand, neutron scattering from aluminum is almost entirely coherent, and Fig. 1.1Ala shows rich, detailed structure in Q, and Bragg diffractions along the x-axis at $E = 0$. (The $E = 0$ corresponds to elastic scattering, which is essential to diffraction.)

Figures 1.1b-e show the different contributions to the experimental data. These cannot be measured, but they can be computed with reasonable reliability today. From Fig. 1.1Vb and 1.1Alc we can see that the dominant contribution to the scattering is incoherent scattering for V, and coherent scattering for Al. The coherent scattering from V and the incoherent scattering from Al are both small.

At higher temperatures, or with higher momentum transfers of the neutron, there is an increasing possibility of having two or more phonons created by one scattering event. In essence, the scattering is becoming more "classical," where the scattering is considered to produce heat rather than individual phonons. These measurements of Fig. 1.1 were at low temperatures and with low neutron energies, however, so the multiphonon contribution makes only small changes in panels d. Nevertheless, it is easy to see that multiphonon scattering adds intensity at larger values of $Q > 7\,\text{Å}^{-1}$ for both vanadium or aluminum. (in Fig. 1.1, panels d are similar to panels c for $Q < 7\,\text{Å}^{-1}$). Finally, the samples were thin, so multiple scattering was unlikely as the neutron traversed the sample. Both multiphonon scattering and multiple scattering can be seen to add some elastic intensity over a broad range of Q.

In this book, all these mechanisms, and more, will be described quantitatively. Methods for optimizing one type of scattering or another will be described, along with experimental practices for measuring them. Different materials have different internal excitations, and some important types of magnetic and phonon excitations will be explained with enough care so that the reader is ready to take the next steps and learn how to calculate them in detail. There is a lot to know about inelastic neutron scattering, and a lot that it can do.

[1] With multiple scattering this can be possible, but one of the scatterings is fully incoherent, and the other is fully coherent.

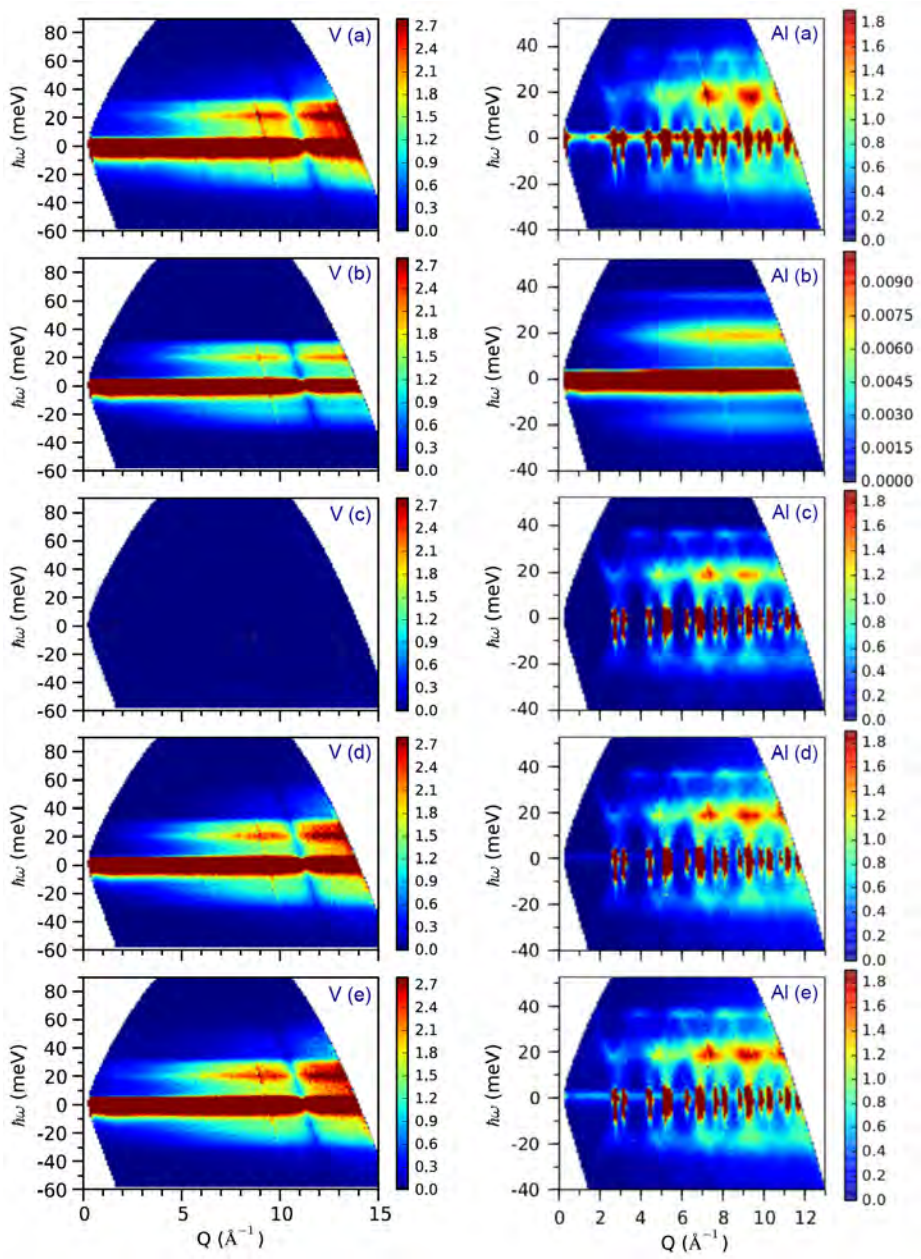


Figure 1.1: See description in text. The vertical columns are scattering from bcc vanadium at left and fcc aluminum at right. For phonon scattering, V is nearly fully incoherent, and the Al is nearly fully coherent. Simulations were performed with MCViNE, as described in various parts of this book. (a) experimental data from polycrystalline V and Al. (b) incoherent phonon scattering (notice small scale for Al) (c) coherent phonon scattering (d) multiphonon scattering added to both incoherent and coherent scattering from b and c (e) multiple scattering added to the result of d. This should be comparable to the experimental data in a, since all effects are included.

1.2 Coherence and Incoherence

Diffraction requires "coherent scattering," characterized by a precise relationship between the phases of the incident and scattered waves. The scattered wave is the sum of component waves, "wavelets" as we call them, emanating from the different atoms in the sample. In diffraction, phase differences between these outgoing wavelets cause constructive or destructive interferences at different angles around the sample, e.g., the appearance of Bragg diffraction peaks.

1.2.1 Wavefunctions

Phase

A wavefunction $\psi(x,t)$ describes the structure of a wave (its crests and troughs) along position x, at any time t. The mathematical form $\psi(kx-\omega t)$ accounts for how the wave amplitude shifts in position with increasing time. The argument of the wavefunction, $kx-\omega t$, is called the "phase" of the wave. It includes two constants: k (the wavevector), and ω (the angular frequency). The phase $kx-\omega t$ is dimensionless, so it can be used as the argument of a sine function or a complex exponential, for example. Our mathematical form causes the entire structure of the wave $\psi(kx-\omega t)$ to move to more positive x with increasing t. This is clear if we recognize that a particular wavecrest in ψ exists at a particular value of phase, so for larger t, the wave amplitude moves to larger x for the same value of $kx-\omega t$.[2]

One-Dimensional Wave

One-dimensional waves are simple because they have no vector character. Suppose the wave is confined a region of length L. The wavefunction and its intensity are:

$$\psi_{1\mathrm{D}}(x,t) = \frac{1}{\sqrt{L}}\,\mathrm{e}^{+\mathrm{i}(kx-\omega t)}\,, \tag{1.1}$$

$$I_{1\mathrm{D}} = \psi_{1D}(x,t)\,\psi^{*}_{1D}(x,t) = \frac{1}{\sqrt{L}}\,\mathrm{e}^{+\mathrm{i}(kx-\omega t)}\,\frac{1}{\sqrt{L}}\,\mathrm{e}^{-\mathrm{i}(kx-\omega t)}\,, \tag{1.2}$$

$$I_{1\mathrm{D}} = \frac{1}{L}\,. \tag{1.3}$$

If $\psi_{1\mathrm{D}}(x,t)$ were an electron wavefunction, the intensity, $I_{1\mathrm{D}}$, would be a probability density. The prefactor in (1.1) ensures proper normalization, giving a probability of 1 for finding the electron in the interval of length L:

$$P = \int_0^L I_{1D}\,\mathrm{d}x = \int_0^L \frac{1}{L}\,\mathrm{d}x = 1\,. \tag{1.4}$$

[2] We say $\psi(kx-\omega t)$ travels to the right with a "phase velocity" of ω/k. The wave $\psi(kx+\omega t)$ travels to the left.

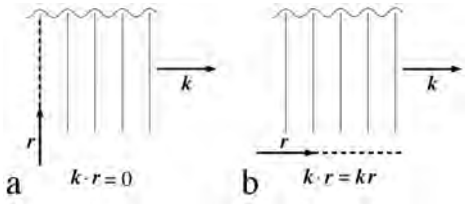


Figure 1.2: Plane wave with $\vec{k}$ oriented to the right, with orientations of $\vec{r}$ being (**a**) along the wave crests, perpendicular to $\vec{k}$, (**b**) parallel to $\vec{k}$.

Plane Wave

In three dimensions, a plane wave is:

$$\psi_{\rm 3Dpl}(\vec{r},t) \;=\; \frac{1}{\sqrt{V}}\,{\rm e}^{+{\rm i}(\vec{k}\cdot\vec{r}-\omega t)}\;, \tag{1.5}$$

which has an intensity and a normalization analogous to those for the one-dimensional wavefunction. The spatial part of the phase, $\vec{k}\cdot\vec{r}$, is illustrated for a snapshot in time in Fig. 1.2 for two orientations of $\vec{r}$: with $\vec{k}\perp\vec{r}$, and with $\vec{k}\parallel\vec{r}$. Along the direction of $\vec{r}$ in Fig. 1.2a there is no change in the phase of the wave (here $\psi_{\rm 3Dpl}(\vec{r},t) = 1/\sqrt{V}\,{\rm e}^{+{\rm i}(0-\omega t)}$), whereas in Fig. 1.2b the phase changes most rapidly along $\vec{r}$ (here $\psi_{\rm 3DPlan}(\vec{r},t) = 1/\sqrt{V}\,{\rm e}^{+{\rm i}(kr-\omega t)}$). The dot product $\vec{k}\cdot\vec{r}$ for the phase in (1.5) gives the plane wave its anisotropy in space.

Spherical Wave

By placing the origin of a spherical coordinate system at the center of the spherical wave, the spherical wave has its simplest form:

$$\psi_{\rm 3Dsph}(\vec{r},t) \;=\; \frac{1}{\sqrt{V}}\,\frac{{\rm e}^{+{\rm i}(kr-\omega t)}}{r}\;. \tag{1.6}$$

If the center of the spherical wave is the distance $\vec{r}'$ away from the origin of the coordinate system:

$$\psi_{\rm 3Dsph}(\vec{r},t) \;=\; \frac{1}{\sqrt{V}}\,\frac{{\rm e}^{+{\rm i}(k|\vec{r}-\vec{r}'|-\omega t)}}{|\vec{r}-\vec{r}'|}\;. \tag{1.7}$$

Figure 1.3 shows a vector construction for $\vec{r}-\vec{r}'$, which can be obtained by connecting the tail of $-\vec{r}'$ to the arrow of $\vec{r}$. At distances far from the scattering center, where the curvature of the spherical wave is not important, it is often useful to approximate the spherical wave as a plane wave with $\vec{r}-\vec{r}'$ pointing along the direction of $\vec{k}$.[3]

[3]This is often useful because real scatterers typically emit spherical waves, but Fourier transforms require plane waves.

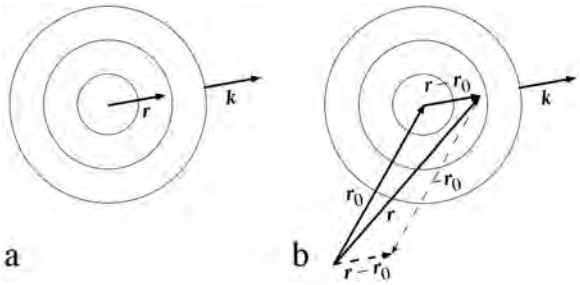


Figure 1.3: Spherical wave with $\vec{k}$ oriented away from the center of wave emission. (**a**) with coordinate system for $\vec{r}$ having its origin at center of wave emission. (**b**) with coordinate system for $\vec{r}$ having an arbitrary origin.

Phase Factor

A phase factor, $e^{-i\vec{\Delta k}\cdot\vec{R}}$ or $e^{-i(\vec{\Delta k}\cdot\vec{R}+\omega t)}$, has the mathematical form of a plane wave (1.5), and is associated with a particular wavelet, but beware. *A phase factor is not a wave.* A phase factor proves handy when two or more wavelets are scattered from different points in space at $\{\vec{R}_j\}$, typically separated by some atomic distances. What is important after the long path to the detector is how the wavelets interfere with each other – constructively or destructively – and this is accounted for by sums of phase factors like this:

$$\psi_{\rm phf}(\vec{\Delta k}) = \sum_{\{\vec{R}_j\}} e^{-i\vec{\Delta k}\cdot\vec{R}_j} \ . \tag{1.8}$$

The definition $\vec{\Delta k} \equiv \vec{k} - \vec{k}_0$ (illustrated in the chapter title image) is repeated a number of times in this book. This $\vec{\Delta k}$ is a difference in the wavevectors of two actual waves. Dot products like $\vec{\Delta k}\cdot\vec{R}_j$ give phase differences between wavelets, but $\vec{\Delta k}$ is not an actual wavevector. Chapter 5 in Fultz and Howe develops these concepts, but the reader is hereby warned that exponentials containing $\vec{\Delta k}$ are not waves, but phase factors.

1.2.2 Coherent and Incoherent Scattering

Coherent scattering preserves the relative phases of the wavelets, $\{\psi_{\vec{r}_j}\}$, scattered from different locations, $\{\vec{r}_j\}$, in a material. For coherent scattering, the total scattered wave, $\Psi_{\rm coh}$, is constructed by adding the *amplitudes* of the scattered wavelets:

$$\Psi_{\rm coh} = \sum_{\vec{r}_j} \psi_{\vec{r}_j} \ . \tag{1.9}$$

The total coherent wave therefore depends on the constructive and destructive interferences of the wavelet amplitudes. Diffraction experiments measure the total coherent *intensity*, $I_{\rm coh}$:

$$I_{\rm coh} = \Psi^*_{\rm coh}\Psi_{\rm coh} = \left|\sum_{\vec{r}_j} \psi_{\vec{r}_j}\right|^2 \ . \tag{1.10}$$

On the other hand, "incoherent scattering" does not preserve a phase relationship between the incident wave and the scattered wavelets. For incoherent scattering it is incorrect to add the amplitudes of the scattered wavelets, $\{\psi_{\vec{r}_j}\}$. Incoherently-scattered wavelets do not maintain phase relationships, so they cannot interfere constructively or destructively. The total intensity of incoherent scattering, I_{inc}, is the sum of individual scattered intensities:

$$I_{\text{inc}} = \sum_{\vec{r}_j} I_{\vec{r}_j} = \sum_{\vec{r}_j} \left|\psi_{\vec{r}_j}\right|^2 \ . \tag{1.11}$$

Because measurable intensities are added in incoherent scattering, the angular distribution of incoherent scattering from a group of N identical atoms is the same as for a single atom, irrespective of how these N atoms are positioned in space. The total intensity is simply N times larger. Some types of incoherent scattering occur with a transfer of energy from the wave to the material, and these processes can be useful for spectroscopic analysis of the atom species in a material.

It is important to emphasize the difference between the right-hand sides of (1.10) and (1.11). Because the intensity of coherent scattering in (1.10) first involves the addition of wave amplitudes, coherent scattering depends on the relative phases of the scattered wavelets and the relative positions of the N atoms in the group. Coherent scattering is useful for diffraction experiments. Incoherent scattering is not. This chapter describes in sequence the four types of scattering having coherent components that allow for diffraction experiments on materials:

- x-rays, which are scattered when they cause the atomic electrons to oscillate and re-radiate,
- electrons, which are scattered by Coulomb interactions when they penetrate the positively-charged atomic core,
- neutrons, which are scattered by nuclei (or unpaired electron spins), and
- γ-rays, which are scattered when they resonantly excite a nucleus, which later re-radiates.

1.2.3 Elastic and Inelastic Scattering

Besides being "coherent" or "incoherent," scattering processes are "elastic" or "inelastic" when there is, or is not, a change in energy of the wave after scattering. We can therefore construct four word pairs:

(coherent elastic) (coherent inelastic)
(incoherent elastic) (incoherent inelastic)

Diffraction experiments need coherent elastic scattering, whereas spectroscopies that measure intensity versus energy often use incoherent inelastic scattering. The case of incoherent elastic scattering is also common, and occurs, for example, when phase relationships between scattered wavelets are disrupted by

disorder in the material. Incoherent elastic intensity does not show the sharp diffractions associated with crystalline periodicities, but has a broad angular dependence. Finally, coherent inelastic scattering is used in neutron scattering studies of excitations in materials, such as phonons (vibrational waves) or magnons (spin waves), that have precise energy-wavevector relationships. In some phonon studies, a neutron loses energy when creating a phonon (so it is inelastic), but the scattering amplitude depends on the phases of the atom movements in the phonon with respect to the neutron wavevectors (so it is coherent).

A deeper and more rigorous distinction between coherent and incoherent scattering involves our knowledge about the internal coordinates of the scatterer:

- Consider a simple oscillator (a bound electron, for example) that is driven by an incident wave and then re-radiates. There is a transfer of energy from the incident wave to the oscillator, and then to the outgoing wave. Suppose we know in full detail how the coordinates of the oscillator respond to the incident wave. Since the scattering process is determined fully, the phases of all outgoing wavelets have a precise and known relationship to the phase of the incident wave. The scattering is coherent.

- On the other hand, suppose the coordinates of this oscillator were coupled to another system within the material (a different electron, for example), and furthermore suppose there is freedom in how the oscillator can interact with this other system. (Often differing amounts of energy can be transferred from the oscillator to the other system because the transfer occurs by a quantum mechanical process that is not deterministic.) If this energy transfer is different for different scatterings, we cannot predict reliably the phase of the scattered wavelet. The scattering is incoherent.

It is therefore not surprising that incoherence is often associated with inelastic scattering, since inelastic scattering involves the transfer of energy from the scatterer to another component of the material. Incoherence does not imply inelastic scattering, however, and inelastic scattering is not necessarily incoherent.

1.2.4 Wave Amplitudes and Cross-Sections

Cross-Sections

X-rays, electrons, neutrons, and γ-rays are detected one-at-a-time in scattering experiments. For example, the energy of an x-ray is not sensed over many positions, as are ripples that spread to all edges of a pond of water. Either the entire x-ray is detected or not within the small volume of a detector. For x-ray scattering by an individual atomic electron as described in the next section, the scattering may or may not occur, depending on a probability for the x-ray–electron interaction.

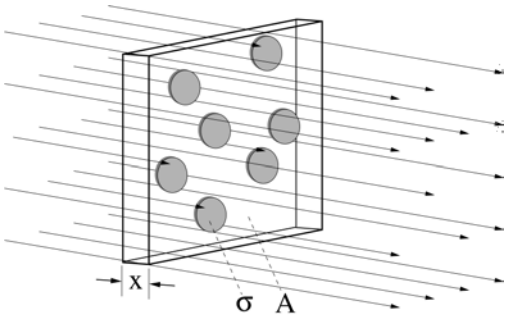


Figure 1.4: These 7 scatterers occupy the fraction 0.2 of the sample area, A, and therefore remove the fraction 0.2 of the rays from the incident beam. From (1.12): $\sigma = (0.2/7)A$. In the thin sample limit, the number of scatterers and the amount of scattering increase in proportion to thickness, x, but σ remains constant.

An important quantity for scattering problems is the "cross-section," σ, which is the effective "target area" presented by each scatterer. With cross-sections it is handy to think of a number, N, of scatterers in a sample of area A as in Fig. 1.4. The probability of scattering is equal to the fraction of sample area "blocked" by all N scatterers. For thin samples when the scatterers do not overlap, the N scatterers block an area equal to $N\sigma$. The fraction of rays removed from the incident beam is the blocked area divided by the total area:

$$N\,\frac{\sigma}{A} = N\,\frac{\sigma x}{Ax} = \rho\,\sigma x\,. \tag{1.12}$$

Here the density of scatterers, $\rho \equiv N/(Ax)$ has units [scatterers cm^{-3}].

To illustrate a salient feature of coherent scattering, consider the elastic scattering of neutrons through the interaction of their spin with the spin polarization of electrons in an antiferromagnet. The total cross-section depends on the total number of unpaired electrons in the material. As mentioned after (1.11), for incoherent scattering the picture would then be complete – the spatial distribution of the scattered intensity is obtained by adding the intensities from independent scattering events from different atoms.

Coherent scattering requires further consideration of the wave amplitudes before calculating the cross-section. A hierarchy of wave interference processes can occur between spin structures on different length scales:

- the unpaired electrons in the same atom (atomic form factor),
- the atoms in the unit cell of the crystal (structure factor),
- the unit cells in the crystal (shape factor), and
- density variations across a material (small angle scattering).

The spatial redistribution of scattered intensity can be spectacularly large in the case of Bragg diffractions, but the total coherent cross-section remains constant. By rearranging the atom positions in a material, the constructive and destructive interferences of coherent scattering are altered and the angles of scattering are redistributed, but for the same incident flux the scattered energy

is conserved (for x-rays or γ-rays), or the total number of scattered particles remains the same (electrons and neutrons).

The flux of scattered x-rays, electrons, neutrons, or γ-rays at the distance $\vec{r}$ from the scatterer decreases as $1/r^2$ along $\hat{\vec{r}}$. A scattered photon carries energy, so the radiated energy flux also decreases as $1/r^2$ from the scatterer. The energy of a photon is proportional to E^*E, so the electric field, E, has an amplitude that must decrease as $1/r$ from the center of scattering. For scattered x-rays, we relate the electric field along $\widehat{\vec{r}}$ to the incident electric field at the scatterer, E_0:

$$E(\vec{r}) \propto \frac{E_0}{r} \ , \tag{1.13}$$

where the constant of proportionality would include any angular dependence. The electric fields $E(\vec{r})$ and E_0 in (1.13) have the same units, of course, so the constant of proportionality has units of length. The square of this "scattering length" is the cross section per steradian, as we next show for electron scattering (but the argument pertains to all waves).

The problem is similar and actually simpler for neutron scattering, for which this "coherent scattering length" is a constant, b. This b can be converted to a cross-section by first squaring it (actually taking b^*b because it is often convenient for b to be a complex number). The second step is to integrate over all solid angle, which is 4π. For all intents and purposes the nucleus that scatters a typical 1 Å neutron is a point and has isotropic scattering, so the relationship between the scattering length and the total coherent cross-section, σ_{coh}, is:

$$\sigma_{\text{coh}} = 4\pi\left|b\right|^2 \ . \tag{1.14}$$

Cross-Section for Wave Scattering

Here we find the cross-section for wave scattering. Imagine a large sphere of radius R around the scatterer, and consider the total flux, $J_{\text{sc}}(R)$, scattered through a unit area of surface of this sphere. The incident beam has a flux J_{in} over an area A. The ratio of all scattered electrons to incident electrons, $N_{\text{sc}}/N_{\text{in}}$, is:

$$\frac{N_{\text{sc}}}{N_{\text{in}}} = \frac{J_{\text{sc}}(R)\ 4\pi R^2}{J_{\text{in}}\ A} = \frac{v\ |\psi_{\text{sc}}(R)|^2\ 4\pi R^2}{v\ |\psi_{\text{in}}|^2\ A} \ . \tag{1.15}$$

We consider elastic scattering for which the incident and scattered electrons have the same velocity, v, but for inelastic scattering these factors do not cancel. We use the spherical wave (1.6) for $\psi_{\text{sc}}(R)$ and the plane wave (1.5) for ψ_{in}. For both waves, the exponential phase factors, multiplied by their complex conjugates, give the factor 1. The normalization factors also cancel, so (1.15) becomes:

$$\frac{N_{\text{sc}}}{N_{\text{in}}} = \frac{|f_{\text{el}}|^2\ 4\pi R^2}{R^2\ A} \ , \tag{1.16}$$

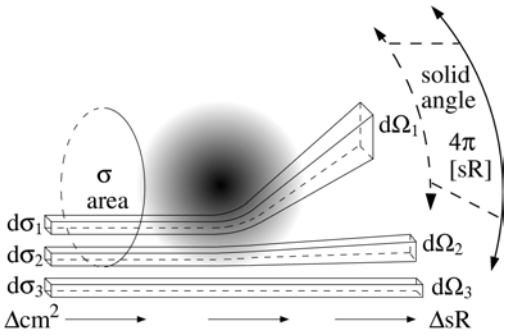


Figure 1.5: The differential scattering cross-section, $d\sigma/d\Omega$, for three paths past a scatterer. The third path, $d\sigma_3/d\Omega_3$, misses the scatterer and contributes only to the forward beam. The paths with areas $d\sigma_1$ and $d\sigma_2$ make contributions to the total cross-section for scattering, σ, and these contributions are included when the intensity is integrated over the differential solid angles $d\Omega_1$ and $d\Omega_2$.

where f_{el}/R is the fraction of the incident electron amplitude that is scattered into a unit area of the sphere at radius R. Figure 1.4 helps demonstrate the fact that the ratio of the cross-section σ to the area A of the incident beam equals the ratio of scattered to incident electrons, N_{sc}/N_{in}:

$$\frac{\sigma}{A} = \frac{N_{sc}}{N_{in}} = \frac{4\pi|f_{el}|^2}{A} , \tag{1.17}$$

$$\sigma = 4\pi|f_{el}|^2 . \tag{1.18}$$

The scattering of an x-ray by a single atomic electron can be treated in the same way, but we need to account for the electric dipolar pattern of x-ray radiation with a factor of 2/3 in the cross-section,

$$\sigma_{x1e} = \frac{8\pi}{3}\left|f_{x1e}\right|^2 , \tag{1.19}$$

where f_{x1e} is the scattering length. This f_{x1e} is the actual constant of proportionality to convert (1.13) into an equality.

Anisotropic scattering is the rule rather than the exception, however, so simple cross-sections like those of (1.18) are usually inadequate, even if altered by factors like the 2/3 used in (1.19). The "differential scattering cross-section," written as $d\sigma/d\Omega$, contains the angular detail missing from the total cross-section, σ.

The differential scattering cross-section, $d\sigma/d\Omega$*, is the piece of area offered by the scatterer,* $d\sigma$*, for scattering an incident x-ray (or electron or neutron) into a particular increment in solid angle,* $d\Omega$*.*

The concept of $d\sigma/d\Omega$ is depicted Fig. 1.5. Note that $d\sigma/d\Omega$ relates an increment in area (on the left) to an increment in solid angle (on the right).

For the simple case of isotropic scattering,

$$\frac{d\sigma}{d\Omega} = \left|f\right|^2 , \tag{1.20}$$

which is a constant. For anisotropic scattering, (1.20) is generalized with a scattering length, $f(\vec{k}_0, \vec{k})$, that depends on the directions of the incident and outgoing wavevectors, $\vec{k}_0$ and $\vec{k}$, respectively:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = \left| f(\vec{k}_0, \vec{k}) \right|^2 . \tag{1.21}$$

We recover the total cross-section, σ, by integrating $\mathrm{d}\sigma/\mathrm{d}\Omega$ over all solid angle,

$$\sigma = \int_{\text{sphere}}^{4\pi} \frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} \mathrm{d}\Omega . \tag{1.22}$$

As a check, substituting the constant (1.20) into (1.22) and integrating gives (1.18), as expected.

Special Characteristics of Coherent Scattering

Compare the differential scattering cross-sections for coherent x-ray scattering by a single electron at $\vec{r}_j$, $\mathrm{d}\sigma_{\mathrm{x1e},\vec{r}_j}/\mathrm{d}\Omega$, and an atom having Z electrons, $\mathrm{d}\sigma_{\mathrm{atom}}/\mathrm{d}\Omega$:

$$\frac{\mathrm{d}\sigma_{\mathrm{x1e},\vec{r}_j}}{\mathrm{d}\Omega}(\vec{k}_0, \vec{k}) = \left| f_{\mathrm{x1e},\vec{r}_j}(\vec{k}_0, \vec{k}) \right|^2 , \tag{1.23}$$

$$\frac{\mathrm{d}\sigma_{\mathrm{atom}}}{\mathrm{d}\Omega}(\vec{k}_0, \vec{k}) = \left| f_{\mathrm{atom}}(\vec{k}_0, \vec{k}) \right|^2 . \tag{1.24}$$

In coherent scattering we sum wave amplitudes (cf., (1.9)), so for coherent scattering we sum the scattering lengths of all Z electrons to obtain the scattering length of an atom:

$$f_{\mathrm{atom}}(\vec{k}_0, \vec{k}) = \sum_{\vec{r}_j}^{Z} f_{\mathrm{x1e},\vec{r}_j}(\vec{k}_0, \vec{k}) . \tag{1.25}$$

Note that (1.25) is a sum of the $f_{\mathrm{x1e},\vec{r}_j}$, but (1.24) is the square of this sum. Equation (1.24) can predict that the coherent x-ray scattering from an atom with Z electrons is Z^2 times stronger than for a single electron, and this proves to be true in the forward direction. However, the total cross-section for coherent scattering must increase linearly with the number of scatterers (here the number of electrons, Z). Consequently the coherent scattering is suppressed in other directions if a scaling with Z^2 is allowed in special directions. The angular distribution of coherent scattering must be different for the atom and for the single electron. That is, $f_{\mathrm{x1e}}(\vec{k}_0, \vec{k})$ and $f_{\mathrm{atom}}(\vec{k}_0, \vec{k})$ must have different shapes (they must depend differently on $\vec{k}_0$ and $\vec{k}$). The following is an inequality for coherent scattering (although its analog for incoherent scattering is an equality):

$$\frac{\mathrm{d}\sigma_{\mathrm{atom,coh}}}{\mathrm{d}\Omega}(\vec{k}_0, \vec{k}) \neq \sum_{\vec{r}_j}^{Z} \frac{\mathrm{d}\sigma_{\mathrm{x1e},\vec{r}_j,\mathrm{coh}}}{\mathrm{d}\Omega}(\vec{k}_0, \vec{k}) . \tag{1.26}$$

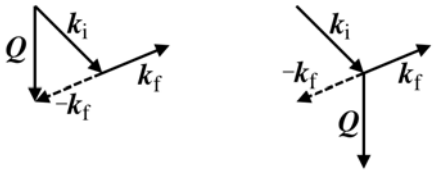


Figure 1.6: Vector difference of $\vec{Q} = \vec{k_i} - \vec{k_f}$, showing inverted $-\vec{k_f}$ with dashed line, and $\vec{Q}$ positioned more conventionally in image at right.

Integrating (1.26) gives an equality for coherent (and incoherent) scattering:

$$\int_{\text{sphere}}^{4\pi} \frac{d\sigma_{\text{atom,coh}}}{d\Omega}(\vec{k_0},\vec{k})d\Omega = \int_{\text{sphere}}^{4\pi} \sum_{\vec{r_j}}^{Z} \frac{d\sigma_{\text{x1e},\vec{r_j},\text{coh}}}{d\Omega}(\vec{k_0},\vec{k})d\Omega \,, \tag{1.27}$$

because with (1.22) we see that (1.27) equates the individual electron cross-sections to the total cross-section of the atom:

$$\sigma_{\text{atom,coh}} = Z\sigma_{\text{x1e,coh}} \,. \tag{1.28}$$

The process of actually performing the sum in (1.25) evidently requires delicacy in accounting for the phase relationships between the x-ray wavelets scattered into different angles, and knowledge about the electron density of the atom. This is the subject of atomic form factor calculations (see Chapter 3 in Fultz and Howe, for example, or (1.76)).

1.3 Momentum Transfer and Phase Shifts

1.3.1 Momentum and Energy

All processes of scattering redirect the incident wave, causing a change in wavevector, $\vec{Q}$. Momentum conservation then requires a transfer of momentum, $\Delta\vec{p}$, from the incident wavefunction to the sample

$$\vec{Q} \equiv \vec{k_i} - \vec{k_f} \,, \tag{1.29}$$

$$\Delta\vec{p} = \hbar\vec{Q} \,, \tag{1.30}$$

where "i" and "f" denote initial and final. Equation 1.29, shown graphically in Fig. 1.6 is familiar from the elastic scattering in diffraction experiments, where $|\vec{k_f}|=|\vec{k_i}|$ (and where the quantity $\Delta\vec{k} \equiv -\vec{Q}$ is often used). The transfer of momentum to the sample (Eq. 1.30) occurs in diffraction experiments, but is rarely considered.

There can be a transfer of energy accompanying a transfer of momentum. Consider the scattering of a particle such as an electron or neutron. The transfer

of energy E is

$$E = \frac{p_{\rm i}^2}{2M} - \frac{p_{\rm f}^2}{2M} , \tag{1.31}$$

$$E = \frac{\hbar^2}{2M}\left(k_{\rm i}^2 - k_{\rm f}^2\right) , \tag{1.32}$$

$$E = -\frac{\hbar^2}{2M}\left(Q^2 - 2\vec{k_{\rm i}}\cdot\vec{Q}\right) , \tag{1.33}$$

where the last line was obtained after substituting $\vec{k_{\rm f}} = \vec{k_{\rm i}} - \vec{Q}$. These relationships between E and $\vec{Q}$ are called the "kinematics" of scattering. Three cases of Eq. 1.33 are noteworthy

1. When $\vec{k_{\rm i}} = \vec{k_{\rm f}}$ and $\vec{Q} = 0$, there is no momentum transfer, and there is no scattering, either. The particle continues with no deflection along the path of the forward beam. Here $E = 0$, of course.

2. When $\vec{Q} = \vec{k_{\rm i}}$, and $\vec{k_{\rm f}} = 0$, the incident particle is stopped by the scatterer. This gives maximum energy transfer $E = +\frac{\hbar^2 Q^2}{2M}$.[4]

3. When $|\vec{k_{\rm f}}|=|\vec{k_{\rm i}}|=k$ and $Q = 2k\sin\theta$, the energy transfer $E = 0$ because the scattered particle has its original wavevector k, and all of its original kinetic energy. For this to occur, the mass of the scatterer must be high, so the scatterer undergoes no recoil.

So one reason why Eq. 1.33 is not used to interpret diffraction experiments is that many atoms participate in the scattering. The mass that absorbs the momentum is therefore large, perhaps the weight of a crystal. This causes E to be small owing to the large M in the denominator of Eq. 1.33. There is little energy in the recoil of a heavy object[5] when it is hit by a light particle, even if there is a big change in momentum of the lightweight particle. We know that many atoms participate in the diffraction process because the diffraction pattern arises from wave interferences from coherent scattering from many different atoms.

[4] Three examples of $\{\vec{Q} = \vec{k_{\rm i}}\ ,\ \vec{k_{\rm f}} = 0\}$ are noteworthy:

- In forward scattering by a scatterer of equal mass, the scatterer continues in the forward direction and the incident particle stops.
- For a lightweight particle scattered by a heavy mass, this would correspond to a $\vec{k_{\rm f}}$ perpendicular to $\vec{k_{\rm i}}$, but this is suppressed because there is no perpendicular component of momentum in the problem.
- This condition of $\vec{Q} = \vec{k_{\rm i}}$ is also a result when all possible orientations of $\vec{Q}$ are averaged over the 4π steradians over a sphere. This is typical for incoherent scattering of a light particle from a heavy scatterer.

[5] The sample then transfers its momentum to the even more massive Earth, which undergoes a negligible change in its orbit.

What is more interesting is when the scattering of the particle causes an excitation within the sample. The classical analog is when the scattering generates heat, and is called "inelastic scattering" because kinetic energy is not conserved. The proper accounting for the internal heat allows total energy to be conserved. Our interest is when the internal heat is quantized, however, and the creation of individual magnons or phonons is used in the energy balance. It is appropriate to consider this as "inelastic scattering," because the incident particle or wave transfers energy to or from the sample, much as a classical particle creates heat.

1.3.2 Phase Shifts

A transfer of momentum always occur in scattering. Here we show why all scattering processes include a factor $\exp(i\vec{Q}\cdot\vec{R})$ in the scattering amplitudes (where $\vec{Q} = \Delta\vec{p}/\hbar$).

Both elastic scattering and inelastic scattering involve momentum transfer to the sample, $\Delta\vec{p}$, as the incident wave (or wavefunction) changes its wavevector from $\vec{k}_\mathrm{i}$ to $\vec{k}_\mathrm{f}$

$$\Delta\vec{p} = \hbar\,\vec{Q} \quad \text{where} \quad \vec{Q} \equiv \vec{k}_\mathrm{i} - \vec{k}_\mathrm{f} \quad \text{and} \quad \vec{k}_\mathrm{f} = \vec{k}_\mathrm{i} - \vec{Q}\,, \tag{1.34}$$

where i and f designate "initial" and "final." To accommodate this change in momentum, it is more convenient to work with the momentum representation of a neutron state $\phi_\alpha(\vec{p})$, rather than the usual spatial representation $\psi_\alpha(\vec{x})$. These are related by Fourier transformation

$$\phi_\alpha(\vec{p}) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\vec{p}\cdot\vec{x}/\hbar} \psi_\alpha(\vec{x})\,\mathrm{d}\vec{x}\,, \tag{1.35}$$

$$\psi_\alpha(\vec{x}) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \mathrm{e}^{+\mathrm{i}\vec{p}\cdot\vec{x}/\hbar} \phi_\alpha(\vec{p})\,\mathrm{d}\vec{p}\,. \tag{1.36}$$

The α denotes the state of the neutron, and changes after scattering.

Phase Shifts in Elastic Scattering

Consider first an elastic scattering, where the incident and final waves are plane waves, and $|k_\mathrm{f}| = |k_\mathrm{i}|$. After scattering, the final wave $\phi_\mathrm{f}(\vec{p})$ is shifted in

momentum by $\Delta\vec{p} = -\vec{Q}/\hbar$ with respect to the initial $\phi_{\rm i}(\vec{p})$, so Eq. 1.36 becomes

$$\psi_{\rm f}(\vec{r}) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} {\rm e}^{+{\rm i}(\vec{p}/\hbar)\cdot\vec{r}}\, \phi_{\rm i}(\vec{p} + \hbar\vec{Q})\, {\rm d}^3\vec{p}\,, \tag{1.37}$$

$$\psi_{\rm f}(\vec{r}) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} {\rm e}^{+{\rm i}(\vec{p}'/\hbar - \vec{Q})\cdot\vec{r}}\, \phi_{\rm i}(\vec{p}')\, {\rm d}^3\vec{p}'\,, \tag{1.38}$$

$$\psi_{\rm f}(\vec{r}) = \frac{{\rm e}^{-{\rm i}\vec{Q}\cdot\vec{r}}}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} {\rm e}^{+{\rm i}(\vec{p}'/\hbar)\cdot\vec{r}}\, \phi_{\rm i}(\vec{p}')\, {\rm d}^3\vec{p}'\,, \tag{1.39}$$

$$\psi_{\rm f}(\vec{r}) = {\rm e}^{-{\rm i}\vec{Q}\cdot\vec{r}}\, \psi_{\rm i}(\vec{r})\,, \tag{1.40}$$

The change of variable from $\vec{p}$ to to $p' = \vec{p} + \hbar\vec{Q}$ from Eq. 1.37 to Eq. 1.38 has no effect on the infinite limits of integration, the phase factor ${\rm e}^{-{\rm i}\vec{Q}\cdot\vec{r}}$ is constant in $\vec{p}'$ for Eq. 1.39, and Eq. 1.36 was recognized in going to Eq. 1.40.

Equation 1.40 gives a general effect of momentum transfer in elastic scattering, sometimes described as the effect of "Gallilean invariance" of the scattering process. The amplitude for elastic scattering at a selected $\vec{Q}$ is defined as $f(\vec{Q})$, where

$$f(\vec{Q}) = K \int_{-\infty}^{\infty} \psi_{\rm f}^*(\vec{r})\, \psi_{\rm i}(\vec{r})\, {\rm d}^3\vec{r}\,, \tag{1.41}$$

$$f(\vec{Q}) = K \int_{-\infty}^{\infty} {\rm e}^{{\rm i}\vec{Q}\cdot\vec{r}} \psi_{\rm i}^*(\vec{r})\, \psi_{\rm i}(\vec{r})\, {\rm d}^3\vec{r}\,, \tag{1.42}$$

The constant K depends on the physical process of scattering. For neutrons, K might be the coherent scattering length b of a nucleus. For electron scattering by an atom, K might be $2Z/(a_0Q^2)$, where a_0 is the Bohr radius, Z is atomic number, and the Q^{-2} originates with the Fourier transform of the Coulomb potential. This $f(\vec{Q})$ is often called the atomic form factor for elastic scattering.

Phase Shifts in Inelastic Scattering

After scattering, the final state of the neutron is β, which always differs from the initial state α owing to a change in energy and wavelength after inelastic scattering. We seek the inelastic form factor, $f(\vec{Q}, E)$, which is the amplitude of ψ_β for the incident amplitude ψ_α. For coherent neutron scattering, this scales with the scattering length b, and with the integral $\langle\beta|\alpha\rangle$. In what follows, we integrate over momentum coordinates, but the initial and final coordinates are not the same. The final wavefunction has a shift in its momentum coordinates (complementary to that of the sample in Eq. 1.34)

$$-\vec{Q} = \vec{p_{\rm f}}/\hbar - \vec{p_{\rm i}}/\hbar\,. \tag{1.43}$$

We can use the same momentum coordinate for the final state of the neutron if we make a shift of all momentum coordinates by $-\hbar\vec{Q}$. Equation 1.36 becomes

$$\psi_{\beta{\rm f}}(\vec{x}) = \frac{1}{\sqrt{2\pi\,\hbar}} \int_{-\infty}^{\infty} {\rm e}^{+{\rm i}(\vec{p}\cdot\vec{x}/\hbar - \vec{Q}\cdot\vec{x})} \phi_{\beta{\rm i}}(\vec{p})\, {\rm d}\vec{p}\,. \tag{1.44}$$

The $\vec{Q}$ is fixed by our selection of detector angle and energy transfer, so its exponential can be removed from the integration over $\vec{p}$

$$\psi_{\beta f}(\vec{x}) = \frac{1}{\sqrt{2\pi\hbar}} e^{-i\vec{Q}\cdot\vec{x}} \int_{-\infty}^{\infty} e^{+i\vec{p}\cdot\vec{x}/\hbar} \phi_{\beta i}(\vec{p})\, d\vec{p}\,. \tag{1.45}$$

Comparing Eq. 1.45 to Eq. 1.36 we find that the final state of the sample after momentum transfer is

$$\psi_{\beta f}(\vec{x}) = e^{i\vec{Q}\cdot\vec{x}} \psi_{\beta i}(\vec{x})\,. \tag{1.46}$$

The amplitude of the scattering, or form factor, depends on the overlap of the initial and final states as

$$\langle\beta|\alpha\rangle = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} e^{i\vec{Q}\cdot\vec{x}} \psi^*_{\beta i}(\vec{x})\, \psi_{\alpha i}(\vec{x})\, d\vec{x}\,. \tag{1.47}$$

For scattering in general

$$f(\vec{Q}) = K\langle f|e^{i\vec{Q}\cdot\vec{r}}|i\rangle\,, \tag{1.48}$$

an expression that is often stated with little proof or context. The phase factor $e^{i\vec{Q}\cdot\vec{r}}$, a complex number of modulus unity, is central to all types of scattering. It can be extended to assemblages of many atoms, such as a crystal in Eq. 2.48 below. As a word of warning, explained in Chapter 2, for quantum mechanical scattering by exciting of phonons or magnons, the phase factor becomes an operator expression, where $\vec{Q}$ and $\vec{r}$ are replaced with operators $\underline{\vec{Q}}$ and $\underline{\vec{r}}$. Evaluation of the exponential requires its expansion in a power series. Unfortunately, the operators $\underline{\vec{Q}}$ and $\underline{\vec{r}}$ do not commute, so subtlety is required when doing the quantum mechanics (Sect. 4.1).

1.4 Born Approximation

Almost without a second thought, we treat neutron scattering as a wave phenomenon with the neutron wavefunction satisfying the Schrödinger wave equation. A neutron diffraction pattern, with its sharp peaks, is certainly evidence of wave behavior. The interpretation of the neutron wavefunction is different from that of a simple wave, however. Suppose we were to turn on a monochromatic neutron beam and watch the formation of a diffraction pattern, using an area detector capable of displaying the impacts of individual neutrons. When the neutron beam is turned on, individual counts are recorded at different pixels in the detector array. With time, an obvious bias appears, where the points of detection are most frequently at the positions of the rings and spots of the diffraction pattern. This behavior motivates the interpretation of the neutron wavefunction in terms of probabilities – specifically, the neutron probability is

the neutron wavefunction times its complex conjugate (which gives a real number). Usually this probabilistic interpretation can be ignored when we consider a diffraction pattern from many neutrons, and we can consider neutron diffraction as the diffraction of any other type of wave. When individual neutron events are considered, however, we may have to recall the probabilistic interpretation of the neutron wavefunction because individual neutron detections look like particles rather than waves.

Another point to remember is that the wave behavior is a characteristic of an *individual* neutron. When considering a diffraction pattern involving multiple neutrons, we do not add the amplitudes of multiple wavefunctions. Neutrons are fermions, and do not form coherent states as in Bose condensation, for example. At the viewing screen, we add the intensities of individual neutrons. The interactions between different neutrons are not coherent.

Our picture of scattering begins with one neutron as a wave incident on an atom. This wave looks like a plane wave because it comes from a distant source. The wave interacts with the nucleus or magnetic electron cloud of the atom, and an outgoing wave is generated. This outgoing wave is something like a spherical wave originating at the atom, although its intensity is not isotropic. Figure 1.7 shows the geometry, wavevectors and position vectors for our neutron scattering problem. Here both $\vec{r}$ and $\vec{r}'$ are large compared to the size of the scatterer. Our plane wave incident from the left, $\Psi_{\rm inc}$, is of the standard form:

$$\Psi_{\rm inc} = e^{i(\vec{k}_i \cdot \vec{r}' - \omega t)} \ . \tag{1.49}$$

In what follows we neglect the time dependence to emphasize the manipulations of the spatial coordinates. We later recover the time-dependence by multiplying our results by $e^{-i\omega t}$. A spherical wave, $\Psi_{\rm sc}$, travels outwards from the center of scattering. The scattered wave has the form:

$$\Psi_{\rm sc} = f(\vec{k}_i, \vec{k}_f) \frac{e^{ik_f|\vec{r}-\vec{r}'|}}{\left|\vec{r}-\vec{r}'\right|} \ , \tag{1.50}$$

where the scattering length $f(\vec{k}_i, \vec{k}_f)$ of Sect. 1.2.4 varies with the orientation of $\vec{k}_i$ and $\vec{k}_f$, $\vec{r}'$ is now used to locate the center of the scatterer, and the difference, $\vec{r}-\vec{r}'$, is the distance from the scatterer to the detector. The intensity of $\Psi_{\rm sc}$ falls off with distance as $1/r^2$, as we expect:

$$I_{\rm sc} = \Psi^*_{\rm sc}\Psi_{\rm sc} = \left|f(\vec{k}_i, \vec{k}_f)\right|^2 \frac{e^{-ik_f|\vec{r}-\vec{r}'|}}{\left|\vec{r}-\vec{r}'\right|} \frac{e^{ik_f|\vec{r}-\vec{r}'|}}{\left|\vec{r}-\vec{r}'\right|} \ , \tag{1.51}$$

$$I_{\rm sc} = \left|f(\vec{k}_i, \vec{k}_f)\right|^2 \frac{1}{\left|\vec{r}-\vec{r}'\right|^2} \ . \tag{1.52}$$

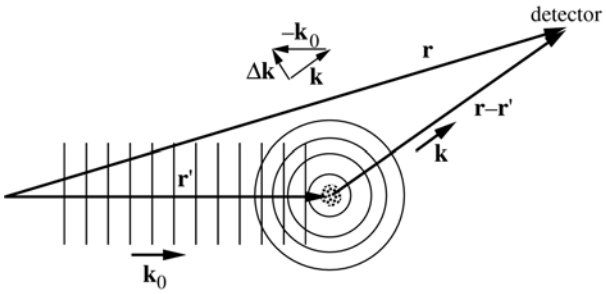


Figure 1.7: Wavevectors and position vectors for neutron scattering.

1.4.1 Green's Function

To obtain the scattering length $f(\vec{k}_i, \vec{k}_f)$, we must solve the Schrödinger equation for the incident neutron inside the scattering atom (the mass of the neutron is m, and its coordinates in the atom are $\vec{r}$):

$$-\frac{\hbar^2}{2m}\nabla^2\Psi(\vec{r}) + V(\vec{r})\,\Psi(\vec{r}) = E\,\Psi(\vec{r})\,, \tag{1.53}$$

$$\frac{\hbar^2}{2m}\nabla^2\Psi(\vec{r}) + E\,\Psi(\vec{r}) = V(\vec{r})\,\Psi(\vec{r})\,, \tag{1.54}$$

which we write as:

$$\left(\nabla^2 + k_i^2\right)\Psi(\vec{r}) = U(\vec{r})\,\Psi(\vec{r})\,, \tag{1.55}$$

after having made the two definitions:

$$k_i^2 \equiv \frac{2mE}{\hbar^2}\,, \tag{1.56}$$

$$U(\vec{r}) \equiv \frac{2mV(\vec{r})}{\hbar^2}\,. \tag{1.57}$$

The formal approach to finding the solution of the Schrödinger equation in this problem makes use of Green's functions. A Green's function, $G(\vec{r}, \vec{r}')$, provides the response at $\vec{r}$ for a point scatterer at $\vec{r}'$:

$$\left(\nabla^2 + k_i^2\right)G(\vec{r}, \vec{r}') = \delta(\vec{r}')\,. \tag{1.58}$$

We find the Green's function in a quick way by starting with an identity:

$$\nabla^2\frac{e^{ikr}}{r} = e^{ikr}\nabla^2\frac{1}{r} - k^2\frac{e^{ikr}}{r}\,, \tag{1.59}$$

$$\left(\nabla^2 + k^2\right)\frac{e^{ikr}}{r} = e^{ikr}\nabla^2\frac{1}{r}\,, \tag{1.60}$$

Recall that:

$$\nabla^2\frac{1}{r} = -4\pi\delta(r)\,, \text{ so} \tag{1.61}$$

$$\left(\nabla^2 + k^2\right)\frac{e^{ikr}}{r} = -e^{ikr}4\pi\delta(r)\,. \tag{1.62}$$

The δ-function simplifies the right-hand side, forcing it to zero everywhere except at $r = 0$. At $r = 0$, however, $e^{ikr} = 1$. From our identity (1.59) we therefore obtain:

$$\left(\nabla^2 + k^2\right) \frac{e^{ikr}}{r} = -4\pi\delta(r) \; . \tag{1.63}$$

We make a shift of the origin: $\vec{r} \to \vec{r} - \vec{r}'$ (so we can see more easily how the outgoing wave originates at the scatterer – see Fig. 1.7). After doing so, we identify our Green's function by comparing (1.58) and (1.63):

$$G(\vec{r}, \vec{r}') = -\frac{1}{4\pi} \frac{e^{ik_i|\vec{r}-\vec{r}'|}}{|\vec{r} - \vec{r}'|} \; . \tag{1.64}$$

With our Green's function in hand, we construct $\Psi_{\text{scatt}}(\vec{r})$ by integrating. The idea is that to obtain the total wave amplitude at $\vec{r}$, we need to add up the spherical wavelet amplitudes emanating from all $\vec{r}'$ (each of form (1.64)), weighted by their strengths. This weight is the right-hand side of (1.55). Formally, the limits of integration cover all of space, but in fact it is only important to extend them over the $\vec{r}'$ where $U(\vec{r}')$ is non-zero (approximately the volume of the atom).

$$\Psi_{\text{sc}}(\vec{r}) = \int U(\vec{r}')\, \Psi(\vec{r}')\, G(\vec{r}, \vec{r}')\, d^3\vec{r}' \; . \tag{1.65}$$

The total wave at $\vec{r}$, $\Psi(\vec{r})$, has both incident and scattered components:

$$\Psi = \Psi_{\text{inc}} + \Psi_{\text{sc}} \; , \tag{1.66}$$

$$\Psi(\vec{r}) = e^{i\vec{k}_i \cdot \vec{r}} + \frac{2m}{\hbar^2} \int V(\vec{r}')\, \Psi(\vec{r}')\, G(\vec{r}, \vec{r}')\, d^3\vec{r}' \; . \tag{1.67}$$

1.4.2 First Born Approximation

Up to here our solution is exact. It is, in fact the Schrödinger equation itself, merely transformed from a differential equation to a integral equation appropriate for scattering problems. The problem with this integral equation (1.65) is that Ψ appears both inside and outside the integration, so an approximation is generally required to proceed further. The approximation that we use is the "first Born approximation." It amounts to using a plane wave, the incident plane wave, for Ψ in the integral:

$$\Psi(\vec{r}') \simeq e^{i\vec{k}_i \cdot \vec{r}'} \; . \tag{1.68}$$

The first Born approximation assumes that the wave is undiminished and scattered only once by the material. This assumption is valid when the scattering is weak.

We simplify (1.64) by making the approximation that the detector is far from the scatterer. This allows us to work with plane waves at the detector, rather

than outgoing spherical waves. To do so we align the outgoing wavevector $\vec{k}_f$ along $(\vec{r} - \vec{r}')$ as shown in Fig. 1.7. The product of scalars, $k_f \left|\vec{r} - \vec{r}'\right|$, in the exponential of a spherical wave emitted from $\vec{r}'$, is then equal to $\vec{k}_f \cdot (\vec{r} - \vec{r}')$ of a plane wave,

$$G(\vec{r}, \vec{r}') \simeq -\frac{1}{4\pi} \frac{e^{i\vec{k}_f \cdot (\vec{r} - \vec{r}')}}{\left|\vec{r}\right|} \ . \tag{1.69}$$

In (1.69) we also assumed that the origin is near the scatterer, so $\left|\vec{r}\right| \gg \left|\vec{r}'\right|$, simplifying the denominator of our Green's function.[6]

Returning to our exact integral equation (1.67), we obtain the approximate scattered wave (the first Born approximation for the scattered wave) by using (1.68) and (1.69) in (1.67):

$$\Psi(\vec{r}) \simeq e^{i\vec{k}_i \cdot \vec{r}} - \frac{m}{2\pi\hbar^2} \int V(\vec{r}') \, e^{i\vec{k}_i \cdot \vec{r}'} \frac{e^{i\vec{k}_f \cdot (\vec{r} - \vec{r}')}}{\left|\vec{r}\right|} d^3\vec{r}' \ , \tag{1.70}$$

$$\Psi(\vec{r}) = e^{i\vec{k}_i \cdot \vec{r}} - \frac{m}{2\pi\hbar^2} \frac{e^{i\vec{k}_f \cdot \vec{r}}}{\left|\vec{r}\right|} \int V(\vec{r}') \, e^{i(\vec{k}_i - \vec{k}_f) \cdot \vec{r}'} d^3\vec{r}' \ . \tag{1.71}$$

If we define:[7]

$$\vec{Q} \equiv \vec{k}_i - \vec{k}_f \ , \tag{1.72}$$

$$\Psi(\vec{r}) = e^{i\vec{k}_i \cdot \vec{r}} - \frac{m}{2\pi\hbar^2} \frac{e^{i\vec{k}_f \cdot \vec{r}}}{\left|\vec{r}\right|} \int V(\vec{r}') \, e^{i\vec{Q} \cdot \vec{r}'} d^3\vec{r}' \ . \tag{1.73}$$

The scattered part of the wave is:

$$\Psi_{sc}(\vec{Q}, \vec{r}) = \frac{e^{i\vec{k}_f \cdot \vec{r}}}{\left|\vec{r}\right|} f(\vec{Q}) \ , \text{ where :} \tag{1.74}$$

$$f(\vec{Q}) \equiv -\frac{m}{2\pi\hbar^2} \int V(\vec{r}') \, e^{i\vec{Q} \cdot \vec{r}'} d^3\vec{r}' \ . \tag{1.75}$$

The factor $f(\vec{Q})$ is the scattering factor of (1.50), which we have found to depend on the incident and outgoing wavevectors only through their difference, $\vec{Q} \equiv \vec{k}_i - \vec{k}_f$. We recognize the integral of (1.75) as the Fourier transform of the potential seen by the incident neutron as it goes through the scatterer. In the first Born approximation:

The scattered wave is proportional to the
Fourier transform of the scattering potential.

[6] If we neglect a constant prefactor, this assumption of $\left|\vec{r} - \vec{r}'\right| = \left|\vec{r}\right|$ is equivalent to assuming that the scatterer is small compared to the distance to the detector.

[7] Sadly, the diffraction vector for elastic scattering is defined as $-\vec{Q} = \Delta\vec{k} \equiv \vec{k}_f - \vec{k}_i$

The factor $f\left(\vec{Q}\right)$ of (1.75) is given various names, depending on the potential $V(\vec{r})$ (we changed notation: $\vec{r}' \to \vec{r}$). When $V(\vec{r})$ is the potential of a single atom, $V_{\rm at}(\vec{r})$, we define $f_{\rm at}\left(\vec{Q}\right)$ as the "atomic form factor":

$$f_{\rm at}\left(\vec{Q}\right) \equiv -\frac{m}{2\pi\hbar^2}\int V_{\rm at}(\vec{r})\, e^{i\vec{Q}\cdot\vec{r}} d^3\vec{r}\,. \tag{1.76}$$

Alternatively, we can use the potential for the entire crystal for $V(\vec{r})$ in (1.75) (and develop the kinematical theory of diffraction). When $V(\vec{r})$ refers to the entire crystal, however, the first Born approximation of 1.73 is sometimes unreliable because multiple scattering invalidates the assumption of (1.68). This assumption is, nevertheless, the basis for the "kinematical theory of diffraction," which we develop for its clarity and its qualitative successes. It is possible to transcend formally the single scattering approximation, and develop a "dynamical theory" of neutron diffraction by considering higher-order Born approximations, but this has not proved a particularly fruitful direction. Modern dynamical theories take a completely different path.

1.4.3 Higher-Order Born Approximations

Nevertheless, it is not difficult in principle to extend the Born approximation to higher orders. Instead of using an undiminished plane wave for $\Psi(\vec{r}')$, we could use a $\Psi(\vec{r}')$ that has been scattered once already. Equation (1.67) gives the second Born approximation if we use do not use the plane wave of (1.68) for $\Psi(\vec{r}')$, but rather:

$$\Psi(\vec{r}') = e^{i\vec{k}_i\cdot\vec{r}'} + \frac{2m}{\hbar^2}\int V(\vec{r}'')\,\Psi(\vec{r}'')\,G(\vec{r}',\vec{r}'')\,d^3\vec{r}''\,, \tag{1.77}$$

where we now use a plane wave for $\Psi(\vec{r}'')$:

$$\Psi(\vec{r}'') \simeq e^{i\vec{k}_i\cdot\vec{r}''}\,. \tag{1.78}$$

The second Born approximation involves two centers of scattering. The first is at $\vec{r}''$ and the second is at $\vec{r}'$ (as shown in Fig. 1.8). The second and higher Born approximations are not used very frequently. If the scatterer is strong enough to violate the condition of weak scattering used in the first Born approximation, the scattering will also violate the assumptions of the second Born approximation.

1.5 Correlation Function for Elastic Scattering – The Patterson Function

1.5.1 Overview

In much of Chapter 2, scattering theory has been developed by calculating the amplitude of the wave scattered from crystals with excitations or disorder. The

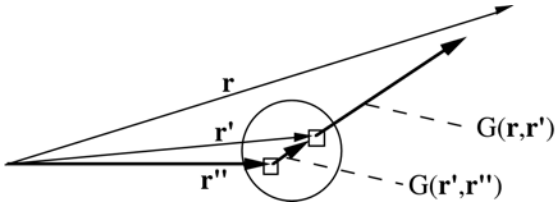


Figure 1.8: Coordinates for the second Born approximation. The neutron path is shown as the dark arrows, which are labeled by the relevant Green's functions.

amplitude of the diffracted wave, ψ, is the sum of phase factors of wavelets emitted from individual atoms. For elastic scattering, which we consider presently, the phase information in $\psi(\vec{Q})$ includes details of atom positions, which can be obtained by inverse Fourier transformation, $F^{-1}\psi$. We then calculate the intensity $I(\vec{Q}) = \psi^*\psi$.

This Sect. 1.5 takes a different approach of calculating directly the diffracted intensity $I(\vec{Q})$, rather than calculating it as $\psi^*\psi$. In this new approach, the real space information is obtained with the Fourier inversion $F^{-1}I$, rather than $F^{-1}\psi$, but this sacrifices some information about atom positions. Nevertheless, the intensity is the actual quantity measured in a diffraction experiment, so this new approach offers a more rigorous understanding of what structural information is available from diffraction experiments. Furthermore, in cases of severely disordered materials, there may be no obvious way to obtain the atom positions needed for a calculation of $\psi(\vec{Q})$. For problems involving severe structural disorder, another advantage of direct manipulations of $I(\vec{Q})$ is that a convenient reference state proves to be a homogeneous distribution of scatterers, or uncorrelated scatterers as in an ideal gas. A powerful tool for calculating diffraction intensities from such materials (and regular crystals too) is the "Patterson function," defined in Sect. 1.5.3 as an autocorrelation function of the scattering factor distribution.

Whereas the diffracted wave, $\psi(\vec{Q})$, is the Fourier transform of the scattering factor distribution, the diffracted intensity, $I(\vec{Q})$, is the Fourier transform of the Patterson function of the scattering factor distribution.

The Patterson function is a function in real space, with argument $\vec{r}$. The Patterson function is a convolution, so the reader should be familiar with convolutions and the convolution theorem (Sect. A.1) before reading the present chapter. The presentation here of real-space correlation functions is good preparation for the discussion that follows on space-time correlation functions. We begin by proving the emphasized statement above. The subsequent section uses the Patterson function to explain diffraction phenomena involving displacements of atoms off of periodic positions owing to temperature.

1.5.2 Atom Centers at Points in Space

The most important results in this chapter are obtained by assuming the scatterers are points. At each point, $\vec{r}_j$, resides the scattering strength of one entire atom, $f_{\vec{r}_j}$ (or one unit cell). The actual shape of the atom is included later by convolution, and does not change the main results obtained with point atoms.

It proves convenient to consider a distribution of scatterers, $f(\vec{r})$, with a continuous variable, $\vec{r}$, rather than a sum over discrete points, $\{\vec{r}_j\}$. We change variables as:

$$\psi(\vec{Q}) = \sum_{\vec{r}_j}^{N} f_{\vec{r}_j} \mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\vec{r}_j} = \int_{-\infty}^{\infty} f(\vec{r})\, \mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\vec{r}} \mathrm{d}^3\vec{r}\,. \tag{1.79}$$

To equate a continuous integral to a discrete sum requires that $f(\vec{r})$ is not a smooth function of position. Over most of space $f(\vec{r})$ is zero, but at atom centers such as $\vec{r} = \vec{r}_i$, $f(\vec{r}_i)$ is a Dirac delta function times a constant, $f_{\vec{r}_i}$:

$$f(\vec{r}_i) = f_{\vec{r}_i}\delta(\vec{r} - \vec{r}_i)\ . \tag{1.80}$$

Recall the important property of the Dirac delta function:

$$y(x') = \int_{-\infty}^{\infty} \delta(x - x')\, y(x)\, \mathrm{d}x\ . \tag{1.81}$$

Equation (1.81) requires that $\delta(x - x')$ is zero everywhere, except at the point $x = x'$. At this point the delta function is infinitely high, but of unit area, so the integral of (1.81) picks out only the value of $y(x)$ at x'. To extend (1.80) to include many atom centers, we take the sum over $\vec{r}_j$:

$$f(\vec{r}) = \sum_{\vec{r}_j}^{N} f_{\vec{r}_j}\delta(\vec{r} - \vec{r}_j)\ , \tag{1.82}$$

so we satisfy the equality in (1.79) between points in space, $\{\vec{r}_j\}$, and a continuous function of $\vec{r}$. We include the shape of the atomic form factor, $f_{\mathrm{at}}(\vec{r})$, in Sect. 1.5.5.

1.5.3 Definition of the Patterson Function

We define the "Patterson function," $P(\vec{r})$:

$$P(\vec{r}) \equiv \int_{-\infty}^{\infty} f^*(\vec{r}')f(\vec{r} + \vec{r}')\, \mathrm{d}^3\vec{r}'\ . \tag{1.83}$$

Equation (1.83) is a convolution. Since the function $f(\vec{r})$ is not inverted in the usual way for a convolution, we write:

$$P(\vec{r}) = f^*(\vec{r}) * f(-\vec{r})\ , \tag{1.84}$$

This is a specific type of convolution known as an "autocorrelation function," sometimes denoted with a special symbol:

$$P(\vec{r}) = f(\vec{r}) \circledast f(\vec{r}) \; . \tag{1.85}$$

The most important feature of the Patterson function is that its Fourier transform is the diffracted intensity in kinematical theory. To show this, we use (1.79) to write $I(\vec{Q}) = \psi^* \psi$ as:

$$I(\vec{Q}) = \int_{-\infty}^{\infty} f^*(\vec{r}\,) \, \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}} \, \mathrm{d}^3\vec{r} \int_{-\infty}^{\infty} f(\vec{r}\,') \, \mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\vec{r}\,'} \, \mathrm{d}^3\vec{r}\,' \; . \tag{1.86}$$

Since $\vec{r}$ and $\vec{r}\,'$ are independent variables:

$$I(\vec{Q}) = \int_{-\infty}^{\infty} \Big(\int_{-\infty}^{\infty} f^*(\vec{r}\,) f(\vec{r}\,') \, \mathrm{e}^{-\mathrm{i}\vec{Q}\cdot(\vec{r}\,'-\vec{r}\,)} \mathrm{d}^3\vec{r}\,' \Big) \, \mathrm{d}^3\vec{r} \; . \tag{1.87}$$

Define $\vec{r} \equiv \vec{r}\,' - \vec{r}$, and change variables $\vec{r}\,' \rightarrow \vec{r} + \vec{r}$. In so doing, the limits of integration for $\vec{r}$ are shifted by $-\vec{r}$, but this is not of concern for integrations performed over all of space:

$$I(\vec{Q}) = \int_{-\infty}^{\infty} \Big(\int_{-\infty}^{\infty} f^*(\vec{r}\,) f(\vec{r}+\vec{r}\,) \, \mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\vec{r}} \mathrm{d}^3\vec{r} \Big) \, \mathrm{d}^3\vec{r} \; , \tag{1.88}$$

$$I(\vec{Q}) = \int_{-\infty}^{\infty} \Big(\int_{-\infty}^{\infty} f^*(\vec{r}\,) f(\vec{r}+\vec{r}\,) \mathrm{d}^3\vec{r} \Big) \mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\vec{r}} \mathrm{d}^3\vec{r} \, . \tag{1.89}$$

Using the definition of (1.83), we rewrite (1.89):

$$I(\vec{Q}) = \int_{-\infty}^{\infty} P(\vec{r}) \, \mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\vec{r}} \mathrm{d}^3\vec{r} \, . \tag{1.90}$$

Equation (1.90) shows that the diffracted intensity is the Fourier transform of the Patterson function:

$$I(\vec{Q}) = FP(\vec{r}\,) \; , \tag{1.91}$$

and by the inverse transformation we must have:

$$P(\vec{r}\,) = F^{-1} I(\vec{Q}) \; . \tag{1.92}$$

For comparison, the diffracted wave, $\psi(\vec{Q})$ of (1.79), is the Fourier transform of the scattering factor distribution, $f(\vec{r}\,)$. We therefore have another relationship between $I(\vec{Q})$ and $f(\vec{r}\,)$:

$$I(\vec{Q}) = \psi^*(\vec{Q}) \, \psi(\vec{Q}) \; , \tag{1.93}$$

$$I(\vec{Q}) = \Big(Ff(\vec{r}\,) \Big)^* \, Ff(\vec{r}\,) = \big| Ff(\vec{r}\,) \big|^2 \; . \tag{1.94}$$

Comparing (1.91) and (1.94):

$$FP(\vec{r}) = \left|Ff(\vec{r})\right|^2 . \tag{1.95}$$

Equation (1.95) is consistent with the convolution theorem of Sect. A.1 – a convolution in real space (the Patterson function of (1.83)) corresponds to a multiplication in Fourier space (right-hand side of (1.95)). Note how (1.94) shows the effects of the flip and the complex conjugation of $f(\vec{r})$ in the convolution of (1.83):

$$F\Big[f^*(\vec{r}) * f(-\vec{r})\Big] = \Big(Ff(\vec{r})\Big)^* \, Ff(\vec{r}) = \left|f(\vec{Q})\right|^2 , \tag{1.96}$$

as compared to:

$$F\Big[f(\vec{r}) * f(\vec{r})\Big] = Ff(\vec{r}) \, Ff(\vec{r}) = \Big(f(\vec{Q})\Big)^2 . \tag{1.97}$$

1.5.4 Properties of Patterson Functions

It is instructive to illustrate the steps in constructing a Patterson function (1.83). The steps in any convolution are shift, multiply, and integrate, and are shown in Fig. 1.9. Figure 1.9a shows the overlap of a function shifted by the distance $\vec{r}$ against the original position shown as a dashed curve. To obtain the Patterson function in Fig. 1.9b, at each shift the function was multiplied by its shifted counterpart, then integrated.

Note that the peaks of the Patterson function in Fig. 1.9b are broader than the peaks in the scattering factor distribution of Fig. 1.9a. Since the peaks in Fig. 1.9a are Gaussian functions of equal width, the peaks in the Patterson function are broadened by a factor of $\sqrt{2}$. Second, the periodicity of the Patterson function is one lattice constant, a. This is expected, since the overlap of the peaks in the function of Fig. 1.9a is maximized each time the shift equals an integral number of lattice constants. The intensities of these primary maxima are proportional to $A^2 + B^2$. There are secondary maxima that occur at shifts of ± 0.3 when the large peak overlaps the small peak. The intensities of these secondary maxima are proportional to AB. The Patterson function has a peak at each distance corresponding to a separation between the peaks in Fig. 1.9a.

The Patterson function, $P(\vec{r})$ of Fig. 1.9b, has a higher symmetry than the $f(\vec{r})$ of Fig. 1.9a. Identical secondary peaks occur in $P(\vec{r})$ when the large peak is shifted to the right by $+0.3a$ and overlaps the small peak, or when the small peak is shifted to the left by $-0.3a$ and overlaps the large peak. For this reason, even when $f(\vec{r})$ has no center of inversion, $P(\vec{r})$ has inversion symmetry. The Patterson function is unchanged if the original function is inverted.[8] Equation (1.92) shows that the measured x-ray diffraction intensity provides the Patterson function, not the scattering factor distribution. We therefore have "Friedel's law":

[8] You can obtain the same $P(\vec{r})$ by taking the mirror image of the $f(\vec{r})$ in Fig. 1.9a (with the small peak to the immediate left of the large peak), and repeating the construction.

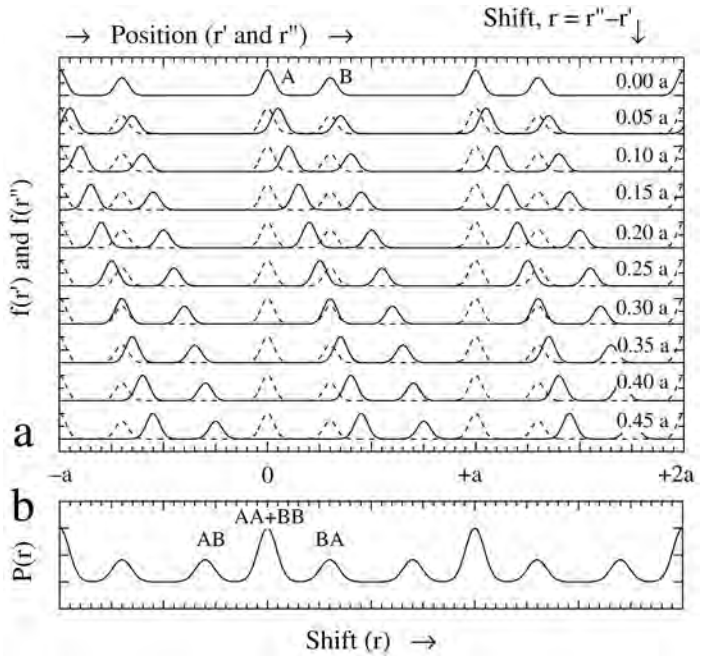


Figure 1.9: (**a**) Shifts of a function of period, a, with respect to itself. The shift, $r = r'' - r'$, is labeled at right in units of a. (**b**) The Patterson function, obtained by integrating the product of the solid and dashed curves for all shifts, r.

Diffraction experiments cannot distinguish between an atom arrangement and the atom arrangement when it is inverted.

This is sometimes called the "phase problem" in structure determination, since the phase of the diffracted wave $\psi(\vec{Q})$ is not measured, only its intensity, $\psi^*\psi$.

1.5.5 Perfect Crystals

In working problems with Patterson functions, it is often convenient to write the scattering factor distribution for an entire crystal, $f(\vec{r})$, in the following way:

$$f(\vec{r}) = f_{\text{at}}(\vec{r}) * \sum_{\vec{R}_n} \delta(\vec{r} - \vec{R}_n) \; . \tag{1.98}$$

Here $f_{\text{at}}(\vec{r})$ is the form factor of one atom. In (1.98) the form factor of the atom is convoluted with a sum of delta functions, each centered at a different atom site, $\vec{R}_n$. We evaluate (1.98) by first writing explicitly the convolution:

$$f(\vec{r}) = \int_{-\infty}^{\infty} f_{\text{at}}(\vec{r}') \sum_{\vec{R}_n} \delta\Big(\vec{r} - (\vec{r}' - \vec{R}_n)\Big) \mathrm{d}^3\vec{r}' \; . \tag{1.99}$$

Rearranging the operations on independent variables:

$$f(\vec{r}) = \sum_{\vec{R}_n} \int_{-\infty}^{\infty} f_{\text{at}}(\vec{r}') \, \delta\Big(\vec{r} - (\vec{r}' - \vec{R}_n)\Big) \mathrm{d}^3\vec{r}' \ . \tag{1.100}$$

The integral of (1.100) serves to pick out the value of $f_{\text{at}}(\vec{r}')$ at the location of the delta function, cf., (1.81). By shifting the delta function continuously by $\vec{r}'$, the shape of $f_{\text{at}}(\vec{r})$ is generated around the center of each delta function. These centers are each atom site, $\vec{R}_n$, so after the integration of (1.100):

$$f(\vec{r}) = \sum_{\vec{R}_n} f_{\text{at}}(\vec{r} - \vec{R}_n) \ . \tag{1.101}$$

Please compare (1.98) and (1.101).

The Patterson function of an infinite one-dimensional perfect crystal, $P_0(x)$, is:

$$P_0(x) = f^*(x) * f(-x) \ , \tag{1.102}$$

which we write using (1.98) for N atoms:

$$P_0(x) = \Big(f^*_{\text{at}}(x) * \sum_{n'=-\infty}^{+\infty} \delta(x - n'a)\Big) * \Big(f_{\text{at}}(-x) * \sum_{n''=+\infty}^{-\infty} \delta(n''a - x)\Big) \ . \tag{1.103}$$

Convolutions are commutative and associative, so we rearrange (1.103):

$$\begin{aligned} P_0(x) = \Big(f^*_{\text{at}}(x) * f_{\text{at}}(-x)\Big) \quad & * \quad \Big(\sum_{n'=-\infty}^{+\infty} \delta(x - n'a)\Big) \\ & * \quad \Big(\sum_{n''=+\infty}^{-\infty} \delta(n''a - x)\Big) \ . \end{aligned} \tag{1.104}$$

Recall that a convolution of two functions requires a shift, overlap, multiplication, and integration. Because the δ-functions are infinitesimally narrow, there is zero overlap of the two series of δ-functions unless the shift, x, satisfies the condition $x = na$, where n is an integer. Therefore:

$$\Big(\sum_{n'=-\infty}^{\infty} \delta(x - n'a)\Big) * \Big(\sum_{n'=-\infty}^{\infty} \delta(x - n'a)\Big) = N'\Big(\sum_{n=-\infty}^{\infty} \delta(x - na)\Big) \ . \tag{1.105}$$

Here $N' = \infty$, which is as expected for an infinite number of overlaps of an infinite chain of atoms. For a chain of N atoms, the Patterson function is:

$$P_0(x) = N\Big(f^*_{\text{at}}(x) * f_{\text{at}}(-x)\Big) * \Big(\sum_{n=-\infty}^{\infty} \delta(x - na)\Big) \ , \tag{1.106}$$

The Fourier transformation of $P_0(x)$ provides the diffracted intensity, $I(Q)$. By the convolution theorem of Sect. A.1, the two convolutions and one multiplication of (1.106) become, after Fourier transformation, two multiplications and one convolution. Using (1.96):

$$I(Q) = N\big|f_{\rm at}(Q)\big|^2 * F\Big[\sum_{n'=-\infty}^{\infty} \delta(x - n'a)\Big] . \tag{1.107}$$

The Fourier transform of the δ-function series is:

$$F\Big[\sum_{n'=-\infty}^{\infty} \delta(x - n'a)\Big] = \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}Qx} \sum_{n'=-\infty}^{\infty} \delta(x - n'a)\, \mathrm{d}x . \tag{1.108}$$

The condition $Qa = 2\pi h$ (where h is an integer) must be satisfied, or the integration over an infinite range of x is zero. The k-space Fourier transform is therefore zero except when $Q = 2\pi h/a$ precisely, so:

$$F\Big[\sum_{n'=-\infty}^{\infty} \delta(x - n'a)\Big] = N\sum_{h=-\infty}^{\infty} \delta(Q - 2\pi h/a) = N\sum_{g} \delta(Q - g) . \tag{1.109}$$

Here again N is the number of terms in the sum in (1.109). In a formal problem, N becomes a mathematical infinity, but it is useful to keep the N because it shows the proportionality to the size of the crystal. The diffraction intensity of (1.107) is:

$$I(Q) = N^2\big|f_{\rm at}(Q)\big|^2\Big[\sum_{h=-\infty}^{\infty} \delta(Q - 2\pi h/a)\Big] . \tag{1.110}$$

Equation (1.110) is a familiar result in a new form. The series of δ-functions gives the centers of the Bragg peaks from the crystal. These peaks are still sharp, but are attenuated at large Q by the atomic form factor intensity, $|f_{\rm at}(Q)|^2$.

1.5.6 Deviations from Periodicity

In many cases of interest, a scattering factor distribution, $f(\vec{r})$, can be expressed as the sum of a perfectly periodic function, $\langle f(\vec{r})\rangle$, plus a deviation function, $\Delta f(\vec{r})$, which provides the random or semi-random deviations from perfect periodicity. We know that the perfectly periodic function, $\langle f(\vec{r})\rangle$, provides sharp Bragg diffractions, but how does the deviation function, $\Delta f(\vec{r})$, affect the diffracted intensity? To find out, we calculate the Patterson function of $f(\vec{r})$:

$$f(\vec{r}) = \langle f(\vec{r})\rangle + \Delta f(\vec{r}) , \tag{1.111}$$

$$P(\vec{r}) \equiv f^*(\vec{r}) * f(-\vec{r}) , \tag{1.112}$$

$$\begin{aligned} P(\vec{r}) = {} & \langle f^*(\vec{r})\rangle * \langle f(-\vec{r})\rangle + \langle f^*(\vec{r})\rangle * \Delta f(-\vec{r}) \\ & + \Delta f^*(\vec{r}) * \langle f(-\vec{r})\rangle + \Delta f^*(\vec{r}) * \Delta f(-\vec{r}) . \end{aligned} \tag{1.113}$$

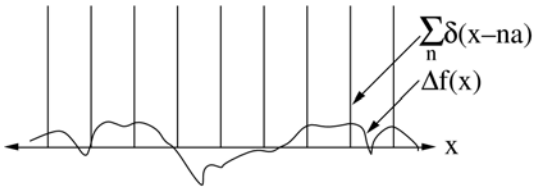


Figure 1.10: Overlap of periodic delta functions, $\sum_n \delta(x - na)$, with a random function of zero mean, $\Delta f(x)$. Since the deviation function $\Delta f(\vec{r})$ has zero mean and is non-periodic, the periodic delta functions overlap $\Delta f(-\vec{r})$ at as many positive values as negative values, demonstrating (1.116).

Look at the second term in (1.113). We rewrite it with the aid of (1.95):

$$\langle f^*(\vec{r})\rangle * \Delta f(-\vec{r}) = \Big[\langle f^*_{\rm at}(\vec{r})\rangle * \sum_{\vec{R}_n} \delta(\vec{r} - \vec{R}_n)\Big] * \Delta f(-\vec{r}) \ . \tag{1.114}$$

Convolutions are associative, so we can group the second and third factors in (1.114), and consider the new convolution:

$$\sum_{\vec{R}_n} \delta(\vec{r} - \vec{R}_n) * \Delta f(-\vec{r}) = \sum_{\vec{R}_n} \Delta f(-\vec{R}_n) \ , \tag{1.115}$$

where we used (1.81) in the same way as for (1.100)–(1.101). We assume that the deviation function, $\Delta f(-\vec{R}_n)$, has zero mean value.[9] Therefore:

$$\sum_{\vec{R}_n} \delta(\vec{r} - \vec{R}_n) * \Delta f(-\vec{r}) = 0 \ . \tag{1.116}$$

The second term for $P(\vec{r})$ in (1.113) is therefore zero (see also Fig. 1.10). Because $\vec{R}_n$ has precise periodicity over an infinite distance, (1.116) also holds true when $\Delta f(\vec{r})$ has short-range structure. By the same argument, the third term in (1.113) is also zero. Equation (1.113) becomes:

$$P(\vec{r}) = \langle f^*(\vec{r})\rangle * \langle f(-\vec{r})\rangle + \Delta f^*(\vec{r}) * \Delta f(-\vec{r}) \ . \tag{1.117}$$

The Patterson function for an alloy with disorder is reduced to two parts defined as the two terms in (1.117): 1) a Patterson function from the average crystal, $P_{\rm avge}(\vec{r})$, and 2) a Patterson function from the deviation crystal, $P_{\rm devs}(\vec{r})$:

$$P(\vec{r}) = P_{\rm avge}(\vec{r}) + P_{\rm devs}(\vec{r}) \ . \tag{1.118}$$

The diffracted intensity is the Fourier transform of the Patterson function of the alloy:

$$I(\vec{Q}) = F\Big[P_{\rm avge}(\vec{r}) + P_{\rm devs}(\vec{r})\Big] \ , \tag{1.119}$$

[9] This does not restrict generality because any non-zero mean could have been transferred into $\langle f(\vec{r})\rangle$ in (1.111).

and since Fourier transforms are distributive:

$$I(\vec{Q}) = F\Big[P_{\text{avge}}(\vec{r})\Big] + F\Big[P_{\text{devs}}(\vec{r})\Big] \,. \tag{1.120}$$

Equation (1.120) shows that the diffraction patterns from the average crystal, $\langle f(\vec{r})\rangle$, and the deviation crystal, $\Delta f(\vec{r})$, are additive. In terms of the diffracted waves from these average and deviation crystals (cf., (1.95)):

$$I(\vec{Q}) = \Big|F\left\langle f(\vec{r})\right\rangle\Big|^2 + \Big|F\Big[\Delta f(\vec{r})\Big]\Big|^2 \,. \tag{1.121}$$

We are familiar with the first term in (1.121), $|F\left\langle f(\vec{r})\right\rangle|^2$, which gives the sharp Bragg diffractions from the average crystal.

The second term in (1.121), $\Big|F\Big[\Delta f(\vec{r})\Big]\Big|^2$, is new. It is often a broad, diffuse intensity, as we show next. We will also show that with increasing disorder and larger $\Delta f(\vec{r})$, the sharp Bragg diffractions become weaker, and the diffuse intensity becomes stronger. Two important sources of $\Delta f(\vec{r})$ in a crystalline alloy are atomic displacement disorder and chemical disorder. Atomic displacement disorder comprises small deviations of atoms from the sites of a perfect crystal. These displacements may be static, or dynamic as in the case of thermal motion. Chemical disorder exists when there is randomness in the species of atoms that occupy the sites of a crystal. We consider these two types of disorder in sequence.

1.5.7 Coherent and Incoherent Scattering

The same argument of Section 1.5.6 can be used to understand the balance of coherent and incoherent neutron scattering. The idea is that the scattering length b_i from atom i contains a part that is the same for all nuclei, $\langle b\rangle$, and a deviation part, δb_i. Since this δb_i is a deviation from the average, it is both positive and negative. When averaged over all atoms, the sum of the deviations is zero, i.e.,

$$\langle \delta b_i\rangle = 0 \,. \tag{1.122}$$

The scattering length at atom i is

$$b_i = \langle b\rangle + \delta b_i \,. \tag{1.123}$$

The scattered neutron wavefunction is constructed as usual

$$\psi(\vec{Q}) = \sum_i b_i \, \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}_i} \,, \tag{1.124}$$

$$\psi(\vec{Q}) = \sum_i \left[\langle b\rangle + \delta b_i\right] \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}_i} \,. \tag{1.125}$$

The intensity, $\psi^*\psi$, is

$$I(\vec{Q}) = \sum_i [\langle b\rangle + \delta b_i]\, e^{-i\vec{Q}\cdot\vec{r}_i} \sum_j \left[\langle b\rangle + \delta b_j\right] e^{i\vec{Q}\cdot\vec{r}_j} , \tag{1.126}$$

$$I(\vec{Q}) = \sum_i \sum_j \left[\langle b\rangle^2 + (\delta b_i + \delta b_j)\langle b\rangle + \delta b_i \delta b_j\right] e^{i\vec{Q}\cdot(\vec{r}_j - \vec{r}_i)} . \tag{1.127}$$

We define the pair distance $\vec{r}_k \equiv \vec{r}_j - \vec{r}_i$. Averaged over all pairs of atoms, the middle term gives zero. The last term $\delta b_i \delta b_j$ is also zero, except for the special case when $i = j$, which denotes the deviation at the same atom (so $\vec{r}_k = 0$)

$$I(\vec{Q}) = N \sum_k \langle b\rangle^2 e^{i\vec{Q}\cdot\vec{r}_k} + \sum_i \delta b_i^2 e^{i\vec{Q}\cdot 0} , \tag{1.128}$$

$$I(\vec{Q}) = N \sum_k \langle b\rangle^2 e^{i\vec{Q}\cdot\vec{r}_k} + \sum_i \delta b_i^2 . \tag{1.129}$$

Equation (1.129) separates the intensities of coherent scattering and incoherent scattering. The first term depends on the interference of waves scattered by the average atom. It gives a series of Bragg peaks. This is wave-like behavior from coherent scattering. The second term is independent of $\vec{Q}$ and does not depend on wave interference. The intensity of the second term increases in proportion to the number of atoms in the material, but the flat shape of the intensity in $\vec{Q}$ is the same for one atom or for N atoms. This is particle-like behavior from incoherent scattering.

Sometimes the second term in (1.129) is rewritten as follows

$$\langle b^2\rangle = \left\langle \left(\langle b\rangle + \delta b_i\right)^2 \right\rangle , \tag{1.130}$$

$$\langle b^2\rangle = \left\langle \langle b\rangle^2 + 2\langle b\rangle \delta b_i + \delta b_i^2 \right\rangle , \tag{1.131}$$

$$\langle b^2\rangle = \langle b\rangle^2 + 2\langle b\rangle\langle \delta b_i\rangle + \langle \delta b_i^2\rangle , \tag{1.132}$$

$$\langle b^2\rangle = \langle b\rangle^2 + \langle \delta b_i^2\rangle , \tag{1.133}$$

$$\langle \delta b_i^2\rangle = \langle b^2\rangle - \langle b\rangle^2 . \tag{1.134}$$

Substituting (1.134) into (1.129), recognizing that the last term in (1.129) can be expressed as the average of δb_i^2 over all N atoms

$$I(\vec{Q}) = N \sum_k \langle b\rangle^2 e^{i\vec{Q}\cdot\vec{r}_k} + N\left(\langle b^2\rangle - \langle b\rangle^2\right) . \tag{1.135}$$

We see that the scattered intensity includes a contribution that depends on the pair separations $\{\vec{r}_k\}$, and a featureless diffuse scattering. Written as a continuous variable, the pair distribution is the Patterson function $P(\vec{r})$, so

$$I(\vec{Q}) = N\langle b\rangle^2 \int_{-\infty}^{\infty} P(\vec{r})\, e^{i\vec{Q}\cdot\vec{r}}\, d^3\vec{r} + N\left(\langle b^2\rangle - \langle b\rangle^2\right) . \tag{1.136}$$

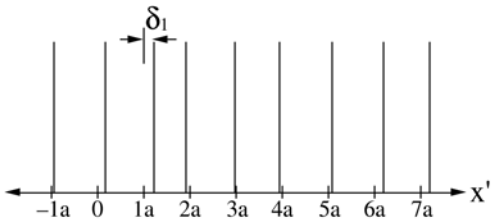


Figure 1.11: Atomic displacement disorder in a one-dimensional crystal.

1.5.8 Uncorrelated Displacements

Atomic displacement disorder exists when atoms do not sit precisely on the periodic sites of a crystal. Atomic size differences in an alloy cause static displacements from lattice sites, and thermal vibrations cause dynamic displacement disorder. Both cause diffuse scattering. Here we consider a simple type of displacement disorder where each atom has a small, random shift, δ, off its site of a periodic lattice as shown in Fig. 1.11.

For now we assume there are no correlations between the displacements, δ_j, of neighboring atoms.[10] The Patterson function, $f(x) * f(-x)$, for this displacement distribution is shown in Fig. 1.12a. To understand this Patterson function, consider the overlap of the atom center distribution with itself after a shift of $x = na + \xi$, where a is the lattice parameter, n is an integer, and ξ is a small distance (typically $\xi < a$). With no correlation between the displacements of neighboring atoms, the probability of overlap of two atom centers is the same for a shift of the crystal by many lattice constants, $na + \xi$, as it is for a shift of one lattice constant, $1a + \xi$. The important exception occurs around $x = 0$, i.e., when $n = 0$. All the atom centers overlap perfectly with themselves when ξ is exactly zero, but even for the smallest shift, $\xi \neq 0$, there are zero overlaps of atom centers.

The best way to work with the Patterson function in Fig. 1.12a is to break it into periodic and non-periodic parts (1.118), as shown in the two plots in Fig. 1.12b. The diffracted intensity from our crystal with displacement disorder is obtained from (1.120) as the sum of the Fourier transforms of these two functions, $P_{\text{avge}}(x)$ and $P_{\text{devs}}(x)$. The Fourier transform of $P_{\text{avge}}(x)$ is the well-known series of Bragg peaks. These peaks are suppressed at large values of Q owing to the breadths of the peaks in $P_{\text{avge}}(x)$ caused by displacement disorder (see top of Fig. 1.13). This suppression of the Bragg peaks at large Q is similar to the suppression caused by the atomic form factor, which also broadens the scattering centers of the atoms.

The Fourier transform of $P_{\text{devs}}(x)$ is new for us. To understand its contribution to the diffraction intensity, we split $P_{\text{devs}}(x)$ into two parts, $P_{\text{devs1}}(x)$ and $P_{\text{devs2}}(x)$ (Fig. 1.12c). The first, $P_{\text{devs1}}(x)$, is a Dirac delta function, whose Fourier transform is a constant in k-space ($F\left[P_{\text{devs1}}(x)\right]$ in Fig. 1.13). The second part, $P_{\text{devs2}}(x)$, is a short, broadened function with negative sign. (In Sect. 1.5.9 we

[10] For example, we assume that if one atom is displaced to the left, its neighbor to the right is equally likely to be displaced to the left or to the right.

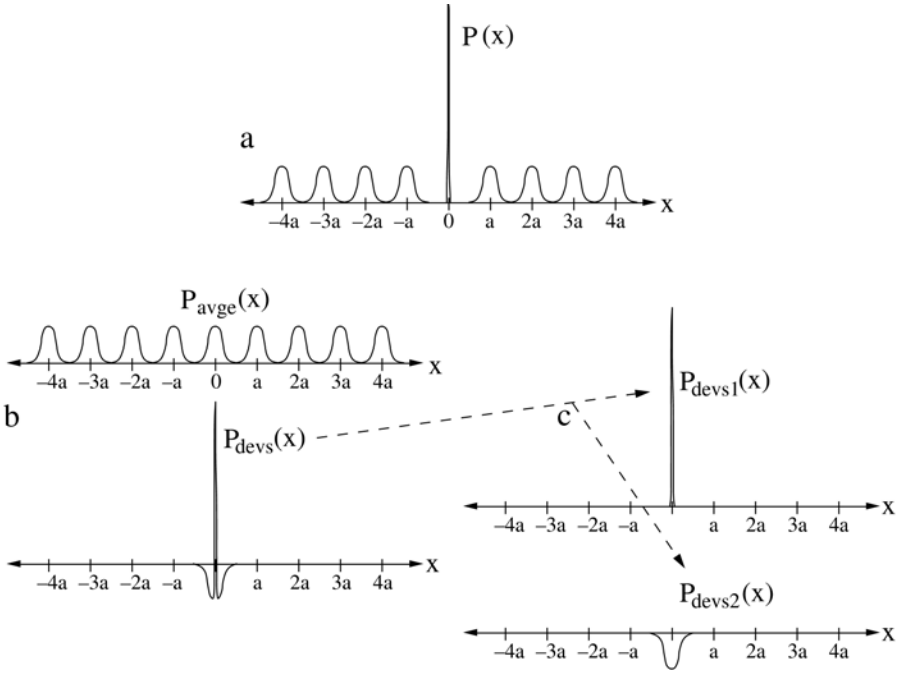


Figure 1.12: (**a**) Patterson function for the random displacements of Fig. 1.11 and (1.118). (**b**) The Patterson function at top is the sum of $P_{\text{avge}}(x)$ and $P_{\text{devs}}(x)$. (**c**) $P_{\text{devs}}(x)$ is the sum of $P_{\text{devs1}}(x)$ and $P_{\text{devs2}}(x)$.

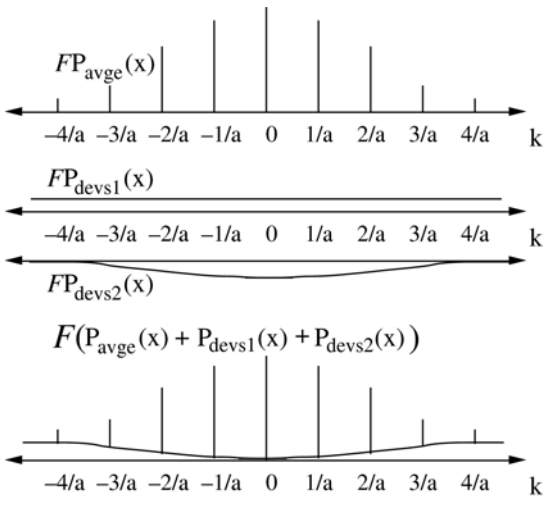


Figure 1.13: The Fourier transform of the Patterson functions of Fig. 1.12. Fourier transform of $P_{\text{avge}}(x)$ (top), Fourier transforms of the two components of $P_{\text{devs}}(x)$ (middle). The sum of all three components (bottom) is the diffraction intensity from our linear crystal with Gaussian displacement disorder.

will consider it to be a Gaussian function.) Its Fourier transform, $F\Big[P_{\text{devs2}}(x)\Big]$, is also shown in Fig. 1.13. The areas of these two parts, $P_{\text{devs1}}(x)$ and $P_{\text{devs2}}(x)$, are equal, since both arise from the same total number of atom-atom overlaps (equal to the number of atoms, N). This has an important consequence for the diffracted intensity at $Q = 0$:

$$I(Q=0) = \int_{-\infty}^{\infty} P_{\text{devs}}(x)\, \mathrm{e}^{-\mathrm{i}0x}\, \mathrm{d}x = \int_{-\infty}^{\infty} P_{\text{devs}}(x)\, \mathrm{d}x\ , \tag{1.137}$$

which is simply the area of the Patterson function, $P_{\text{devs}}(x)$. Since $P_{\text{devs1}}(x)$ and $P_{\text{devs2}}(x)$ have equal and opposite areas, at $Q = 0$ there is zero diffuse scattering from atomic displacement disorder.

The $F\Big[P_{\text{devs2}}(x)\Big]$ has a negative sign that decreases in magnitude with Q. The diffuse scattering therefore increases with Q, as the flat contribution originating from $F\Big[P_{\text{devs1}}(x)\Big]$ increasingly dominates over $F\Big[P_{\text{devs2}}(x)\Big]$. The function $P_{\text{devs2}}(x)$, incidentally, has the same shape as the individual peaks in $P_{\text{avge}}(x)$. In this case the Q-dependence of the rolloff of the Bragg peaks is the same as the Q-dependence of the diffuse scattering. The effects of displacement disorder increase with the characteristic size of the displacements, δ_j. The larger the characteristic δ, the faster the rolloff of the Bragg peaks with Q, and the greater the intensity of the diffuse scattering.

1.5.9 Temperature

During thermal vibrations, the distances between atoms undergo rapid changes with time. It is useful, however, to think of each x-ray scattering event as taking an instantaneous snapshot of the atom positions. The diffraction data are averages of many different instantaneous atom configurations. Over a large crystal, however, each instantaneous snapshot looks approximately the same. The same argument of the previous section on atomic displacement disorder is then appropriate for understanding the diffraction effects caused by thermal disorder in atom positions. This section uses a simple model of atom vibrations to calculate two effects of temperature:

- the Debye–Waller factor that causes the Bragg peaks to lose intensity,
- the thermal diffuse scattering, which is where the "lost" intensity reappears.[11]

A detailed analysis of thermal vibrations is not simple, because it should be performed with knowledge of the polarizations and numbers of all phonon modes. A complete analysis considers the contribution of each phonon to the relative separation of each atom-atom pair in the solid. In phonons with long

[11] The total coherent cross-section remains constant.

wavelengths, for example, neighboring pairs of atoms tend to move together.[12] In contrast, high frequency phonons affect strongly the mutual displacements of neighboring atoms. In addition, it is important to know how the atom motions within each phonon are oriented with respect to $\vec{Q}$ – atom motions nearly perpendicular to $\vec{Q}$ have weak effects on the scattering. Calculating the Patterson function from the densities of phonons with all polarizations is a problem for computers, and is beyond the scope of this book.

Thermal vibrations broaden the Patterson function of the scattering factor distribution. To develop a simple analytical model, we assume each atom center has a thermal spread around its crystal site that is a Gaussian function of characteristic width, σ. (A plausibility argument for a Gaussian function is provided in Appendix A.11.) For any n^{th} neighbor pair of atoms, we expect the vibrations of both atoms to affect the Patterson function. Suppose we place a stationary atom 1 at a point in space. When an atom 2 vibrates with respect to atom 1, there is a probability distribution for their interatomic separation, x:

$$p_{\text{atom2}}(x) = \frac{1}{\sqrt{\pi}\,\sigma}\,\mathrm{e}^{-x^2/\sigma^2}\ . \tag{1.138}$$

This function is shown schematically in Fig. 1.14a. Now let atom 1 vibrate. For every interatomic separation provided by the thermal motion of the atom 2, make a thermal distribution for the position of atom 1. To obtain the distribution of separations between atoms 1 and 2, the displacement distribution of atom 2 serves to weight the displacement distribution of atom 1. The various weights are shown in Fig. 1.14b, and the weighted sum of the net thermal distribution of atom 1 with respect to atom 2 is shown schematically in Fig. 1.14c as $p * p(x)$. The procedure we followed was in fact a convolution: the distribution of atom 1 was shifted, $p_{\text{atom1}}(x - x')$, multiplied by $p_{\text{atom2}}(x')$, and summed (integrated) over all values of x':

$$P_{\text{therm}}(x) = \int_{-\infty}^{\infty} p_{\text{atom2}}(x')\, p_{\text{atom1}}(x - x')\,\mathrm{d}x'\ . \tag{1.139}$$

For n^{th} neighbor pairs of atoms, the Patterson function of the thermal spread, $P_{\text{therm}}(x)$, is the convolution of the Gaussian thermal spread functions of both atoms (cf., (8.24)):

$$P_{\text{therm}}(x) = \left(\frac{1}{\sqrt{\pi}\sigma}\mathrm{e}^{-x^2/\sigma^2}\right) * \left(\frac{1}{\sqrt{\pi}\sigma}\mathrm{e}^{-x^2/\sigma^2}\right) \text{ when } n \neq 0\ . \tag{1.140}$$

$$P_{\text{therm}}(x) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-x^2/(2\sigma^2)} \text{ when } n \neq 0\ . \tag{1.141}$$

A detailed analysis treats closer pairs of atoms differently from more distant pairs, but here we ignore this difference except for the case of $n = 0$. In the

[12] Another aspect of the problem is that a crystal has fewer long-wavelength than short-wavelength vibrational modes. However, the lower energy of the long-wavelength modes means that their occupancy is higher at all temperatures, especially low temperatures.

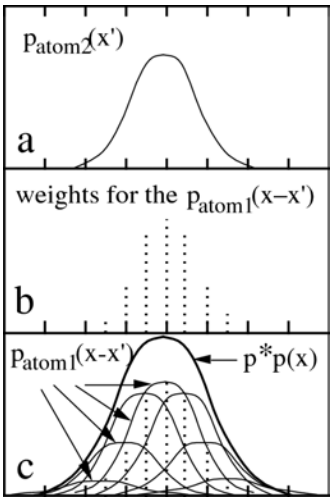


Figure 1.14: (**a**) The thermal spread of centers for atom 2. (**b**) Weights for the centers of the thermal distribution of the atom 1. (**c**) The distribution of all thermal separations of atom 1 with respect to atom 2.

special case of $n = 0$, we are considering the autocorrelation function between the positions *of the individual atoms with themselves*. Each atom sees itself as being at rest, so the Patterson function for the thermal spread is:

$$P_{\text{therm}}(x) = N\delta(x) \text{ when } n = 0\ . \tag{1.142}$$

We now obtain the Patterson function for the entire crystal by convoluting the thermal spread function, $P_{\text{therm}}(x)$, with the Patterson function of the perfect crystal (1.106). The Patterson function, $P(x)$, for the crystal with thermal displacement disorder is the following modification of (1.106). Note the special treatment of the $n = 0$ term, which provides the δ-function:

$$\begin{aligned} P(x) = \quad & N\Big[f^*_{\text{at}}(x) * f_{\text{at}}(-x)\Big] \\ & *\Big[\ \delta(x) + \Big(\sum_{n\neq 0; n=-\infty}^{n=\infty} \delta(x - na)\Big) * \Big(\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}\Big)\Big]\ . \end{aligned} \tag{1.143}$$

We rewrite the sum in (1.143) by adding and subtracting the $n = 0$ term (the same trick used in Fig. 1.12 to give an uninterrupted infinite series for $P_{\text{avge}}(x)$):

$$\begin{aligned} P(x) = \quad & N\Big[f^*_{\text{at}}(x) * f_{\text{at}}(-x)\Big] * \Big[\ \delta(x) - \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)} \\ & + \Big(\sum_{n=-\infty}^{n=\infty} \delta(x - na)\Big) * \Big(\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)}\Big)\Big]\ . \end{aligned} \tag{1.144}$$

The diffracted intensity is the Fourier transform of the Patterson function of (1.144). The transformation from (1.144) to (1.145) follows that from (1.106) to

(1.110), plus the fact that the Fourier transform of a Gaussian is a Gaussian:

$$I(Q) = N\left|f_{\text{at}}(Q)\right|^2\Big[\quad (1 - \mathrm{e}^{-\sigma^2(Q)^2/2}) + \mathrm{e}^{-\sigma^2(Q)^2/2}\sum_h \delta(Q - 2\pi h/a)\Big]\,. \tag{1.145}$$

The last term in the square brackets is the expected set of sharp Bragg peaks, but attenuated at larger values of Q by the "Debye–Waller factor," $D(Q)$:

$$D(\sigma, Q) = \mathrm{e}^{-\sigma^2(Q)^2/2}\,. \tag{1.146}$$

The Debye–Waller factor suppresses the intensity of Bragg peaks at high Q, as does the size of the atom through the factor $|f_{\text{at}}(Q)|^2$ of Sect. 3.3.2, so the Debye–Waller factor can be considered a "thermal fattening of the atoms." The intensity lost from the Bragg peaks reappears[13] as the first term in brackets in (1.145), $1 - \mathrm{e}^{-\sigma^2(Q)^2/2}$, which is the "thermal diffuse scattering." The thermal diffuse scattering has no distinct peaks, but usually has gradual modulations that increase with Q as shown in Fig. 1.13.

The Debye–Waller factor can provide quantitative information about the mean-squared displacement, $\left\langle x^2\right\rangle$, during thermal motion of the atoms. The larger is $\left\langle x^2\right\rangle$, the smaller the Debye–Waller factor (and the larger the suppression of the Bragg diffractions). We first relate $\left\langle x^2\right\rangle$ to the σ^2 in the thermal spread function of the individual atoms. This is the second moment of the Gaussian function:

$$\left\langle x^2\right\rangle = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{\pi}\,\sigma}\,\mathrm{e}^{-x^2/\sigma^2}\mathrm{d}x = \frac{1}{2}\sigma^2\,, \tag{1.147}$$

so from (1.146) and (5.20):

$$D(\sigma, Q) = \mathrm{e}^{-\left\langle x^2\right\rangle (Q)^2} = \mathrm{e}^{-\left\langle x^2\right\rangle (4\pi \sin\theta/\lambda)^2}\,. \tag{1.148}$$

At modest temperatures and small Q we can often linearize the exponential to predict a suppression of the Bragg peaks that is quadratic in Q:

$$D(\sigma, Q) \simeq 1 - \left\langle x^2\right\rangle \left(\frac{4\pi \sin\theta}{\lambda}\right)^2\,. \tag{1.149}$$

Physically, the Debye–Waller factor accounts for the loss of constructive interference in diffraction when the mean-squared atomic displacements become comparable to the x-ray wavelength. The Debye–Waller factor always suppresses the intensity of Bragg peaks.

[13] Never forget that the total cross-section for coherent scattering is constant.

Equations (1.148) or (1.149) can be used to determine $\langle x^2 \rangle$ from experimental data on diffraction intensities.[14] Conversely, it is often important to predict the Debye–Waller factor for a material at a known temperature. In essence, $\langle x^2 \rangle$ is proportional to the potential energy of a harmonic oscillator, and scales linearly with temperature, T. Although $\langle x^2 \rangle$ can be calculated easily for the single oscillator in the Einstein model, it is more handy to express the Debye–Waller factor in terms of a Debye temperature, θ_D, since tabulations of θ_D are conveniently available. For the Debye model the Debye–Waller factor has been worked out, and at temperatures comparable to the Debye temperature or higher, the Debye–Waller factor is:

$$D(T,Q) \simeq \exp\Big[\frac{-12h^2T}{m\,k_B\,\theta_D^2}\Big(\frac{\sin\theta}{\lambda}\Big)^2\Big], \tag{1.150}$$

$$D(T,\theta) \simeq 1 - \frac{22,800\,T}{m\,\theta_D^2}\Big(\frac{\sin\theta}{\lambda}\Big)^2 \tag{1.151}$$

Here the units of mass are the atomic weight (e.g., 55.847 for Fe), T and θ_D are in Kelvin, and λ is in Å. For use in (1.148) and (1.149), in the Debye model:

$$\langle x^2 \rangle = 144.38\,\frac{T}{m\,\theta_D^2}\ . \tag{1.152}$$

Although the Debye–Waller factor pertains to the thermal spread of distances between pairs of atoms, a Debye–Waller factor is often assigned to the scattering from a single atom. With this approximation, the atomic form factor, f, of each atom is replaced with $f\exp(-M)$. The Debye–Waller factor for the intensity is therefore $\exp(-2M)$. Also defined is the parameter B, related to $\langle x^2 \rangle$. Standard relationships are:

$$2M = \langle x^2 \rangle\Big(\frac{4\pi\sin\theta}{\lambda}\Big)^2, \tag{1.153}$$

$$M = B\Big(\frac{\sin\theta}{\lambda}\Big)^2. \tag{1.154}$$

In the case of an alloy, it is typical to assign different Debye–Waller factors for each type of atom, A or B, written as e^{-M_A} and e^{-M_B}:

$$\psi(\vec{Q}) = \sum_{\vec{r}}\Big[\,e^{-M_A} f_A\,\delta_A(\vec{r}) + e^{-M_B} f_B\,\delta_B(\vec{r})\Big]e^{i\vec{Q}\cdot\vec{r}}\ . \tag{1.155}$$

Here the δ-functions are Kroneker delta functions indicating the presence of an A or B atom at $\vec{r}$.

At temperatures below approximately half the Debye temperature, and especially below a quarter of the Debye temperature, (1.150) is no longer reliable

[14] Note that $\langle x^2 \rangle$ is along the direction of Q. In an isotropic material $\langle x^2 \rangle$ would equal 1/3 of the mean-squared atomic displacement.

for calculating the Debye–Waller factor. Two quantum effects are important at low temperatures. First, owing to Bose–Einstein phonon population statistics, the higher frequency phonons are not excited in simple proportion to the ratio kT/ε, where ε is the phonon energy. Second, at temperatures below about half the Debye temperature, the "zero-point" vibrations of the solid account for an increasingly large fraction of the atom displacements. Owing to zero-point vibrations, the thermal diffuse scattering can never be eliminated, even by cooling to arbitrarily low temperature.

The derivation of (1.145) was clean because we assumed the same Gaussian thermal spread for all interatomic correlations. For long wavelength phonons, however, adjacent atoms tend to move together in a group. In general, the nearest-neighbor pair correlations are less broadened than the correlations for more distant neighbor pairs. If atoms tend to move in groups, as in acoustic modes, the displacement has long-range modulations, and the thermal diffuse scattering intensity is concentrated near the reciprocal lattice points. The detailed shape of the thermal diffuse scattering depends on the lattice dynamics of the crystal vibrations [9.2]. With a Born–von Kármán model of lattice dynamics, for example, it is possible to calculate the projected components of the atom movements normal to the diffracting planes, and obtain a more accurate $P_{\text{therm}}(x)$ of (1.141). Alternatively, the phonon spectrum of the crystal can be deduced from measurements of the thermal diffuse scattering, at least in principle. In practice, such measurements require careful correction for other sources of diffuse intensity (such as atomic size and displacement effects).

Chapter 2

Inelastic Scattering

Elastic scattering involves momentum transfers and positions $\{\vec{Q}, \vec{r}\}$, which are complementary variables in quantum mechanics: $\vec{Q} \leftrightarrow \vec{r}$. Inelastic scattering is an extension into energy and time $\{\vec{Q}, E, \vec{r}, t\}$, which provides pairs of complementary variables: $(\vec{Q}, E) \leftrightarrow (\vec{r}, t)$. The amplitude of the scattered wave, ψ, is the sum of phase factors of wavelets emitted from individual atoms, but now we allow for a time variation. The phase information in $\psi(\vec{Q}, E)$ includes details of atom positions and their motions. Recall the case for elastic scattering, where we obtained by inverse Fourier transformation: $f(\vec{r}) = F^{-1}_{\vec{Q}}\psi(\vec{Q})$. For inelastic scattering the analogous result is: $f(\vec{r}, t) = F^{-1}_{\vec{Q}}F^{-1}_{E}\psi(\vec{Q}, E)$.

An inelastic experiment measures an intensity and not a wave amplitude. The experimental information on $\{\vec{r}, t\}$ can be obtained directly from the scattered intensity $I(\vec{Q}, E)$ by Fourier inversion $F^{-1}_{\vec{Q}}F^{-1}_{E}I(\vec{Q}, E)$. As for the case with the Patterson function for diffraction experiments, when we Fourier transform the intensity instead of the wave itself, there is a similar loss of information about atom positions and the phases of their motions. Nevertheless, the intensity is the actual quantity measured in a scattering experiment and we must make do with it. Inelastic scattering has an important analog to the Patterson function of elastic scattering, the "Van Hove function," defined in Sect. 2.1.1 as an autocorrelation function of the scattering factor distribution in space and time.

Whereas the inelastically scattered wave, $\psi(\vec{Q}, E)$, is the double Fourier transform of the moving scattering factor distribution, the scattered intensity, $I(\vec{Q}, E)$, is the double Fourier transform of the Van Hove correlation function of the scattering factor distribution.

After proving this emphasized statement, subsequent sections use the Van Hove function to explain scattering phenomena involving various dynamical excitations. Then starting anew with Fermi's Golden Rule, inelastic scattering is

calculated in a more precise and general way. The Van Hove space-time correlation function is developed by taking careful account of the non-commutivity of the position and momentum operators. A proper treatment of magnetic scattering is then presented. These latter sections are parallel to similar sections in the books by Squires and Lovesey, which are recommended to all practitioners of inelastic neutron scattering, especially persons inclined towards theory.

2.1 Correlation Function for Inelastic Scattering – The Van Hove Function

2.1.1 Concept of Van Hove Function

Preliminary: Atom Centers at Points in Space and Time

As was the case for elastic scattering, most of the essential results for inelastic scattering can be obtained by assuming the scatterers are points. The point scatterer emits a wavelet from position $\vec{r}_j$ at time t_k. This $\vec{r}_j$ and t_k provide the phase of the wavelet from the point emitter relative to wavelets from other point emitters. This phase is $\vec{Q}\cdot\vec{r}_j - \omega t_k$. The amplitude of the scattering is given by the scattering strength of the point emitter, $f(\vec{r}_j, t_k)$, which is an amplitude at a point in space and an instant in time. For a nucleus we may consider f as being at a point, although for magnetic spin distributions the shape of electron orbitals may be included later by convolution. For the distribution in time, we will usually consider a Fourier series with different frequencies, or energies $E = \hbar\omega$.

It proves convenient to consider a distribution of scatterers, $f(\vec{r}, t)$, with continuous variables, $\vec{r}$ and t, rather than a sum over discrete points, $\{\vec{r}_j\}$, and snapshots in time t_k. We change variables as:

$$\psi(\vec{Q}, E) = \sum_{\vec{r}_j}^{N} \sum_{t_k} f_{\vec{r}_j, t_k} \mathrm{e}^{-\mathrm{i}(\vec{Q}\cdot\vec{r}_j - \omega t)} = \iint_{-\infty}^{+\infty} f(\vec{r}, t)\, \mathrm{e}^{-\mathrm{i}(\vec{Q}\cdot\vec{r} - \omega t)} \mathrm{d}^3\vec{r}\,\mathrm{d}t \ . \tag{2.1}$$

To equate a continuous integral to a discrete sum requires that $f(\vec{r}, t)$ is not a smooth function of position or time. Over most of space and time, $f(\vec{r}, t)$ is zero, but when the scattering amplitude exists at $\vec{r} = \vec{r}_i$ and $t = t_k$, $f(\vec{r}_i, t_k)$ is a Dirac delta function times a constant, $f_{\vec{r}_j, t_k}$:

$$f(\vec{r}_j, t_k) = f_{\vec{r}_j, t_k} \delta(\vec{r} - \vec{r}_i)\delta(t - t_k) \ . \tag{2.2}$$

To extend (2.2) to include many atom centers, we take the sum over $\vec{r}_j$ and t_k:

$$f(\vec{r}, t) = \sum_{\vec{r}_j}^{N} \sum_{t_k} f_{\vec{r}_j, t_k} \delta(\vec{r} - \vec{r}_j)\delta(t - t_k) \ , \tag{2.3}$$

so we satisfy the equality in (2.1) between points in space and time, $\{\vec{r}_j, t_k\}$, and continuous functions of $\vec{r}, t$.

Definition of the Van Hove Function

We define the "Van Hove function," $G(\vec{r},t)$:

$$G(\vec{r},t) \equiv \iint_{-\infty}^{+\infty} f^*(\vec{r}',t')f(\vec{r}+\vec{r}',t+t')\,\mathrm{d}^3\vec{r}'\,\mathrm{d}t' \ . \tag{2.4}$$

Equation (2.4) is a double autocorrelation function – a space-time correlation function with limits of integration over all space and all time.

A most important feature of the Van Hove function is that its Fourier transform is the scattered intensity in kinematical theory. To show this, we use (2.1) to write $I(\vec{Q},E) = \psi^*(\vec{Q},E) \times \psi(\vec{Q},E)$ as:

$$\begin{aligned} I(\vec{Q},E) &= \iint_{-\infty}^{+\infty} f^*(\vec{r}',t')\,\mathrm{e}^{\mathrm{i}(\vec{Q}\cdot\vec{r}'-\omega t')}\mathrm{d}^3\vec{r}'\,\mathrm{d}t' \\ &\times \iint_{-\infty}^{+\infty} f(\vec{r}'',t'')\,\mathrm{e}^{-\mathrm{i}(\vec{Q}\cdot\vec{r}''-\omega t'')}\mathrm{d}^3\vec{r}''\mathrm{d}t'' \ . \end{aligned} \tag{2.5}$$

Since $\vec{r}'$ and $\vec{r}''$ are independent variables, as are t' and t'':

$$\begin{aligned} I(\vec{Q},E) &= \iint_{-\infty}^{+\infty} \Big(\iint_{-\infty}^{+\infty} f^*(\vec{r}',t')f(\vec{r}'',t'') \\ &\qquad \times \mathrm{e}^{-\mathrm{i}[\vec{Q}\cdot(\vec{r}''-\vec{r}')-\omega(t''-t')]}\ \mathrm{d}^3\vec{r}''\mathrm{d}t''\Big)\,\mathrm{d}^3\vec{r}'\,\mathrm{d}t' \ . \end{aligned} \tag{2.6}$$

Define $\vec{r} \equiv \vec{r}'' - \vec{r}'$ and $t \equiv t'' - t'$, and change variables $\vec{r}'' \to \vec{r}+\vec{r}'$ and $t'' \to t+t'$. In so doing, the limits of integration for $\vec{r}$ are shifted by $-\vec{r}'$ and $-t'$, but this is not of concern for integrations performed over all of space and all of time:

$$\begin{aligned} I(\vec{Q},E) &= \iint_{-\infty}^{+\infty} \\ &\times \Big(\iint_{-\infty}^{+\infty} f^*(\vec{r}',t')f(\vec{r}+\vec{r}',t+t')\,\mathrm{e}^{-\mathrm{i}(\vec{Q}\cdot\vec{r}-\omega t)}\,\mathrm{d}^3\vec{r}\mathrm{d}t\Big)\mathrm{d}^3\vec{r}'\,\mathrm{d}t' \ , \end{aligned} \tag{2.7}$$

$$\begin{aligned} I(\vec{Q},E) &= \iint_{-\infty}^{+\infty} \Big(\iint_{-\infty}^{+\infty} f^*(\vec{r}',t')f(\vec{r}+\vec{r}',t+t')\mathrm{d}^3\vec{r}'\,\mathrm{d}t\Big) \\ &\qquad \times \mathrm{e}^{-\mathrm{i}(\vec{Q}\cdot\vec{r}-\omega t)}\,\mathrm{d}^3\vec{r}\mathrm{d}t \ . \end{aligned} \tag{2.8}$$

Using the definition of (2.4), we rewrite (2.8):

$$I(\vec{Q},E) = \iint_{-\infty}^{+\infty} G(\vec{r},t)\,\mathrm{e}^{-\mathrm{i}(\vec{Q}\cdot\vec{r}-\omega t)}\,\mathrm{d}^3\vec{r}\mathrm{d}t \ . \tag{2.9}$$

Equation (2.9) shows that the scattered intensity is the Fourier transform of the Van Hove function:

$$I(\vec{Q}, E) = F_{\vec{r}} F_t G(\vec{r}, t) \ , \tag{2.10}$$

and by the inverse transformation we must have:

$$G(\vec{r}, t) = F_{\vec{Q}} F_E I(\vec{Q}, E) \ . \tag{2.11}$$

For comparison, the scattered wave, $\psi(\vec{Q}, E)$ of (2.1), is the Fourier transform of the scattering factor distribution, $f(\vec{r}, t)$. We therefore have another relationship between $I(\vec{Q}, E)$ and $f(\vec{r}, t)$:

$$I(\vec{Q}, E) = \psi^*(\vec{Q}, E)\, \psi(\vec{Q}, E) \ , \tag{2.12}$$

$$I(\vec{Q}, E) = \left(F_{\vec{r}} F_t f(\vec{r}, t)\right)^* F_{\vec{r}} F_t f(\vec{r}, t) = \left|F_{\vec{r}} F_t f(\vec{r}, t)\right|^2 \ . \tag{2.13}$$

Comparing (2.10) and (2.13):

$$F_{\vec{r}} F_t G(\vec{r}, t) = \left|F_{\vec{r}} F_t f(\vec{r}, t)\right|^2 \ . \tag{2.14}$$

Equation (2.14) is consistent with the convolution theorem of Sect. A.1 – a (double) convolution in real space (the Van Hove function of (2.4)) corresponds to a multiplication in Fourier space (right-hand side of (2.14)).

2.1.2 Examples of Van Hove Functions

In this section we examine the scattering from a simple form of the Van Hove function from 2.9 with one spatial dimension:

$$I(Q, E) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(x, t)\, \mathrm{e}^{-\mathrm{i}(Qx - \omega t)}\, \mathrm{d}x\, \mathrm{d}t \ . \tag{2.15}$$

First assume coherent scattering so there are predictable phase relationships between different scattering centers. Section 2.2 develops further the cases where incoherent averaging of the space and time correlations produce a "self-correlation function" and the "Patterson function." We later treat the incoherent case by following the "self-correlation function" of the individual scatterers.

Traveling Wave

Suppose an elementary excitation in a solid provides a periodic modulation of the scattering factor in space and time:

$$f(x, t) = f_0\, \mathrm{e}^{\mathrm{i}(qx - \omega_0 t)} \ . \tag{2.16}$$

We use (2.4) to obtain its Van Hove function:

$$G(x,t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_0^* \, \mathrm{e}^{-\mathrm{i}(qx'-\omega_0 t')} \, f_0 \, \mathrm{e}^{\mathrm{i}[q(x+x')-\omega_0(t+t')]} \, \mathrm{d}x' \mathrm{d}t' \ , \tag{2.17}$$

$$G(x,t) = |f_0|^2 \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}qx} \mathrm{d}x' \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega_0 t} \, \mathrm{d}t' = \mathrm{e}^{\mathrm{i}qx} \int_{-\infty}^{\infty} \mathrm{d}x' \ \mathrm{e}^{-\mathrm{i}\omega_0 t} \int_{-\infty}^{\infty} \mathrm{d}t' \ , \tag{2.18}$$

$$G(x,t) = |f_0|^2 \, \mathrm{e}^{\mathrm{i}(qx-\omega_0 t)} \ . \tag{2.19}$$

If we ignore the prefactor $|f_0|^2$, we find $G(x,t) = f(x,t)$. In this case where $G(x,t)$ has the form of a wave (2.19), (2.15) becomes:

$$I(Q,E) = |f_0|^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}(qx-\omega_0 t)} \, \mathrm{e}^{-\mathrm{i}(Qx-\omega t)} \, \mathrm{d}x \, \mathrm{d}t \ , \tag{2.20}$$

$$I(Q,E) = |f_0|^2 \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}(q-Q)x} \, \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}(\omega-\omega_0)t} \, \mathrm{d}t \ , \tag{2.21}$$

$$I(Q,E) = |f_0|^2 \, \delta(q-Q-g) \, \delta(\omega-\omega_0) \ . \tag{2.22}$$

Equation (2.22) shows that the energy transfer, $\hbar\omega$, must match that of the energy of the excitation in the material, $\hbar\omega_0$. Furthermore, the momentum transfer must match that of the wave, modulo a reciprocal lattice vector, g (which provides a factor of $\mathrm{e}^{\mathrm{i}gx} = 1$ in the integrand).

An elementary excitation in a solid with unique $\{q, \omega_0\}$ provides intensity at a single point in energy, and at a distance of Q away from each reciprocal lattice point. This of course assumes coherent scattering – incoherent scattering places a restriction on ω only.

Dispersive Excitations

Figure 2.1 presents simple sine waves of different wavelengths, but the same velocity. Such is the case for long-wavelength sound waves in solids. This is the "hydrodynamic limit," where the wavelengths are so large that the material behaves as a continuum (i.e., atomic features are not important). Three such waves are presented in Fig. 2.1a, but these are not special ones, and we expect that all intermediate wavelengths are possible. Each wave can be understood individually with the same analysis of the traveling wave as in Sect. 2.1.2.

The simple wave of Eq. 2.16 had a scattered intensity that was a delta function at one frequency and one wavevector, Eq. 2.22. If we allow a continuous range of wavevectors, we will have a continuous range of energies of the waves. There will be a continuous set of delta functions on a plot of the scattered intensity in Fig. 2.1b. These are related as $\omega = v\,k$, where v is the wave

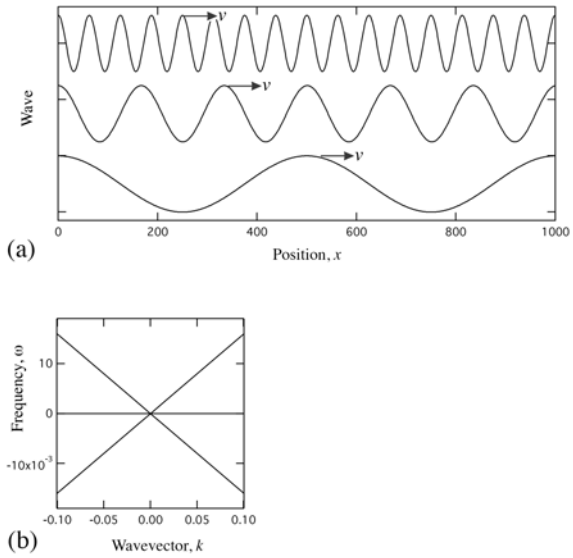


Figure 2.1: (**a**) Variations of a scattering function in space and time. Simple cosine waves are assumed, moving with equal velocities v. Their Van Hove functions are also cosine waves of the same wavelengths and frequencies. (**b**) The scattered intensity.

velocity. Note also that $v = \mathrm{d}\omega/\mathrm{d}k$.[1] Each traveling wave contributes a point to Fig. 2.1b, but for numerous wavelengths these points will make a straight line in the first quadrant of positive ω and k.

There are also excitations where the scattering factor distribution moves to the left in Fig. 2.1a, so we expect a mirror symmetry in $-k$. Also, we expect both creations and annihilations of these excitations, so in the scattered intensity we expect mirror symmetry in $-\omega$, too. The resulting intensity appears as straight lines are in all four quadrants of Fig. 2.1b.

Local Excitations

It is instructive to develop the Van Hove function and the scattered intensity with a pictorial approach, as in Fig. 2.2. The scattering factor distribution is shown in Fig. 2.2a as it undergoes a full cycle of oscillation.[2] The construction of the Van Hove function $G(x,t)$ parallels that of the Patterson function in Fig. 1.9. The detailed steps of shifting and integrating in Fig. 1.9 are not shown here. Suffice to say that situation along the x-dimension is quite analogous to that of Fig. 1.9, but a different instantaneous structure exists at each different time. For the case where $t = 0$, for example, it is necessary to average the shifting and integrating of Fig. 1.9 for all nine times shown in Fig. 2.2a. Notice that although some of the scattering factor distributions have zero intensity for separations

[1] The slopes of these straight lines $\omega(k)$ are expected to decrease as k gets large, comparable to a reciprocal lattice vector, and atomic-scale effects become important. Nevertheless, it is useful to compare sound velocities obtained from macroscopic measurements to the slopes of phonon dispersions from inelastic scattering. These two velocities generally agree.

[2] The convention is that times t_k with larger subscripts are later times. The pattern is repeated after 8 time steps.

that are at odd multiples of a, the result for $G(x,t)$ always has intensity at these locations (although weaker). The time dimension of $G(x,t)$ is obtained in a similar way. Consider the first time interval for time differences of t_1. For this case it is necessary to match all pairs of the scattering factor distributions that differ by one in their vertical stacking in Fig. 2.2a. These eight cases are then overlapped and shifted along the x-axis as before, and their results are averaged. The other time intervals require taking pairs separated further in time, but always it was eight cases whose shift and overlap are evaluated for constructing Fig. 2.2b.

To see how some of the trends in $G(x,t)$ in Fig. 2.2b come from the $f(x,t)$ in Fig. 2.2a, first identify the shifts in x and t that give optimal overlaps.[3] There is always optimal overlap at the origin, i.e., $(x,t) = (0,0)$. Other strong overlaps occur for shifts of x by $2na$ (n is an integer), and $t = 0$. There is also a strong overlap for shifts of t by mt_8 (m is an integer) and $x = 0$. The $G(x,t)$ in in Fig. 2.2b has maxima at these special shifts, and for any combination ($x = 2na$, $t = mt_8$). There is a second set of maxima, however, where the shifts in x and t are correlated as ($x = 2(n+1)a$, $t = (m+1/2)t_8$). Peaks in $G(x,t)$ in in Fig. 2.2b can be found for these combinations, too.

The final step is to obtain the $I(Q,\omega)$ of Fig. 2.2c. In this case the situation is fairly simple. There is only one time frequency in Fig. 2.2a, equal to $2\pi/t_8$. If we allow creation and annihilation of such an excitation, we expect intensity at frequencies $\omega = \pm 2\pi/t_8$. The spatial periodicity of the problem includes a superlattice periodicity, so peaks appear at intervals of $\pm n\pi/a$, where n is either even or odd.

Propagating Excitations

The next example in Fig. 2.3a is essentially the scattering factor distribution shown in Fig. 1.9, but now moves continuously to the right with increasing time. Its Van Hove function for zero time therefore has the same inversion symmetry along x as shown in Fig. 1.9, which was used to demonstrate Friedel's law. In assessing the overlap of the scattering factor distributions, for all cases of zero time shift (i.e., the overlap of the scattering factor with itself before it moves any further), the largest intensity occurs at lattice translation vectors. For most other time shifts t, the evaluation of $G(x,t)$ sums the overlaps of pairs of $f(x,t_k)$ in Fig. 2.3a that are separated in time. The best overlaps of these time-shifted $f(x,t_k)$ generally do not occur at lattice translations. Figure 2.3b shows that the shape of $G(x,t)$ is constant, but moves to the right with increasing time shift.

Peaks in the scattered intensity $I(Q,\omega)$ are shown in Fig. 2.3c. Along the time dimension of Figs. 2.3a,b, the pattern repeats itself with a periodicity of eight time snapshots in Figs. 2.3a. The inelastic part of the scattering therefore occurs at $\pm 2\pi/t_8$, assuming that excitations of the type shown in Fig. 2.3a can be both created and annihilated. There is another point worth noting about the time dependence in Fig. 2.3c. Along the time dimension we encounter sets

[3] We assume that the space and time periodicities repeat outside the range shown in Fig. 2.2.

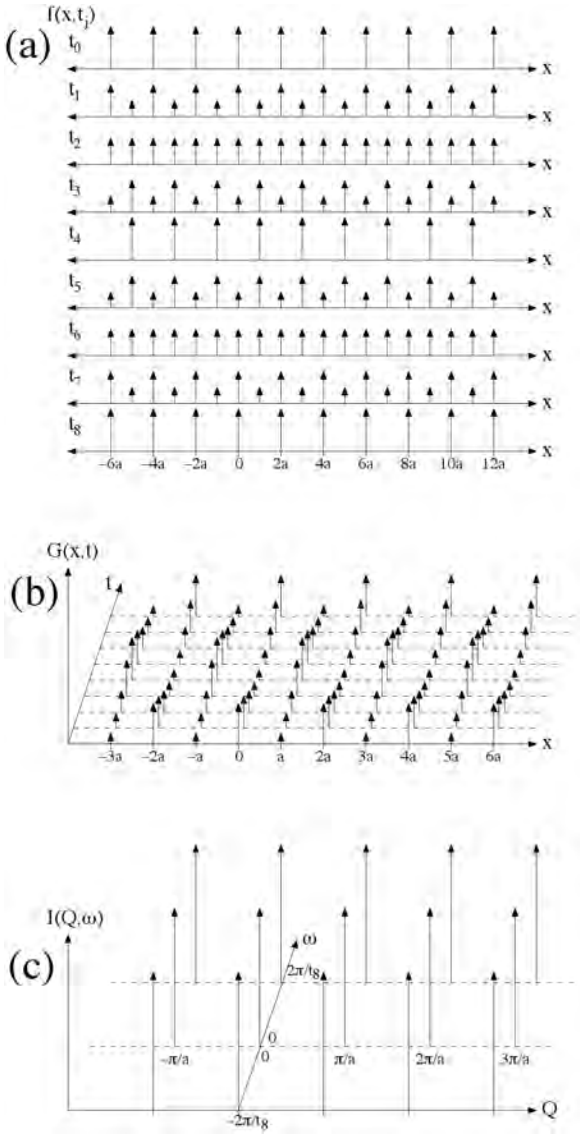


Figure 2.2: (**a**) Variations of a scattering function $f(x, t_i)$ in space and time. Nine snapshots are shown vertically for nine t_i. (**b**) The Van Hove function, $G(x, t)$, obtained by overlapping all pairs of scattering factor distributions in (a), and summing the resulting product of overlaps for all pairs separated by the same number of time intervals. (**c**) The scattered intensity, $I(Q, \omega)$. It is assumed possible to both create and annihilate an excitation as shown in (a), hence points at $\pm\omega$.

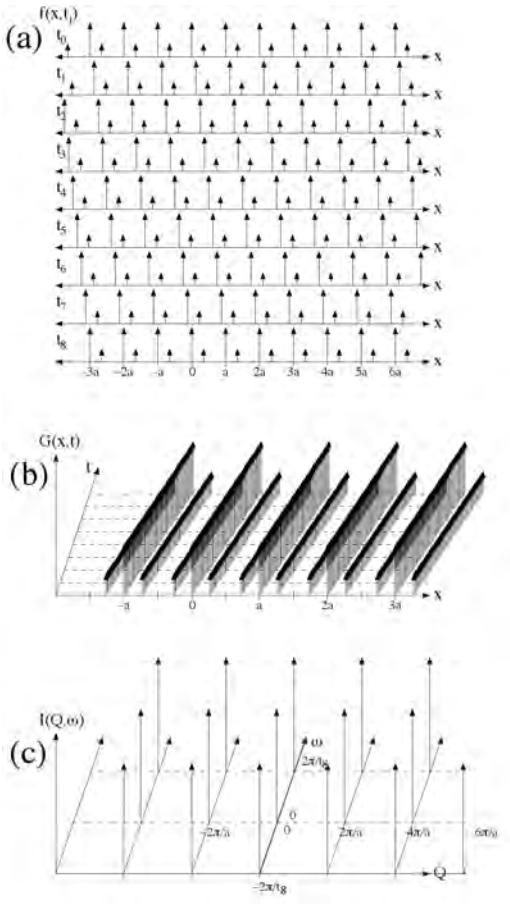


Figure 2.3: (**a**) Variations of a scattering function in space and time. Nine snapshots in time are shown vertically. (**b**) The Van Hove function, obtained by overlapping all pairs of scattering factor distributions in (a), and summing the resulting product of overlaps for all pairs separated by the same number of time intervals. It was assumed that the scattering factor distribution moves to the right continuously between the time snapshots shown in (a). (**c**) The scattered intensity. It is assumed possible to both create and annihilate an excitation as shown in (a), hence points at $\pm\omega_0$.

of three δ-functions for each fundamental periodicity of the wave. The basic periodicity of the δ-functions, and their grouping into threes, requires that we have higher-order Fourier components in the frequency domain – formally, we expect a whole series of excitations of energies: $\pm n\hbar\omega_0 = \pm n\hbar 2\pi/t_8$. The incident energy of the wave may be insufficient to excite some of these with higher n, however. This is a bit different from the situation of elastic scattering in k-space, where it is common to measure a series of Bragg peaks of multiple orders, for example.

Disordered Excitations

Another example is presented in Fig. 2.4. This example is a space-time analog of displacement disorder presented in Sect. 1.5.8. Here we assume that the scattering factor distribution is initially periodic, but is set in motion. Each atom is first displaced to the right, but each atom oscillates with a slightly different frequency. We assume that the frequencies have a narrow spread $\Delta\omega$ about a central frequency ω_0. Likewise, we assume that the amplitudes of the displacements are not large, so each arrow does not travel far from its lattice site. Eventually the oscillation damps away, and the scatterers are back to their initial periodic configuration. It is easiest to first analyze the situation for long times. Here the scatterers are in periodic positions, and have stopped moving. As the time interval becomes long with respect to the damping time, most of the spatial correlations will involve correlations in a precisely periodic structure. For long times, $G(x,t)$ will exhibit a set of peaks at equal intervals of $x = a$.

At short time separations, the individual arrows in Fig. 2.4b have not displaced much. At the shortest times after the neutron impact, the scatterers have nearly the same displacements since they move nearly in phase. It is incorrect, however, to assume that the peaks in $G(x,t)$ are displaced to the right. Recall that it is an average over many different time-separated snapshots of $f(x,t)$ in Fig. 2.4a, many of which are moving to the left. For small t, $G(x,t)$ is therefore similar to that of a perfect crystal in its spatial periodicities, with sharp peaks at each lattice translation.

At intermediate times, however, this this regularity is lost, even between neighboring arrows in this example. Figure 2.4c shows the intensity from this $G(x,t)$. For $\omega = 0$ the situation is as in Fig. 1.13, but for $\omega = 2\pi/t_8$ the intensity is broadened along the ω axis, owing the spread in frequencies. It is also expected that the disorder along the time domain will produce extra incoherent scattering between the sharp peaks in $G(x,t)$.

The initial coherence of the arrows is lost at intermediate times (assuming that the dephasing time $2\pi/\Delta\omega$ is less than the damping time). If we assume that there are no short-range spatial correlations, we have precisely the situation considered in Sect. 1.5.8. Figures 1.12 and 1.13 show the situation for long times, or equivalently for $\omega = 0$ in the $I(Q,\omega)$ of Fig. 2.4.

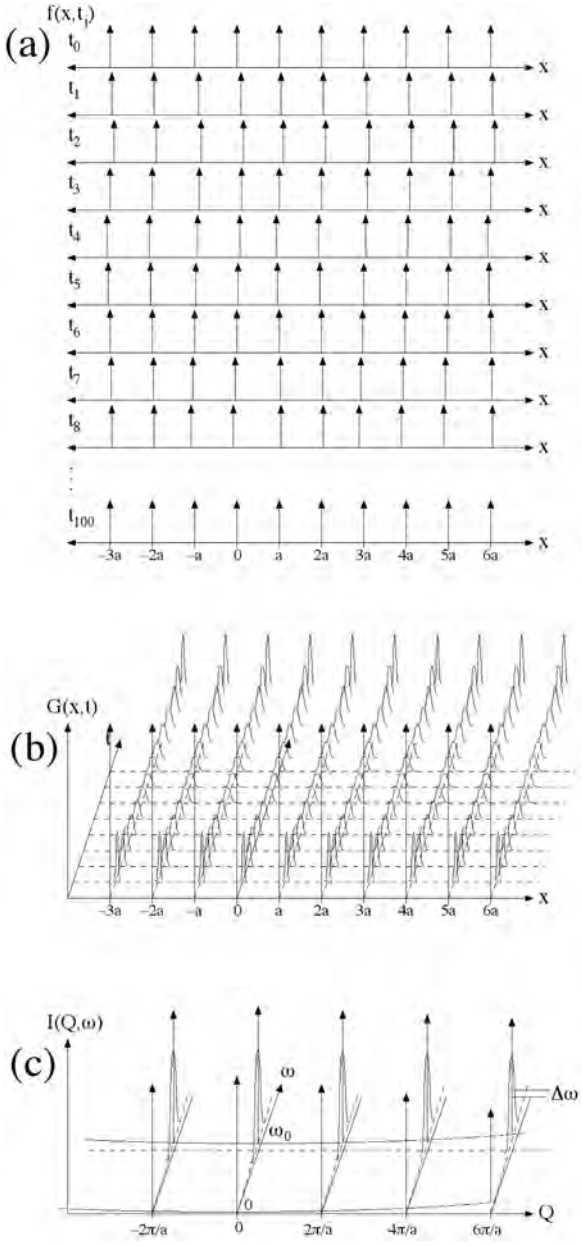


Figure 2.4: (**a**) Variations of a scattering function in space and time. Nine snapshots in time are shown vertically. (**b**) The Van Hove function, obtained by overlapping all pairs of scattering factor distributions in (a), and summing the resulting product of overlaps for all pairs separated by the same number of time intervals. (**c**) The scattered intensity. It is assumed possible to both create and annihilate an excitation as shown in (a), hence points at $\pm\omega_0$.

2.2 Autocorrelation Functions

The Van Hove function is worthy of deep respect, because it includes all the information available from all four types of scattering:
(coherent, inelastic), (coherent, elastic),
(incoherent, inelastic), (incoherent, elastic).
Subsets of this total information are often important, perhaps because the scattering is primarily incoherent, or is primarily elastic, for example. In particular, two important, although less general, correlation functions are:

- The Patterson function, $P(\vec{r})$, contains the spatial information obtained from diffraction measurements with *elastic coherent* scattering.
- The "self-correlation" function, $G_s(t)$, contains the time information obtained from measurements of *incoherent inelastic* scattering. It involves an averaging over Q.

2.2.1 Autocorrelation Functions from Van Hove Functions

Concept of Patterson and Self-Corrrelation Functions

Figure 2.5 shows the essential ideas. A "snapshot in time" is obtained by horizontal sampling across the figure. In a real experiment the samplings will be at many different times, but here we see that all such snapshots have the same periodic structure. As we know from Sect. 1.5.3, the diffraction intensity for this time snapshot is obtained from the Patterson function, whose periodicity in space gives the Bragg diffractions. In this particular case of a sine wave, however, we have Bragg peaks only at $Q = 0$ and $Q = \pm g$. If we had a series of δ-functions with the same periodicity, however, we would have a series of Bragg peaks to arbitrarily large orders of $Q = ng$. To obtain the Patterson function, the horizontal lines in Fig. 2.5a, each at a particular t_k, are convoluted with themselves, and the results are averaged for all t_k. The $f(x, t_k)$ for dfferent t_k in Fig. 2.5a do not interact with each other until after the convolution is performed. Specifically:

$$P(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f^*(x', t')\, f(x' + x, t')\, \mathrm{d}x'\, \mathrm{d}t' \,, \tag{2.23}$$

where we use the same value of t' in both arguments of the scattering factor. The $P(x)$ includes an averaging over time, and does not consider correlations in time. A $P(\vec{r})$ is shown in Fig. 2.5 as a horizontal line at a specific time. This information is probed by elastic scattering experiments, which do not allow for measuring the time variations and hence energy transfers.

Likewise we can take a "snapshot in space" by sampling the moving wave, at a specific $\vec{r}_1$ for example (although the results for all other values of $\vec{r}_j$ are the same in this example). Notice how the wave in Fig. 2.5 shows a time periodicity

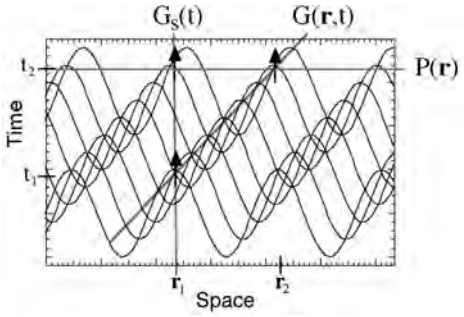


Figure 2.5: A simple sine wave, moving to the right with increasing time. Three methods of sampling the intensity of the wave in space and time are shown as straight lines, as explained in the text. The relevant correlation functions for each sampling are shown for each line.

along a vertical line, labeled as $G_s(t)$. The time periodicity of our moving wave is, of course, ω, in the wave function: $e^{iQx-\omega t}$. The restriction to a single value of Q eliminates the possibility of obtaining any spatial information on the excitation, but the ω-dependence is retained. We obtain the time correlations for a specific position:

$$G_s(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f^*(x',t')\, f(x',t'+t)\, dx'\, dt' \,, \tag{2.24}$$

where we use the same value of x' in both arguments of the scattering factor. The $G_s(t)$ includes an averaging over position, but does not consider correlations in position. This information is probed by incoherent inelastic scattering experiments, which do not allow for measuring the Q-dependences and hence spatial information.

More information comes by identifying correlations in both time and space, however, as is indicated by the diagonal line in Fig. 2.5. This requires the full Van Hove function, $G(\vec{r},t)$. This $G(\vec{r},t)$ cannot be obtained by simply summing the correlation functions of different space or time shapshots, i.e., summing $G_s(t)$ or $P(\vec{r})$. Nevertheless, it is instructive to evaluate the Patterson function and self-correlation function for each of the three examples shown in Figs. 2.2, 2.3, 2.4 (see Problem 3).

Local Excitations

Along the horizontal axis of $\omega = 0$ in Fig. 2.2c, we obtain a diffraction pattern with strong fundamentals at $Q = \pm n\pi/a$, where n is an even integer, and somewhat weaker superlattice diffractions for odd n. The same result is true for two other values of $\omega = \pm 2\pi/t_8$. This periodic structure in Q is evident from the Van Hove function $G(x,t)$ in Fig. 2.2b. The spatial correlation function of (2.23), $P(x) = \iint f^*(x',t_k)\, f(x'+x,t_k) dx' dt_k$ has spatial periodicities that are similar at any time t_k, although different weights for the superlattice diffractions. The average over all t_k is of course what makes the final result for $I(Q,\omega=0)$ in Fig. 2.2c.

Now look at the correlation functions along lines parallel to the time axis for fixed x' in $G(x', t)$ of Fig. 2.2b. The time periodicity of $G(x', t)$ in Fig. 2.2b differs in phase for even and odd n in $x' = \pm na$. All time periodicities are the same, however, and include only the frequencies $-2\pi/t_8, 0, -2\pi/t_8$.[4] The frequency spectrum of $G(x', t)$, measured along $x' = 0$, is generally the same as the inelastic scattering along any other $x' = na$. If we average the intensities along these individual x', we lose the information on phases of scattering between the scatterers at different x'. The mixed constructive and destructive interference between these different scatterers results in intensities between individual atoms only. The result is a "self-correlation" function of (2.24), or a case of incoherent inelastic scattering, $G_s(t) = \iint f^*(x_j, t')\, f(x_j + x, t' + t)\mathrm{d}x_j\mathrm{d}t'$. In this particular case, the incoherent average of these (obtained by summing the intensities at all Q) looks generally the same as the spectrum along a slice along any particular Q.

The Patterson and self-correlation functions of Fig. 2.3 can be understood in much the same way. The largest horizontal periodicity in Fig. 2.3b is a, so this Patterson function gives diffraction peaks at multiples of $2\pi/a$. The self-correlation function likewise has the simple periodicities of $-2\pi/t_8, 0, -2\pi/t_8$, neglecting the detailed structure in cuts along the time axis that would give higher-order Fourier components.

Acoustic Excitations

More interesting are the dispersions of different excitations that have the same wave speed (presented in Fig. 2.1b). All these excitations would be measured in an incoherent inelastic scattering experiment, since the averaging over all x' of $G(x', t)$ would preserve all time periodicities. An incoherent inelastic scattering experiment would collapse all the inelastic intensity onto a single ω-axis, and lose the relationship between ω and Q that is so evident from the coherent inelastic scattering shown in Fig. 2.1b. For dispersive excitations, the self-correlation function has considerably less information compared to the full Van Hove function. On the other hand, if one is interested in obtaining an accurate density of states (i.e., a total energy spectrum), irrespective of the values of Q, the self-correlation function may be advantageous. It does not require accurate measurements of intensities over all values of Q and ω.

Disordered Excitations

Finally we consider the Patterson function and self-correlation function of Fig. 2.4. The scatterers oscillate about their lattice sites, and the Patterson function is therefore as expected of a perfect crystal. For the time average, there is displacement disorder in the arrangement, however, so we expect a diffuse

[4] Incidentally, if the arrows were to oscillate through zero in Fig. 2.2a, we would have no value of $\omega = 0$, and no elastic scattering. This corresponds to half the scatterings occurring out-of-phase with the other half.

background, increasing with Q as shown in Fig. 1.13. The self-correlation function for incoherent inelastic scattering is the same for all x, since the scatterers are assumed independent in their motions, without positional correlations that may case neighbors to move together in phase. In this case we expect the same time structure for $G_s(t)$ and $G(x,t)$, and hence inelastic incoherent scattering will have the same energy spectrum as the coherent inelastic scattering for all Q. Both will show a broadening in frequency of the excitation around ω_0. The broadening arises from two effects. First is the frequency spread of the oscillators, postulated to be $\Delta\omega$. The second effect is the damping of the oscillations. The damping time of τ provides a broadening in energy of $\hbar/\tau$, known as "lifetime broadening." In the present problem we have assumed that the broadening of the excitation at ω_0 is dominated by the frequency spread $\Delta\omega$, since the dephasing time was assumed shorter than the damping time.

2.2.2 Inelastic Scattering from Diffusion

Atom Correlations from Diffusion Equation

Here we pick a simple case for uncorrelated atom movements in diffusion. We begin with the Gaussian concentration profile, $c(x,t)$, which is a standard solution of the 1D diffusion equation

$$c(x,t) = \frac{c'}{\sqrt{4\pi Dt}}\, e^{-\frac{x^2}{4Dt}} \ . \tag{2.25}$$

The Van Hove function of Eq. 2.4 is

$$G_{\rm diff}(x,t) = \int_{-\infty}^{+\infty} \Big\langle c(x',0)\, c(x'+x,t) \Big\rangle \, dx' \ . \tag{2.26}$$

An important point is that there are no space or time correlations between different atoms in diffusion – two different atoms do not know about each other's jumps, and all atoms follow the same concentration profile.[5] Ensemble averaging is therefore unnecessary.

$$G_{\rm diff}(x,t) = \int_{-\infty}^{+\infty} c(x',0)\, c(x'+x,t)\, dx' \ . \tag{2.27}$$

As $t \to 0$, the function $c(x', t \to 0)$ of Eq. 2.25 becomes a Dirac delta function of unit area

$$G_{\rm diff}(x,t) = \int_{-\infty}^{+\infty} \delta(x')\, c(x',t)\, dx' \ . \tag{2.28}$$

[5] This is unlike phonon scattering, for example, where the motion of one atom follows the motion of another as a vibrational wave moves from one atom to the other.

so the spatial integral simply returns the function $c(x,t)$ of Eq. 2.25

$$G_{\text{diff}}(x,t) = \frac{c'}{\sqrt{4\pi D t}}\, \mathrm{e}^{-\frac{x^2}{4Dt}} \ . \tag{2.29}$$

We calculate the $S(Q,\omega)$ by using Eq. 2.29 in the Fourier transform relationship of Eq. 2.9

$$S(Q,\omega) = \frac{c'}{2\pi\hbar} \int_{t=0}^{+\infty} \int_{x=-\infty}^{+\infty} \frac{\mathrm{e}^{-\frac{x^2}{4Dt}}}{\sqrt{4\pi D t}}\, \mathrm{e}^{-\mathrm{i}(Qx-\omega t)}\, \mathrm{d}x\, \mathrm{d}t \ , \tag{2.30}$$

$$S(Q,\omega) = \frac{c'}{2\pi\hbar} \int_{t=0}^{+\infty} \mathrm{e}^{-Q^2 D t}\, \mathrm{e}^{\mathrm{i}\omega t}\, \mathrm{d}t \ , \tag{2.31}$$

$$S(Q,\omega) = \frac{c'}{2\pi\hbar}\, \frac{DQ^2}{(DQ^2)^2 + \omega^2} \ , \tag{2.32}$$

where we used a standard result that the Fourier transform of a Gaussian is a Gaussian to obtain Eq. 2.31, and that the Fourier transform of a decaying exponential is a Lorentzian function to obtain Eq. 2.32. The Lorentzian function is a symmetrical, peaked function with a maximum at $\omega = 0$, or $\Delta E = 0$. Its half-width at half-maximum is at $\omega = DQ^2$, or

$$\Delta E = \hbar\, D\, Q^2 \ . \tag{2.33}$$

Physical interpretation – quasielastic scattering

We can interpret Eq. 2.32 in terms of how inelastic scattering is sensitive to the atom jumps in diffusion. The energy width of the inelastic intensity of Eq. 2.33 corresponds to a time $\Delta\tau$ by the uncertainty relationship

$$\Delta\tau = \frac{\hbar}{\Delta E} = \frac{1}{D\, Q^2} \tag{2.34}$$

An important result attributed to Einstein relates the diffusion coefficient D to the atom jump frequency Γ and the jump length a. A factor of 1/6 is appropriate for 3D diffusion

$$D = \frac{\Gamma\, a^2}{6} \ , \tag{2.35}$$

where Γ is the atom jump frequency and a is its jump length. The time uncertainty becomes

$$\Delta\tau = \frac{1}{\left(\frac{\Gamma a^2}{6}\right)\left(\frac{2\pi}{a}\right)^2} \ , \tag{2.36}$$

$$\Delta\tau = \frac{0.30}{\Gamma} \ , \tag{2.37}$$

for a representative Q of $2\pi/a$. In other words, without diffusion ($D = 0$) the inelastic spectrum would be a sharp peak at $E = 0$, corresponding to elastic scattering only. When the atoms jump frequently, there is time-broadening of the energy $\Delta E = \hbar\Gamma$ at a characteristic Q. This is typical of wave scattering that is interrupted by a jump of the scattering center, which shortens the coherence time for the emitted wave, broadening it in energy. This process of scattering by moving atoms is often called "quasielastic scattering" because it can be considered as elastic scattering (no overall energy shift), but with an energy broadening as the moving atoms interrupt the coherence of the elastic line.

2.2.3 Relationships Between Intensities, Correlation Functions, Waves, and Scattering Lengths

Much of scattering science involves relationships between the wavefunction $\psi(Q, \omega)$ and physical scattering lengths $f(r, t)$ in the sample. Many important functional relationships are obtained by:

- Fourier transformation F
- autocorrelation $\circledast$
- multiplication $\|^2$
- averaging $\langle\rangle$

A map of the important physical functions and how they are transformed into one another is presented in Fig. 2.6. It is a map of the universe of scattering functions.

When using Fig. 2.6, reverse transformations are possible in all cases, but information is lost by averaging over an argument of a function. For example, $I(\omega)$ can be transformed to $S(Q, \omega)$ in the incoherent approximation, but dispersive information about $\omega(Q)$ cannot be obtained. (Of course, if the scattering is incoherent, there is no dispersive information to begin with.) Averaging over ω is not shown in Fig. 2.6 because a more common manipulation in scattering science is to identify the elastic scattering within a set of inelastic data at $E = 0$. Finally, the functions $f(r, \omega)$ and $\psi(Q, t)$ are not shown. They may prove useful as intermediate steps in some calculations, but they mix the phase information about the wave and the scatterer.

Some definitions are:

- $S(Q, \omega)$ is the "scattering law."
- $Y(Q, \tau)$ is the momentum-time corrrelation function. It is called the "intermediate scattering function."
- $\Gamma(R, \omega)$ is the dynamical pair distribution function.
- $G(R, \tau)$ is the Van Hove space-time correlation function.

- $P(R)$ is the Patterson function.
- $\mathcal{M}(\tau)$ is the "memory function," a time correlation function for dynamics at a site (sometimes called $G_s(\tau)$, a self-correlation function).
- $I(Q)$ is a diffraction pattern.
- $I(\omega)$ is an inelastic spectrum.
- $f(r,t)$ is a scattering length density.
- $\psi(Q,\omega)$ is a wavefunction.

Two correlation functions in Fig. 2.6 have special uses. The function $\Gamma(R,\omega)$, the dynamical pair distribution function, is used for projections of excitations onto specific sites. More precisely, it gives the vibrational spectrum between a particular pair of atoms. The function $Y(Q,\tau)$ is useful for studies of transient phenomena that may occur after an impulsive perturbation.

2.3 Essence of Coherent Inelastic Neutron Scattering

2.3.1 Spherical Waves from Point Scatterers in Motion

An intuitive shortcut from (1.58) to (1.65) is to regard $(\nabla^2 + \vec{k}_i^2)$ as a scattering operator that generates a scattered wavelet proportional to $U(\vec{r}')\Psi(\vec{r}')$. The scattered wavelet must have the properties of (1.50) for its amplitude and phase versus distance. The scattered wavelet amplitude emitted from a small volume, $d^3\vec{r}'$, centered about $\vec{r}'$ is:

$$d\Psi_{sc}(\vec{r},\vec{r}') = U(\vec{r}')\Psi(\vec{r}')\frac{e^{i(k_f|\vec{r}-\vec{r}'|-\omega_0 t)}}{|\vec{r}-\vec{r}'|}d^3\vec{r}' \ . \tag{2.38}$$

This is an expression for a spherical wave at $\vec{r}$ originating from the small volume $d^3\vec{r}'$ surrounding $\vec{r}'$. It is isotropic – note that the exponential factor $k_f|\vec{r}-\vec{r}'|$ is a product of scalars.

We obtain the total scattered wave by integrating around all volume of the scatterer. The incident plane wave, $\propto e^{i\vec{k}_i\cdot\vec{r}'}$ (1.68), helps sets the phase of the scattering at each volume interval. The phase of the outgoing wave also depends on the orientation of the outgoing $\vec{k}_f$ with respect to the position of the scattering point, $\vec{r}'$. The *relative* phase from each scattering point depends on the change in wavevector, $\vec{Q} \equiv \vec{k}_i - \vec{k}_f$, as $e^{i\vec{Q}\cdot\vec{r}'}$.

$$\Psi_{sc}(\vec{Q},\vec{r}) \quad = \quad -\frac{e^{i(\vec{k}_f\cdot\vec{r}-\omega_0 t)}}{|\vec{r}|}\frac{m}{2\pi\hbar^2}\int V(\vec{r}')\,e^{i\vec{Q}\cdot\vec{r}'}\,d^3\vec{r}' \ . \tag{2.39}$$

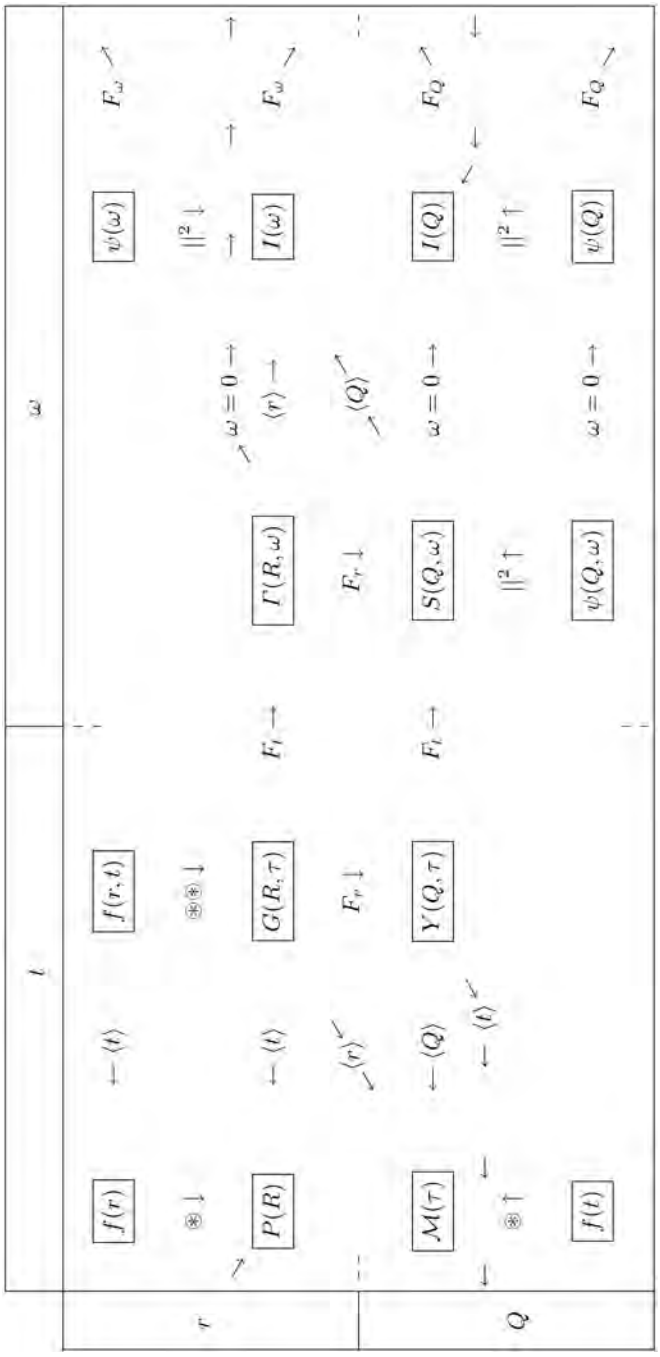


Figure 2.6: Interrelationships between Correlation Functions, Scattering Length Densities, Intensities, Wavefunctions

In arriving at (2.38) we have repeated, in an intuitive way, the steps to (1.74) and (1.75). We note that the momentum transfer from the neutron to the sample is $\hbar\vec{Q}$.

The trick now is to replace the potential, $V(\vec{r})$, with a suitable potential for neutron scattering. For nuclear scattering, relevant to phonon measurements, we use the "Fermi pseudopotential," which places all the potential at a point nucleus:

$$V_{\rm nuc}(\vec{r}) = 4\pi\frac{\hbar^2}{2m}\, b\, \delta(\vec{r}) \;, \tag{2.40}$$

where b is a simple constant (perhaps a complex number). For thermal neutrons, the δ-function is an appropriate description of the shape of a nucleus.[6] Fermi pseudopotentials are placed at crystal translation vectors, $\{\vec{R}_j\}$, accounting for all N nuclei in the crystal

$$V_{\rm n}(\vec{r}') = 4\pi\frac{\hbar^2}{2m_{\rm n}}\sum_{j=1}^{N} b_j\, \delta(\vec{r}' - \vec{R}_j) \;. \tag{2.41}$$

This sum of δ-functions transforms the integration of Eq. 2.39 into a sum over points at $\vec{R}_j$. For laboratory distances (large compared to the size of the scatterer), $\vec{r} \gg \vec{R}_j$, the outgoing spherical wave can be treated as a plane wave, allowing rearrangement

$$\Psi_{\rm sc}(\vec{r}, t') = \frac{{\rm e}^{{\rm i}(\vec{k}_{\rm f}\cdot\vec{r} - \omega_{\rm f}t')}}{\left|\vec{r}\right|}\;\sum_j b_j\, {\rm e}^{{\rm i}(\vec{k}_{\rm i}-\vec{k}_{\rm f})\cdot\vec{R}_j}\; {\rm e}^{{\rm i}(\omega_{\rm f}-\omega_{\rm i})t} \;. \tag{2.42}$$

The factor ${\rm e}^{{\rm i}\vec{Q}\cdot\vec{r} - \omega_{\rm f}t')}$ gives the precise phase at the time and place of neutron detection. It has modulus unity, and we ignore it because we measure the intensity of scattering. We use Eq. 5.44 for $\vec{R}_j$,

$$\vec{R}_j = \vec{r}_l + \vec{r}_\kappa + \vec{u}_{l\kappa}(t) \;, \tag{2.43}$$

where the three terms are the lattice vector, basis vector of the unit cell (containing $\mathcal{R}$ atoms), and the displacement vector. We allow the scattered wave to emanate from the nucleus at later times

$$\Psi_{\rm sc}(\vec{r}) = \frac{1}{r}\sum_{l,\kappa} b_\kappa\, {\rm e}^{{\rm i}\vec{Q}\cdot(\vec{r}_l+\vec{r}_\kappa)} {\rm e}^{{\rm i}\vec{k}_{\rm i}\cdot\vec{u}_{l\kappa}(0)} {\rm e}^{-{\rm i}\vec{k}_{\rm f}\cdot\vec{u}_{l\kappa}(t)} {\rm e}^{{\rm i}(\omega_{\rm f}-\omega_{\rm i})t} \;, \tag{2.44}$$

where we note that because all unit cells are identical, b depends on the basis vector only.

[6] Typically of a nuclear scale of 10^{-12} cm or so, sometimes with an imaginary component to account for absorption. For magnetic scattering, however, the δ-function should be convoluted with a real-space form factor for the magnetic electrons. This could be done at the end of the calculation by multiplying the k-space result with a form factor.

It is easiest to understand the case when the scattering is a fast process, such as for x-ray scattering, where an x-ray traverses the crystal more rapidly than typical vibrational periods of atoms. Incoherent neutron scattering is also fast, where the thermal neutron is scattered from individual atomic nuclei without concern for phase relationships between nuclei. This motivates a first approximation; ignore the time dependence of the potential by setting $\vec{u}_{l\kappa}(t) = \vec{u}_{l\kappa}(0)$, and take the time average

$$\Psi_{\rm sc}(\vec{Q},\vec{r}) = \frac{1}{r}\sum_{l,\kappa} b_\kappa\, {\rm e}^{{\rm i}\vec{Q}\cdot(\vec{r}_l+\vec{r}_\kappa)}\, {\rm e}^{{\rm i}\vec{Q}\cdot\vec{u}_{l\kappa}(0)}\, \frac{1}{\tau}\int_0^\tau {\rm e}^{{\rm i}(\omega_{\rm f}-\omega_{\rm i})t}{\rm d}t\ , \tag{2.45}$$

$$\Psi_{\rm sc}(\vec{Q},\vec{r}) = \frac{1}{r}\sum_{l,\kappa} b_\kappa\, {\rm e}^{{\rm i}\vec{Q}\cdot(\vec{r}_l+\vec{r}_\kappa)}\, {\rm e}^{{\rm i}\vec{Q}\cdot\vec{u}_{l\kappa}(0)}\, \delta(\omega_{\rm f}-\omega_{\rm i})\ . \tag{2.46}$$

The δ-function enforces no change of frequency or energy, so Eq. 2.46 pertains to elastic scattering, where the modes of atom vibrations are unchanged. We need to perform an ensemble average over the different values of $\vec{u}_{l\kappa}(0)$, which evaluates as[7]

$$\langle {\rm e}^{{\rm i}\vec{Q}\cdot\vec{u}(0)}\rangle_N = {\rm e}^{-\frac{1}{2}\langle[\vec{Q}\cdot\vec{u}(0)]^2\rangle_N} \equiv {\rm e}^{-W}\ . \tag{2.47}$$

Equation 2.46 for elastic scattering becomes

$$\Psi^{\rm el}_{\rm sc}(\vec{Q},\vec{r}) = \frac{{\rm e}^{-W}}{r}\sum_\kappa b_\kappa\, {\rm e}^{{\rm i}\vec{Q}\cdot\vec{r}_\kappa}\sum_l {\rm e}^{{\rm i}\vec{Q}\cdot\vec{r}_l} \equiv \frac{{\rm e}^{-W}}{r}\mathcal{F}(\vec{Q})\,\mathcal{S}(\vec{Q})\ . \tag{2.48}$$

Here we separated the sums over basis and lattice vectors into a structure factor, $\mathcal{F}(\vec{Q})$, and a shape factor, $\mathcal{S}(\vec{Q})$. Equation 2.48 can be used to calculate the neutron diffraction pattern from a crystal, giving the average static structure. The factor ${\rm e}^{-W}$ is the "Debye–Waller factor," which becomes appreciably less than 1 when the atom displacement $\vec{u}_\kappa$ is comparable to $1/Q \sim \lambda/2\pi$, the "wavelength" associated with the scattering angle and energy. This is consistent with the origin of the Debye–Waller factor from the destructive interference of scattered wavelets that are emitted from different spatial positions of the displaced nuclei. Note that the ensemble average of Eq. 2.47 does not reveal the energy of an atom vibration, only how displaced atom positions alter the phase.

Next consider a process of scattering that is slow compared to the atomic vibrational period. Such is the case for coherent scattering of neutrons by phonons, and for nuclear resonant scattering of photons by phonons. For

[7] An approximate demonstration of this is by expanding $\exp({\rm i}\vec{Q}\cdot\vec{u}) = 1 + {\rm i}\vec{Q}\cdot\vec{u} + \ldots$ and taking either an ensemble average or a time average. The average of the second term, ${\rm i}\vec{Q}\cdot\vec{u}$ is zero owing to positive and negative excursions of $\vec{u}$, leaving the quadratic term, consistent with Eq. 2.47. The average of the subsequent cubic term also vanishes, conveniently, but unfortunately the quartic term has the wrong sign. Nevertheless, it can be proved that Eq. 2.47 is exact if the probability distribution of $\vec{u}$ is Gaussian about the origin, and this is often reasonable for lattice dynamics.

clarity, we pick a single phonon with wavevector $\vec{k}$. The form of $\vec{u}_{l\kappa\vec{k}}(t)$ is obtained later in Eq. 5.56, but is presented here (mostly for interest, with normalization for a quantum vibration)

$$\vec{u}_{l\kappa\vec{k}}(t) \quad = \quad \sqrt{\frac{2\hbar}{N\, m_\kappa\, \omega_{\vec{k}}}}\, \vec{e}_\kappa(\vec{k})\, e^{i(\vec{k}\cdot\vec{r}_l - \omega_{\vec{k}} t)} \ . \tag{2.49}$$

Equation 2.49 shows that $\vec{u}_{l\kappa\vec{k}}(t) = \vec{u}_{l\kappa\vec{k}}(0)e^{-i\omega_{\vec{k}}t}$. We need to take a time average of the displacement factors in Eq. 2.44

$$\langle e^{i\vec{k}_i\cdot\vec{u}_{l\kappa\vec{k}}(0)}\, e^{-i\vec{k}_f\cdot\vec{u}_{l\kappa\vec{k}}(t)}\rangle_t = \langle e^{i\vec{k}_i\cdot\vec{u}_{l\kappa\vec{k}}(0)}\, e^{-i\vec{k}_f\cdot\vec{u}_{l\kappa\vec{k}}(0)e^{-i\omega_{\vec{k}}t}}\rangle_t \tag{2.50}$$

Coherent phonon scattering involves correlations of displacements $\vec{u}_{l\kappa\vec{k}}(t)$ at different atom positions at different times. This Van Hove space-time correlation function **?**) is a rich approach to studying dispersive excitations in solids. The proper quantum mechanical treatment of time and temperature averages is taken up in Section 4, but Eq. 2.51 below is the proper result for incoherent scattering. Furthermore, it is possible to average the coherent scattering over all $\vec{Q}$ to recover this incoherent result **???**), and the incoherent approximation serves us well in the following analysis of multiphonon scattering. The loss of phase information by averaging the coherent scattering over all $\vec{Q}$ brings us back to consideration of a fast scattering process.

2.3.2 One-Phonon and Multiphonon Scattering

Equation 2.48 is the elastic scattering only. It does not conserve the nuclear scattering cross section. Nuclear interactions do not necessarily depend on the displacements of atoms, so we need a compensating factor $e^{+\langle(\vec{Q}\cdot\vec{u})^2\rangle}$ to conserve the cross-section and account for the rest of the scattering intensity. Expanding this new factor, we calculate the intensity as $\Psi^*_{sc}\Psi_{sc}$, and for clarity we normalize the intensity

$$\begin{aligned}\langle\Psi^*_{sc}\Psi_{sc}\rangle_N \quad &= \quad \frac{1}{\mathcal{R}}\sum_\kappa^{\mathcal{R}} e^{-\langle(\vec{Q}\cdot\vec{u}_\kappa)^2\rangle} \\ &\times \quad \left(1 + \langle(\vec{Q}\cdot\vec{u}_\kappa)^2\rangle + \frac{1}{2}(\langle(\vec{Q}\cdot\vec{u}_\kappa)^2\rangle)^2 + ...\right)\end{aligned} \tag{2.51}$$

The first and second terms in parentheses of Eq. 2.51 are the fractions of elastic (0-phonon) and 1-phonon scattering. The third term with the $\frac{1}{2}(\langle(\vec{Q}\cdot\vec{u}_\kappa)^2\rangle)^2$ accounts for 2-phonon scattering processes. Only the 1-phonon scattering proves useful for obtaining a phonon DOS, but it is important to have an understanding of the higher-order terms (called "multiphonon scattering") for designing an experiment, or performing quantitative corrections of experimental data.

The incoherent approximation proves more robust than it may seem to deserve. Energy and momentum conservation suppress the possibilities for

cross-terms between scattering by 0-phonon and 1-phonon processes. The frequencies of the incident and scattered waves are mismatched. Integrated over many periods of the neutron wavefunction, these mismatched waves will not add constructively or destructively, so there is no coherent interaction between these two scattering processes. [8]

A more detailed treatment of 1-phonon scattering includes the thermal factors for phonon populations, giving the double-differential scattering cross section[9]

$$\begin{aligned}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} &= \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^3}{2V_0}\sum_{\vec{g},\vec{k}}\frac{n(\vec{k})+\frac{1}{2}\pm\frac{1}{2}}{\omega_{\vec{k}}} \\ &\quad\times\Big|\sum_\kappa \frac{b_\kappa}{\sqrt{m_\kappa}}\,\mathrm{e}^{-\langle(\vec{Q}\cdot\vec{u}_\kappa)^2\rangle}\left[\vec{Q}\cdot\vec{e}_\kappa(\vec{k})\right]\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}_\kappa}\Big|^2 \\ &\quad\times\delta(\vec{Q}+\vec{k}\pm\vec{g})\,\delta(\omega-\omega_{\vec{k}})\;, \end{aligned} \tag{2.52}$$

where the phonon occupancy factor, $n(\vec{k})$, is the Planck distribution of Eq. 5.16 at the temperature of interest and $\varepsilon_i = \hbar\omega_{\vec{k}}$.

$$n(\vec{k}) = \frac{1}{\mathrm{e}^{\hbar\omega(\vec{k})/k_\mathrm{B}T}-1}\;. \tag{2.53}$$

The signs in the factor $n(\vec{k})+\frac{1}{2}\pm\frac{1}{2}$ in Eq. 2.52 are such that it is $n(\vec{k})$ for phonon annihilation, and $n(\vec{k})+1$ for phonon creation – it is always possible to create a phonon, even at $T=0$ when no phonon excitations are present. The $\delta(\vec{Q}+\vec{k}\pm\vec{g})$ and $\delta(\omega-\omega(\vec{k}))$ enforce the conservation of momentum and energy on the neutron scattering process. The phonons on different branches must be considered separately in Eq. 2.52, and it is traditional to add a "branch index," sometimes denoted j, to $\omega(\vec{k})$, since more than one ω_j may correspond to a specific $\vec{k}$.

Why should the expansion of an evidently classical exponential in Eq. 2.51 produce the series of neutron-phonon interactions? In the classical limit, the scattering of every neutron is inelastic, generating heat composed of a number of phonons with total energy, E, the energy transfer from the neutron. For classical scattering, an atom of mass m will recoil with velocity $v = \hbar Q/m$, with momentum transfer $\hbar Q$, and energy transfer

$$E_R = \frac{\hbar^2 Q^2}{2m}\;. \tag{2.54}$$

[8] Perhaps they exist as quantum beats of the wavetrains, but we ignore them because we do not have sufficient time resolution at the detector to sense them.

[9] A new factor is the ratio $k_\mathrm{f}/k_\mathrm{i}$, which expresses the effect on flux caused by the rate at which neutrons leave the sample. Compared to an elastically-scattered neutron, fewer neutrons per second will pass into an increment of solid angle $\mathrm{d}\Omega$ if they are slowed to smaller values of k_f. Also, the use of solid angle eliminates the need for the factor r^{-2}.

In the quantum limit where Q is small, the energy recoil may or may not occur, depending on whether phonons are created or not. The total scattering, $S(Q,E)$, is the sum of components, $\sum_{n=0}^{\infty} S_n(Q,E)$, from neutrons scattered after creating different numbers, n, of phonons in the sample.[10] The thermal energy is

$$k_{\mathrm{B}}T = 2m\omega^2\langle u^2\rangle \tag{2.55}$$

for one harmonic mode of frequency ω. From Eqs. 2.54 and 2.55

$$Q^2\langle u^2\rangle = \frac{E_{\mathrm{R}}}{\hbar\omega}\,\frac{k_{\mathrm{B}}T}{\hbar\omega} \tag{2.56}$$

This is the quantity in which which the multiphonon expansion is performed (Eq. 2.51). From Eq. 2.56 we see that the relative intensities of the sequence of phonon scattering terms scale with 1) the ratio of the recoil energy to the oscillator energy, and 2) the ratio of temperature to the oscillator energy. When the first ratio is small, quantum mechanics requires that some recoils excite the oscillator, but most do not. The same is true for temperature (and at low temperatures we of course have low heat capacity because the oscillator is usually not excited).

It remains to get the spectral shape of each order of the multiphonon scattering. The spectrum for one-phonon scattering weights more heavily the low-energy modes because they have larger amplitudes of motion, providing a factor of $g(\varepsilon)/\varepsilon$. The number of phonons is the Planck distribution $1/[\exp(\varepsilon/k_{\mathrm{B}}T) - 1]$, and the two factors provide the shape of the 1-phonon profile, $A_1(\varepsilon)$:

$$A_1(\varepsilon) = \frac{g(\varepsilon)}{\varepsilon}\,\frac{1}{\mathrm{e}^{\varepsilon/k_{\mathrm{B}}T} - 1}\ . \tag{2.57}$$

When two phonons are created simultaneously, the total spectrum of energies is the convolution of the 1-phonon profile with the 1-phonon profile.[11] The 2-phonon spectrum is the convolution of two 1-phonon profiles,

$$A_2(E) = A_1 * A_1 = \int_{-\infty}^{\infty} A_1(E - E')\,A_1(E')\,\mathrm{d}E'\ , \tag{2.58}$$

The n-phonon profile is the convolution of another 1-phonon profile with the $(n-1)$-phonon profile

$$A_n(\varepsilon) = A_1 * A_{n-1} = \int_{-\infty}^{\infty} A_1(\varepsilon - \varepsilon')\,A_{n-1}(\varepsilon')\,\mathrm{d}\varepsilon'\ . \tag{2.59}$$

[10] Phonon annihilation is handled by extending the range of E to negative values for each $S_n(Q,E)$.

[11] Consider each phonon excitation to be a random variable with a probability distribution of A_1. The sum of two random variables has a distribution that is the convolution of the probability distributions $A_1 * A_1$.

Starting with A_1, we can generate the spectral shapes of all orders of multiphonon scattering by the systematic application of Eq. 2.59.

The total scattering is the sum of these spectral profiles, weighted by the corresponding terms of Eq. 2.51

$$S(Q,E) = \sum_{n=0}^{\infty} \exp(-2W) \frac{(2W)^n}{n!} A_n(E) \ , \tag{2.60}$$

where we have defined $S(Q,E)$ as the double-differential cross-section $\mathrm{d}^2\sigma/\mathrm{d}\Omega\mathrm{d}E$ without prefactors.[12]

In much experimental work, we seek to isolate the single scattering profile, $A_1(E)$, from the measured $S(Q,E)$, and then determine the phonon DOS $g(E)$ by Eq. 2.57. One way to correct for multiphonon scattering is by iteration on an initial guess of the phonon DOS. With an initial guess at the DOS or $A_1(E)$ (often the $S(Q,E)$ with a constant background removed), the various $A_2(E)$, $A_3(E)$, $A_4(E)$... are calculated and subtracted from $S(Q,E)$ to give a better $A_1(E)$. Iteration is complete when there is a good match between the measured $S(Q,E)$ and the $S(Q,E)$ recalculated with Eq. 2.60. When the multiphonon scattering is weak, perhaps only two iterations are necessary. When the multiphonon scattering is strong, the procedure becomes difficult and slow. If there is strong multiphonon scattering, however, the experiment is moving into the limit of classical scattering, so isolating the single phonon spectrum becomes more challenging. Section 4.2.2 develops further these concepts of multiphonon scattering.

2.3.3 Neutron Weighting

Strictly speaking, the phonon DOS from the procedure outlined above is not the true phonon DOS, but rather the neutron-weighted DOS. For an elemental scatterer, the neutron-weighted DOS is identical to the phonon DOS. This is not the usual case for a sample that contains more than one type of atom. The neutron-weighted phonon DOS is

$$g_{\mathrm{NW}}(\varepsilon) \propto \sum_d g_d(\varepsilon) \exp(-2W_d) \exp(2W) \frac{\sigma_d}{m_d} \tag{2.61}$$

where $\exp(-2W_d)$, σ_d and M_d are the Debye-Waller factor, total scattering cross-section and mass of atom d. The $g_d(\varepsilon)$ are the phonon partial DOS functions of Eq. 5.62. The Debye-Waller factor is an explicit function of $g_d(\varepsilon)$. The term $\exp(2W)$ is the average Debye-Waller correction; it is calculated from the self-consistent neutron-weighted DOS. At low temperatures, the factor $\exp[2(W - W_d)]$ is approximately unity.

For the case where σ_d/m_d is the same for all species d, $g_{\mathrm{NW}}(\varepsilon) \approx g(\varepsilon)$, but in general obtaining the true phonon DOS from the neutron-weighted phonon

[12] This $S(Q,E)$, called the "scattering law," is normalized so that it is the double Fourier transform of the Van Hove correlation function $G(\vec{r},t)$.

DOS requires a full analysis of the lattice dynamics. This can be performed by simulational procedures or by some types of analytical methods. The neutron-weight correction as well as other approximations involved in the data analysis can be overcome by fitting a dynamics model to the neutron-scattering data directly. Although this approach is both scientifically and computationally demanding, we foresee no better method for extracting the vibrational dynamics from inelastic neutron scattering measurements.

Chapter 3

Inelastic Electron Scattering and Spectroscopy

3.1 Inelastic Electron Scattering

Principles

This Chapter 3 first describes how high-energy electrons are scattered inelastically by materials, and then explains how electron energy-loss spectrometry (EELS) is used in materials research. Inelastic scattering occurs by the processes listed below in order of increasing energy loss, E. Although the scattering is called "inelastic," energy is conserved – the spectrum of energy gains by the sample is mirrored in the spectrum of energy losses of the high-energy electrons.

Electrons undergoing energy losses to crystal vibrations, quantized as *phonons* with $E \sim 10^{-2}$ eV, are indistinguishable from elastically scattered electrons, given the present state-of-the-art for EELS in a TEM.

With modern instrumentation, it is possible to measure *interband transitions* of electrons from occupied valence bands to unoccupied conduction bands of semiconductors and insulators. With $E \sim 2$ eV, these spectral features are quite close in energy to the intense zero-loss peak from elastic scattering, so resolving them has been a challenge.

In many solids, especially metals, the bonding electrons can be understood as a gas of free electrons. When a high-energy electron suddenly passes through this electron gas, *plasmons* may be created. Plasmons are brief oscillations of the free electrons, giving broadened peaks in EELS spectra. Plasmon energies ($E \sim 10$ eV) increase with electron density, so plasmon spectra can be used to estimate free electron density. Plasmon spectra are also useful for measuring the thickness of a TEM specimen because more plasmons are excited as the electron traverses a thicker specimen.

Electrons that ionize atoms by causing *core excitations* are used for microchemical analysis. Chemical spectroscopy with EELS measures the intensities

of "absorption edges," which are jumps in spectral intensity at the threshold energies for ejecting core electrons from atoms in the material ($10^2 < E < 10^4$ eV). After a core electron has been excited from the atom, the remaining "core hole" decays quickly, often by the emission of a characteristic x-ray. Characteristic x-rays with energies from $> 10^2$ to $> 10^4$ eV are used in energy dispersive x-ray spectrometry (EDS) for chemical analysis.

Methods

"Analytical transmission electron microscopy" uses EDS or EELS to identify the elements in a specimen, and to measure elemental concentrations or spatial distributions. To quantify chemical concentrations, a background is subtracted to isolate the heights of absorption edges (EELS) or the intensities of peaks in an x-ray energy spectrum (EDS). These isolated intensities are then compared for the different elements in the specimen, and often converted into absolute concentrations with appropriate constants of proportionality. The accuracy of quantification depends on the reliability of these constants, so significant effort has been devoted to understanding them.

In this chapter, after brief descriptions of an EELS spectrometer and features of a typical EELS spectrum, plasmon energies are discussed with a simple model of a free electron gas. The section on "core excitations" provides a higher-level treatment of how a high-energy electron can cause a core electron to be ejected from the atom. It turns out that the probability of a core electron excitation is proportional to the square of the Fourier transform of the product of the initial and final wavefunctions of the excited electron. The cross-section for inelastic scattering also has an angular dependence that must be considered when making quantitative measurements with EELS. Some experimental aspects of EELS measurements are presented, including energy-filtered TEM imaging.

This chapter then presents the principles of EDS in the TEM, which involves more physical processes than EELS. Interestingly, the cross-section for core ionization decreases with atomic number, but the cross-section for x-ray emission increases with atomic number in an approximately compensating way. This gives EDS spectrometry a balanced sensitivity for most elements except the very lightest ones.

3.2 Electron Energy-Loss Spectrometry (EELS)

3.2.1 Instrumentation

Spectrometer

After electrons have traversed a TEM specimen, a significant minority of them have lost energy to plasmons or core excitations, and exit the specimen with energies less than E_0, the energy of the incident electrons. (E_0 may be $200,000 \pm 0.5$ eV, for example.) To measure the energy spectrum of these losses, an EELS spectrometer can be mounted after the projector lenses of a TEM. The heart of a

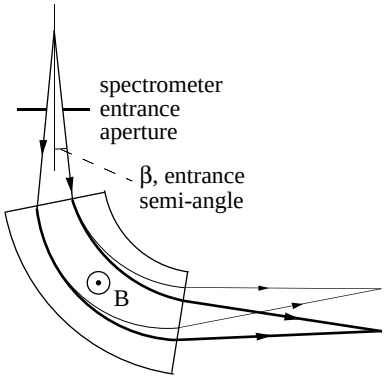


Figure 3.1: Some electron trajectories through a magnetic sector with uniform magnetic field, B. The light curves are trajectories for lower-energy electrons (those with larger energy loss, E), and heavier curves are for higher-energy electrons. The collection semi-angle of the spectrometer is β.

transmission EELS spectrometer is a magnetic sector, which serves as a prism to disperse electrons by energy. In the homogeneous magnetic field of the sector, Lorentz forces bend electrons of equal energies into arcs of equal curvature. Some electron trajectories are shown in Fig. 3.1.

The spectrometer must allow an angular range for electrons entering the magnetic sector, both for reasons of intensity and for measuring how the choice of scattering angle, ϕ, affects the spectrum (cf., (3.44)). A well-designed magnetic sector provides good focusing action. Focusing in the plane of the paper (the equatorial plane) is provided by the magnetic sector of Fig. 3.1 because the path lengths of the outer trajectories are longer than the path lengths of the inner trajectories. It is less obvious, but also true, that the fringing fields at the entrance and exit boundaries of the sector provide an axial focusing action. With good electron optical design, the magnetic sector is "double-focusing" so that the equatorial and axial focus are at the same point on the right of Fig. 3.1. Since the energy losses are small in comparison to the incident energy of the electrons, the energy dispersion at the focal plane of typical magnetic sectors is only a few microns per eV.

Electrons that lose energy to the sample move more slowly through the magnetic sector, and are bent further upwards in Fig. 3.1.[1] In a "serial spectrometer," a slit is placed at the focal plane of the magnetic sector, and a scintillation counter (see Sect. **??**) is mounted after the slit. Intensity is recorded only from those electrons bent through the correct angle to pass through the slit. A range of energy losses is scanned by varying the magnetic field in the spectrometer. A "parallel spectrometer," shown in the chapter title image, covers the focal plane of the magnetic sector with a scintillator and a position-sensitive photon detector such as a photodiode array. The post-field lenses Q1–Q4 magnify the energy dispersion before the electrons reach the scintillator. A parallel spectrometer has an enormous advantage over a serial spectrometer in its rate of data acquisition, but it requires calibrations for variations in pixel sensitivity.

The optical coupling of a magnetic sector spectrometer to the microscope usually puts the object plane of the spectrometer at the back focal plane of

[1] Their longer time in the magnetic field overcomes the weaker Lorentz forces.

the final projector lens (Fig. **??**). This back focal plane contains the diffraction pattern of the sample when the microscope is in image mode. When the microscope is operated in image mode, the spectrometer is therefore said to be "diffraction-coupled" to the microscope. With diffraction coupling, the collection angle, β, of the spectrometer is controlled by the objective aperture of the microscope. Alternatively, when the microscope is operated in diffraction mode, the back focal plane of the projector lens contains an image, and the spectrometer is said to be "image-coupled" to the microscope. With image-coupling, the collection angle, β, is controlled by an aperture at the entrance to the spectrometer (at top of Fig. 3.1).

Monochromator

The typical energy resolution for EELS spectrometers was about 1 eV or so for many years, but recent developments have allowed energy resolutions better than 0.1 eV on commercial microscopes. This is accomplished by starting with a field emission gun, often a Schottky effect gun (Sect. **??**), followed by an electron monochromator, often a Wien filter as described here. Electrons traveling through a Wien filter encounter a region with crossed electric and magnetic fields that induce competing forces on the electron. In a Wien filter these electric and magnetic forces are tuned to cancel for electrons of one velocity, v_0, which avoid deflection and pass through the exit aperture of the filter. Specifically, for an electron with velocity v_z down the optic axis along $\vec{z}$, a magnetic field oriented along $\vec{y}$ produces a force along $\vec{x}$, $F_x^{\mathrm{mag}} = ev_zB_y$. A Wien filter has an electric field along $\vec{x}$ in the same region, generating a force on the electron of $F_x^{\mathrm{el}} = -eE_x$. The special condition of cancelling forces, $F_x^{\mathrm{mag}} = -F_x^{\mathrm{el}}$, can be true for electrons of only one velocity, v_0:

$$\begin{aligned} ev_0B_y &= eE_x \,, \\ v_0 &= \frac{E_x}{B_y} \,. \end{aligned} \tag{3.1}$$

Electrons with velocities differing from v_0 are deflected, and do not pass through the exit aperture of the Wien filter. In practice, it is typical to operate the Wien filter at a voltage close to that of the electron gun itself, so the electron velocity through the filter will be slow enough that sub-eV resolution is possible with reasonable values of electric field, magnetic field, and aperture size. Biasing the Wien filter assembly near –100 or –200 keV can be challenging, however.

The Wien filter first disperses electrons of different energies into different angles, and then allows electrons of only a selected energy to pass through its exit aperture. Monochromatization therefore discards a substantial fraction of electrons – perhaps 80% of the electrons are lost when monochromating to 0.1 eV. When operating in STEM mode, the electron current is also reduced substantially when forming the smallest electron probes. It is typical to make compromises between the brightness of the image, the electron monochromatization, and the size of the probe – an increase in one usually requires a decrease

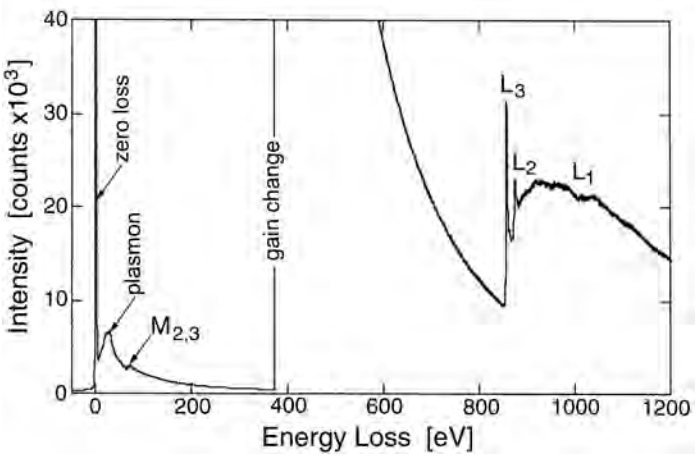


Figure 3.2: EELS spectrum of Ni metal, showing zero-loss peak, bulk plasmon, and L-edge with white lines at the edge. After [5.1].

in another. Manufacturers are constantly trying to find ways to improve these aspects of microscope performance.

3.2.2 General Features of EELS Spectra

A typical EELS spectrum is presented in Fig. 3.2. The enormous "zero-loss peak" is from electrons of 200,000 eV that passed through the specimen without any energy loss. The sharpness of this peak indicates that the energy resolution is about 1.5 eV. The next feature is at the energy loss $E = 25$ eV, from electrons having energies of 199,975 eV. It is the "first plasmon peak," caused by the excitation of one plasmon in the sample. With thicker specimens there may also be peaks at multiples of 25 eV from electrons that excited two or more plasmons in the specimen. The small bump in the data at 68 eV is not a plasmon peak, but rather a core loss. Specifically it is a Ni $M_{2,3}$ absorption edge caused by the excitation of $3p$ electrons out of Ni atoms. An enormous feature is seen at an energy loss of about 375 eV, but it is an artifact of the serial data acquisition method, and not a feature of the material. (At 375 eV the detector operation was changed from measuring an analog current to the counting of individual electron events.)

The background in the EELS spectrum falls rapidly with energy (the denominator of Δk^2 in (3.28) is partially responsible for this), and the next feature in the Ni spectrum of Fig. 3.2 is a core loss edge at 855 eV. This feature is caused by the excitation of $2p^{3/2}$ electrons out of the Ni atom, and is called the "L_3 edge." The L_2 edge at 872 eV is caused by the excitation of $2p^{1/2}$ electrons out of the atom. Right at the L_2 and L_3 edges are sharp, intense peaks known as "white lines" that originate from the excitation of $2p$ electrons into unoccupied $3d$ states at a Ni atom. Such features are typical of transition metals and their alloys as described in Section 3.2.3. More generally, unoccupied states such as antibonding orbitals are often responsible for sharp peaks at core edges.

Compared to plasmon excitations, the cross-sections for inner-shell ioniza-

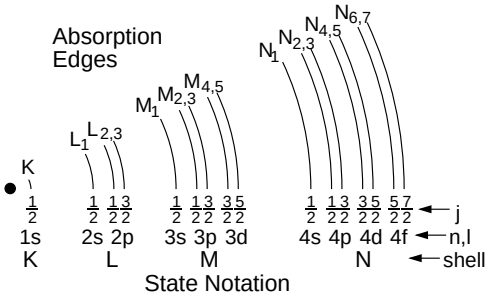


Figure 3.3: Possible edges of inner-shell ionizations and their nomenclature.

tions are relatively small, and become smaller at larger energy losses. To obtain good intensities, for many elements it is preferable to use absorption edges at lower energy losses (e.g., L and M). Some of the nomenclature of electronic transitions was given previously in Sect. **??**. Figure 3.3 shows an orbital representation and associated nomenclature for EELS edges.

3.2.3 * Fine Structure

Near-Edge Fine Structure

The region in an EELS spectrum around a core-loss edge often shows clear and reproducible structure that can be used to identify the local chemical environment. This "electron energy-loss near-edge structure" (ELNES) depends on the number and energy of unoccupied states at the excited atom. Chemists call these low-lying unoccupied states "lowest unoccupied molecular orbitals," and they include antibonding orbitals. Physicists call them "states above the Fermi energy," and they include the conduction band. A core electron can be excited into these unoccupied states, and the energy gained by the core electron during this transition is mirrored in the energy-loss spectrum of the high-energy electron. Simple metals with nearly-free electrons show core edges in EELS spectra that are smooth and without sharp features. On the other hand, materials with high densities of states just above the Fermi level, such as transition metals and rare-earth metals, have sharp features at their absorption edges associated with transitions into unoccupied d- and f-states, respectively. These features do not appear at all absorption edges, owing to the dipole selection rule where the angular momentum must change by ± 1. This selection rule allows transition metals with unoccupied d-states to have intense white lines at their $L_{2,3}$ edges, which involve excitations from core p-electrons, but not at their L_1 edges, which involve excitations from s-electrons (see Fig. 3.2).

The intensity of the white lines at the $L_{2,3}$ edge of Ni in Fig. 3.2 can be understood with the inelastic cross section for core shell ionizations (3.37), where ψ_β is an unoccupied $3d$ state and ψ_α is an occupied $2p$ core state, both centered at the Ni atom. The intensities of the white lines are larger when there are more unoccupied $3d$ states (the factor $\rho(E)$ in (3.37)). If the integral in (3.37) is evaluated, integrated intensities of the white lines can be used

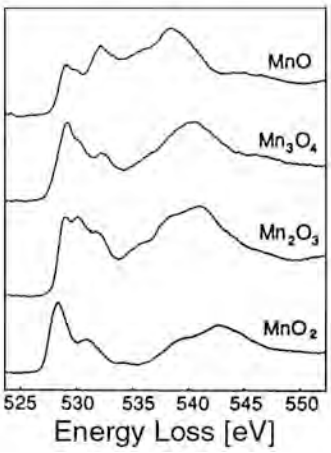


Figure 3.4: Oxygen *K*-edges from various manganese oxides, showing a variety of ELNES spectral features. After [5.2].

to quantify $\rho(E)$, the number of unoccupied 3*d* states at Ni atoms, and how this number changes with alloying or chemical bonding. Likewise, rare earth metals with unoccupied *f*-states have sharp features at their $M_{4,5}$ edges, which involve core *d*-electrons (but not at their M_2 or M_3 edges, which involve *p*-electrons). Semiconductors and insulators usually show distinct structure at their absorption edges, owing to the excitation of core electrons into unoccupied states above the band gap.

Because the number of unoccupied states is sensitive to the chemical and structural environment around the excited atom, ELNES can be used as a "fingerprint" of its local environment, even when the experimental systematics are not simple, or when electronic structure calculations are not possible. Figure 3.4 shows that the oxygen *K*-edge ELNES is sensitive to the local environment around the O atom in a variety of manganese oxides. The structure around 527-532 eV is dominated by the effects of chemical bonding on the density of electron states at the O atom, but the peak from 537-545 eV is more sensitive to the local positions of Mn atoms near the O atom – it is part of the "extended fine structure," described below.

Changes to the chemical environment around an atom alter the energy of the lowest unoccupied state, and therefore shift the onset energy of the core edge. Chemical shifts of absorption edges therefore reflect changes in the energies of the unoccupied states. More subtly, however, they also reflect changes in the energies of the core states. Any change to the outer atomic electrons, as caused by changes in chemical bonding for example, alters the intra-atomic electron-electron interactions. The energies of core electrons are therefore altered by changes in the outer electrons. For example, if an outer electron of a Li atom is transferred to a neighboring F atom, one may expect a lower-energy unoccupied state about the Li, and a shift of the Li *K* edge to lower energy. In fact, however, the loss of this electron in Li^+ reduces the screening of the core 1*s* electrons, causing them to be more tightly bound to the nucleus. This causes the absorption edge to shift to higher energies. Lithium has only three

electrons, so this effect is anomalous, but chemical shifts of absorption edges for all elements depend in part on the shifts in energies of the core electrons caused by intra-atomic screening.

Finally, we note that the core hole itself alters the energies of the atomic electrons. It is sometimes assumed that the removal of a core electron serves to increase the effective nuclear charge from Z to $Z + 1$, but the effects of a core hole on the energy levels of an unstable atom are not easy to understand.

Extended Fine Structure

Extended electron energy-loss fine structure (EXELFS) starts at energies where the outgoing electron state can be considered free of the atom, perhaps about 30 eV beyond the absorption edge. The state of the outgoing electron from the "central atom" is affected by the surrounding atoms, and self-interference occurs as the outgoing electron is backscattered from the nearest-neighbor shells of atoms. This process is illustrated schematically in Fig. 3.5. With changes in the wavelength of the outgoing electron, constructive and destructive interference occurs, causing the EXELFS signal, χ, to be oscillatory:

$$\chi(k) = \sum_j \frac{N_j f_j(k)}{r_j^2 k} N_j f_j(k) \, \mathrm{e}^{-2r_j/\lambda} \, \mathrm{e}^{-2\sigma_j^2 k^2} \sin\bigl(2kr_j + \delta_0 + \delta_j\bigr) \; . \tag{3.2}$$

Equation (3.2) includes a number of different effects, and its factors are best justified one-by-one. The sine function is the oscillatory interference of the outgoing electron wavefunction with itself as it travels the distance $2r_j$ from the central (excited) atom to a neighboring atom at r_j and back again. The phase of this electron wave is shifted by the amount δ_j upon scattering by the neighboring atom at the distance r_j, and by δ_0 from the central atom. These phase shifts generally depend on the electron wavevector, and this k-dependence must be known for quantitative work. The other factors in (3.2) are the number and backscattering strength of the neighboring atoms, N_j and $f_j(k)$, a qualitative decay factor to account for the finite lifetime of the outgoing electron state, $\mathrm{e}^{-2r_j/\lambda}$ (where λ is the electron mean-free-path), and a Debye-Waller factor, $\mathrm{e}^{-2\sigma_j^2 k^2}$, that attenuates $\chi(k)$. Here σ_j^2 is a mean-squared displacement of the central atom relative to its neighboring atoms, typically originating with temperature or disorder in the local structure. The sum in (3.2) is over the neighboring atoms, and typically includes the first- and second-nearest-neighbor (1nn and 2nn) shells around the central atom.

Figure 3.6 shows some steps in a typical EXELFS analysis, in this case for the $L_{2,3}$ edge from a slightly-oxidized sample of bcc Fe metal. Figure 3.6a shows the absorption edge after correction for the pre-edge background. The region of interest begins above the L_3 and L_2 edges. Unfortunately, the L_1 edge ($2s$ excitation) appears as a feature in the region of interest, so it is best to work with data at energies beyond the L_1 edge. The useful data range did include the oscillations with broad peaks at about 920 and 1000 eV (barely visible in Fig.

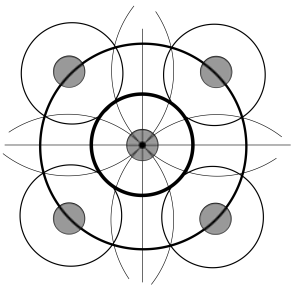


Figure 3.5: Pictorial representation of the electron interference that gives rise to EXELFS. Crests of the electron wavefunction emanating from a central atom are drawn with an amplitude that diminishes with distance. For this particular wavevector and phase shifts, the electron wave crest backscattered from the four neighboring atoms is in phase with the wave crest emanating from the central atom, giving constructive interference and an enhanced probability for the emission of the electron.

3.6). Extracting these small oscillations from the monotonic decay characteristic of an isolated atom is usually done by fitting a cubic spline function through the EXELFS oscillations. Subtracting this spline fit reveals the oscillations in energy, which are converted to k-space as in Fig. 3.6b, using the wavevector dependence on energy above the absorption edge, E_a (where k is in Å^{-1}):

$$E - E_a = \frac{\hbar^2 k^2}{2m_e} = 3.81\, k^2\,[\text{eV}]\ , \tag{3.3}$$

Real space periodicities are obtained from the data of Fig. 3.6b by taking their Fourier transform.[2] The periodicities in real space are not affected significantly if $\chi(k)$ is multiplied by a power of k, and doing so helps to sharpen the peaks in the real space data. The real-space function of Fig. 3.6c, called a "pseudo-" or "raw-" radial distribution function, was obtained by taking the Fourier transform of $k\chi(k)$. The peak at 2.25 Å corresponds approximately to the position of the 1nn shell of Fe atoms in bcc Fe (2.02 Å), but a discrepancy is expected because the phase shifts δ_j and δ_0 of (3.2) were not included in the data analysis. For comparative work with similar specimens, however, this simple Fourier transform method may be adequate.

Better-known than EXELFS is EXAFS (extended x-ray absorption fine structure) spectroscopy, performed with tuneable synchrotron radiation. EXAFS is identical to EXELFS, except that the excitation of the central atom is caused by a photon. The energy of the incident photon is tuned from below an absorption edge to well above it. The self-interference of the backscattered photoelectron is seen in the data as decreased or increased photon transmission through the sample (or electron yield in another variant of the EXAFS technique). The

[2]It is an approximation to ignore the slight phase difference between the L_1 and L_2 EXELFS oscillations, and to neglect the L_1 EXELFS, but the approximation is not too bad.

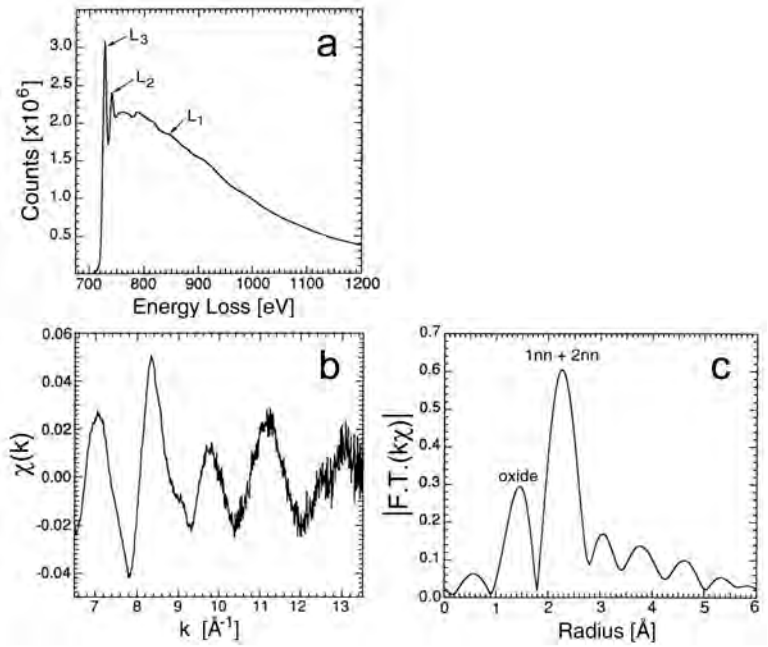


Figure 3.6: (**a**) Fe *L*-edge from pure Fe metal at 97 K. Pre-edge background was subtracted, but no corrections were performed for plasmon excitations, which do not affect the gradual EXELFS structure. (**b**) Fe $L_{2,3}$ edge EXELFS extracted from data in **a**. (**c**) Magnitude of Fourier transform of data in **b**. After [5.3].

analysis of the $\chi(k)$ data is identical to that of EXELFS, and (3.2) was originally proposed for EXAFS.

There is a stronger E-dependence for EXELFS spectra than for EXAFS spectra, causing EXELFS to be more practical than EXAFS for energies below about 2 keV. Nevertheless, EXAFS is more practical at higher energies, and higher energies have two advantages. Atomic levels at higher energies are better separated in energy, making it easier to obtain wide ranges of energy where the extended fine structure can be measured without interruption from other absorption edges. The second advantage of EXAFS is its ability to work with K-shell excitations of many elements, whose simpler structure allows their $\chi(k)$ to be interpreted more reliably. On the other hand, EXELFS can be performed readily on local regions of material identified in TEM images. Synchrotron beamline optics including x-ray mirrors and Fresnel zone plates now allow EXAFS measurements on areas smaller than ~ 1 μm, however.

3.3 Plasmon Excitations

3.3.1 Plasmon Principles

A fast electron jolts the free electrons when it passes through a material. The displaced charge creates an electric field to restore the equilibrium distribution of electrons, but the charge distribution oscillates about equilibrium for a number of cycles. These charge oscillations, called "plasmons," are quantized in energy. Larger energy losses correspond to the excitation of more plasmons, not to an increase in the energy of a plasmon. In most EELS spectra, the majority of inelastic scattering events are plasmon excitations.

To find the characteristic oscillation frequency of a plasmon, consider the rigid translation of a wide slab of electron density by a small amount, x, as in Fig. 3.7. At the bottom surface of the slab all the electrons are removed, but at the top the electron density is doubled. This charge disturbance therefore sets up the electric field, $\mathcal{E}$, of a parallel-plate capacitor:

$$\mathcal{E} = 4\pi\sigma_s \,, \tag{3.4}$$

where σ_s is the surface charge density equal to the electron charge, e, times ρ, the number of electrons per unit volume, times the displacement, x:

$$\sigma_s = e\rho x \,. \tag{3.5}$$

The field, $\mathcal{E}$, provides the restoring force to move the electron slab back to its original position. The electrostatic restoring force *per unit volume* of slab is:

$$F = -e\rho\,\mathcal{E} \,. \tag{3.6}$$

Substituting (3.5) into (3.4), and then into (3.6) gives a restoring force linear in the displacement:

$$F = -e^2\rho^2 4\pi x \,. \tag{3.7}$$

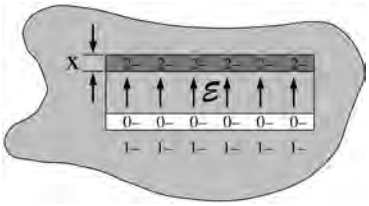


Figure 3.7: Displacement of a slab of electric charge, leading to doubling of the charge density at the top of the slab over thickness x, and depletion of charge at the bottom. A wide, flat slab idealizes the problem as one dimensional.

The Newtonian equation of motion per unit volume of the electron slab is:

$$F = \rho\, m_{\rm e} \frac{{\rm d}^2 x}{{\rm d}t^2} \ . \tag{3.8}$$

Substituting (3.7) into (3.8) gives:

$$\frac{{\rm d}^2 x}{{\rm d}t^2} = -\left(\frac{4\pi e^2 \rho}{m_{\rm e}}\right) x \ . \tag{3.9}$$

Equation (3.9) is the equation of motion for an undamped harmonic oscillator with the characteristic frequency:

$$\omega_{\rm p} = \sqrt{\frac{4\pi e^2 \rho}{m_{\rm e}}} = 5.64 \times 10^4 \ \sqrt{\rho} \ , \tag{3.10}$$

where the units of ρ are [electrons cm^{-3}], and $\omega_{\rm p}$ is [Hz]. With analogy to a mechanical oscillator, the electron density provides the stiffness. The higher the electron density, the higher the plasmon frequency. For metals, assuming an approximate free electron density of $\rho = 10^{23}$ electrons cm^{-3}, $\omega_{\rm p} \simeq 2 \times 10^{16}$ Hz.[3] The characteristic energy of such an oscillation is the plasmon energy, given by:

$$E_{\rm p} = \hbar\, \omega_{\rm p} \ , \tag{3.11}$$

and for our example $E_{\rm p} = (6.6{\times}10^{-16}\ {\rm eV\,s})\,(2{\times}10^{16}\ {\rm s}^{-1}) \approx 13\ {\rm eV}$.

In EELS, intense plasmon peaks are prominent at energy losses of 10–20 eV. Plasmons are not long-lived, however, often because they promote excitations of electrons near the Fermi energy. Plasmon peaks therefore tend to be broadened in energy.[4] Free electron metals such as aluminum have sharper plasmon peaks than transition metals, which have a high density of states at the Fermi energy. Compared to core electron excitations, however, plasmon excitations do not provide much detailed information about the individual atom species in the material.

[3] The present one-dimensional approach is not necessarily reliable for relating the free electron density to the plasmon energy. A more general approach uses the imaginary component of the dielectric constant of the material, and the theory uses the mathematics discussed in Problem 5.6.

[4] "Lifetime broadening" is understandable from the uncertainty principle: $\Delta E \Delta t \simeq \hbar$. A short lifetime Δt comes with a large uncertainly in energy, ΔE.

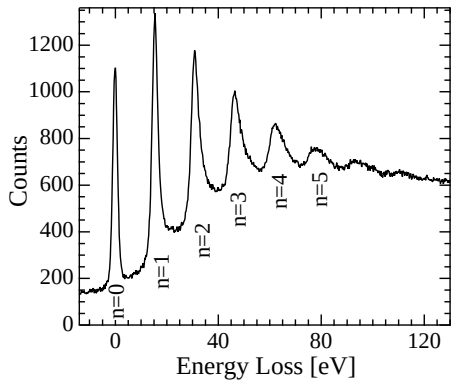


Figure 3.8: Low-loss spectrum taken from a thick sample of ~ 120 nm Al metal on C using 120 keV electrons and β = 100 mrad. Plasmon peaks are visible at energies of $n \times 15$ eV, where n is the number of plasmons excited in the sample. After [5.4].

3.3.2 * Plasmons and Specimen Thickness

The characteristic length or "mean free path," $\overline{\lambda}$, over which a 100 keV electron excites one plasmon is about 100 nm in metals and semiconductors. This is an average length, so in a TEM specimen of even 50 nm, some electrons excite one, two, or more plasmons. The probability P_n for the excitation of n plasmons in a sample of thickness, t, is determined by the statistics of Poisson processes:

$$P_n = \frac{1}{n!}\left(\frac{t}{\overline{\lambda}}\right)^n \mathrm{e}^{-t/\overline{\lambda}} = \frac{I_n}{I_\mathrm{t}} \ , \tag{3.12}$$

where I_n is the number of counts in the n^{th} plasmon peak, and I_t is the number of counts in all plasmon peaks with $n \geq 0$ (I_t includes the zero-loss peak for which $n = 0$). The EELS spectrum in Fig. 3.8 shows distinct plasmon peaks. After subtracting a background from other inelastic processes (originating with the Al L-edge and a contribution from oxides and the substrate), P_n is obtained as the fractional area of the n^{th} plasmon peak.

Good samples for TEM imaging are several times thinner than the sample used for Fig. 3.8, but plasmon peak areas still offer a practical way to determine thicknesses of thin samples. Setting $n = 0$ in (3.12), the thickness, t, is:

$$\frac{t}{\overline{\lambda}} = \ln\!\left(\frac{I_\mathrm{t}}{I_0}\right) \ . \tag{3.13}$$

Measurement of I_t and I_0 (the zero-loss or the $n = 0$ plasmon peak) involves the choice of the energies ε, δ and Δ, which define the limits of integration, as illustrated in Fig. 3.9.[5] The lower limit ($-\varepsilon$) of the zero-loss region can be taken anywhere to the left of the zero-loss peak where the intensity has decreased to zero, the separation point δ between the zero-loss and inelastic

[5] If a collection aperture limits the angles recorded by the spectrometer to a maximum angle β, $\overline{\lambda}$ in (3.13) must be interpreted as an effective mean free path, $\overline{\lambda}(\beta)$.

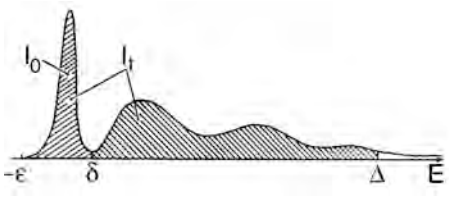


Figure 3.9: The integrals and energies involved in measuring specimen thickness by the log-ratio method. After [5.5].

regions may be taken as the first minimum in intensity, and $\Delta \approx 100\,\text{eV}$ is usually sufficient to include most of the inelastic scattering in relatively thin, low Z materials (for high Z and/or thick specimens, several hundred eV should be used since the scattered intensity is shifted to higher energy loss). Equation (3.13) has been shown to give 10 % accuracy for relative thickness measurements on samples as thick as $t = 5\overline{\lambda}$. Some deviations from the intensities of (3.12) are of course expected when the electron beam passes through regions of non-uniform thicknesses or composition, and when other contributions are present in the low-loss spectrum.

Absolute determinations of specimen thickness require values for the total inelastic mean free path. For materials of known composition, it is possible to calculate a value for the mean free path according to the semi-empirical equation [5.5]:

$$\overline{\lambda} \approx \frac{106\,F}{\ln(2\beta E_0/E_\text{m})}\frac{E_0}{E_\text{m}}\,, \tag{3.14}$$

where units for $\overline{\lambda}$ are [nm], β is the collection semiangle [mrad], E_0 is the incident energy [keV], E_m is a mean energy loss [eV] that depends on the composition of the sample, and F is a relativistic factor:

$$F = \frac{1 + E_0/1022}{(1 + E_0/511)^2}\,, \tag{3.15}$$

and $F = 0.768$ for $E_0 = 100\,\text{keV}$ and 0.618 for $E_0 = 200\,\text{keV}$. For a specimen of average atomic number Z, E_m can be obtained from the semi-empirical formula:

$$E_\text{m} \approx 7.6Z^{0.36}\,. \tag{3.16}$$

For large collection apertures, i.e., $\beta > 20\,\text{mrad}$ for $E_0 = 100\,\text{keV}$ or $\beta > 10\,\text{mrad}$ at 200 keV, (3.14) becomes inapplicable and the mean free path saturates at a value independent of β.

Specimen thickness measurement by this EELS plasmon technique has some advantages over other techniques for measuring specimen thickness (such as CBED) because it can be applied for a wide range of specimen thicknesses, including very thin specimens, and for specimens that are highly disordered or amorphous. Table 3.1 lists some values for calculated (with (3.10) and (3.11)) and measured plasmon energies E_p, widths of the plasmon peaks ΔE_p, characteristic scattering angles for plasmons ϕ_{E_p}, and calculated mean free paths $\overline{\lambda}$, for 100 keV incident electrons.

Table 3.1: Plasmon data for selected materials [5.5]

Material	E_p (calc.) (eV)	E_p (expt.) (eV)	ΔE_p (eV)	ϕ_{E_p} (mrad)	$\overline{\lambda}$ (nm)
Li	8.0	7.1	2.3	0.039	233
Be	18.4	18.7	4.8	0.102	102
Al	15.8	15.0	0.5	0.082	119
Si	16.6	16.5	3.7	0.090	115

Table 3.2: Energy notation

variable	definition
E	energy transfer from incident electron to atomic electron
E_0	energy of incident electron (T + mass energy) e.g., 100.00 keV
T	incident kinetic energy (low E_0: $T \simeq E_0$) (high E_0: $T = mv^2/2 < E_0$)
E_α	energy of bound atomic electron
$E_{\alpha\beta}$	difference in energy between atomic states α and β
E_a	energy of atomic absorption edge (e.g., E_K), $E_a \simeq -E_\alpha$
E_p	plasmon energy
E_m	mean energy loss
sign	all variables are positive except E_α

3.4 Core Excitations

3.4.1 Scattering Angles and Energies – Qualitative

When a high-energy electron undergoes *inelastic* scattering, its energy loss, E, is actually a *transfer* of energy to the sample. When this energy is transferred to an atomic electron, the atomic electron may find an unoccupied electron state about the same atom, or it may leave the atom entirely (i.e., the atom is ionized). The total energy and the total momentum are conserved, but the scattering redistributes the energy and momentum between the high-energy electron and the atomic electron. The two electrons have coupled behavior. In particular, the probabilities and energies for the allowed excitations of the atomic electron are mirrored in the spectrum of energy losses of the high-energy electron. Relevant energies and their notation are listed in Table 3.2.

When a high-energy electron transfers energy to a core electron, the wavevector of the high-energy electron is changed in both magnitude and direction. The change in energy is obtained from the change in *magnitude* of the wavevector. The change in momentum is obtained from both the change in *direc-*

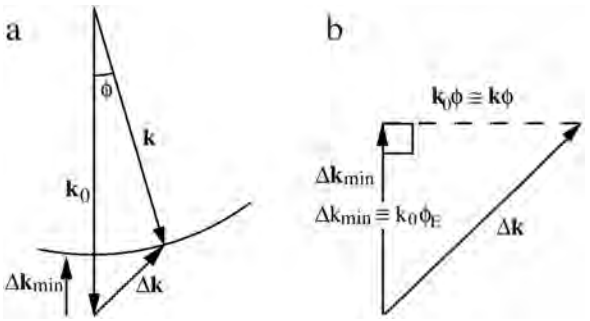


Figure 3.10: Kinematics of inelastic electron scattering. (**a**) Definitions, with sphere of constant E. (**b**) Enlargement valid for small ϕ, or equivalently for small Δk.

tion and the change in *magnitude* of the wavevector. Total momentum is conserved, and before scattering the total momentum is with the incident electron, $\vec{p}_0 = m_e\vec{v}_0 = \hbar\vec{k}_0$. After scattering, the momentum transfer to the atomic electron must be $\hbar\Delta\vec{k} = \hbar(\vec{k} - \vec{k}_0)$. This same $\Delta\vec{k} \equiv \vec{k} - \vec{k}_0$ is used for elastic scattering (Fig. 1.5), but inelastic scattering has an extra degree of freedom because $k \neq k_0$. Figure 3.10a shows that increasing ϕ, the scattering angle, gives larger values of Δk for the same E. Momentum conservation requires that the head of the wavevector $\Delta\vec{k}$ lies along the circle of radius k. The scattering vector, $\Delta\vec{k}$, can be zero only when both $\phi = 0$ and $E = 0$.[6] When $E = 0$ but $\phi \neq 0$, Δk cannot be zero – this is the case for elastic scattering in diffraction experiments.

We first consider general features of how inelastic scattering depends on E and $\Delta\vec{k}$. For E only slightly larger than an absorption edge energy, E_a, the inelastic scattering is forward-peaked with a maximum intensity at the smallest Δk. Figure 3.10a shows that when $\phi = 0$ and $E \neq 0$, there is a nonzero minimum value of Δk, corresponding to inelastic scattering in the forward direction: $\Delta\vec{k}_{min} \equiv (|\vec{k}| - |\vec{k}_0|)\widehat{\vec{k}_0}$. In a particle model, these low-angle scatterings correspond to soft collisions with large impact parameters (meaning that the high-energy electron does not pass close to the center of the atom). The energy transfer is still large ($E \simeq E_a$), unlike most soft classical collisions,[7] but the outgoing core electron carries insignificant kinetic energy and Δk is small. For small Δk, the scattering is sensitive to the large r (long-range) features of the scattering potential.

On the other hand, at larger energy losses ($E \gg E_a$), the scattering is at higher angles, corresponding to hard collisions with small impact parameters. The outgoing core electron carries significant kinetic energy (equal to $E - E_a$), and

[6] This is the case for no scattering, or for elastic forward scattering, which involves a phase shift.

[7] A classical analogy can be contrived. Suppose a fast billiard ball passes near a second ball located in a pit, and some of the kinetic energy of the fast ball is used to lift the second ball out of the pit. If the second ball leaves the pit with minimal velocity, momentum conservation allows little change in direction of the fast ball as it slows down. Quantum mechanics uses the same energy and momentum arguments, but Sect. 3.4.2 provides the probabilistic mechanism for "lifting the ball out of the pit." This mechanism provides an additional dependence on E and $\Delta\vec{k}$.

the momentum transfer deflects the high-energy electron. For sufficiently large E, we expect the momentum transfer and energy transfer to be understandable by collisional kinematics, with little influence from the characteristics of the atom such as E_a. In fact, for larger energy transfers the inelastic intensity becomes concentrated around a specific value of Δk such that:

$$\Delta k_B = \frac{\sqrt{2mE}}{\hbar} \ . \tag{3.17}$$

This peak in Δk corresponds to the momentum transfer in classical "elastic" scattering of a moving ball (electron) by another ball initially at rest. This peak is called the "Bethe peak," and in a two-dimensional plot of inelastic scattering intensity versus Δk and E, these peaks become a "Bethe ridge" (cf., Fig. 3.11). Substituting into (3.17) a handy expression involving the Bohr radius, a_0, and the Rydberg energy, $E_R = \hbar^2(2ma_0^2)^{-1}$, we obtain for Δk_B:

$$(\Delta k_B a_0)^2 \approx \frac{E}{E_R} \ , \tag{3.18}$$

for which the equivalent scattering angle for the Bethe ridge, ϕ_r, is:

$$\phi_r \approx \sqrt{\frac{E}{E_0}} \ . \tag{3.19}$$

The results of (3.17)–(3.19) are valid for small ϕ and non-relativistic electrons.

Experimentally, we count electrons. The energy spectrum of these electrons, $\rho(E)\,\mathrm{d}E$, varies with solid angle, Ω. With reference to Fig. 1.5, the three different $\mathrm{d}\Omega_j$ will have different energy spectra. The most detailed experimental measurements would provide an energy spectrum at each differential solid angle, $\mathrm{d}\Omega$. The number of electrons detected in a range $\mathrm{d}\Omega$ around Ω and a range $\mathrm{d}E$ around E is proportional to the "double-differential cross-section," $\mathrm{d}^2\sigma/\mathrm{d}\Omega\,\mathrm{d}E$. In practice, there is often cylindrical symmetry around the forward beam, so we may need only the ϕ-dependence (where $\phi = 2\theta$ in scattering angle). Experimental EELS spectra are measurements of intensity versus energy loss, $I(E)$, over a finite range of scattering angles, ϕ.

Theoretically, we calculate the probability that a transfer of energy, E, and momentum, $\hbar\,\vec{\Delta k}$, occurs between a high-energy electron and an atomic electron. To relate the theory to measured EELS spectra, we then need:

- The variation of the inelastic scattering over the parameter space of (ϕ,E). This is given by a double-differential cross-section, $\mathrm{d}^2\sigma_{in}/\mathrm{d}\phi\,\mathrm{d}E$, described in Sect. 3.4.3 ($\mathrm{d}\phi$ refers to rings of solid angle). This $\mathrm{d}^2\sigma_{in}/\mathrm{d}\phi\,\mathrm{d}E$ includes the "generalized oscillator strength" of the specific atom.

- EELS spectra, $I(E)$, are measured over a range of ϕ, so we need to integrate $\mathrm{d}^2\sigma_{in}/\mathrm{d}\phi\,\mathrm{d}E$ over angle to obtain the differential cross-section, $\mathrm{d}\sigma_{in}/\mathrm{d}E$, described in Sect. 3.4.5.

- Compositional analysis by EELS uses total intensities, given by the total cross-section, σ (or more typically by partial cross-sections, corresponding to a finite range in energy), as described in Sect. 3.4.6. This total probability for ionizing an atom is also needed for understanding EDS spectra, which measure x-ray emissions after the atom is ionized.

3.4.2 ‡ Inelastic Form Factor

Here we calculate the probability of an inelastic scattering process involving the excitation of a core electron. In this process, a high-energy electron excites a second electron from a bound atomic state into a state of higher energy. Since two electrons are involved, for conciseness we employ the Dirac bra and ket notation.[8] The high-energy electron, "electron 1," is initially in a plane wave state $\left|\vec{k}_0\right\rangle$, and after scattering it is in the state $\left|\vec{k}\right\rangle$. The atomic electron, "electron 2," is initially in the bound state $|\alpha\rangle$. After scattering, electron 2 is in the state $|\beta\rangle$, which may be either a bound state that is initially unoccupied, or a spherical (or plane) wave state if electron 2 is ejected from the atom. For inelastic scattering, $\left|\vec{k}\right| \neq \left|\vec{k}_0\right|$ and $\alpha \neq \beta$. The Schrödinger equation with the initial state is written as:

$$H_0 \left|\vec{k}_0, \alpha\right\rangle = (E_0 + E_\alpha) \left|\vec{k}_0, \alpha\right\rangle \ . \tag{3.20}$$

So long as the two electrons are far apart and therefore non-interacting, the two-electron system obeys the unperturbed Hamiltonian:

$$H_0 = -\frac{\hbar^2}{2m_\mathrm{e}}\nabla_1^2 - \frac{\hbar^2}{2m_\mathrm{e}}\nabla_2^2 + V(\vec{r}_2) \ . \tag{3.21}$$

The coordinates of the high-energy electron 1 are $\vec{r}_1$, and the coordinates of the atomic electron 2 are $\vec{r}_2$. With different coordinates, each Laplacian in (3.21) acts on only one of the two electrons, and the potential energy term involves only electron 2. In such problems we can express the initial state as a product of one-electron wavefunctions: $\left|\vec{k}_0, \alpha\right\rangle = \left|\vec{k}_0\right\rangle |\alpha\rangle$, and the final state as: $\left|\vec{k}, \beta\right\rangle = \left|\vec{k}\right\rangle |\beta\rangle$. When using a product wavefunction in (3.21), the factor for electron 2, $|\beta\rangle$, is a constant under the operations of ∇_1^2, and $\left|\vec{k}\right\rangle$ is a constant under the operations of ∇_2^2 and $V(\vec{r}_2)$. A "constant factor" does not affect the solution of the Schrödinger equation for the other wavefunction of the product. The Hamiltonian of (3.21) is therefore equivalent to two independent Hamiltonians for two independent electrons. This is as expected when the two electrons have no interaction.

[8] Recall that Dirac notation is free of spatial coordinates and explicit functional forms of wavefunctions, but these are obtained with the position operator for coordinate set 1, $\vec{r}_1$, as: $\vec{r}_1 \left|\vec{k}\right\rangle = \psi(\vec{r}_1)$. Actual evaluations of integrals require expressions such as: $\langle a| H |a\rangle = \int \psi_a^* H \psi_a \mathrm{d}^3 r$. When $|\alpha\rangle$ is an eigenstate of H, $\langle\alpha| H |\alpha\rangle = E_\alpha \langle\alpha|\alpha\rangle = E_\alpha$, since the state functions are normalized. State functions are orthonormal, so $\langle\alpha|\beta\rangle = 0$ and $\langle\alpha|\alpha\rangle = 1$.

As the high-energy electron approaches the atom, we must consider two perturbations of our two-electron system. One perturbation is the Coulombic interaction of electron 2 with the electron 1, which is $+e^2/\left|\vec{r}_1 - \vec{r}_2\right|$. The second perturbation is the interaction of the high-energy electron 1 with the potential from the rest of the atom,[9] $V(\vec{r}_1)$. The perturbation Hamiltonian, H', is:

$$H' = \frac{e^2}{\left|\vec{r}_1 - \vec{r}_2\right|} + V(\vec{r}_1) \; . \tag{3.22}$$

This perturbation H' couples the initial and final states of the system. The stronger the coupling, the more probable is the transition from the initial state $\left|\vec{k}_0\right\rangle |\alpha\rangle$ to the final state $\left|\vec{k}\right\rangle |\beta\rangle$. It is a result from time-dependent perturbation theory that the wavefunction of the scattered electron 1 is an outgoing spherical wave times the form factor, $f(\vec{k},\vec{k}_0)$ (cf., (1.50)), where:

$$f(\vec{k},\vec{k}_0) = \frac{-m_e}{2\pi\hbar^2} \langle\beta| \left\langle\vec{k}\right| H' \left|\vec{k}_0\right\rangle |\alpha\rangle \; . \tag{3.23}$$

Substitution of (3.22) into (3.23) gives:

$$\begin{aligned} f(\vec{k},\vec{k}_0) = \frac{-m_e}{2\pi\hbar^2}\Big[\, e^2 \quad & \langle\beta| \left\langle\vec{k}\right| \frac{1}{\left|\vec{r}_1 - \vec{r}_2\right|} \left|\vec{k}_0\right\rangle |\alpha\rangle \\ & + \langle\beta| \left\langle\vec{k}\right| V(\vec{r}_1) \left|\vec{k}_0\right\rangle |\alpha\rangle \,\Big] \; . \end{aligned} \tag{3.24}$$

When evaluating the second term of (3.24), the coordinates of electron 2 appear only in the atomic wavefunctions $|\alpha\rangle$ and $\left|\beta\right\rangle$, so these wavefunctions are moved out of the integral involving the coordinates of electron 1:

$$\begin{aligned} f(\vec{k},\vec{k}_0) = \frac{-m_e}{2\pi\hbar^2}\Big[\, e^2 \quad & \langle\beta| \left\langle\vec{k}\right| \frac{1}{\left|\vec{r}_1 - \vec{r}_2\right|} \left|\vec{k}_0\right\rangle |\alpha\rangle \\ & + \langle\beta|\alpha\rangle \left\langle\vec{k}\right| V(\vec{r}_1) \left|\vec{k}_0\right\rangle \,\Big] \; . \end{aligned} \tag{3.25}$$

For inelastic scattering we have $\alpha \neq \beta$, so the second term[10] is zero by the orthogonality of the atomic wavefunctions. To be explicit in notation, we denote the inelastic contribution to $f(\vec{k},\vec{k}_0)$ as $f_{in}(\vec{k},\vec{k}_0)$, and call it the "inelastic form factor." To calculate $f_{in}(\vec{k},\vec{k}_0)$, we use spatial coordinate representations

[9] For the potential from the rest of the atom, we could use the potential of an atom without a core electron, since we consider electron 2 separately.

[10] For elastic scattering there is no transfer of energy from the high-energy electron (electron 1) to the atomic electron (electron 2), so $\alpha = \beta$. By the orthonormality of the atomic wavefunctions we know that $\langle\alpha|\alpha\rangle = 1$, so this second term is nearly equal to the right hand side of (1.76). The difference is that the scattering potential from electron 2 is considered separately as the first term in (3.25), but together the two terms in (3.25) account for the scattering from the entire atom.

for our wavefunctions. The non-zero first term of (3.25) is:

$$f_{\text{in}}(\vec{k},\vec{k}_0) = \frac{-m_e e^2}{2\pi\hbar^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \mathrm{e}^{-\mathrm{i}\vec{k}\cdot\vec{r}_1}\mathrm{e}^{\mathrm{i}\vec{k}_0\cdot\vec{r}_1}\frac{1}{\left|\vec{r}_1-\vec{r}_2\right|} \times \psi^*_\beta(\vec{r}_2)\,\psi_\alpha(\vec{r}_2)\,\mathrm{d}^3\vec{r}_2\,\mathrm{d}^3\vec{r}_1\ . \tag{3.26}$$

We change variables: $\vec{r} \equiv \vec{r}_1 - \vec{r}_2$ (so $\vec{r}_1 = \vec{r} + \vec{r}_2$), and $\Delta\vec{k} \equiv \vec{k} - \vec{k}_0$, and separate the integrations:

$$f_{\text{in}}(\vec{k},\vec{k}_0) = \frac{-m_e e^2}{2\pi\hbar^2} \int_{-\infty}^{+\infty}\mathrm{e}^{-\mathrm{i}\Delta\vec{k}\cdot\vec{r}}\frac{1}{\left|\vec{r}\right|}\mathrm{d}^3\vec{r} \times \int_{-\infty}^{+\infty}\mathrm{e}^{-\mathrm{i}\Delta\vec{k}\cdot\vec{r}_2}\,\psi^*_\beta(\vec{r}_2)\,\psi_\alpha(\vec{r}_2)\,\mathrm{d}^3\vec{r}_2\ . \tag{3.27}$$

Equation (3.27) shows that the only dependence of f_{in} on $\vec{k}$ and $\vec{k}_0$ is through their difference, $\Delta\vec{k}$. The integral over $\vec{r}$ is $4\pi/\Delta k^2$ (A.27):

$$f_{\text{in}}(\Delta\vec{k}) = \frac{-2m_e e^2}{\hbar^2\Delta k^2}\int_{-\infty}^{+\infty}\mathrm{e}^{-\mathrm{i}\Delta\vec{k}\cdot\vec{r}_2}\psi^*_\beta(\vec{r}_2)\psi_\alpha(\vec{r}_2)\mathrm{d}^3\vec{r}_2\ . \tag{3.28}$$

This inelastic form factor, $f_{\text{in}}(\Delta\vec{k})$, is the amplitude of the outgoing high-energy electron wavefunction along the direction $\vec{k} = \vec{k}_0 + \Delta\vec{k}$ when the high-energy electron excites the atomic transition $\psi_\alpha \rightarrow \psi_\beta$. The inelastic form factor has many similarities to the elastic form factor, $f_{\text{el}}(\Delta\vec{k})$, of (1.76). Specifically, the second term of (**??**) for $f_{\text{el}}(\Delta\vec{k})$, which describes elastic scattering from the atomic electron density, $\rho(\vec{r})$, has the same form as (3.28). It is convenient to think of both the inelastic and elastic form factors in a common way. Along the direction $\vec{k} = \vec{k}_0 + \Delta\vec{k}$, wavelets are emitted from all sub-volumes, $\mathrm{d}^3\vec{r}_2$, of the atom. Each wavelet has a relative phase $\mathrm{e}^{-\mathrm{i}\Delta\vec{k}\cdot\vec{r}_2}$, and its amplitude for elastic scattering is proportional to an electron density. The full wave is the coherent sum (integration) of wavelets from all volumes of the atom, weighted by an electron density. For elastic scattering the electron density is the usual electron density, $\rho(\vec{r}) = \psi^*_\alpha(\vec{r})\psi_\alpha(\vec{r})$. For inelastic scattering, however, this "density" is the overlap of the initial and final wavefunctions, $\rho'(\vec{r}) = \psi^*_\beta(\vec{r})\psi_\alpha(\vec{r})$.

Note the common prefactors of $f_{\text{el}}(\Delta\vec{k})$ of (**??**) and $f_{\text{in}}(\Delta\vec{k})$ of (3.28). Recall that the factor of Δk^{-2} originates with the Fourier transform of the Coulomb potential (**??**). Using the definition of the Bohr radius, $a_0 = \hbar^2/(m_e e^2)$, this prefactor is $2/(a_0\Delta k^2)$, which has dimensions of length. We now obtain the

differential cross-section for inelastic scattering, $\mathrm{d}\sigma_{\mathrm{in}}/\mathrm{d}\Omega$, as $f^*_{\mathrm{in}}\, f_{\mathrm{in}}$ (1.21)[11]:

$$\frac{\mathrm{d}\sigma_{\mathrm{in}}(\vec{\Delta k})}{\mathrm{d}\Omega} = \frac{4}{a_0^2 \Delta k^4} \left| \int_{-\infty}^{+\infty} \mathrm{e}^{-\mathrm{i}\Delta\vec{k}\cdot\vec{r}_2} \psi^*_\beta(\vec{r}_2)\psi_\alpha(\vec{r}_2)\mathrm{d}^3\vec{r}_2 \right|^2 . \tag{3.29}$$

Although energy is transferred from the high-energy electron 1 to the atomic electron 2, the total energy is conserved. In the transition $\left|\vec{k}_0\right\rangle|\alpha\rangle \rightarrow \left|\vec{k}\right\rangle|\beta\rangle$, the total energy before scattering equals the total energy after scattering:

$$E_0 + E_\alpha = (E_0 - E) + E_\beta , \tag{3.30}$$

$$E = E_\beta - E_\alpha \equiv E_{\alpha\beta} . \tag{3.31}$$

A spectrum of electron energy losses shows enhanced intensity when $E = E_{\alpha\beta}$. Owing to the Pauli principle, however, the state ψ_β must be initially empty for it to be allowed as a final state for electron 2. The EELS spectrum usually shows a jump in intensity, or "edge jump," when $E_{\alpha\beta} = E_{\mathrm{a}}$, where E_{a} corresponds to the lowest energy of an unoccupied state ψ_β. Enhanced intensity extends for $E > E_{\mathrm{a}}$, because other unoccupied states of higher energy are available to the atomic electron 2.

With actual wavefunctions for ψ_α and ψ_β, we could use (3.29) to calculate the strength of this inelastic scattering,[12] and the measured intensity of the electron energy-loss spectrum at the various energies $E_{\alpha\beta} > E_{\mathrm{a}}$. To do this, however, we must first relate the experimental conditions to the cross-section of (3.29). Specifically, we need to know how experimental detector angles, ϕ, select Δk at various E. This is the topic of the next subsection.

3.4.3 ‡ * Double-Differential Cross-Section, $\mathrm{d}^2\sigma_{\mathrm{in}}/\mathrm{d}\phi\,\mathrm{d}E$

In EELS, we measure the spectrum of energy losses from electrons in some range of Δk, set by the angle, β, of a collection aperture (see Fig. 3.1). To understand the intensity of core-loss spectra, we need to know how the inelastic scattering depends on both scattering angle, ϕ, and energy loss, E. This dependence of the intensity on ϕ and E is provided by a double-differential cross-section, $\mathrm{d}^2\sigma_{\mathrm{in}}/\mathrm{d}\phi\,\mathrm{d}E$. We start with the ϕ-dependence for fixed E. For small Δk we can approximate, as shown in Fig. 3.10b:

$$\Delta k^2 = k^2\phi^2 + \Delta k^2_{\mathrm{min}} . \tag{3.32}$$

[11] A correction factor at high energy losses accounts for how the outgoing flux of scattered electrons is reduced when the electron is slowed (cf. (1.15)), but we safely ignore this effect for energy losses of a few keV.

[12] There is a subtle deficiency of (3.28) and (3.29). The excitation of a core electron changes the electronic structure of the atom. It is not necessarily true that atomic wavefunctions are appropriate for ψ_α or ψ_β when a core hole is present. The atomic electrons change their positions somewhat is response to the core hole, so the second term in (3.25) may not be strictly zero by orthogonality.

The increment in solid angle covered by an increment in ϕ (making a ring centered about $\vec{k}_0$) is:

$$\mathrm{d}\Omega = 2\pi \sin\phi \, \mathrm{d}\phi \; . \tag{3.33}$$

By differentiating (3.32) (for fixed E, Δk_{min} is a constant):

$$\phi \, \mathrm{d}\phi = \frac{\Delta k}{k^2} \mathrm{d}\Delta k \; , \tag{3.34}$$

so for the small ϕ of interest:

$$\mathrm{d}\Omega = 2\pi \frac{\Delta k}{k^2} \mathrm{d}\Delta k \; . \tag{3.35}$$

Substituting (3.35) into (3.29), and re-defining $\vec{r}_2 \rightarrow \vec{r}$, provides:

$$\begin{aligned} \frac{\mathrm{d}\sigma_{\mathrm{in}}(\Delta k)}{\mathrm{d}\Delta k} &= \frac{\mathrm{d}\sigma_{\mathrm{in}}}{\mathrm{d}\Omega} \frac{\mathrm{d}\Omega}{\mathrm{d}\Delta k} \\ &= \frac{8\pi}{a_0^2 k^2 \Delta k^3} \left| \int_{-\infty}^{+\infty} \mathrm{e}^{-\mathrm{i}\vec{\Delta k}\cdot\vec{r}} \psi_\beta^*(\vec{r}) \psi_\alpha(\vec{r}) \, \mathrm{d}^3\vec{r} \right|^2 \; , \end{aligned} \tag{3.36}$$

where the right-hand side is averaged for all $\vec{\Delta k}$ of the detected electrons.

When ψ_β is a bound state of the atom, (3.36) can be used directly to obtain an EELS intensity at the energy corresponding to the transition $\alpha \rightarrow \beta$. In the more typical case, ψ_β lies in a continuum of states, such as free electron states when the atomic electron leaves the atom with considerable energy, or a band of unoccupied antibonding states for energies E that are close to the absorption edge energy, E_{a}. We then need to scale the result of (3.36) by the number of states in the energy interval of the continuum, which is $\rho(E)\mathrm{d}E$. Here $\rho(E)$ is the "density of unoccupied states" available to the atomic electron when it is excited. Accounting for the density of states of ψ_β gives the double-differential cross-section:

$$\frac{\mathrm{d}^2\sigma_{\mathrm{in}}(\Delta k,E)}{\mathrm{d}\Delta k \, \mathrm{d}E} = \frac{8\pi}{a_0^2 k^2 \Delta k^3} \, \rho(E) \left| \int_{-\infty}^{+\infty} \mathrm{e}^{-\mathrm{i}\vec{\Delta k}\cdot\vec{r}} \psi_\beta^*(\vec{r}) \psi_\alpha(\vec{r}) \, \mathrm{d}^3\vec{r} \right|^2 \; . \tag{3.37}$$

The convention is to rewrite (3.37) to isolate the scattering properties of the atom. This is done by defining the "generalized oscillator strength," GOS, or $G_{\alpha\beta}(\vec{\Delta k},E)$:

$$G_{\alpha\beta}(\vec{\Delta k},E) \equiv E_{\alpha\beta} \frac{2m_{\mathrm{e}}}{\hbar^2 \Delta k^2} \left| \int_{-\infty}^{+\infty} \mathrm{e}^{-\mathrm{i}\vec{\Delta k}\cdot\vec{r}} \psi_\beta^*(\vec{r}) \psi_\alpha(\vec{r}) \, \mathrm{d}^3\vec{r} \right|^2 \; . \tag{3.38}$$

Here $E_{\alpha\beta}$ is the difference between the energies of the states ψ_α and ψ_β. Using (3.38) in (3.37):[13]

$$\frac{\mathrm{d}^2\sigma_{\mathrm{in}}(\Delta k,E)}{\mathrm{d}\Delta k\,\mathrm{d}E} = \frac{2\pi\hbar^4}{a_0^2 m_\mathrm{e}^2 E_{\alpha\beta} T}\frac{1}{\Delta k}\rho(E)\,G_{\alpha\beta}(\Delta k,E)\ . \tag{3.41}$$

To make connection to experimental EELS spectra, we convert the Δk-dependence of (3.41) into a dependence on the scattering angle ϕ of Fig. 3.10. We do so by arranging (3.34) as a relationship between $\mathrm{d}\Delta k$ and $\mathrm{d}\phi$, and substituting into (3.41):

$$\frac{\mathrm{d}^2\sigma_{\mathrm{in}}(\Delta k,E)}{\mathrm{d}\phi\,\mathrm{d}E} = \frac{2\pi\hbar^4}{a_0^2 m_\mathrm{e}^2 E_{\alpha\beta} T}\frac{k_0^2\phi}{\Delta k^2}\rho(E)\,G_{\alpha\beta}(\Delta k,E)\ . \tag{3.42}$$

Figure 3.10b shows the definition of $\phi_E \equiv \Delta k_{\mathrm{min}}/k_0$ and the approximation:

$$\Delta k^2 \simeq k_0^2\left(\phi^2+\phi_E^2\right)\ , \tag{3.43}$$

from which we obtain a useful expression:

$$\frac{\mathrm{d}^2\sigma_{\mathrm{in}}(\phi,E)}{\mathrm{d}\phi\,\mathrm{d}E} = \frac{2\pi\hbar^4}{a_0^2 m_\mathrm{e}^2 E_{\alpha\beta} T}\frac{\phi}{\phi^2+\phi_E^2}\rho(E)\,G_{\alpha\beta}(\Delta k,E)\ . \tag{3.44}$$

From Newtonian mechanics we would expect ϕ_E, which is a ratio of k-vectors, to depend on the energies in the collision problem as $\sqrt{E/T}$. There is, however, a change in mass energy equivalent of the high-energy electron after scattering. This energy loss from the change in mass is significant, so the wavelength change is considerably smaller than the non-relativistic prediction. The result from relativistic kinematics is:

$$\phi_E = \frac{E}{2\gamma T} \simeq \frac{E}{2E_0}\ . \tag{3.45}$$

As an example, for the C K-edge at 284 eV in a 200 kV microscope $\phi_E = 0.7$ mrad.

3.4.4 * Scattering Angles and Energies – Quantitative

We revisit the angular dependence of the inelastic scattering. At lower energy losses (E slightly larger than E_a), and at smaller scattering angles, the main

[13] For accuracy, we have written (3.41) with the incident kinetic energy, T, that differs from the incident energy, E_0, as:

$$T \equiv \frac{1}{2}m_\mathrm{e}v^2 = \frac{E_0}{2}\frac{1+\gamma}{\gamma^2}\ , \tag{3.39}$$

owing to the relativistic correction:

$$\gamma \equiv \frac{1}{\sqrt{1-(v/c)^2}} = 1+\frac{E_0}{mc^2}\ , \tag{3.40}$$

($\gamma \approx 1.4$ for 200 keV electrons).

angular dependence in (3.44) is from the Lorentzian factor, $(\phi^2 + \phi_E^2)^{-1}$, peaked at $\phi = 0$, with ϕ_E (3.45) as the half-width of the angular distribution. (The factor ϕ in the numerator of (3.44) merely accounts for the larger radius of a ring at larger ϕ.)

We first compare this characteristic angle for inelastic scattering, ϕ_E, to the characteristic angle for elastic scattering, ϕ_0. The elastic angle ϕ_0 is associated with the atomic form factor, which is a measure of the size of the atom. For convenience we select r_0, the Bohr radius of the Thomas–Fermi atom as this size (**??**), and obtain ϕ_0 as:

$$\phi_0 = \frac{1}{k_0 r_0} \, . \tag{3.46}$$

Putting typical values into (3.45) and (3.46), we find that ϕ_E is generally a few tenths of a milliradian while ϕ_0 is a few tens of milliradians, i.e., $\phi_0 \approx 100\phi_E$. The inelastic scattering is concentrated into a much smaller range of angles about the forward beam than the elastic scattering, especially when $E \simeq E_a$. Section 3.4.1 discussed the other extreme case where $E \gg E_a$, and the collision kinematics are insensitive to the shape of the atom – recall that the intensity became bunched into angles characteristic of classical "billiard-ball" collisions.

The generalized oscillator strength, $G_{\alpha\beta}(\Delta k, E)$ of (3.38), helps complete the picture of how the inelastic intensity varies between these two extremes of $E \simeq E_a$ and $E \gg E_a$. The generalized oscillator strength, $G_{\alpha\beta}(\Delta k, E)$, is the probability of the transition $\alpha \rightarrow \beta$, normalized by a factor related to the energy and momentum transfer. Figure 3.11 shows the GOS on the two-dimensional space of $\{\ln(\phi), E\}$ in a plot known as a "Bethe surface." The individual curves in Fig. 3.11 show the angular dependence of the inelastic scattering for each energy loss above the carbon *K*-edge. Likewise, the energy dependence of the GOS may be obtained by taking sections through the Bethe surface at constant scattering angle. The Bethe ridge is marked on Fig. 3.11. Although distinct at large E, the Bethe peak is less well-defined at energy transfers closer to E_a (the C *K*-edge threshold at 0 eV in Fig. 3.11).

In EELS measurements, an entrance aperture having an acceptance semi-angle β is placed around the forward beam (Fig. 3.1). This aperture cuts off the scattering beyond a certain ϕ. The measured spectrum of intensity versus energy is therefore an integration of the scattering intensity over combinations of E and $\vec{\Delta k}$ that fall below this cutoff. At energies significantly above the edge, Fig. 3.11 shows that a substantial portion of the intensity is concentrated in the Bethe ridge at larger scattering angles. A relatively large objective aperture (> 10 mrad or so) is needed to include this intensity in the EELS spectrum. On the other hand, at energies just above the edge, a small aperture will collect most of the intensity. This small aperture may be useful for removing background intensity at large Δk that originates from tails of other elements with lower E_a.

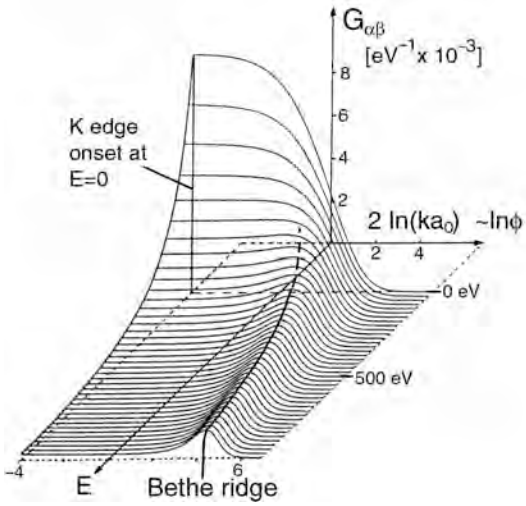


Figure 3.11: Bethe surface for K-shell ionization of C, calculated using a hydrogenic model. The GOS is zero for energy losses below the ionization threshold $E_K = E_{\alpha\beta}$, or $E < 0$. The horizontal coordinate increases with scattering angle. The Bethe ridge is most distinct at large E towards the front of the figure. After [5.5].

3.4.5 ‡ * Differential Cross-Section, $\mathrm{d}\sigma_{\mathrm{in}}/\mathrm{d}E$

Ignoring any truncation of the scattered inelastic intensity caused by the spectrometer entrance aperture, β, we integrate (3.44) over all possible scattering angles, ϕ, from 0 to π. This provides the total inelastic differential cross-section, $\mathrm{d}\,\sigma_{\mathrm{in},\alpha\beta}(E)/\mathrm{d}E$ for exciting an atomic electron from state $|\alpha\rangle$ to state $\left|\beta\right\rangle$:

$$\frac{\mathrm{d}\sigma_{\mathrm{in},\alpha\beta}(E)}{\mathrm{d}E} = \frac{2\pi\hbar^4}{a_0^2 m_{\mathrm{e}}^2 E_{\alpha\beta} T}\rho(E) G_{\alpha\beta}(\Delta k,E) \int_0^{\pi} \frac{\phi}{\phi^2+\phi_E^2}\mathrm{d}\phi \,. \tag{3.47}$$

Here we have ignored the ϕ-dependence of the GOS, $G_{\alpha\beta}(\Delta k,E)$. With the reasonable approximation that $\phi_E \ll \pi$, the integration of (3.47) gives:

$$\frac{\mathrm{d}\sigma_{\mathrm{in},\alpha\beta}(E)}{\mathrm{d}E} = \frac{\pi\hbar^4}{a_0^2 m_{\mathrm{e}}^2 E_{\alpha\beta} T}\rho(E) G_{\alpha\beta}(\Delta k,E) \ln\left(\frac{\pi^2}{\phi_E^2}\right) . \tag{3.48}$$

With (3.48) and (3.45) we obtain the inelastic differential cross-section:

$$\frac{\mathrm{d}\sigma_{\mathrm{in},\alpha\beta}(E)}{\mathrm{d}E} = \frac{2\pi\hbar^4}{a_0^2 m_{\mathrm{e}}^2 E_{\alpha\beta} T}\rho(E) G_{\alpha\beta}(\Delta k,E) \ln\left(\frac{2\pi\gamma T}{E}\right) . \tag{3.49}$$

Figure 3.12 shows a plot of the energy-differential cross-section for K-shell ionization of C ($E_{\mathrm{a}} = 284\,\mathrm{eV}$), calculated for different collection semiangles β using hydrogenic wavefunctions.[14] Logarithmic axes are used to illustrate the approximate behavior:

$$\frac{d\sigma_{\mathrm{in},\alpha\beta}(E)}{\mathrm{d}E} \propto E^{-r} , \tag{3.50}$$

[14] A hydrogenic atom uses the wavefunctions of a hydrogen atom, but with radial coordinates rescaled by a larger nuclear charge. There are no electron-electron interactions for a hydrogenic atom, but analytical expressions for the wavefunctions are available.

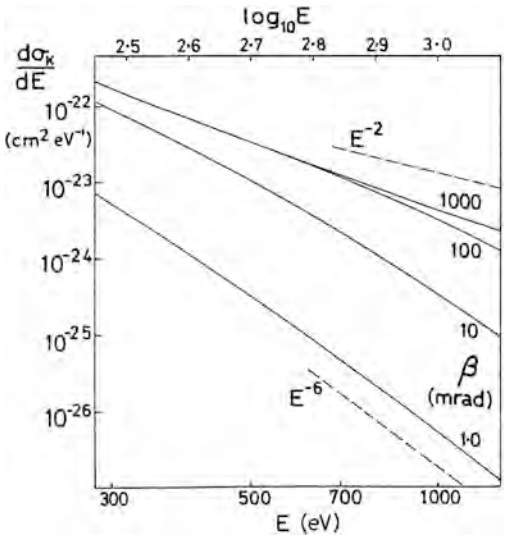


Figure 3.12: Energy-differential cross-section for K-shell ionization of C ($E_{\alpha\beta} = E_K = 284\,\text{eV}$) calculated for different collection semi-angles β. After [5.5].

where r is the downward slope in Fig. 3.12 and is constant over various ranges in energy loss. The value of r depends on the size of the collection aperture.

For large β, when most of the inner-shell scattering contributes to the energy loss spectrum, r is typically about 3 at the ionization edge, decreasing towards 2 with increasing energy loss. The asymptotic E^{-2} behavior occurs because for $E \gg E_a$, practically all of the scattering lies within the Bethe ridge. It approximates Rutherford scattering from a free electron (**??**), for which $\mathrm{d}\sigma_{\mathrm{in},\alpha\beta}(E)/\mathrm{d}E \propto \Delta k^{-4} \propto E^{-2}$.

For small β, r increases with increasing energy loss, the largest value (just over 6) corresponding to large E and small β. The breaks in slope in Fig. 3.12 correspond to the condition where E is large enough so that the Bethe ridge moves to angles outside the collection aperture. It is usually important to avoid this transition in experimental practice because it complicates the E-dependence of the measured intensity. It may be a good idea to calculate ϕ_r with (3.19), and use a collection angle, β, a few times larger than this, as mentioned in the context of (3.53).

3.4.6 ‡ Partial and Total Cross-Sections, σ_{in}

In quantitative elemental analysis, the inelastic intensity measured with an aperture angle β is integrated over an energy range of width δ beyond an absorption edge. Assuming a thin specimen with negligible multiple scattering, the integrated intensity above E_a, is:

$$I_a(E_a, \delta, \beta) = N\, I_0\, \sigma_{\mathrm{in,a}}(E_a, \delta, \beta)\ , \tag{3.51}$$

where N is the number of atoms per unit specimen area, and I_0 is the integrated zero-loss intensity. In (3.51), the "partial cross-section" $\sigma_{\mathrm{in,a}}(E_a, \delta, \beta)$ is

the integral of (3.44) over the range of collection angle and energy:

$$\sigma_{\text{in,a}}(E_\text{a}, \delta, \beta) = \int_0^\beta \int_{E_\text{a}}^{E_\text{a}+\delta} \frac{\mathrm{d}^2\sigma_{\text{in}}(\phi, E)}{\mathrm{d}\phi\, \mathrm{d}E}\, \mathrm{d}E\, \mathrm{d}\phi \ . \tag{3.52}$$

For numerical integration of $\mathrm{d}^2\sigma_{\text{in}}(\phi, E)/\mathrm{d}E\mathrm{d}\phi$, it is sometimes convenient to use the power-law behavior of (3.50).

Figure 3.13 shows the calculated angular dependence of K-shell partial cross-sections for the first-row (second-period) elements. The figure illustrates the dependence of the cross-sections on collection angle β, incident electron energy E_0, and ionization energy E_K, for constant integration width δ. The cross-sections saturate at large values of β, i.e., above the Bethe ridge angle, ϕ_r, owing to the fall-off in $G_{\alpha\beta}(\Delta k,E)$ at large Δk. The median scattering angle (for energy losses in the range E_a to $E_\text{a} + \delta$) corresponds to a partial cross-section equal to one-half of the saturation value, and is typically $5\left\langle\phi_E\right\rangle$, where:

$$\left\langle\phi_E\right\rangle = \frac{E_K + \delta/2}{2\gamma T} \ , \tag{3.53}$$

with $\gamma \equiv (1 - v^2/c^2)^{-1/2}$. Figure 3.13 shows that the saturation cross-sections decrease with increasing incident electron energy, although the low-angle values increase. This is because a small collection aperture accepts a greater fraction of the scattering when the incident energy is high and the scattering is more strongly forward-peaked.

For a very large range of energy integration δ, the partial cross-section becomes the "integral cross-section" $\sigma_{\text{in},K}(E_K, \beta)$ for inner shell scattering up to β and all permitted values of energy loss. By setting $\beta = \pi$, the integral cross-section becomes the "total cross-section" $\sigma_{\text{in},K}(E_K)$ for inelastic scattering from the K-shell. An approximate expression for $\sigma_{\text{in},K}(E_K)$ is the "Bethe asymptotic cross-section":

$$\sigma_{\text{in},K}(E_K) = 4\pi a_0^2 N_K b_K \frac{E_R^2}{T E_K}\, \ln\!\left(\frac{c_K T}{E_K}\right) \ , \tag{3.54}$$

where N_K is the number of electrons in the K-shell (2, but for the L and M shells this would be 8 and 18, respectively), $E_R \equiv \hbar^2(2m_\text{e}a_0^2)^{-1}$, $b_K \approx f_K/N_K$, $c_K \approx 4E_K/\langle E\rangle$, where $f_K \approx 2.1 - Z/27$ is the dipole oscillator strength for K-shell ionization and typically $\langle E\rangle \approx 1.5E_K$. A similar expression, useful for calculating the integral cross-section as a function of collection angle β, is given in Problem 5.10.

Computer programs are available to calculate differential cross-sections for K, L and M shell ionizations using various atomic models [5.5]. Figures 3.14a and 3.14b compare experimental N-K and Cr-L edges to those calculated with the widely-used SIGMAK and SIGMAL programs of Egerton [5.5]. These programs calculate inelastic cross-sections for individual, isolated atoms with

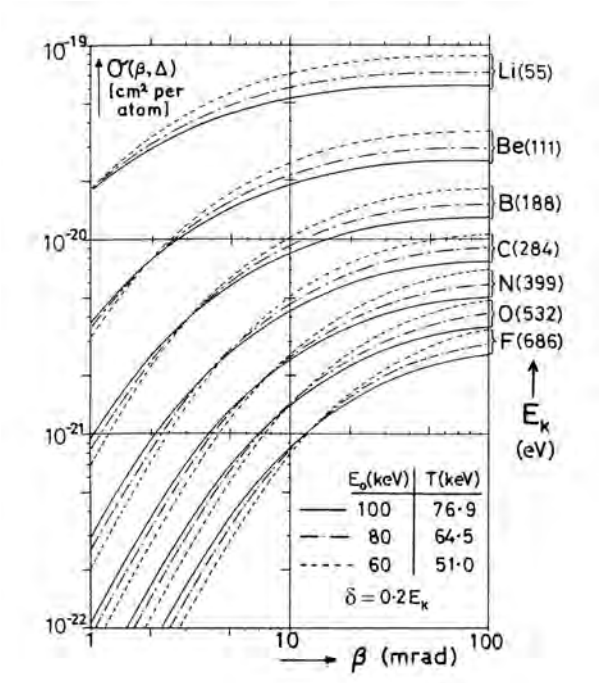


Figure 3.13: Partial cross-sections for *K*-shell ionization of second-period elements, calculated for an integration width δ equal to one-fifth of the edge energy, assuming hydrogenic wavefunctions and non-relativistic kinematics. After [5.5].

hydrogenic wavefunctions. As shown in Fig. 3.14, their integrated intensities are generally reliable, but they cannot provide information about local chemical effects in the near-edge region of the spectrum as discussed in Sect. 3.2.3. Figure 3.14b shows that the *L*-shell calculation with the SIGMAL program is good on the average, but it cannot model the white line peaks at the edge onsets. It does estimate their average intensity, however, based on the number of unoccupied *d*-states of the element.

*** Dipole Approximation and X-Ray Absorption Edges**

For energy losses near the absorption edge (small $E - E_a$) where most of the intensity occurs with small ϕ and small Δk, it is sometimes convenient to use the "dipole approximation" for the integral in (3.38). The dipole approximation is obtained by approximating $e^{-i\vec{\Delta k}\cdot\vec{r}} \simeq 1 - i\vec{\Delta k}\cdot\vec{r}$, and recognizing that the integration of the first term of 1, i.e., $\langle\beta|1|\alpha\rangle$, is zero owing to the orthogonality of ψ_α and ψ_β. The dipole approximation therefore amounts to replacing the factor $e^{-i\vec{\Delta k}\cdot\vec{r}}$ in the integrand of (3.38) with the simpler factor $-i\vec{\Delta k}\cdot\vec{r}$. Electric dipole radiation is the dominant transition process in EELS, but non-dipole transitions are observed at large Δk when higher order terms must be considered in the expansion $e^{-i\vec{\Delta k}\cdot\vec{r}} \simeq 1 - i\vec{\Delta k}\cdot\vec{r} - (\vec{\Delta k}\cdot\vec{r})^2/2 + ...$

For atomic transitions induced by x-rays, the GOS for inelastic x-ray scatter-

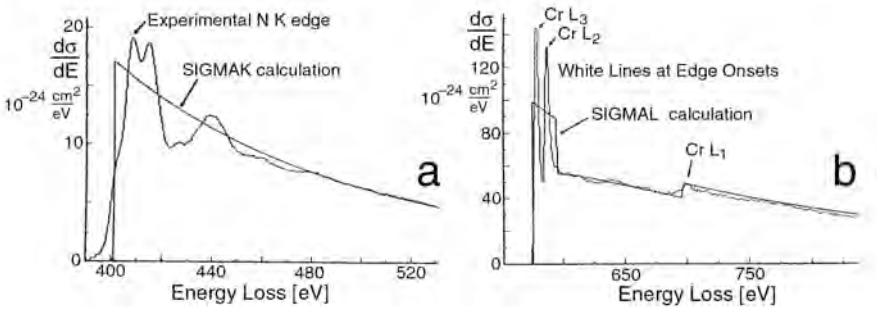


Figure 3.14: (**a**) Comparison of an experimental N *K*-edge and a hydrogenic fit to the edge using the SIGMAK program, and (**b**) comparison of an experimental Cr $L_{2,3}$-edge and a modified hydrogenic approximation to the edge using the SIGMAL program. After [5.6].

ing differs from (3.38) in that the exponential, $e^{-i\Delta\vec{k}\cdot\vec{r}}$, is replaced by the dipole operator, $e\vec{r}$. For small values of Δk, the integral in (3.38) is identical for both electron and photon inelastic scattering, and x-ray and electron absorption edges look very similar. Although the dipole approximation provides the same selection rules for the allowed atomic transitions for both EELS and for inelastic x-ray scattering, the E-dependence of EELS spectra is significantly different from that for inelastic scattering spectra of photons. This difference originates from the nature of electron scattering by a Coulomb potential, whose Fourier transform causes (3.28) to decrease strongly with Δk. Since large energy losses, E, are associated with the larger Δk, it becomes difficult to acquire EELS spectra at $E > 4$ keV. In practice, inelastic x-ray scattering, using a synchrotron radiation source for example, is performed for energies from 5–50 keV or so, whereas EELS experiments measure energy losses less than 5 keV.

3.4.7 Quantification of EELS Core Edges

The energies of absorption edges in EELS are quick and reliable indicators of the elements in a material, but quantifying the chemical composition requires more effort. The absorption edge must first be isolated from the background. The background originates primarily from other core edges, sometimes the high-energy tails of plasmon peaks, and occasionally from artifacts of the spectrometer. The background under an absorption edge is often modeled by a simple function such as AE^{-r}, where A and r are constants obtained by fitting the pre-edge background. This background function is extrapolated under the absorption edge and subtracted from the data. The edge jump is often complicated by fine structure effects of chemical bonding, which while of interest in their own right, may interfere with quantification with the use of (3.55) below.

After background removal, the intensity above an elemental absorption

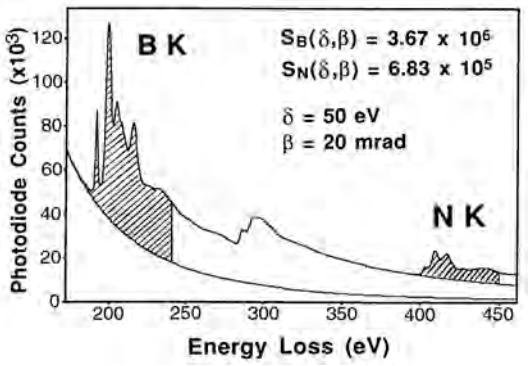


Figure 3.15: Chemical composition determination of BN. Background fits are shown, as are the 50 eV ranges of energy integration, δ. Using a collection aperture angle of 20 mrad, the partial ionization cross-sections were calculated with (3.52). Agreement was within 4 % of the known composition of B–50 at.% N. After [5.7].

edge is integrated over an energy range, δ. This range, δ, is determined by two competing requirements. The energy range should be close to the edge, because the intensity is strongest and the background correction is most accurate. On the other hand, chemical effects in the near-edge structure may not be understood, especially in an unknown material. Furthermore, some edges display "delayed maxima," where the maximum spectral intensity occurs at energies beyond the onset of the edge. Most quantification software uses an atomic wavefunction and a free electron wavefunction for the GOS of (3.38) (or $f_{\text{in}}(\vec{\Delta k})$ of (3.28)). Such software predicts smooth near-edge structures without solid-state or chemical effects (e.g., without the white lines of transition metals shown in Fig. 3.14b). Although it may seem desirable to start the range of energy integration above the near-edge region, it is common practice to start the energy integration at the edge, and ignore solid-state effects. Some solid-state effects average out when their peaks and valleys are averaged over a range in energy. An example of chemical analysis of BN is presented in Fig. 3.15. With the partial ionization cross-sections calculated from (3.52), the fractions of boron and nitrogen, c_{B} and c_{N}, were obtained from the integrated intensities, $I_{\text{BK}}(E_K, \delta, \beta)$ and $I_{\text{NK}}(E_K, \delta, \beta)$, which are the shaded areas in Fig. 3.15. Using 1s wavefunctions for ψ_α, and outgoing plane waves for ψ_β in (3.38), a software package (using 3.44 and (3.52)) was used to obtain a correction factor for the measured intensities:

$$\frac{c_{\text{B}}}{c_{\text{N}}} = \frac{I_{\text{BK}}(E_K, \delta, \beta)}{I_{\text{NK}}(E_K, \delta, \beta)} \frac{\sigma_{\text{in,NK}}(E_K, \delta, \beta)}{\sigma_{\text{in,BK}}(E_K, \delta, \beta)} \,. \tag{3.55}$$

For specimens of even modest thickness, core edges in EELS spectra are distorted by effects of multiple scattering. The problem occurs when a high-energy electron undergoes an inelastic scattering from a core excitation plus a second inelastic scattering by a plasmon excitation. (Two core excitations are relatively unlikely.) The probability of a plasmon excitation can be determined by examining the low-loss part of the EELS spectrum in the region of the plasmon (energies from 0 to 30 eV in Fig. 3.2). By treating this low loss region as an "instrument function," it is possible to use procedures analogous to those of Sect. 9.2 to deconvolute the effects of multiple scattering from the core loss

spectrum. If only one plasmon peak is visible in the EELS spectrum and its height is small relative to the zero-loss peak, the sample is probably thin enough so that multiple scattering can be neglected.

3.5 Energy-Filtered TEM Imaging (EFTEM)

3.5.1 Spectrum Imaging

Chemical mapping of the element distributions in samples is possible when EELS is performed in STEM mode (Fig. **??**). The electron beam is focused into a small probe and EELS spectra are acquired from a two-dimensional grid of points across the sample. Each "pixel" in the image can contain an entire EELS spectrum. The data set, called a "spectrum image," contains a wealth of information on chemical variations across the specimen. Analysis of this chemical information follows procedures of EELS spectrometry described in Sects. 4.2–4.4. Unfortunately, it may take hours to acquire a complete spectrum image. Exposing a sample to a high-intensity probe beam for a long time often causes problems with contamination and radiation damage. Over long times the specimen may also drift in position, blurring the image.

3.5.2 Energy Filters

Another method for chemical mapping has become popular, but it requires specialized instrumentation. Because an EELS spectrum consists of electrons that pass through the specimen and through the optical system of the TEM, the optical system can be used to make images of inelastically-scattered electrons. A conventional TEM uses all electrons that pass through the sample, but an instrument component known as an "energy filter" allows image formation with electrons that have undergone selected energy losses in the specimen. The technique of "energy-filtered TEM" (EFTEM), detects "chemical contrast" in specimens by adjusting an energy filter to pass electrons that have lost energy to core ionizations of selected elements.[15] These "energy-filtered images" (EFI) can reveal chemical contrast with sub-nanometer spatial resolution.

Alternatively, an energy filter can pass only zero-loss electrons, thereby removing all inelastic scattering from a conventional image or diffraction pattern. By using pure elastically-scattered electrons, chromatic aberration and the inelastic background are eliminated, so improved contrast is possible for thicker specimens, and more reliable interpretations of both images and diffraction patterns are possible. Energy filters will probably find a wider use for chemical analysis at the near-atomic scale, however, so this is the focus of the present discussion.

Figure 3.16a shows a modification of the magnetic prism EELS spectrometer shown previously in Fig. 3.1. Section 3.2.1 described how the magnetic sector operates as a focusing lens. Its optical analog is shown in Fig. 3.16b. By

[15] Plasmon images may also be useful.

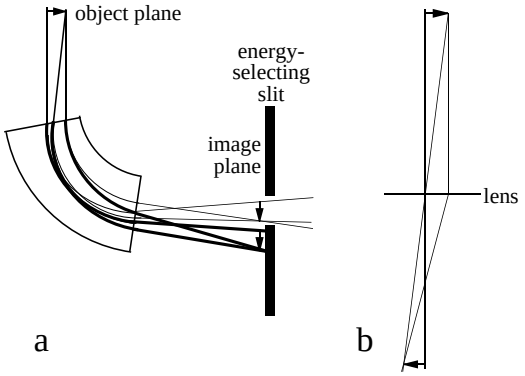


Figure 3.16: (**a**) An energy filter based on a magnetic sector like that in Fig. 3.1. (**b**) Optical analogy for monochromatic electrons.

comparing the ray paths in Figs. 3.16a and b, it is evident that the magnetic sector effectively bends the ray diagram of Fig. 3.16b. As in Fig. 3.1, the thin rays in Fig. 3.16a correspond to trajectories of electrons that have experienced energy losses, and the thick rays correspond to the zero-loss electrons. Notice how the thin and thick rays have the same pattern, but the thin rays are bent by an additional angle. The aperture at the image plane in Fig. 3.16a has been positioned to remove the zero-loss electrons. Additional lenses to the right of the energy-selecting slit in Fig. 3.16a (such as Q1–Q4 in the title image of Chapter 5) are used to project the energy-filtered image (EFI) onto an area detector such as a scintillator with a CCD camera.

Chapter 2 described how lenses produce dispersions of positional information (images) and angular information (diffraction patterns). The energy filter of Fig. 3.16 focuses an arrow on the object plane (a diffraction pattern or an image at the back focal plane of the final projector lens) onto an arrow at the image plane. In this case the image plane happens to be the same plane as the energy dispersion plane (as in Fig. 3.1). Selecting a narrow range of energy therefore limits the field of view of the image, so at the energy-selecting slit it may be appropriate to form a small image with a low magnification (or a small diffraction pattern) and magnify it with subsequent lenses.

Ensuring a good resolution of Δk, good spatial resolution, and high energy selectivity is a challenge for an energy filter, because it must also allow for measurements over wide ranges of these variables Δk, x, and E. Its performance is degraded by various types of aberrations. Recall (Sect. **??**) that spherical aberration mixed angular and spatial information – spherical aberration caused errors in position (x) at the focal plane for rays leaving the specimen at different angles (Δk). An energy filter has aberrations that mix energy, angular, and spatial information. For example, not all locations on an EFI may correspond to the same energy loss. In analogy to the apertures used to suppress problems with spherical aberration, energy filters require apertures to limit their field of view, their angular acceptance, or their acceptance window of energy losses.

To optimize imaging performance, it is best for electrons of only one energy to pass through the objective lens. Energy filters are integrated into the

microscope electronics to achieve this by allowing the energy filter system to control the high voltage at the electron gun. For making images with electrons that have undergone an energy loss of $e\Delta V$, the high voltage is increased by the amount ΔV. With such control over the high voltage system, the focus can then be adjusted only once, and focus will be maintained for images of electrons that have undergone different losses in the sample. The change in incident energy of the electrons then requires that the condenser lens currents are also tuned by the electronics of the energy filter system. This ensures a consistent intensity of illumination on the specimen, important for quantitative work.

Chemical analysis by EFTEM usually identifies elements by their core-loss spectra. For the thin specimens needed for EFTEM, however, only a small fraction of the incident electrons can ionize atoms in the sample. A central concern, therefore, is collecting as many inelastically-scattered electrons in as short a time as possible to minimize the effects of specimen drift, contamination, or radiation damage. This motivates the use of large apertures and wide energy windows, pushing the limits of filter performance. For example, a wider energy window gives more intensity, but a wider range of energies exacerbates chromatic aberration, leading to a loss of spatial resolution. Fortunately, in spite of these challenges, many types of chemical analyses with sub-nanometer resolution are now possible with EFTEM.

3.5.3 Chemical Mapping with Energy-Filtered Images

The ability to quantify chemical information at a near-atomic scale (including light elements such as C, O and N) makes EFTEM an important tool for materials characterization. Chemical information is usually obtained from the increase (or "jump") in EELS spectral intensity at an absorption edge. Unfortunately, absorption edges reside on large, sloping backgrounds from plasmons or absorption edges of other elements. Chemical mapmaking therefore requires EFIs at energy losses both above and below an absorption edge. Data processing is then needed to isolate the chemical information from the background. For example, the chemical contrast in an image acquired above the absorption edge, the "post-edge" image, can be better seen by subtracting or dividing by an image of the background obtained from "pre-edge" EFIs.[16] Two types of elemental maps are typically used:

- a "jump-ratio image," where a post-edge image is divided by a pre-edge background image,
- a "three-window image," where intensities in two pre-edge images are extrapolated to the energy of the post-edge image, and subtracted from it.

Jump-ratio images have the advantage that variations in specimen thickness and diffraction contrast are largely cancelled by the division. The three-window

[16]For example, the counts in each pixel of a background image could be subtracted from the counts in each corresponding pixel of the post-edge EFI.

image provides better elemental quantification, however. Unfortunately, three-window images are generally noisier than jump-ratio images owing to the background subtraction procedure, and may require longer times for measurement.

A first step in obtaining a chemical map with EFTEM is acquiring an EELS spectrum to locate the edges of interest, decide on placement of the energy windows, and determine the suitability of the specimen thickness. Another preparatory step is tilting the sample or incident beam to minimize the diffraction contrast in the bright-field TEM image. Since elastic scattering can be much stronger than inelastic scattering, diffraction contrast can dominate the appearance of energy-filtered images. To increase the relative amount of chemical contrast, the sample should be tilted away from strong diffraction conditions. Six EFIs of the same region are useful for making chemical maps. These images (with approximate energy windows) are:

- Unfiltered (bright field) image (all energies),
- Zero-loss image (5–10 eV),
- Low-loss (plasmon) image (20 eV),
- Pre-edge image 1 (10–20 eV),
- Pre-edge image 2 (10–20 eV),
- Post-edge image (10–20 eV).

For thin samples, the intensity of the inelastic spectrum is proportional to the sample thickness. Variations in sample thickness do not distort ratios of element concentrations if they are obtained as ratios of images from different elemental edges. When examining the chemical map of one element, however, it is important to know how the thickness varies across the sample. Since the unfiltered image contains both the elastically- and inelastically-scattered electrons, while the zero-loss image contains only elastically-scattered electrons, it is possible to divide the unfiltered image by the zero-loss image and take the logarithm of the result to obtain a thickness map in units of t/λ (as described in Sect. 3.3.2). This thickness map can be used to identify irregular sample surfaces, and permits conversion from areal densities of atoms to absolute concentrations. Similarly, three-window images can be corrected for the effects of thickness by dividing by the low-loss image.

For EFIs, the best specimen and microscope parameters are usually similar to those for EELS. For example, as in EELS, the specimen should be very thin, i.e., $t/\lambda < 0.5$, and ideally about half this thickness.[17] A rule-of-thumb is that the plasmon peak should be no more than one-fifth the height of the zero-loss peak. Similarly, a small collection angle (5-10 mrad) is preferable because it usually increases the signal/background ratio of an edge (see Problem 5.9). Elements

[17] Deconvolution of plural scattering is not possible in EFTEM because a full spectrum is not acquired.

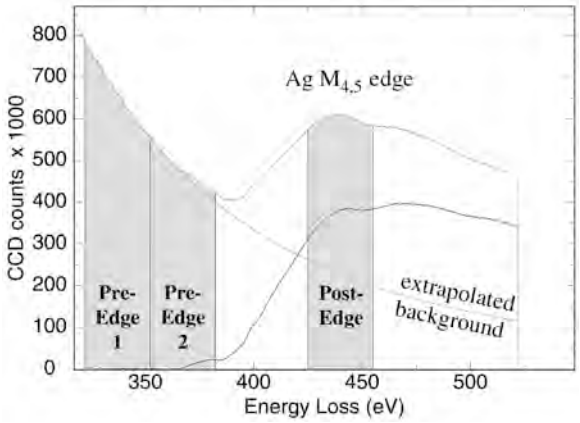


Figure 3.17: A portion of an EELS spectrum showing the Ag $M_{4,5}$ edge and the placement of the pre-edge 1, pre-edge 2, and post-edge energy windows used for the EFIs in Fig. 3.18. An extrapolated background and a background-stripped Ag $M_{4,5}$ edge are also shown. After [5.8].

of atomic number greater than 12 allow a choice of edge energy for elemental analysis. It is best to use major edges, and those with threshold energies from 100 to 1000 eV. At lower energies diffraction contrast and a steep background complicate quantification. Spatial resolution is also impaired for energy losses below 50 eV because ionization becomes delocalized, not necessarily occurring at the atoms nearest the high-energy electron. At energies above 1000 eV the intensity becomes inconveniently low. This often requires a wider window for energy selection, leading to problems with chromatic aberration. Effects of chromatic aberration can be suppressed by using smaller objective apertures, but this restricts the range of Δk, and hence the spatial resolution (cf. (**??**)).

3.5.4 Chemical Analysis with High Spatial Resolution

Figures 3.17 and 3.18 illustrate several aspects of making chemical maps by EFTEM imaging. The experiment confirmed that Ag enrichment was responsible for the conventional contrast of two planes at the interface between an Ω-phase precipitate and an Al-rich matrix. The Ag $M_{4,5}$ edge has a delayed maximum, peaking about 50 eV above the edge onset (Fig. 3.17). For best intensity, the post-edge window should incorporate the intensity maximum of the absorption edge, but this required the pre-edge 2 image and the post-edge image to be recorded with a fairly large energy separation. This is not optimal for EFTEM imaging. Better detectability and spatial resolution can be achieved when the windows labeled "pre-edge 2" and "post-edge" abut together at the onset of an abrupt absorption edge. Abrupt *K*-edges or intense, sharp white lines at the *L*-edges of transition metals offer this possibility. Additionally, spatial resolution can be improved by using energy windows narrower than the 30 eV windows used in this example, provided sufficient signal is available. Nevertheless, for Ag the energy windows were chosen as in Fig. 3.17.

Figure 3.18 shows a set of EFTEM images of an Ω precipitate plate in an Al-Cu-Mg-Ag alloy. To suppress diffraction contrast, the sample was tilted off the exact [011] zone-axis of the α-phase matrix, but the $\alpha|\Omega$ interfaces were

still parallel to the electron beam. Two dark lattice-fringes can be seen at the $\alpha|\Omega$ interfaces on both sides of the Ω plate in the zero-loss image (Figs. 3.18a,b). These fringes are the width of two {111} Al planes (0.46 nm). At each interface, both the Ag jump-ratio image and three-window image show high-intensity lines. The three-window image is noisier than the jump-ratio image, but provides a better estimate of the high enrichment of Ag. The background is featureless in both images, indicating that diffraction contrast and thickness effects are negligible. Two line profiles, acquired from the boxes shown in each image, are shown below the jump-ratio and three-window images. Both reveal segregation of Ag to the precipitate. These Ag layers are only 0.46 nm wide, showing the outstanding spatial resolution of the technique. The flatness of this internal interface was helpful for detecting chemical contrast at high spatial resolution.

3.6 Energy Dispersive X-Ray Spectrometry (EDS)

3.6.1 Electron Trajectories Through Materials

This section explains how high-energy electrons traverse thin TEM specimens and generate x-ray emissions from atoms. Some issues of instrumentation and artifacts are discussed. The following Sect. 3.7 describes procedures for quantitative analysis of x-ray spectra to obtain chemical concentrations in the sample. We first consider the trajectories of high-energy electrons through the sample because these paths determine where the x-rays come from. For thin samples, most electrons go straight through. Some electrons undergo high-angle deflections from elastic Rutherford scattering, so the electron beam broadens as it traverses the sample. Along an electron trajectory we need to know the probabilities, or at least the relative probabilities, that the electron will ionize atoms of different types – this was the topic of Sect. 3.4. After an atom is ionized, it is important to know the probability that it will emit an x-ray, and the probability that the x-ray will leave the sample and be counted by the detector.

Large-angle scatterings of electrons are primarily elastic in origin,[18] and occur when the high-energy electron passes close to an atomic nucleus. In these scatterings the shielding effects of the atomic electrons can be ignored, and the result is the Rutherford scattering cross-section, $\mathrm{d}\sigma_\mathrm{R}/\mathrm{d}\Omega$, of (**??**), written with $2\theta \equiv \phi$ as:

$$\frac{\mathrm{d}\sigma_\mathrm{R}}{\mathrm{d}\Omega} = \frac{Z^2\mathrm{e}^4}{16T^2}\frac{1}{\sin^4(\phi/2)} \ . \tag{3.56}$$

Equation (3.56) is also useful for understanding the occurrence of electron backscattering from the sample. "Backscattered electrons" are defined as elec-

[18]The discussion of ionization cross-sections in Sect. 3.4.4 showed that the electron energy-loss spectrum tends to be forward-peaked, especially at small energy losses, owing to the ϕ-dependence of (3.44).

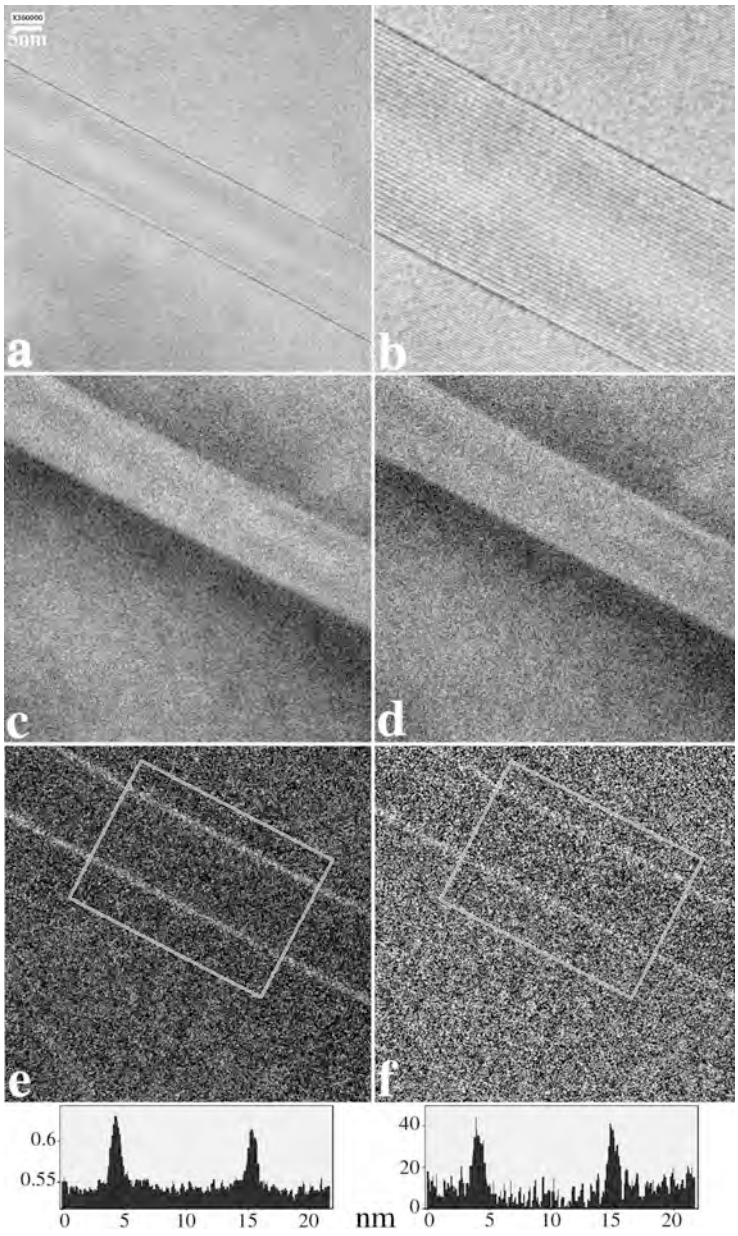


Figure 3.18: A set of EFIs from an edge-on Ω-phase plate-shaped precipitate in an Al-Cu-Ag-Mg alloy (α-phase), acquired with the sample tilted so the habit-plane interfaces were parallel to the electron beam, but the sample was not directly on the zone axis. Some energy windows are shown in Fig. 3.17. (**a**) zero-loss image, (**b**) enlargement of part **a**, (**c**) pre-edge 2 image, (**d**) post-edge image, (**e**) Ag $M_{4,5}$ jump-ratio image, and (**f**) Ag $M_{4,5}$ three-window chemical map. Line profiles corresponding to **e** and **f** should be aligned along the short edge of the boxes enclosing the $\alpha|\Omega$ interfaces. (β = 18 mrad). After [5.8].

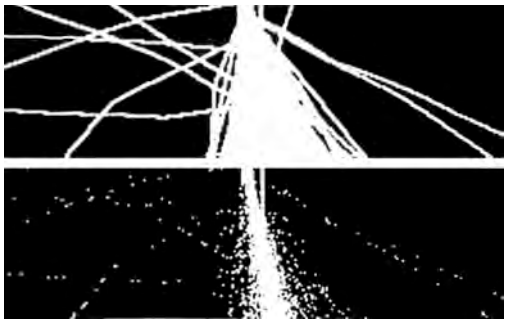


Figure 3.19: Monte Carlo simulations of electron trajectories (top), and assumed locations of x-ray emission (bottom). (In reality, most individual paths generate zero x-rays.) After [5.9].

trons scattered by angles so large that they reverse direction and go back out through the same surface they entered. Because of the T^{-2} dependence in (3.56), electron backscattering is relatively rare for electrons of several hundred keV passing through thin specimens.[19]

Electron trajectories are typically calculated individually with a Monte Carlo algorithm. The computer code allows for random occurrences of scattering events, consistent with a user-specified density of nuclei of charge Ze, electron energy, and Rutherford cross-section of (3.56). The electrons move along straight paths between these elastic collisions, which occur with randomness in the path length and scattering angle.

Along the straight paths between the Rutherford scattering events, the electron is assumed to lose energy at random to inelastic processes, both core excitations and plasmons. The core electron excitations are the ionization events that enable the subsequent emission of x-rays. Sections 3.4.1–3.4.6 described how the probability for ionizing an atom depends on the scattering angle, ϕ, and energy loss, E, of the incident electron. To calculate x-ray emission, we need to integrate over all ϕ and E to obtain a total cross-section for inelastic scattering by core electron ionization, σ_{in}. This was obtained as (3.49) and (3.54), which depend on the energy of the incident electron as: $1/\tau \ln(\tau)$, where $\tau \simeq 2\pi\gamma T/E_{\alpha\beta} \gg 1$. We therefore expect that as the high-energy electron loses energy in a thick specimen, the inelastic scattering events become more frequent, at least until its kinetic energy, T, becomes too small. Monte Carlo codes have been developed to model electron trajectories in solids with all the physical phenomena mentioned in this section, and typical results from a Monte Carlo simulation are presented in Fig. 3.19.

A schematic map of the electron trajectories in a thick bulk specimen is shown in Fig. 3.20a. The deep penetration and lateral broadening of high-energy electrons in bulk material causes the region of x-ray emission to be approximately 1 μm in diameter. This is a typical spatial resolution of an

[19] Backscattered electrons are much more common in scanning electron microscopy, which uses incident electrons of a few keV. Although these electrons tend to be multiply-scattered, backscattered electrons provide some chemical analysis capability to the SEM image; the factor of Z^2 in (3.56) causes the backscattered electron image (BEI) to be brighter in regions containing heavier elements.

electron microprobe, for example. Specimens used in TEM may be only tens of nm in thickness, however. A thin specimen, as depicted in Fig. 3.20b, lacks the bulk of the material where most broadening of the electron beam occurs. Spatial resolution in an analytical TEM is therefore much better than in an electron microprobe.[20] As a rule of thumb, the spatial resolution is significantly smaller than the width of the probe beam plus the thickness of the sample. Monte Carlo simulations that implement the model of elastic–inelastic scattering described in this section provide an approximation for the beam broadening, b, in [cm]:

$$b = 6.25 \times 10^5 \frac{Z}{E_0} \sqrt{\frac{\rho\, t^3}{A}}\ , \tag{3.57}$$

where A is the atomic weight of the element [g/mole], ρ is density [g cm^{-3}], t is thickness [cm], and E_0 is incident energy [eV].

"Secondary electron" emission is especially important in scanning electron microscopy (SEM). A secondary electron is an electron that is weakly bound to the sample and is ejected with a few (at most tens of) electron volts of energy. Since these electrons have little energy, they can traverse only short distances through a material (less than about 100 Å), and therefore originate from the near-surface region. The detected secondary electrons are highly sensitive to surface topography, being more likely to emerge from the peaks than the valleys of the surface drawn in Fig. 3.21. Secondary electron imaging (SEI) is the main technique of SEM, and can be performed in much the same way in the TEM. The instrument is operated in scanning mode with a secondary electron detector attached to the microscope column as illustrated in Fig. 3.22. The number of secondary electrons emitted per incident electron is defined as the "secondary electron yield," and can be either less than or greater than one. For incident electrons with energies less than 1 keV, the secondary electron yield increases with incident energy, but reaches a maximum (1–3 secondaries/incident electron) at an energy of order 1 keV. The yield is lower at higher energies because the incident electrons penetrate too deeply into the material, and the secondary electrons cannot escape.

3.6.2 Fluorescence Yield

After a core electron has been emitted from an atom, the ionized atom quickly decays from its excited state. This may occur by "radiative" or "non-radiative" processes, in which the atom emits either an x-ray or an Auger electron, respectively. Both processes (described in Sect. **??**) compete for the atomic decay. For a K-shell ionization, for example, the "fluorescence yield," $\overline{\omega}_K$, is defined as the fraction of decays that occur by the emission of a K-shell x-ray. A calculation of $\overline{\omega}_K$ requires knowledge of the relative rates of decay of the atom by Auger

[20] On the other hand, x-ray emission from the large volume on the left of Fig. 3.20 provides much greater intensity. This high intensity, and the higher current of the incident electron beam, allows electron microprobes to use wavelength dispersive x-ray spectrometers, which have low collection efficiency but excellent energy resolution.

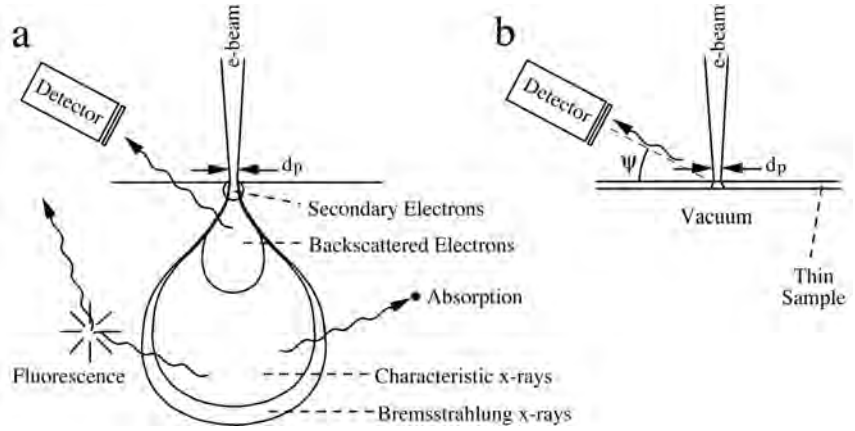


Figure 3.20: Differences in beam broadening in a bulk specimen (**a**), and a thin film (**b**). Part **a** shows regions of electron penetration, electron escape, and x-ray emission. For high-energy electrons, dimensions of regions of x-ray emission are typically a few microns, microns for backscattered electrons, tens of Å for secondary electrons. The larger dimensions do not exist for the thin specimen in **b**.

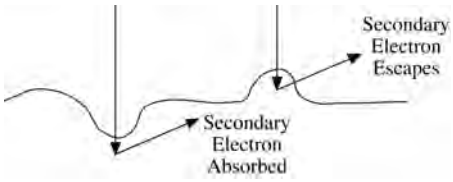


Figure 3.21: The escape probability of a secondary electron depends on the surface topography.

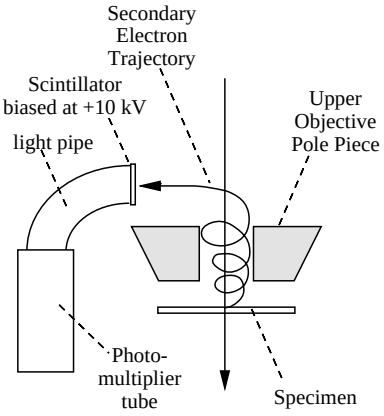


Figure 3.22: Everhart-Thornley detector and its configuration in a TEM. The secondary electron typically follows a spiral path along the magnetic field lines through the upper pole piece of the objective lens.

and by x-ray processes. The rate of x-ray emission is calculated for an electric dipole transition between the two atomic states of the atom, $|\alpha\rangle$ and $|\beta\rangle$. The x-ray emission rate is proportional to factors like $|\langle\alpha|e\vec{r}|\beta\rangle|^2$. The rate of Auger electron emission involves two electrons, and is calculated for a Coulomb interaction between them. The Auger electron emission rate is proportional to factors like $|\langle k|\langle\beta|e^2/(\vec{r}_1 - \vec{r}_2)|\alpha\rangle|\gamma\rangle|^2$, where $|\alpha\rangle$, $|\beta\rangle$, and $|\gamma\rangle$ are atomic states. The state $|k\rangle$ is that of a free electron with the Auger energy (the difference in energy between the states $|\alpha\rangle$ and $|\beta\rangle$, minus the binding energy of state $|\gamma\rangle$). The fluorescence yield is the ratio of the x-ray rate to the total rate, where the total rate is the sum of x-ray plus Auger rates. Empirically, for a K-shell emission, $\overline{\omega}_K$ depends approximately on atomic number, Z, as:

$$\overline{\omega}_K = \frac{Z^4}{10^6 + Z^4} \; . \tag{3.58}$$

Heavier elements tend to emit x-rays, and lighter elements tend to emit Auger electrons.[21] The K-fluorescence yield of the elements is presented in Fig. 3.23. The fluorescence yield increases rapidly with Z. On the other hand, the K-

[21] Approximately, the Auger emission probability is independent of Z, whereas the x-ray emission probability increases strongly with Z. Unfortunately, it is generally impractical to use a TEM for chemical analysis by measuring the energies of Auger electrons. Auger electrons lose a significant fraction of their energy through nanometer distances in a material. Auger energies characteristic of atomic transitions are obtained only for those few atoms at the very surface of a sample.

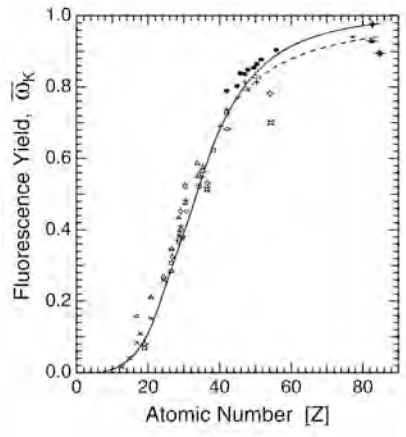


Figure 3.23: *K*-shell fluorescence yield of the elements. The difference, $1-\overline{\omega}_K$, is the yield of Auger electrons. After [5.10].

shell ionization cross-section decreases strongly with Z. This decrease in total ionization cross-section, denoted Q_K but equal to $\sigma_{\rm in}$ of Sect. 3.4.6, can be obtained from (3.54), or can be calculated with actual wavefunctions as in (3.38) (substituted into (3.44) and (3.52)). It also can be approximated as Z^{-4}. This Z-dependence of Q_K is opposite to that of $\overline{\omega}_K$ in (3.58). The probability of generating an x-ray depends on the product of $\overline{\omega}_K$ and Q_K, and this product turns out to be relatively constant in the energy range from 1–20 keV. The EDS method therefore has a well-balanced sensitivity to the elements from Na to Rh.

The detection of x-ray fluorescence radiation is the most widely-used technique for microchemical analysis in a TEM. A solid state detector, whose characteristics were described in Sect. **??** (Fig. **??**), is positioned near the specimen. The energy spectrum of the x-rays emitted from the specimen is acquired in a multichannel analyzer (Sect. **??**, Fig. **??**). A typical EDS spectrum, in this case from SiC, is presented in Fig. 3.24. The widths of the peaks are set by the energy resolution of the detector, and not by the atoms in the specimen. The detector characteristics also affect the intensities of the peaks. Notice that in spite of the equiatomic stoichiometry of SiC, the intensity of the C peak in the spectrum is much less than that of the Si peak.

Factors for converting x-ray intensities to elemental concentrations are a necessary part of quantitative EDS measurements. Fortunately, the thinness of a TEM sample simplifies the conversion process – for a particular sample geometry these conversion factors can often be regarded as a set of constants (Sect. 3.7.1). Simple constants of conversion are not appropriate when there is significant x-ray absorption and secondary x-ray fluorescence events in the sample, as illustrated in Fig. 3.20a, and this is typically the case for measurements on bulk samples in an electron beam microprobe or a scanning electron

Unfortunately, the vacuum in a TEM is not particularly good, and the sample is heated under the electron beam. The surfaces of a TEM specimen quickly become contaminated, even if they are not oxidized already.

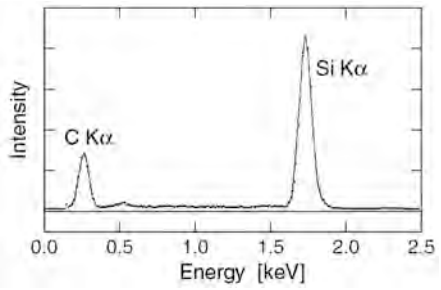


Figure 3.24: EDS spectrum from SiC, acquired with a Si[Li] detector having an ultra-thin window. After [5.11].

microscope. The thinness of the TEM sample minimizes problems with x-ray absorption and fluorescence (as illustrated in Fig. 3.20b), and quantitation is often straightforward.

3.6.3 EDS Instrumentation Considerations

Beam-Specimen-Detector Geometry

In general, the largest detector, located as closely as possible to the specimen, has the best geometry for efficient x-ray detection. Another important geometrical parameter is the take-off angle, ψ, which is the angle between the specimen surface and the line taken by the x-rays to the center of the detector (see Fig. 3.20b). From (6.42), sometimes known as Beer's law:

$$\frac{I}{I_0} = \mathrm{e}^{-(\mu/\rho)\rho x} \,, \tag{3.59}$$

where I/I_0 is the fraction of x-rays transmitted through a thickness, x, of a material with density ρ. Here μ/ρ is the mass absorption coefficient, which is tabulated as a function of Z and the energy of the x-ray (as in Appendix A.2). The likelihood of x-ray absorption in the sample depends on the length of the "escape path," or "absorption path," through the sample. The absorption path depends on two factors: 1) the depth of x-ray generation in the sample, t, and 2) the take-off angle, ψ (the larger ψ, the shorter the absorption path). Equation (3.59) becomes:

$$\frac{I}{I_0} = \mathrm{e}^{-(\mu/\rho)\rho t \csc\psi} \,. \tag{3.60}$$

This geometrical factor for x-ray escape probability is plotted in Fig. 3.25 for various realistic μt. A rule of thumb, perhaps consistent with Fig. 3.25, is that the x-ray emission increases from 0° to 30° and then levels off.[22] In most microscopes, samples are typically tilted about 30° toward the x-ray detector,

[22]Very thick samples may generate more x-rays with higher angles of tilt, since the angle of incidence between the electron beam and the sample surface affects the average depth of the interaction volume. The smaller this angle of incidence, the closer is the interaction volume to the surface, and the shorter the absorption path for emitted x-rays.

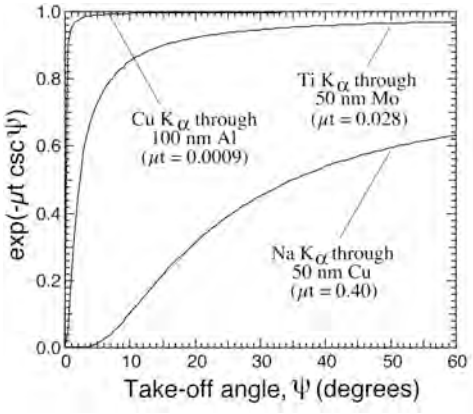


Figure 3.25: Fraction of x-rays that leave the sample, I/I_0, versus detector take-off angle, ψ, for various characteristic depths in the sample, μt, where t is the depth of the primary ionization. The full intensity is an integration of x-rays originating from all depths.

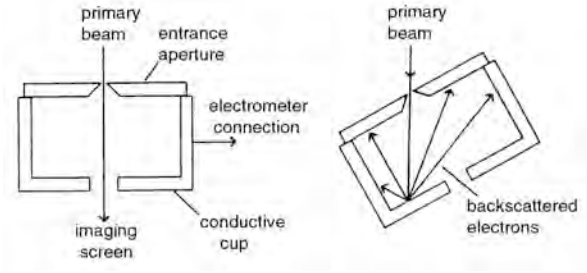


Figure 3.26: Faraday cage. Left: orientation for calibration of zero current and positioning the beam. Right: orientation for current measurement. After [5.9].

or the detector may be mounted at a high angle. For some horizontal detectors, the Si[Li] detector is tilted 20–30° toward the sample, so little specimen tilting is required.

Probe Diameter, Current and Convergence Angle

The probe diameter, $d_{\min}$, current, i, and convergence angle, α, all affect the x-ray emission process. Accurate quantitative work with EDS usually requires knowledge of these parameters. Techniques for determining them are described here. As suggested in Sect. 2.3.3 (and analyzed in Sect. 7.5.1), α can be measured directly from the diameter of the disks in the diffraction pattern. A straightforward method to measure the incident current, i, is with a Faraday cage (Fig. 3.26). For some TEM's, a Faraday cage is mounted on a removable specimen rod. Once calibrated against measurements with a Faraday cage, the beam current can be determined approximately from readings of the emission current meter on the instrument console. Another method of measuring the probe current is with an EELS spectrometer, by counting the total number of electrons in the spectrum acquired in a known time, and converting this to amperes.

A common way to measure the probe size, $d_{\min}$, is to record an image of the probe (preferably with a linear detector such as a charge-coupled device (CCD) camera) at high magnification by forming a focused image at the eucentric

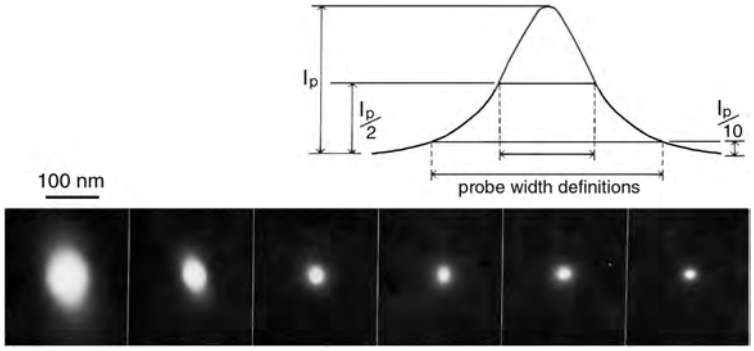


Figure 3.27: A series of images showing the electron intensity distribution on the TEM screen at 100,000 X magnification in a Philips EM400T with a LaB_6 filament for six different spot size settings. The FWHM probe sizes calculated from these images were 200, 100, 40, 20, 10 and 10 nm, respectively. After [5.9].

height, and then focusing the probe on the viewing screen using the second condenser lens, C2. If the probe is assumed to be Gaussian, the size can be arbitrarily defined as the full-width-at-half-maximum (FWHM) containing 75 % of the current and indicated as $I_p/2$ in Fig. 3.27, or the full-width-at-tenth-maximum (FWTM) indicated as $I_p/10$ in the figure. The FWTM definition is probably a better criterion since this part of the beam contains most of the current, and the tails associated with very small probes may be non-Gaussian and quite wide.

In the TEM, the probe size is determined by the current through the first condenser lens, C1 (often called the "spot size" control), and the convergence angle is determined primarily by the size of the C2 aperture (with an additional adjustment of the objective lens prefield, or condenser "minilenses," for some microscopes). The probe current on the specimen in a typical TEM can be varied over two orders of magnitude, depending on the probe size. If spatial resolution is not the main consideration, then a large probe size with a high current provides the best x-ray counting statistics. If high spatial resolution is needed, however, there is a trade-off between resolution and probe current. Theoretically, the probe size should be independent of the size of the C2 aperture. In practice, the C2 aperture affects the probe size because high convergence angles ($> 10^{-2}$ rad) are typically used to form very small probes, and this can lead to wide tails in the probe that are truncated by the aperture.

3.6.4 Artifacts in EDS Measurements

Ideally, when an electron probe is placed on a specimen for chemical analysis, the x-rays come only from the primary interaction of the beam and the specimen. Unfortunately, there are two other sources of spurious x-radiation in the TEM. One is x-ray generation in the illumination system of the microscope

from electron interactions with the column components. If not eliminated by shielding and design, these x-rays (and/or stray electrons) "rain down" on the specimen through the upper objective lens pole piece. X-rays are also generated by electrons or x-rays that scatter from the specimen, and subsequently strike the objective lens components or specimen holder. These spurious x-ray signals must be subtracted from the measured spectrum to perform quantitative x-ray analyses. There are different ways of recognizing and dealing with these effects, depending on the type of sample. We describe three typical situations below. Manufacturers of TEM instruments are well aware of these issues, and go to great lengths in the design of their microscope columns and holders to minimize them, but this does not mean that spurious x-rays cannot be a problem for some measurements.

In the case of an electropolished thin foil, where absorption and fluorescence can be neglected, the situation is straightforward. After measurements have been made on the specimen area of interest, the probe is placed just off the specimen in the hole and a spectrum is acquired under identical conditions. The x-ray signal from the hole, often called the "hole count," is subtracted from the specimen spectra to effectively remove any spurious signal generated by the microscope or holder. For most modern instruments the hole count from thin specimens is almost negligible compared to the specimen spectra, so corrections are small and quantification is straightforward. It must be remembered, however, that sending the electron beam through the hole is not the same as sending it through the specimen, from which electrons and x-rays are emitted, enabling secondary x-ray generation. The hole count is not exactly the same as the actual background, but hopefully close to it. For thin foils, high-energy x-rays striking the specimen from above generally pass right through it and do not produce any significant background.

At the other extreme, consider a typical sample prepared by focused ion-beam milling. Such samples comprise a thin membrane approximately 30 μm wide and extending 10 μm into the sample, surrounded by residual bulk material that is approximately 20 μm or more in thickness, supported on a C-shaped washer. In this case, there is ample opportunity for x-rays raining down on the sample, and for scattered electrons and x-rays generated by the sample, to strike bulk sample and support materials, generating many spurious x-rays. If this spurious x-radiation is coming from x-rays produced in the illumination system, the ratio of the $K\alpha$ x-rays to the $L\alpha$ x-rays for elements like Cu will be very high, because the high-energy x-rays will fluoresce the $K\alpha$ x-rays much more than the $L\alpha$ x-rays. As a rule, for electron excitation of x-rays from medium atomic number elements like Cu and Ni, the $L\alpha$ peaks are usually as high or higher than the $K\alpha$ peaks, so an abnormal ratio is an indication of x-ray fluorescence. It is always a good idea to acquire a hole count from such a specimen, but it is also important to realize that the signal generated from scattered electrons and x-rays in this type of specimen may be significantly greater when the probe is placed on the membrane than in the hole, so the hole count subtraction may not be so reliable as for thin-foil samples. It is often useful to place the probe on a C (carbon) layer at the edge of the specimen, or

on a coating layer (e.g., if Pt was used to preserve the surface during milling), and compare these spectra to the hole count and specimen spectra to make the quantification process as accurate as possible.

A common situation that falls in between these two extremes is small particles supported on a thin carbon film, which is supported on a metal grid (often Cu). Spurious x-rays from the grid can be eliminated by using Be or polymeric grids, which do not produce significant x-ray signals. If one stays away from the metal grid bars, x-rays produced from the stray radiation above the sample are generally not a problem, but spurious x-rays can still be produced from electron-specimen interactions. These are often seen as Cu peaks (assuming a Cu grid) in the x-ray spectra. It is common practice to place the probe on the C support just next to the particle being analyzed, and subtract this from the specimen spectrum. Again this is an approximation, since electron scattering and x-ray generation from the C film are not the same as from the particle. Thought and judgement are needed, as always, for the best experimental work.

3.7 Quantitative EDS

3.7.1 Thin-Film Approximation

Cliff–Lorimer Factors

Microchemical analysis by EDS begins by removing the background from the measured x-ray spectrum. The background originates primarily from bremsstrahlung radiation, which we found in Sect. **??** to depend weakly on energy, especially for thin specimens where multiple scatterings of the high-energy electron are unlikely. In the analysis of an EDS spectrum such as that in Fig. 3.28, a power series in E is typically used to model the background. With two or more adjustable parameters, the background can be modeled well. Subtracting the background from the spectrum provides peaks that can be either integrated numerically (with the procedure of Fig. **??**), or fit to analytical functions such as Gaussian functions. The peak areas can be treated individually, and this would be acceptable in the simple case of Fig. 3.28. When overlaps of peaks occur, it is preferable to work with sets of peaks (such as $K\alpha$, $K\beta$, L-series, etc.) with the energies and relative intensities expected for each element (including the sensitivity of the EDS spectrometer). Either method provides a set of peak intensities, $\{I_j\}$, where j denotes a particular chemical element. These $\{I_j\}$ are converted to a set of elemental concentrations, $\{c_j\}$, as described next.

In thin foil specimens, it is unlikely that an x-ray emitted from one atom will be absorbed by a second atom (cf., Fig. 3.20). Such double-scattering processes are neglected in the "thin-film approximation." This simplifies enormously the task of determining the $\{c_j\}$ from the $\{I_j\}$. In the thin-film approximation, the ratio of x-ray peak intensities from the elements A and B, I_A/I_B, is simply

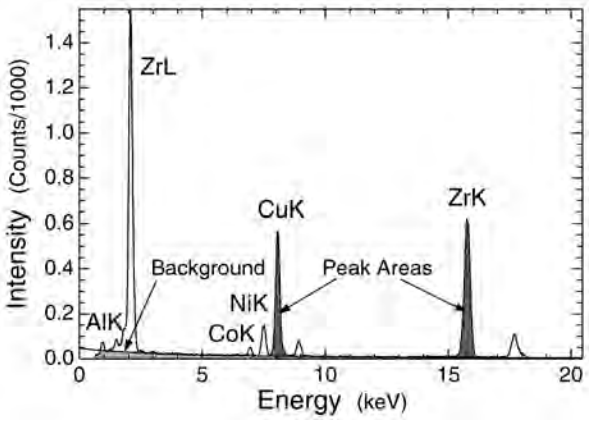


Figure 3.28: EDS spectrum of a Zr-based metallic glass, showing a fitted background and two peak areas above the background. After [5.12].

proportional to the corresponding weight-fraction ratio, c_A/c_B:

$$\frac{c_A}{c_B} = k_{AB}\frac{I_A}{I_B} , \tag{3.61}$$

where k_{AB} is a constant for a given accelerating voltage and a specific EDS spectrometer, and is independent of specimen thickness and composition. This constant k_{AB} is often called a "Cliff–Lorimer factor." It accounts for the efficiency of x-ray production for different accelerating voltages and the efficiency of the detector at the relevant x-ray energies.

A convenient feature of EDS is that for a large number of elements, the k_{AB} factor for their $K\alpha$ x-ray peaks is approximately 1. The ratio of the peak intensities (or even peak heights) therefore gives a good approximation of the sample composition, making for a simple, semi-quantitative EDS analysis. This approximation holds for elements from about Mg to Zn. Below or above this range of atomic numbers the k_{AB} factor gradually increases, but for elements of similar atomic number it is still reasonable to estimate their concentrations by comparing intensities of their $K\alpha$ peaks.

A normalization procedure:

$$\sum_j c_j = 1 , \tag{3.62}$$

is used to convert the ratios of the weight fractions to weight percentages (or, alternatively, atomic fractions to atomic percentages). That is, if k_{AB} for elements A and B in a binary system is known, quantification is based on the measured ratio of I_A and I_B (3.61), and using (3.62):

$$c_A + c_B = 1 . \tag{3.63}$$

For a ternary system with elements A, B and C, the following equations are

used:

$$\frac{c_C}{c_A} = k_{CA}\frac{I_C}{I_A} , \tag{3.64}$$

$$\frac{c_C}{c_B} = k_{CB}\frac{I_C}{I_B} , \tag{3.65}$$

$$c_A + c_B + c_C = 1 . \tag{3.66}$$

For a ternary alloy we have one more unknown, but one more independent peak ratio and another equation (3.65). In general, as we add more elements we can still use a set of linear equations like (3.64) and (3.65), plus (3.62) to complete the alloy chemistry.

The Cliff–Lorimer factors are mutually related. This is seen by dividing (3.65) by (3.64):

$$\frac{c_A}{c_C}\frac{c_C}{c_B} = \frac{k_{CB}}{k_{CA}}\frac{I_C}{I_B}\frac{I_A}{I_C} . \tag{3.67}$$

By the definition in (3.61), $k_{CA} = 1/k_{AC}$, so:

$$\frac{c_A}{c_B} = k_{AC}\,k_{CB}\frac{I_A}{I_B} . \tag{3.68}$$

Comparing (3.61) and (3.68), we obtain a general relationship between the Cliff–Lorimer factors:

$$k_{AB} = k_{AC}\,k_{CB} . \tag{3.69}$$

Cliff–Lorimer factors, or "k-factors," are often stored in a look-up table in the EDS software.

k-Factor Determination

Considerable effort is devoted to obtaining accurate Cliff–Lorimer factors, k_{AB}, since the accuracy of the EDS analysis depends on them. The k-factors are a combination of specimen and detector properties. Consider a k_{AB} coefficient for $K\alpha$ x-ray emission from elements A and B. The thin film approximation assumes both types of x-rays originate in the same region, and take direct paths through the specimen. We therefore expect the k_{AB} coefficient to be the ratio:

$$k_{AB} = \frac{A_A\overline{\omega}_B a_B Q_{KB}}{A_B\overline{\omega}_A a_A Q_{KA}}\,\mathrm{e}^{(\mu^A_{Be}-\mu^B_{Be})t} , \tag{3.70}$$

where A_i is the atomic weight of element i, (needed when the k_{AB} are for determining mass fractions), $\overline{\omega}_i$ is its fluorescence yield, a_i is its fraction of $K\alpha$ emission (for which $K\beta$ emission competes, but $a_i = 1$ for $Z < 19$), and μ^i_{Be} is the "effective" mass-absorption coefficient for the x-ray from element i and the detector window of effective thickness t (comprising, for example, the Be

window, the Si dead layer, and the Au conductive film). The Q_{Ki} are the K-shell ionization cross-sections (which could in principle be obtained from the total cross-section of (3.54), but better results are available).

There are essentially three ways to determine k_{AB}: 1) determine it experimentally using standards, 2) use values available in the literature, or 3) calculate it from first principles. The first method is the most reliable. Experimental k_{AB} values are determined for a specific microscope, detector and operating conditions. Use of calculated and/or experimental k_{AB} values from the literature is possible, but errors are expected owing to differences in the characteristics of the specimen, microscope, detector, and experimental geometry (including the tilt of the sample). The agreement between experimental and calculated k_{AB} values is typically good to 5 % for $Z > 14$, and for these elements it is often sufficient to calculate k_{AB} values for a given detector and accelerating voltage. The disagreement between theory and experiment for low Z may be due to a combination of inadequate theory, absorption of low-energy x-rays within the specimen, contamination on the detector window, or the loss of light elements during electron irradiation. For routine analysis it is common to use the k_{AB} values provided by the software of the EDS spectrometer system. Performing similar measurements on experimental standards of known composition can provide correction procedures to improve quantification for specimens of similar compositions.

3.7.2 * ZAF Correction

In an EDS spectrum, the x-ray peaks from different elements have intensities that depend on: 1) the path and energy of the high-energy electron passing through the sample, 2) the ionization cross-sections of the elements, 3) the fluorescence yields, and 4) the probabilities that emitted x-rays are seen by the detector. The thin film approximation collects all these effects into a constant factor for each type of characteristic x-ray. In the thin-sample limit, all peaks in an EDS spectrum increase in intensity with increased sample thickness, but the ratios of peak intensities remain unchanged. This permits the use of (3.61) for samples of all thickness. For thicker samples, however, the peak intensity ratios are altered. In TEM, the generation of characteristic x-rays from different elements is not altered by changes in the incident beam as it passes through a sample of moderate thickness. The thickness effects originate with the scattering of the characteristic x-rays by the different elements in the sample. As the samples become thicker and the x-ray exit paths through the sample become longer, these inelastic x-ray scattering processes involve a larger fraction of the x-rays, altering the ratios of peak intensities. Correction for these inter-element interactions is performed by considering the atomic number, Z, the absorption, A, and fluorescence, F, in procedures called "ZAF corrections."

Table 3.3: Limits to the thin foil approximation caused by absorption. Thickness limit is for a 3 % error in the k_{AB} factor [5.9]

material	thickness [nm]	absorbed x-ray(s)
Al–7 %Zn	94	Al $K\alpha$
NiAl	9	Al $K\alpha$
Ag_2Al	10	Al $K\alpha$, Ag $L\alpha$
FeS	50	S $K\alpha$
FeP	34	P $K\alpha$
Fe–5 %Ni	89	Ni $K\alpha$
CuAu	11	Cu $K\alpha$, Au $M\alpha$
MgO	25	Mg $K\alpha$, O $K\alpha$
Al_2O_3	14	Al $K\alpha$, O $K\alpha$
SiO_2	14	Si $K\alpha$, O $K\alpha$
SiC	3	Si $K\alpha$, C $K\alpha$

*** X-Ray Absorption Within the Specimen**

X-ray absorption follows Beer's Law (3.59). Since x-rays are generated throughout the foil thickness, evaluating the average absorption generally requires an integration of (3.60) over the sample thickness. Fortunately, for thin foils we can linearize the exponential in (3.60) as: $e^{-x} \simeq 1 - x$, and take the average depth of x-ray emission as $t/2$, where t is the sample thickness. In this case absorption alters the x-ray intensity ratio I_A/I_B from the ratio recorded for an infinitely-thin specimen, I_{A0}/I_{B0}:

$$\frac{I_A}{I_B} \simeq \frac{I_{A0}}{I_{B0}} \left(\frac{1 - \frac{\mu_A}{\rho_A} \frac{t}{2} \rho_A \csc \psi}{1 - \frac{\mu_B}{\rho_B} \frac{t}{2} \rho_B \csc \psi} \right), \tag{3.71}$$

$$\frac{I_A}{I_B} \simeq \frac{I_{A0}}{I_{B0}} \left(1 + (\mu_B - \mu_A) \frac{t}{2} \csc \psi \right). \tag{3.72}$$

Equation (3.72) shows the importance of the *difference* in absorption coefficients for the x-rays of elements A and B – if they have similar μ, the intensity ratios I_A/I_B are unaffected. Table 3.3 shows thicknesses at which the thin-film approximation is no longer valid due to absorption effects in specific materials.[23]

[23] To make an absorption correction, however, it is necessary to know the mean x-ray path length within the specimen, and this is difficult to determine from wedge-shaped or irregular specimens.

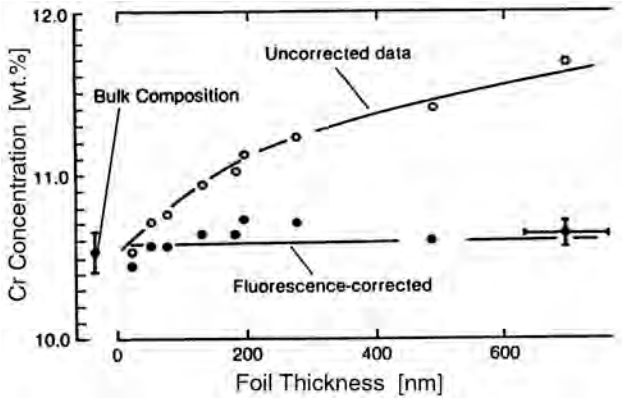


Figure 3.29: Experimental data showing an increase in the apparent Cr concentration with thickness in thick specimens of a Fe-10% Cr alloy, owing to fluorescence of Cr $K\alpha$ by Fe $K\alpha$ x-rays. After [5.9].

*** Characteristic Fluorescence Correction**

Characteristic x-rays from a heavier element can photoionize atoms of lighter elements, causing them to fluoresce. This enhances the number of x-rays detected from the light element, and suppresses the number from the heavier element. Fluorescence effects in thin foils are much weaker than in bulk samples (Fig. 3.20). Nevertheless when strong fluorescence does occur, e.g., Cr $K\alpha$ fluorescence under Fe $K\alpha$ radiation, quantitative microchemical analysis of TEM specimens may require a fluorescence correction (Fig. 3.29). Several fluorescence corrections for thin films have been developed, and a successful model [5.13] uses an enhancement factor, X_A, for the element, A, that undergoes fluorescence:

$$X_A = c_B \overline{\omega}_{KB} \frac{r_A - 1}{r_A} \frac{A_A}{A_B} \mu_{BA} \frac{U_B \ln U_B}{U_A \ln U_A} \frac{t}{2} \left| 0.923 - \ln(\mu_B t) \right| , \tag{3.73}$$

where U_i is the overvoltage ratio (ratio of incident electron energy to K-edge energy) for element i, A_i is its atomic weight, c_i is its weight fraction, r_i is its absorption-edge jump-ratio (the fractional countrate change across the EELS absorption edge), $\overline{\omega}_{Ki}$ is a fluorescence yield, and μ_{BA} is the mass absorption coefficient of element B in element A. For a sample in which the element B causes fluorescence of element A, the measured composition is corrected by:

$$\frac{c_B}{c_A} = k_{BA} \frac{I_B}{I_A} (1 + X_A) . \tag{3.74}$$

3.7.3 * Limits of Microanalysis

There are three quantifiable limits to microanalysis: 1) the absolute accuracy of quantification, 2) the minimum detectable mass (fraction), and 3) the spatial resolution. Of course there are other practical limits including contamination, insensitivity to low Z in EDS, and specimen preparation and geometry, but here

we discuss the first two quantifiable limits. Limits on spatial resolution were discussed in Sects. **??** and 3.6.3.

The accuracy of quantification is limited by the counting statistics of the x-ray spectra. For strong peaks on a weak background, the standard deviation, σ, is given by:

$$\sigma = \sqrt{N} \,, \tag{3.75}$$

where N is the number of counts in the peak after background subtraction.[24] Once the standard deviation is known, different confidence limits can be set for the value of N, i.e. 68 % confidence that N will lie in $N \pm \sigma$, 95 % in $N \pm 2\sigma$ and 99 % in $N \pm 3\sigma$. The value of 3σ, taken to be the 99 % confidence level in the value of I_A, is often used to estimate the error in the peak intensity:

$$\text{Error}(\%) = \pm \frac{3\,\sigma}{N} \times 100 = \pm \frac{3}{\sqrt{N}} \times 100 \,. \tag{3.76}$$

The larger is N, the lower the error in the analysis. For a 1 % accuracy at the 99 % confidence level, one needs 10^5 counts in a peak, or 10^4 counts for 1 % accuracy at the 68 % confidence level. The error in I_A/I_B is:

$$\text{Error}(\%) = \pm \left(\frac{3}{\sqrt{N_A}} + \frac{3}{\sqrt{N_B}} \right) \times 100 \,. \tag{3.77}$$

When using (3.61) for composition analysis, to the error of (3.77) we must add any error in k_{AB}, which is again the sum of the errors in I_A and I_B for the standard.

If Gaussian statistics are assumed, there is a simple statistical criterion that can be used to define the minimum mass fraction (MMF). A peak containing I_B counts from element B in a matrix of A is considered statistically real and not a random fluctuation in the background intensity, I_B^b, when:

$$I_B \geq 3\sqrt{2I_B^b} \,. \tag{3.78}$$

The MMF of B that can be detected in a binary material of elements A and B, c_B(MMF) in at.%, is obtained using (3.61) and (3.78):

$$c_B(\text{MMF}) = 3\sqrt{2I_B^b}\,\frac{c_A k_{BA}}{I_A - I_A^b} \,. \tag{3.79}$$

In practice, a MMF of approximately 0.1 wt.% can be obtained in EDS if enough counts are collected. Similarly, the minimum detectable mass (MDM) is predicted to be around 10^{-20} g for a range of Z from 10 to 40.

[24] When the background is a substantial fraction of the peak height, this argument is invalid for reasons stated in Problem 1.9. For weak peaks it is more accurate to use the background counts over the width of the peak, N_b, to obtain $\sigma = \sqrt{N_b}$ for use in (3.75).

These statistical analyses give the accuracy for quantification of a single measurement. In many cases, it is possible to obtain only a limited number of counts in a spectrum owing to factors such as beam damage or specimen drift. In such situations, it is possible to reduce the error in quantification (or at least assess it) by combining the results from n different measurements of the intensity ratio I_A/I_B. The total absolute error in I_A/I_B at a given confidence value is obtained using the Student-t distribution. In this approach, the error of the estimate E is given by:

$$E < \frac{t_{\alpha/2} S}{\sqrt{N}} , \tag{3.80}$$

where $t_{\alpha/2}$ is the Student-t value such that the normal curve area to its right equals $\alpha/2$ with a probability of $1 - \alpha$, S is the standard deviation for n measurements of the intensity N_i, given by:

$$S = \sqrt{\sum_{i=1}^{n} \frac{(N_i - \langle N_i \rangle)^2}{n-1}} , \tag{3.81}$$

which contain on average $\langle N_i \rangle$ counts. By increasing the number of measurements, one can reduce the error of measurement. In other words, if we estimate μ by means of a random sample size of n, we can assert with a probability of $1 - \alpha$ (where $1 - \alpha = 0.95$ for a 95 % confidence level for example) that the error in the measurement $E = \left| \langle N_i \rangle - \mu \right|$ is less than $(t_{\alpha/2} S)/\sqrt{n}$, at least for sufficiently large values of n. Equation (3.80) can also be rearranged and solved for n to determine the number of measurements n that must be taken to achieve a mean $\langle N_i \rangle$ which is in error by less than E.

Further Reading

The contents of the following are described in the Bibliography.

C. C. Ahn, Ed.: *Transmission Electron Energy Loss Spectrometry in Materials Science and the EELS Atlas* 2nd Ed. (Wiley-VCH, Weinheim, 2004).

C. C. Ahn and O. L. Krivanek: *EELS Atlas* (Gatan, Inc., Pleasanton, CA 1983).

M. M. Disko, C. C. Ahn and B. Fultz, Eds.: *Transmission Electron Energy Loss Spectroscopy in Materials Science* (Minerals, Metals & Materials Society, Warrendale, PA 1992).

R. F. Egerton: *Electron Energy-Loss Spectroscopy in the Electron Microscope* 2nd Ed. (Plenum Press, New York 1996).

J. J. Hren, J. I. Goldstein and D. C. Joy, Eds.: *Introduction to Analytical Electron Microscopy* (Plenum Press, New York 1979).

D. C. Joy, A. D. Romig, Jr. and J. I. Goldstein, Eds.: *Principles of Analytical Electron Microscopy* (Plenum Press, New York 1986).

H. Raether: *Excitations of Plasmons and Interband Transitions by Electrons* (Springer-Verlag, Berlin and New York 1980).

L. Reimer, Ed.: *Energy-Filtering Transmission Electron Microscopy* (Springer-Verlag, Berlin 1995).

L. Reimer: *Transmission Electron Microscopy: Physics of Image Formation and Microanalysis*, 4th Ed. (Springer-Verlag, New York 1997).

P. Schattschneider: *Fundamentals of Inelastic Electron Scattering* (Springer-Verlag, Vienna, New York 1986).

D. B. Williams: *Practical Analytical Electron Microscopy in Materials Science* (Philips Electron Instruments, Inc., Mahwah, NJ 1984).

D. B. Williams and C. B. Carter: *Transmission Electron Microscopy: A Textbook for Materials Science* (Plenum Press, New York 1996).

Problems

5.1

Use Moseley's laws (Sect. 1.2.2) to determine the $K\alpha$ and $K\beta$ x-ray energies for the elements Cu, Al, Mg, Zn, Be, Li and Ni. Which of these can be detected with a typical Be-window EDS detector? Explain.

5.2

(**a**) Using equations provided in the text, plot the mean free path, $\overline{\lambda}$, for inelastic scattering of electrons in Cu with an accelerating potential of 200 kV as a function of collection angles β ranging from 0.1 to 20 mrad.

(**b**) Using the same equations, plot the inelastic mean free path, $\overline{\lambda}$, as a function of the average atomic number of a material Z. Explain the graphs.

5.3

Use the K, L, α, β, etc., notation to name the X-rays generated by the following electron transitions:

(**a**) a hole in the K shell is filled by an electron from the L_{III} shell,
(**b**) a hole in the K shell is filled by an electron from the M_{II} shell,
(**c**) a hole in the K shell is filled by an electron from the O_{III} shell,
(**d**) a hole in the L_{III} shell is filled by an electron from the M_{I} shell,
(**e**) a hole in the L_{II} shell is filled by an electron from the N_{IV} shell,
(**f**) a hole in the L_{I} shell is filled by an electron from the O_{III} shell.

5.4

Find the thickness (in nm) of the Al sample in the EELS spectrum of Fig. 3.30. Assume 100 keV electrons.

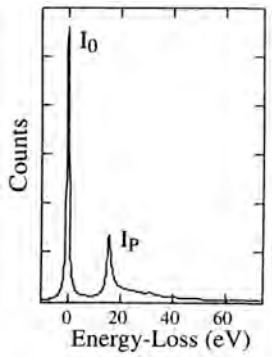


Figure 3.30: EELS low-loss spectrum of thin foil of Al for Problem 5.4. Assume the zero-loss peak is a triangle of 2.0 eV width at half-height.

5.5

A TEM specimen undergoes a type of radiation damage known as "knock-on damage" when a high-energy electron transfers enough energy to an atom to displace it from its crystallographic site. For a given electron energy, knock-on damage tends to be most severe for elements of low atomic number.

(**a**) In a direct ("head-on") collision between a high-energy electron and an atom, show that the energy transfer scales inversely with the atomic weight of the atom. (For simplicity, you may assume that the incident electron is scattered elastically by an angle of 180°.)

(**b**) If a Li atom requires 10 eV to leave its crystal site, calculate the threshold energy for an incident electron to induce knock-on damage. Do the same calculation for Al, Cu, and Au.

5.6

This problem presents two mathematical tricks for working with Dirac δ-functions. Calculations of electronic energies or scattering intensities often employ sums of Dirac δ-functions because δ-functions are handy for representing energy eigenvalues. For example, it is possible to write a distribution function for an energy spectrum, $n(E)$, as:

$$n(E) = \frac{1}{N} \sum_{\alpha}^{N} \delta(E - \varepsilon_\alpha) \; . \tag{3.82}$$

The idea behind this equation is that if N is large so there are numerous states (or transitions), each of energy ε_α, the discrete sum on the right becomes a continuum. To integrate the number of states up to some energy E', each δ-function on the right side contributes 1 to the sum when $E' > \varepsilon_\alpha$. A direct calculation of this type can be clumsy, however. Two expressions for δ-functions

can be of assistance in mathematical work:

$$\delta(E-\varepsilon_\alpha) = -\lim_{\delta\varepsilon\to 0}\frac{1}{\pi}\,\mathrm{Im}\left(\frac{1}{E+i\delta\varepsilon-\varepsilon_\alpha}\right), \tag{3.83}$$

$$\delta(E-\varepsilon_\alpha) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}(E-\varepsilon_\alpha)t}\,\mathrm{d}t\,. \tag{3.84}$$

Prove, or convince yourself, that these two equations are appropriate ways to represent a δ-function.

5.7

Suppose that samples containing mixtures of elements A,B and A,C were used to obtain the Cliff–Lorimer constants, k_{AB} and k_{AC}. Suppose the x-ray intensity of element C was less reliably correlated to an independent determination of composition of the samples, and the error in k_{AB} was estimated as 1 %, whereas the error in k_{AC} was estimated as 10 %.

(**a**) Estimate the error in a calculated k_{BC}.

(**b**) Estimate the error in the absolute concentration of B for a material of A and B elements, with nominal compositions 10 % B, 50 % B, and 90 % B.

5.8

The EDS data in Figs. 3.31b–d below were obtained from the Al-Ag precipitate shown in a. The number of counts in the peaks and background beneath them in d are:

Al $K\alpha$: 14,986 in peak, 1,969 in background
Ag $K\alpha$: 10,633 in peak, 1,401 in background

Given that $k_{\alpha\mathrm{AgAl}} = 2.3$ for the microscope conditions used, and that the thin-film approximation is valid, what is the composition of the precipitate?

5.9

In an EELS experiment, suppose we seek to measure the near-edge region from an element with a low concentration in the specimen. To optimize the ratio of edge jump to background, is it better to use a large or small acceptance angle for the EELS spectrometer? Why?

(*Hint*: Assume the angular dependence of the background is that of a single absorption edge that lies at lower energy than the edge of the dilute element.)

(*Further Hint*: Look at the Bethe surface.)

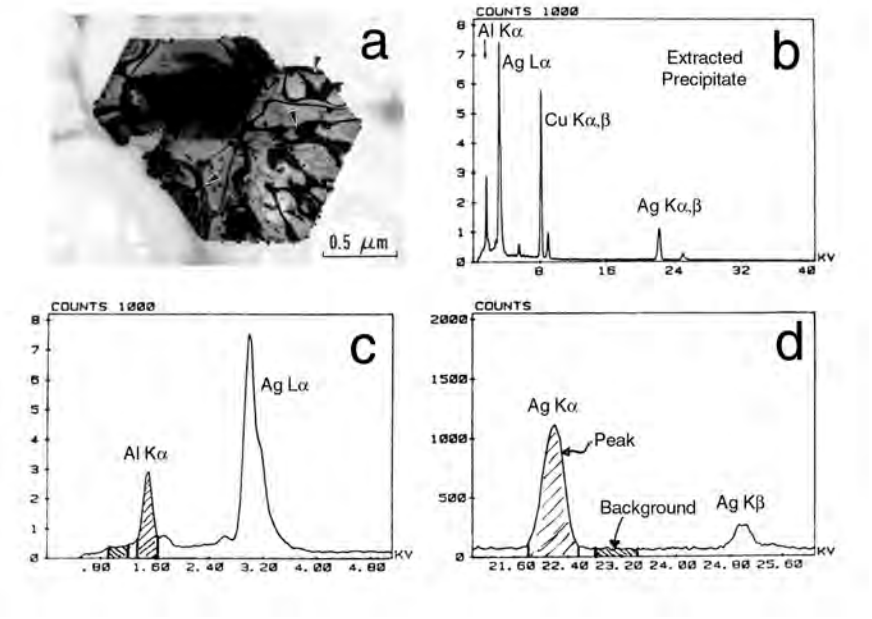


Figure 3.31: **a–d.** EDS data (b–d) from extracted precipitate on holey C support film in (**a**) (for Problem 5.8). After [5.14].

5.10

The integral inelastic cross-section for a K edge, $\sigma_{\text{in},K}(\beta)$, as a function of collection angle, β, can be predicted with moderate accuracy using the formula [5.5]:

$$\sigma_{\text{in},K}(\beta) = 4\pi a_0^2 \frac{E_R^2}{T\langle E\rangle} f_K \ln\left[1 + \left(\frac{\beta}{\langle\phi_E\rangle}\right)^2\right], \tag{3.85}$$

where $\langle\phi_E\rangle = \langle E\rangle (2\gamma T)^{-1}$, $T = m_e v^2/2$, $\gamma = (1 - v^2/c^2)^{-1/2}$, $f_K = 2.1 - Z/27$, $E_R = h^2/(2m_e a_0)^2 = 13.6\,\text{eV}$ is the Rydberg energy, $\langle E\rangle = 1.5E_K$, E_K is the K-edge energy, $a_0 = 52.92 \times 10^{-12}$ m is the Bohr radius, m_e is the rest mass of the electron, v is the electron velocity, c is the speed of light, and Z is the atomic number. Using this equation, plot $\sigma_{\text{in},K}(\beta)$ versus β for carbon with a K-edge energy of 284 eV, for collection angles β ranging from 0.1 to 20 mrad, assuming an accelerating potential of 200 kV.

5.11

Consider the probability of inelastic scattering, p_i, and elastic scattering, p_e, through a thin layer of material. We set $p_e + p_i = p$, where p is the total

probability of scattering from the incident beam.

(**a**) For thin samples of n layers, show that the x-ray mass absorption factor, $\mu = (n/x)p$, where x is the thickness of one layer.

Subsequent layers have the same probabilities, so for n thin layers we expect:

$$(p_e + p_i)^n = p^n \ . \tag{3.86}$$

(**b**) For thin samples, show that the ratio of double inelastic scattering, p_{2i}, to single inelastic scattering, p_i, is: $p_{2i}/p_i = (n/2)p_i$.

(*Hint*: Perform a binomial expansion of (6.54) and consider the physical meaning of the individual terms.)

5.12

When a hole in the core shell of an atom decays by an Auger process, one electron falls into the core hole and a second electron carries energy from the atom. A proper treatment of the Auger effect accounts for the indistinguishability of the two electrons. For electrons of the same spin, this includes antisymmetrizing the two-electron wavefunction of the initial state:

$$\psi_{\alpha\gamma} = \frac{1}{\sqrt{2}}\Big[\psi_\alpha(\vec{r}_1)\psi_\gamma(\vec{r}_2) - \psi_\gamma(\vec{r}_1)\psi_\alpha(\vec{r}_2)\Big] \ . \tag{3.87}$$

(**a**) Write integral expression(s) for the matrix element, $\langle\beta k|H'|\alpha\gamma\rangle$, with $H' = e/(|\vec{r}_1 - \vec{r}_2|)$, using (3.87) for $\psi_{\alpha\gamma}$ and a similar expression for $\psi^*_{\beta k}$.

(**b**) Show that the rate of the Auger transition,

$$\Gamma = \frac{2\pi}{\hbar} \int \psi^*_{\beta k} \frac{e}{|\vec{r}_1 - \vec{r}_2|} \psi_{\alpha\gamma} \, \mathrm{d}^3\vec{r}_1 \, \mathrm{d}^3\vec{r}_2 \ , \tag{3.88}$$

involves the difference of two matrix elements, one for the transition $|\alpha\gamma\rangle \rightarrow |\beta k\rangle$ and the other for the "exchange transition," $|\alpha\gamma\rangle \rightarrow |k\beta\rangle$.

(**c**) The inelastic scattering of a high-energy electron by a core electron transition involves two electrons, the perturbation $H' = e/(|\vec{r}_1 - \vec{r}_2|)$, and the same calculation as in a and b (compare (3.24) and (3.88)). Is the exchange transition important for the scattering of a high-energy electron? Why?

Chapter 4

General Formulation of Thermal Neutron Scattering

A more general treatment of nuclear scattering is possible than was presented in the previous Sect. 2.3. It starts with Fermi's golden rule, and avoids the explicit use of wavefunctions for the scatterer (i.e., it does not require phonon solutions like $\vec{u}_{l,\kappa}(\vec{q},t)$ of (2.49)).

4.0.1 Fermi's Golden Rule

Fermi's golden rule gives the transition rate from an initial state to a final state, W, at time t':

$$W(t') = \frac{2\pi}{\hbar}|\langle\Psi_{\rm f}(\vec{r},t')|V(\vec{r},t')|\Psi_{\rm i}(\vec{r},t')\rangle|^2 \ . \tag{4.1}$$

For nuclear scattering of a neutron, the states $|\Psi\rangle$ include coordinates for the neutron and coordinates for the crystal. The nuclear forces of interaction are very short range, compared to the atom motions in the crystal, so the potential in (4.1) is $V_{\rm nuc}(\vec{r},t')$, which involves the coordinates of the neutron and nuclei only. It includes the crystal only insofar as the positions of the nuclei alter the location of the δ-functions of the Fermi pseudopotential of (2.40). Moving the atoms will connect the neutron and crystal coordinates through the conservation of momentum and energy only. The electronic interactions between the atoms are not affected by the neutron, however, and the forces between the nucleus and the neutron and the nucleus are not affected by the atom positions. We can therefore separate the state $|\Psi_{\rm i}(\vec{r},t)\rangle$ into a lattice part $|\lambda_{\rm i}\rangle$ and a neutron part $|\vec{k}_{\rm i}\rangle$:

$$|\Psi_{\rm i}(\vec{r},t)\rangle = |\lambda_{\rm i}(\vec{r}_{\rm nu},t)\rangle\,|\vec{k}_{\rm i}(\vec{r}_{\rm ne},t)\rangle \ , \tag{4.2}$$

where the independent coordinates $\vec{r}_{\rm nu}$ and $\vec{r}_{\rm ne}$ refer to the postions of the nucleus and neutron. We have assumed the neutron states are plane-wave

states characterized by the wavevector $\vec{k}_\mathrm{f}$ (as in (1.49) and (1.68), but here $k_\mathrm{i} \neq k_\mathrm{f}$). In practice, the $\langle||\rangle$ in (4.1) denotes an integration over all positional coordinates at the instant t' when W is evaluated. To get the total probability of the transition, $P_{\mathrm{i}\to\mathrm{f}}$, we integrate this over all times when the two states interact in the presence of the perturbing potential $V(\vec{r},t)$:

$$\begin{aligned} P_{\mathrm{i}\to\mathrm{f}} &= \frac{2\pi}{\hbar}\int \langle\lambda_\mathrm{i}(\vec{r},t)|\,\langle\vec{k}_\mathrm{i}(\vec{r},t)|V^*(\vec{r},t)|\lambda_\mathrm{f}(\vec{r},t)\rangle\,|\vec{k}_\mathrm{f}(\vec{r},t)\rangle \\ &\qquad \langle\lambda_\mathrm{f}(\vec{r},t)|\,\langle\vec{k}_\mathrm{f}(\vec{r},t)|V(\vec{r},t)|\lambda_\mathrm{i}(\vec{r},t)\rangle\,|\vec{k}_\mathrm{i}(\vec{r},t)\rangle\,\mathrm{d}t\,, \end{aligned} \tag{4.3}$$

where we have substituted (4.2) into (4.1), written out the $||^2$, and integrated over all times.

Recall that the time evolution of the state of the scatterer is:

$$|\psi(t)\rangle = \mathrm{e}^{-\mathrm{i}\underline{H}t/\hbar}|\psi(t=0)\rangle\,, \tag{4.4}$$

$$\langle\psi(t)| = \langle\psi(t=0)|\mathrm{e}^{+\mathrm{i}\underline{H}t/\hbar}\,, \tag{4.5}$$

We assume that the states of the crystal $\{|\lambda_\mathrm{j}\rangle\}$ are eigenstates with specific energies $\{\varepsilon_\mathrm{j}\}$, so we can simplify (4.4) and (4.5) using $\varepsilon_\mathrm{i} = \hbar\omega_\mathrm{i}$ as:

$$|\lambda_\mathrm{i}(t)\rangle = \mathrm{e}^{-\mathrm{i}\omega_\mathrm{i}t}|\lambda_\mathrm{i}(t=0)\rangle\,, \tag{4.6}$$

$$\langle\lambda_\mathrm{i}(t)| = \langle\lambda_\mathrm{i}(t=0)|\mathrm{e}^{+\mathrm{i}\omega_\mathrm{i}t}\,, \tag{4.7}$$

The general formulation makes use of these expressions because they allow us to work with states of the crystal at $t = 0$ such as $|\lambda_1(0)\rangle$ and $|\lambda_2(0)\rangle$, which are constant and can be pulled out of any integration over time. Similarly for the position evolution of plane-wave states of the neutron $\{|\vec{k}\rangle\}$, which have constant momentum:[1]

$$|\vec{k}_\mathrm{i}(\vec{r})\rangle = \mathrm{e}^{-\mathrm{i}\vec{k}_\mathrm{i}\cdot\underline{\vec{r}}}|\vec{k}_\mathrm{i}(\vec{r}=0)\rangle\,, \tag{4.8}$$

$$\langle\vec{k}_\mathrm{i}(\vec{r})| = \langle\vec{k}_\mathrm{i}(\vec{r}=0)|\mathrm{e}^{+\mathrm{i}\vec{k}_\mathrm{i}\cdot\underline{\vec{r}}}\,. \tag{4.9}$$

Returning to the evaluation of (4.3), into it we substitute (4.6), (4.7) and (4.8), (4.9):

$$\begin{aligned} P_{\mathrm{i}\to\mathrm{f}} &= \frac{2\pi}{\hbar}\int \langle\lambda_\mathrm{i}(0)|\mathrm{e}^{\mathrm{i}\omega_\mathrm{i}t}\,\langle\vec{k}_\mathrm{i}(0)|\mathrm{e}^{\mathrm{i}\vec{k}_\mathrm{i}\cdot\underline{\vec{r}}}\,V^*(\vec{r},t)\,\mathrm{e}^{-\mathrm{i}\vec{k}_\mathrm{f}\cdot\underline{\vec{r}}}|\vec{k}_\mathrm{f}(0)\rangle\,\mathrm{e}^{-\mathrm{i}\omega_\mathrm{f}t}|\lambda_\mathrm{f}(0)\rangle \\ &\qquad \langle\lambda_\mathrm{f}(0)|\mathrm{e}^{\mathrm{i}\omega_\mathrm{f}t}\,\langle\vec{k}_\mathrm{f}(0)|\mathrm{e}^{\mathrm{i}\vec{k}_\mathrm{f}\cdot\underline{\vec{r}}}\,V(\vec{r},t)\,\mathrm{e}^{-\mathrm{i}\vec{k}_\mathrm{i}\cdot\underline{\vec{r}}}|\vec{k}_\mathrm{i}(0)\,\rangle\mathrm{e}^{-\mathrm{i}\omega_\mathrm{i}t}|\lambda_\mathrm{i}(0)\rangle\,\mathrm{d}t\,, \end{aligned} \tag{4.10}$$

where the operators $\underline{\vec{r}}$ refer to the neutron coordinates, and the (0) refer to $t = 0$ and $\vec{r} = 0$ (although we are not concerned about the time-dependence of the neutron wavefunction or the position-dependence of the crystal excitation).

The next step in simplifying (4.10) is to substitute the scattering potential for V. As before, we use the sum of Fermi pseudopotentials at all crystal sites

[1] Plane waves prove their convenience here. For other states we would have to integrate over k-space.

(i.e., (2.41)). These δ-functions are most convenient in selecting the $\vec{r}$ where the neutron sees the nuclei, because the integrations over spatial coordinates become trivial — we delete the integral over position, and in each matrix element we substitute the nuclear positions with $\vec{R}_j$ and $\vec{R}_k$, and the times t_j and t_k. We sum over all nuclei in the crystal:

$$P_{\mathrm{i}\to\mathrm{f}} = \frac{2\pi}{\hbar}\int \sum_j \langle\lambda_\mathrm{i}(0)|\mathrm{e}^{\mathrm{i}\omega_\mathrm{i}t_j}\,\langle\vec{k}_\mathrm{i}(0)|\mathrm{e}^{\mathrm{i}\vec{k}_\mathrm{i}\cdot\underline{\vec{R}}_j}\,b_j^*\,\mathrm{e}^{-\mathrm{i}\vec{k}_\mathrm{f}\cdot\underline{\vec{R}}_j}|\vec{k}_\mathrm{f}(0)\rangle\,\mathrm{e}^{-\mathrm{i}\omega_\mathrm{f}t_j}|\lambda_\mathrm{f}(0)\rangle$$
$$\sum_k \langle\lambda_\mathrm{f}(0)|\mathrm{e}^{\mathrm{i}\omega_\mathrm{f}t_k}\,\langle\vec{k}_\mathrm{f}(0)|\mathrm{e}^{\mathrm{i}\vec{k}_\mathrm{f}\cdot\underline{\vec{R}}_k}\,b_k\,\mathrm{e}^{-\mathrm{i}\vec{k}_\mathrm{i}\cdot\underline{\vec{R}}_k}|\vec{k}_\mathrm{i}(0)\rangle\mathrm{e}^{-\mathrm{i}\omega_\mathrm{i}t_k}|\lambda_\mathrm{i}(0)\rangle\,\mathrm{d}t\,. \tag{4.11}$$

The notation (0) refers to both $t = 0$ and $\vec{r} = 0$, since we made use of the relations (4.6)–(4.9). At $t = 0$), $\vec{r} = 0$, the phase factors, $\mathrm{e}^{\mathrm{i}(\vec{(}Q)\cdot\vec{r}-\omega t)}$, of the plane-wave states of (1.49) and (1.68) are equal to 1. Therefore:

$$\langle\vec{k}_\mathrm{i}(0)|\vec{k}_\mathrm{f}(0)\rangle\langle\vec{k}_\mathrm{f}(0)|\vec{k}_\mathrm{i}(0)\rangle = 1\,. \tag{4.12}$$

We define the difference in frequency as ω:

$$\omega_i - \omega_f \equiv \omega\,. \tag{4.13}$$

Substituting (4.12) and (4.13) into (4.11):

$$P_{\mathrm{i}\to\mathrm{f}} = \frac{2\pi}{\hbar}\int \sum_j \mathrm{e}^{\mathrm{i}\omega t_j}b_j^*\langle\lambda_\mathrm{i}(0)|\,\mathrm{e}^{\mathrm{i}\vec{k}_\mathrm{i}\cdot\underline{\vec{R}}_j(t_j)}\mathrm{e}^{-\mathrm{i}\vec{k}_\mathrm{f}\cdot\underline{\vec{R}}_j(t_j)}|\lambda_\mathrm{f}(0)\rangle$$
$$\sum_k \mathrm{e}^{-\mathrm{i}\omega t_k}b_k\langle\lambda_\mathrm{f}(0)|\mathrm{e}^{\mathrm{i}\vec{k}_\mathrm{f}\cdot\underline{\vec{R}}_k(t_k)}\,\mathrm{e}^{-\mathrm{i}\vec{k}_\mathrm{i}\cdot\underline{\vec{R}}_k(t_k)}|\lambda_\mathrm{i}(0)\rangle\,\mathrm{d}t\,, \tag{4.14}$$

The integration is over all times, so we can redefine t as a difference between scattering times:

$$t \equiv t_j - t_k\,, \tag{4.15}$$

and likewise we define the scattering vector $\vec{Q}$:

$$Q \equiv \vec{k}_\mathrm{i} - \vec{k}_\mathrm{f}\,. \tag{4.16}$$

Substituting these differences into (4.14):

$$P_{\mathrm{i}\to\mathrm{f}} = \frac{2\pi}{\hbar}\int \mathrm{e}^{\mathrm{i}\omega t}\sum_j b_j^*\langle\lambda_\mathrm{i}(0)|\,\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\underline{\vec{R}}_j(t)}|\lambda_\mathrm{f}(0)\rangle$$
$$\sum_k b_k\langle\lambda_\mathrm{f}(0)|\mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\underline{\vec{R}}_k(0)}|\lambda_\mathrm{i}(0)\rangle\,\mathrm{d}t\,, \tag{4.17}$$

The next step is to consider changes in the state of the crystal, the $\{|\lambda(0)\rangle\}$. First consider the final states, $\{|\lambda_\mathrm{f}(0)\rangle\}$. We do not have control over which final

state is obtained, and in prinicple all acceptable final states will occur over the duration of a long experiment. In effect, an experiment sums over final states, but this sum is special. Since the final states are assumed to form a complete set:

$$\sum_{\mathrm{f}} |\lambda_{\mathrm{f}}(0)\rangle\langle\lambda_{\mathrm{f}}(0)| = 1 \ , \tag{4.18}$$

for each term in the double sum (4.17). The initial states $\{|\lambda_{\mathrm{i}}(0)\rangle\}$ can be controlled by thermodynamics. Temperature will alter the distribution of intial states by the appropriate thermodynamic distribution, i.e., a Bose–Einstein distribution for phonons. Instead of writing this distribution $n(\varepsilon)$ directly, we define the brackets $\left\langle\right\rangle$ to mean the thermodynamic average:

$$P_{\mathrm{i}\to\mathrm{f}} = \frac{2\pi}{\hbar}\int \mathrm{e}^{\mathrm{i}\omega t}\sum_j\sum_k b_j^* b_k \left\langle \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{\underline{R}}_j(t)}\,\mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\vec{\underline{R}}_k(0)}\right\rangle \mathrm{d}t \ , \tag{4.19}$$

$$P_{\mathrm{i}\to\mathrm{f}} = \frac{2\pi}{\hbar}\int \mathrm{e}^{\mathrm{i}\omega t}\sum_j\sum_k b_j^* b_k \left\langle \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{\underline{R}}_j(t)}\,\mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\vec{\underline{R}}_k(0)}\right\rangle \mathrm{d}t \ , \tag{4.20}$$

We make use of the prefactors described in Sect. 2.3 to convert the transition probability into a cross-section:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} = \frac{k_{\mathrm{f}}}{k_{\mathrm{i}}}\frac{(2\pi)^3}{2V_0}\sum_j\sum_k b_j^* b_k \int \mathrm{e}^{\mathrm{i}\omega t}\left\langle \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{\underline{R}}_j(t)}\,\mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\vec{\underline{R}}_k(0)}\right\rangle \mathrm{d}t \ . \tag{4.21}$$

Equation (4.21) is the most general result that can be obtained from Fermi's Golden Rule. At its heart is a thermodynamic average of phase factors from scattering by atoms j and k at their different positions in space and at different snapshots in time. Note that the space coordinates for the two atoms are generally not evaluated at the same time. If there were no time dependence to the atom positions, the Fourier transform would yield a delta function $\delta(\omega - 0)$, indicating pure elastic scattering. If there were no spatial periodicities, the thermodynamic average over all phase factors would not produce any constructive or destructive interferences at different angles, so there would be no structure of the cross section in solid angle, Ω. The topic of this book is inelastic scattering, so we assume motion of the scatterers, and we will be sensitive to the how the two phase factors change with time. The upcoming approximation will address the thermal spread of atom positions over time. Specifically we will assume this is a Gaussian function, or at least it is small. For the spatial periodicities of the scatterers, we will assume the translational periodicity of a crystal. This is a more restrictive assumption about the sample, and can be misleading in cases of disordered solids. It is therefore sometimes important to return to (4.21) for guidance on the scattering problem, since the only assumption is that the neutron is scattered one time.

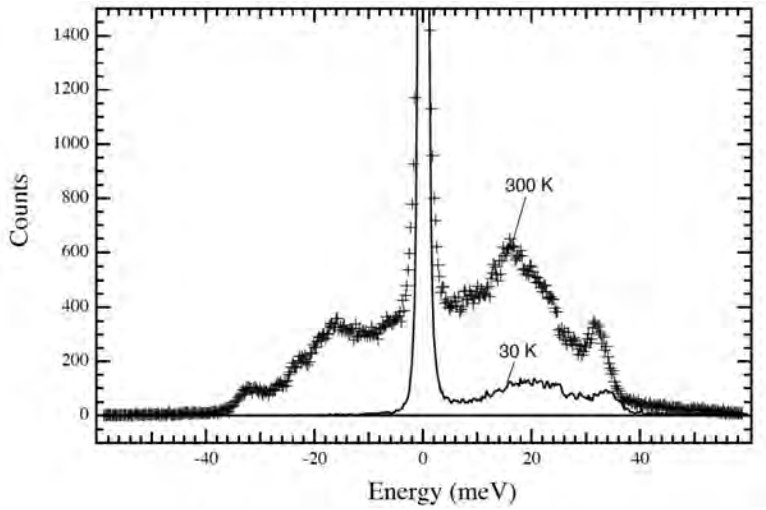


Figure 4.1: Incoherent inelastic scattering of phonons in Invar, fcc $Fe_{64}Ni_{36}$. Data were for equal incident flux, but different temperatures as labeled. Elastic peak at $\varepsilon = 0$ rises to 15,000 counts in both spectra. Compare the relative intensites at $\pm\varepsilon$, which are related by detailed balance. (Incidentally, some multiphonon scattering is evident in the 300 K data above 37 meV.) These spectra were from nuclear resonant inelastic x-ray scattering of ^{57}Fe with a high resolution monochromator at Sector 3 of the Advanced Photon Source (courtesy S.H. Lohaus and P. Guzman). The principles are similar to incoherent inelastic neutron scattering, but these are inelastic x-ray measurements.

4.0.2 Detailed Balance

The intensities of inelastic spectra depend on the ratio of energy transfer to temperature, at least in the usual case where the sample is in thermodynamic equilibrium before scattering. There is a detailed balance between the rates of two scattering processes, one with the creation of an excitation, and the other with the annihilation of the same excitation. Consider the temperature dependence of the positive and negative energy transfers between the neutron and the specimen. One extreme is when the sample is at a very low temperature (in practice, where $k_B T$ is much smaller than the energy resolution of the instrument). In this case there are no excitations present in the sample, so no scattering can occur with the annihilation of excitations. Excitations can be created by transfer of energy from the incident neutron, of course. At low temperatures the inelastic spectrum will have intensity on one side of the elastic peak, but no intensity on the other side. This is seen in Fig. 4.1, at 30 K, where there is no intensity for $S(-E)$, but a weak but reasonable spectrum is seen at $S(+E)$.

The other extreme is for very high temperatures where $k_B T$ is much larger

than the energy transfer $|E|$. Because the probability of creating or annihilating one additional excitation such as a phonon makes little difference to the energy of the sample, we expect the measured intensities to be more similar at $\pm E$. Scattering processes involving the creation or annihilation of excitations occur with similar probabilities when the sample is at high temperatures.

Each creation process has an inverse annihilation process, and the intensities of these two processes in an inelastic spectrum are related by a thermal factor that we show to be a simple Boltzmann factor. We will compare intensities at pairs of energies $\pm E$, i.e., at $S(-E)$ and $S(+E)$. The systematics to be explained can be seen with Fig. 4.1.

The energy equivalence to 300 K is k_B 300=25 meV, and at 30 K it is 2.5 meV. The heights of each spectrum differ by a factor of e at these energies, i.e., at 300 K, with E in meV, $S(+25) = 2.71828\, S(-25)$, and at 30 K, $S(+2,5) = 2.71828\, S(-2,5)$. This ratio is visible at ±25 meV for the 300 K spectrum but the 30 K spectrum is covered by the tails of the elastic line at 2.5 meV. Notice also the differences in the shapes of the two spectra on the phonon creation side of the elastic peak $+E$. The 30 K spectrum is actually close to the shape of the phonon DOS itself, whereas the 300 K spectrum is enhanced at energies below approximately 25 meV, owing to the large phonon occupancy factor.

A picture inspired by quantum mechanics can account for the detailed balance of the intensities of spectral features at $\pm E$ in Fig. 4.1. The density of states for excitations is symmetric, i.e., $S(-E) = S(+E)$, because an excitation can be created or annihilated in each mode at E. Suppose that the observed intensities are proportional to the thermal occupancies before scattering, which are the Planck factors $n_{E,T}$

$$n_{E,T} = \frac{1}{e^{\frac{E}{k_B T}} - 1} \,. \tag{4.22}$$

The idea follows stimulated emission in optics, where both phonon absorption and phonon emission to the neutron increase with $n_{E,T}$, but emission (phonon creation) has an extra spontaneous contribution of +1. This relates the intensities at $\pm E$ as

$$\frac{I(+E)}{I(-E)} = \frac{n_{E,T} + 1}{n_{E,T}} \,. \tag{4.23}$$

This Eq. 4.23 is easy to remember, and readily shows that as $n_{E,T} \to 0$ at low T, only the right side of the spectrum will be observed (see 30 K spectrum in Fig. 4.1).

However, Eq. 4.23 is not quite so quantum mechanical as it may seem. Rearranging

$$\begin{aligned} \frac{I(+E)}{I(-E)} &= \left(\frac{1}{e^{\frac{E}{k_B T}} - 1} + \frac{e^{\frac{E}{k_B T}} - 1}{e^{\frac{E}{k_B T}} - 1} \right) \left(e^{\frac{E}{k_B T}} - 1 \right) \,, \\ \frac{I(+E)}{I(-E)} &= e^{\frac{E}{k_B T}} \,, \end{aligned} \tag{4.24}$$

and the intensities for phonon creation and annihilation are in the ratio of the simple Boltzmann factor $\exp(\frac{E}{k_B T})$. The Planck occupancy does express the quantum statistics of bosons, of course. Furthermore, a simple classical analysis might give the incorrect ratio of $I(+E)/I(-E) = \exp(\frac{2E}{k_B T})$ because spectral components at $\pm E$ are separated in energy by $2E$.

The condition of detailed balance follows from two assumptions:

- The probability of the initial state of the sample, the $|\lambda_i\rangle$ in (4.2), is as expected for thermodynamic equilibrium.
- The interaction operator for the transition probability, the $V(\vec{r}, t')$ of (4.1), is Hermitian. (This is certainly true for the delta function (2.40) for the Fermi pseudopotential.)

To show the essence of the derivation of the detailed balance condition, assume the initial state $|\lambda_1\rangle$ exists, and consider the probability, $W'_{1\to 2}$, for a transition to a final state $|\lambda_2\rangle$, and the probability for the reverse transition, $W'_{2\to 1}$ from a pre-existing state $|\lambda_2\rangle$:

$$W'_{1\to 2} = |\langle\lambda_2|V|\lambda_1\rangle|^2 , \tag{4.25}$$

$$W'_{2\to 1} = |\langle\lambda_1|V|\lambda_2\rangle|^2 , \tag{4.26}$$

which are both products of a number with its complex conjugate:

$$W'_{1\to 2} = \langle\lambda_2|V|\lambda_1\rangle \left(\langle\lambda_2|V|\lambda_1\rangle\right)^* , \tag{4.27}$$

$$W'_{2\to 1} = \left(\langle\lambda_1|V|\lambda_2\rangle\right)^* \langle\lambda_1|V|\lambda_2\rangle . \tag{4.28}$$

For a Hermitian operator, recall that $V = (V^{\mathrm{T}})^* \equiv V^\dagger$. We use the transpose to operate on the other side of V, for which we use the complex conjugates of the bras and kets:

$$W'_{1\to 2} = \langle\lambda_2|V|\lambda_1\rangle \langle\lambda_1|V|\lambda_2\rangle , \tag{4.29}$$

$$W'_{2\to 1} = \langle\lambda_2|V|\lambda_1\rangle \langle\lambda_1|V|\lambda_2\rangle , \tag{4.30}$$

so:

$$W'_{1\to 2} = W'_{2\to 1} \equiv W' . \tag{4.31}$$

This result (4.31) is true so long as V is Hermitian. Starting in the known states $|\lambda_1\rangle$ and $|\lambda_2\rangle$, the transition probabilites between these states are equal.

Now assume that the states $|\lambda_i\rangle$ are populated in thermodynamic equilibrium, differing by a Boltzmann factor. The measured cross sections are proportional to $W_{1\to 2}$ and $W_{2\to 1}$:

$$W_{1\to 2} = \frac{\mathrm{e}^{-E_1/k_B T}}{Z} W'_{1\to 2} = \frac{\mathrm{e}^{-E_1/k_B T}}{Z} W' , \tag{4.32}$$

$$W_{2\to 1} = \frac{\mathrm{e}^{-E_2/k_B T}}{Z} W'_{2\to 1} = \frac{\mathrm{e}^{-E_2/k_B T}}{Z} W' , \tag{4.33}$$

where Z is the partition function. Now that we have taken into consideration the fact that the initial states have probabilities as expected from thermodynamic equilibrium, we can relate the observed intensity for the transition $1 \rightarrow 2$ to the observed intensity for its reverse transition $2 \rightarrow 1$:

$$W_{1\rightarrow 2} = \mathrm{e}^{(E_2-E_1)/k_\mathrm{B}T}\, W_{2\rightarrow 1}\ . \tag{4.34}$$

Suppose the state $|\lambda_2\rangle$ has an energy higher (more positive) than $|\lambda_1\rangle$, because the state $|\lambda_2\rangle$ has an extra excitation in the sample. The transition $1 \rightarrow 2$ is uphill energetically, and requires the neutron to transfer energy to create an excitation sample. Nevertheless, this transition is more intense experimentally because the initial state $|\lambda_1\rangle$ is more probable thermodynamically. Equation (4.34) shows that the intensity $W_{1\rightarrow 2} > W_{2\rightarrow 1}$ because $E_2 > E_1$ and the exponential is greater than 1. To clarify (4.34), we recognize that the difference in energy, $E = E_2 - E_1$, is the energy of the excitation in the solid. It is more traditional to write the condition of detailed balance as:

$$S(E) = \mathrm{e}^{E/k_\mathrm{B}T}\, S(-E)\ . \tag{4.35}$$

where E is the energy of the excition in the solid, and the argument $-E$ corresponding to $W_{2\rightarrow 1}$ signifies that $S(-E)$ is on the phonon annihilation side of the elastic peak in the spectrum. Detailed balance remains valid when a single scattering creates or annihilates multiple excitations – a detailed balance between forward and reverse processes still exists because the thermodynamic probabilities of the required initial states are set by E.

A practical use of detailed balance is to check the quality of experimental data. For example, in a spectrum measured at 300 K, equivalent to 25 meV, the intensities at ±25 meV on the two sides of the elastic peak must be in the ratio of e^{-1}. If this were not true, we might suspect instrument artifacts, such as differences in sensitivity or resolution. A noise background could also cause measured data to violate the condition of detailed balance, and perhaps detailed balance could be exploited to help subtract some sources of background from experimental data. We warn the reader, however, about such simple interpretations with data from a time-of-flight chopper spectrometer. The value of Q varies across the energy scale of data from a particular detector, and the relationship is not symmetric, i.e., $Q(E) \neq Q(-E)$. When multiphonon scattering is strong, and the Debye–Waller factor is significant, detailed balance will not be observed in the experimental data unless they are rebinned for constant Q.

4.0.3 Crystalline Periodicity

The translational periodicity of crystals allows the reduction of the double sum in (4.21) to a single sum. We separate the atom position operators, $\underline{\vec{R}}_j(t)$, into static, $\underline{\vec{x}}_{l,\kappa}$, and time-varying, $\underline{\vec{u}}_{l,\kappa}(t)$, components:

$$\underline{\vec{R}}_j(t) \quad = \quad \underline{\vec{x}}_l + \underline{\vec{x}}_\kappa + \underline{\vec{u}}_{l,\kappa}(t)\ , \tag{4.36}$$

where the static positions were broken into lattice vectors (index l) and basis vectors (index κ) in (4.36). The exponentials in (4.21) refer to pairs of atoms separated by a distance $\vec{x}_j - \vec{x}_k$. In an infinite crystal, where all unit cells are equivalent, these exponentials cannot depend on the absolute position of the unit cell, but only on the lattice translation vector $\vec{x}_l$. Any one of the N unit cells can be taken as the origin, and the terms from the l^{th} neighbor must be the same. This distance, $\vec{x}_l$, has no time dependence, and is a constant. It therefore commutes through the other distance operators, and we can write:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} = \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^3 N|b|^2}{2V_0}\sum_l \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{x}_l}\int \mathrm{e}^{\mathrm{i}\omega t}\left\langle \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\underline{\vec{u}}_l(t)}\,\mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\underline{\vec{u}}_0(0)}\right\rangle \mathrm{d}t\ . \tag{4.37}$$

Here, for simplicity we have also assumed all nuclei are the same so $b_j = b_k \equiv b$, and we have assumed a simple lattice without a basis. The result in (4.37) accounts for the interactions of waves scattered from pairs of nuclei, recognizing that their instantaneous displacements may differ by a phase factor. This difference in phase factor gives a constant prefactor $\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{x}_l}$. The other exponentials are not constants, however, and need to be handled with care. The temptation would be to combine the exponentials into a single factor such as $\mathrm{e}^{\mathrm{i}\vec{Q}\cdot(\underline{\vec{R}}_j(t)-\underline{\vec{R}}_j(0))}$. Unfortunately, this tempting step would be incorrect, except for classical systems. The quantum argument, described next, is subtle.

4.1 Quantum Behavior

4.1.1 Commutation Relations

Calculating the expectation value of an operator $\underline{A}$ using the left- and right-hand sides of (4.4) and (4.5) gives:

$$\langle\psi(t)|\underline{A}|\psi(t)\rangle = \langle\psi(t=0)|\mathrm{e}^{+\mathrm{i}\underline{H}t/\hbar}\underline{A}\mathrm{e}^{-\mathrm{i}\underline{H}t/\hbar}|\psi(t=0)\rangle\ . \tag{4.38}$$

Evidently the time-dependence of the matrix element can be transferred from the state functions ψ to the operator $\underline{A}$ by replacing the operator by $\mathrm{e}^{+\mathrm{i}\underline{H}t/\hbar}\underline{A}\mathrm{e}^{-\mathrm{i}\underline{H}t/\hbar}$. This moves us into the "Heisenberg picture" of quantum mechanics where state functions are fixed, but the time-dependence is in the operator. The motivation for putting the dynamics into the operators is as follows. In passing from classical mechanics to quantum mechanics, we replace the position vector $\vec{R}(t)$ with an operator, denoted $\underline{\vec{R}}(t)$. In particular, we will alter soon work with phase factors that are a time-dependent operators:

$$\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{R}} \longrightarrow \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\underline{\vec{R}}}\ . \tag{4.39}$$

Changing $\vec{R}$ to the operator $\underline{\vec{R}}$ leads to a subtlety in calculating the intensity from the wave amplitude. It turns out that the operator $\underline{\vec{R}}(0)$ does not commute

with the operator $\underline{\vec{R}}(t)$ for the same atom at a different time. The operators are related as:

$$\underline{\vec{R}}(t) = e^{+i\underline{H}t/\hbar}\underline{\vec{R}}(0)e^{-i\underline{H}t/\hbar} , \tag{4.40}$$

$$e^{-i\underline{H}t/\hbar}\vec{R}(t) = \underline{\vec{R}}(0)e^{-i\underline{H}t/\hbar} , \tag{4.41}$$

and likewise for the exponentiated operators:

$$e^{i\vec{Q}\cdot\underline{\vec{R}}(t)} = e^{+i\underline{H}t/\hbar}e^{i\vec{Q}\cdot\underline{\vec{R}}(0)}e^{-i\underline{H}t/\hbar} , \tag{4.42}$$

$$e^{-i\underline{H}t/\hbar}e^{i\vec{Q}\cdot\underline{\vec{R}}(t)} = e^{i\vec{Q}\cdot\underline{\vec{R}}(0)}e^{-i\underline{H}t/\hbar} . \tag{4.43}$$

We cannot interchange the order of $\underline{\vec{R}}$ and $\underline{H}$ because $\underline{H}$ includes the momentum operator.

To get better expressions for calculating with these operators, we need a result about exponentiations of non-commuting operators. Operators in exponential functions are defined in terms of the power series expansion of the exponential. Consider the exponentiation of two operators $\underline{A}$ and $\underline{B}$:

$$\exp(\underline{A}) = 1 + \underline{A} + \frac{1}{2}\underline{A}\,\underline{A} + \dots , \tag{4.44}$$

$$\exp(-\underline{B}) = 1 - \underline{B} + \frac{1}{2}\underline{B}\,\underline{B} + \dots , \tag{4.45}$$

Now take the product, including all terms to the second order:

$$\exp(\underline{A})\exp(-\underline{B}) = \left[1 + \underline{A} + \frac{1}{2}\underline{A}\,\underline{A} + \dots\right]\left[1 - \underline{B} + \frac{1}{2}\underline{B}\,\underline{B} + \dots\right] . \tag{4.46}$$

Group the terms in order of the nth power of operators $\underline{A}$ and $\underline{B}$ as O^n, and stop at the second order. (This truncations may seem like an approximation now, but it proves exact after a physical fact is brought out later.)

$$\exp(\underline{A})\exp(-\underline{B}) = O^0 + O^1 + O^2 , \tag{4.47}$$

$$O^0 = 1 , \tag{4.48}$$

$$O^1 = \underline{A} - \underline{B} , \tag{4.49}$$

$$O^2 = \frac{1}{2}\underline{AA} - \underline{AB} + \frac{1}{2}\underline{BB} , \tag{4.50}$$

$$\exp(\underline{A})\exp(-\underline{B}) = 1 + \underline{A} - \underline{B} + \frac{1}{2}\underline{AA} - \underline{AB} + \frac{1}{2}\underline{BB} , \tag{4.51}$$

where we have been fastidious about keeping the operator $\underline{A}$ to the left of the operator $\underline{B}$, because they may not necessarily commute.

Now evaluate the exponentiation of the sum $\underline{A} + \underline{B}$, again to second order in the operators:

$$\exp(\underline{A} - \underline{B}) = 1 + (\underline{A} - \underline{B}) + \frac{1}{2}(\underline{A} - \underline{B})\,(\underline{A} - \underline{B}) , \tag{4.52}$$

$$\exp(\underline{A} - \underline{B}) = 1 + \underline{A} - \underline{B} + \frac{1}{2}\left[\underline{A}\,\underline{A} - \underline{A}\,\underline{B} - \underline{B}\,\underline{A} + \underline{B}\,\underline{B}\right] , \tag{4.53}$$

Note that (4.51) and (4.53) are unequal when $\underline{A}$ does not commute with $\underline{B}$. We need one more exponentiation expression to make an equality. It is, to second order in the operators

$$\exp(\frac{1}{2}[\underline{B},\underline{A}]) = 1+\frac{1}{2}[\underline{B},\underline{A}]\,, \tag{4.54}$$

$$\exp(\frac{1}{2}[\underline{B},\underline{A}]) = 1+\frac{1}{2}(\underline{BA}-\underline{AB})\,. \tag{4.55}$$

Multiplying the expressions for (4.53) and (4.55)

$$\begin{aligned}\exp(\underline{A}-\underline{B})\,\exp(\frac{1}{2}[\underline{B},\underline{A}]) &= \Big[1+\underline{A}-\underline{B}+\frac{1}{2}\big(\underline{A}\,\underline{A}-\underline{A}\,\underline{B}-\underline{B}\,\underline{A}+\underline{B}\,\underline{B}\big)\Big]\\ &\times \Big[1+\frac{1}{2}(\underline{BA}-\underline{AB})\Big]\end{aligned} \tag{4.56}$$

and considering terms only to second order in the operators

$$\exp(\underline{A}-\underline{B})\,\exp(\frac{1}{2}[\underline{B},\underline{A}]) = 1+\underline{A}-\underline{B}+\frac{1}{2}\underline{AA}-\underline{AB}+\frac{1}{2}\underline{BB}\,, \tag{4.57}$$

This is exactly (4.51), so we conclude

$$\exp(\underline{A}-\underline{B})\,\exp(\frac{1}{2}[\underline{B},\underline{A}]) = \exp(\underline{A})\exp(-\underline{B})\,. \tag{4.58}$$

This is the "Weyl identity."

Not so fast, you may say. What about the terms O^3 and higher? The next subsection shows how these higher terms can be identically zero when the commutator $[\underline{B},\underline{A}]$ is a constant, and gives a practical case where this happens. If you are satisfied with the approach in this section, you can skip the next section because it is not crucial for the development.

4.1.2 An Operator Identity (Optional Section)

Another useful identity can be obtained from an expansion for projection operations. We take the expression $e^{+\underline{A}}\underline{B}e^{-\underline{A}}$ and write out expansions for the exponentiated operators

$$e^{+\underline{A}} = 1+\underline{A}+\frac{1}{2!}\underline{A}^2+\frac{1}{3!}\underline{A}^3+\ldots\,, \tag{4.59}$$

$$e^{-\underline{A}} = 1-\underline{A}+\frac{1}{2!}\underline{A}^2-\frac{1}{3!}\underline{A}^3+\ldots\,, \tag{4.60}$$

and develop a scheme for working with all terms to higher order beyond the cubic. Expanding the typical projection expression in the Heisenberg representation

$$e^{+\underline{A}}\underline{B}e^{-\underline{A}} = \Big(1+\underline{A}+\frac{1}{2!}\underline{A}^2+\frac{1}{3!}\underline{A}^3+\ldots\Big)\underline{B}\Big(1-\underline{A}+\frac{1}{2!}\underline{A}^2-\frac{1}{3!}\underline{A}^3+\ldots\Big)\,, \tag{4.61}$$

we seek to group all terms in the same order of operators, denoted O^0, O^1, O^2, ... such as the lowest ones that are easy to see in Eq. 4.61

$$O^0 = \underline{B} , \tag{4.62}$$

$$O^1 = \underline{BA} - \underline{AB} = [\underline{B}, \underline{A}] . \tag{4.63}$$

It turns out that the big trick is already evident in the expression (4.63) for O^1. To go from O^0 to O^1, we commute O^0 with $\underline{A}$. This works to all higher orders. For example, go from O^1 to O^2, first laboriously by writing out all the terms to second order

$$O^2 = \underline{\left(1 + A + \frac{1}{2!}A^2\right)\underline{B}\left(1 - \underline{A} + \frac{1}{2!}\underline{A}^2\right)} , \tag{4.64}$$

$$O^2 = \frac{1}{2!}\underline{BA}^2 - \underline{ABA} + \frac{1}{2!}\underline{A}^2\underline{B} , \tag{4.65}$$

where the ends of the underlines in (4.64) denote the product factors that contribute terms to second order in $\underline{A}$ in (4.65). Equation (4.65) can be written alternatively as

$$O^2 = \frac{1}{2!}\left(\underline{BA}^2 - \underline{ABA} - \underline{ABA} + \underline{A}^2\underline{B}\right) , \tag{4.66}$$

$$O^2 = \frac{1}{2!}\left((\underline{BA} - \underline{AB})\underline{A} - \underline{A}(\underline{BA} - \underline{AB})\right) , \tag{4.67}$$

$$O^2 = \frac{1}{2!}\left[[\underline{B}, \underline{A}], \underline{A}\right] , \tag{4.68}$$

The same trick works for going from O^2 to O^3

$$O^3 = \frac{1}{3!}\left[\left[[\underline{B}, \underline{A}], \underline{A}\right], \underline{A}\right] , \tag{4.69}$$

What we do by commuting with $\underline{A}$, perhaps clear in (4.67) is making a term with a power of $\underline{A}$ to the right, and a term with the opposite sign with $\underline{A}$ to the left. This is just what we need when working the expansion to higher orders, using the method of underlines in (4.64). It is a good exercise to obtain O^3 by the method of (4.64), but to the next order there will be four underlines that denote the terms for O^3.

This elegance can be hard to use, but it is valuable in our case of position and momentum operators, $\underline{X}$ and $\underline{P}$. The important point is that the commutation of these operators is a constant,

$$[\underline{X}, \underline{P}] = i\hbar . \tag{4.70}$$

Any later commutation with a constant is zero, because the two terms in the commutator are equal. Therefore, all terms beyond O^1 vanish identically. For

example, consider

$$e^{-i\underline{PX'}/\hbar}\underline{X}e^{+i\underline{PX'}/\hbar} = \underline{X} + [\underline{X}, i\underline{PX'}/\hbar] , \tag{4.71}$$

$$e^{-i\underline{PX'}/\hbar}\underline{X}e^{+i\underline{PX'}/\hbar} = \underline{X} + \frac{i}{\hbar}\Big(\underline{XPX'} - \underline{PX'}\,\underline{X}\Big), \tag{4.72}$$

$$e^{-i\underline{PX'}/\hbar}\underline{X}e^{+i\underline{PX'}/\hbar} = \underline{X} + \frac{i}{\hbar}\Big((i\hbar + \underline{PX})\underline{X'} - \underline{PX'}\,\underline{X}\Big), \tag{4.73}$$

$$e^{-i\underline{PX'}/\hbar}\underline{X}e^{+i\underline{PX'}/\hbar} = \underline{X} - \underline{X'} , \tag{4.74}$$

where we substituted the commutation result (4.70) in the small ( ) in (4.73). This is an exact result when the commutator $[\underline{B}, \underline{A}]$ is a constant – the higher terms obtained by commuting $[\underline{B}, \underline{A}]$ with $\underline{A}$ give zero identically.

4.1.3 Gaussian Thermal Averages

Now that we have the expression (4.58), we can use it to rearrange (4.37) into a form that shows intensity contributions from different numbers of phonon excitations. To clarify the next steps, it is traditional to make the definitions:

$$\underline{\vec{U}} \equiv -i\vec{Q}\cdot\underline{\vec{u}}_0(0) , \tag{4.75}$$

$$\underline{\vec{V}} \equiv i\vec{Q}\cdot\underline{\vec{u}}_l(t) . \tag{4.76}$$

Now we substitute (4.75) and (4.76) into (4.58):

$$\exp(\underline{\vec{U}})\exp(\underline{\vec{V}}) = \exp(\underline{\vec{U}} + \underline{\vec{V}})\ \exp\Big(\frac{1}{2}\Big[\underline{\vec{U}}\,\underline{\vec{V}} - \underline{\vec{V}}\,\underline{\vec{U}}\Big]\Big) . \tag{4.77}$$

At this point we need to take a thermal average (the $\langle\rangle$ in (4.37)). In doing so, we make the assumption that the vibrational atom displacements are distributed with a Gaussian spread. The Gaussian distribution function can be used to average a squared displacement, X^2:

$$\langle X^2\rangle = \int_{-\infty}^{\infty} X^2 \frac{1}{\sqrt{\pi\sigma^2}} e^{-X^2/\sigma^2}\, dX , \tag{4.78}$$

$$\langle X^2\rangle = \frac{1}{2}\sigma^2 , \tag{4.79}$$

a standard result.

For comparison, we next average an exponential e^X:

$$\langle e^X\rangle = \int_{-\infty}^{\infty} e^X \frac{1}{\sqrt{\pi\sigma^2}} e^{-X^2/\sigma^2}\, dX , \tag{4.80}$$

by completing the square:

$$-(X/\sigma - \sigma/2)^2 = -\Big(X^2/\sigma^2 - X + \sigma^2/4\Big) , \tag{4.81}$$

so by adding and subtracting the last term of (4.81) in the exponential:

$$\langle e^X \rangle = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi\sigma^2}} e^{-(X/\sigma-\sigma/2)^2} e^{\sigma^2/4}\, dX\,, \tag{4.82}$$

$$\langle e^X \rangle = e^{\sigma^2/4} \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi\sigma^2}} e^{-(X/\sigma-\sigma/2)^2}\, dX\,, \tag{4.83}$$

$$\langle e^X \rangle = e^{\sigma^2/4}\,, \tag{4.84}$$

where the last result was obtained by integrating the normalized Gaussian function. By comparing (4.79) and (4.84), we obtain:

$$\langle e^X \rangle = e^{X^2/2}\,. \tag{4.85}$$

Using (4.85), we obtain a factor with our operators $\underline{\vec{U}}$ and $\underline{\vec{V}}$:

$$\langle \exp(\underline{\vec{U}} + \underline{\vec{V}}) \rangle = \exp\left(\frac{1}{2}\left(\underline{\vec{U}} + \underline{\vec{V}}\right)^2\right). \tag{4.86}$$

This is the "Bloch identity" for Gaussian thermal averages, usually written more simply with $\underline{C} = \underline{\vec{U}} + \underline{\vec{V}}$. We use this result to obtain the thermal average of (4.77):

$$\langle \exp \underline{\vec{U}} \exp \underline{\vec{V}} \rangle = \exp\left(\frac{1}{2}\left(\underline{\vec{U}} + \underline{\vec{V}}\right)^2\right) \exp\left(\frac{1}{2}\langle \underline{\vec{U}}\,\underline{\vec{V}} - \underline{\vec{V}}\,\underline{\vec{U}} \rangle\right), \tag{4.87}$$

$$\langle \exp \underline{\vec{U}} \exp \underline{\vec{V}} \rangle = \exp\left(\frac{1}{2}\langle \underline{\vec{U}}^2 + \underline{\vec{V}}^2 + \underline{\vec{U}}\underline{\vec{V}} + \underline{\vec{V}}\underline{\vec{U}} + \underline{\vec{U}}\underline{\vec{V}} - \underline{\vec{V}}\underline{\vec{U}} \rangle\right), \tag{4.88}$$

$$\langle \exp \underline{\vec{U}} \exp \underline{\vec{V}} \rangle = \exp\left(\frac{1}{2}\langle \underline{\vec{U}}^2 + \underline{\vec{V}}^2 \rangle\right) \exp\langle \underline{\vec{U}}\underline{\vec{V}} \rangle\,. \tag{4.89}$$

By examining (4.75) and (4.76), we can see that:

$$\langle \underline{\vec{U}}^2 \rangle = \langle \underline{\vec{V}}^2 \rangle\,, \tag{4.90}$$

because the average vibrational amplitudes do not change over time, and all unit cells are equivalent in the crystal. This allows a final simplification of (4.89):

$$\langle \exp \underline{\vec{U}} \exp \underline{\vec{V}} \rangle = \exp\langle \underline{\vec{U}}^2 \rangle \exp\langle \underline{\vec{U}}\underline{\vec{V}} \rangle\,. \tag{4.91}$$

Equation (4.91) was obtained with the one assumption of a Gaussian thermal spread of atom positions. Even if this in not quite the case, (4.91) is expected to be valid when the atom displacements are small. By expanding both sides of (4.85), and recognizing that the thermal average of odd powers of X are zero, (4.85) seems a reasonable approximation.

Finally, we use (4.91) to rewrite (4.37) in a way that will later let us identify individual phonon scatterings. Making use of the definitions (4.75) and (4.76):

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} = \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^3 N|b|^2}{2V_0}\sum_l \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{x}_l}\int \mathrm{e}^{\mathrm{i}\omega t}\exp\langle\underline{\vec{U}}^2\rangle\exp\langle\underline{\vec{U}\vec{V}}\rangle\,\mathrm{d}t\ . \tag{4.92}$$

Taking the thermal averages has removed much of the time dependence in our factors. The only remaining time dependence is within the $\langle\underline{\vec{U}\vec{V}}\rangle$ factor, so we rewrite:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} = \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^3 N|b|^2}{2V_0}\sum_l \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{x}_l}\exp\langle\underline{\vec{U}}^2\rangle\int \mathrm{e}^{\mathrm{i}\omega t}\exp\langle\underline{\vec{U}\vec{V}}\rangle\,\mathrm{d}t\ . \tag{4.93}$$

Using (4.75), the factor $\exp\langle\underline{\vec{U}}^2\rangle$ in (4.93) becomes $\exp(-\langle[\vec{Q}\cdot\underline{\vec{u}}_l(0)]^2\rangle)$. We recognize this as a Debye–Waller factor. To its right in (4.93) is a sum over pairs of atoms separated by $\vec{x}_l$. There is a phase associated with the atom separation in the pair. Suppose for the moment that there were no displacements of the atoms from their lattice sites. In this case the final exponential $\exp\langle\underline{\vec{U}\vec{V}}\rangle$ would equal $\mathrm{e}^0 = 1$. This is the case of elastic scattering, and the sum of phase factors of (4.93) would reduce to to the elastic scattering result of (2.48). It is the final factor, the Fourier transform of $\exp\langle\underline{\vec{U}\vec{V}}\rangle$, that makes (4.93) interesting, and it is in a convenient form for further development.

4.2 Practical Expressions for Phonon Scattering

4.2.1 Impulse Approximation

It is relatively easy to adapt the general result (4.21) to the case where the energy of the incident neutron is much higher than the characteristic excitations in the solid. We therefore ignore the interatomic interactions, and consider the collision of the neutron with a single nucleus at $\vec{R}$. This approaches the problem of hitting a ball with a classical projectile, so we lose some features of wave mechanics.[2] We do have to account for momentum and energy transfer, of course. The impulse approximation begins with the replacement of the operator:

$$\underline{\vec{R}}_j(t) \quad\longrightarrow\quad \underline{\vec{R}}_j(0) + t\underline{\vec{v}}_j\ , \tag{4.94}$$

$$\underline{\vec{R}}_j(t) \quad\longrightarrow\quad \underline{\vec{R}}_j(0) + \frac{t}{M_j}\underline{\vec{p}}_j\ , \tag{4.95}$$

[2] Another viewpoint is that we expect multiple excitations to occur in the solid. The effects of coherence are generally washed out when multiple excitations occur. For incoherent inelastic scattering as in Sect. 2.2, we expect to analyze the scattering by considering only a single nucleus at a time. We expect the excitations to be incoherent.

where we expect t to be small since the neutron is moving fast. Because the incoherent character of the multiple excitations suppresses the phase relationships between different scatterers, and we consider terms $j = k$ in (4.21). It is tempting to substitute (4.95) directly into (4.21) to obtain:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} \quad ? = ? \quad \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^3}{2V_0}\sum_j\sum_k b_j^* b_k \delta_{j,k} \times \int \mathrm{e}^{\mathrm{i}\omega t}\left\langle \mathrm{e}^{\mathrm{i}\vec{Q}\cdot(\vec{R}_j(0)+t/M_j\vec{p}_j)}\,\mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\vec{R}_k(0)}\right\rangle \mathrm{d}t\,, \tag{4.96}$$

and with $\underline{\vec{R}}_j(0) = \underline{\vec{R}}_k(0)$:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} \; ? = ? \; \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^3}{2V_0}\sum_j |b_j|^2 \int \mathrm{e}^{\mathrm{i}\omega t}\left\langle \mathrm{e}^{\mathrm{i}t/M_j\vec{Q}\cdot\vec{p}_j}\right\rangle \mathrm{d}t\,. \tag{4.97}$$

The missing piece in (4.97) is a phase factor associated with the energy gain of the scatterer. The energy gain is kinetic, $E_\mathrm{kin} = \frac{p^2}{2M_j} = \frac{\hbar^2Q^2}{2M_j}$. Fermi's Golden Rule, which connects the wavefunctions before and after the scattering, is senstitive to the phases of the initial and final states of the scatterer. The transfer of energy causes a relative change in the phase of these two states by the factor: $\exp(\mathrm{i}E_\mathrm{kin}\hbar^{-1}\,t)$. This phase relationship for the total energy transfer leaves a minimal amount of quantum mechanics in the scattering problem (which we expect to go away at very high incident energies):

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} = \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^3}{2V_0}\sum_j |b_j|^2 \int \mathrm{e}^{\mathrm{i}\omega t}\mathrm{e}^{\mathrm{i}\left(\frac{\hbar^2Q^2}{2M_j}\right)\hbar^{-1}\,t}\left\langle \mathrm{e}^{\mathrm{i}t/M_j\vec{Q}\cdot\vec{p}_j}\right\rangle \mathrm{d}t\,. \tag{4.98}$$

In Sect. 4.1.3 we calculated the thermal average, $\langle\rangle$, when the displacements of the scatterer had a Gaussian thermal spread. To apply this result to (4.98), we use the result: $\langle \mathrm{e}^{\mathrm{i}X}\rangle = \mathrm{e}^{-X^2/2}$:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} = \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^3}{2V_0}\sum_j |b_j|^2 \int \mathrm{e}^{\mathrm{i}\omega t}\mathrm{e}^{\mathrm{i}\left(\frac{\hbar^2Q^2}{2M_j}\right)\hbar^{-1}\,t}\mathrm{e}^{-(\langle\vec{Q}\cdot\vec{p}_j\rangle t/M_j)^2/2}\,\mathrm{d}t\,. \tag{4.99}$$

We complete the square in the exponential, defining:

$$x^2 \equiv \left(\sqrt{a}t - \frac{b}{2\sqrt{a}}\right)^2 = at^2 - bt + \frac{b^2}{4a}\,, \tag{4.100}$$

and we obtain a result:

$$\int_{-\infty}^{\infty} \mathrm{e}^{-at^2+bt}\,\mathrm{d}t = \int_{-\infty}^{\infty}\frac{\mathrm{e}^{-x^2}}{\sqrt{a}}\,\mathrm{d}x = \sqrt{\frac{\pi}{a}}\,\mathrm{e}^{b^2/4a}\,. \tag{4.101}$$

With the associations for a and b:

$$a = \frac{1}{2}\left(\frac{\langle \vec{Q}\cdot\vec{\underline{p}}_j\rangle}{M_j}\right)^2 , \tag{4.102}$$

$$b = \mathrm{i}\left(\omega + \frac{\hbar Q^2}{2M_j}\right) , \tag{4.103}$$

(4.99) becomes:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} = \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^{7/2}}{2V_0}\sum_j |b_j|^2\left(\frac{M_j}{\langle \vec{Q}\cdot\vec{\underline{p}}_j\rangle}\right)^2 \exp\left(-\frac{\left(\frac{\hbar Q^2}{2M_j}+\omega\right)^2}{2\left(\frac{\langle \vec{Q}\cdot\vec{\underline{p}}_j\rangle}{M_j}\right)^2}\right) . \tag{4.104}$$

The differential scattering cross section in (4.104) has the shape of a Gaussian function, centered at an energy $\hbar\omega$:

$$\hbar\omega = -\frac{\hbar^2 Q^2}{2M_j} . \tag{4.105}$$

The center of the Gaussian is simply the energy transfer from a scattering with a single-particle recoil. This result could be obtained by classical mechanics. The spread of this Gaussian is obtained from from the denoninator in the Gaussian of (4.104). This width increases with Q and with the mean-squared velocity of the scatterers. The ratio of shift to width grows larger with Q, however. In the classical limit of very large Q, the width is negligible, so the energy and momentum transfers are well-defined.

4.2.2 Multiphonon Expansion

We return to develop (4.93) with the "multiphonon expansion," which is obtained from the expansion of the exponential $\exp\langle\underline{\vec{U}}\underline{\vec{V}}\rangle$. This is seen most easily in the incoherent approximation, where we replace:

$$\exp\langle\underline{\vec{U}}\underline{\vec{V}}\rangle \longrightarrow \exp\langle\underline{\vec{U}}\rangle\langle\underline{\vec{V}}\rangle = \exp\langle\underline{\vec{U}}\rangle\langle\underline{\vec{U}}\rangle . \tag{4.106}$$

so that, treating $\underline{\vec{U}}$ as a displacement and not an operator:

$$\begin{aligned}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} &= \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^3 N|b|^2}{2V_0}\sum_l \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{x}_l}\exp\langle\underline{\vec{U}}^2\rangle \\ &\quad\times\int \mathrm{e}^{\mathrm{i}\omega t}\exp\left(-\mathrm{i}\vec{Q}\cdot\vec{u}\,\mathrm{i}\vec{Q}\cdot\vec{u}\right)\mathrm{d}t .\end{aligned} \tag{4.107}$$

The harmonic oscillator energy $M\omega^2 u_{\max}^2$, is assumed quantized in units of $\hbar\omega$. From $\hbar\omega \propto M\omega^2 u_{\max}^2$, we replace the displacement $u = \sqrt{\hbar/(2M\omega)}$, for which we expect a time-dependence:

$$\begin{aligned}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} &= \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^3 N|b|^2}{2V_0}\sum_l \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{x}_l}\exp\langle\underline{\vec{U}}^2\rangle \\ &\quad\times\int \mathrm{e}^{\mathrm{i}\omega t}\exp\!\left(\frac{Q^2\hbar}{2M\omega}\Upsilon(t)\right)\mathrm{d}t\ , \end{aligned} \tag{4.108}$$

The function $\Upsilon(t)$ serves to convert between phonon energy and nuclear displacement, and varies with the phonon DOS, $g(E)$, and temperature through the Planck factor, $n(E)$. It can be shown to be [Lovesey Vol 1. Eq. 4.238]

$$\Upsilon(t) = \int_{-\infty}^{\infty}\frac{g(E)}{E}\,n(E)\,\mathrm{e}^{-\mathrm{i}(E/\hbar)t}\,\mathrm{d}E\ . \tag{4.109}$$

Assuming the argument of the exponential in (4.108) is small, we expand this exponential

$$\begin{aligned}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} &= \frac{k_\mathrm{f}}{k_\mathrm{i}}\frac{(2\pi)^3 N|b|^2}{2V_0}\sum_l \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{x}_l}\exp\langle\underline{\vec{U}}^2\rangle \\ &\quad\times\int \mathrm{e}^{\mathrm{i}\omega t}\left(1+\frac{\frac{\hbar^2Q^2}{2M}}{\hbar\omega}\Upsilon(t)+\frac{1}{2}\left(\frac{\frac{\hbar^2Q^2}{2M}}{\hbar\omega}\Upsilon(t)\right)^2+\ldots\right)\mathrm{d}t\ . \end{aligned} \tag{4.110}$$

The terms in the expansion of (4.110) are recognized as a series in powers of $E_\mathrm{Recoil}/(\hbar\omega)$ — the recoil energy, $p^2/(2M)$, divided by the energy of the phonon, $\hbar\omega$. It is instructive to compare (4.110) to (2.51), which is re-written here. Note that it did not include a Debye–Waller factor:

$$\begin{aligned}\Psi_\mathrm{sc}(\vec{Q},\vec{r}) &= -\frac{\mathrm{e}^{\mathrm{i}(\vec{k}_\mathrm{f}\cdot\vec{r}-\omega_0 t)}}{|\vec{r}|} \\ &\quad\times\sum_{l,\kappa} b_\kappa\,\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{x}_{l,\kappa}}\left(1+\mathrm{i}\vec{Q}\cdot\vec{u}_{l,\kappa}(t)-\frac{1}{2}(\vec{Q}\cdot\vec{u}_{l,\kappa}(t))^2+\ldots\right)\ . \end{aligned} \tag{4.111}$$

Analyzing the two series in (4.110) and (4.111), term-by-term, we find:

- The first term, the 1, provides a Fourier transform of a constant, which is a delta function, $\delta(\omega)$. Since the excitation energy is therefore zero, this is an elastic scattering process. Note that the sum of phase factors $\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{x}_l}$ over lattice sites and the Debye–Waller factor $\mathrm{e}^{\langle\underline{\vec{U}}^2\rangle}$ are as expected for diffracted neutron wavefunctions. Elastic nuclear scattering has no Q dependence, except through the Debye–Waller factor.

- The second term, involving $\frac{\hbar^2 Q^2}{2M}/(\hbar\omega)$, is the inelastic scattering that occurs by exciting one phonon. We found in (2.52) that its Fourier transform led to a delta function $\delta(\omega - \omega(\vec{q}))$ from the conservation of energy. A conservation of momentum led to a similar delta function, $\delta\Big((\vec{q} - \vec{Q}) - \vec{g}\Big)$. Single phonon scattering increases with Q as Q^2.

- The third term, involving $\left(\frac{\frac{\hbar^2 Q^2}{2M}}{\hbar\omega}\right)^2$, is the scattering that involves the excitation of two phonons. This occurs in one scattering event, not through the creation of two phonons by two different deflections of the neutron. (The latter is "multiple scattering.") When the time function is squared, such as $\cos^2(\omega t) = 1/2\Big(1 + \cos(2\omega t)\Big)$, the frequency is doubled, and energy conservation provides the delta function $\delta(\omega - 2\omega(\vec{q}))$. Two-phonon scattering involves twice the energy transfer as a one-phonon process. Note how it increases rapidly with Q, going as Q^4.

- Higher order terms, involving $+ \ldots$, involve the excitation of many phonons in one scattering of the neutron. These higher-order terms approach classical behavior. Typically the scattering of a particle with a large a large momentum transfer causes the excitation of many phonons, sometimes better described as the creation of heat.

We note that the analysis here has assumed ignored phase relationships in multiphonon scattering. This is typical of even more sophisticated treatments of the problem. Multiphonon scattering is usually analyzed in the incoherent approximation.

4.3 Magnetic Scattering

4.3.1 Magnetic Form Factor and Scattering Amplitude

Magnetic scattering originates with the interaction between the spin of the neutron and the spins of electrons and/or the motions of electrons. Magnetic scattering is inherently more complicated than nuclear scattering because the potentials have vector character. The magnetic forces are also long range.

The scattering potential can be written in the general form $V = -\vec{\mu}_n \cdot B$, where B, which originates with the electrons, has a spin component B_S and an orbital component B_L:

$$B = B_S + B_L \ . \tag{4.112}$$

These components have different mathematical forms[3]:

$$B_{\mathrm{L}} \propto \frac{\widehat{\vec{R}\times\vec{p}}}{R^2}\,, \tag{4.113}$$

$$B_{\mathrm{S}} \propto \vec{\nabla}\times\left(\frac{\vec{s}\times\widehat{\vec{R}}}{R^2}\right). \tag{4.114}$$

From (1.75) we obtain the scattering amplitude in the first Born approximation as the Fourier transform of the scattering potential $V = -\vec{\mu}_{\mathrm{n}}\cdot B$

$$f_{\mathrm{mag}}(\vec{Q},E) = \sqrt{\frac{k_{\mathrm{f}}}{k_{\mathrm{i}}}}\frac{2m}{\hbar^2}\frac{1}{4\pi}\int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}}\,(-\vec{\mu}_{\mathrm{n}}\cdot B)\,\mathrm{d}\vec{r}\,. \tag{4.115}$$

$$f_{\mathrm{mag}}(\vec{Q},E) = \sqrt{\frac{k_{\mathrm{f}}}{k_{\mathrm{i}}}}(-\gamma r_{\mathrm{e}})\int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}}\,\vec{\sigma}\cdot\left(\frac{1}{\hbar}\frac{\widehat{\vec{R}\times\vec{p}}}{R^2}+\vec{\nabla}\times\left(\frac{\vec{s}\times\widehat{\vec{R}}}{R^2}\right)\right)\mathrm{d}\vec{r}\,, \tag{4.116}$$

where $\vec{\sigma}$ is the neutron spin, later to be the spin operator. Many constants were combined into the "classical electron radius," $r_{\mathrm{e}} = e^2/(m_{\mathrm{e}}c^2)$, and γ is the gyromagnetic ratio of the neutron, $\gamma = 1.913$.

The evaluation of (4.116) is most expedient with two mathematical tricks:

$$\int_{-\infty}^{\infty}\frac{\widehat{\vec{R}}}{R^2}\,\mathrm{e}^{-\mathrm{i}\vec{Q}\cdot\vec{R}}\,\mathrm{d}\vec{R} = -4\pi\mathrm{i}\,\frac{\widehat{\vec{Q}}}{Q}\,, \tag{4.117}$$

$$\vec{\nabla}\times\left(\frac{\vec{s}\times\widehat{\vec{R}}}{R^2}\right) = \frac{1}{2\pi^2}\int_{-\infty}^{\infty}\widehat{\vec{q}}\times(\vec{s}\times\widehat{\vec{q}})\,\mathrm{e}^{-\mathrm{i}\vec{q}\cdot\vec{R}}\,\mathrm{d}\vec{q}\,. \tag{4.118}$$

The derivation of (4.117) is not difficult – it involves transformation to spherical coordinates where the R^2 in the denominator is cancelled by an R^2 in the differential volume element.[4] Unfortunately, (4.118) is much more work to obtain unless one is rather clever with, or accepting of, vector identities. A sketch of its derivation is given in Appendix B.2 of Squires' book.

In using the tricks (4.117) and (4.118), the necessary steps are:

- With the neutron at $\vec{r}$ and the i^{th} electron at $\vec{r}_i$, the position for the magnetic field, $\vec{R}$ in (4.113) and (4.114), is the distance separating $\vec{r}$ and $\vec{r}_i$: $\vec{R} = \vec{r} - \vec{r}_i$.

[3]Note that B_{L} has the form of the Biot-Savart law for the electron current (with electron momentum $\vec{p}$) about the atom. The B_{S} can be written as a curl of a vector potential, $B_{\mathrm{S}} = \vec{\nabla}\times A$, if $A = \mu_{\mathrm{e}}\times\vec{R}/R^2$, and $\mu_{\mathrm{e}} = -e\hbar/m_{\mathrm{e}} = -2\mu_{\mathrm{B}}\vec{s}$.

[4]Note that (4.117) is the inverse transformation of (A.27) in the Appendix, with the interchange of $\vec{R}$ and $\vec{Q}$.

- The Fourier transform of $\widehat{\vec{R}}\times\vec{p}/R^2$ first involves replacing the exponential $e^{-i\vec{Q}\cdot\vec{R}} = e^{-i\vec{Q}\cdot\vec{r}}e^{+i\vec{Q}\cdot\vec{r}_i}$, in order to use coordinates of the neutron and the i^{th} electron. The second phase factor $e^{+i\vec{Q}\cdot\vec{r}_i}$, is a constant, and is removed from the integration over all space. The remaining spatial integration becomes, from (4.117), $-i4\pi\,\widehat{\vec{Q}}\times\vec{p}/Q$.

- Notice that the right-hand side of (4.118) has its only dependence on $\vec{R}$ in the exponential (which we again write as $e^{-i\vec{q}\cdot\vec{R}} = e^{-i\vec{q}\cdot\vec{r}}e^{+i\vec{q}\cdot\vec{r}_i}$). When is (4.118) substituted into (4.116), a double integral (over $\vec{r}$ and $\vec{q}$) is obtained. After the phase factor, $e^{i\vec{q}\cdot\vec{r}_i}$, is isolated, the $\vec{r}$-integral is of the form $\int e^{i(\vec{Q}-\vec{q})\cdot\vec{r}}d\vec{r}$. The $\vec{r}$-integral therefore gives a three-dimensional δ-function, $(2\pi)^3\delta(\vec{Q}-\vec{q})$.

The result from these manipulations with (4.117) and (4.118) is:

$$\begin{aligned} f_{\text{mag}}(\vec{Q},E) &= \sqrt{\frac{k_f}{k_i}}(-\gamma r_0) \\ &\quad \times\vec{\sigma}\cdot\left(\int_{-\infty}^{\infty} e^{i\vec{q}\cdot\vec{r}_i}\,\frac{1}{2\pi^2}\widehat{\vec{q}}\times(\vec{s}\times\widehat{\vec{q}})\,(2\pi)^3\delta(\vec{Q}-\vec{q})\,d\vec{q}\right. \\ &\quad \left. -\frac{i4\pi}{\hbar}\,e^{i\vec{Q}\cdot\vec{r}_i}\,\frac{\widehat{\vec{Q}}\times\vec{p}}{Q}\right). \end{aligned} \tag{4.119}$$

Integrating over the δ-function forces $\vec{q}\longrightarrow\vec{Q}$:

$$\begin{aligned} f_{\text{mag}}(\vec{Q},E) &= \sqrt{\frac{k_f}{k_i}}(-\gamma r_0)\,e^{i\vec{Q}\cdot\vec{r}_i} \\ &\quad \times 4\pi\vec{\sigma}\cdot\left(\widehat{\vec{Q}}\times(\vec{s}\times\widehat{\vec{Q}}) - \frac{i}{\hbar}\,\frac{\widehat{\vec{Q}}\times\vec{p}}{Q}\right). \end{aligned} \tag{4.120}$$

With the definition of $\widetilde{\vec{M}}_\perp(\vec{Q})$, which involves a sum over all electrons in the sample:

$$\widetilde{\vec{M}}_\perp(\vec{Q}) \equiv \sum_{\vec{r}_i} e^{i\vec{Q}\cdot\vec{r}_i}\left(\widehat{\vec{Q}}\times(\vec{s}\times\widehat{\vec{Q}}) - \frac{i}{\hbar}\,\frac{\widehat{\vec{Q}}\times\vec{p}}{Q}\right), \tag{4.121}$$

$$f_{\text{mag}}(\vec{Q},E) = 4\pi\sqrt{\frac{k_f}{k_i}}(-\gamma r_0)\,\vec{\sigma}\cdot\widetilde{\vec{M}}_\perp(\widehat{\vec{Q}})\,. \tag{4.122}$$

Equation (4.122) is a general expression for the magnetic scattering from the spin and orbital moment of the electrons. It includes a sum over the phase

factors for electrons at all $\{\vec{r}_i\}$. Unfortunately, the spin and orbital terms in the large parentheses in (4.121) have different forms, and to see more cearly the vectorial character of magnetic scattering it is convenient to make them equivalent using the expression (proved with some effort in Squires Appendix H.1)

$$\widetilde{\vec{M}}_{\perp L} \equiv \frac{\mathrm{i}}{\hbar Q}\sum_{\vec{r}_i} \mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}_i}(\widehat{\vec{Q}}\times\vec{p}) = \frac{1}{2\mu_\mathrm{B}}\widehat{\vec{Q}}\times\left(\widetilde{\vec{M}}_L(\vec{Q})\times\widehat{\vec{Q}}\right), \tag{4.123}$$

where the Fourier transform of the magnetic form factor from the spatial distribution of electron currents is:

$$\widetilde{\vec{M}}_L(\vec{Q}) \equiv \int \vec{M}_L(\vec{r})\,\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}_i}\mathrm{d}\vec{r}\,. \tag{4.124}$$

It is natural to write (4.121) as:

$$\widetilde{\vec{M}}_\perp \equiv \widetilde{\vec{M}}_{\perp s} + \widetilde{\vec{M}}_{\perp L}\,, \tag{4.125}$$

$$\widetilde{\vec{M}}_\perp = \frac{1}{2\mu_\mathrm{B}}\widehat{\vec{Q}}\times\left(\widetilde{\vec{M}}(\vec{Q})\times\widehat{\vec{Q}}\right), \tag{4.126}$$

where $\widetilde{\vec{M}}(\vec{Q})$ is the Fourier transform of the spatial distribution of all magnetization (as in (4.124)):

$$\widetilde{\vec{M}}(\vec{Q}) \equiv \int \vec{M}(\vec{r})\,\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}_i}\mathrm{d}\vec{r}\,, \tag{4.127}$$

We arrive at the cleaner expression for the magnetic scattering factor

$$f_\mathrm{mag}(\vec{Q},E) = 4\pi\sqrt{\frac{k_\mathrm{f}}{k_\mathrm{i}}}(-\gamma r_0)\sum_{\vec{r}_i}\mathrm{e}^{\mathrm{i}\vec{Q}\cdot\vec{r}_i}\vec{\sigma}\cdot\left(\widehat{\vec{Q}}\times(\widetilde{\vec{M}}\times\widehat{\vec{Q}})\right), \tag{4.128}$$

4.3.2 Vector Orientations in Magnetic Scattering

Equation (4.128) shows that the magnetic scattering is proportional to the vector $\widehat{\vec{Q}}\times\widetilde{\vec{M}}\times\widehat{\vec{Q}}$. The maximum magnetic scattering therefore occurs when the direction of the spin, $\vec{S}$, or magnetization, $\vec{M}(\vec{r})$, is perpendicular to the scattering vector, $\vec{Q}$. This is illustrated in Fig. 4.3.2, which shows intensity contours measured about the forward beam in a small-angle scattering experiment. A magnetic field of 8 kG was applied to the specimen, perpendicular to the direction of the incident beam. This field should was sufficient to saturate the magnetic moment of the sample (a soft magnetic material), aligning all its spins. Notice that the contours are oriented perpendicularly to the direction of the applied magnetic field. The scattering along the direction of the magnetic field is non-zero, however, because Ni-Fe has strong nuclear scattering. By

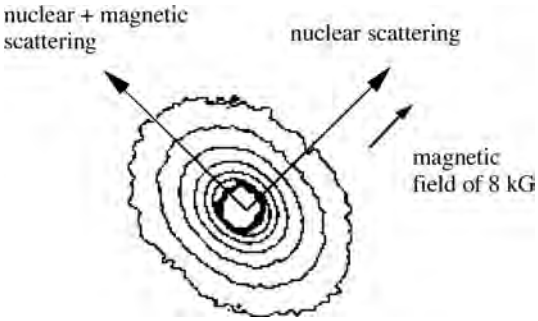


Figure 4.2: Experimental intensity contours from small-angle neutron scattering (SANS) from fcc Ni-Fe in the presence of an 8 kG applied magnetic field. The forward beam was perpendicular to the plane of the paper. The intensity decreases with angle from the forward beam, but more rapidly in the direction of the applied magnetic field.

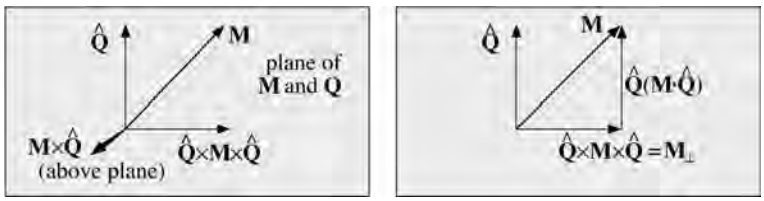


Figure 4.3: Important vectors for magnetic scattering. (**left**) The gray plane, parallel to the paper, is defined by the vectors $\widetilde{\vec{M}}$ and $\widehat{\vec{Q}}$. The vector product $\widetilde{\vec{M}} \times \widehat{\vec{Q}}$ is perpendicular to the plane of the paper and $\widetilde{\vec{M}}_\perp \equiv \widehat{\vec{Q}} \times \widetilde{\vec{M}} \times \widehat{\vec{Q}}$ again lies in the plane. (**right**) The vector of length $\widetilde{\vec{M}} \cdot \widehat{\vec{Q}}$ along the direction $\widehat{\vec{Q}}$.

comparing intensities parallel and perpendicular to the applied magnetic field, it is possible to extract individual profiles for magnetic and nuclear scattering.

The relationship between the generalized magnetization, $\widetilde{\vec{M}}$, its projection, $\widetilde{\vec{M}}_\perp$, and the scattering vector, $\vec{Q}$, is illustrated with the help of Fig. 4.3.2 and its caption. By comparing the two parts of this figure, we find

$$\widehat{\vec{Q}} \times \widetilde{\vec{M}} \times \widehat{\vec{Q}} = \widetilde{\vec{M}} - \widehat{\vec{Q}}(\widetilde{\vec{M}} \cdot \widehat{\vec{Q}}) \; . \tag{4.129}$$

The cross-section for magnetic scattering is proportional to $|f_{\rm mag}|^2$. We need to take the product of $\widetilde{\vec{M}}_\perp$ with its Hermitian adjoint to obtain the intensity

$$\widetilde{\vec{M}}_\perp^\dagger \widetilde{\vec{M}}_\perp = \left(\widetilde{\vec{M}}^\dagger - \widehat{\vec{Q}}(\widetilde{\vec{M}}^\dagger \cdot \widehat{\vec{Q}})\right)\left(\widetilde{\vec{M}} - \widehat{\vec{Q}}(\widetilde{\vec{M}} \cdot \widehat{\vec{Q}})\right) \; . \tag{4.130}$$

When distributing the product in (4.130), the two middle terms have the same form but opposite sign to the fourth term, so

$$\widetilde{\vec{M}}_{\perp}^{\dagger}\widetilde{\vec{M}}_{\perp} = \widetilde{\vec{M}}^{\dagger}\widetilde{\vec{M}} - (\widetilde{\vec{M}}^{\dagger}\cdot\widehat{\vec{Q}})(\widetilde{\vec{M}}\cdot\widehat{\vec{Q}})\,. \tag{4.131}$$

We resolve $\widetilde{\vec{M}}$ and $\vec{Q}$ into Cartesian components

$$\widetilde{\vec{M}}_{\perp}^{\dagger}\widetilde{\vec{M}}_{\perp} = \sum_{\alpha,\beta} S_{\alpha}^{\dagger}S_{\alpha} - S_{\alpha}^{\dagger}\widehat{Q}_{\alpha}S_{\beta}\widehat{Q}_{\beta}\,, \tag{4.132}$$

$$\widetilde{\vec{M}}_{\perp}^{\dagger}\widetilde{\vec{M}}_{\perp} = \sum_{\alpha,\beta}\left(\delta_{\alpha\beta} - \widehat{Q}_{\alpha}\widehat{Q}_{\beta}\right)S_{\alpha}^{\dagger}S_{\beta}\,. \tag{4.133}$$

4.3.3 Averaging over Neutron Polarizations

To obtain an experimental cross-section from (4.128), we need to average over the spin orientations of the incident neutrons (orientations of $\vec{\sigma}$). We expect to write the cross-section as $|f_{\mathrm{mag}}|^2$

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} = (\gamma r_{\mathrm{e}})^2\frac{k_f}{k_i}\left|\langle\lambda_f,\sigma_f|\vec{\sigma}\cdot\widetilde{\vec{M}}_{\perp}|\lambda_i,\sigma_i\rangle\right|^2\delta(E_f - E_i + \hbar\omega)\,. \tag{4.134}$$

Compared to nuclear scattering, (4.134) includes additional coordinates in the matrix element, σ_i and σ_f, to account for the change in spin of the neutron after scattering. Again, the λ_i and λ_f refer to states of the scatterer. For magnetic scattering, a change in λ may originate with the creation or annihilation of an excitation of the electron spins. Further progress with these coordinates will require a magnetic dynamics model. The total cross-section requires that we sum over all initial and final states of the neutron:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} = (\gamma r_{\mathrm{e}})^2\frac{k_f}{k_i}\sum_i\sum_f\left|\langle\lambda_f,\sigma_f|\vec{\sigma}\cdot\widetilde{\vec{M}}_{\perp}|\lambda_i,\sigma_i\rangle\right|^2\delta(E_f - E_i + \hbar\omega)\,. \tag{4.135}$$

The coordinates σ_i and σ_f describe to the neutron spin, which we expect to be up or down (sometimes $\uparrow,\downarrow$). Equation (4.134) is the cross-section for one scattering process, but in an experiment we expect to average over the spins of many neutrons. The total cross-section should include weighting functions such as p_{σ_i} to account for different probabilities of initial spin up and spin down neutrons, as for example with polarized beam experiments. The final spin states are assumed unbiased, however, and so have no associated weight function.

The operator in (4.135) can be resolved into its vector components:

$$\vec{\sigma}\cdot\widetilde{\vec{M}}_{\perp} = \sigma_x\widetilde{M}_{\perp x} + \sigma_y\widetilde{M}_{\perp y} + \sigma_z\widetilde{M}_{\perp z}\,. \tag{4.136}$$

The coordinates of σ_α pertain to the neutrons only, and the coordinates of $\widetilde{M}_{\perp\alpha}$ pertain only to the electrons. The products in (4.136) serve to group the factors involving electrons and neutrons, but they separate as for example:

$$\langle\lambda_f,\sigma_f|\sigma_x\widetilde{M}_{\perp x}|\lambda_i,\sigma_i\rangle = \langle\lambda_f|\widetilde{M}_{\perp x}|\lambda_i\rangle\,\langle\sigma_f|\sigma_x|\sigma_i\rangle\ . \tag{4.137}$$

The cross-section of (4.135) can be separated into nine different terms by use of (4.136) and (4.137) (note the sequencing of subscripts x and y):

$$\begin{aligned}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} \;=\; & (\gamma r_\mathrm{e})^2\frac{k_f}{k_i}\sum_i\sum_f \\ & \Big(\langle\lambda_i|\widetilde{M}_{\perp x}|\lambda_f\rangle\,\langle\sigma_i|\sigma_x|\sigma_f\rangle\langle\sigma_f|\sigma_x|\sigma_i\rangle\langle\lambda_f|\widetilde{M}_{\perp x}|\lambda_i\rangle \\ & +\langle\lambda_i|\widetilde{M}_{\perp x}|\lambda_f\rangle\,\langle\sigma_i|\sigma_x|\sigma_f\rangle\langle\sigma_f|\sigma_y|\sigma_i\rangle\langle\lambda_f|\widetilde{M}_{\perp y}|\lambda_i\rangle \\ & +\langle\lambda_i|\widetilde{M}_{\perp y}|\lambda_f\rangle\,\langle\sigma_i|\sigma_y|\sigma_f\rangle\langle\sigma_f|\sigma_x|\sigma_i\rangle\langle\lambda_f|\widetilde{M}_{\perp x}|\lambda_i\rangle+\ldots\Big) \\ & \delta(E_f-E_i+\hbar\omega)\ .\end{aligned} \tag{4.138}$$

By closure, $\sum_f|\sigma_f\rangle\langle\sigma_f| = 1$, there is simplification of the spin factors

$$\begin{aligned}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} \;=\; & (\gamma r_\mathrm{e})^2\frac{k_f}{k_i}\sum_i\sum_f \\ & \Big(\langle\lambda_i|\widetilde{M}_{\perp x}|\lambda_f\rangle\,\langle\sigma_i|\sigma_x\sigma_x|\sigma_i\rangle\langle\lambda_f|\widetilde{M}_{\perp x}|\lambda_i\rangle \\ & +\langle\lambda_i|\widetilde{M}_{\perp x}|\lambda_f\rangle\,\langle\sigma_i|\sigma_x\sigma_y|\sigma_i\rangle\langle\lambda_f|\widetilde{M}_{\perp y}|\lambda_i\rangle \\ & +\langle\lambda_i|\widetilde{M}_{\perp y}|\lambda_f\rangle\,\langle\sigma_i|\sigma_y\sigma_x|\sigma_i\rangle\langle\lambda_f|\widetilde{M}_{\perp x}|\lambda_i\rangle+\ldots\Big) \\ & \delta(E_f-E_i+\hbar\omega)\ .\end{aligned} \tag{4.139}$$

The spin operators have the properties[5]

$$\begin{aligned}\sigma_x|\uparrow\rangle &= |\downarrow\rangle\quad,\quad \sigma_y|\uparrow\rangle = +\mathrm{i}|\downarrow\rangle\quad,\quad \sigma_z|\uparrow\rangle = +|\uparrow\rangle\ , \\ \sigma_x|\downarrow\rangle &= |\uparrow\rangle\quad,\quad \sigma_y|\downarrow\rangle = -\mathrm{i}|\downarrow\rangle\quad,\quad \sigma_z|\downarrow\rangle = -|\downarrow\rangle\ .\end{aligned} \tag{4.140}$$

Equation (4.139) includes terms with the factors $\langle\uparrow|\sigma_x\sigma_y|\uparrow\rangle$ and $\langle\uparrow|\sigma_y\sigma_x|\downarrow\rangle$. Evaluating them with (4.140) gives:

$$\langle\uparrow|\sigma_x\sigma_y|\uparrow\rangle = \langle\uparrow|\sigma_x(+\mathrm{i})|\downarrow\rangle = \langle\uparrow|+\mathrm{i}|\uparrow\rangle = +\mathrm{i}\langle\uparrow|\uparrow\rangle = +\mathrm{i}\ , \tag{4.141}$$

$$\langle\uparrow|\sigma_y\sigma_x|\uparrow\rangle = \langle\uparrow|\sigma_y|\downarrow\rangle = \langle\uparrow|-\mathrm{i}|\uparrow\rangle = -\mathrm{i}\langle\uparrow|\uparrow\rangle = -\mathrm{i}\ , \tag{4.142}$$

where the last result used the normalization $\langle\uparrow|\uparrow\rangle = 1$. It is not surprising that (4.141) and (4.142) give opposite results because the spin operators σ_x and σ_y

[5]These relations are obtained, for example, from explicit forms of the Pauli spin matrices.

do not commute. The consequence is that there is a pairwise cancellation of the six terms in (4.139) that involve the subscripts xy, yx, xz, zx, yz, zy.

For unpolarized neutrons with $|\uparrow\rangle$ and $|\downarrow\rangle$ of equal probabilities, we expect no bias for the three remaining terms of (4.139), which becomes

$$\begin{aligned}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} &= (\gamma r_\mathrm{e})^2\frac{k_f}{k_i}\sum_i\sum_f \\ &\Big(\langle\lambda_i|\widetilde{M}_{\perp x}|\lambda_f\rangle\,\langle\sigma_i|\sigma_x^2|\sigma_i\rangle\langle\lambda_f|\widetilde{M}_{\perp x}|\lambda_i\rangle \\ &+\langle\lambda_i|\widetilde{M}_{\perp y}|\lambda_f\rangle\,\langle\sigma_i|\sigma_y^2|\sigma_i\rangle\langle\lambda_f|\widetilde{M}_{\perp y}|\lambda_i\rangle \\ &+\langle\lambda_i|\widetilde{M}_{\perp z}|\lambda_f\rangle\,\langle\sigma_i|\sigma_z^2|\sigma_i\rangle\langle\lambda_f|\widetilde{M}_{\perp z}|\lambda_i\rangle\Big)\,\delta(E_f-E_i+\hbar\omega)\,. \qquad (4.143)\end{aligned}$$

The three remaining terms have factors $\langle\sigma_i|\sigma_x^2|\sigma_i\rangle$, $\langle\sigma_i|\sigma_y x^2|\sigma_i\rangle$, and $\langle\sigma_i|\sigma_z^2|\sigma_i\rangle$. These terms evaluate to 1 as for example

$$\langle\uparrow|\sigma_x^2|\uparrow\rangle = \langle\uparrow|\sigma_x|\downarrow\rangle = \langle\uparrow|\uparrow\rangle = 1\,. \qquad (4.144)$$

simplifying (4.143)

$$\begin{aligned}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} &= (\gamma r_\mathrm{e})^2\frac{k_f}{k_i}\sum_i\sum_f \\ &\Big(\langle\lambda_i|\widetilde{M}_{\perp x}|\lambda_f\rangle\,\langle\lambda_f|\widetilde{M}_{\perp x}|\lambda_i\rangle \\ &+\langle\lambda_i|\widetilde{M}_{\perp y}|\lambda_f\rangle\,\langle\lambda_f|\widetilde{M}_{\perp y}|\lambda_i\rangle \\ &+\langle\lambda_i|\widetilde{M}_{\perp z}|\lambda_f\rangle\,\langle\lambda_f|\widetilde{M}_{\perp z}|\lambda_i\rangle\Big)\,\delta(E_f-E_i+\hbar\omega)\,. \qquad (4.145)\end{aligned}$$

In (4.145) we recognize that that the three terms in parentheses are the projections onto a set of final states of the three components of the vector $\widetilde{\vec{M}}_\perp$. This can be obtained as $\langle\lambda_f|\widetilde{\vec{M}}_\perp|\lambda_f\rangle\langle\lambda_f|\widetilde{\vec{M}}_\perp|\lambda_i\rangle$. We therefore make use of (4.133) to rewrite (4.145) as

$$\begin{aligned}\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\,\mathrm{d}E} &= (\gamma r_\mathrm{e})^2\frac{k_f}{k_i}\sum_{\alpha,\beta}(\delta_{\alpha\beta}-\widetilde{M}_\alpha\,\widetilde{M}_\beta) \\ &\sum_i\sum_f\langle\lambda_i|\widetilde{M}_{\perp\alpha}|\lambda_f\rangle\,\langle\lambda_f|\widetilde{M}_{\perp\beta}|\lambda_i\rangle\,\delta(E_f-E_i+\hbar\omega)\,. \qquad (4.146)\end{aligned}$$

Further Reading

The contents of the following are described in the Bibliography.

M. Born and K. Wang: *Dynamical Theory of Crystal Lattices* (Oxford Classic series, 1988).

B. Fultz and J. M. Howe: *Transmission Electron Microscopy and Diffractometry of Materials, Fourth Edn.* (Springer–Verlag, Heidelberg, 2013).

S. W. Lovesey: *Theory of Neutron Scattering from Condensed Matter, Vol. 1* (Oxford, 1984).

S. W. Lovesey: *Theory of Neutron Scattering from Condensed Matter, Vol. 2* (Oxford, 1984).

A. A. Maradudin, E. W. Montroll, G. H. Weiss and I. P. Ipatova: *The Theory of Lattice Dynamics in the Harmonic Approximation, Second Edn.* (Academic Press, New York, 1971)

V. F. Sears: *Neutron Optics* (Oxford, 1989).

G. L. Squires: *Introduction to the Theory of Thermal Neutron Scattering* (Dover, Mineola, New York 1996).

Problems

1.) Suppose the scattering factor distribution is a rectangular function in time:

$$f(\tau) = \frac{1}{t} f_0 \quad \text{when } 0 < \tau < t \text{ , and} \tag{4.147}$$

$$f(\tau) = 0 \quad \text{otherwise.} \tag{4.148}$$

Calculate the self-correlation function $G_s(\tau)$, and the scattered intensity, $I(\omega)$.

(*Hint*: See section 9.4.1 in main text. You can sample time in intervals δt if you like.)

2.) Is it possible to have a self-correlation function $G_s(\tau)$ of the form:

$$G_s(\tau) = \frac{1}{t} g_0 \quad \text{when } 0 < \tau < t \text{ , and} \tag{4.149}$$

$$G_s(\tau) = 0 \quad \text{otherwise?} \tag{4.150}$$

(**a**) If yes, give an example. If not, why not?

(**b**) For very long t and short τ the practical situation may be different. If so, how?

3.) Make graphs of the self-correlation functions $G_s(t)$ for the three cases of Figs. 2.2, 2.3, 2.4.

4.) Perhaps with the help of Fig. 2.6,

(**a**) Propose two paths to get from the Van Hove function $G(R, \tau)$ to the diffraction pattern $I(Q)$.

(**b**) Suppose you have $f(r)$ and $f(t)$. Why can you not obtain a rigorous $f(r,t)$ using some combination of $f(r)$ and $f(t)$?

(**c**) Suppose you have $I(Q)$ and $I(\omega)$. Why can you not obtain a rigorous $S(Q,\omega)$ using some combination of $I(Q)$ and $I(\omega)$? List a couple of things you can learn about $S(Q,\omega)$ if you have both $I(Q)$ and $I(\omega)$.

(**d**) Suppose you have $I(Q)$ and $I(\omega)$, and suppose that the scattering is incoherent. How could you construct an estimate of the measured intensity $I(Q,\omega)$?

Chapter 5

Dynamics of Materials and Condensed Matter

5.1 Phonon Thermodynamics

5.1.1 Phonon Statistical Mechanics

Phonons are useful for understanding the collective motions of atoms in crystals because as normal modes of a solid, phonon modes can be treated independently. Independent contributions to thermodynamic functions are then possible, simplifying analyses of the vibrational partition function and vibrational entropy, for example.

Phonons are bosons, so there is no limit to how many of them can be present in each mode, or oscillator. Consider a set of oscillators N, all of the same energy ε, and assume these oscillators can exchange energy. Suppose there are a number m of phonons that are free to distribute between these different oscillators. We calculate the phonon entropy from the equation carved in stone in Fig. 5.1

$$S = k_{\mathrm{B}} \ln \Omega \,, \tag{5.1}$$

where there are a number of ways Ω, to arrange the m phonons in the N oscillators.

Figure 5.2 illustrates five possibilities for distributing 4 phonons among 4 oscillators. The top configuration in Fig. 5.2 has all four phonons in the left oscillator. The bottom configuration places one phonon in each oscillator. The middle three are other possibilities, such as three phonons in the first oscillator, one in the second, and none in the other two.

What is important about Fig. 5.2 is that it inspires the trick for calculating Ω combinatorially. The trick is to consider the three vertical bars and the four circles as distributed at random over the seven positions in each row. The number Ω becomes the number of ways to distribute m indistinguishable

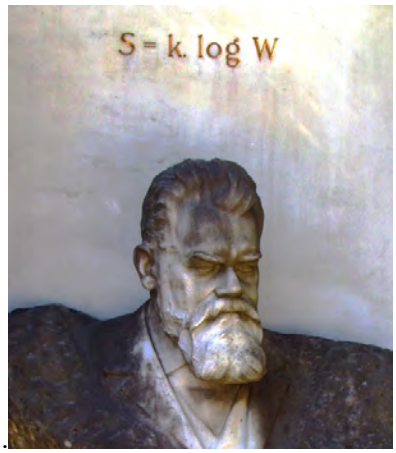


Figure 5.1: The Boltzmann monument in Vienna. The constant k is related to k_B by the factor 2.3026 if log denotes $\log_{10}$. Our notation uses Ω instead of W to avoid conflict with Debye–Waller factors.

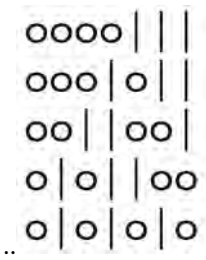


Figure 5.2: Five configurations of placing 4 (m) phonons into 4 (N) oscillators. Vertical lines demark oscillators; circles denote phonons.

phonons over $(N-1)+m$ possible sites (equal to the probability of distributing $N-1$ indistinguishable bars over $(N-1)+m$ possible sites). The result is a binomial coefficient.[1] For systems of many oscillators we can replace $N-1$ with N, of course:

$$\Omega = \frac{(N+m)!}{N!\,m!} \; . \tag{5.2}$$

Substituting (5.2) into (5.1), and using the Stirling approximation of $\ln(x!) \simeq x\ln x - x$

$$S_{\rm osc} = k_{\rm B}\Big[(N+m)\ln(N+m) - N\ln N - m\ln m\Big] \; . \tag{5.3}$$

We now consider the entropy per oscillator, $S_{\rm osc}/N$, and the occupancy of each oscillator, $n \equiv m/N$.

$$\frac{S_{\rm osc}}{N} = k_{\rm B}\Big[(1+n)\ln(N(1+n)) - \ln N - n\ln(nN)\Big] \; , \tag{5.4}$$

$$\frac{S_{\rm osc}}{N} = k_{\rm B}\Big[(1+n)\ln(1+n) + (1+n)\ln N - \ln N - n\ln n - n\ln N\Big] \; , \tag{5.5}$$

$$\frac{S_{\rm osc}}{N} = k_{\rm B}\Big[(1+n)\ln(1+n) + (1+n)\ln N - (1+n)\ln N - n\ln n\Big] \; , \tag{5.6}$$

$$\frac{S_{\rm osc}}{N} = +k_{\rm B}\Big[(1+n)\ln(1+n) - n\ln n\Big] \; . \tag{5.7}$$

It is interesting to compare (5.7) to the analogous equation for fermions, where m electrons are distributed over N states of energy ε, for example. For fermions, the combinatorics give $\Omega = N!\,/\,[(N-m)!\,m!]$. Fermion entropy, incidentally, is the same as the entropy of mixing of atoms on a crystal lattice, were each lattice site can hold only one atom:

$$\frac{S_{\rm mix}}{N} = -k_{\rm B}\Big[(1-c)\ln(1-c) + c\ln c\Big] \; . \tag{5.8}$$

Equation (5.7) and (5.8) have two differences – all signs are reversed, and n can exceed 1, whereas $0 \le c \le 1$.

5.1.2 Phonon Free Energy

With (5.7) for the entropy per oscillator, we construct a Helmholtz free energy for a system of oscillators, where the phonon energy is simply the number of phonons per oscillator times the phonon energy, $n\,\varepsilon$

$$F_{\rm osc} = E_{\rm osc} - TS_{\rm osc} \; , \tag{5.9}$$

$$F_{\rm osc} = n\,\varepsilon - T\,k_{\rm B}\Big[(1+n)\ln(1+n) - n\ln n\Big] \; . \tag{5.10}$$

[1] The first phonon can be placed in $N-1+m$ possible sites, the second has one less, i.e. $N-1+m-1$ possibilities, and the $m^{\rm th}$ phonon can be placed in N ways. The product of these independent probabilities is $(N+m-1)!\,/\,(N-1)!$. There is, however, an overcounting because we put the phonons down individually, labeling the first, second, etc., but the phonons are indistinguishable. The overcounting is the number of ways of distributing m phonons over m sites, which is $m!$, so we divide our result by $m!$ in (5.2).

We seek the equilibrium number of phonons per oscillator at temperature T by calculating the minimum of $F_{\rm osc}$ with respect to n:

$$\begin{aligned}
\frac{\partial F_{\rm osc}}{\partial n} &= 0 \,, && (5.11)\\
0 &= \varepsilon - T\,k_{\rm B}\Big[\,\ln(1+n) + 1 - \ln n - 1\Big] \,, && (5.12)\\
-\frac{\varepsilon}{k_{\rm B}\,T} &= \ln\frac{n}{1+n} \,, && (5.13)\\
{\rm e}^{-\beta\varepsilon} &= \frac{n}{1+n} \,, && (5.14)\\
{\rm e}^{-\beta\varepsilon} &= n(1-{\rm e}^{-\beta\varepsilon}) \,, && (5.15)\\
n(T) &= \frac{1}{{\rm e}^{+\beta\varepsilon}-1} \,, && (5.16)
\end{aligned}$$

where $\beta \equiv (k_{\rm B}\,T)^{-1}$. Equation (5.16) is the Planck distribution for phonons, or the Bose-Einstein distribution (for zero chemical potential). It is a consequence of the statistics of configurations illustrated in Fig. 5.2.

5.1.3 Phonon Entropy and Some Chemical Trends

Here we extend the thermodynamics to a collection of phonon modes with different energies, and we illustrate a handy calculational approach with the partition function. A quantum harmonic oscillator has energies that increase by ε for each phonon, and the partition function for a single harmonic oscillator is

$$\begin{aligned}
\mathcal{Z}_i &= \sum_n^{\infty} {\rm e}^{-\beta(n+1/2)\epsilon_i} \,, && (5.17)\\
\mathcal{Z}_i &= \frac{{\rm e}^{-\beta\epsilon_i/2}}{1-{\rm e}^{-\beta\epsilon_i}} \,, && (5.18)
\end{aligned}$$

where (5.18) was obtained by identifying (5.17) as a geometric series times the constant factor $\exp(-\beta\epsilon_i/2)$. The partition function for a harmonic solid with N atoms and $3N$ independent oscillators is the product of these single oscillator partition functions,

$$\mathcal{Z}_N = \prod_i^{3N} \frac{{\rm e}^{-\beta\epsilon_i/2}}{1-{\rm e}^{-\beta\epsilon_i}} \,, \qquad (5.19)$$

from which we can calculate the phonon free energy as $F = -k_{\rm B}T\ln\mathcal{Z}$,

$$F_{\rm ph} = \frac{1}{2}\sum_i^{3N}\epsilon_i + k_{\rm B}T\sum_i^{3N}\ln\left(1-{\rm e}^{-\beta\epsilon_i}\right) \,, \qquad (5.20)$$

and the phonon entropy by differentiating F with respect to T:

$$S_{\rm ph} = k_{\rm B} \sum_{i}^{3N} \left[-\ln\left(1 - {\rm e}^{-\beta\epsilon_i}\right) + \frac{\beta\epsilon_i}{{\rm e}^{\beta\epsilon_i} - 1} \right] . \tag{5.21}$$

It is often useful to work with a phonon density of states (DOS), $g(\epsilon)$, where $3Ng(\epsilon)\,{\rm d}\epsilon$ phonon modes are in an energy interval ${\rm d}\epsilon$. For a DOS acquired as digital data in m intervals of width $\Delta\epsilon$ (so $\epsilon_j = j\Delta\epsilon$), the partition function can be computed numerically as:

$$\mathcal{Z}_N = \prod_{j=1}^{m} \left(\frac{{\rm e}^{-\beta\epsilon_j/2}}{1 - {\rm e}^{-\beta\epsilon_j}} \right)^{3Ng(\epsilon_j)\Delta\epsilon} . \tag{5.22}$$

Given the phonon DOS, $g(\epsilon)$, using (5.7) it is possible to obtain this useful expression for the phonon entropy of a harmonic material at any temperature

$$S_{\rm ph} \;=\; 3k_{\rm B} \int_0^\infty g(\epsilon) \left[\Big(n(\epsilon)+1\Big) \ln\Big(n(\epsilon)+1\Big) - n(\epsilon) \ln\Big(n(\epsilon)\Big) \right] {\rm d}\epsilon \;, \tag{5.23}$$

where $g(\epsilon)$ is normalized to 1 and $n(\epsilon)$ is the Planck distribution for phonon occupancy at the temperature of interest. A handy expression for the high-temperature limit of the difference in phonon entropy between two harmonic phases, α and β can be obtained readily from (5.21):

$$S_{\rm ph}^{\beta-\alpha} = 3k_{\rm B} \int_0^\infty (g^\alpha(\epsilon) - g^\beta(\epsilon)) \ln(\epsilon)\, {\rm d}\epsilon \;. \tag{5.24}$$

The DOS is central to understanding the phonon contributions to thermodynamics, but $g(\epsilon)$ must be known to high accuracy because small differences in DOS curves are often important. For example, if we apply (5.24) to a case where the phonon DOS curves of the α and β phases have the same shape, but differ in energy scaling by 10%, we obtain a change in phonon entropy of $\Delta S_{\rm ph} = 3k_{\rm B}\ln(1.1) \simeq 0.3k_{\rm B}$/atom. This change in phonon entropy is almost half of the maximum possible change in configurational entropy of the order-disorder transformation of a binary alloy, which is $\ln 2\, k_{\rm B}$/atom = $0.69\, k_{\rm B}$/atom. When comparing the phonon entropies of different alloy phases, a 1% accuracy in the difference in logarithmic-averaged phonon energy is often required. Although the role of phonon entropy on phase transformations has been discussed for many years [1,2], only in recent times have measurements and calculations become adequate for assessing its thermodynamic importance to phase stability [3-5]. Today it is clear that phonon entropy usually differs enough between alloy phases and compositions to have thermodynamic importance.

At low temperatures it is possible to use electronic structure calculations with the local density approximation to calculate phonon frequencies with reasonable accuracy, but a 1% accuracy remains a challenge. Inelastic neutron

scattering faces its challenges too.A triple axis spectrometer at a reactor neutron source is highly accurate in its energy and momentum measurements, and when a full set of phonon dispersions are measured on single crystals of ordered compounds, the phonon DOS derived from a lattice dynamics model is highly reliable. This technique is not so reliable for disordered solid solutions, unfortunately, because the dispersions are inherently broadened, and this broadening need not be symmetrical in energy. Direct measurements of the full phonon DOS of disordered alloys are possible with chopper spectrometers at pulsed neutron sources, but other difficulties arise. In general, the neutron scattering from phonon excitations does not receive an equal contribution from the different elements in an alloy, and the displacement amplitudes of the different elements may be different in different phonons. This causes a "neutron-weight" that is difficult to correct, as discussed in Sect. 2.3.3, although differential measurements are often possible.

Our knowledge about the phonon entropy of alloy phases is being organized today, but some trends are clear, and some correlations between phonon entropy and alloy properties are known [5,6]. The Hume–Rothery rules of alloy thermodynamics are based on the atomic properties of: 1) electronegativity, 2) metallic radius, and 3) electron-to-atom ratio, with the first factors being the most important. For phonon entropy, we should perhaps add 4) atom mass, to this list. It is found that across the periodic table, where atomic weight varies from 1 to 238 and beyond, for matrix having atoms of mass M_0, and solutes having mass M_S, the phonon entropy upon alloying tends to scale with $\ln(M_0/M_S)$ [7]. This is a general consequence of the relationship for an oscillator of $\omega = \sqrt{k/M}$, where the frequency, ω, scales as the inverse root of the mass, M, assuming that the spring constant, k, is in fact a constant. There is a large scatter in this correlation, however, and it is not a reliable one for most alloys. The spring constants, or more specifically the interatomic force constants $\Phi_{\alpha\alpha' l\kappa l'\kappa'}$ (Eq. 5.47), are in fact not constant in alloys. In particular, a larger solute atom will cause a local compression. Since interatomic forces generally become stiffer under compression, alloying with an atom having both a large size and a large mass will produce an uncertain result for the vibrational entropy.

Van de Walle and Ceder [8] proposed a model based upon a bond-stiffness versus bond-length argument. In their model, an atom pair in different local atomic configurations will have different bond stiffnesses, with greater stiffness for shorter interatomic distances. These characteristics seem transferable when atom pairs are in different crystal structures. The model is useful for semi-quantitative arguments, and when calibrated for specific elements by ab-initio calculations, for example, this model can be used for comparing phonon entropies of different alloy phases.

In alloying there are chemical trends of phonon entropy. In a systematic study of transition metal solutes in vanadium, it was found that the phonon entropy had a robust correlation with the difference in electronegativity between the solute atom and the vanadium atom. Results, shown in Fig. 5.3, indicate that this correlation works well for solutes from the 3*d*, 4*d*, and 5*d*

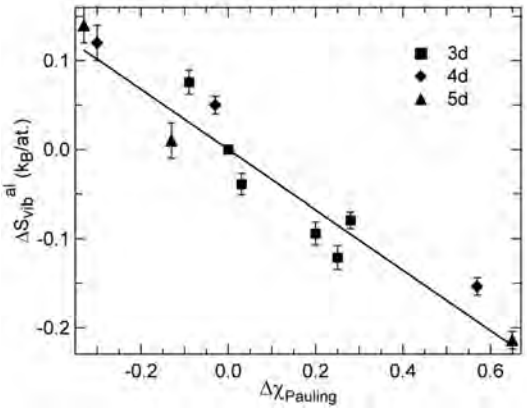


Figure 5.3: Vibrational entropy of alloying of 6 at% solutes in bcc vanadium, plotted versus Pauling's electronegativity of the solute element [9].

..

rows of the periodic table, in spite of their large differences in mass and atomic size. Electronic structure calculations on transition metals have shown that the large differences in electronegativity cause large charge transfers between the solute and its first-nearest-neighbor atoms. These alter the charge screening sufficiently to vary the interatomic forces, as implied by a substantial decrease in electron states at the Fermi level [5, 6, 9, 10]. A simple explanation for transition metals is that with large charge transfers, the bonding tends to become more ionic, leading to fewer metallic-like states near the Fermi level that can become excited in response to atom displacements. The crystal becomes more difficult to distort, and the interatomic forces become stiffer. For alloying of free electron metals, the solute may donate electrons to the band, stiffening it. There is also an effect from hybridizing the occupied local states at the host atoms and the unoccupied local states at the solute, which can cause effects at least as large [11].

5.2 Heat Capacity

5.2.1 Harmonic Heat Capacity

First consider a harmonic model where C_V is obtained with a fixed set of oscillator frequencies, or a fixed phonon DOS. The temperature derivative of the Planck occupancy distribution (Eq. 5.16) with respect to T, weighted by the energy of the mode, ε_i, shows that each mode of energy ε_i contributes to the heat capacity at T the amount

$$C_{V,i}(T) = k_B \left(\frac{\varepsilon_i}{k_B T}\right)^2 \frac{\exp(\varepsilon_i/k_B T)}{(\exp(\varepsilon_i/k_B T) - 1)^2} \,. \tag{5.25}$$

In the high temperature (classical) limit where $k_B T \gg \varepsilon_i$, the heat capacity becomes $1\,k_B$ per mode. [2] Equation 5.25 is consistent with Eq. 5.21 if we integrate over T for all modes

$$S_{V,\mathrm{vib}}(T) = \int_0^T \sum_i^{3N} \frac{C_{V,i}(T')}{T'}\,\mathrm{d}T' \ . \tag{5.26}$$

(Usually we integrate against the phonon DOS, $g(\varepsilon)$, to obtain $S_{V,\mathrm{vib}}(T)$.)

A Debye model predicts some general features of the heat capacity. Figure 5.4a shows heat capacity curves for two materials with similar Debye temperatures. The curves look similar at low T, and are essentially identical at very high T. Their most pronounced difference is found at approximately one-quarter of the Debye temperature. The integrand of Eq. 5.26 has a maximum below about a fifth of the Debye temperature, so experimental studies of vibrational entropy by heat capacity should include this range of temperatures. The difference between these two curves, shown in Fig. 5.4b, is typical of the shape of a differential heat capacity curve obtained from real phonon DOSs, although there are differences in detail. The Debye model assumes a phonon DOS that is a quadratic function of ε, and increases monotonically up to a cutoff energy ε_D. The actual DOS often has a more abrupt rise at low energies that comes from the lower sound velocity of the low transverse modes, which are also weighted at their cutoff energy by the effects of dispersion (e.g., from a high DOS at the Brillouin zone boundary). Plots of ε_D versus T typically show a dip and a rise from a mimimum at roughly a tenth of the Debye temperature $\theta_D = \varepsilon_D/k_B$, but this varies from material to material.

5.2.2 Quasiharmonic Thermodynamics

More typically measured by calorimetry is the amount of heat going into the solid at constant pressure, C_p, defined as

$$C_p(T) = T\frac{\mathrm{d}S}{\mathrm{d}T}\bigg)_p \ . \tag{5.27}$$

The difference between C_p and C_V is a classical thermodynamic relationship

$$C_p - C_V = 9Bv\alpha^2 T \ , \tag{5.28}$$

where B is bulk modulus, α is the linear coefficient of thermal expansion, v is specific volume, and 9 is the square of three dimensions. Equation 5.28 can

[2] Even with phonon softening this could remain true. In the high-temperature limit, with phonon occupancies of $k_B T/\varepsilon_i$ per mode, the softening of the modes gives an increase in occupancy proportional to $1/\varepsilon_i$, but a decrease in energy proportional to ε_i. The product of these two factors produces a heat capacity C_V at high temperatures that is unaffected by mode softening, so C_V can remain as $3k_B$/atom at high temperatures. This is not so if there are other phenomena at work, such as a buildup of elastic energy as a crystal expands.

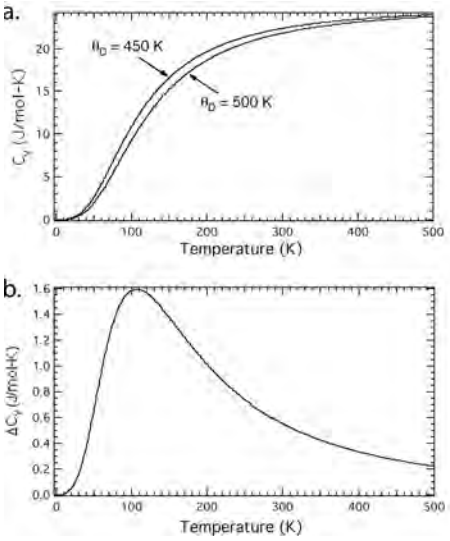


Figure 5.4: (**a**) Heat capacity versus temperature for two Debye solids with Debye temperatures of 450 and 500 K. (**b**) Difference of the two curves in part a.

be derived from formal thermodynamic manipulations, aided for example by techniques of Jacobian transformations.

A physical derivation of Eq. 5.28 can be performed by considering the free energy of a crystal having only 1) phonons, and 2) thermal expansion. There are energy terms from the phonons and from the elastic expansion, and an entropy from the phonons (an average thermal expansion does not generate entropy itself). The free energy is

$$F(T) = E_{\rm elas}(T) + E_{\rm ph}(T) - TS_{\rm ph} \ . \tag{5.29}$$

The elastic energy of thermal expansion, using $\delta V/V_0 = 3\alpha T$ is

$$\begin{aligned} E_{\rm elas} &= \frac{1}{2} B \frac{(\delta V)^2}{V_0} \ , \\ E_{\rm elas} &= \frac{9}{2} B V_0 \, \alpha^2 T^2 \ , \end{aligned} \tag{5.30}$$

The phonon energy in the classical limit is

$$E_{\rm ph}(T) = 3Nk_{\rm B}T \ . \tag{5.31}$$

In the high-temperature limit, the phonon entropy is (cf., Eq. 5.21 for large β)

$$S_{\rm ph}(T) = k_{\rm B} \sum_j^{3N} \ln\left(\frac{k_{\rm B}T}{\hbar\omega_j}\right) \ . \tag{5.32}$$

For small $\delta V/V^0$ (where the superscript 0 denotes a reference value at zero temperature and pressure), we expect a typical phonon frequency ω_j to shift as

$$\omega_j = \omega_j^0(1 - \gamma\delta V/V_0) \,, \tag{5.33}$$

$$\ln(\omega_j) \simeq \ln(\omega_j^0) - \gamma\frac{\delta V}{V^0} \,, \tag{5.34}$$

where γ is the Grüneisen parameter. Using the approach of Eq. 5.34 with and $\delta V/V_0 = 3\alpha T$ for the thermal expansion

$$S_{\rm ph}(T) = k_{\rm B}\sum_j^{3N}\ln\left(\frac{k_{\rm B}T}{\hbar\omega_j^0(1-3\gamma\alpha T)}\right) \,, \tag{5.35}$$

$$S_{\rm ph}(T) = \left[k_{\rm B}\sum_j^{3N}\ln\left(\frac{k_{\rm B}T}{\hbar\omega_j^0}\right)\right] + 9N\gamma\alpha k_{\rm B}T \,, \tag{5.36}$$

$$S_{\rm ph}(T) = S_{\rm ph}^0(T) + 9N\gamma\alpha k_{\rm B}T = S_{\rm ph}^0 + S_{\rm q} \,, \tag{5.37}$$

where the first term, $S_{\rm ph}^0(T)$, is the harmonic phonon entropy with phonon frequencies characteristic of $T = 0$, and unchanged with temperature. Using Eqs. 5.30, 5.31 and 5.37 in the free energy expression for the expanded crystal with phonons, Eq. 5.29

$$F(T) = \frac{9}{2}BV_0\alpha^2T^2 + 3Nk_{\rm B}T - TS_{\rm ph}^0(T) - 9N\alpha\gamma k_{\rm B}T^2 \,. \tag{5.38}$$

First we obtain the equilibrium thermal expansion coefficient, α, by calculating

$$\frac{\partial F(T)}{\partial\alpha} = 0 = 9BV_0\alpha T^2 - 9N\gamma k_{\rm B}T^2 \,,$$

$$\alpha = \frac{C_V\gamma}{3BV_0} \,, \tag{5.39}$$

where we have used the classical result $C_V = 3Nk_{\rm B}$.[3] If the electronic entropy depends on volume, an additional term, $C_{{\rm el}V}\,\gamma_{\rm el}/(3BV_0)$ is added to the right-hand side of Eq. 5.39.

Substituting Eq. 5.39 for the thermal expansion into 5.38 for the free energy

$$F(T) = F^0(T) - \frac{9}{2}BV_0\alpha^2T^2 \,, \tag{5.40}$$

where $F^0(T)$ is the harmonic free energy at finite temperature with zero thermal expansion ($\alpha = 0$). We use Eq. 5.39 to express the quasiharmonic entropy $S_{\rm q}$ of Eq. 5.36 or 5.37 as

$$S_{\rm q} = 9BV_0\alpha^2T \tag{5.41}$$

[3]The effect of phonon energy is small as discussed in Footnote 2, although there is also a small effect from a reduction in zero-point phonon energy as the solid expands.

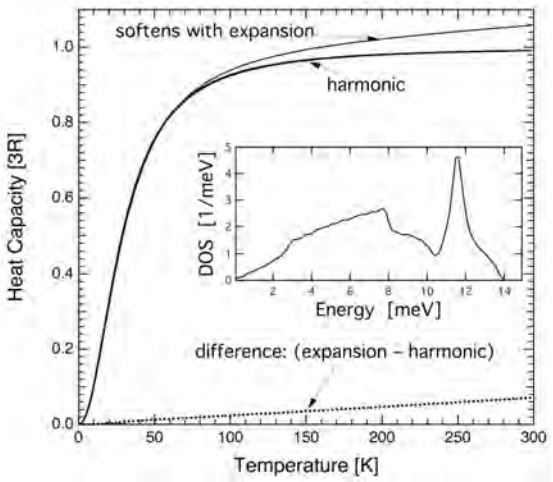


Figure 5.5: Heat capacity versus temperature for trial models of cerium metal. Inset is the phonon DOS at 0 K, and the harmonic heat capacity curve was obtained from this DOS ($3R$ is the classical limit). The elastic energy from thermal expansion was added to the harmonic curve to obtain the quasiharmonic one. Using experimental α and B in Eq. 5.39, a γ was obtained from Eq. 5.39 that predicted the phonon DOS would soften by 6.5 % from 0 to 300 K. (The real heat capacity also includes an electronic contribution.)

(consistent with differentiating Eq. 5.40 with respect to T). Finally, if we substitute Eq. 5.41 into Eq. 5.27, we can recover Eq. 5.28. Typical effects on heat capacity are shown in Fig. 5.5. This S_q of Eq. 5.41 should always be considered when comparing the vibrational thermodynamics of different materials at elevated temperatures.

5.2.3 Anharmonic Heat Capacity

In mechanics, the word "anharmonic" describes any oscillator with generalized forces that deviate from linearity with generalized coordinates. In vibrational thermodynamics, "anharmonic" is used more restrictively. The word "quasiharmonic" accounts for the effects of thermal expansion on phonons just considered (e.g., Eqs. 5.28, 5.41), effects that are not strictly harmonic. An implicit assumption of quasiharmonic theory is that an oscillation has a well-defined, long-lived, harmonic-like frequency that changes modestly and predictably with changes in T and P.

Equation 5.39 is useful for separating anharmonic from quasiharmonic behavior. Suppose we know (by inelastic scattering measurements for example) the thermal softening of the phonon modes, from which we calculate an average Grüneisen parameter γ. With the phonon DOS we can also calculate C_V (or its molar quantity), and conventional measurements can provide B, V_0 (or

molar volume v), and α, accounting for all unknowns in Eq. 5.39. If Eq. 5.39 proves to be true with the γ measured from the phonon softening, it is possible that the solid is "quasiharmonic."[4]

Anharmonic behavior is identified as phonon softening or stiffening that is inconsistent with the Grüneisen parameter γ needed for equality in Eq. 5.39. Anharmonicity involves phenomena beyond those of independent phonons. Higher-order phonon-phonon interactions are the source of anharmonic behavior as discussed in Sect. 5.8.1. Section 5.8.2 discusses electron-phonon interactions, which also vitiate Eq. 5.39.[5] Substantial deviations from quasiharmonic behavior are known for Cr **?**), V **?**), Ce **?**), Mo, W, Si, Ge **?**), and anharmonic or non-harmonic effects could be important in many other systems.

5.2.4 Thermodynamic Entropy

The temperature dependence of the entropy

$$\frac{\mathrm{d}S(T,V)}{\mathrm{d}T} = \frac{\partial S}{\partial T}\bigg)_V + \frac{\partial S}{\partial V}\bigg)_T \frac{\mathrm{d}V}{\mathrm{d}T} \ . \tag{5.42}$$

contains harmonic, quasiharmonic, and anharmonic contributions, along with non-harmonic effects from electron-phonon interactions. For the harmonic part of the vibrational entropy we use Eq. 5.23, with a phonon DOS, $g_0(\varepsilon)$, from low temperature measurement or calculation (ideally $T = 0\,\mathrm{K}$). The harmonic part contributes to the first term on the right-hand-side of Eq. 5.42, i.e., $\partial S/\partial T)_V$. Harmonic phonons undergo no change in frequency with T or V. The temperature dependence of $S_\mathrm{h}(T)$ originates with the Planck occupancy factors $n(\varepsilon, T)$.[6]

Quasiharmonic phonons have frequencies that depend on volume only, and make a contribution through the second term on the right-hand-side of Eq. 5.42. At a fixed volume, however, they behave as harmonic oscillators. Their frequencies can change with temperature, but only because thermal expansion alters the volume of the solid.

Anharmonic entropy, S_a, may contribute to either of the two terms on the right-hand side of Eq. 5.42, but especially the first, a pure temperature-dependence of the vibrational entropy at fixed volume caused by changes in the interatomic force constants with temperature. Consider for example a quartic term in the interatomic potential, which alters the force constants with larger thermal displacements. The quartic term does not affect the thermal expansion, but it does change the phonon frequencies. It causes a type of "phonon-phonon

[4]However, some phonon modes may be anharmonic and others quasiharmonic, or the quasiharmonic modes may have different Grüneisen parameters.

[5]It remains a challenge to differentiate between these sources of non-harmonic behavior **?**), so sometimes the effects of electron-phonon interactions are called "anharmonic." More often "anharmonic" is reserved for phonon-phonon interactions only.

[6]For harmonic solids at high temperatures, differences in the phonon DOS will cause a difference in entropy for different phases, but this difference in entropy does not change with temperature. (At high T, S increases with T by the same amount for each phase.)

interaction," so called because the existence of phonons affects the energy to create new phonons. Other temperature-dependences of force constants at constant volume may arise if the nature of the bonding changes as atoms are displaced further from their equilibrium positions, or perhaps thermal excitations of electrons alter the sensitivity of electronic energy to atom displacements (Section 5.8.2 addresses such effects from adiabatic electron-phonon interactions). The second term in Eq. 5.42 is associated with changes of the interatomic force constants with volume. The quasiharmonic contribution is expected to account for most of it, but the practice of using harmonic entropy expressions, while correct to first order **?**), is not fully verified. The cubic term associated with thermal expansion also shortens phonon lifetimes, causing energy spreads of phonon linewidths. This may not alter the vibrational entropy to first order, but again the robustness of the harmonic approximation is not fully tested.

References

[1] Slater, J.C. (1939) *Introduction to Chemical Physics*, McGraw-Hill, New York, Chapter 13.
[2] Booth, C. and Rowlinson, J.S. (1955) *Trans. Faraday Soc.*, **51**, 463.
[3] Anthony, L., Okamoto, J.K. and Fultz B. (1993) *Phys. Rev. Lett.*, **70**, 1128.
[4] Wallace, D.C. (2003) *Statistical Physics of Crystals and Liquids: A Guide to Highly Accurate Equations of State*, World Scientific, Singapore.
[5] Fultz, B. (2010) "Vibrational Thermodynamics of Materials", *Progress in Materials Science* **55**, 247-352.
[6] Fultz, B. (2014) *Phase Transitions in Materials*, Cambridge University Press, Cambridge.
[7] Bogdanoff, P.D., Fultz, B., Robertson, J.L. and Crow, L. (2002) *Phys. Rev. B* **65**, 014303.
[8] van de Walle, A. and Ceder G. (2002) *Revs. Mod. Phys.* **74**, 11.
[9] Delaire, O., Fultz, B. (2006) *Phys. Rev. Lett.* **97**, 245701.
[10] Delaire, O., Lucas, M.S., Muñoz, J.A., Kresch, M. and Fultz, B. (2008) "Adiabatic Electron-Phonon Interaction and High-Temperature Thermodynamics of the A15 Compounds V_3X", Phys. Rev. Lett. **101**, 105504.
[11] Muñoz, J.A, Lucas, M.S. Mauger, L., Horwath, J., Semiatin, S.I., Xiao, Y., Stone, M.B., Abernathy, D.L. and Fultz, B. (2013) "Electronic structure and vibrational entropies of fcc Au-Fe alloys," Phys. Rev. B **87**, 014301.
[12] Wallace, D.C. (1998) *Thermodynamics of Crystals*, Dover, Mineola, New York.
[13] Trampenau, J., Petry, W., Herzig, C. (1993) *Phys. Rev. B* **47**, 3132.
[14] Bogdanoff, P.D., Fultz, B., Robertson, J.L. and Crow, L. (2002) *Phys. Rev. B* **65**, 014303.
[15] Manley, M.E., McQueeney, R.J., Fultz, B., Osborn, R., Kwei, G.H. and Bogdanoff, P.D. (2002) *Phys. Rev. B* **65**, 144111.

5.3 Lattice Dynamics

5.3.1 Atomic Force-Constants

Since most of the atom mass is in the nucleus, for lattice dynamics we consider the Hamiltonian for nuclear motions in a crystal

$$H_{\rm n} = \sum_{l,\kappa} \frac{\vec{p}^{\,2}_{l\kappa}}{2m_\kappa} + \Phi \ , \tag{5.43}$$

and the characteristics of the atom vibrations require attention to the potential energy, Φ. Following the development of Maradudin, et al. **?**) for a crystal lattice with a basis, the basis vectors are $\{\vec{r}_\kappa, \kappa = 1, 2...\mathcal{R}\}$. Each atom κ, in the unit cell l, is allowed to vibrate about equilibrium with the displacement $\vec{u}_{l\kappa}(t)$. The instantaneous position $\vec{R}_{l\kappa}(t)$ of atom $l\kappa$ at time t is

$$\vec{R}_{l\kappa}(t) = \vec{r}_l + \vec{r}_\kappa + \vec{u}_{l\kappa}(t) \ . \tag{5.44}$$

We use Cartesian components $u_{\alpha l\kappa}$ for the displacement vector, where $\alpha = \{x, y, z\}$. The total potential energy of the crystal, Φ, is a function of the instantaneous positions of all the atoms in the crystal, but it is expanded in a Taylor series of the atomic displacements about $\vec{u}_{l\kappa}(t) = 0$ (i.e., the equilibrium positions of the atoms $\vec{R}_{l\kappa} = \vec{r}_l + \vec{r}_\kappa$)

$$\begin{aligned} \Phi &= \Phi_0 + \sum_{\alpha l\kappa} \Phi_{\alpha l\kappa}\, u_{\alpha l\kappa} \\ &\quad + \frac{1}{2} \sum_{\alpha l\kappa} \sum_{\alpha' l'\kappa'} \Phi_{\alpha\alpha' l\kappa l'\kappa'}\, u_{\alpha l\kappa}\, u_{\alpha' l'\kappa'} + \ldots \ , \end{aligned} \tag{5.45}$$

where the coefficients of the Taylor series are the derivatives of the potential with respect to the displacements:

$$\Phi_{\alpha l\kappa} = \left. \frac{\partial \Phi}{\partial u_{\alpha l\kappa}} \right)_0 , \tag{5.46}$$

$$\Phi_{\alpha\alpha' l\kappa l'\kappa'} = \left. \frac{\partial^2 \Phi}{\partial u_{\alpha l\kappa}\, \partial u_{\alpha' l'\kappa'}} \right)_0 , \tag{5.47}$$

where the subscript zero means that derivatives are evaluated in the equilibrium configuration (all displacements equal to zero) and Φ_0 is the static potential energy of the crystal. Because the force on any atom must vanish in the equilibrium configuration, we have **?**)

$$\Phi_{\alpha l\kappa} = 0 \qquad \forall \ \alpha, l, \kappa \ . \tag{5.48}$$

In the harmonic approximation of lattice dynamics we keep only the terms of the series written explicitly in Eq. 5.45 – we neglect terms of order three and

higher in the displacements

$$H_{\rm n} = \sum_{l\kappa} \frac{\vec{p}^{\,2}_{l\kappa}}{2m_\kappa} + \Phi_0 + \frac{1}{2}\sum_{\alpha l\kappa}\sum_{\alpha' l'\kappa'} \Phi_{\alpha\alpha' l\kappa l'\kappa'}\, u_{\alpha l\kappa}\, u_{\alpha' l'\kappa'} \ , \tag{5.49}$$

where m_κ is the mass of the atom at basis index κ of the unit cell. We rewrite the Hamiltonian in matrix form

$$H_{\rm n} = \sum_{l\kappa} \frac{\vec{p}^{\,2}_{l\kappa}}{2m_\kappa} + \Phi_0 + \frac{1}{2}\sum_{l\kappa}\sum_{l'\kappa'} \vec{u}^{\,T}_{l\kappa}\, \underline{\Phi}_{l\kappa l'\kappa'}\, \vec{u}_{l'\kappa'} \ , \tag{5.50}$$

where a 3×3 force-constant sub-matrix is defined for each atom pair $(l\kappa; l'\kappa')$

$$\underline{\Phi}_{l\kappa l'\kappa'} = [\Phi_{\alpha\alpha' l\kappa l'\kappa'}] \ . \tag{5.51}$$

If $(l,\kappa) \neq (l',\kappa')$, Eq. 5.47 applies. If $(l,\kappa) = (l',\kappa')$, $\Phi_{\alpha\alpha' l\kappa l\kappa}$ is a "self-force constant," derived from the requirement of no overall translation of the crystal

$$\underline{\Phi}_{l\kappa l\kappa} = -\sum_{(l',\kappa')\neq(l,\kappa)} \underline{\Phi}_{l\kappa l'\kappa'} \ . \tag{5.52}$$

Because $-\underline{\Phi}_{l\kappa l'\kappa'}\vec{u}_{l'\kappa'}$ is the force acting upon atom $(l\kappa)$ when atom $(l'\kappa')$ is displaced by $\vec{u}_{l'\kappa'}$, it follows that $\underline{\Phi}_{l\kappa l'\kappa'}$ must be a real symetric matrix:

$$\underline{\Phi}_{l\kappa l'\kappa'} = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix} . \tag{5.53}$$

Any crystal is invariant when translated by a lattice vector, so the force constant matrices must also have the following property:

$$\underline{\Phi}_{l\kappa l'\kappa'} = \underline{\Phi}_{0\kappa(l'-l)\kappa'} = \underline{\Phi}_{(l-l')\kappa 0\kappa'} \tag{5.54}$$

5.3.2 Equations of Motion

In the harmonic approximation, the equations of motion for all nuclei are

$$m_\kappa \ddot{\vec{u}}_{l\kappa}(t) = -\sum_{l',\kappa'} \underline{\Phi}_{l\kappa l'\kappa'}\, \vec{u}_{l'\kappa'}(t) \qquad \forall\ l,\kappa \ . \tag{5.55}$$

In Eq. 5.55 there are $3\times\mathcal{R}\times N^{\rm cell}$ equations of motion to solve for a finite crystal containing $N^{\rm cell}$ unit cells. We seek solutions having the form of plane waves[7] of wavevector $\vec{k}$, angular frequency $\omega_{\vec{k}j}$, and "polarization" $\vec{e}_{\kappa j}(\vec{k})$ (where j is a

[7] In general, the motion of an atom $\vec{u}_{l\kappa}(t)$ will be a sum over many phonons $\vec{u}_{l\kappa\vec{k}j}(t)$.

"branch index" discussed below)

$$\vec{u}_{l\kappa\vec{k}j}(t) = \sqrt{\frac{2\hbar}{N\, m_\kappa\, \omega_{\vec{k}j}}}\; \vec{e}_{\kappa j}(\vec{k})\, \mathrm{e}^{\mathrm{i}(\vec{k}\cdot\vec{r}_l - \omega_{\vec{k}j} t)} \; . \tag{5.56}$$

$$\vec{u}_{l\kappa\vec{k}j}(t) = \hbar\sqrt{\frac{2n(\varepsilon_{\vec{k}j}, T) + 1}{N\, m_\kappa\, \varepsilon_{\vec{k}j}}}\; \vec{e}_{\kappa j}(\vec{k})\, \mathrm{e}^{\mathrm{i}(\vec{k}\cdot\vec{r}_l - \omega_{\vec{k}j} t)} \; , \tag{5.57}$$

where we take the real part to obtain physical displacements. The phase factor, $\mathrm{e}^{\mathrm{i}\vec{k}\cdot\vec{r}_l}$, provides all the long-range spatial modulation of $\vec{u}_{l\kappa\vec{k}j}(t)$. The dependence on κ, a short-range basis vector index, is taken out of the phase factor and placed in the complex constant $\vec{e}_{\kappa j}(\vec{k})$. It is convenient for the $\vec{e}_{\kappa j}(\vec{k})$ of Eq. 5.56 to have modulus unity, as does the exponential. The prefactors are consistent with the quantization of energy for one phonon, $\hbar\omega = m\,\omega^2\langle u^2\rangle = \frac{1}{2}\, m\,\omega^2\, u^2_{\max}$, or for the thermal population of phonons in mode $\varepsilon_{\vec{k}j}$, having energy $E = \varepsilon_{\vec{k}j}[n(\varepsilon_{\vec{k}j}, T) + \frac{1}{2}]$. (For root-mean-squared displacements, delete the $\sqrt{2}$ in Eqs. 5.56 and 5.57.)

We impose periodic boundary conditions on the finite crystal. These require the set of possible wavevectors $\{\vec{k}\}$ to have N^{cell} values, a large number that gives a very fine mesh of $\vec{k}$-points in reciprocal space. With $\mathcal{R}$ atoms in the basis of the unit cell, the crystal has $3 \times \mathcal{R} \times N^{\text{cell}}$ vibrational modes, in agreement with its total number of mechanical degrees of freedom. Each wavevector $\vec{k}$ is associated *a-priori* with $3\times\mathcal{R}$ types of vibrational modes, identified by a branch index, j. Each of the $3\mathcal{R}$ different modes corresponds to a different polarization vector $\vec{e}_{\kappa j}(\vec{k})$ and angular frequency $\omega_{\vec{k}j}$ ($1 \le j \le 3\mathcal{R}$), although degeneracies can be induced by symmetry.

5.3.3 The Eigenvalue Problem for the Polarization Vector

The polarization vector, $\vec{e}_{\kappa j}(\vec{k})$, is a characteristic of each vibrational mode $\vec{k}$, j. The vector $\vec{e}_{\kappa j}(\vec{k})$ contains information on the excursion of each atom κ in the unit cell for the phonon mode mode $\vec{k}$, j. Specifically, it gives 1) the displacement direction of the atom, and 2) its phase lag in time with respect to the other atoms. The vectors $\vec{e}_{\kappa j}(\vec{k})$ for all the atoms in the basis ($1 \le \kappa \le \mathcal{R}$) and their associated angular frequencies $\omega_{\vec{k}j}$ can be calculated by diagonalizing the "dynamical matrix" $\underline{D}(\vec{k})$. The dynamical matrix is obtained by substituting Eq. 5.56 into 5.55. It has the dimensions ($3N\times3N$) and is constructed from (3×3) submatrices $\underline{D}_{\kappa\kappa'}(\vec{k})$

$$\underline{D}(\vec{k}) = \begin{pmatrix} \underline{D}_{11}(\vec{k}) & \cdots & \underline{D}_{1N}(\vec{k}) \\ \vdots & \ddots & \vdots \\ \underline{D}_{N1}(\vec{k}) & \cdots & \underline{D}_{NN}(\vec{k}) \end{pmatrix} . \tag{5.58}$$

Each sub-matrix $\underline{D}_{\kappa\kappa'}(\vec{k})$ is the Fourier transform of the force-constant matrix $\underline{\Phi}_{l\kappa l'\kappa'}$, considered as a function of $(l' - l)$:

$$\underline{D}_{\kappa\kappa'}(\vec{k}) = \frac{1}{\sqrt{m_\kappa m_{\kappa'}}} \sum_{l'} \underline{\Phi}_{0\kappa l'\kappa'} \, \mathrm{e}^{i\vec{k}\cdot(\vec{r}_{l'} - \vec{r}_0)} \, , \tag{5.59}$$

where we set $l = 0$ because the summation is over all l' and the origin is arbitrary. By similarly collecting the polarization vectors into a vector of size $3 \times \mathcal{R}$, we rewrite the system of differential equations (Eq. 5.55) with the plane wave solutions (Eq. 5.56) as an eigenvalue problem:

$$\underline{D}(\vec{k}) \, \vec{e}_j(\vec{k}) = \omega^2_{\vec{k}j} \, \vec{e}_j(\vec{k}) \, , \tag{5.60}$$

where

$$\vec{e}_j(\vec{k}) = \begin{pmatrix} e_{x1j}(\vec{k}) \\ e_{y1j}(\vec{k}) \\ e_{z1j}(\vec{k}) \\ e_{x2j}(\vec{k}) \\ \vdots \\ e_{zNj}(\vec{k}) \end{pmatrix} . \tag{5.61}$$

It can be shown that the $(3\mathcal{R}\times 3\mathcal{R})$ dynamical matrix $\underline{D}(\vec{k})$ is hermitian (for any value of $\vec{k}$). It is fully diagonalizable, and the $\omega^2_{\vec{k}j}$ are real. The $3\mathcal{R}$ eigenvectors and eigenvalues of the dynamical matrix evaluated at a particular wavevector $\vec{k}$ then correspond to the $3\mathcal{R}$ eigenmodes of vibration of the crystal for that wavevector.

5.3.4 Calculation of the Phonon Density of States

To calculate the phonon density of states (DOS) of the crystal, $g(\varepsilon)$, the dynamical matrix is diagonalized at a large number of points in reciprocal space (typically covering the first Brillouin zone). The diagonalization of $\underline{D}(\vec{k})$ at each $\vec{k}$ point returns $3\mathcal{R}$ eigenvalues of angular frequency $\omega_{\vec{k}j}$ $(1 \leq j \leq 3\mathcal{R})$, which are then binned into the DOS histogram.

A phonon partial DOS, $g_d(\varepsilon)$, is a similar quantity, but it gives the spectral distribution of motion by one atom, the species d in the unit cell. Unlike the total DOS, $g(\varepsilon)$, the eigenvalues of the crystal are not weighted equally in the partial DOS

$$g_d(\varepsilon) = \sum_{\vec{k}} \sum_{\alpha\kappa j} \delta_{d\kappa} \, |e_{\alpha\kappa j}(\vec{k})|^2 \, g(\varepsilon) \, , \tag{5.62}$$

where the Kroneker delta is zero unless the atom species d is at the site κ (when $\delta_{dd} = 1$). Because the eigenvalues of the dynamical matrix are normalized for each $\vec{k}$ as

$$\sum_{\alpha\kappa j} |e_{\alpha\kappa j}(\vec{k})|^2 = 1 \ , \tag{5.63}$$

the total DOS is the sum of the partial DOSs of all atoms in the unit cell,

$$g(\varepsilon) = \sum_d g_d(\varepsilon) \ . \tag{5.64}$$

To calculate a partial DOS, for each diagonalization of the dynamical matrix at a specific $\vec{k}$ point, the partial DOS histogram for atom d is incremented at frequency $\omega_{\vec{k}j}$ by the amount $\sum_\alpha e^*_{\alpha dj}(\vec{k})\, e_{\alpha dj}(\vec{k})$. The partial DOS for atom d is large at energies where there are many modes with large displacements of atom d. The histogram for the phonon partial DOSs are computed simultaneously with the total DOS.

5.3.5 Symmetry Constraints on the Force-Constant Matrices

References

[1] A. A. Maradudin, E.W. Montroll, G. H. Weiss and I. P. Ipatova: *The Theory of Lattice Dynamics in the Harmonic Approximation, Second Edn.* (Academic Press, New-York, 1971)
[2] G. Venkataraman, L. A. Feldkamp and V. Sahni: *Dynamics of Perfect Crystals* (MIT Press, Cambridge, 1975)
[3] M. Born and K. Huang: *Dynamical Theory of Crystal Lattices* (Oxford Classic series, 1988)
[4] P. Brüesch: *Phonons: Theory and Experiment I, Lattice Dynamics and Models of Atomic Forces* (Springer-Verlag, Berlin, 1982)
[5] D. C. Wallace: *Thermodynamics of Crystals* (Wiley, 1972)

5.4 Computer Simulations of Lattice Dynamics

5.4.1 Density Functional Theory

A large branch of computational materials science owes its predictive power to developments in density functional theory (DFT), in conjunction with advances in computing power. There are excellent reviews and books on density functional theory and its applications to materials, some of which offer a broader context of condensed matter theory **?**)**?**)**?**). This rich field is far too extensive to cover here. The techniques for calculating phonons from first principles also deserve a much more extensive presentation **?**)**?**)**?**)**?**)**?**)**?**)**?**)**?**). This brief overview of some key methods and concepts is presented with apologies that

numerous intellectual achievements in this rapidly-moving field cannot be covered.

Density functional theory elevates the importance of the electron density with respect to the electron wavefunctions, relying on a theorem that all properties of an electronic system can be obtained from the ground state electron density **?**)**?**). A functional relationship must exist between the density and the physical quantity of interest, but finding this relationship is a challenge. The successful approach developed by Kohn and Sham **?**) assumes that the ground state electron density of a system of interacting electrons can be equal to that of a non-interacting system of electrons, with an interacting electron density. Working with independent electrons makes calculations practical, freeing us from time-ordered-fermion field operators for electron-electron interactions. In particular, the electron exchange and Coulomb correlation energy, which are nonlocal multibody interactions, can be successfully approximated with functionals of the local density, or the density and its local gradients, for example. The density is obtained from the electron wavefunctions, so computational algorithms require iteration to self consistency of the potential and the wavefunctions. The local density functional theory has been surprisingly successful, although problems occur for strongly-correlated electron systems.

5.4.2 First-Principles Phonon Calculations

Principles for calculating phonon dispersions from the electronic energy were outlined in the classic work of Ziman **?**) and others **?**). Two approaches are used. The direct methods impose displacements of atoms, and calculate energies of distorted configurations. In the supercell method, pairs of atoms are displaced by $\vec{u}$ and $\vec{u}'$ along directions of symmetry, for example, and from the energies of the distorted structures, $E(\vec{u},\vec{u}')$, the force constants $\underline{\Phi}_{l\kappa l'\kappa'}$ of Eq. 5.55 can be calculated as

$$\Phi_{\alpha\alpha' l\kappa l'\kappa'} = \left.\frac{\partial^2 E(\vec{u},\vec{u}')}{\partial u_{\alpha l\kappa}\,\partial u_{\alpha' l'\kappa'}}\right)_0 , \tag{5.65}$$

(cf., Eq. 5.47). With these ab-initio force constants, the lattice dynamics can be calculated **?**). The Hellman-Feynman theorem shows that errors in the wavefunctions that cause only second-order errors in energy are responsible for first-order errors in forces, so energy convergence is stringent **?**)**?**). With the frozen phonon method, the potential energy of an individual phonon is calculated directly by generating periodic distortions of a cell or supercell **?**)**?**). Calculations with direct methods are versatile, and can account for unusual structures. Frozen phonons of different amplitudes are a natural way to account for anharmonic potential energies **?**)**?**)**?**), although a good sampling of all phonons is challenging. When periodic boundary conditions are used, the cell should be comfortably larger than the longest-range interatomic force, and the phonon wavevector should be commensurate with the structure. With many atoms and electrons in a large supercell, however, the calculations can become

unwieldy, especially since the computation time often scales as the cube of the number of electrons.

The second approach calculates the sensitivities to small density modulation, usually with a linear response theory **?**)**?**)**?**). These are second derivatives of the energy, and the simplest approach to the problem is with second order perturbation theory. For periodic crystals with both electrons and phonons expressed as Bloch functions, the density susceptibility function for the Fourier coefficient ($\vec{k}$) of the electron density depends on the Fourier coefficient of the phonon modulation ($\vec{k}'$) as

$$\chi(\vec{k},\vec{k}') = 2\sum_i \sum_j \frac{\langle i|e^{-i\vec{k}\cdot\vec{r}}|j\rangle\langle j|e^{i\vec{k}'\cdot\vec{r}}|i\rangle}{\epsilon_i - \epsilon_j} \ , \tag{5.66}$$

where the intermediate electronic states $|j\rangle$ are unoccupied, and the states $|i\rangle$ are occupied. A considerable simplification is that $\vec{k}$ and $\vec{k}'$ must be related by a reciprocal lattice vector, reducing the number of terms in Eq. 5.66, although for more realistic interacting systems, the situation is not so simple. A variational approach for obtaining the energy to second order has been developed to obtain a linear response to periodic perturbations **?**). In this method, the dynamical matrix can be calculated directly **?**). It is also possible to recover real-space force constants by Fourier transformation of the dynamical matrix for appropriate selections of $\vec{k}$.

5.4.3 Molecular Dynamics

Molecular dynamics simulations integrate equations of motion to follow in time the positions and velocities of all atoms. Many molecular dynamics simulations are classical or semi-classical, using forces obtained as gradients of model potential energy functions, for example. The atoms are allowed to move under Newton's law for a short time, and the forces are recalculated for the new configuration of atoms. These simulations can be extended beyond a microcanonical ensemble by allowing contact of a small simulation system with a reservoir, accounting for fluctuations of temperature or pressure, for example. Realistic thermostatic or barostatic control can be achieved by extending the dynamical system with degrees of freedom that represent the external system **?**)**?**). Molecular dynamics is not restricted to harmonic forces, and large-amplitude displacements allow investigations of anharmonic behavior. Accurate force fields that account for all detail of the local atomic environment can be a challenge. Such force fields can be obtained from quantum mechanics, but often they are empirical.

The velocity-velocity autocorrelation function $\langle v(t)|v(0)\rangle$ can be calculated from the trajectory of each atom in a molecular dynamics simulation.[8] The

[8]The notation indicates that the velocity history of each atom is multiplied by the velocity history shifted by t, and integrated over all time.

Fourier transform of the velocity-velocity autocorrelation function gives the phonon DOS **?**)

$$g(\omega) = \int e^{i\omega t} \frac{\langle v(t)|v(0)\rangle}{\langle v(0)|v(0)\rangle} dt \ . \tag{5.67}$$

In another approach, it is possible to obtain differences in thermodynamic quantities such as phonon free energy by parameterizing a Hamiltonian to allow the system to switch from one structure to another, and integrating along the path of this switching parameter.

It is possible to solve for the forces on the moving nuclei by performing a full DFT calculation at each timestep of the simulation, but this is cumbersome. The Car–Parrinello algorithm makes practical quantum molecular dynamics simulations of greater complexity **?**)**?**). It uses the electron degrees of freedom as dynamical variables in a set of coupled equations of motion for the nuclei and the electrons. The nuclei have a classical kinetic energy of $M/2\ [du_{l\kappa}/dt]^2$. The trick to the algorithm is that the electrons are given a fictitious kinetic energy of $\mu/2\ \int [d\psi/dt]^2 d\vec{r}$, with a mass μ that is an adjustable parameter (that helps with the stability of the numerical calculations). With this coupling of nuclear and electronic motions, constrained by the orthogonality of the electron wavefunctions, the electrons stay near their ground states during the simulation. The electron wavefunctions do not need to be recalculated at each step of the simulation, offering an improvement in efficiency. The Car–Parrinello algorithm is an adiabatic approximation.

5.5 Group Theory and Lattice Dynamics

5.5.1 Real Space

The paper by A. A. Maradudin and S. H. Vosko "Symmetry Properties of the Normal Modes of a Crystal", Reviews of Modern Physics **40**, 1-37 (1968), shows how group theory can be used to understand the normal modes of crystal vibrations. The focus of their substantial manuscript is on using the symmetry of the reciprocal lattice to find the eigenvectors of the dynamical matrix and classify them. The degeneracies of the different normal modes are addressed rigorously, and their association with the symmetry elements of the crystal are useful in labeling them. Projection operator methods offer the possibility of finding eigenvectors and eigenvalues without diagonalizing the dynamical matrix.

The paper by J. L. Warren "Further Considerations on the Symmetry Properties of the Normal Modes of a Crystal", Reviews of Modern Physics **40**, 38-76 (1968), also describes the point group of the bond. The bond means the interactions between an atom and its neighbors. The group of the bond refers to the real space symmetries of the interatomic interactions, i.e., how the force constants transform under the point group operations at a central atom.

Consider the space group operator in Seitz notation:

$$\underline{\mathcal{S}} \equiv \{\mathbf{S} \,|\, \vec{v}(S) + \vec{x}(m)\} \,, \tag{5.68}$$

where $\mathbf{S}$ is a rotation operation, $\vec{x}(m)$ is a lattice translation, and $\vec{v}(S)$ is an additional displacement that is required for non-symmorphic crystals (i.e., crystals with screw axes or glide planes).[9] Apply this operator to a displacement vector u (the displacement of an atom off its site during a vibration):

$$u'_\alpha(L,K) = \sum_\beta S_{\alpha\beta} u_\beta(l,\kappa) \,, \tag{5.69}$$

where the operation serves to mix the Cartesian components (denoted subscript α) of the displacement vector for the atom at l,κ, and translates the vector to the atom at L,K. The potential energy of the crystal depends quadratically on the u_α, times a force constant matrix involving the spatial derivatives of the potential along the Cartesian axes of the u_α. The potential energy of the crystal is invariant under symmetry operations of (5.68). This leads to the equality:

$$\Phi_{\mu\nu}(LK, L'K') = \sum_{\alpha\beta} S_{\mu\alpha} S_{\nu\beta} \Phi_{\alpha\beta}(l\kappa, l'\kappa') \,. \tag{5.70}$$

The matrix equation (5.70) must be true for all valid force constant matrices $\Phi_{\alpha\beta}(l\kappa, l'\kappa')$. It is instructive to set $\mathbf{S} = \mathbf{I}$, the identity, so there is no rotation, and only translations are imposed. This can be used to show that force constants can depend only on relative separations between atoms, not absolute positions.

A number of other tests of force constants can be performed with (5.70). Imposing rotations that mix the Cartesian axes can be used to demonstrate that some force constants are equal. Force constants that must equal their negative are found – this result means that the force constant is zero. Equation (5.70) could be handy for determining allowed force constants if a set of space group matricies, $\underline{\mathcal{S}}$, are available.

5.5.2 k-space

Quantum Mechanics

In quantum mechanics, group theory addresses the symmetry of the hamiltonian, H. If an operator, R, commutes with H,

$$RH = HR \,, \tag{5.71}$$

$$\sum_j R_{ij} H_{jk} = \sum_j H_{ij} R_{jk} \,, \tag{5.72}$$

it is convenient to select basis functions for which R is diagonal, so the single remaining terms in the sums are:

$$R_{ii} H_{ik} = H_{ik} R_{kk} \,, \tag{5.73}$$

$$(R_{ii} - R_{kk}) H_{ik} = 0 \,. \tag{5.74}$$

[9] Note that this vector $\vec{v}(S)$ mixes the rotation and translation operations.

Equation (5.74) shows that when $i \neq k$, then the off-diagonal elements $H_{ik} = 0$. *The basis functions for which R is diagonal are the same for which H is diagonal.*

For example, consider the rotational symmetry of a hydrogen atom about the $\widehat{z}$-axis. The Abelian group of rotations has 1-dimensional representations in terms of functions $e^{im\phi}$. Done. Obtaining the ϕ-dependence this way was much easier than solving the Schrödinger equation in spherical coordinates.

Lattice Dynamics

The analogous question for lattice dynamics is, "Can we use symmetry to get the eigenvectors of the dynamical matrix without solving the eigenvalue problem in detail?" The dynamical matrix is:

$$D_{\alpha\beta}(\kappa\kappa'|\vec{k}) = \frac{1}{\sqrt{M_\kappa M_{\kappa'}}} \sum_{l'} \Phi_{\alpha\beta}(l\kappa; l'\kappa') \exp[-i\vec{k} \cdot (\vec{x}(l) - \vec{x}(l'))] \,, \tag{5.75}$$

and the relevant eigenvalue equation is:

$$\sum_{\beta\kappa'} D_{\alpha\beta}(\kappa\kappa'|\vec{k}) u_\beta(\kappa') = \omega^2 u_\alpha(\kappa) \,. \tag{5.76}$$

All functions in (5.76) live in reciprocal space. The entire paper of Maradudin and Vosko is in reciprocal space, or k-space.

Consider the matrix $\mathbf{\Gamma}_{\underline{S}}(\vec{k})$ that performs symmetry operations $\underline{S}$ on the dynamical matrix. Doing two such operations is tricky, since the $\vec{k}$ used in constructing the matrix for the second operation depends on the result of the first operation. Maradudin and Vosko choose instead to work with the space group of a single wavevector $\vec{k}$, and designate the symmetry operations by $\mathbf{R}_i = \{\mathbf{R} \,|\, \vec{v}(S) + \vec{x}(m)\}$. These symmetry operations on $\vec{k}$ produce equivalent wavevectors. In particular, the rotational part of $\mathbf{R}_i$ therefore must generate a $\vec{k}$ that differs only by a reciprocal lattice vector $\vec{g}$:

$$\mathbf{R}\vec{k} = \vec{k} - \vec{g} \,. \tag{5.77}$$

This is quite restrictive, and means that $\vec{g}$ is zero for most vectors $\vec{k}$. The situation is much like the allowed k-vectors for ordered structures in the Landau-Khachaturyan formalism of second-order phase transitions. The operations of (5.77) eliminate the equivalent vectors from the star of $\vec{k}$. Inequivalent $\vec{k}$ can exist only at special points such as the boundaries of Brillouin zones.

So although:

$$\mathbf{D}(\underline{S}\vec{k}) = \mathbf{\Gamma}^{-1}_{\underline{S}}(\vec{k}) \mathbf{D}(\vec{k}) \mathbf{\Gamma}_{\underline{S}}(\vec{k}) \,, \tag{5.78}$$

and therefore in general:

$$\mathbf{D}(\vec{k}) \neq \mathbf{\Gamma}^{-1}_{\underline{S}}(\vec{k}) \mathbf{D}(\vec{k}) \mathbf{\Gamma}_{\underline{S}}(\vec{k}) \,, \tag{5.79}$$

Table 5.1: Elementary Variables Used by Maradudin and Vosko

l, l'	indices of unit cells
κ, κ'	indices of atom in basis
r	number of atoms in basis
$\vec{x}_{l,\kappa}$	atom site in a crystal
$\vec{x}_l$	lattice site in a crystal
$\vec{x}_\kappa$	basis vector
M_κ	mass of atom in basis
α, β	Cartesian indicies $\{x, y, z\}$
$\vec{u}(\kappa)$ and $u_\alpha(\kappa)$, $u_\beta(\kappa)$	atom displacement in a phonon, and Cartesian components
$\Phi_{\alpha\beta}(l\kappa; l'\kappa')$	force constant between two atoms
$\vec{e}(\vec{k}, \sigma, \lambda)$	phonon polarization vector (with components for all κ in unit cell)
σ (or s)	denotes (or index for) a distinct value of ω_j^2
λ (or a)	denotes (or index for) independent eigenvector for each σ or ω_j^2
f_σ	number of eigenvectors for each degenerate energy ($1 \le a \le f_\sigma$)
$\underline{S} \equiv \{\mathbf{S} \mid \vec{v}(S) + \vec{x}(m)\}$	Seitz space group operator
h	order of the group (number of elements)
$\mathbf{S}$	matrix of rotation or improper rotation
$\vec{v}(S)$	little displacement vector for a screw axis operation
$\vec{x}(m)$	lattice translation vector
$\mathbf{\Gamma}_{\underline{S}}(\vec{k}; \{\mathbf{S} \mid \vec{v}(S) + \vec{x}(m)\})$	$3r \times 3r$ unitary matrix of symmetry operator (in recip. space)
$\mathbf{R}_i \equiv \{\mathbf{R} \mid \vec{v}(S) + \vec{x}(m)\}$	symmetry operator for a single wavevector $\vec{k}$

with the restriction to the group of the wavevector $\vec{k}$, the unitary matricies $\mathbf{\Gamma}_{\mathbf{R}_i}$ commute with $\mathbf{D}$:

$$\mathbf{D}(\vec{k}) = \mathbf{\Gamma}^{-1}_{\mathbf{R}_i}(\vec{k})\mathbf{D}(\vec{k})\mathbf{\Gamma}_{\mathbf{R}_i}(\vec{k}) \ . \tag{5.80}$$

These $\{\mathbf{\Gamma}_{\mathbf{R}_i}(\vec{k})\}$ actually form a $3r$-dimensional unitary representation of the space group of the wavevector, $\vec{k}$. They could be used to develop the symmetry properties of the eigenvectors of the dynamical matrix, but this is not the approach of Maradudin and Vosko.

Multiplier Representation

Maradudin and Vosko define a new matrix $\mathbf{T}(\vec{k};\mathbf{R}_i)$, which differs from the $\mathbf{\Gamma}_{\mathbf{R}_i}(\vec{k})$ by an exponential phase factor of modulus unity:

$$\begin{aligned}\mathbf{T}(\vec{k};\{\mathbf{R}_i \,|\, \vec{v}(S) + \vec{x}(m)\}) \quad &= \quad \exp[\mathrm{i}\vec{k}\cdot(\vec{v}(S) + \vec{x}(m))] \\ &\quad \times\mathbf{\Gamma}_{\mathbf{R}_i}(\vec{k};\{\mathbf{R} \,|\, \vec{v}(S) + \vec{x}(m)\}) \ ,\end{aligned} \tag{5.81}$$

This choice of $\mathbf{T}(\vec{k};\mathbf{R}_i)$ is somewhat unusual because these matricies do not form a group in the usual sense – the product of two of them is not an element of the group:

$$\mathbf{T}(\vec{k};\mathbf{R}_i)\mathbf{T}(\vec{k};\mathbf{R}_j) = \exp[\mathrm{i}\vec{g}\cdot\vec{v}(\mathbf{R}_j)]\mathbf{T}(\vec{k};\mathbf{R}_i\mathbf{R}_j) \ . \tag{5.82}$$

(where the product of exponentials of two cases of (5.81) were simplified by noting the restrictive condition (5.77) on the allowed $\mathbf{R}$, giving the condition $\vec{g}\cdot\vec{x}(m) = 2\pi m$, and exponential phase factors of +1). The phase factor $\exp[\mathrm{i}\vec{g}\cdot\vec{v}(S)]$ can differ from +1, *although it differs from +1 only for crystals with screw axes AND special k-points*.[10]

Matrices $\{\mathbf{T}(\vec{k};\mathbf{R}_i\}$ that obey a multiplication rule of (5.82) are said to form a "multiplier representation" of the group of $\{\mathbf{R}_i\}$. The phase factor is the "multiplier." For most crystals this is +1, and we are back to ordinary group representations. Even if the multiplier is complex, however, if we can get the eigenvectors of these $\mathbf{T}$, we can get the eigenvectors of the dynamical matrix, $\mathbf{D}$.

We rewrite (5.76):

$$\mathbf{D}(\vec{k})\vec{e}(\vec{k}\sigma\lambda) = \omega^2_\sigma\vec{e}(\vec{k}\sigma\lambda) \ , \quad \lambda = 1,2,3\ldots f_\sigma \ , \tag{5.83}$$

noting that:

$$\mathbf{D}(\vec{k})[\mathbf{T}(\vec{k};\mathbf{R}_i)\vec{e}(\vec{k}\sigma\lambda)] = \omega^2_\sigma[\mathbf{T}(\vec{k};\mathbf{R}_i)\vec{e}(\vec{k}\sigma\lambda)] \ . \tag{5.84}$$

[10] Is it important to study phonons (or magnons) at special k-points in crystals with screw axes (non-symmorphic crystals)?

The $\{\mathbf{T}\}$ mix the degenerate $\vec{e}(\vec{k}\sigma\lambda)$ (i.e., they mix those λ that go with the same σ). Showing this more clearly, we write:

$$\mathbf{T}(\vec{k};\mathbf{R}_i)\vec{e}(\vec{k}\sigma\lambda) = \sum_{\lambda'}^{f_\sigma} \boldsymbol{\tau}^{(\sigma)}_{\lambda'\lambda}(\vec{k};\mathbf{R}_i)\vec{e}(\vec{k}\sigma\lambda') \ , \tag{5.85}$$

for every $\mathbf{R}_i$ in the group of the wavevector $\vec{k}$. The f_σ functions transform among themselves under $\mathbf{T}(\vec{k};\mathbf{R}_i)$. Furthermore, from (5.80) we can show that $\mathbf{T}(\vec{k};\mathbf{R}_i)$ commutes with $\mathbf{D}$. These conditions are sufficient to show that the $\{\boldsymbol{\tau}^{(\sigma)}(\vec{k};\mathbf{R}_i)\}$ provide an f_σ-dimensional *irreducible* multiplier representation of the point group of the wavevector $\vec{k}$.

Now make a vector of all eigenvectors:

$$\vec{\mathbf{e}}(\vec{k}) = \left[\vec{e}(\vec{k}\sigma_1\lambda_1), \vec{e}(\vec{k}\sigma_1\lambda_2), \ldots \vec{e}(\vec{k}\sigma_2\lambda_1), \vec{e}(\vec{k}\sigma_2\lambda_2), \ldots\right] , \tag{5.86}$$

which transforms as:

$$\mathbf{T}(\vec{k};\mathbf{R}_i)\vec{\mathbf{e}}(\vec{k}) = \vec{\mathbf{e}}(\vec{k})\boldsymbol{\Delta}(\vec{k};\mathbf{R}_i) \ , \tag{5.87}$$

and the $3r \times 3r$ dimensional matrix $\boldsymbol{\Delta}(\vec{k};\mathbf{R}_i)$ has the block-diagonal form:

$$\boldsymbol{\Delta}(\vec{k};\mathbf{R}_i) = \begin{bmatrix} \boldsymbol{\tau}^{(1)}(\vec{k};\mathbf{R}_i) & 0 & 0 & \ldots \\ & \boldsymbol{\tau}^{(2)}(\vec{k};\mathbf{R}_i) & 0 & \ldots \\ 0 & 0 & \boldsymbol{\tau}^{(3)}(\vec{k};\mathbf{R}_i) & \ldots \\ . & & & \end{bmatrix} . \tag{5.88}$$

The unitary matrices $\boldsymbol{\tau}^{(\sigma)}(\vec{k};\mathbf{R}_i)$ are known for all 230 space groups. In a formal sense, the problem is solved.

Projection Operators

In a practical sense, we can generate the eigenfunctions by projection operator machinery. Maradudin and Vosko show that:

$$\underline{\mathcal{P}}^{(s)}_{\lambda\lambda'}(\vec{k}) = f_s/h \sum_{\mathbf{R}_i}^{h} \boldsymbol{\tau}^{(s)}_{\lambda'\lambda}(\vec{k};\mathbf{R}_i)^* \ \mathbf{T}(\vec{k};\mathbf{R}_i) \tag{5.89}$$

is a projection operator.

The projection operator method is considered standard practice, and is not developed further by Maradudin and Vosko. Perhaps it is useful to review it in the context of group theory in quantum mechanics.

Suppose a set of $\{\phi_\lambda^{(j)}\}$ are partner functions in the representation (j), and transform among themselves under operation of the symmetry operator $\underline{P}_{\mathbf{R}_i}$. Explicitly:

$$\underline{P}_{\mathbf{R}_i}\phi_\kappa^{(j)} = \sum_{\lambda=1}^{l_j} \phi_\lambda^{(j)} \mathbf{\Gamma}_{\lambda\kappa}^{(j)}(\mathbf{R}_i) \; . \tag{5.90}$$

Multiply by the complex conjugates of all representation matrices and sum over all elements in the group:

$$\sum_{\mathbf{R}_i}^{h} \mathbf{\Gamma}_{\lambda'k'}^{(i)}(\mathbf{R}_i)^* \underline{P}_{\mathbf{R}_i}\phi_\kappa^{(j)} = \sum_{\mathbf{R}_i}^{h}\sum_{\lambda=1}^{l_j} \mathbf{\Gamma}_{\lambda'k'}^{(i)}(\mathbf{R}_i)^* \mathbf{\Gamma}_{\lambda k}^{(j)}(\mathbf{R}_i)\phi_\kappa^{(j)} \; , \tag{5.91}$$

$$\sum_{\mathbf{R}_i}^{h} \mathbf{\Gamma}_{\lambda'k'}^{(i)}(\mathbf{R}_i)^* \underline{P}_{\mathbf{R}_i}\phi_\kappa^{(j)} = \delta_{ij}\delta_{\kappa\kappa'}\phi_{\lambda'}^{(j)} \; , \tag{5.92}$$

where the last line was obtained through the Great Orthogonality Theorem of group representation theory. Equation (5.92) shows how a projection operator, $\underline{\mathcal{P}}_{\lambda k}^{(j)}$:

$$\underline{\mathcal{P}}_{\lambda k}^{(j)} = l_j/h \sum_{\mathbf{R}_i} \mathbf{\Gamma}_{\lambda'\kappa'}^{(i)}(\mathbf{R}_i)^* \underline{P}_{\mathbf{R}_i} \; , \tag{5.93}$$

pulls out the function $\phi_{\lambda'}^{(j)}$.

For the eigenvectors of the dynamical matrix for a particular $\vec{k}$, the projection operator of (5.89) does the same thing. Since the matricies $\boldsymbol{\tau}^{(\sigma)}(\vec{k};\mathbf{R}_i)$ are available, one could use projection operator techniques to operate on an arbitrary vector and generate all eigenfunctions of the dynamical matrix. Labels can be assigned to various phonons based on group theory designations. Another step could be to substitute the $\vec{e}(\vec{k}\sigma\lambda)$ into (5.83) and obtain ω_σ^2 by matrix multiplication rather than by matrix inversion. Is this useful, or just elegant? Projection operator methods will reveal all degeneracies of normal modes for a particular $\vec{k}$. In such cases where a set of partner eigenvectors is found, it is not possible to obtain them directly. This requires diagonalizing the dynamical matrix to solve for the eigenvectors, or more accurately a block diagonal form of the dynamical matrix as in (5.88).

Compatibility Relations

The paper by J. L. Warren "Further Considerations on the Symmetry Properties of the Normal Modes of a Crystal", Reviews of Modern Physics **40**, 38-76 (1968), adds to the discussion of Maradudin and Vosko. Compatibility relationships are described in more detail in Warren's paper. Compatibility relationships refer to the new degeneracies that appear when special high symmetry directions in

the Brillouin zone come together at special points, such as at the point Γ at the center of the zone, or at the points {X, W, K, L} on the surface of the zone.

For example, when moving along a line in k-space along the (100) direction, an abrupt change in symmetry occurs when the edge of the zone (X) or center of the zone (Γ) is reached. The source of this new symmetry is the fact that at these special points, translational vectors of the reciprocal lattice cause the points to become equivalent (identical) to other points in the reciprocal lattice.[11] For example, at an arbitrary position along the (100) direction, the symmetry elements are $\{E, C_4^2, 2C_4, 2\sigma_v, 2\sigma_d\}$. This fourfold axis obviously does not have threefold symmetry elements, but the situation changes at the origin, point Γ. The (111) direction, which includes a threefold axis, also converges at Γ. Some representations along the (100) axis are expected to be compatible with some representations along the (111) axis when the two axes meet at the origin. The larger symmetry group at the orgin, point Γ, contains the elements of the threefold and fourfold axes as subgroups. Some representations of the groups from the threefold and fourfold axes are compatible at Γ, while some are not. Those that are compatible at Γ must have eigenvalues that are degenerate in energy. Their associated phonon branches rigorously converge in energy at Γ.

The way to test for compatibility is to inspect the characters, $\chi(\mathbf{R}_i) = \mathrm{Tr}[\Gamma(\mathbf{R}_i)]$, of the representations of the three groups. This can be done with the decomposition formula of group representation theory in exactly the same way as is done for analyzing the degeneracies in crystal field theory. (After all, we are considering only the point group of the wavevector.) The decomposition formula for the characters is:

$$a_i = \frac{1}{h} \sum_{\mathbf{R}_i} \chi(\mathbf{R}_i)^* \chi(\mathbf{R}) \ , \tag{5.94}$$

where a_i is the number of times that a particular representation of the subgroup (e.g., of the threefold axis) appears in the group representation of the special point (e.g., Γ).

Equation (5.94) is most easily used by inspection of character tables.[12] For example, if all characters of a representation along the fourfold axis match the characters of a representation at Γ, these representations are compatible. In a more complicated situation where there is only one $a_i = 1$, the sum of the characters from the representations along the fourfold axis must equal the character of a representation at Γ. For the simplest example, the characters of all symmetry elements in the Δ_1 representation of the group of the fourfold axis are +1. The same symmetry elements have characters +1 for the Γ point for the representation Γ_1. The threefold axis has a different set of symmetry elements than the fourfold axis, but these elements are also +1 for the Γ_1 representation, and are +1 only for the Λ_1 representation. We deduce that the Δ_1

[11] These k-vectors are not identical when they are just slightly displaced off these special points, and related eigenvectors have to be considered independently.

[12] These are found in the classic paper by L. P. Bouckaert, R. Smoluchowski, and E. P. Wigner, Phys. Rev. **50**, 58 (1936), and an excellent description is provided in Chapter 8 of M. Tinkham's book on *Group Theory and Quantum Mechanics*.

representation of the (100) fourfold axis and the Λ_1 representation of the (111) threefold are compatible at the point Γ. These two representations have the same basis functions as Γ_1, but the symmetry operators at Γ interchange these basis functions because they are degenerate in energy. The significance that the basis functions of the Δ_1 and Λ_1 representations must have exactly the same energy at the point Γ in k-space. They are not degenerate just away from the Γ point (unless accidentally, or in the case of no crystal potential), but group theory says that they are rigorously degenerate at Γ. This is a result of potential use in the analysis of phonon data. We can determine the required conditions for the merging of dispersion curves at special point in k-space. (Incidentally, s-type functions are suitable basis functions for the symmetric representations in this simple example.)

5.5.3 Time-reversal symmetry in the dynamical matrices

Time reversal symmetry must be considered in magnetic systems. Consider the spin as a rotation in a particular direction. The spin is flipped upon time reversal. Time reversal symmetry must also be considered in certain cases for non-magnetic systems, and results in additional degeneracies. In most cases time-reversal symmetry does not result in additional degeneracies, however. (Perhaps we could consider additional degeneracies as "accidental," but get the software to flag them so we can inspect for this later.)

When the Hamiltonian is invariant by the time-reversal operation (the direction of momentum, or spin is inverted), the system is said to have the time-reversal symmetry. The system without the external magnetic field has generally the time-reversal symmetry. The time-reversal symmetry often gives the nontrivial physical significance like the additional degeneracy of the system. One of most well-known examples is the Kramers degeneracy, i.e. a system of an odd number of electrons has the extra degeneracy, but an even number does not. Of course, the present ionic lattice system has such symmetry, too. Therefore, it needs to be checked if there occur additional degeneracies or not due to the symmetry, which can be done by investigating the symmetry of the dynamical matrices.

Maradudin and Vosko begins the discussion by considering the additional rotational element $\mathbf{S}\_\vec{k} = -\vec{k}$ in the point group of the crystal. In the same way as $\mathbf{R}$ (in the the point group G_0), with each element $\mathbf{S}\_\mathbf{R}$ (in the coset $\mathbf{S}\_G_0$) we associate a new matrix operator $T(\vec{k}; \mathbf{S}\_\mathbf{R})$, which is defined anti-unitary as it should be. The analysis for $T(\vec{k}; \mathbf{R})$ in section 4.4.3 can be extended straightforwardly to include the anti-unitary matrix operator $T(\vec{k}; \mathbf{A})$ (the element $\mathbf{S}\_\mathbf{R}$ will be denoted by $\mathbf{A}$ hereafter). It is highly desirable to derive the condition for the existence of extra degeneracies due to the time-reversal symmetry, by looking into the linear dependence of the eigenvectors $\vec{e}(\vec{k}sa\lambda)$ and $T(\vec{k}; \mathbf{A})\vec{e}(\vec{k}sa\lambda)$ (it is clear both should be eigenvectors because the system has the time-reversal

symmetry). We define

$$\vec{\bar{e}}(\vec{k}sa\lambda) = T(\vec{k};\mathbf{A})\vec{e}(\vec{k}sa\lambda) \tag{5.95}$$

and consider and compare the transformation properties of $\vec{e}(\vec{k}sa\lambda)$ and $\vec{\bar{e}}(\vec{k}sa\lambda)$ under $T(\vec{k};\mathbf{R})$

$$T(\vec{k};\mathbf{R})\vec{e}(\vec{k}sa\lambda) = \sum_{\lambda'}^{f_s} \tau^{(s)}_{\lambda'\lambda}(\vec{k};\mathbf{R})\vec{e}(\vec{k}sa\lambda') \ , \tag{5.96}$$

$$T(\vec{k};\mathbf{R})\vec{\bar{e}}(\vec{k}sa\lambda) = \sum_{\lambda'}^{f_s} \bar{\tau}^{(s)}_{\lambda'\lambda}(\vec{k};\mathbf{R})\vec{\bar{e}}(\vec{k}sa\lambda') \ , \tag{5.97}$$

where it is found that the irreducible multiplier representations $\{\bar{\tau}^{(s)}(\vec{k};\mathbf{R})\}$ and $\{\tau^{(s)}(\vec{k};\mathbf{R})\}$ belong to the same factor system. They can be either equivalent or inequivalent. In particular, they can be separated into three types depending on the relationship between two irreducible multiplier representations.

First, we think of a case the irreducible multiplier representations $\bar{\tau}^{(s)}$ and $\tau^{(s)}$ are inequivalent, which will be referred to as of the *third* type. In this case, the eigenvectors $\vec{e}(\vec{k}sa\lambda)$ and $\vec{\bar{e}}(\vec{k}sa\lambda)$ are orthogonal and the f_s-fold degeneracy in the dynamical matrix is doubled by $2f_s$ by the time-reversal symmetry.

On the other hand, it is a bit complicated for an alternative situation $\bar{\tau}^{(s)}$ and $\tau^{(s)}$ are equivalent. In the case, two representations are related by a similarity transformation, $\bar{\tau}^{(s)}(\vec{k};\mathbf{R}) = \beta^{-1}\tau^{(s)}(\vec{k};\mathbf{R})\beta$, where β is a unitary matrix. Thanks to the irreducibility of the matrices $\{\tau^{(s)}(\vec{k};\mathbf{R})\}$, the uniqueness up to a phase factor is followed for the matrix β. By Schur's Lemmas, the matrix $\beta\beta^*\tau^{(s)}(\vec{k};\mathbf{A}_0^{-2})$ commuting with all the matrices $\{\tau^{(s)}(\vec{k};\mathbf{R})\}$ of an irreducible multiplier representation of the point group must be proportional to the unit matrix. This results in two cases;

$$\beta\beta^* = \phi(\vec{k};\mathbf{A}_0;\mathbf{A}_0)\tau^{(s)}(\vec{k};\mathbf{A}_0^2) \ , \tag{5.98}$$

$$\beta\beta^* = -\phi(\vec{k};\mathbf{A}_0;\mathbf{A}_0)\tau^{(s)}(\vec{k};\mathbf{A}_0^2) \ , \tag{5.99}$$

where $\phi(\vec{k};\mathbf{A}_0;\mathbf{A}_0)$ is a "multiplier". Further, as is often the case, $\beta\beta^* = \pm 1$ for a suitable element $\mathbf{A}_0$ in $\mathbf{S}\_G_0$. Now we can turn our attention to the problem of liner independence of $\vec{e}(\vec{k}sa\lambda)$ and $\vec{\bar{e}}(\vec{k}sa\lambda)$. For the purpose, it would be easy to investigate the scalar product of two eigenvectors. For the necessary algebra, it is important to note the scalar products of a unitary and an anti-unitary transformation on vectors are distinguished as

$$\langle T(\vec{k};\mathbf{R})\varphi, T(\vec{k};\mathbf{R})\psi\rangle = \langle\varphi,\psi\rangle \ ,$$

$$\langle T(\vec{k};\mathbf{A}_0)\varphi, T(\vec{k};\mathbf{A}_0)\psi\rangle = \langle\psi,\varphi\rangle \ ,$$

and also the properties for the successive transformations under $T(\vec{k};\mathbf{R})$ and $T(\vec{k};\mathbf{A}_0)$ on the eigenvectors $\vec{e}(\vec{k}sa\lambda)$ and $\vec{\tilde{e}}(\vec{k}sa\lambda)$. After a little algebra, it is finally found that, adopting $\beta\beta^* = \pm 1$,

$$\langle\vec{\tilde{e}}(\vec{k}sa\lambda'), \vec{e}(\vec{k}sa\lambda)\rangle = \pm\langle\vec{\tilde{e}}(\vec{k}sa\lambda'), \vec{e}(\vec{k}sa\lambda)\rangle \ , \tag{5.100}$$

where the upper plus sign is corresponding to a case in Eq.(4) and the lower minus sign to a case in Eq.(5), respectively. Clearly, if $\beta\beta^*$ satisfies Eq.(5), then

$$\langle\vec{\tilde{e}}(\vec{k}sa\lambda'), \vec{e}(\vec{k}sa\lambda)\rangle = 0 \ .$$

This means that two eigenvectors are linearly independent and the minimum dimension of the subspace that is invariant under $\{T(\vec{k};\mathbf{R})\}$ is $2f_s$. In the case, the set of eigenvectors transforms according to an irreducible multiplier representation of the *second* type under the symmetry operations. The other case where $\beta\beta^*$ satisfies Eq.(4) cannot say anything about the linear independence of two eigenvectors. If they are linearly dependent, they can differ by at most an arbitrary phase factor and there occurs no additional degeneracy. On the other hand, if they are linearly independent, this can be referred to as an accidental degeneracy. This is a case under the irreducible multiplier representations called the *first* type.

The above discusses the criteria for establishing the type of representation by dealing with the relation between $\{\bar{\tau}^{(s)}(\vec{k};\mathbf{R})\}$ and $\{\tau^{(s)}(\vec{k};\mathbf{R})\}$. Before closing this section, it may be heuristic to introduce the analogous criterion for irreducible multiplier representations. Starting from the orthogonality theorem,

$$\begin{aligned}\sum_{\mathbf{R}} \bar{\tau}^{(s)}_{\mu\mu'}(\vec{k};\mathbf{R})^* \tau^{(s)}_{\nu\nu'}(\vec{k};\mathbf{R}) &= (h/f_s)\beta^{-1^*}_{\mu\nu}\beta^*_{\nu'\mu'} \ , \quad \text{first or second type,} \\ &= 0 \ , \qquad\qquad\quad \text{third type,}\end{aligned}$$

where h is the order of the group G_0. Using the relation between $\bar{\tau}^{(s)}_{\mu\mu'}(\vec{k};\mathbf{R})$ and $\tau^{(s)}_{\mu\mu'}(\vec{k};\mathbf{A}_0^{-1}\mathbf{R}\mathbf{A}_0)$ and Eqs.(4) and (5), we arrive at another interesting criterion

$$\begin{aligned}\sum_{\mathbf{R}} \phi(\vec{k};\mathbf{A}_0\mathbf{R},\mathbf{A}_0\mathbf{R})\tau^{(s)}_{\lambda\lambda'}(\vec{k};\mathbf{A}_0\mathbf{R}\mathbf{A}_0\mathbf{R}) &= +(h/f_s)\delta_{\lambda\lambda'} \ , \quad \text{first type,} \\ &= -(h/f_s)\delta_{\lambda\lambda'} \ , \quad \text{second type,} \\ &= 0 \ , \qquad\qquad \text{third type.}\end{aligned}$$

It is often customary to express the criterion in terms of the characters χ^s, i.e. $\chi^s(\vec{k};\mathbf{R}) = \mathrm{Tr}\tau^s(\vec{k};\mathbf{R})$,

$$\begin{aligned}\sum_{\mathbf{R}} \phi(\vec{k};\mathbf{A}_0\mathbf{R},\mathbf{A}_0\mathbf{R})\chi^{(s)}(\vec{k};\mathbf{A}_0\mathbf{R}\mathbf{A}_0\mathbf{R}) &= +h \ , \quad \text{first type,} \\ &= -h \ , \quad \text{second type,} \\ &= 0 \ , \quad \text{third type.}\end{aligned}$$

The criterion can be also expressed as a sum over the elements $\mathbf{A}_0$ of the coset $\mathbf{S}\_G_0$ instead of $\mathbf{R}$ of the point group G_0

$$\begin{aligned}\sum_{\mathbf{A}_0} \phi(\vec{k};\mathbf{A}_0,\mathbf{A}_0)\chi^s(\vec{k};\mathbf{A}_0^2) &= +h\,, \qquad \text{first type,}\\ &= -h\,, \qquad \text{second type,}\\ &= 0\,, \qquad \text{third type.}\end{aligned}$$

5.5.4 Implementation in *DANSE*

Some results from group theory seem both useful and practical to implement in software. Others are not such good value:

- The point group of the bond can be used to fill out the force constant matrix, to check for inconsistencies of the force constants, and to help in generating them. This should, in principle, be facilitated by the matrix operations that are available as Pythonized routines in *Computational Crystallographic Toolbox* from the Lawrence Berkeley National Lab [?].
- The group of the wavevector can be assesed. If the Brillouin zone has been assessed previously by group theory, we can use the compatibility relationships to determine which dispersions are degenerate in energy at special points. This is important information when assessing the behavior of fuzzy data. Knowing that a convergence of curves is expected puts constraints on data, and allows a larger number of counts to be analyzed simultaneously, improving reliability.
- Time reversal symmetry might be useful to consider in magnetic systems. It seems plausible that one could assess the merging of magnon dispersions at special points, but we need to see how well this has been worked out by theorists. It would probably be too much to actually calculate character decompositions on the computer for individual cases.
- I think that actually reducing the dynamical matrix by symmetry is too much to do. The procedure may be elegant to implement, but the net gain does not seem large enough when compared to brute force diagonalization, which works rather well so far.

5.6 Spin Dynamics in Solids

5.6.1 Spin as a Source of Magnetism

Electrons are the source of magnetism in materials. Although the atomic nuclei also have spins and magnetic moments, nuclear moments are smaller by a factor of 1,000. The electron contributes to magnetism in two separate ways. One is

from orbital motion – the classical origin of magnetism. The Hamiltonian for electron motion in an electromagnetic field is

$$\mathcal{H} = \frac{1}{2m}\left[\vec{p} + \frac{e}{c}\vec{A}(\vec{r})\right]^2 . \tag{5.101}$$

Here we assume that the magnetic field $\vec{H}$ is uniform and take the Coulomb gauge of $\nabla \cdot \vec{A} = 0$, consistent with $\vec{A}(\vec{r}) = \frac{1}{2}(\vec{H} \times \vec{r})$:

The first order term with respect to $\vec{A}(\vec{r})$ gives

$$\mathcal{H}^{(1)} = \frac{e}{2mc}\vec{H} \cdot (\vec{r} \times \vec{p}) \tag{5.102}$$

$$= \mu_B \vec{H} \cdot \vec{L} , \tag{5.103}$$

where $\vec{L} = \vec{r} \times \vec{p}$ (orbital angular momentum) and $\mu_B = e\hbar/2mc$ (Bohr magneton). Incidentally, the second order term in (5.101) is:

$$\mathcal{H}^{(2)} = \frac{e^2}{8mc^2}(\vec{r} \times \vec{H})^2 , \tag{5.104}$$

which is simply the magnetic induction, i.e. Lenz's law. The first order term (5.103) gives the paramagnetism leading to the magnetization parallel to the magnetic field, while the second order term (5.104) gives the diamagnetism leading to the magnetization antiparallel to the magnetic field. For atoms with closed shells (i.e. $\vec{L} = 0$) or superconducting metals, diamagnetism plays the dominant role. Except for such cases, however, paramagnetism is usually more important than diamagnetism.

There is another important contribution to magnetism from the electron besides its orbital motion. Electron spin angular momentum is the essential ingredient of magnetism in most systems having interesting magnetic properties. In metallic systems where the electron wavefunction is well approximated as a plane wave, it can be shown easily that $\vec{L} = 0$. The angular momentum also vanishes in insulators when the force the electron feels in a solid deviates much from the spherical symmetry the electron would feel in an isolated atom. In these cases, an orbital motion of the electron cannot contribute to the magnetism.

Electron spins can be localized or itinerant, and these two cases provide different types of magnetism. Usually, localized spins are found at ions in magnetic insulators and itinerant spins are on conduction electrons in metallic systems, although there are other important magnetic systems with both itinerant and localized spins that show unusual properties. The following sections describe the characteristic magnetic properties of localized spins, itinerant spins, localized spins immersed in itinerant spins, and strongly correlated electrons.

5.6.2 Localized Spins

Consider a system of localized spins, where the spins are arranged periodically on a crystal, and interact with each other through an exchange interaction. To understand effects of the spin-spin exchange interaction in the system, several kinds of spin models are available. These textbook models have a long history. Depending on the spin dimensionality, there are the Heisenberg model, the XY model, and the Ising model or n-state Potts model. The commonly-used Heisenberg model treats three-dimensional spins, the XY model treats two-dimensional spins, and the Ising model or n-state Potts model treats one-dimensional spins. This section discusses magnetic properties of localized spins within the ferromagnetic Heisenberg model.

For an external magnetic field $h\hat{\vec{z}}$, we have the spin interaction Hamiltonian ($J > 0$) on the square lattice[13]

$$\mathcal{H} = -J \sum_{i \neq j} \vec{S}_i \cdot \vec{S}_j - g\mu_B \sum_i h S_i^z \,, \tag{5.105}$$

where the first term gives the spin-spin interaction intrinsic to the Heisenberg model. Now introduce the "mean field" (or "molecular field") approximation. Details of interactions between neighboring spins are ignored, and the Hamiltonian $\mathcal{H}$ is replaced by:

$$\mathcal{H} = -J \sum_{i \neq j} \vec{S}_i \cdot \langle \vec{S}_j \rangle - g\mu_B h \sum_i S_i^z \,. \tag{5.106}$$

The magnetization $\langle M_z \rangle = -g\mu_B \sum_i \langle S_{iz} \rangle = -g\mu_B N \langle S_z \rangle$ can then be obtained through $\langle S_z \rangle = \sum S_z e^{-\beta\mathcal{H}}/Z$, where $Z = \sum e^{-\beta\mathcal{H}}$. The evaluation of $\langle S_z \rangle$ is simple because $\mathcal{H}$ now includes only scalar quantities. Following the definition of the susceptibility $\chi = \lim_{h\to 0} \langle M_z \rangle / h$,

$$\chi = \frac{C}{T - T_C} \,, \tag{5.107}$$

where T_C and C are called the Curie temperature and the Curie constant, respectively. The temperature-dependence of χ in (5.107) is characteristic of a localized spin system.

For understanding dynamical properties of a magnetic system, a starting point is its elementary excitations. For the case of three-dimensional spins in the Heisenberg model, the elementary excitation is known as a spin wave (or magnon). We consider the operator $O_{\vec{q}}$ that creates a specific excitation in the system,

$$[\mathcal{H}, O_{\vec{q}}] = \omega_{\vec{q}} O_{\vec{q}} \,, \tag{5.108}$$

[13] For geometrical reasons, on some lattices the spins cannot have long-range order.

where $\omega_{\vec{q}}$ is the energy of the excitation. The operator $O_{\vec{q}}$ raises the spin wave and provides the spin wave energy. Introducing the Fourier transformation of spin operators, we reexpress the Heisenberg Hamiltonian as:

$$\mathcal{H} = -\sum_{\vec{q}} J(\vec{q})\vec{S}(\vec{q})\cdot\vec{S}(-\vec{q}) \tag{5.109}$$

$$= -\sum_{\vec{q}} J(\vec{q})\left[S_z(\vec{q})S_z(-\vec{q}) + \frac{1}{2}\Big(S_+(\vec{q})S_-(-\vec{q}) + S_-(\vec{q})S_+(-\vec{q})\Big)\right], \tag{5.110}$$

where we introduce $\vec{S}(\vec{q}) = \sum_i \vec{S}_i e^{i\vec{q}\cdot\vec{R}_i}$ and $S_\pm(\vec{q}) = \sum_i S^i_\pm e^{i\vec{q}\cdot\vec{R}_i}$ ($S^i_\pm = S^i_x \pm iS^i_y$). Assuming the ground state is ferromagnetic, the low energy excited states should correspond to states with spins slightly deviated from perfect alignment. $S_-(\vec{q})$ is an operator that performs this role – it creates the spin wave excitation (magnon). The excitation energy of magnon is determined from (5.108):

$$[\mathcal{H}, S_-(\vec{q})] = \omega_{\vec{q}} S_-(\vec{q}), \tag{5.111}$$

which is not, however, exactly solvable and requires an approximation. The most immediate approach is the mean field approximation for $\mathcal{H}$, which yields:

$$\omega_{\vec{q}} = 2\langle S_z\rangle[J(0) - J(\vec{q})] \tag{5.112}$$

$$\approx 2(J/N)\langle S_z\rangle a^2\vec{q}^2, \tag{5.113}$$

where for the three-dimensional cubic lattice:

$$J(\vec{q}) = 2(J/N)[\cos(q_x a) + \cos(q_y a) + \cos(q_z a)]. \tag{5.114}$$

At low temperatures, $\langle S_z\rangle \approx NS$ and the magnon energy is

$$\omega_{\vec{q}} = 2JSa^2\vec{q}^2. \tag{5.115}$$

The magnon energy spectrum is gapless, i.e., $\omega_{\vec{q}} \to 0$ as $\vec{q} \to 0$. The zero temperature magnetization (M_0) then decreases with increasing T as magnons are created. At small T:

$$M(T) = M_0 - g\mu_B \sum_{\vec{q}} n_{\vec{q}}, \tag{5.116}$$

where $n_{\vec{q}}$ is the Bose distribution function for magnons. A very famous result is the decrease of magnetization at low T is:

$$\frac{M(T)}{M_0 - 1} \propto T^{3/2}, \tag{5.117}$$

which is obtained by a simple integration of (5.116).

5.6.3 Itinerant Spins

Both itinerant electrons in metals and localized spins in insulators exhibit interesting magnetism. The magnetization of the system, $\langle M_z \rangle$, in an external magnetic field h is:

$$\langle M_z \rangle = -\mu_B \sum_{\vec{k}} \left[\langle c^\dagger_{\vec{k}\uparrow} c_{\vec{k}\uparrow} \rangle - \langle c^\dagger_{\vec{k}\downarrow} c_{\vec{k}\downarrow} \rangle \right] , \tag{5.118}$$

where $c^\dagger_{\vec{k}\uparrow}$ ($c_{\vec{k}\uparrow}$) is an electron creation (annihilation) operator with a momentum $\vec{k}$ and spin up. In the noninteracting electron gas, the magnetic susceptibility (called the Pauli susceptibility $\chi_{\rm P}$) is:

$$\chi_{\rm P} = 2\mu_B^2 N(\epsilon_F)[1 + O(T/\epsilon_F)^2] , \tag{5.119}$$

where $N(\epsilon_F)$ is the electron density of states at the Fermi level and ϵ_F is the Fermi energy. $\chi_{\rm P}$ is almost constant at low temperatures and has a very weak temperature dependence.

Now consider a ferromagnetic metal. In ferromagnetic metals, the exchange interaction between electrons can raise the ferromagnetic order by overcoming the competition between kinetic energy and exchange energy. The susceptibility, χ_S, depends on the electron exchange interaction $\bar{V}$, called the Stoner susceptibility. It can be obtained by extension of $\chi_{\rm P}$ as:

$$\chi_S = \frac{\chi_{\rm P}}{1 - \frac{1}{2\mu_B^2}\bar{V}\chi_{\rm P}}. \tag{5.120}$$

Noting that $\chi_{\rm P} \approx 2\mu_B^2 N(\epsilon_F)$ at low temperatures, χ_S becomes:

$$\chi_S = 2\mu_B^2 \frac{N(\epsilon_F)}{1 - \bar{V}N(\epsilon_F)} . \tag{5.121}$$

From (5.121), $1 - \bar{V}N(\epsilon_F) \geq 0$ is a ferromagnetic instability condition at $T = 0$, called the "Stoner condition." For transition metals, the d-band is narrow and $N(\epsilon_F)$ can be large, so the Stoner condition is often satisfied (as for Fe, Co, Ni).

The temperature dependence of χ_S is very weak at room temperature, especially compared to the Curie–Weiss susceptibility. This can be easily understood because only electrons near the Fermi level can participate in thermally-driven magnetic excitations. Nevertheless, the temperature dependence of the Stoner susceptibility provides context for a longstanding puzzle of the ferromagnetism observed in Fe ($T_{\rm C}$ =1044 K), Ni (627 K), and Co (1388 K). Putting $\chi_{\rm P} = 2\mu_B^2 N(\epsilon_F)[1 - aT^2]$, and $\chi_S(T_{\rm C}) = \infty$ at the transition temperature $T_{\rm C}$, we obtain for $\chi_S(T)$:

$$\chi_S = \frac{2\mu_B^2/\bar{V}a}{T^2 - T_{\rm C}^2} = \frac{2\mu_B^2/\bar{V}a}{(T + T_{\rm C})(T - T_{\rm C})} , \tag{5.122}$$

and, in particular, near T_C:

$$\chi_S \approx \frac{\mu_B^2/\bar{V}aT_C}{T - T_C} \ . \tag{5.123}$$

The Stoner susceptibility therefore shows a Curie–Weiss behavior near T_C. However, it is well known that ferromagnetic metals like Fe, Ni, and Co show Curie–Weiss behavior over much wider temperature ranges. The observed Curie–Weiss behavior has been one of the most important sources of controversy over whether spins in ferromagnetic transition metals are itinerant or localized. Improving the Stoner spin susceptibility within a picture of itinerant spins remains an open problem.

Dynamical responses of spins in metals are also very different from those in localized spin systems. For conduction electrons, magnetic responses of their spins are directly related to details of the electron band structure. In the noninteracting electron gas, $\chi_0(\vec{q}, \omega)$ governing the magnetic responses of the system is

$$\chi_0(\vec{q}, \omega) = 2\mu_B^2 \sum_{\vec{k}} \frac{n_{\vec{k}} - n_{\vec{k}+\vec{q}}}{\varepsilon_{\vec{k}+\vec{q}} - \varepsilon_{\vec{k}} - \omega} \ , \tag{5.124}$$

where $n_{\vec{k}}$ is the Fermi distribution function and $\varepsilon_{\vec{k}}$ is the band energy at $\vec{k}$. We note that the Pauli susceptibility χ_P corresponds to $\chi_0(\vec{q}, \omega)$ at $\vec{q} = 0$ and $\omega = 0$. In the same way as was found for non-interacting systems, for interacting electron systems in the paramagnetic state, $\chi(\vec{q}, \omega)$ is:

$$\chi(\vec{q}, \omega) = 2\mu_B^2 \frac{F(\vec{q}, \omega)}{1 - \bar{V}(\vec{q})F(\vec{q}, \omega)} \ . \tag{5.125}$$

The susceptibility $\chi(\vec{q}, \omega)$ is directly related to the neutron scattering cross section as:

$$S(\vec{q}, \omega) = \frac{2}{1 - e^{-\beta\omega}} \, \mathrm{Im}\chi(\vec{q}, \omega) \ . \tag{5.126}$$

It is curious that for temperatures well above T_C, neutrons are still scattered magnetically by metallic spins. A further consideration makes this even more curious. In thermal neutron scattering experiments, the neutron velocity v_n is typically smaller than the electron velocity v_F by a factor of 10^{-3}. This means that a neutron interacts with approximately 10^3 electrons as it moves across the magnetic moment of an atom. The average spin of 10^3 electrons will be zero at $T > T_C$. The answer to this puzzle is provided by further analysis of (5.125). Neutrons are scattered by collective motions of electron spins (called spin-fluctuations)[14], not by individual electron spins. More precisely, the paramagnetic scattering of neutrons is caused by fluctuations of the magnetic force

[14]The original discussion of spin-fluctuations is provided by T. Izuyama, D.J. Kim, and R. Kubo, J. Phys. Soc. Jpn. **18**, 1025 (1963).

on neutrons exerted by spins of itinerant electrons. On the other hand, in the ferromagnetic state ($T < T_C$), we need to consider effects from the spin polarization from the splitting of the electron band states into up and down spin states. In this case of itinerant band electrons, the elementary excitations are like the spin wave and Stoner excitation in the ferromagnetic states, obtained from the condition:

$$1 = \bar{V}(\vec{q}) \sum_{\vec{k}} \frac{n_{\vec{k}\uparrow} - n_{\vec{k}+\vec{q}\downarrow}}{\varepsilon_{\vec{k}+\vec{q}\downarrow} - \varepsilon_{\vec{k}\uparrow} - \omega} \,. \tag{5.127}$$

Even though the spin wave is usually discussed in the context of systems with localized spins, it is interesting that spin waves are also excited in the ferromagnetic ordered state of itinerant spins. Like the insulating case, we find $M(T)/M_0 - 1 \propto T^{3/2}$, as in (5.117) for ferromagnetic metals. This experimental finding strongly implies the existence of spin waves in metallic magnetism.

Returning to the problem of the ferromagnetic transition metals Fe, Co, and Ni, we ask again, "*Are these spins itinerant?*" or "*Can Stoner theory properly explain their magnetism?*" The answers are still extremely unclear. There have been attempts to incorporate the electron-phonon interaction within the itinerant picture. These have improved the Stoner suceptibility somewhat and have led to a Curie–Weiss-like behavior with a smaller T_C than obtained from band theory, more consistent with experiment. Nevertheless, many features consistent with local moments are observed experimentally. The observed change in the specific heat of Fe at T_C, for example, is associated with an entropy of order $k_B \ln 2$, as expected for localized electrons. Spin-resolved photoemission experiments show that the exchange splitting does not vanish at $T > T_C$, strong experimental indication of local moments in ferromagnetic transition metals. These controversies demonstrate that a better theory is needed to properly account for electron correlations.[15]

5.6.4 Localized Spins Embedded in Itinerant Spins

A mixed magnetic system with localized spins (or magnetic impurities) embedded in itinerant spins provides many exotic magnetic properties that differ from the properties of from systems having localized or itinerant spins, but showing features of both. Consider two cases:

- Mn atoms in Cu metal, where Mn $(3d)^5$ electrons play the role of a magnetic impurity of $S = 5/2$, and Cu gives a sea of conduction electrons,
- An f-electron system, for example Ce which has a localized f-level and a $5d$ conduction band.

[15] A compromise model that includes both itinerant and localized features with "correlated delocalized electrons" is available, where the magnetic moments at different sites fluctuate in magnitude and direction at finite temperatures.

The governing Hamiltonian[16] for both cases can be written as

$$\mathcal{H} = t \sum_{ij} c^{\dagger}_{i\sigma} c_{j\sigma} + J \sum_{il} \vec{S}_l \cdot \sigma_i, \tag{5.128}$$

where σ is the spin of conduction electrons and $\vec{S}$ is the localized spin. Although both cases can be explained by $\mathcal{H}$, they correspond to different limits of $\mathcal{H}$.

First, for Mn impurities in Cu, we set $t \gg J$, so the Hamiltonian expresses an effective indirect interaction of localized spins mediated by conduction electrons. This interaction is called the RKKY (Ruderman-Kittel-Kasuya-Yosida) interaction:

$$\mathcal{H}_{\mathrm{RKKY}} = J(k_{\mathrm{F}}|\vec{r}_i - \vec{r}_j|)\, \vec{S}_i \cdot \vec{S}_j\ , \tag{5.129}$$

and $J(x)$ has the functional form:

$$J(x) \propto -\frac{\cos x}{x^3} + \frac{\sin x}{x^4}\ . \tag{5.130}$$

The RKKY interaction is also observed in magnetic multilayer systems composed of layers of Fe/Cu, for example, where Cu plays the role of a nonmagnetic metallic spacer. Here it is found that the exchange coupling constant oscillates and has both positive and negative values depending on the thickness of the Cu layer, consistent with the k_{F} for the conduction electron sea in Cu.

The second case is the opposite limit of $t \ll J$, and approaches the Kondo model for heavy fermion systems. In the past decades, very unusual low temperature behaviors have been observed in rare earth metals (e.g., Ce) and actinides (e.g., U). The linear specific heat at low temperatures shows an unusually high coefficient, γ, of order 1 J/mol K^2, in contrast to a 1 mJ/mol K^2 typical of ordinary metals. High values of γ are typical of "heavy fermion" systems.[17] Furthermore, these heavy fermion systems have an electrical resistivity of $\rho_0 + AT^2$ at low T with huge values of A of order $10\mu\Omega\mathrm{cmK}^{-2}$, whereas A is of order $10^{-5}\mu\Omega\mathrm{cmK}^{-2}$ or less for ordinary metals. Heavy fermion systems exhibit another basic and universal magnetic property. Below a characteristic temperature T^*,[18] heavy fermion systems show Fermi liquid behavior with a huge effective mass, and then constant (but very high) Pauli susceptibility. Above T^*, they show a Curie–Weiss susceptibility that originates from the localized f electrons. At T^*, the quasi-particles are screened by conduction electrons and a singlet (nonmagnetic) state is formed, owing to strong electron correlations. Although heavy fermion systems show differences in their properties, the disappearance of a magnetic moment is a common feature.

[16] This is called the "Kondo lattice model" (KLM). Recently, the ferromagnetic KLM ($J < 0$) has been frequently adopted to describe the colossal magnetoresistance in manganese oxides, where J is the Hund coupling.

[17] Values of γ for UPt_3, $CeAl_3$, $CeCu_2Si_2$, UBe_{13} are 0.45 J/mol K^2, 1.6 J/mol K^2, 1.1 J/mol K^2, and 1.1 J/mol K^2, respectively.

[18] T^* is not necessarily same as T_K, which is the usual Kondo temperature for a single Kondo ion, and in most cases we see $T^* < T_K$ owing to the condensation of Kondo ions.

Another unusual property originating with magnetic scattering is the Kondo effect. The Kondo effect is a logarithmic increase of the electrical resistivity of the Kondo system when the temperature is reduced. It is caused by spin flip scatterings of conduction electrons at magnetic impurities.

5.6.5 Strongly Correlated Electrons

Since the discovery of high T_c superconductors,[19] enormous theoretical effort has been made to understand the two-dimensional Hubbard model, which was originally introduced to understand the metal-insulator transition in transition metals. The Hubbard Hamiltonian is:

$$\mathcal{H} = t \sum_{ij} c^\dagger_{i\sigma} c_{j\sigma} + U \sum_i n_{i\uparrow} n_{i\downarrow} \ . \tag{5.131}$$

In high T_c superconductors, $c^\dagger_{i\sigma}(c_{i\sigma})$ is the creation (annihilation) operator of the electron (or the hole) at the highest antibonding orbital (predominantly $d_{x^2-y^2}$) of Cu^{2+} positioned at i in CuO_2 plane. Approproximately, $t \sim 0.1$ eV and $U \sim 1$ eV. It is immediately clear that the energy scale of the Coulomb correlation energy, U, is too high compared to the interesting energy scales, which are around 10 meV for critical temperature of 100 K. The low energy excitation relevant to superconductivity is therefore believed to originate with spins, not with charge, and effort has been made to derive an effective spin Hamiltonian that depends on the hole concentration (doping).

In the undoped case (half-filling), there is one electron at each site and the hopping into the nearest neighbor costs the energy U. In the limit of $t \ll U$, the electrons look localized at each site because of the high energy barrier U. The Hubbard model can be transformed into the Heisenberg spin Hamiltonian:

$$\mathcal{H} = J \sum_{ij} \vec{S}_i \cdot \vec{S}_j \ , \tag{5.132}$$

$$J = 4t^2/U \ . \tag{5.133}$$

In this limit of $t \ll U$, the system acts as an antiferromagnetic insulator.

By introducing a small fraction, x, of holes, the superconducting phase appears. Typical fractions for $Bi_2Sr_2CaCu_2O_{8+x}$ (Bi2212) are $x \sim 0.1$ (this may be underdoped, but overdoping occurs for $x \sim 0.2$). The effective Hamiltonian for a small number of holes has also been reduced from the Hubbard model, and is called the $t-J$ model:

$$\mathcal{H}_{t-J} = t \sum_{ij} a^\dagger_{i\sigma} a_{j\sigma} + J \sum_{ij} \vec{S}_i \cdot \vec{S}_j \ , \tag{5.134}$$

where $a^\dagger_{i\sigma} = c^\dagger_{i\sigma}(1 - n_{i-\sigma})$ and $a_{i\sigma} = c_{i\sigma}(1 - n_{i-\sigma})$. In the limit of half-filling, $\mathcal{H}_{t-J}$ is the Hamiltonian of a Heisenberg antiferromagnet. Research interests are in

[19] Examples of some high T_c materials are $La_{2-x}Sr_xCuO_4$ ($T_c \sim 40$ K), $YBa_2Cu_3O_7$ ($T_c = 92$ K), $Bi_2Sr_2Ca_2Cu_3O_{10}$ ($T_c = 110$ K), and $Tl_2Sr_2Ca_2Cu_3O_{10}$ ($T_c \sim 125$ K).

small deviations from half-filling, where holes move in the antiferromagnetic lattice and counteract antiferromagnetic long-range order. This leads to antiferromagnetic spin-fluctuations peaked at a momentum near $\vec{Q} = (\pi, \pi)$. In the last decades, spin fluctuations have been found to play fundamental roles in high T_c superconductors. In the antiferromagnetic Fermi liquid model, many key properties of superconducting cuprates have been understood by a strong interaction between quasi-particles and spin-fluctuations. For example, the spin-fluctuation model has been successful in explaining transport properties such as electrical resistivity ($\rho \propto T$), the magnitude of T_c, results from nuclear magnetic resonance (NMR) experiments, angle-resolved photoemission spectroscopy (ARPES) experiments, and other experimental results. The detailed phenomenological propagator for spin-fluctuations is:

$$\chi_{\rm sf}(\vec{q}, \omega) = \frac{\chi_Q}{1 + \xi^2(\vec{q} - \vec{Q})^2 - {\rm i}\omega/\omega_{\rm sf}} \,, \tag{5.135}$$

which has proved applicable for interpreting NMR measurements and neutron scattering experiments. Here $\omega_{\rm sf}$ is the spin-fluctuation energy and ξ is the antiferromagnetic correlation length. Through the value of ξ, the spin-fluctuation depends on the hole concentration. In the limit of $\xi \to \infty$, the system in the pure magnon region, i.e. the undoped antiferromagnetic insulator with long-range order. An interaction between quasi-particles and spin-fluctuations is especially strong in underdoped materials. It has been shown recently that strong anisotropies in ARPES data for underdoped Bi2212 originate with a strong coupling between quasi-particles and spin-fluctuations. Incidentally, for other superconducting materials, i.e., $La_{2-x}Sr_xCuO_4$ and $YBa_2Cu_3O_{7-x}$, data from inelastic neutron scattering and measurements of the spin-lattice relaxation rate by NMR indicate that the spin-fluctuations induce an opening of the gap in the spin excitation[20] in the CuO_2 plane, with an energy comparable to the BCS gap. The apparent gap develops well above T_c.

5.7 Simulations of Spin Dynamics

5.7.1 Monte Carlo Method

Section 5.6 introduced different spin systems and their excitations. Several approaches are available for calculating the states and state evolution of these spin systems. Perhaps the most conventional approach, although not a simulation *per se*, is the diagrammatic expansion of the Green's function. This formal approach is not practical for many complex systems, however. Instead, one may use the local density approximation (LDA), or LDA+U to study band magnetism, or the small-size exact diagonalization for more localized spin systems. However, for more complete descriptions of the dynamics of spins, the most suitable and reliable method is often a Monte Carlo simulation. There

[20]The spin gap corresponds to the pseudogap of a small energy scale.

are many variants of Monte Carlo simulations developed and optimized for specific problems. This section explains the Monte Carlo approach for calculating the states and dynamics of classical lattice spin systems. Monte Carlo simulations of classical systems are easier to understand than quantum Monte Carlo simulations, which are mentioned at the end of this section.

A Monte Carlo simulation is a Markovian process. Such processes will reach a steady state of a system that is independent of the initial configuration. Unfortunately, especially at low temperatures, this final state of equilibrium may require a very long time to achieve. If one knows the ground state *a priori*, at low temperatures it may be appropriate to start with the system in a ground state configuration, and allow temperature to produce disorder in this configuration. The ground state is often unknown, however, and this is typical of more complicated systems. It may be possible to start with several candidate ground state configurations, and identify the true ground state structure as the one that does not evolve with time.

An alternative approach is sometimes called "simulated annealing," where the simulation begins with the system in a fully random state characteristic of infinite temperature. Equilibrium at lower temperatures is achieved by gradually reducing the temperature, and allowing the system to relax under the spin-spin interactions. Some delicacy is required for balancing the slowness of cooling with the need to minimize the time of the simulation.

Finally, systems that undergo spontaneous symmetry breaking may pose special problems. Starting from the random spin configuration at $T = \infty$ in the isotropic ferromagnet, one may find that the ordering direction would rotate without any preferred direction even below T_C. In such a case, therefore, one may include an infinitesimal anisotropy in the Hamiltonian to induce the symmetry breaking, or one may start with a symmetry-broken ground state configuration[21] at $T = 0$ and raising the temperature. Without an appled bias, the formation of local domains, each with its own direction of spin alignment, is a common feature. Elimination of the domain boundaries is favorable energetically, but may take a long time in practice.

5.7.2 Spin Updates in Monte Carlo Simulations

Metropolis Algorithm

In the Metropolis algorithm, a new configuration is generated from an existing one by using a transition probability that depends on the energy difference between the two configurations. The state of thermodynamic equilibrium satisfies the detailed balance between two states n and m,

$$P_n W_{n\to m} = P_m W_{m\to n} \ , \tag{5.136}$$

where P_n is the probability of the system being in the state n and $W_{n\to m}$ is the transition rate from $n \to m$. The Metropolis algorithm selects the simplest

[21] For example, by initially aligning all the spins along a certain direction.

choice of the transition rate that is consistent with detailed balance

$$W_{n\to m} = \begin{cases} \exp(-\Delta E/k_B T) & \Delta E > 0 \\ 1 & \Delta E < 0 \,, \end{cases} \tag{5.137}$$

where $\Delta E = E_n - E_m$. The Metropolis algorithm updates one spin at a time in a given configuration at temperature T:

1. select the spin at site i,
2. evaluate ΔE by updating the spin at i,
3. generate a random number η, where $0 < \eta < 1$,
4. accept the updated configuration if $\eta < \mathrm{e}^{-\Delta E/k_\mathrm{B}T}$, or reject otherwise,
5. return to step 1 for a different spin at site $i + 1$.

For the model of continuous spins of $|\mathbf{S}_i| = 1$, oriented in three dimensions, one may update the spin at site i by generating two random numbers η_1 and η_2, such that $\zeta^2 = \zeta_1^2 + \zeta_2^2 < 1$ such that $\zeta_1 = 1 - \eta_1$ and $\zeta_2 = 1 - \eta_2$ (where $0 < \eta_1, \eta_2 < 1$)

$$S_x = 2\zeta_1 \sqrt{1-\zeta^2} \,, \tag{5.138}$$

$$S_y = 2\zeta_2 \sqrt{1-\zeta^2} \,, \tag{5.139}$$

$$S_z = 1 - 2\zeta^2 \,. \tag{5.140}$$

The Metropolis algorithm ensures that the steady state of the system is the actual state of thermodynamic equilibrium. This is proved by assuming that the system is in thermodynamic equilibrium, and then showing that the Metropolis algorithm has the transition rates needed to keep it there. In thermodynamic equilibrium, the probability P_n is the Boltzmann factor normalized by the partition function, $\mathrm{e}^{-E_n/k_\mathrm{B}T}/Z$, a central result of statistical physics. Substituting into (5.136)

$$\mathrm{e}^{-E_n/k_\mathrm{B}T}\, W_{n\to m} = \mathrm{e}^{-E_m/k_\mathrm{B}T}\, W_{m\to n} \,, \tag{5.141}$$

$$\mathrm{e}^{-(E_n-E_m)/k_\mathrm{B}T} = \frac{W_{m\to n}}{W_{n\to m}} \tag{5.142}$$

Note that $W_{n\to m} = 1$ (since the transition $n \to m$ is downhill energetically, it will always occur according to (5.137)). Equation (5.142) becomes:

$$W_{m\to n} = \mathrm{e}^{-(\Delta E)/k_\mathrm{B}T} \,, \tag{5.143}$$

which is the rate used by the Metropolis algorithm in step 4 or (5.137).

All Markovian processes converge to a steady state[22], and the Metropolis algorithm assures us that thermodynamic equilibrium will be achieved between all pairs of spins in the system. It does not ensure that equilibration will occur in a reasonable time, however.

[22] Or a cyclic state in anomalous cases.

5.7.3 Low Temperatures

At low temperatures, a more sophisticated update algorithm is necessary. Most random updates cause a large energy exponent $\Delta E/k_{\rm B}T$, so the Metropolis algorithm will reject most changes to the spin configuration. The equilibration procedure can then become far too slow to be practical. To speed things up, the randomly-selected changes in spin can be made smaller, for example. This can be done by attenuating the random changes of spin by a factor δ ($0 \le \delta \le 1$), i.e. $\Delta\mathbf{S}$ should be replaced by $\delta\Delta\mathbf{S}$. The factor δ can be adjusted so that the acceptance rate is around 50% on the average.

An actual implementation could be based on a parameterized temperature-dependent solid angle, $\Omega(T)$, of spin i with respect to its original orientation $\mathbf{S}_i$. This is begun by defining $\alpha = \tan^{-1} S_i^y/S_i^x$ and $\beta = \cos^{-1} S_i^z$, from which we define the rotation matrices $R_z(\alpha)$ and $R_y(\beta)$:

$$R_z(\alpha) = \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{5.144}$$

$$R_y(\beta) = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix}. \tag{5.145}$$

The updated orientation of spin i, $\mathbf{S}_i'$ is

$$\mathbf{S}_i' = R_z^{-1}(\alpha)\, R_y^{-1}(\beta)\, \mathbf{S}_i'' \,, \tag{5.146}$$

where

$$\mathbf{S}_i'' = (\sin\theta\cos\phi,\ \sin\theta\sin\phi,\ \cos\theta) \,, \tag{5.147}$$

$$\theta = \Omega(T)\ \eta_1 \,, \tag{5.148}$$

$$\phi = 2\pi\ \eta_2 \,. \tag{5.149}$$

A useful form[23] of $\Omega(T)$ may be

$$\Omega(T) = \pi \tanh(\xi T) \,. \tag{5.150}$$

With this form of $\Omega(T)$, we can adjust ξ so that the acceptance rate is around 50%. Note the two extreme limits: $\Omega(T) \to 0$ as $T \to 0$ and $\Omega(T) \to \pi$ as $T \to \infty$.

Overrelaxation Technique

The overrelaxation technique is used in combination with the Metropolis algorithm for improving the rate of relaxation to the equilibrium configuration, especially at low temperatures. Let us assume that we are treating a system of isotropic continuous spins in the Heisenberg model, with the Hamiltonian

[23] The form of $\Omega(T)$ given here is just the simplest example. One may introduce a more complicated form to give a more desirable performance.

$\mathcal{H} = J\sum_{ij}\mathbf{S}_i\cdot\mathbf{S}_j$. In the overrelaxation method, a spin is precessed about the full interaction field, which it can do without any change of energy. When the angle of precession, θ, is as large as π, this alteration of the spin structure can promote quicker changes in the orientation of other spins during subsequent Monte Carlo steps.

At the site i, the full interaction field for $\mathbf{S}_i$ is $\mathbf{S}_{\mathbf{nn}} \equiv J\sum_{j\neq i}\mathbf{S}_j$. The overrelaxed spin, $\mathbf{S}_i'$, is evaluated by successive rotations using rotations by Euler angles. One needs two angles of α and β to define the direction of $\mathbf{S}_{\mathbf{nn}}$:

$$\alpha = \tan^{-1} S^y_{\mathbf{nn}}/S^x_{\mathbf{nn}}\,, \tag{5.151}$$

$$\beta = \cos^{-1} S^z_{\mathbf{nn}}/|\mathbf{S}_{\mathbf{nn}}|\,, \tag{5.152}$$

$$\mathbf{S}_i' = R_z^{-1}(\alpha)\,R_y^{-1}(\beta)\,R_z(\pi)\,R_y(\beta)\,R_z(\alpha)\,\mathbf{S}_i. \tag{5.153}$$

In an actual simulation, a single "hybrid" Monte Carlo update of the spin configuration could include a Metropolis sweep and two overrelaxation sweeps.

Equilibration (Thermalization)

Repeating the hybrid Monte Carlo steps, spins are updated by sweeping the whole spin lattice, eventually reaching the equilibrium configuration at a given T. It is an important problem to minimize the number of Monte Carlo steps needed to obtain the fully relaxed equilibrium state. The optimization differs, however, depending on the specific problem or the simulation conditions such as temperature and lattice size, for example. It is essential to perform calculations of simple quantities like susceptibilities or magnetization, and test if the system is in the equilibrium state. In more advanced validations, one may check the autocorrelation time. Substantial gains in efficiency are often found for hybrid Monte Carlo simulations. For the square lattice Heisenberg model, the nonhybrid Monte Carlo method typically requires at least $O(10^4)$ steps at the temperature range of $O(0.1J)$, whereas the hybrid method requires $O(10^3)$ steps.

5.7.4 Time Evolution of Spins

For a system of spins in the Heisenberg model, the equation of motion for the spin degrees of freedom is

$$\frac{\partial\mathbf{S}_i}{\partial\tau} = -\mathbf{S}_i\times\frac{\partial\mathcal{H}}{\partial\mathbf{S}_i} = -J\mathbf{S}_i\times\sum_{j\neq i}\mathbf{S}_j\,. \tag{5.154}$$

The N spins are coupled to each other, and the time-evolution of the spins is obtained by solving the coupled differential equations. There are many integration algorithms to solve the differential equations. Here we introduce the Suzuki-Trotter decomposition, which is especially suitable for the lattice problem. Taking the example of the simple square lattice with the nearest-neighbor spin coupling, one can divide the lattice into two sublattices $\mathcal{A}$ and $\mathcal{B}$

by checkerboard decomposition. Employing the Suzuki-Trotter decomposition up to $O(\mathrm{d}\tau^5)$,

$$\{\mathbf{S}_i(\tau + \mathrm{d}\tau)\} = \mathrm{e}^{(\mathbf{A}+\mathbf{B})\mathrm{d}\tau}\{\mathbf{S}_i(\tau)\} \ , \tag{5.155}$$

and $\mathrm{e}^{(\mathbf{A}+\mathbf{B})\mathrm{d}\tau}$ is decomposed using

$$\mathrm{e}^{(\mathbf{A}+\mathbf{B})\mathrm{d}\tau} = \prod_{i=1}^{5} \mathrm{e}^{p_i \mathbf{A}\mathrm{d}\tau/2}\, \mathrm{e}^{p_i \mathbf{B}\tau}\, \mathrm{e}^{p_i \mathbf{A}\mathrm{d}\tau/2} + O(\mathrm{d}\tau^5) \ , \tag{5.156}$$

with

$$p_1 = p_2 = p_4 = p_5 = p = 1/(4 - 4^{1/3}) \tag{5.157}$$

and

$$p_3 = 1 - 4p \ . \tag{5.158}$$

Here $\mathbf{A}$ and $\mathbf{B}$ are the rotation generators of the sublattices $\mathcal{A}$ and $\mathcal{B}$, with fixed $\{\mathbf{S}_{i\mathcal{B}}\}$ and $\{\mathbf{S}_{i\mathcal{A}}\}$, respectively.

The following algorithmic explanation may be more clear.

$\mathrm{e}^{\mathbf{A}\delta\tau}\{\mathbf{S}_i\}$: time evolution of sublattice $\mathcal{A}$

for $i = 1$ to N
when $i \in \mathcal{A}$
$\mathbf{S_{nn}} = J \sum_{j \neq i (j \in \mathcal{B})} \mathbf{S}_j$
$\mathrm{e}^{\mathbf{A}\delta\tau}\{\mathbf{S}_i\} = \left[(\mathbf{S_{nn}} \cdot \mathbf{S}_i)\mathbf{S_{nn}}/|\mathbf{S_{nn}}|^2\right](1 - \cos(|\mathbf{S_{nn}}|\delta\tau))$
$+ \mathbf{S}_i \cos(|\mathbf{S_{nn}}|\delta\tau) + [(\mathbf{S_{nn}} \times \mathbf{S}_i)/|\mathbf{S_{nn}}|] \sin(|\mathbf{S_{nn}}|\delta\tau)$

$\mathrm{e}^{\mathbf{B}\delta\tau}\{\mathbf{S}_i\}$: time evolution of sublattice $\mathcal{B}$

for $i = 1$ to N
when $i \in \mathcal{B}$
$\mathbf{S_{nn}} = J \sum_{j \neq i (j \in \mathcal{A})} \mathbf{S}_j$
$\mathrm{e}^{\mathbf{B}\delta\tau}\{\mathbf{S}_i\} = \left[(\mathbf{S_{nn}} \cdot \mathbf{S}_i)\mathbf{S_{nn}}/|\mathbf{S_{nn}}|^2\right](1 - \cos(|\mathbf{S_{nn}}|\delta\tau))$
$+ \mathbf{S}_i \cos(|\mathbf{S_{nn}}|\delta\tau) + [(\mathbf{S_{nn}} \times \mathbf{S}_i)/|\mathbf{S_{nn}}|] \sin(|\mathbf{S_{nn}}|\delta\tau)$

In this algorithm, the time evolution of spins is performed by combinations of $\mathrm{e}^{-\mathbf{A}\delta\tau}$ and $\mathrm{e}^{-\mathbf{B}\delta\tau}$, just as given by the Suzuki-Trotter decomposition.

5.7.5 Observables

Inelastic neutron scattering, especially experiments on single crystal samples, can probe directly the spin-spin correlation function $S(\mathbf{q}, \omega)$ that describes the dynamics of spins

$$S(\mathbf{q}, \omega) = \frac{1}{2\pi}\frac{1}{N^2} \sum_{ij} \mathrm{e}^{\mathrm{i}\mathbf{q}\cdot(\mathbf{r}_i - \mathbf{r}_j)} \int_0^{\tau_{\max}} \mathrm{e}^{\mathrm{i}\omega\tau} \langle \mathbf{S}_i(\tau) \cdot \mathbf{S}_j(0) \rangle \, \mathrm{d}\tau \ . \tag{5.159}$$

This $S(\mathbf{q},\omega)$ can be calculated by Monte Carlo simulation. $S(\mathbf{q},\omega)$ is obtained by a simple Fourier transformation of the time-dependent correlation function

$$C_{ij}(\tau) = \langle \mathbf{S}_i(\tau) \cdot \mathbf{S}_j(0) \rangle \,, \tag{5.160}$$

which is readily evaluated in a Monte Carlo simulation. Pole structures of $S(\mathbf{q},\omega)$, $\omega(\mathbf{q})$, can provide information on magnetic excitations and relaxations, which are fundamental to understanding the spin systems. Incidentally, the quasi-elastic response $S(\mathbf{q},0)$ tells us the kinds of magnetic fluctuations (correlations) that dominate at a given T. For instance, if $S(\mathbf{Q},0) \gg S(\mathbf{0},0)$ with say $\mathbf{Q} = (\pi,\pi)$, we conclude that antiferromagnetic exchange interactions are dominant over ferromagnetic ones at the given T.

For powder or polycrystal samples, the measured spectra are in the angle-integrated form

$$S(|\mathbf{q}|,\omega) = \frac{2}{N^2} \sum_{ij} \frac{\sin(|\mathbf{q}||\mathbf{r}_i - \mathbf{r}_j|)}{|\mathbf{q}||\mathbf{r}_i - \mathbf{r}_j|} \int_0^{\tau_{\max}} \mathrm{e}^{\mathrm{i}\omega\tau} \langle \mathbf{S}_i(\tau) \cdot \mathbf{S}_j(0) \rangle \, \mathrm{d}\tau \,. \tag{5.161}$$

In some experiments, or in some stages of data analysis, scattered neutrons with a rather wide distribution of momenta are collected into a single energy bin. For such data, the resulting spectra $S(\omega)$ probes the local responses of spins as

$$S(\omega) = \int S(\mathbf{q},\omega) \, \mathrm{d}\mathbf{q} = \frac{1}{2\pi} \frac{1}{N} \sum_i \int_0^{\tau_{\max}} \mathrm{e}^{\mathrm{i}\omega\tau} \langle \mathbf{S}_i(\tau) \cdot \mathbf{S}_i(0) \rangle \, \mathrm{d}\tau \,. \tag{5.162}$$

The quasi-elastic local response, $S(0)$, is also an interesting quantity. It is directly related to the spin-lattice relaxation rate T_1 of a local probe as in nuclear magnetic resonance (NMR) or Mössbauer spectrometry.

5.7.6 Comments on Quantum Monte Carlo Simulations

An essential difference between classical and quantum Monte Carlo simulations is in how the spin configuration is updated, that is, how the Metropolis algorithm is implemented. The classical Monte Carlo method is easier in that the equilibrium probability has the proportionality

$$P(\{\mathbf{S}_i\}) \propto \mathrm{e}^{-\beta \mathcal{H}(\{\mathbf{S}_i\})} \tag{5.163}$$

for the particular configuration $\{\mathbf{S}_i\}$, that is: $\mathcal{H}(\{\mathbf{S}_i\}) = E(\{\mathbf{S}_i\})$. On the other hand, for the quantum Heisenberg model, for which

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{V} \tag{5.164}$$

with

$$\mathcal{H}_0 = J\sum_{ij} S_i^z\, S_j^z\,, \tag{5.165}$$

$$\mathcal{V} = J\sum_{ij} (S_i^x\, S_j^x + S_i^y\, S_j^y)\,, \tag{5.166}$$

$$\left[\mathcal{H}_0, \mathcal{V}\right] \neq 0\,, \tag{5.167}$$

we apply the Trotter formula[24]

$$\mathrm{e}^{-\beta\mathcal{H}} \approx [\mathrm{e}^{-\beta\mathcal{H}_0/m}\mathrm{e}^{-\beta\mathcal{V}/m}]^m\,, \tag{5.168}$$

$$Z = \sum_{\alpha}\langle\alpha|\mathrm{e}^{-\beta\mathcal{H}}|\alpha\rangle\,, \tag{5.169}$$

$$Z = \sum_{\{\alpha_k\}}\prod_{k=1}^{m}\langle\alpha_k|\mathrm{e}^{-\beta\mathcal{H}_0/m}|\alpha_k'\rangle\langle\alpha_k'|\mathrm{e}^{-\beta\mathcal{V}/m}|\alpha_{k+1}\rangle\,, \tag{5.170}$$

where $|\{\alpha\}\rangle$ can be the eigenstate of $\mathcal{H}_0$ and $|\alpha_1\rangle = |\alpha_{m+1}\rangle$. Then the matrix element $\langle\alpha|\mathrm{e}^{-\beta\mathcal{V}/m}|\alpha'\rangle$ is evaluated classically, leading to $\mathrm{e}^{-\beta\mathcal{V}(\alpha,\alpha')/m}$. A d-dimensional quantum spin problem therefore always corresponds to an effective $(d+1)$-dimensional problem.[25]

Besides the quantum (localized) spin model, another important problem is that of electrons with itineracy. One may introduce the Hubbard model as the simplest example,

$$\mathcal{H} = t\sum_{ij}\sum_{\sigma} c_{i\sigma}^\dagger c_{j\sigma} + U\sum_i n_{i\uparrow}n_{i\downarrow}\,. \tag{5.171}$$

For this problem, we need to apply another kind of quantum Monte Carlo method that incorporates the path integral formalism. In this method, the itinerant degrees of freedom of electrons are completely integrated out by the path integral formalism[26] and the remaining problem is then cast as an Ising spin problem in $(d+1)$-dimensions.

[24] cf. Suzuki-Trotter decomposition for the time evolution integrator

[25] One additional dimension comes from the Trotter decomposition.

[26] The basic idea is to transform the Hubbard model into a quadratic form by introducing the additional Ising-type bosonic field through the Trotter decomposition and the Hubbard-Stratonovich transformation. Fields of the quadratic action can be always integrated out, that is, $\mathrm{Tr}\left[\mathrm{e}^{-\sum_{ij} c_i^\dagger A_{ij} c_j}\right] = \det(1+\mathrm{e}^{-A})$, where A carries the Ising-type auxiliary field.

5.8 Interactions between Thermal Excitations of Electrons and Phonons

5.8.1 Formalism of Phonon-Phonon Interactions

The full Hamiltonian of the crystal, H, is

$$H = H_{\rm n} + H_{\rm e} + H_{\rm ep} \tag{5.172}$$

where $H_{\rm e}$ and $H_{\rm ep}$ are the contributions from the electron energy (including electron-electron interactions and thermal electronic excitations), and from electron-phonon interactions (EPI). To understand phonon-phonon interactions, we consider in detail the Hamiltonian for nuclear motions $H_{\rm n}$, extending the Taylor series of Eq. 5.49

$$\begin{aligned} H_{\rm n} = \Phi_0 \quad &+ \quad \sum_{\kappa} \frac{\vec{p}_\kappa^{\,2}}{2m} + \frac{1}{2!}\sum_{\kappa\alpha}\sum_{\kappa'\alpha'} \Phi_{\alpha\alpha'\kappa\kappa'}\, u_{\alpha\kappa}\, u_{\alpha'\kappa'} \\ &+ \quad \frac{1}{3!}\sum_{\kappa\alpha}\sum_{\kappa'\alpha'}\sum_{\kappa''\alpha''} \Phi_{\alpha\alpha'\alpha''\kappa\kappa'\kappa''}\, u_{\alpha\kappa}\, u_{\alpha'\kappa'}\, u_{\alpha''\kappa''} \\ &+ \quad \frac{1}{4!}\sum_{\kappa\alpha}\sum_{\kappa'\alpha'}\sum_{\kappa''\alpha''}\sum_{\kappa'''\alpha'''} \Phi_{\alpha\alpha'\alpha''\alpha'''\kappa\kappa'\kappa''\kappa'''}\, u_{\alpha\kappa}\, u_{\alpha'\kappa'}\, u_{\alpha''\kappa''}\, u_{\alpha'''\kappa'''} + \ldots \end{aligned} \tag{5.173}$$

where we have simplified the development to one atom per unit cell, so $u_{\alpha\kappa}$ is the displacement of atom κ along Cartesian direction α.

The transformation to normal coordinates for an infinite periodic crystal is

$$\vec{U}_{\vec{k}_i} = \frac{1}{\sqrt{N\,m}} \sum_{\vec{r}_\kappa} \vec{u}_{\vec{r}_\kappa} \mathrm{e}^{-\mathrm{i}\vec{k}_i\cdot\vec{r}_\kappa} \ , \tag{5.174}$$

and by Fourier inversion

$$\vec{u}_{\vec{r}_\kappa} = \frac{1}{\sqrt{N\,m}} \sum_{\vec{k}_i} \vec{U}_{\vec{k}_i} \mathrm{e}^{+\mathrm{i}\vec{k}_i\cdot\vec{r}_\kappa} \ . \tag{5.175}$$

Substituting Eq. 5.175 into Eq. 5.173, we identify the Fourier transform of $\Phi_{\alpha\alpha'\kappa\kappa'}$, and use it to define the dynamical matrix of Eqs. 5.58 and 5.59. Our interest now is in the higher-order terms of Eq. 5.173, which cause anharmonic behavior when the atoms are vibrating with substantial amplitude. Using Eqs. 5.173 - 5.175 we can assess phonon-phonon interactions with a classical approach, which is valid at high temperatures.

There is, however, a clarity to be gained in working with the quantized phonon field (a displacement field in the crystal), especially when considering excitations of nonlocal electron states simultaneously with phonon excitations. The phonon field operator is $A_{\vec{k}_i}$

$$A_{\vec{k}_i} = a_{\vec{k}_i} + a^\dagger_{-\vec{k}_i} = A^\dagger_{-\vec{k}_i} \ . \tag{5.176}$$

The raising and lowering operators $a^\dagger_{\vec{k}_i}$ and $a_{\vec{k}_i}$ create and annihilate the phonon $\vec{k}_i$. They are constructed from the momentum and position operators of the Hamiltonian when a plane wave solution is considered for one $\vec{k}_i$. The reader should have references showing the properties of $a^\dagger$ and a for a simple harmonic oscillator **?**), and the $a^\dagger_{\vec{k}_i}$ and $a_{\vec{k}_i}$ work the same way for a single normal mode of a harmonic crystal. The formalism is now in the "occupation number representation," where n phonons of wavevector $\vec{k}_i$ are created with n raising operations as $\left(a^\dagger_{\vec{k}}\right)^n |0\rangle_{\vec{k}} = (n!)^{-1/2}|n\rangle_{\vec{k}}$. The phonon field operators are related to the displacements (now operators) as Fourier coefficients

$$\vec{u}(\vec{r}_\kappa) = \sum_{\vec{k}_i} \sqrt{\frac{\hbar}{2N\,m\,\omega_{\vec{k}_i}}}\,\vec{e}(\vec{k}_i)\,\mathrm{e}^{\mathrm{i}\vec{k}_i\cdot\vec{r}_\kappa}\,A_{\vec{k}_i}\ . \tag{5.177}$$

Note the similar roles of the phonon field operator $A_{\vec{k}_i}$ in Eq. 5.177 and the vector $\vec{U}_{\vec{k}_i}$ in Eq. 5.175.

By substituting Eq. 5.177 into Eq. 5.173, the dynamical matrix is recovered from the quadratic term, with prefactors such as $\hbar/(4\omega_{\vec{k}_i}\,\omega_{\vec{k}_j})$. We define similarly the transformed higher-order terms, so Eq. 5.173 becomes

$$\begin{aligned}
\hat{H}_\mathrm{n} &= \Phi_0 + \sum_{\vec{k}_i} \hbar\,\omega_{\vec{k}_i}\left(a^\dagger_{\vec{k}_i}\,a_{\vec{k}_i} + \frac{1}{2}\right) \\
&+ \sum_{\vec{k}_i}\sum_{\vec{k}_j}\sum_{\vec{k}_k} V(\vec{k}_i,\vec{k}_j,\vec{k}_k)\,A_{\vec{k}_i}\,A_{\vec{k}_j}\,A_{\vec{k}_k} \\
&+ \sum_{\vec{k}_i}\sum_{\vec{k}_j}\sum_{\vec{k}_k}\sum_{\vec{k}_l} V(\vec{k}_i,\vec{k}_j,\vec{k}_k,\vec{k}_l)\,A_{\vec{k}_i}\,A_{\vec{k}_j}\,A_{\vec{k}_k}\,A_{\vec{k}_l} + \ldots\ ,
\end{aligned} \tag{5.178}$$

where the first sum on the right-hand side includes both the kinetic energy and the harmonic part of the potential energy (following standard development **?**)), and the phonon number operator is $a^\dagger_{\vec{k}_i}\,a_{\vec{k}_i}$. The V are related to the Φ as Fourier transforms. For example,

$$\begin{aligned}
V(\vec{k}_i,\vec{k}_j,\vec{k}_k) &= \frac{1}{3!}\sqrt{\frac{1}{N\,m^3}\left(\frac{\hbar}{2}\right)^3\frac{1}{\omega_{\vec{k}_i}\omega_{\vec{k}_j}\omega_{\vec{k}_k}}}\;\delta(\vec{k}_i+\vec{k}_j+\vec{k}_k-\vec{g}) \\
&\times \sum_{\vec{r}_\kappa} \vec{e}(\vec{k}_i)\vec{e}(\vec{k}_j)\vec{e}(\vec{k}_k)\mathrm{e}^{+\mathrm{i}(\vec{k}_i+\vec{k}_j+\vec{k}_k)\cdot\vec{r}_\kappa}\,\Phi_{\alpha,\alpha',\alpha''\kappa,\kappa',\kappa''}\ ,
\end{aligned} \tag{5.179}$$

where the factor δ signifies that the sum $\vec{k}_i+\vec{k}_j+\vec{k}_k$ equals a reciprocal lattice vector, $\vec{g}$. For terminology, the terms with $V(\vec{k}_i,\vec{k}_j,\vec{k}_k)$ and $V(\vec{k}_i,\vec{k}_j,\vec{k}_k,\vec{k}_l)$ in Eq. 5.178 give the energies of "phonon-phonon interactions" because they alter

the phonon energies when larger vibrational displacements are present in the crystal.

Details of the phonon dispersions are immediately relevant when attempting to calculate anharmonic behavior, and several concepts have been developed in detail. For example, if the phonon dispersion relation $\varepsilon(\vec{k})$ were linear, it would be easy to add two wavevectors $\vec{k}+\vec{k}' = \vec{k}''$, and simultaneously conserve energy $\varepsilon + \varepsilon' = \varepsilon''$. The number of three-phonon processes increases as k^2, at least for small k. Many phonon dispersions $\varepsilon(\vec{k})$ are concave downwards, however, so simultaneous energy and momentum conservation is often impossible unless at least one phonon lies on a different branch. Phonon processes that satisfy momentum and energy conservation depend on the symmetry of the dispersion relations, and on the crystal structure. For larger phonon wavevectors $\vec{k}$, momentum conservation is possible by adding a reciprocal lattice vector. The idea is that the momentum transferred to the entire crystal occurs with zero energy because of the large mass of the crystal. Such "umklapp" processes allow many more three-phonon interactions, but the phonon wavevectors must be of length comparable to the reciprocal lattice vector if the vector additions are possible.

An important question is how many terms are needed in Eq. 5.178 to account accurately for anharmonic behavior. We expect the higher-order terms in Eq. 5.173 to be progressively smaller, but the requirement of energy conservation and the factor δ for momentum conservation in Eq. 5.179 restricts the allowable three-phonon processes, owing to restrictions imposed by the phonon dispersions $\varepsilon(\vec{k})$. Fourth-order anharmonicity can therefore play a role as significant as the third. Another complexity is that four-phonon processes can be generated from three-phonon processes. The idea is that two three-phonon process such as $\vec{k} + \vec{k}' \rightarrow \vec{k}''$ and $\vec{k}'' \rightarrow \vec{k}''' + \vec{k}''''$ can be combined through intermediate states $\{\vec{k}''\}$ to become a four-phonon process in second-order perturbation theory, which involves matrix elements as $\sum_{\vec{k}''} \langle \vec{k}''', \vec{k}'''' | H''' | \vec{k}'' \rangle \langle \vec{k}'' | H''' | \vec{k}, \vec{k}' \rangle$. With effort, this field theoretic approach has been used for some physical predictions as described below. Approximate summations of the perturbation series have been performed, and some results for anharmonic heat capacity were reported for solids with Lennard-Jones potentials **?**). This approach also has conceptual value for understanding the meaning of anharmonicity.

5.8.2 Formalism of Electron-Phonon Interactions

The large difference in energy scales of phonons and electrons can motivate the separation of the Hamiltonian of the solid into a term with nuclear coordinates for the phonons H_{n}, and a term with electron coordinates H_{e} (Eq. 5.172). The energy of the crystal deformation caused by a phonon originates with the electrons, of course, but although this potential energy of deformation is

electronic in origin, it transfers to kinetic energy in the motion of the nuclei.[27] If the electrons were always in their ground states, all this energy of deformation would be associated with $H_{\rm n}$ alone (i.e., the phonons). Treating the electron system and the phonon system as two independent thermodynamic systems becomes inconsistent at finite temperature, however, because the presence of phonons alters the thermal excitations of electrons.

Very generally, the electron-phonon interaction (EPI) requires the coordinates of the electrons $\{\vec{r}^{\,\rm el}_\lambda\}$, and coordinates of the nuclei $\{\vec{r}^{\,\rm n}_j\}$

$$H_{\rm ep} = \sum_{\lambda,j} v(\vec{r}^{\,\rm el}_\lambda, \vec{r}^{\,\rm n}_j) \; . \tag{5.180}$$

First assume the nuclear motions are slow enough so the electron levels adapt continuously to the evolving structure, i.e., the Born–Oppenheimer approximation. This "adiabatic approximation" does not allow the nuclear kinetic energy to alter the electron states.[28] The adiabatic EPI increases with the number of phonons, but it also scales with the thermal excitations of electrons. (Effects from how the interatomic potential depends on phonon populations alone are already counted as phonon-phonon interactions.)

In this adiabatic picture, unique electron eigenstates exist for a snapshot of the nuclear thermal displacements, but these evolve continuously as the nuclei move. Taking for reference a ground state with the nuclear positions of a perfectly periodic crystal and the electronic structure at 0 K, we seek to account for how the displaced nuclear positions alter the thermally-excited electronic energy levels (changing the electronic excitations induced by temperature). For a given electronic structure, the adiabatic electron-phonon interaction is proportional to the number of phonons, $n(\varepsilon, T) + 1/2$, times the difference of electron occupancy with respect to the ground state, $f(T) - f(0)$, where $f(T)$ is the Fermi-Dirac distribution. When there are sharp features in the electronic DOS near the Fermi level, there will be other temperature-dependences associated with the thermal sampling of the excited electron states.

How the electron eigenstates change with nuclear displacements is a problem that generally requires considerable detail about the states of electrons and phonons. A simple approach is to consider an electronic band formed from isotropic *s*-electrons, and a uniform dilation as may be associated with longitudinal phonons of long wavelength,

$$H^{\rm D}_{\rm ep} = -\sum_\lambda D\Delta(\vec{r}^{\,\rm el}_\lambda) \; . \tag{5.181}$$

where $\Delta(\vec{r}^{\,\rm el}_\lambda)$ is the fractional change in volume at $\vec{r}^{\,\rm el}_\lambda$, and D is a "deformation

[27] For a harmonic phonon, the electron-phonon interaction affects equally the energies of the electrons and the phonons.

[28] The "non-adiabatic" electron-phonon interaction accounts for how the electronic states are altered by the nuclear kinetic energy, not the potential energy of displaced nuclei as for the adiabatic case. The non-adiabatic EPI does require thermal activation, and can be responsible for superconductivity.

potential," typically a few eV.[29] This simple approach is convenient because the $\Delta(\vec{r}_\lambda^{\,\mathrm{el}})$ is related to the displacement $\vec{u}$ as through its divergence: $\Delta(\vec{r}_\lambda^{\,\mathrm{el}}) = \vec{\nabla} \cdot \vec{u}$, so

$$H_{\mathrm{ep}}^{\mathrm{D}} = -D \sum_{\lambda} \vec{\nabla} \cdot \vec{u}(\vec{r}_\lambda^{\,\mathrm{el}}) \; . \tag{5.182}$$

For longitudinal phonons

$$\vec{e}(\vec{k}_i) = \frac{\vec{k}_i}{k_i} \; , \tag{5.183}$$

and by taking the divergence of Eq. 5.177 we have, now in the occupation number representation

$$\hat{H}_{\mathrm{ep}}^{\mathrm{D}} = -\mathrm{i}D \sum_{\vec{r}_\lambda^{\,\mathrm{el}}} \sum_{\vec{k}_i} \sqrt{\frac{\hbar}{2N\,m\,\omega_{\vec{k}_i}}}\, |\vec{k}_i|\, \mathrm{e}^{\mathrm{i}\vec{k}_i \cdot \vec{r}_\lambda^{\,\mathrm{el}}}\, A_{\vec{k}_i} \tag{5.184}$$

Consistent with the second quantization formalism for phonons, fermion field operators are used for the electrons

$$\Psi^\dagger(\vec{r}_\lambda^{\,\mathrm{el}}) = \sum_{\vec{k}_\lambda} C_{\vec{k}_\lambda}^\dagger\, \mathrm{e}^{-\mathrm{i}\vec{k}_\lambda \cdot \vec{r}_\lambda^{\,\mathrm{el}}} \; , \tag{5.185}$$

which when applied to a multi-electron state, places an electron in the state $\vec{k}_\lambda$ (the operator $C_{\vec{k}_\lambda}$ removes an electron from this state). When small, $\hat{H}_{\mathrm{ep}}^{\mathrm{D}}$ in Eq. 5.184 is a perturbation that mixes electronic states as $\langle \Psi_f | \hat{H}_{\mathrm{ep}}^{\mathrm{D}} | \Psi_i \rangle$. This integral can be evaluated by using the fermion field operators C and $C^\dagger$ (Eq. 5.185), where the exponential phase factors cancel the k-space integration unless there is a conservation of wavevector. This forces the same total $\vec{k}_\lambda + \vec{k}_i$ for the creation operators as for the annihilation operators

$$\hat{H}_{\mathrm{ep}}^{\mathrm{D}} = -\mathrm{i}D \sqrt{\frac{\hbar}{2N\,m}} \sum_{\vec{k}_\lambda}^{\text{electron}} \sum_{\vec{k}_i}^{\text{phonon}} \sqrt{\frac{k_i}{c_L}}\, A_{\vec{k}_i} C_{\vec{k}_\lambda + \vec{k}_i}^\dagger C_{\vec{k}_\lambda} \; , \tag{5.186}$$

plus an analogous term with $A_{\vec{k}_i}^\dagger C_{\vec{k}_\lambda + \vec{k}_i} C_{\vec{k}_\lambda}^\dagger$ for phonon creation. For Eq. 5.186 we also used a linear dispersion relationship $\omega_L(k_i) = c_L k_i$, appropriate for long-wavelength, longitudinal acoustic phonons. In accounting for the effects of phonons on the electron states, it is usually necessary to consider a more general case than the isotropic deformation potential of Eq. 5.181 as

$$I \propto \left\langle \vec{k}_\lambda + \vec{k}_\lambda \middle| \vec{e}_j(\vec{k}_j) \cdot \nabla_{\vec{r}_j^{\,\mathrm{n}}} v(\vec{r}_\lambda^{\,\mathrm{el}},\, \vec{r}_j^{\,\mathrm{n}}) \middle| \vec{k}_\lambda \right\rangle \; , \tag{5.187}$$

[29] It is not typical to have electron energies shifted by several eV under high compressions, so the D typically pertain to small elastic distortions over a finite volume of crystal.

where the gradient is with respect to the nuclear coordinates.[30]

In general, there are two lower order terms that are used to describe the electron-phonon interaction. The first is a generalization of the previous result Eq. 5.186

$$\hat{H}^1_{\rm ep} = \sum_{\vec{k}_i} \sum_{\vec{k}_\lambda} V^{\rm ep}(\vec{k}_i, \vec{k}_\lambda)\, A_{\vec{k}_i} C^\dagger_{\vec{k}_\lambda + \vec{k}_i} C_{\vec{k}_\lambda} \ . \tag{5.188}$$

It accounts for the processes where an electron is excited from state $\vec{k}_\lambda$ to $\vec{k}_\lambda + \vec{k}_i$, simultaneously with the annihilation of a phonon in state $\vec{k}_i$, or creation of a phonon in state $-\vec{k}_i$. Momentum conservation restricts significantly these three-phonon processes. The second low-order term for electron-phonon coupling is $\hat{H}^2_{\rm ep}$, and it includes a factor from second-order perturbation theory, $\sum_{\vec{k}'} \langle \vec{k}|\Omega|\vec{k}'\rangle\langle\vec{k}'|\Omega|\vec{k}\rangle$. Note that the initial and final states of the electron have the same $\vec{k}$, so there is no change of electron state by this process. Nevertheless, the scattering into the virtual states $\{\vec{k}'\}$ serve to alter the self-energy of the electron.

For calculating the thermodynamic effects of electron-phonon coupling, it is more useful to know how the electron-phonon coupling depends on energy, rather than momentum as in Eq. 5.188 for $\hat{H}^1_{\rm ep}$. This requires averaging over all electron states near the Fermi surface separated by $\vec{k}_j$ and differing in energy by a selected $\hbar\omega$. Most work on electron-phonon coupling has focused on superconductivity, where $\hbar\omega$ is rather small, and the electron states are close enough to the Fermi surface that it is reasonable to use ground state Fermi surface properties. The Éliashberg coupling function $\alpha^2 g(\omega)$, where $g(\omega)$ is the phonon DOS, accounts for all scattering between electron states on the Fermi surface. The function $\alpha^2 g(\omega)$ often looks similar $g(\omega)$, although some parts of the spectrum are given different weights. An important moment is calculated by weighting by ω^{-1}, and can be shown to be

$$2\int_0^{\omega_{\rm max}} \frac{\alpha^2 g(\omega)}{\omega} {\rm d}\omega = \frac{N(E_{\rm F})\langle I^2\rangle}{M\langle\omega^2\rangle} = \lambda \tag{5.189}$$

where $\langle I^2\rangle$ is an average of I^*I of Eq. 5.187 over all directions, $\langle\omega^2\rangle$ is a second moment of the phonon DOS, $N(E_{\rm F})$ is the electron DOS at the Fermi surface, and λ is the "electron-phonon coupling parameter." The strength of the electron-phonon coupling scales with λ.

Further Reading

The contents of the following are described in the Bibliography.

[30] "Frölich coupling" is another example of an EPI, where longitudinal optical phonons create an electric field in an ionic solid. In this case, the Fourier transform of a Coulomb potential gives a factor of $4\pi k_i^{-2}$, and with two ions involved, $\hat{H}^{\rm Fr}_{\rm ep}$ goes as k_i^{-1} (instead of the $k_i^{+1/2}$ for $\hat{H}^{\rm D}_{\rm ep}$ in Eq. 5.186). Transition rates, or scattering rates, go as the square of these quantities.

Phillip R. Bevington: *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill, New York, 1969). See especially Ch. 4.

L. P. Bouckaert, R. Smoluchowski, and E. P. Wigner: Phys. Rev. **50**, 58 (1936).

A. A. Maradudin and S. H. Vosko: 'Symmetry Properties of the Normal Modes of a Crystal', Reviews of Modern Physics **40**, 1-37 (1968).

M Tinkham: *Group Theory and Quantum Mechanics* (McGraw-Hill, New York, 1964) Chapter 8.

G. L. Squires: *Introduction to the Theory of Thermal Neutron Scattering* (Dover, Mineola, New York 1996).

J. L. Warren: 'Further Considerations on the Symmetry Properties of the Normal Modes of a Crystal', Reviews of Modern Physics **40**, 38-76 (1968).

Chapter 6

Chopper Spectrometers

6.1 Concept of a Chopper Spectrometer

Enrico Fermi was unusual. It is a remarkable and profound honor that the name "fermion," with a lower case f, is standard terminology. The field of experimental neutron scattering offers him a smaller honor with the term "Fermi chopper" (Fig. 6.1). A generic Fermi chopper is depicted in Figure 6.2. The ARCS instrument is in fact a Wide Angular Range Direct Geometry Fermi Chopper Spectrometer.[1]

The Fermi chopper is discussed at length in Section 6.4. For now we note that it works as a fast shutter. By aligning its slot at the right instant after a neutron burst leaves the moderator, it selects a bunch of neutrons that have a particular velocity. We know the time of the neutron burst and the distance between the neutron source and the Fermi chopper, so we therefore know the neutron velocity and its energy, $E = \frac{1}{2}mv^2$. The idea is best illustrated with the time and distance diagram in Figure 6.3.

The slopes of the lines in Figure 6.3 are velocities of neutrons. Neutrons of many velocities are emitted from the moderator. The figure depicts two neutron pulses from the moderator, separated in time by perhaps 1/60 sec. The Fermi chopper selects a narrow range of velocities, corresponding to a time window of tens of microseconds. Of course the narrower this window, the more neutrons are blocked by the chopper, and the lower the flux of neutrons on the sample. On the other hand, the narrower this window, the more precisely selected are the velocities, and hence better energy resolution is achieved. The Fermi chopper therefore controls the incident energy, the intensity, and the energy resolution of the neutrons incident on the sample.

The sample usually transmits most of the incident neutrons without scattering, and most of the scattered neutrons are scattered elastically. There is a strong tendency for the lines in Figure 6.3 to have unchanged slopes through

[1] Unfortunately for Enrico Fermi, the acronym ARFCS is less euphonious than ARCS.

Figure 6.1: Enrico Fermi works with an electronic control for a neutron chopper during his Argonne days.

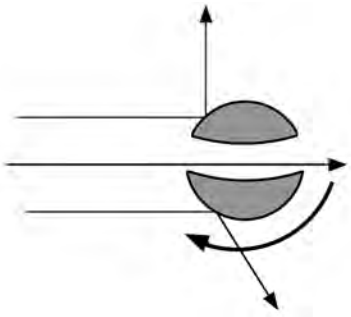


Figure 6.2: Fermi chopper, comprising a spinning cylinder with a slot in it. Neutrons are transmitted only when the slot is aligned properly along the path of the beam. The chopper works by scattering neutrons out of the forward beam, or by absorbing them by the inner surfaces of the slot. The ARCS choppers work primarily by absorption.

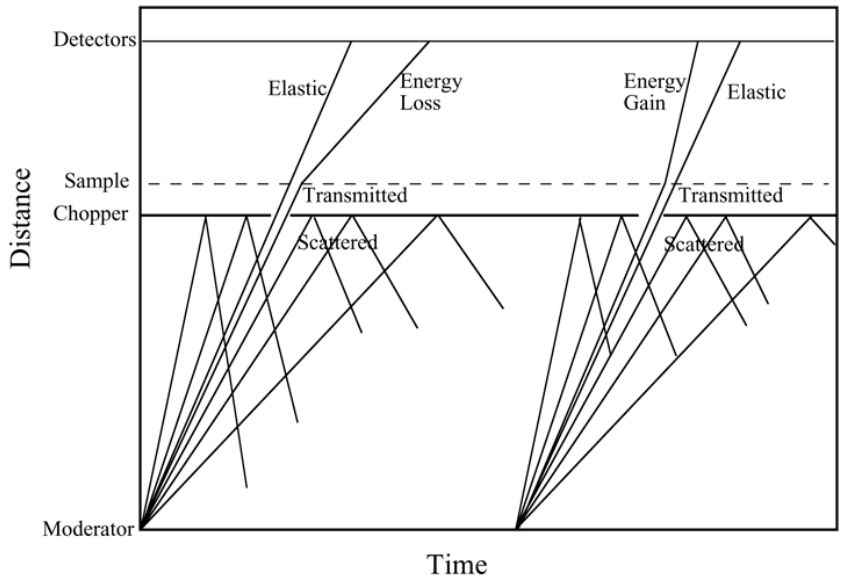


Figure 6.3: Distance–time diagram for inelastic scattering by a direct geometry chopper spectrometer. Positions of the moderator, chopper, sample, and detectors are marked on the vertical axis. Two moderator pulses occur at points on the x-axis. (Real pulses are not instantaneous, and this issue is discussed below.) When neutrons reach the position of the chopper, they are transmitted, absorbed, or scattered.

the sample.[2] It is straightforward to know the total distance from the neutron source to the sample, and from the sample to the detector tube. Given the velocity selected by the Fermi chopper, we can figure out the time of arrival of the elastic neutrons at the detectors. Experimentally, we observe an intense elastic peak in the time spectrum at any of the detectors. This is typically quite close to the predicted arrival time, but the experimental time is used to identify the elastic scattering. Those neutrons that are scattered inelastically, however, have lines with kinks at the sample, and after the sample their slopes in Fig. 6.3 are either steeper or shallower than for the incident beam. The neutrons that arrive earlier have gained energy from the sample, and those that arrive later have lost energy to the sample. The inelastic spectrum is obtained from the histogram of neutron arrival times. A Fermi chopper spectrometer works entirely by timing.

6.2 Neutron Sources

6.2.1 Spallation

It is a challenge to produce "free neutrons," meaning neutrons that are "free of the nucleus." "Spallation" is one way to make them. The "spallation" process got its name by analogy to using a hammer to "chip" pieces off a heavy stone. Here the the hammer is the particle beam, the chips are neutrons, and the stone is the nucleus. Other things come out of the nucleus besides neutrons, especially γ radiation. High-energy protons (of order 1 GeV) are preferred for the particle beam because protons produce more neutrons, and less heat and photons than are produced by electron beams. The major component of a spallation neutron source is therefore a high-energy, high-current proton accelerator. It often includes a linear accelerator followed by a buncher ring to compress the proton pulse into a short burst in time. Neutron yields are largest for nuclei of high atomic number, since these are neutron-rich.[3] Tungsten and mercury are good choices in practice. Uranium is even better, but it tends to give problems at higher power levels.

6.2.2 Moderation and Moderator Spectrum

The neutrons fresh from a spallation reaction have energies of order MeV, but the neutrons used in inelastic scattering have energies of order 100 meV. The excessive energies of the spalled neutrons need to be "moderated." The "moderator" makes the transition between the neutron target and the neutron instrument by delivering a useful spectrum of neutrons to the instrument.

[2] Elastic scattering will change the direction of the neutron, but this cannot be depicted on Fig. 6.3.

[3] An excess of neutron over protons is required for stability. Too many protons means too much Coulombic repulsion, but neutrons can help overcome this instability and keep the nucleus together. Nevertheless, too many neutrons will "drip" out of the nucleus, and neutron-rich isotopes near the neutron "drip line" are good candidates for the target material.

How does the moderator reduce the neutron energy by a huge factor of 10^7? By inelastic collisions with nuclei in the moderator. In an inelastic collision between a neutron and a nucleus of the same mass, hydrogen of course, up to half of the kinetic energy of the collision can be transferred to the hydrogen nuclei. The hydrogen in the moderator can be in various chemical forms such as liquid hydrogen, water, or solid methane, depending on the desired temperature and density, for example. The number of collisions required for moderation, n, is

$$2^n \simeq 10^7 , \tag{6.1}$$

$$n \simeq 23 . \tag{6.2}$$

For the moderator of the ARCS spectrometer, the process of moderation therefore involves a relatively small number of collisions, and is therefore a statistical process. As an approximation it is often assumed that the neutrons leaving the moderator consist of two components. The first is a fully moderated spectrum of thermal neutrons. These neutrons have the Maxwell-Boltzmann spectrum, with probabilities based on Boltzmann factors of $\exp[(-m_{\mathrm{n}}v^2)/(kT)]$, where T is the temperature of the moderator. The second subspectrum is called the "epithermal" neutron spectrum. It is a broad spectrum with a tail that goes to very high energies. Epithermal neutrons have not undergone enough interactions with the moderator to acquire a thermal distribution. This approximation of two subspectra has some semblance of the truth.

A better moderator spectrum can be calculated by Monte Carlo neutron transport codes, such as the one that produced the spectra of Figure 6.4a. The peak intensity for the "decoupled, poisoned moderator" is around 40 meV, but there is useful intensity over a broad range of energies beyond 1 eV. Figure 6.4 shows similar calculations for a moderator that uses liquid hydrogen. Its spectrum is peaked at a much lower energy, perhaps 3 meV. The total number of neutrons is somewhat better, but by bunching up their energy into a narrow range, the usable intensity is much higher. "Cold neutrons"' are a new frontier in neutron scattering research.

6.2.3 Decoupled, Poisoned Moderator

An important consideration for moderator design is "coupling," meaning the connection between the target that is hit by the proton beam, and the material of the moderator. After slowing down in the moderator, the slower neutrons can be absorbed by materials such as Cd and Gd. Putting a layer of Cd or Gd between the moderator and the target can suppress the transmission of the slower neutrons back into the target. This is a "decoupled" moderator. (On the other hand, the fast neutrons from the target are not absorbed as they enter the moderator.) In contrast, a "coupled" moderator allows transmission of all neutrons between the target and the moderator, and less absorption. A coupled moderator produces more neutrons, but it has a disadvantage in the time structure of its neutron pulse. The slow neutrons traveling around the

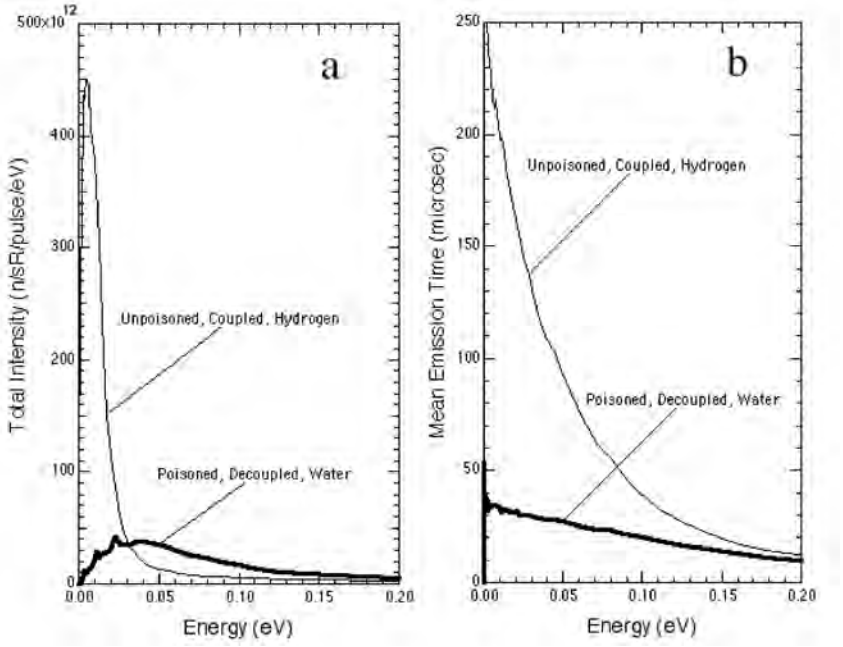


Figure 6.4: Monte Carlo simulations of intensity spectra and time spectra for two SNS moderators. ARCS is on the decoupled, poisoned water moderator. For comparison a cryogenic hydrogen moderator is also shown. (**a**) Total intensity. (**b**) Mean emission time.

coupled moderator take some time before entering the neutron instrument. This time delay degrades the energy resolution of the Fermi chopper spectrometer, which works by timing.

Another trick to producing short neutron bursts from the moderator is to put a neutron absorber such as Cd or Gd at some depth inside the moderator itself. Again, the idea is to absorb the slow neutrons that do not follow an efficient path out of the moderator. This is called "poisoning" the moderator. A "poisoned, decoupled moderator" generates neutron bursts that are short in time, and this is the moderator of choice for the ARCS instrument, for example. For neutrons having energies below 100 meV, the moderator for ARCS emits neutron bursts of approximately 20 microseconds. In contrast, an "unpoisoned, coupled moderator" generates more neutrons in each burst, although these neutrons are emitted over a longer time.

An additional complexity is that the neutrons of different energies are emitted from the moderator at different times. Of course the neutrons with highest energies, which do not undergo enough collisions with the nuclei in the moderator, are emitted in the shortest times after the proton pulse hits the target. In general, the lower-energy neutrons leave the moderator at later times, with a broader spread in their emission times. The details of this time-energy correlation are not simple, however, and are best understood by Monte Carlo simulations and experimental measurements. Nevertheless, the neutron emission times affect the energy resolution of the spectrometer, as discussed in Section 6.4. Figure 6.4b shows this energy dependence of the pulse emission times for two SNS moderators. It also shows clearly that the emission time is shorter for the decoupled and poisoned ARCS moderator than for the coupled, unpoisoned hydrogen moderator.

6.3 Neutron Guides

6.3.1 Geometrical Optics

Here we develop the scattering of neutrons at interfaces between two homogeneous media. In a homogeneous potential, a neutron wavefunction propagates forward without deflection, but it has a wavelength that depends on the potential. This is much like the propagation of light through glass, for example. The only deflections occur at interfaces, or at changes in the "density." With geometrical optics we can readily utilize the familiar constructions of light optics, scaled appropriately for neutrons. In the present analysis we justify the "geometrical optics" approach to analyzing neutron scattering from macroscopic objects, especially mirrors.

Recall from (1.74) and (1.75), or (2.39) the integral form of the Schrödinger equation in the Born approximation:

$$\Psi_{sc}(\vec{Q}) \quad = \quad -\frac{m}{2\pi\hbar^2}\int V(\vec{r}')\,e^{i\vec{Q}\cdot\vec{r}'}\,d^3\vec{r}' \;, \tag{6.3}$$

where we have ignored the standard form of the outgoing spherical wave that properly multiplies (6.3) (and sets the relationship between $f(\vec{Q})$ and $\Psi_{\rm sc}(\vec{Q},\vec{r})$). For nuclear scattering we use the "Fermi pseudopotential" of (2.40), which places all the scattering potential at a point nucleus at $\vec{r}$:

$$V_{\rm nuc}(\vec{r}) = 4\pi\frac{\hbar^2}{2m}\, b\, \delta(\vec{r}) \;, \tag{6.4}$$

where b is a simple constant (perhaps a complex number). The next step is to place numerous Fermi pseudopotentials at the positions of all N nuclei in the material, $\{\vec{r}_i\}$:

$$V(\vec{r}) = 4\pi\frac{\hbar^2}{2m}\, \bar{b} \sum_{\vec{r}_i}^{N} \delta(\vec{r}-\vec{r}_i) \;, \tag{6.5}$$

where $\bar{b}$ is the average scattering length per nucleus (assuming a mix of isotopes or elements). Substituting (6.5) into (6.3), we notice the handy cancellation of many constant prefactors:

$$\Psi_{\rm sc}(\vec{Q}) \;\; = \;\; -\bar{b}\int \sum_{\vec{r}_i}^{N} \delta(\vec{r}\,' - \vec{r}_i)\, {\rm e}^{{\rm i}\vec{Q}\cdot\vec{r}\,'}\, {\rm d}^3\vec{r}\,' \;. \tag{6.6}$$

In a homogeneous medium we can consider wave motion only in the forward direction, for which $\vec{Q}=0$, In the forward direction, ${\rm e}^{{\rm i}\vec{Q}\cdot\vec{r}\,'}=1$, simplifying (6.6):

$$\Psi_{\rm sc}(\vec{Q}) \;\; = \;\; -\bar{b}\int \sum_{\vec{r}_i}^{N} \delta(\vec{r}\,' - \vec{r}_i)\, {\rm d}^3\vec{r}\,' \;, \tag{6.7}$$

$$\Psi_{\rm sc}(\vec{Q}) \;\; = \;\; -N\bar{b} \;. \tag{6.8}$$

Note how the details of the positions $\{\vec{r}_i\}$ are lost for forward scattering. We can obtain the same equation (6.8) if, instead of placing δ-functions at all nuclei, we we use a homogeneous potential throughout the entire material. Instead of using (6.5), equation (6.8) can be obtained by using the following homogeneous potential for $V(\vec{r}\,')$ in (6.3):

$$v_0(\vec{r}) = 4\pi\frac{\hbar^2}{2m}\, \bar{b}\rho \;, \tag{6.9}$$

where ρ is the number of nuclei per unit volume.

This is an important observation. Using (6.9) instead of (6.5), we make the transition from individual scatterings by atomistic Fermi pseudopotentials to geometrical optics. The neutron wave is now considered to travel through a homogeneous potential. The neutron is treated as propagating without scattering, although with a different wavevector depending on the "density," $\bar{b}\rho$.

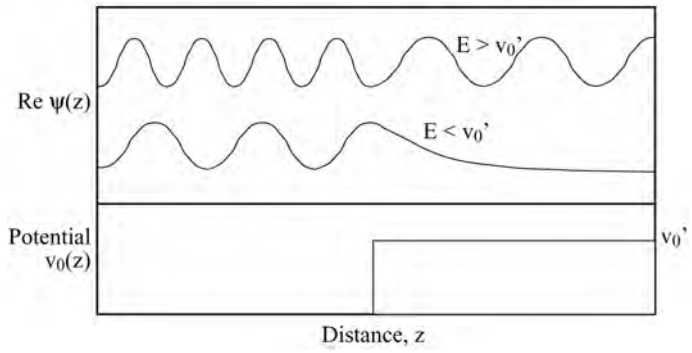


Figure 6.5: Snapshots of neutron wavefunctions near an interface, which has a step in its potential at the transition between the vacuum (left) and the material (right). The total energies of the wavefunctions are either larger or smaller than the homogeneous potential inside the material.

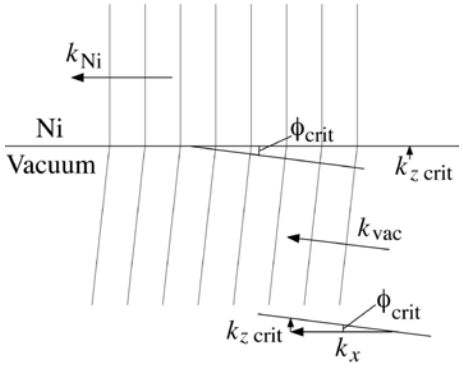


Figure 6.6: Wavevectors and wave crests at a Ni/vacuum interface. The critical condition has the wavevector in the Ni layer parallel to the interface as shown.

6.3.2 Total Reflection

When a neutron of energy E enters a region where the potential energy $v_0(\vec{r})$ is positive, its kinetic energy is reduced and its wavelength is increased (see Fig. 6.5). The kinetic energy cannot go below zero, of course, and a consequence is that some neutrons may not have enough energy to enter a material having a positive $v_0(\vec{r})$ (except for some surface penetration). Substituting typical numbers into (6.9), we find that $v_0(\vec{r}) = 3{\times}10^{-4}$ meV, corresponding to a neutron wavelength of 500 Å. Reflection will be total, since the neutron cannot penetrate into the solid, but the neutron is conserved. This result of 500 Å pertains to neutrons arriving normal to a surface. Neutrons of wavelength longer than this critical wavelength will be reflected by the surface. A critical wavevector is shown in Fig. 6.6.

It is possible to get total reflection of more energetic neutrons if they arrive

at shallow angles to the surface. The effect is analogous to the case of total internal reflection of light, where a light ray moving in a medium of high index of refraction can be totally reflected at an interface with a medium of low index of refraction, if it reaches this interface at an angle below a critical angle. Such a case is shown in Fig. 6.6. The neutron travels a bit slower in the Ni than in the vacuum. In the Ni, the neutron has a higher potential energy, a lower kinetic energy, and a longer wavelength. Notice the matching of the wave crests across the interface. This continuity of the neutron wavefunction forces a change in direction of the wavevector across the interface – this allows for differences in wavelengths in the vacuum and in the Ni.

The relationship between the neutron wavelength and the critical angle for a homogeneous potential $v_0(\vec{r})$ can be derived from the Schrödinger equation (6.14) by separation of variables. Start with the neutron wavefunction for a neutron moving as a plane wave in the x-z plane, where $\hat{\vec{z}}$ is normal to the interface between, say vacuum and nickel metal. (For neutrons, the nickel metal has the lower index of refraction than the vacuum.)

$$\Psi_{sc}(\vec{k},\vec{r}) = e^{i\vec{k}\cdot\vec{r}} , \tag{6.10}$$

$$\Psi_{sc}(\vec{k},x,z) = e^{i(k_x\hat{\vec{x}}+k_z\hat{\vec{z}})\cdot(x\hat{\vec{x}}+z\hat{\vec{z}})} , \tag{6.11}$$

$$\Psi_{sc}(\vec{k},x,z) = e^{ik_x x}\, e^{ik_z z} , \tag{6.12}$$

$$\Psi_{sc}(\vec{k},x,z) \equiv \psi_x(x)\ \psi_z(z) . \tag{6.13}$$

Substitute (6.13) into the Schrödinger equation (1.53):

$$-\frac{\hbar^2}{2m}\Big(\frac{\partial^2\psi_x}{\partial x^2}+\frac{\partial^2\psi_z}{\partial z^2}\Big)+V(z)\,\psi_x(x)\psi_z(z)=E\,\psi_x(x)\psi_z(z) , \tag{6.14}$$

where we have assumed the potential varies only along $\vec{z}$, changing only at the interface. Dividing through by $\psi_x(x)\psi_z(z)$ and rearranging:

$$-\frac{\hbar^2}{2m}\frac{\frac{\partial^2\psi_z}{\partial z^2}}{\psi_z}+V(z)-E = +\frac{\hbar^2}{2m}\frac{\frac{\partial^2\psi_x}{\partial x^2}}{\psi_x} , \tag{6.15}$$

$$= -\varepsilon . \tag{6.16}$$

We have separated the z-dependence from the x-dependence. The left-hand side of (6.15) depends only on z, the right only on x, but both z and x can change independently. This means that both sides can only be equal to a constant, which we denoted $-\varepsilon$ in (6.16). The two equations for $\psi_x(x)$ and $\psi_z(z)$ become:

$$\frac{\hbar^2}{2m}\frac{\partial^2\psi_z}{\partial z^2} = \Big[\varepsilon-E+V(z)\Big]\psi_z , \tag{6.17}$$

$$\frac{\hbar^2}{2m}\frac{\partial^2\psi_x}{\partial x^2} = -\varepsilon\,\psi_x . \tag{6.18}$$

To obtain the required a plane wave solution for a propagating neutron, $\psi_x = e^{\pm i(\sqrt{2m\varepsilon})x/\hbar}$, we see that ε must be positive. Our neutron is moving mostly

along the x-direction at a glancing angle to the surface, with kinetic energy E when it is in the vacuum. We therefore know that $\varepsilon - E$, although negative, must be rather small. It is therefore possible for a small positive $V(z)$ to switch the sign of the right-hand side of (6.17) from negative to positive. The consequence is interesting. The solution for $\psi_z(z)$ changes from a propagating wave to a damped exponential function. When $V(z)$ is sufficiently positive, the neutron wavefunction therefore does not propagate into the nickel. It is instead reflected from the interface. Of course the $V(z)$ for nickel is fixed, but we can alter the incident angle to get the same effect as shown in Fig. 6.5. By reference to (6.12) and (6.13), for example, we can see that the change in sign of (6.17) occurs when we select k_z so that:

$$k_{z\,\mathrm{crit}} = \frac{\sqrt{2mv_0}}{\hbar} = \sqrt{4\pi\,\bar{b}\rho}\ . \tag{6.19}$$

6.3.3 Critical Angle

The critical angle for total reflection, ϕ_{crit}, is the ratio of $k_{z\,\mathrm{crit}}$ to the wavevector along x, k_x, (the magnitude of k_x is essentially the same as the magnitude of the incident wavevector k):

$$\phi_{\mathrm{crit}} = \frac{\sqrt{4\pi\,\bar{b}\rho}}{k} = \frac{\lambda}{2\pi}\ \sqrt{4\pi\,\bar{b}\rho}\ , \tag{6.20}$$

$$\phi_{\mathrm{crit}} = \lambda\ \sqrt{\frac{\bar{b}\rho}{\pi}}\ . \tag{6.21}$$

The critical angle increases in proportion to the wavelength of the neutron. Lower-energy neutrons can be reflected from a surface at higher angles than higher-energy neutrons. For natural Ni metal, the evaluation of (6.21) gives the wavelength-dependence of the critical angle, ϕ_{crit}, here converted from radians to degrees:

$$\phi^{\mathrm{Ni}}_{\mathrm{crit}}\,[^\circ] = 0.0991\ \lambda\ [\text{Å}]\ . \tag{6.22}$$

Finally, neutron reflectivity can be used for making spin-polarized neutron beams. The coherent scattering lengths are different when neutrons of up spin or down spin are scattered by a magnetic moment having a component along $\hat{z}$. For Co, in fact, the scattering lengths, $b^{\uparrow}_{\mathrm{Co}}$ ad $b^{\downarrow}_{\mathrm{Co}}$, are of opposite sign. A polarizer can be built by choosing incident angles where one of the neutron polarizations is transmitted into the magnetic material, and the other reflected.

6.3.4 Guide Design

For neutrons incident on a surface below the critical angle, simple ray diagrams can be used to develop conceptual designs of neutron guides. Their characteristics are analogous to cylindrical optical fibers for light, but neutron guides

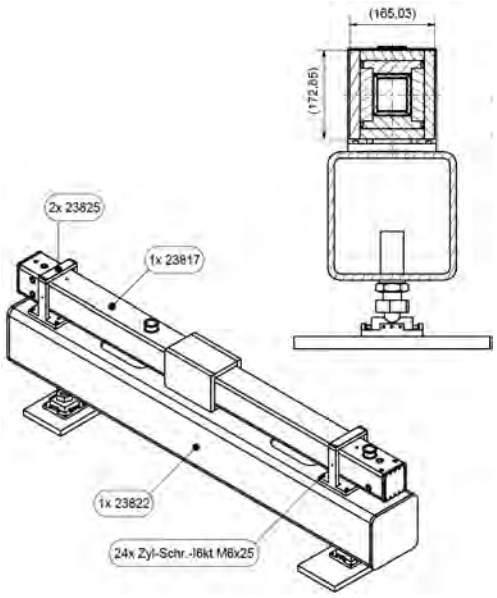


Figure 6.7: End view and overview of the second section of the main neutron guide for the ARCS spectrometer. Rigid mounting and mechanical adjustment screws are required for ensuring that the long glass mirrors are aligned precisely.

are usually made of four long mirrors with a rectangular cross section of order 0.1×0.1 m^2. A section of the ARCS guide is shown in Fig. 6.7. The four mirrors are seen end-on in the drawing at the top left.

The critical angle for Ni metal is a reference standard, but today a "supermirror" can be prepared from multiple layers of metals, giving higher critical angles by a factor m, where $\phi_{\rm crit} = m\phi_{\rm crit}^{\rm Ni}$. Today the upper limit to m has reached 7. Since these angles are quite small for neutrons of meV energy, guides tend to be quite long. Nevertheless, focusing guides can be designed with parabolic or elliptical surfaces, for example, and these are efficient for transporting neutrons over long distances.

Efficient neutron transport is usually the main function of a guide. With a guide in its incident flight path, the instrument can be placed a good distance away from the moderator. This allows more space around the instrument, which can be important when instruments are crowded around a small neutron target. The separation from the moderator also allows the instrument to be placed in a location with lower levels of background radiation. Two other issues are useful for understanding the use of guides for neutron transport.

- By giving the mirrors of the guide a slight curvature along the beam direction, the direction of the neutron beam can be bent away from its original straight-line path out of the moderator. This makes it possible for the specimen to be out of a line-of-sight path from the moderator, and therefore less subject to background from fast neutrons.

- Most importantly, the guide reduces the usual attenuation of neutron flux with distance. Instead of the r^{-2} fall-off of intensity, a good guide

should cause minimal loss of neutrons that enter the guide below the critical angle. (Note that those entering at higher angles would often be blocked by collimators before the specimen anyhow.) A simple straight mirror guide essentially takes the numbers and divergences of neutrons entering the guide, and translates this distribution to the exit of the guide.

6.3.5 Brightness

Inelastic scattering measurements require beams on samples that have:

- High neutron current [neutrons/sec]
- Small size [cm^2]
- High flux [neutrons cm^{-2} s^{-1}]
- Low divergence [α, radians] (This requirement is unimportant if the Q-dependence is not of interest.)

Even if the neutron guide optics were perfect, and did not absorb neutrons or cause unnecessary spread in divergence, for example, compromises will always be required for meeting all these criteria of a good incident beam. These compromises are quantified in the present section.

A fundamental problem with neutron optics is the size of the moderator. With a cross section of order 0.1×0.1 m^2, the moderator is far from being a point source of neutron emission. This gives a fundamental limitation to the "brightness," β, which is depicted with the three sources shown at the top of Fig. 6.8. All three moderators in Fig. 6.8 emit the same neutron current, and they send the same flux (current density) into the guides, which focus the rays on the sample below. The sources to the left have the higher brightness, and sources with higher brightness are better for making the smallest neutron beams on the sample. The focused spot on the specimen is, in fact, an image of the source itself, so it should be easiest to form a small spot when the source is small.

The source of Fig. 6.8c has the lowest brightness. Nevertheless, the focused beams in Figs. 6.8b and 6.8c have the same size. To make a small spot on the specimen with the low brightness source of Fig. 6.8c, however, the guide in Fig. 6.8c must provide stronger focusing, i.e., a larger angle of convergence. (Good focusing with a large angle of convergence requires higher quality optics.)

More quantitatively, the brightness of the source, β, is defined as the flux per solid angle $\left[\text{neutrons}/\left(\text{s cm}^2\text{ sr}\right)\right]$, measured at the source of the neutrons. Brightness is a valuable concept because brightness is a conserved quantity when the subsequent optics are ideal. For example, after a guide focuses the beam as in Fig. 6.8c, the width of the focused neutron beam is reduced by a factor of two compared to the source, but the angle of convergence is increased by a factor of 2. In other words, the flux has increased by a factor of 4, and the

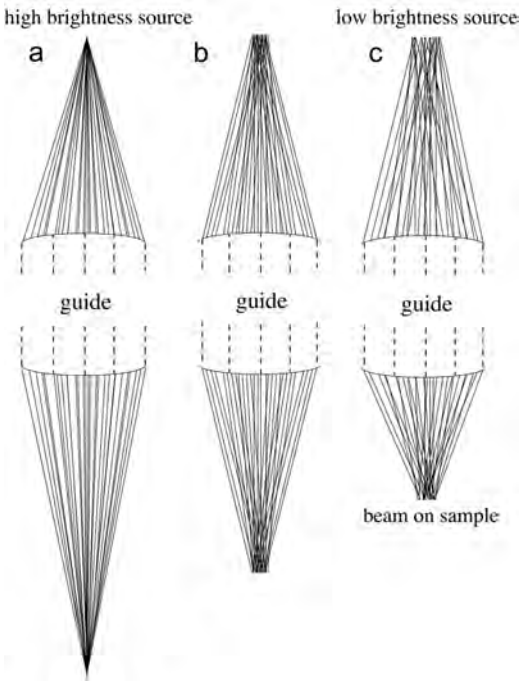


Figure 6.8: **a–c.** Formation of focused beams with sources of differing brightness. For all 3 moderators (at top) the neutron currents (number of lines) are the same, and the fluxes at the guide entrances are the same. The brightness of the sources decreases from left to right, owing to a larger area of the source.

solid angle has increased by a factor of 4, leaving unchanged the flux per solid angle (the brightness is conserved). Where the focused beam hits the specimen:

$$\beta = \frac{j_0}{\alpha_x \alpha_y} \; . \tag{6.23}$$

Here j_0 is the flux (neutrons/cm^2) in the beam on the specimen, α_x and α_y are the the angles of beam convergence in the x- and y-planes. We can relate the beam size to the brightness of the source and the convergence angles of the guide, assuming perfect guide optics. The beam width, d_0, is related to the total neutron current, I_p, by the relationship between current and flux:

$$I_\mathrm{p} = d_x d_y \, j_0 \; . \tag{6.24}$$

For simplicity, assume square cross-sections for the guide and moderator, so $d_x = d_y = d$ and $\alpha_x = \alpha_y = \alpha$. Substituting (6.23) into (6.24) and solving for d_0:

$$d_0 = \frac{\sqrt{\frac{I_\mathrm{p}}{\beta}}}{\alpha} \; . \tag{6.25}$$

Equation 6.25 shows that the beam width d_0 improves (becomes smaller) in proportion to the product $\alpha \sqrt{\beta}$, as suggested by the previous discussion of Fig. 6.8.

The compromises needed in guide design and experimental setups are therefore clear. We have to balance beam width against divergence. A high divergence impairs the Q-resolution of the instrument, but for a moderator of fixed brightness, a high divergence is required for a high neutron flux on a small sample. The trade-off is one of inverse proportionality. If Q-resolution is not an issue, however, small samples become more appropriate.

- To Do: plots of I(E) for ARCS

6.4 Fermi Choppers

A first glance at Figure 6.1 shows electronic equipment that looks a bit complicated, and in fact it is. The electromechanical control of a Fermi chopper is not simple, and requires specialized technology. The energy resolution of a Fermi chopper neutron spectrometer depends on timing, and much timing precision is demanded from the Fermi chopper. It must be open at a precise time after the proton pulse hits the neutron target. Microsecond precisions are needed for Fermi chopper timings, as we now show from an elementary calculation.

A neutron of 200.0 meV energy has a velocity of 6,185 m/s. It travels down a 11.6 m flight path in 1,876 μsec, where it encounters the ARCS Fermi chopper. If the energy resolution is to be 1 %, the velocity resolution needs to be 2 %. The Fermi chopper should therefore be open to the beam for 37 μsec. To ensure that all neutrons passing through the Fermi chopper have energies of 200 meV, this opening must occur at a reproducible time delay after each proton pulse hits the neutron target. Variations in this time delay correspond directly to an energy broadening (with no gain in intensity). The electromechanical control system for the Fermi chopper should ensure that the chopper is at the same angle of rotation after the required time delay, here 1,876 μsec. This "phasing accuracy" needs to be 2 to 4 μsec to be negligibly smaller than opening window of 37 μsec in our example.

This is a stringent demand – the slot in a rotating cylinder must be at the same orientation after each neutron pulse, with only a 2 μsec margin for error. For the ARCS spectrometer, the SNS proton pulse on target has a pulse frequency of 60 Hz, but has some drift over time associated with drift in the Tennessee power grid. An electromechanical feedback control system ensures synchronization of the Fermi chopper rotor with the proton pulse. Today these electronic units are somewhat smaller than the one being caressed by Enrico Fermi in Figure 6.1, but they are far more precise and reliable than his.

Another stringent demand can be understood from the opening of the slot in the rotor. The slot has an effective width that corresponds to perhaps 5° of the rotor circumference. How fast does the rotor need to spin for the opening time of 37 μsec in our example? This rotational frequency is (5/360)/(48 μsec) = 375 Hz. Such fast rotations present mechanical challenges for bearings and

heat dissipation.[4] In practice, Fermi choppers have magnetic bearings, and their rotors spin in a partial vacuum with some helium gas for heat transport. An obvious issue with our calculation of a 375 Hz rotational frequency is that it is not an integral multiple of the 60 Hz proton pulse frequency of the Spallation Neutron Source. In practice, we might select a 360 Hz rotational frequency so that the chopper can be open at a fixed time delay for every proton pulse. Note that the chopper will be open six times for each neutron pulse.[5]

There are different opinions on what to do with the other five pulses that could pass through the Fermi chopper for each proton pulse on target. The relationship between neutron energy and neutron velocity is:

$$E_n = 5.2276{\times}10^{-6}\ v_n^2\ . \tag{6.26}$$

For our hypothetical rotor spinning at 360 Hz (six rotations per proton pulse at 60 Hz), with a phase delay of 1,876 μsec to select a 200.0 meV energy, chopper openings will occur at time delays, τ, of:

$$\tau = 1,876 + n{\times}2,778\ [\mu\text{sec}]\ , \tag{6.27}$$

where $n = \{0, 1, 2, 3, 4, 5\}$, and 2,778 μsec is one-sixth of 1/60 sec. For these opening times and a 11.6 m flight path, the five incident energies are 200.0, 25.9, 9.6, 5.0, 3.0 and 2.0 meV. It might be possible to acquire additional inelastic scattering spectra from these five additional incident energies. This practice is called "repetition rate multiplication." It should be noted that in the simple Fermi chopper configuration proposed here, the energy resolution becomes impractically narrow for the lower incident energies, however.

The ARCS spectrometer includes a second chopper, a "T_0-chopper," located closer to the moderator than the Fermi chopper. This T_0-chopper is much like a Fermi chopper, and is phased to block the incident neutrons that would arrive at the Fermi chopper during our five other openings of the Fermi chopper. The T_0-chopper also serves to block the flash of γ-radiation and fast neutrons that are emitted by the moderator when the proton pulse hits the target.

The slot through a Fermi chopper has a gradual curvature to accommodate the time for neutron passage through the chopper rotor. This idea is shown in Figure 6.9 for a neutron depicted as moving from left to right. The positions of the neutron and the chopper are shown for three snapshots in time, with $t_1 < t_2 < t_3$.

The slot through the rotor of a Fermi chopper contains a stack of slits and slats, which allow the Fermi chopper to have good energy resolution while passing a wide neutron beam. The structure is shown in Figure 6.10 for a simple case with two "slats." The slats typically contain boron, a strong absorber of neutrons. The neutrons traversing the Fermi chopper must travel through

[4] Disk choppers are an alternative to cylindrical Fermi choppers. For a neutron beam of moderate width, the disk needs to be large, and centrifugal stresses become excessive at 375 Hz.

[5] There are actually twelve openings per proton pulse if you count the half-rotations where the back side of the rotor is facing the incident beam. For these cases, however, the rotor is upside-down in the figures below, and the curvature of the slats impedes the neutron transmission.

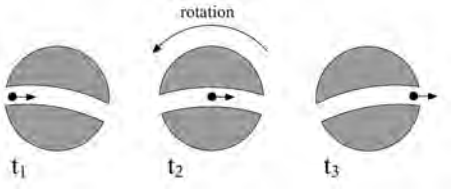


Figure 6.9: Fermi chopper, rotating counterclockwise, as a neutron moves from left to right. Compare the rotation direction and slot curvature to that of Fig. 6.2.

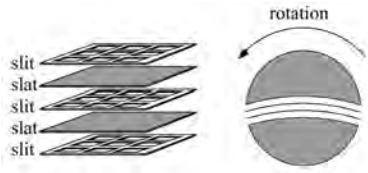


Figure 6.10: Fermi chopper rotor with two slats. The structure at left is an exploded view of the layered slits and slats that fill the curved slot through the rotor.

the narrow gaps called "slits." In practice these slits are plates of aluminum with many holes in them. The aluminum serves to maintain a precise spacing between the slats, a challenge at high rotational frequencies. In practice, a Fermi chopper may contain 10 to 40 slats. With many slats, it is possible for the Fermi chopper to pass a beam of 5 to 6 cm in width while maintaining good energy resolution. Not surprisingly, with narrower slits there is a loss of intensity that accompanies the improvement in energy resolution.

A drawing of the Fermi chopper used in the ARCS instrument is shown in Fig. 6.11. The neutron beam passes through the circular opening at the center of the housing. The electric motor, sensors, and magnetic bearings are above and below the rotor assembly. A corresponding pair of photographs is shown in Fig. 6.12. In the ARCS instrument, two identical chopper systems are mounted on a translation table so either can be moved into the neutron beam. The choppers can be moved while they are spinning, allowing quick changes to the energy of the incident beam of the ARCS instrument.

Rotor assemblies with curved slots, especially those with many slits and slats, are optimized to work at particular combinations of rotational frequencies, neutron velocities, and energy resolutions. These need to be selected in the planning stage of an inelastic neutron scattering experiment.

Finally, it is important to remember that Fermi choppers scatter neutrons out of the incident beam, generating a high intensity of scattered neutrons around a Fermi chopper. Good shielding is required, both for radiation protection for the experimenters, and for minimizing background in the instrument itself. The ARCS T_0 chopper helps alleviate this problem by minimizing the number of neutrons that are scattered out of the beam by the Fermi chopper.

6.5 Detectors

Neutron detectors usually rely on a nuclear transmutation reaction with a large cross-section. Widely used is ^{3}He, which absorbs a neutron and undergoes

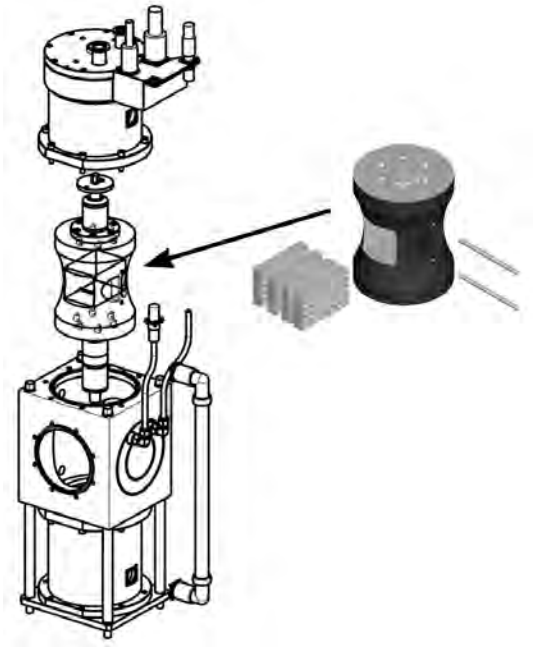


Figure 6.11: Exploded view of ARCS Fermi chopper, showing housing and rotor at left. The slit packages are held into the rotor with two pins, as shown at right.

Figure 6.12: Left: photograph of ARCS Fermi chopper, showing housing, windows, and water cooling. Right: photograph of the slit package, with aluminum slits seen end-on.

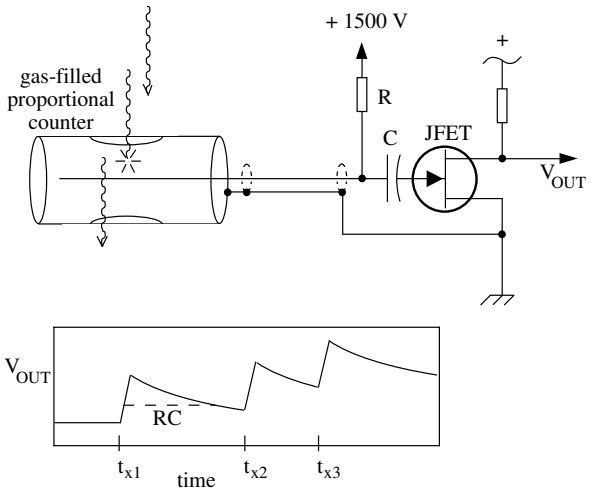


Figure 6.13: Input circuit for a simple charge sensitive preamplifier with a gas-filled proportional counter. The time-dependent voltage across the field effect transistor (FET) is indicated schematically after detection of an x-ray at times t_{x1}, t_{x2}, and t_{x3}.

fission

$$n + {}^3He \rightarrow {}^1H + {}^3H + 0.76\,MeV\ . \tag{6.28}$$

The ^{3}He is often used to fill a gas-filled proportional counter, a device that is old and simple, but is still often the best choice for a neutron detector. The hydrogen (protium) and tritium on the right-hand side of (6.28) form as energetic positive ions, and the free electrons are attracted to the anode wire in the center of the detector tube, which is biased at a high positive voltage. In the strong electric field near the anode wire, these electrons build up enough kinetic energy in a mean free path so they ionize additional gas atoms, and more electrons are created in this process of "gas gain." The gas-filled proportional counter was traditionally considered inexpensive, but more recently the high cost of ^{3}He fill gas has become a concern.

The electronics for the counter tube are configured as shown in Fig. 6.13 with a charge-sensitive preamplifier. The capacitor, C, integrates the negative charge collected on the anode wire, causing a quick rise in resistance across the field effect transistor. A small value of C allows for a large rise in voltage and good sensitivity. On the other hand, small stray capacitances between the detector and the preamplifier can have a detrimental effect on the detector signal, so interconnections between detector and preamplifier are kept as short as possible. The resistor, R, bleeds away the voltage across C with a much longer time constant. Typically, $RC = (10^7\Omega)(10^{-11}\text{F}) = 10^{-4}$ sec.

The ^{3}He gas-filled detector depicted in Fig. 6.13 can be modified to locate the position of the neutron arrival along the length of the detector. The modifications include the use of a resistive wire as the anode, and preamplifiers at both ends of the anode wire. By comparing the electron charges collected on the capacitors at the two ends of the detector, the position of the event can be determined. A neutron that ionizes the gas at one end of the detector tube

Figure 6.14: Installation of detector bank on the ARCS spectrometer, showing 16 tubes in the middle bank. Each tube is 1 m in height and approximately 2.5 cm in diameter.

produces a larger pulse in the preamplifier connected to that end. Such linear position-sensitive detectors require that the resistivity of the anode wire be steady with time, and not affected by contamination from the detector gas, for example. A two-dimensional array of side-by-side ^{3}He linear position sensitive detector tubes is often used as a 2-dimensional area detector. Modules of linear position-sensitive detector tubes are shown in Fig. 6.14, and the detector array of the spectrometer was completed by adding more detectors to the right of those shown.

6.6 Energy Resolution

Energy resolution is an important figure-of-merit for the design of direct geometry chopper spectrometers, and can be the most important figure-of-merit for incoherent scattering. The primary flight paths at the SNS are quite long, and the moderator pulses are short on flight paths 17 and 18, promoting good energy resolution. In evaluating instrument configurations, the secondary flight path is the parameter for adjusting energy resolution. When selecting a distribution of detector positions around the specimen, a spherical locus of detector positions provides energy resolution that is uniform at all angles around the sample.

Basic considerations for energy resolution were presented in the discussion of Fig. 6.3. Here we consider the issue further, starting with the moderator.

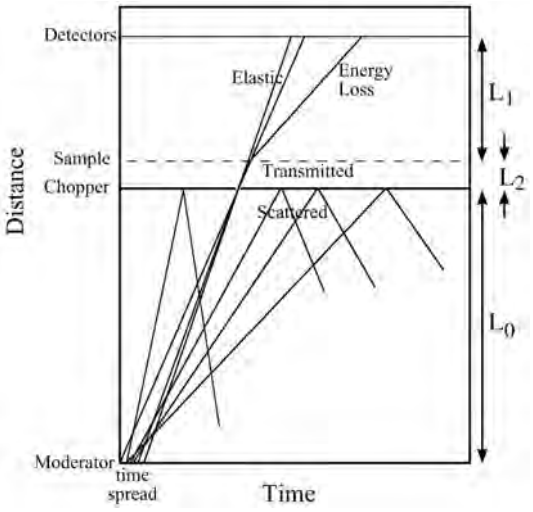


Figure 6.15: Distance–time diagram for a direct geometry chopper spectrometer. The figure is much the same as Fig. 6.3, but there is a spread in times when the neutrons leave the moderator. Positions of the moderator, chopper, sample, and detectors are marked on the vertical axis. Here the Fermi Chopper is open for only a minimal time.

The moderator does not produce an instantaneous pulse of neutrons, but the pulse has a time spread of several microseconds. This time spread of neutron emission, shown in Fig. 6.4b, is a gradual function of energy. The important consequence is a lower bound on the energy resolution. Even by reducing the opening time for the Fermi chopper to nearly zero seconds, a minimum energy resolution is caused by the spread of moderator emission times. Figure 6.15 presents a graphical explanation of the issue. In comparison to Fig. 6.3, note the spread in emission times at the position of the moderator in Fig. 6.15. Even though the Fermi chopper has a minimal open time, the neutrons passing through the chopper have a spread in slopes in Fig. 6.3, a spread in velocities, and hence a spread in energy.[6]

The energy resolution of a Fermi chopper spectrometer is often analyzed in terms of the moderator performance and the chopper performance. For analytical calculations, it is typical to assume that both the moderator and the chopper cause time smearing that has a Gaussian shape about an average time. The advantage to this approach is that one can convolute the different effects of time smearing to get another Gaussian function, whose width is obtained by adding in quadrature the widths from independent broadenings. The resolution is then obtained in terms of the following variables for time, energy, and distance:

[6]The paths from moderator to detectors in Fig. 6.15 are often compared to the ray diagram of a pinhole camera, where the pinhole is at the position of the chopper.

- L_0...distance from the moderator to the Fermi chopper
- L_1...distance from the sample to detector
- L_2...distance from the Fermi chopper to the sample
- t_r...time spread of neutrons passing through the Fermi chopper
- t_m...time spread of neutrons from the moderator
- δ_m...distance the neutrons travel in time t_m
- E_0...incident energy selected by Fermi chopper
- E_1...energy of the neutron leaving the sample

It is typical to express the energy resolution as a fraction of the incident energy, $\Delta E/E$. The Gaussian analysis performed by Windsor [C.G. Windsor] gives:

$$\begin{aligned}\frac{\Delta E}{E} &= \frac{2\delta_m}{L_0}\Bigg[\left(1+\left(\frac{E_1}{E_0}\right)^{3/2}\frac{L_2}{L_1}\right)^2 \\ &\quad +\left(\frac{t_r}{t_m}\right)^2\left(1+\left(\frac{E_1}{E_0}\right)^{3/2}\frac{L_0}{L_1}\left(1+\frac{L_2}{L_0}\right)\right)^2\Bigg],\end{aligned} \tag{6.29}$$

Energy resolution can be presented in various ways, and it is important to be careful when comparing plots for different instruments. Perhaps the most important criterion is the neutron flux on the sample for a given energy resolution, calculated for conditions that optimize the intensity. Figures 6.16 and 6.17 show such plots for the ARCS instrument. The first figure shows the performance for incident neutrons of 63 meV energy, approximately at the peak of the moderator brightness. These neutrons have a longer time spread for emission from the moderator, however, and this limits the ultimate energy resolution to about 2.5% of the incident energy. Figure 6.17 shows the same plot for neutrons of 250 meV incident energy. The ultimate resolution is better, perhaps 2% of the incident energy, in large part owing to moderator performance. The intensity is lower at 250 meV for the modest resolution of 4%, however. The neutron guide is not so effective at 250 meV, and the moderator spectrum is weaker at this energy.

6.7 Q Resolution

Optimization is quite different for Q resolution, however. This has been less of a focus of the ARCS instrument design, however, because historical trends have emphasized the design of instruments for work with polycrystalline samples. Measuring dispersive excitations in single crystals forces the consideration of

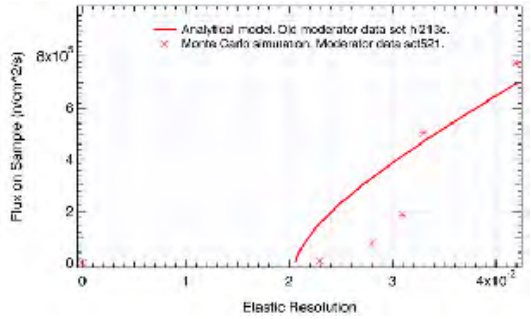


Figure 6.16: Flux on sample for various energy resolutions of the ARCS spectrometer for 63 meV incident neutrons.

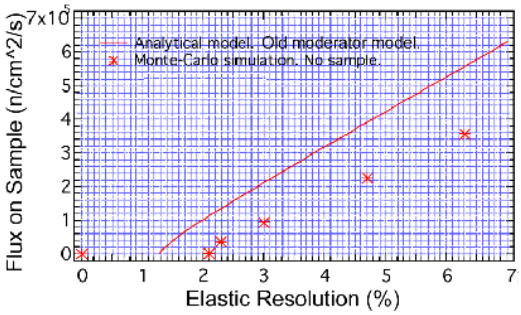


Figure 6.17: Flux on sample for various energy resolutions of the ARCS spectrometer for 250 meV incident neutrons.

Q resolution. Specifically, it is suggested that the instrument be optimized for fractional Q resolution. The figure-of-merit, $R_Q \equiv \Delta Q/Q$, is analogous to the figure-of-merit, $\Delta E/E$ for E resolution.

The parameter space for optimization is broad, and optimization depends on the type of measurement to be performed. Various types of excitations must be considered. The extreme cases are treated below as: 1) elastic scattering, 2) highly dispersive inelastic scattering and 3) dispersionless inelastic scattering.

6.7.1 Summary of Q Broadening

Four sources of Q broadening are considered:

- The finite size of the moderator, even over a long primary flight path, provides a significant angular divergence of neutrons on the specimen. This divergence is about 0.4° without a guide, and will be larger for thermal neutrons that have been reflected along the guide. If the excitations in the specimen do not select among these incident $\vec{k}_i$, the resolution of the experiment will be dominated by incident divergence.
- A 0.2° mosaic spread of a good crystal will cause an angular spread of 3.5 cm at 10 m. This is comparable to detector pixelation along a diagonal direction, which is approximately 2.5 $\sqrt{2}$ = 3.5 cm.
- The energy resolution leads to a coupled smearing in Q. Since the energy resolution is quite good for all flight paths, this source of Q smearing is

not the biggest problem for single crystal work, especially in the forward direction.

- The sin θ-dependence of Q conspires with finite detector pixel size to give a divergent l_3 in the the forward direction. A R_Q of 1.0 % is not possible for scattering angles, 2θ, smaller than 14° if a 10 m secondary flight path is used with 2.54 cm detectors. The problem is proportionately worse for shorter flight paths.

6.8 Optimization for $\Delta Q/Q$ in Elastic Scattering

It is easiest to first analyze the Q resolution for elastic scattering, and some of the issues for elastic scattering pertain directly to inelastic scattering.

6.8.1 Incident Divergence

The finite size of the moderator provides an angular divergence of incident neutrons on the sample. This is approximately 10 cm over 1350 cm, or about 0.4°. This incident divergence may or may not contribute to the divergence in scattered beams, depending on the process involved. For elastic Bragg scattering, for example, the incident divergence will allow intense Bragg peaks from all parts of a sample misoriented to within about 0.4°. Although incident divergence is a potentially serious problem, a perfect crystal will still provide sharp Bragg diffractions. This incident divergence can be the dominant effect on Q-resolution for polycrystalline samples, however.

6.8.2 Mosaic Spread

It is important to consider the limit to Q resolution caused by the sample itself – the instrument need not be much better than the intrinsic resolution of the sample. One issue is that a finite sample subtends a non-zero angle over the secondary flight path, l_3. This can be ignored for a typical 1 cm sample, which is much smaller than the expected size of detector pixels.

The "mosaic spread," denoting the mean mutual misorientations of different subvolumes in a crystal, is an important figure-of-merit for single crystal samples. A good single crystal sample may have a mosaic spread of 0.2 °, for which the angular blurring of an elastic beam at a detector located 10 m away from the sample is 3.4 cm. This is only slightly larger than typical detector pixel resolutions. Better instrument resolution would be a reasonable request, except for the fact that a 10 m sphere of detectors around the sample is prohibited by constraints of cost and space. Nevertheless, we note that a long secondary flight path, $L_1 = 10$ m, could be justified for single crystals of high quality in cases where the incident divergence does not dominate the Q resolution.

6.8.3 Locus of Constant $\Delta Q/Q$

We seek the locus of detectors that provides a constant $\Delta Q/Q$, defined as the resolution, R_Q:

$$R_Q \equiv \frac{\Delta Q}{Q} \, . \tag{6.30}$$

By definition

$$Q \equiv 4\pi \frac{\sin(\theta)}{\lambda} \, , \text{ so} \tag{6.31}$$

$$\Delta Q = 4\pi \frac{\cos(\theta)}{\lambda} \Delta\theta \, , \tag{6.32}$$

giving

$$R_Q = \cot(\theta)\Delta\theta \, . \tag{6.33}$$

When R_Q is plotted in polar coordinates, the detector locus is defined. This is shown in Fig. 6.18. This shape is quite incompatible with the spherical solution for constant $\Delta E/E$. At low angles, the detector placements required for constant $\Delta E/E$ and constant $\Delta Q/Q$ are exactly orthogonal.

The resolution R_Q diverges at small angles, because at small angles $Q \to 0$, so very small ΔQ is needed to maintain a constant $\Delta Q/Q$. A small angular spread for ΔQ is achieved only when the detector pixels are very small or are placed at very large L_1. At the other extreme, the detector distance for constant $\Delta Q/Q$ is infinitesimal when the scattering angle, 2θ, is 180 °, since small variations in angle have no effect on Q for direct backscattering.

6.8.4 Coupling of Q Resolution to E Resolution

One obvious incompatibility of the $\Delta Q/Q$ optimization and the $\Delta E/E$ optimization is that the secondary flight path, L_1, becomes vanishingly small at $\theta = 90°$. (This is strictly true only for the elastic scattering. For a fixed energy transfer, however, there will be a particular angle of inelastic scattering for which $\Delta Q = 0$ in Eq. (6.33).)

Poor energy resolution will cause a smearing in Q, increasing R_Q. For elastic scattering:

$$Q = k_f - k_i = \frac{\sqrt{2mE}}{\hbar} 2\sin\theta \, , \tag{6.34}$$

$$\frac{\mathrm{d}Q}{\mathrm{d}E} = \frac{\sqrt{2m}}{\hbar\sqrt{E}} \sin\theta \, , \tag{6.35}$$

$$\frac{\mathrm{d}Q}{\mathrm{d}E} = \frac{Q}{2E} \sin\theta \, , \tag{6.36}$$

$$\frac{\mathrm{d}Q}{Q} = \frac{\mathrm{d}E}{E} \frac{\sin\theta}{2} \, , \tag{6.37}$$

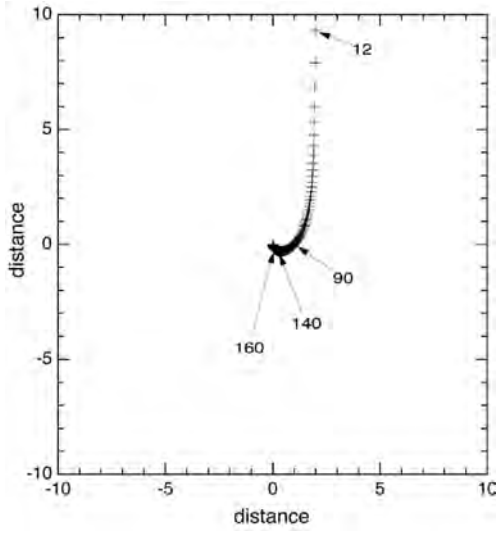


Figure 6.18: Detector locus for constant $\Delta Q/Q$. Crosses are spaced at 2° increments of 2θ . Labels indicate specific 2θ angles. Sample would be located at (0,0). Distance units are arbitrary, but they are equal for the two axes.

so $\Delta Q/Q$ and $\Delta E/E$ are comparable for modest angles.

The obvious solution is to use a spherical detector locus for high scattering angles. The 3.0 m flight path of ARCS gives energy resolutions as small as $\Delta E/E = 1\,\%$. At $2\theta = 40^\circ$ this contribution to $\Delta Q/Q$ is less than 0.2 %, and this is even smaller at lower angles.

6.9 Optimization of $\Delta Q/Q$ for Inelastic Scattering

6.9.1 Ewald Spheres and Incident Divergence

Three important cases are depicted in Fig. 6.2. Case 1, elastic scattering, was the topic of the previous Sect. 6.8. Case 2, dispersive inelastic scattering, is considered first. This problem is presented in two parts, the case for the first Brillouin zone, and the case for higher Brillouin zones. Case 3 of Fig. 6.2 is presented last, and it is more straightforward.

For a fixed energy loss, the value of $\left|\vec{k}_f\right|$ is constant, although $\vec{k}_f$ may take different orientations. The kinematics of this scattering are understood most conveniently by use of the Ewald sphere construction, which is shown for elastic scattering in Fig. 6.3a. The Ewald sphere construction is useful for analysis of diffraction conditions because it identifies the relationship between the wavevector transfer, $\vec{Q}$, and the reciprocal lattice vector, $\vec{\tau}$, showing when the Laue condition, $\vec{Q} = \vec{\tau}$, is satisfied for diffraction. In inelastic scattering the

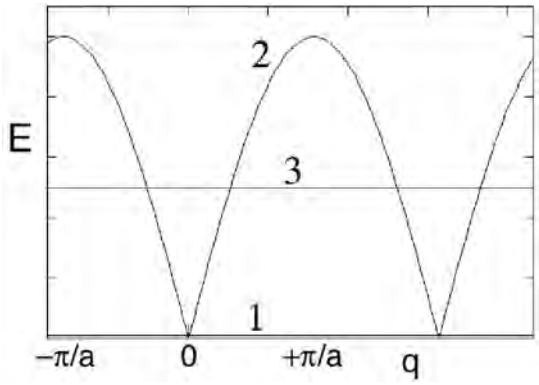


Figure 6.2: Three types of scattering processes (1) elastic scattering, (2) dispersive inelastic scattering, and (3) non-dispersive inelastic scattering.

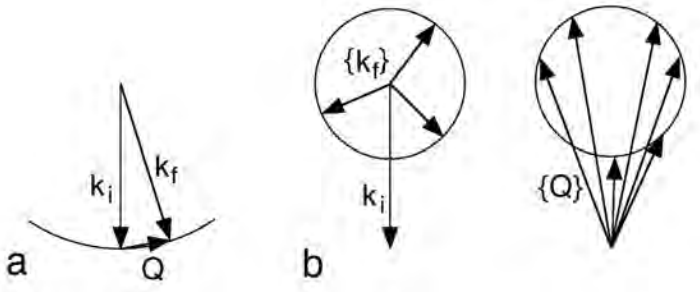


Figure 6.3: (**a**) Ewald sphere construction for elastic scattering, showing allowed $\vec{Q}$ for Bragg diffraction. (**b**) Ewald sphere construction for inelastic scattering, showing allowed $\{\vec{Q}\}$.

momentum of the excitation, $\vec{q}$, plays a similar role to $\vec{\tau}$, especially in the first Brillouin zone where τ is zero.

For the inelastic scattering process shown in Fig. 6.3b, all $\{\vec{k}_f\}$ have the same length (as for the elastic case of Fig. 6.3a), since the energy E_f is identical for all orientations of $\vec{k}_f$. In both Figs. 6.3a and 6.3b, the allowed $\vec{k}_f$ make a sphere, which when placed at the tail of $\vec{k}_i$, defines the allowed $\{\vec{Q}\}$ as shown in the figure (where as usual $\vec{Q} \equiv \vec{k}_f - \vec{k}_i$).

The advantage of the Ewald sphere constructions of Fig. 6.3 is in analyzing tilts of the incident beam, as occurs for incident divergence, for example. The cases of Fig. 6.4 show conditions for elastic scattering before and after tilt. It is evident that the condition for momentum conservation (the Laue condition)

$$\vec{Q} - \vec{\tau} = 0 \tag{6.38}$$

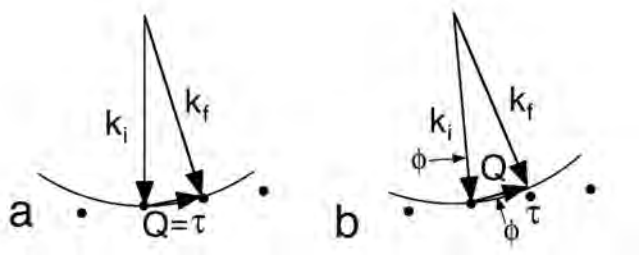


Figure 6.4: Ewald sphere construction for elastic scattering (**a**) before, and (**b**) after tilt of $\vec{k}_i$. Notice that $\vec{Q}$ no longer touches the reciprocal lattice vector $\vec{\tau}$ after tilt.

is generally violated by tilting, because tilts of $\vec{k}_i$ by the angle ϕ cause $\vec{Q}$ to tilt by this same angle ϕ. The consequence of this violation of momentum conservation is that coherent elastic scattering is altered, and Bragg diffractions may be eliminated, for example. Those neutron trajectories with improper $\vec{k}_i$ will not contribute to the Bragg diffraction.

6.9.2 First Brillouin Zone – Soft Dispersions

One case for inelastic scattering is shown in Fig. 6.5a. This case is a scattering from the first Brillouin zone, where the reciprocal lattice vector, $\vec{\tau} = 0$. The diameters of the two circles are set by the energy and momentum transfer to the excitation. In Fig. 6.5a, $\left|\vec{k}_f\right| \simeq \sqrt{2}\left|\vec{k}_i\right|$, indicating that the neutron has lost half its energy to the solid. For the present case where the excitation is of energy $E_i/2$, we further assume a relatively soft dispersion where the momentum of the excitation q is assumed the same as the momentum of the neutron, k_f. A circle is used for the locus of acceptable $\vec{q}$ (a circle is not expected for anisotropic crystals, of course). This circle has a large radius, as may be expected if a dispersive excitation has a soft dispersion relation.

The effect of tilt on the scattering condition is shown in Fig. 6.5b. It is evident by geometry that the tilt of $\vec{k}_i$ and $\vec{q}$ are by the same angle for this case of $\vec{q}$ lying in the first Brillouin zone. This could affect the ability of the neutron to excite the dispersive mode, at least if the dispersive mode is sharp in $\vec{q}$. Conditions where highly anisotropic dispersions have a delicate contact with the locus of $\vec{Q}$ can also be envisioned.

6.9.3 Higher Brillouin Zones – Stiff Dispersions

The case of a high energy excitation is shown in Fig. 6.6. In this case the dispersion is very steep, and the energy is so high that the energy of the excitation is large, even when q is small. In this particular case, there are no excitations in

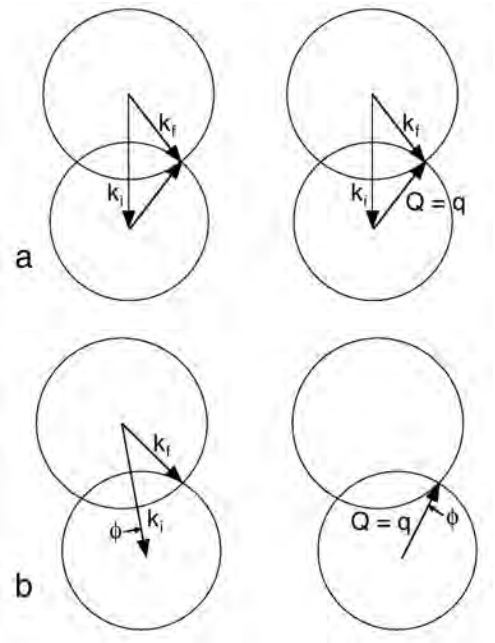


Figure 6.5: Ewald sphere construction for inelastic scattering (**a**) before, and (**b**) after tilt of $\vec{k}_i$. Notice that although $\vec{Q}$ has the same magnitude, it changes orientation by the same tilt angle, ϕ, as the incident beam.

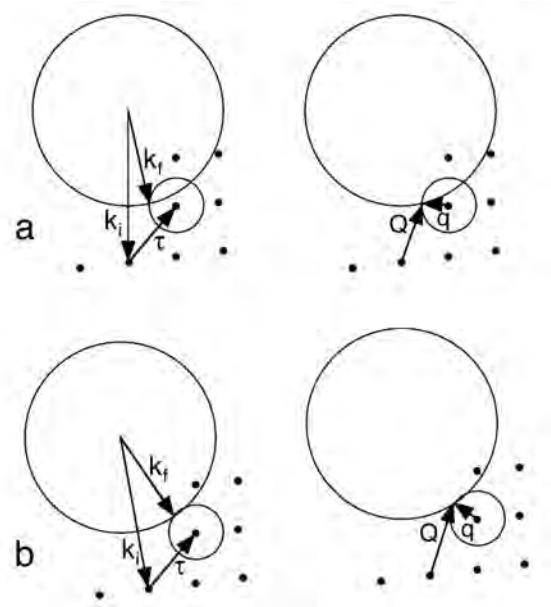


Figure 6.6: Ewald sphere construction for inelastic scattering (**a**) before, and (**b**) after tilt of $\vec{k}_i$. In this case the excitation has a large E for a small q. Notice that although $\vec{Q}$ has the same magnitude, it changes orientation strongly after tilt of the incident beam.

the first Brillouin zone because $q < Q_{min}$.[7] The excitation can occur in a higher Brillouin zone with the help of a reciprocal lattice vector, $\vec{\tau}$ so that

$$\vec{Q} - \vec{q} - \vec{\tau} = 0 \, . \tag{6.39}$$

This case of a high-energy excitation is a higher Brillouin zone is interesting because a tilt of $\vec{k}_i$ does not induce a tilt of $\vec{Q}$ by the same angle. Figure 6.6 shows that small tilts have big effects on the orientation of $\vec{q}$. In essence, the reciprocal lattice vector can amplify the tilt of the incident beam on the rotation of $\vec{q}$. In this case it seems plausible that for sharp, stiff excitations, only a $\vec{k}_i$ of the correct orientation can generate the excitation. The mosaic spread of the sample will be able to pick these acceptable $\vec{k}_i$ from the incident beam, and the Q resolution of the experiment will originate with the mosaic spread of the sample and not the incident divergence.

[7] Experimentally, it might be prudent to use a larger E_i and k_f so that excitations in the first zone are allowed. In this case the result of Sect. 6.9.2 is recovered, that is the tilt angle of $\vec{k}_i$ equals the tilt angle of $\vec{q}$. The present results from higher Brillouin zones remain valid, however.

6.9.4 Non-Dispersive Excitations

The case of non-dispersive excitations in Fig. 6.2 is straightforward to analyze with Figs. 6.5 and 6.6. The point is that the circle of q can be of arbitrary radius, since the energy of the excitation does not depend on q. For this reason, at a fixed energy transfer equal to that of the excitation, there will always be an excitation with an appropriate q to satisfy momentum conservation. All orientations of $\vec{k}_i$ will be useful for generating the excitation, and the incident divergence will dominate over the crystal mosaic spread in setting the Q resolution. A similar result pertains to inelastic incoherent scattering.

6.10 Background

Much of the work in commissioning a new inelastic neutron spectrometer is eliminating sources of background in the measurements. Finding and exorcising the different types of background contributions is a dark art, but some general guidance is possible.

The usual sources of background originate with neutrons that 1) were once good neutrons in the incident beam, or 2) are bad neutrons that entered the instrument uninvited. Unwanted neutrons can enter the instrument from the experiment hall around the instrument. A heavy shielding of borated hydrocarbons around the detector chamber can help protect humans in the experiment hall, but it is at least as important for keeping neutrons out of the instrument.

Perhaps a bigger problem comes from fast neutrons generated in the proton target region. Not all of these are moderated to reasonable velocities, and some of these fast neutrons enter the instrument and cause trouble. Concerning Fast Neutrons:

- Even with a heavy T_0 chopper made from iron and nickel, some 0.1% to 1% of the fast neutrons typically pass through. They then scatter in a number of ways from materials in the sample chamber and detector chamber, and can lose enough energy to be detected efficiently by ^{3}He detectors.
- On the way towards the detector chamber, either before or after the T_0 chopper, fast neutrons can interact with materials in the primary flight path, such as concrete shielding, and lose energy. Background problems from these "moderated" neutrons are tricky to diagnose because they can have a time structure similar to the wanted neutrons in the incident beam. They may, however, scatter from parts of the instrument that are nominally protected from the incident beam and cause unexpected effects.

Neutrons wanted in the incident beam can scatter in unwanted ways. Concerning Neutrons in the Incident Beam:

- The first concern in designing a neutron scattering experiment is to keep all materials out of the incident beam except for the sample itself. Obviously anything exposed to the incident beam (sample support, heating

elements, pressure cell) will scatter neutrons in much the same way as the sample.

- Peculiar effects can occur when neutrons scattered from the sample find their way into another material in the sample chamber or detector chamber, and scatter a second time. Multiple scattering is a problem inside thick samples themselves, but it is even trickier to sort out if a neutron scattered by the sample is again scattered by a pressure cell and then goes into the detectors of the instrument. Calculating multiple scattering is the topic of Section 8.2, but avoiding multiple scattering should be a goal of experiment design.

- There is often a "small-angle elastic scattering" from small-scale density variations in the sample or in the sample environment equipment. The forward beam is broadened as it leaves the sample, and the albedo of the beam may not all enter the "get lost" pipe and beam stop. These neutrons may scatter from the forward-mounted detectors, giving a source of neutrons that shines back into the spectrometer. Radial baffles can help hide detectors from scattering from other detectors across the instrument. A radial collimator around the sample itself also proves effective in suppressing scattering diametrically across the detector array.

Evidently there are numerous risks of background from neutrons that enter the instrument either invited or uninvited. An older viewpoint is that since the energies and directions of these neutrons are hard to understand, they can be modeled as a gas with a wide spectrum of energies. In other words, these unwanted neutrons go everywhere. Covering everything with absorbing material is good practice.

A more refined practice is to use a radial collimator around the sample so that the scattered neutron beams from the sample must take a straight line to the detector array. Neutrons that are scattered by other materials away from the sample position are absorbed by the walls of the collimator and do not reach the detector. (Usually this collimator oscillates by a small angle so it does not cast a shadow, and its absorption characteristics are averaged across the detectors.) Very large improvements in background are possible with an oscillating radial collimator, as shown in Sect. 8.3.4.

Last but definitely not least, another important practice to reduce background is to eliminate parts of the incident neutron beam that cannot illuminate the sample. If the sample is small and intercepts only a fraction of the area of the incident beam, it is often a good idea to eliminate the rest of the incident beam by upstream collimation before the beam reaches the sample. This is less of a concern when the sample is isolated by being hung on a thin wire or mounted at the end of a thin stick, for example. In the case of an isolated sample, the outer neutrons that miss the sample will go straight into the get lost pipe and beam stop. However, if there is anything around the sample that could be hit by the incident beam, the best practice is to remove this extra material, or collimate the incident beam so incident neutrons do not hit it.

6.11 Sample Design

Inelastic neutron scattering experiments require careful attention to the sample itself. Of course the sample must be in the correct state, satisfying basic criteria such as magnetic polarization, chemical composition, crystal structure, crystallographic quality or texture, temperature and pressure. Criteria for these sample states must not be compromised to the detriment of the science. The in-situ control of temperature, pressure, and magnetic field often leads to some compromises in experiment design, and expertise with the available "sample environment equipment" is essential in planning an experiment.

A common problem for inelastic scattering experiments is a sample that is too small. There is always a background in the spectrum from stray neutrons in the instrument, and inelastic neutron scattering experiments may be ruined by poor signal-to-noise ratios whenever this background has an intensity comparable to the scattering from the sample itself. Robust scattering from the sample usually means having a large quantity of sample. Experimenters are often dismayed that inelastic scattering measurements typically two orders-of-magnitude more sample than is typical for neutron diffraction measurements. The ideal sample for inelastic scattering will be as wide as possible so that it fills the cross section of the incident beam. With a wide sample, all neutrons in the beam will be candidates for scattering, and no parts of the beam will pass directly into the beamstop or generate unnecessary background.

The thickness of the sample in the direction of the beam needs to be chosen with care. It is possible to have a sample that is too thick. There are two limits that set upper bounds on the thickness of samples, and we consider each in turn below. The sample should be as thick as possible, provided it does not exceed limits imposed by:

- Multiple scattering. If the scattering cross-section is large and the sample is thick, too many neutrons will be scattered multiple times inside the specimen. Each scattering has its own energy spectrum, and the convolution of these spectra is difficult to sort out in the measured data.

- Absorption. Some nuclei are strong absorbers of neutrons. For thick samples of strong absorbers, few neutrons can leave the samples and enter the detectors.

6.11.1 Scattering and Attenuation

As a neutron beam passes through a material, there is a reduction in the number of neutrons in the forward beam. At the depth x, the increment of thickness of a material, $\mathrm{d}x$, scatters a number of neutrons, $\mathrm{d}I$, removing them from the beam. The number of scattered neutrons, $-\mathrm{d}I(x)$, equals the product of 1) the increment of thickness, $\mathrm{d}x$, 2) the number of neutrons present at x, $I(x)$, and 3)

a material coefficient, s:

$$-\mathrm{d}I(x) = \mathsf{s}\, I(x)\, \mathrm{d}x \,, \tag{6.40}$$

$$\frac{\mathrm{d}I(x)}{\mathrm{d}x} = -\mathsf{s}\, I(x) \,, \tag{6.41}$$

$$I(x) = I_0\, \mathrm{e}^{-\mathsf{s}x} \,. \tag{6.42}$$

If scattering is the only process that removes neutrons from the incident beam as it traverses a sample of thickness t, the loss from the beam equals the intensity of the scattering:

$$I_{\mathrm{scat}}(t) = I_0 \left(1 - \mathrm{e}^{-\mathsf{s}t}\right) \,. \tag{6.43}$$

The product in the exponent, $\mathsf{s}x$ or $\mathsf{s}t$, must be dimensionless, so s has dimensions of [cm^{-1}]. When $\mathsf{s}x$ is small, it equals the fraction of neutrons removed from the incident beam. From Fig. 1.4 we know that this fraction also equals $N\sigma/A$, so:

$$\mathsf{s} = \frac{N\sigma}{Ax} = \frac{N}{V}\sigma \,, \tag{6.44}$$

where N/V has units [atoms cm^{-3}] and σ is the scattering cross-section with units [cm^2].[8]

It is straightforward to calculate the composite mass attenuation coefficient for a compound or an alloy. The point to remember is that the total neutron scattering depends on the number and types of atoms in the path of the beam. The composite attenuation coefficient is obtained from the attenuation coefficients, s_i, for the different elements, i, weighted by their atomic fractions in the material, f_i:

$$\langle \mathsf{s} \rangle = \sum_i f_i\, \mathsf{s}_i \,. \tag{6.45}$$

6.11.2 Multiple Scattering Criterion

It may seem curious that in a well-planned experiment, most of the neutrons are transmitted through the sample without scattering.[9] Using a very thick specimen can cause the data to be uninterpretable, as we now show.

Consider the probability of inelastic scattering, p_{i}, and elastic scattering, p_{e}, through a thin layer of material. We set $p_{\mathrm{e}} + p_{\mathrm{i}} = p$, where p is the total probability of scattering in the layer. For thin samples of n layers, each of

[8] Since density varies with the type of material, "mass attenuation coefficients" are often normalized as ratios s/ρ. Here the density, ρ, has units [$\mathrm{g\,cm}^{-3}$], so the coefficients s/ρ have units [cm^{-1}]/[$\mathrm{g\,cm}^{-3}$]=[$\mathrm{cm}^2\ \mathrm{g}^{-1}$]. Exponents in (6.42) are products $(\mathsf{s}/\rho) \times \rho \times x$, and are, of course, dimensionless.

[9] Often the next largest fraction of neutrons are scattered elastically, and the smallest fraction are scattered inelastically.

thickness x, we have the relationship between numbers of layers and the bulk scattering coefficient, s:

$$\langle \mathsf{s} \rangle x = n\, p \ . \tag{6.46}$$

For consideration of multiple scattering, it is most convenient to work with the numbers of layers and scattering probabilities.

A challenge for multiple scattering calculations is the three-dimensionality of the sample. Neutrons can be scattered initially upwards, followed by a second scattering to the side, and a third scattering out to the detectors. A computer program such as MSCATT or MCViNE is required for accounting for such possibilities. Here we provide an approximate analysis in one dimension.

We assume that all neutrons entering the layered sample eventually leave the back of the specimen and are observed. The probability, p' for a neutron being scattered j times is the product of j of the layers doing a scattering, and $n - j$ of the layers doing no scattering:

$$p' = p^j (1-p)^{n-j} \ . \tag{6.47}$$

We are not keeping track of which particular layer has done the scattering. We keep track of the fractions of neutrons that are scattered j times in our n layers when the scattering probability in each layer is p. The number of ways of arranging the j scatterings over the n layers is therefore the binomial coefficient. The total fraction of neutrons that are observed in our example is 1, and should equal the sum of all scatterings, including $j = 0$ for no scattering:

$$1 = \sum_{j=0}^{n} \frac{n!}{(n-j)!\ j!}\, p^j (1-p)^{n-j} \ . \tag{6.48}$$

We have implicitly assumed $p \ll 1$, so the number of scatterings $j \ll n$, and each layer scatters at most once. (This also ensures that the terms for which $j \sim n$ are negligible in (6.48).) We can always satisfy this assumption $p \ll 1$ by dividing our sample into finer and finer layers, because (6.46) shows us that for a fixed sample, np is a constant. Write out the first few terms of (6.48) in the binomial expansion:

$$\begin{aligned} 1 \quad = \quad & (1-p)^n + np(1-p)^{n-1} + \frac{n(n-1)}{2}\, p^2 (1-p)^{n-2} + \\ & + \frac{n(n-1)(n-2)}{6}\, p^3 (1-p)^{n-3} + \ldots \end{aligned} \tag{6.49}$$

We simplify by using our condition of large n and small p,

$$\begin{aligned} 1 \quad = \quad & (1-p)^n + (np)^1 (1-p)^{n-1} + \frac{1}{2}\, (np)^2 (1-p)^{n-2} + \\ & + \frac{1}{6}\, (np)^3 (1-p)^{n-3} + \ldots \end{aligned} \tag{6.50}$$

The first term in (6.50) is the the probability of zero scattering, the second is the probability of one scattering, the third – two scatterings, etc. We are interested in knowing the ratios of the different terms in this series. Since $1 - p \simeq 1$, and using (6.46), the ratios are:

$$1 \; : \; \mathsf{s}x \; : \; \frac{1}{2}(\mathsf{s}x)^2 \; : \; \frac{1}{6}(\mathsf{s}x)^3 \; ... \tag{6.51}$$

The parameter $\mathsf{s}x$ is dimensionless, and for experiment design we seek $\mathsf{s}x \simeq 0.10$. In other words, we seek a sample that scatters 10% of the incident neutrons. For this particular case, (6.51) gives the ratio of double scattering to single scattering of $1/2\mathsf{s}x = 0.05$. Consider the effect of this double scattering on the inelastic spectrum. The scattering probability is the sum of an inelastic and elastic probability:

$$p = p_\mathrm{e} + p_\mathrm{i} \, , \tag{6.52}$$

$$np = np_\mathrm{e} + np_\mathrm{i} \, , \tag{6.53}$$

$$\langle \mathsf{s} \rangle x = \langle \mathsf{s}_\mathrm{e} \rangle x + \langle \mathsf{s}_\mathrm{i} \rangle x \, . \tag{6.54}$$

For this sample that scatters 10% of the incident neutrons, the ratio of the amount of inelastic single scattering to the amount of inelastic double scattering will be 5% – the same ratio as for the total scattering.

The effect of multiple scattering is to smear out the measured inelastic energy spectrum, and spread it out over twice the energy range of the single-scattering spectrum. Suppose the first scattering has the double-differential cross section of (2.52). Suppose a particular inelastic scattering causes an energy loss $\epsilon = \hbar\omega$. If this scattered neutron with altered energy and wavevector now undergoes a second inelastic scattering, the process may involve another double-differential cross section of (2.52), but with somewhat different parameters. An energy spectrum associated with this second scattering will be associated with every energy loss from the first scattering. Approximately, the energy spectrum for double scattering is the convolution of the single-scattering energy spectrum with itself. A similar type of argument applies to momentum distribution, but this is complicated by the vectorial aspects of the momentum transfer. Today the accepted practice is to ignore multiple scattering by assuming it is a weak background underneath the measured data, and does not contain significant structure. So long as it represents only 5% of the spectral area, this shouldn't be a problem, should it?

A more insidious problem occurs when the sample is near other material that can scatter neutrons, such as a heater or a cold finger for sample temperature control. Depending on geometry, neutrons can be scattered from the sample, to this support material, to the detector. The result may be a new feature in the inelastic spectrum, such as a spurious peak from an excitation in the support material. What is so insidious is that by removing the sample, this spurious peak will disappear. It is therefore tempting to assign this peak to an excitation in the sample, not in the support material. This problem was discussed qualitatively in Sects. 6.10, and a specific example is calculated quantitatively in Sect. 8.3.4.

6.11.3 Absorption Criterion

When the sample contains a strong absorber of neutrons, the analysis is a bit more complicated than for (6.42). The loss of neutrons from the incident beam is now

$$-\mathrm{d}I(x) = (\mathsf{s}+\mathsf{a})\, I(x)\, \mathrm{d}x\ , \tag{6.55}$$

$$I(x) = I_0\, \mathrm{e}^{-(\mathsf{s}+\mathsf{a})x}\ . \tag{6.56}$$

where the parameter a is the absorption coefficient, causing neutrons to "disappear," rather than scatter. The $I(x)$ in (6.56) is the number of neutrons at depth x into the specimen, but Eq. 6.43 does not give the number of scattered neutrons. Furthermore, the scattered neutrons can be absorbed on their way out of the sample. With a one-dimensional model, for a sample of thickness t, the path out has a length $t - x$, so an exponential function with this argument attenuates the outgoing beam. For the scattered neutron intensity, I_{scat} from the increment $\mathrm{d}x$ is

$$\mathrm{d}I_{\mathrm{scat}} = I(x)\, \mathrm{e}^{-\mathsf{a}(t-x)}\, \mathsf{s}\, \mathrm{d}x\ . \tag{6.57}$$

Substituting (6.56) into (6.57), rearranging, and integrating $\mathrm{d}I_{\mathrm{scat}}$:

$$I_{\mathrm{scat}} = I_0\, \mathsf{s}\, \mathrm{e}^{-\mathsf{a}t} \int_0^t \mathrm{e}^{-\mathsf{s}x}\, \mathrm{d}x\ , \tag{6.58}$$

$$I_{\mathrm{scat}} = I_0\, \mathrm{e}^{-\mathsf{a}t}\, (1 - \mathrm{e}^{-\mathsf{s}t})\ . \tag{6.59}$$

We can use (6.59) to calculate the scattering for any combination of absorption coefficient (a) and scattering coefficient (s). Suppose that $\mathsf{a} = 0$, and there is no absorption. For a sample that scatters 10% of the incident neutrons, we obtain (6.43). Incidentally, for such a thin scatterer we can expand the exponential in (6.43) to obtain $I_{\mathrm{scat}} = I_0\, \mathsf{s}\, t$, which is perhaps more intuitive. Now consider the other limit where absorption is strong. A sample that is too thick will have no scattered intensity, since all neutrons are absorbed. There is an optimal thickness t' for maximum scattering that we find by the analysis:

$$\left.\frac{\mathrm{d}I_{\mathrm{scat}}}{\mathrm{d}t}\right|_{t'} = 0 \tag{6.60}$$

$$t' = \frac{1}{\mathsf{s}}\, \ln\left(\frac{\mathsf{s}}{\mathsf{a}} + 1\right)\ . \tag{6.61}$$

For strong absorbers, $\mathsf{a} \gg \mathsf{s}$, and we obtain from (6.61):

$$t' = \frac{1}{\mathsf{a}} \quad \text{optimal thickness for strong absorbers.} \tag{6.62}$$

6.11.4 Hydrogen Criterion

Finally, we consider the inelastic scattering from hydrogen, a unique element. Phonon scattering cross sections are proportional to σ/m. Hydrogen has a

Table 6.1: Neutronics of $LiFePO_4$

	units	Li	Fe	P	O	O_4
σ_{scat}	10^{-24} cm^2	1.37	11.62	3.312	4.232	19.93
σ_{abs}	10^{-24} cm^2	70.5	2.56	0.172	0.0	0.0
σ_{scat}/m	10^{-24} cm^2/A	0.196	0.208	0.107	0.265	1.06

uniquely large σ and an uniquely low m, making it a stronger inelastic scatterer than Zr, by a factor of 1100. A trace of hydrogen in Zr would contribute a large amount of the inelastic scattering, and Zr has a tendency to absorb hydrogen and retain it in a modest vacuum. In many cases the spectral contributions from hydrogen are at high frequencies, and can perhaps be separated from the modes from Zr, for example. Nevertheless, it is important to know the hydrogen concentration in a sample for inelastic scattering experiments, and the best practice is to eliminate hydrogen from the sample.

6.12 Sample Design: Worked Example of $LiFePO_4$

Here are some sample calculations that were used for obtaining a thickness and mass of a typical sample for inelastic scattering. The neutronic properties of the elements of lithium iron phosphate are listed in Table 6.1.

The first step is to calculate the mass, M, of a sample that is a 10% scatterer. We do the calculation for 1 cm^2 of sample area, so the total cross section should be 0.1 cm^2

$$\begin{aligned} \sigma_{\text{tot}} = 0.1\,\text{cm}^2 &= [1.37 + 11.62 + 3.31 + 16.93] \\ &\quad \times [10^{-24}\,\text{cm}^2]\left[\frac{1\text{ mole}}{158\,g}\frac{6.02\times 10^{23}\text{ atoms}}{\text{mole}}\right] M\,. \quad (6.63) \\ M &= 0.79\,\text{g}\,. \quad (6.64) \end{aligned}$$

To fill the 5×5cm beam of the ARCS instrument, the sample mass should be 19 g.

The second check is for absorption. The total absorption for a 1 cm^2 sample is:

$$\begin{aligned} \sigma_{\text{abs tot}} &= [70.5 + 2.56 + 0.176] \\ &\quad \times [10^{-24}\,\text{cm}^2]\left[\frac{1\text{ mole}}{158\,g}\frac{6.02\times 10^{23}\text{ atoms}}{\text{mole}}\right] 0.79\,\text{g}\,. \quad (6.65) \\ \sigma_{\text{abs tot}} &= 0.219\,. \quad (6.66) \end{aligned}$$

This is small enough to be ignored, but if it were much larger we would have to resize the sample to be consistent with (6.62), for example.

Finally, we consider the risk from some residual water in the sample. One H_2O molecule has a ratio of $\sigma_{scat}/m = 170$ for its hydrogen atoms. The σ_{scat}/m in Table 6.1 are nearly 1000 times smaller. We should seek a mole fraction of water that is 10^{-4} or less. Perhaps some water could be removed by heating the sample slightly in the vacuum of the ARCS spectrometer.

Through data analysis procedures described in Chapter 6, it is possible to convert a measured inelastic spectrum to a representation of the phonon DOS. The problem of "neutron weighting" in $LiFePO_4$ can be seen from the last line in Table 6.1 in the values of σ_{scat}/m. For this material, the vibrational modes involving large amplitudes of motion for Li atoms will be much more prominent in the extracted DOS than the modes involving large amplitudes for O atoms. This neutron weight problem is one that can be addressed by lattice dynamics calculations, for example.

6.13 Sample Design: Sachets for Powders

Powder samples can be challenging to mount and accommodate in furnaces and cryostats. Here is an illustrated example of a successful design for a sachet (a small bag with a French name) filled with powdered SrO. It was designed to be heated to 1100°C in a vacuum furnace. Prime concerns were:

- The sample must not spill from the container. Spills of powder inside neutron instruments are serious issues, and require professional decontamination. After exposure to neutrons, many materials become "activated," meaning that they are radioactive. If the powder is also chemically toxic, the spilled powder is called "mixed waste," which is one of the highest categories of hazardous materials. Do not spill powders of samples used for neutron scattering research.

- The sachet should be made of thin sheet material so it does not contribute significantly to the inelastic scattering background. Neutron facilities often supply vanadium sample containers because these are well suited for neutron diffraction. Although vanadium has no coherent scattering (suppressing diffraction peaks), vanadium still has considerable inelastic scattering. For inelastic scattering, niobium can be a good choice for high temperatures, and aluminum is good for low temperatures.

- For experiments in furnaces, the sample should not react with the sachet at high temperatures. In some cases it is possible to examine a phase diagram to see if any compound made from elements in the container and elements in the sample has a low melting temperature. If so, the highest temperature of the experiment should be set with caution. It is usually a good idea to test the combination of the sachet and the sample in a furnace well before going to the neutron experiment.

- The sample thickness should be approximately uniform, and remain stable with handling and mounting. Shaping loose powders into flat plates

Figure 6.7: A partially-complete sachet. A sheet of thin niobium was folded in half, and the edges were folded multiple times in ever sharper triangle shapes. A pair of pliers will be used to crimp the edges.

is challenging, and a technique is shown below.

Figure 6.7 shows a sachet constructed of thin (50 micron) niobium sheet. The edges comprise many thicknesses of material, but they will be shielded from the incident beam. It is useful to draw the shape of the sample sachet in the lab notebook. An important trick shown in Fig. 6.7 is that the opening for the sample is formed around a plate of appropriate thickness. In this case a ruler proved the right thickness, and was used to shape the sachet.

Figure 6.8 shows that it usually takes two pairs of hands to fill the sachet with powder. It takes a bit of skill to ensure that the sachet does not bulge excessively in the middle, or to ensure that the powder does not settle to the bottom with further handling. Figure 6.9 is a reminder to weigh every sachet both before and after loading with powder.

Figure 6.10 shows a boron nitride frame around the sample sachet, which serves to block the incident beam from hitting the edges of the sachet. The frame should face the incident beam, looking upstream from the sample. For this experiment, the sample plane normal was tilted +45° with respect to the beam, so the normal to the sample pointed to detectors at +45°, and the plane of the sample intersected detectors at +135°, giving a low scattered intensity at this angle. It is also appropriate to adjust collimation slits in the incident beam to confine the height and width of the beam to the central region of the sample.

Further Reading

The contents of the following are described in the Bibliography.

Figure 6.8: Loading powder into the sachet. For valuable powders it is appropriate to do the loading over clean aluminum foil so that any spilled powder can be recovered and reused.

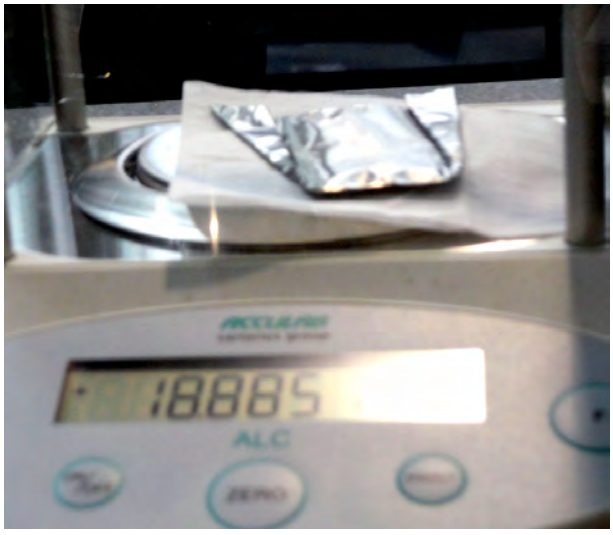


Figure 6.9: Remember to write the mass of the empty and filled sachet in the lab notebook.

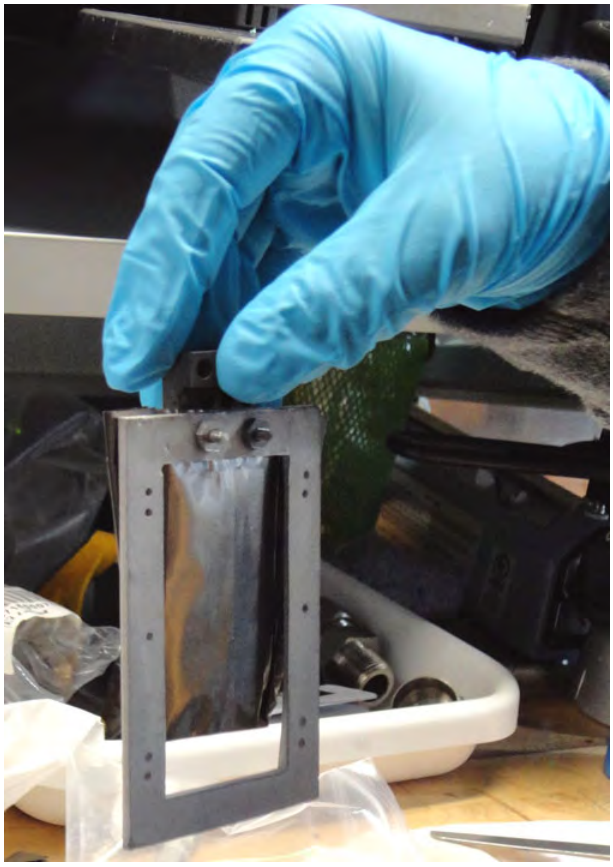


Figure 6.10: Completed sample assembly, ready to mount in the neutron beam.

Varley F. Sears: *Neutron Optics*, (Oxford University Press, New York and Oxford 1989).

G. Shirane, S. M. Shapiro, and J. M. Tranquada: *Neutron Scattering with a Triple-Axis Spectrometer*, (Cambridge University Press, Cambridge 2002).

G. L. Squires: *Introduction to the Theory of Thermal Neutron Scattering*, (Dover, Mineola NY 1978).

C. G. Windsor: *Pulsed Neutron Scattering*, (Taylor and Francis, London 1981).

Chapter 7

Essential Data Processing

This chapter explains procedures for the reduction of data from a time-of-flight inelastic neutron spectrometer. To date the practice of reducing inelastic scattering data from time-of-flight instruments has been poorly-documented, dull, tedious, and error-prone. We hope this chapter transcends these traditional shortcomings.

The present chapter explains the rationale behind the data reduction modules of *DANSE* (Chapter 7 better explains their use). After an overview of a typical data reduction process, the individual steps are explained in detail. Approximately, analysis steps that are further from the raw data are more closely-related to the dynamic processes in the sample, and tend to be more computationally-intensive. These steps (such as corrections for absorption, multiple scattering, and multiphonon scattering) are described later in this chapter. Although the data analysis steps described first are somewhat independent of the previous chapters, understanding some of the later steps (such as corrections for multiphonon scattering) may be easier after reading Chapters 2 and 3.

The treatment in this chapter presupposes a "direct geometry" configuration (monochromation before the sample) and a two-dimensional, pixellated detector system, composed of an array of linear, position-sensitive detectors. This configuration is used in the Pharos and ARCS spectrometers, for example.

The first task is to take a raw set of data from an inelastic neutron scattering experiment (with a powder sample), and convert it into physical units such as meV and Å^{-1} for energy and momentum transfer. Such basic steps are needed before the data can be compared to predictions from theory, or even compared to results from other inelastic spectrometers. Although the specific steps may depend in part on the scientific issues being studied, some data processing steps are common for nearly all experimental work, and are required to make an inelastic neutron scattering data set useful. There may be others steps as well, but most experiments follow most of the steps presented here.

Some notation is listed in Table 7.1. Fig. 7.1 shows the position of the detector tubes with respect to the sample, S, in the plane of the spectrometer,

Table 7.1: Notation

k_i (k_f)	modulus of the initial (final) neutron wavevector
$\vec{Q} = \vec{k}_\mathrm{i} - \vec{k}_\mathrm{f}$	momentum transferred from the neutron to the sample
ϕ	scattering angle ($Q = \sqrt{k_\mathrm{i}^2 + k_\mathrm{f}^2 - 2k_\mathrm{i}k_\mathrm{f}\cos\phi}$)
E_i (E_f)	incident (final) neutron energy
$\hbar\omega = \Delta E = E_\mathrm{i} - E_\mathrm{f}$	Energy transferred from neutron to sample
l_2	distance from the sample to the center of a detector
d	labels a specific detector
h	the height of a pixel in a detector

Fig. 7.2 the position perpendicular to the plane.

Typically one wants to know the dynamic structure factor, or scattering law, $S(Q,\omega)$. The coherent nuclear scattering of (4.93) can be summarized in simplified form:

$$\Big(\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\mathrm{d}E}\Big)_\mathrm{coh} = \frac{\sigma_\mathrm{coh}}{4\pi}\frac{k_\mathrm{f}}{k_\mathrm{i}}NS_\mathrm{coh}(\vec{Q},\omega)\ , \tag{7.1}$$

with similar expressions for incoherent scattering or magnetic scattering. Owing to several complications originating with how the measurement is made, the measured data are actually:

$$\begin{aligned}\Big(\frac{\mathrm{d}^2\sigma}{\mathrm{d}\Omega\mathrm{d}t}\Big)_\mathrm{coh} &= \Phi\Big[\frac{\sigma_\mathrm{coh}}{4\pi}\frac{k_\mathrm{f}}{k_\mathrm{i}}NS_\mathrm{coh}(d,h,t)\times T(d,h,t)\times\epsilon(d,t)\\ &+ S_\mathrm{nuis}(d,h,t) + b(d,h,t)\Big]\ .\end{aligned} \tag{7.2}$$

Here Φ is the incident flux, and often must be known by an independent measurement with a beam monitor, for example. The data are initially histogrammed as a function of neutron time-of-flight (TOF) rather than neutron energy,[1] where the two are related by $E = 1/2m_nv^2 = 5.227(l/t)^2$ (the second equality holds for speeds in mm/μs or equivalent). Background b and nuisance scattering from the sample, S_nuis, pollute the interesting signal.[2]

Other factors modify the interesting term involving S_coh. The sample may absorb neutrons, so only a fraction T of neutrons are transmitted through the sample, and this varies with scattering angle and neutron energy. The efficiency

[1] The introduction of new data acquisition hardware and software might alleviate this by histogramming directly into energy and momentum. That may simplify some tasks, for example eliminating the conversions from TOF to energy and from TOF and angle to momentum. However, this may complicate other tasks such as background subtraction.

[2] This list includes anything that does not interest the experimenter. For example, in a phonon measurement this may include all magnetic scattering, multiphonon and multiple scattering effects.

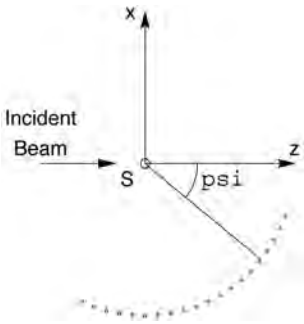


Figure 7.1: Layout of the secondary flight path of a direct geometry time-of-flight spectrometer. The detectors are arrayed in a circle of radius l_2 around the sample (at S). The beam is incident on the sample from the left. The x and z axes show the coordinate system used in the discussion of binning into rings (Sect. 7.1.2), the y-axis of that system points out of the page. ψ is the angle from the z-axis to the center of a given detector (in the x-z plane).

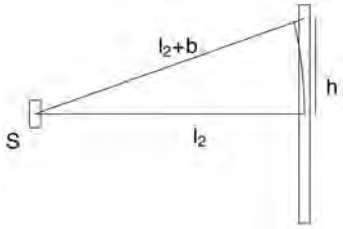


Figure 7.2: Labeling some of the distances in the secondary flight path of a direct geometry time-of-flight spectrometer. The center of the detector is l_2 from the sample (at S), and a given pixel is at height h from the center. The total distance from the sample to a pixel is $l_2 + b(h) = (l_2^2 + h^2)^{1/2}$.

ϵ of the detectors varies with energy and from one detector to another. The jobs ahead are:

- Account for incident flux.
- Remove background.
- Compute incident energy.
- Convert from time to energy.
- Correct for detector efficiency.
- Correct for absorption.
- Subtract similarly treated empty can data.
- Bin into rings of constant scattering angle ϕ.
- Subtract additional nuisance scattering (multiple scattering and so on).
- Convert from angle to momentum.

Errors must be propagated. This is usually done according to a few simple rules, such as when data sets are added or subtracted, the errors add in quadrature; the error for the raw histogram is the square root of the counts in any given bin **?**).

While many experiments could be treated by a similar process, not all of them can. In the later case, we have to know when departures from the routine require intervention in the reduction process. The sequence in which these tasks are presented here is that in which they are usually performed. As a rule, these operations do not commute, and we have to figure out what needs to be adjusted when re-ordering things. For instance, if the background is known as a function of time, either it should be subtracted before the data are transformed into energy, or the background must be transformed into energy as well. The sequence for other tasks, such as conversion to momentum units, is motivated by efficiency (one needs to think about energy when thinking about momentum, but not vice versa, so one might as well do the energy conversion first, followed by the momentum conversion).

7.1 Steps to Transforming Data into a Function of Energy and Momentum

7.1.1 Operations and Data Structures

The basic data structure is a histogram. This implies sets of arrays: the array to store the histogram values *per se*, a similar array for errors, and associated arrays to store the axis values for each dimension.

Operations can be divided into those that change the structure of some of the arrays, and those that change only the contents of the arrays. Those that change the structure include energy rebin (§ 7.1.2), sum into rings (§ 7.1.2), and momentum rebin (§ 7.1.2); those that don't change the array structures include subtracting various artifacts (§ 7.1.2, 7.1.2) and multiplicative corrections (§ 7.1.2, 7.1.2). The structure-changing operations require the creation of one or more new arrays; the rest can be done in place (assuming one is confident of the correction or if the correction can be easily undone). In the sections ahead, changes to the structure of the data will be noted.

7.1.2 A Closer Look at Each Task

Initial Data

The starting point is an array representing $I_0(d, h, t)$, the raw counts in each detector, at each detector position, for each TOF bin. The error is assumed to be the square root of the intensity: $\sigma_0 = \sqrt{I_0(d, h, t)}$. The TOF bins in this histogram usually have a constant time width, so all the bins are specified by some initial time t_0, the bin-width Δt, and the number N of bins.

Initially there are five major arrays representing:

- The actual data. Three dimensional array; if there are T time bins, D detectors, and H positions per tube, the array dimensions are $D \times H \times T$.
- The error. Same dimensions as the data.
- The time of flight. A one dimensional array of length T.
- The detector angles. A one dimensional array of length D.
- The pixel heights. A one dimensional array,[3] length H.

Those arrays are stored in a histogram.

Some additional parameters are needed. These include the incident neutron energy, E_i, and the geometry of the instrument, especially the length l_2 from the sample to each detector, and the angle of each detector. We assume these additional parameters are known.

Normalize by Incident Flux

Typically one divides by the total counts m in a peak in a beam monitor, which gives a proportional measure of the incident flux. Other measures may also be considered in unusual circumstances, for example, microamp-hours of proton beam delivered to the target. Regardless of the source, one still divides each element of the input array by some scalar:

$$I_1(d_j, h_k, t_i) = I_0(d_j, h_k, t_i)/m \ , \tag{7.3}$$

[3] One might allow for the possibility that the pixel heights vary from one tube to another. The pixel position array becomes a two dimensional array, with size $D \times H$.

for all $\{i, j, k\}$, and similarly for the error.

Subtract Background

Physically, the background is another complex subject; computationally, it is not. It can have different origins, from cosmic rays hitting the detectors (probably a small contribution) and neutrons from other experiments, to more serious causes, such as fast neutrons thermalizing in the instrument shielding, or slow neutrons scattering from poorly-masked beamline components. At times it can be difficult to determine what is background and what isn't, especially when looking for diffuse scattering.

A common way to estimate the background is to subtract a constant (in time) background determined from the data. This may have some justification for removing the background that originates with radiations from neighboring instruments, or neutrons from far enough away that any time structure is lost. Another approach subtracts the scattering measured with an empty can. This is better done at a later stage, and is described below.

Because the background could vary with time and detector position, we denote it as $b(d, h, t)$. Thus, we have

$$I_2(d_j, h_k, t_i) = I_1(d_j, h_k, t_i) - b(d_j, h_k, t_i) \; . \tag{7.4}$$

For each element of I_2 one must know about one element of I_1 and one element of b, and the same for the error.

Convert from Time to Energy

This step changes the structure of the data arrays. Prior to this step, the structures are the same as the initial arrays. There are still some more corrections to be made, but those corrections can be done more easily in terms of energy. This is computationally more intricate than the previous steps. In fact, it may be done in two steps.

First, one creates a new set of histogram bins whose boundaries are determined by t_2, the sample-to-detector TOF, and l_2, the distance from sample to detector, via the classical relation

$$\hbar\omega = E_\mathrm{i} - E_\mathrm{f} = 5.227\Big[v_\mathrm{i}^2 - \Big(\frac{l_2}{t_2}\Big)^2\Big] \; . \tag{7.5}$$

Note that pixels at different positions along the length of the detector tube are at different distances from the sample.[4] The effect is to coarsen the energy resolution. An extra 3 cm is a little more than the diameter of the detector tube, so this makes a contribution to the resolution similar to the size of the detector, or the sample size. For 35 meV neutrons and $l_2 = 4$ m the difference in energy is

[4] For a nominal l_2 of 4 m (Pharos), this discrepancy gets as large as about 0.03 m (generally, the extra path length is $b = l_2[(1 + h^2/l_2^2)^{1/2} - 1] \approx h^2/2l_2^2$).

about 1.5%. One way to account for this is to write a separate set of histogram energy bin boundaries for each height in the detector tube:

$$\hbar\omega = E_{\rm i} - E_{\rm f} = 5.227\Big[v_{\rm i}^2 - \Big(\frac{l_2^2 + h^2}{t_2^2}\Big)\Big]\,. \tag{7.6}$$

In making this transition, we need to multiply by the time bin-width and divide by the energy bin-width:[5]

$$I_3(d, \tilde{h}, \omega(h)) = I_2(d, h, t)\Big|\frac{dt}{d\omega}\Big|\,. \tag{7.7}$$

At this step, the array describing the time bins has been replaced by a two-dimensional array describing the energy bins. Here is a list of the arrays at this stage of data reduction:

- Data array. Still $D \times H \times T$.
- Error array. Still $D \times H \times T$.
- Intermediate energy bin boundaries array. Two dimensional, $H \times T$ or so.
- Detector angles array, one-dimensional, length H. Same as before.
- Pixel position array. Same as before.

The energy binwidth is changing from bin to bin, which makes it hard to think about in plotting, fitting, etc. Also, the energy of any given bin is a function of the position of the pixel height. This is why we usually take the second step, rebinning into constant energy bins. Rebinning will also remove the artifact of each detector position having a unique set of energy bin boundaries.

For the second step one creates another set of histogram bin boundaries, spaced uniformly in energy. One then assigns counts from the old bins to the new bins by prorating them. The rules are simple: if an old bin lies entirely within a new bin, one puts all of the old counts into the new bin; in doing so, multiply by the old bin width and divide by the new bin width. If an old bin overlaps two new bins, assign counts to the first new bin based on the fraction of old bin overlapped by the first, and assign counts to the second new bin according to the fraction of the old bin overlapped by the second bin (and multiply by the old bin width while dividing by the new). The same holds true if the old bin overlaps many new bins. One can of course think of this the other way around: each new bin asks "what does each old bin owe me".

One could write:

$$I_3(d, h, \omega) = \sum_i \mathbf{M}(\omega, \tilde{\omega}_i(h)) \times I_3(d, h, \tilde{\omega}_i(h)) \times \Big|\frac{d\tilde{\omega}_i}{d\omega}\Big|\,, \tag{7.8}$$

[5] This is because the histogram reflects counts per some unit (μs or meV); we have the constraint that the *integral* of the cross section be independent of the unit. If the histogram contains counts instead of counter-per-some-unit, then this factor is not necessary.

where $\mathbf{M}(\omega, \tilde{\omega}(h))$ is a matrix with the overlaps between old bins and the bins. One could get clever and drop the $|dt/d\tilde{\omega}|$ in Eq. 7.7 if one remembers to use $|dt/d\omega|$ instead of $|d\tilde{\omega}/d\omega|$ in Eq. 7.8. One typically needs to know something about several (but not all) elements of the old array before he can learn something about one element of the new array.

At the end of this step, three arrays have changed:[6]

- Data array. Now $D \times H \times N_E$.
- Error array. Now $D \times H \times N_E$.
- Energy bin values array. One dimensional, N_E.
- Detector angles array, one-dimensional, length H. Same as before.
- Pixel position array. Same as before.

Detector Efficiency

Detector efficiency varies with energy. These data are provided by the manufacturer. One simply divides by a pre-existing array of numbers:

$$I_4(d_j, h_k, \omega) = \frac{I_3(d_j, h_k, \omega)}{\epsilon(d_j, \omega)} \ . \tag{7.9}$$

The efficiencies are typically known as a function of final neutron energy, so an additional step is required to convert $\epsilon(E_f)$ into $\epsilon(\omega)$.

Absorption correction

The absorption by the sample varies with both neutron energy and mean path length through the sample. For a single energy, the transmission through a material with unit cell volume V and absorption cross section per unit cell σ_{abs}, the transmission probability is $T = \mathrm{e}^{-l\sigma_{abs}/V}$. For a wise choice of units for V and σ_{abs}, the path length l will be in cm. The mean path length through the sample varies with angle, so one needs to compute $T(\phi, \omega)$.

In a real experiment, the length through the sample depends on where in the sample the neutron scatters, to which direction it scatters, and the change in the neutron energy (recall that the absorption cross section per unit cell varies as inversely with neutron speed, so when the speed changes, the probability of absorption per unit length changes). It does not depend on momentum transfer. One must calculate this dependence; let us assume this has been done. Then one divides each element of the old array by one element the array representing $T(\phi, \omega)$

$$I_7(\phi_i, \omega_j) = I_6(\phi_i, \omega_j)/T(\phi_i, \omega_j) \ . \tag{7.10}$$

Section 7.3 of this chapter takes a more detailed look at the absorption correction.

[6] N_E is the number of new energy bins

Subtract empty can data

Having accounted for absorption by the sample, it now makes sense to subtract the scattering measured from the empty can, if the data from the empty can be brought through a similar treatment chain to this stage[7].

Bin into Rings

For a powder sample, the useful spatial information is the scattering angle ϕ. That is to say, for a given Q, E_i, and ω, a powder scatters into a cone of angle ϕ . So the two labels d and h can be condensed into the scattering angle ϕ.

The first step is to identify the ϕ for each pixel. A conventional instrument coordinate system has the sample at the origin and the transmitted beam forming the z-axis. The x-axis runs horizontally from the sample (for Pharos, that's toward FP-15) and perpendicular to the beam, and the y-axis, determined by right-handedness, points upward. (See Fig. 7.1.) It is natural to think of the detector as lying on a cylinder, with the axis of the cylinder coincident with the y-axis just described. So consider two vectors: $\vec{p}$, which runs from the sample to the pixel, and $\vec{a}$, with length l_2 running from the sample in the direction of the transmitted beam. In the instrument coordinate system, $\vec{a} = l_2\hat{z}$, and $\vec{p} = -l_2\sin(\psi)\hat{x} + h\hat{y} + l_2\cos(\psi)\hat{z}$. ψ is the angle between the tranmitted beam and the center of the detector; it lies in the x-z plane. The angle between $\vec{a}$ and $\vec{p}$ is given by $\vec{p}\cdot\vec{a} = |a||p|\cos\phi = a_z p_z$:

$$\begin{aligned}\phi &= \cos^{-1}\left(\frac{a_z p_z}{|a||p|}\right) = \cos^{-1}\left(\frac{p_z}{|p|}\right) = \cos^{-1}\left(\frac{l_2\cos(\psi)}{\sqrt{l_2^2 + h^2}}\right) \\ &= \cos^{-1}\left(\frac{\cos(\psi)}{\sqrt{1 + h^2/l_2^2}}\right). \end{aligned} \tag{7.11}$$

Having found the angle ϕ for each pixel, an efficient method to combine pixels with similar angles is desirable. The simplest method uses three nested `for` loops: one over the new angles, and two over the old detector indices. More abstractly:

$$I_5(\phi_k, \omega) = \sum_{i,j} \mathbf{N}(\phi_k, d_i, h_j) \times I_4(d_i, h_j, \omega)\ , \tag{7.12}$$

where $\mathbf{N}$ is a matrix whose elements are 1 or 0 depending on whether ϕ for the i^{th}, j^{th} pixel is within the bounds defined for the k^{th} element of the new array. It may not be necessary to know all of the points of the old array in order to find one element of the new array.

This operation depends on the instrument configuration (where the detectors are) and the set of angular rings which can be defined by users. After this

[7]Not an identical chain: hopefully your empty can didn't need an absorption correction. (If it was made out of vanadium, it might need that!)

step, all but one of the arrays have changed. Here is a list of the arrays after this step of data reduction:

- Data array: $N_\phi \times N_E$.
- Error array: $N_\phi \times N_E$.
- Energy bin values array: N_E.
- Phi-bin values array, N_ϕ

Subtract Nuisance Scattering

Other nuisance scattering includes all scattering from the sample that is either non-interesting or confusing. This usually includes higher-order scattering processes, and is generally difficult to compute (especially since one has now applied a number of multiplicative factors which must be accounted for). Computationally, this is simply a matter of subtracting two arrays, element-by-element, which brings us to:

$$I_6(\phi, \omega) = I_5(\phi, \omega) - S_{\text{nuis}}(Q, \omega) \tag{7.13}$$

Convert to Momentum Transfer

Ideally, there would be a prorating scheme like that used in going between time and energy. Another approach frequently used is interpolation. Essentially one recognizes that the regularly-spaced angles at which one has measured are an irregularly-spaced array of momentum transfers, Q. This momentum rebinning is typically performed with a convenient interpolation algorithm in, for example, IDL. Like the energy rebinning, one could write this process as a matrix operator acting on a vector:

$$S(Q, \omega_j) = \sum_i \mathbf{M}(Q, \phi_i, \omega_j) I_7(\phi_i, \omega_j) \ . \tag{7.14}$$

The arrays have changed again. After this step in data reduction, the arrays are:

- Data array: 2D, $N_E \times N_q$.
- Error array: 2D, $N_E \times N_q$.
- Energy bin values array: 1D, N_E.
- Q-bin values array: 1D, N_q.

At this point we have isolated the scattering of interest, and converted it into real physical units. This was considered a real achievement in the 1990's. For some experimental work the data analysis can be declared complete, and graphs of the results prepared for publication.

7.2 Transformations and Information

Information is lost in the course of data reduction – the word "reduction" itself implies a loss. Some loss of information is both inevitable and necessary with data from a direct geometry chopper spectrometer – with 10^5 detector pixels of 10^4 time bins each, a data set of 10^9 elements defies human comprehension. As described in the previous sections, the full experimental information on the arrival of each neutron at a specific detector pixel at a specific time is put through a number of transformations to improve comprehension, but also to improve counting statistics by summing events that are expected to be physically equivalent. It should also be noted that many data elements are usually uninteresting, or hold minimal counts, but even this information can be useful for removing extraneous scattering in the "good" data elements.

This section discusses the different physical scattering processes that can be measured experimentally, the possible transformations that can be performed on them, and whether these transformations are warranted, given the information required to do the transformation properly. For example, if the Q-resolution is lost by doing experiments on an incoherent scatterer like hydrogen, by performing measurements on a filter-difference spectrometer, or by rebinning all the Q-dependent data into energy, there is no path back to Q information by direct analysis of data alone. Fortunately, there may be a route back to the Q information through the use of a theoretical model or a simulation, and this is an opportunity provided by `DANSE` software.

Theory and computation play a bigger role when the experimental data require more supplementary information to enable the transformation. For data containing minimal information, it is possible to obtain detailed elementary excitations with theory alone – this is of course what has been done for many years by computational condensed-matter physicists. This dominance of theory over experiment was not the original intent of `DANSE`, but since `DANSE` provides such tools it will be interesting to see the path taken by future users of the `DANSE` system. Some community standards will have to be considered, but it is already clear that some transformations are better described as computational science rather than experimental science, even if experimental data is part of the effort. For example, consider data on elastic incoherent scattering from a polycrystalline sample. Very little information is contained in such data on excitations in solids, but these data could be augmented by ab-initio theory to make predictions of the inelastic coherent scattering from a monocrystal. In this example most of the work would be done by theory, and the experimental data would add little value. For this reason, the `DANSE` software does not support this transformation.[8] The present section lists all possible data transformations, and explains why approximately half of them are supported by `DANSE`.

[8] Such analysis is possible by the ab-initio tools in `DANSE`, but `DANSE` does not offer this as a transformation of experimental data.

Table 7.2: Scattering Processes

Energy	Momentum	Sample
inelastic	coherent	monocrystal
elastic	incoherent	polycrystal

7.2.1 Categorization of Transformations and Information

Table 7.2 lists the fundamental possibilities for a neutron scattering measurement. A real experiment may include combinations of coherent and inelastic scattering processes, but we consider these independently in what follows, because this is the approach used in data analysis. The categories of energy, momentum, and sample are orthogonal, and make the natural diagram of Fig. 7.3. Figure 7.3 is structured so that it the data of maximum detail is near the origin in the box labeled "inelastic coherent mono." (Note: with the word "inelastic," elastic scattering is also assumed included.) This inner cube contains enough information so that reprocessing allows one to move in any of the other three directions using experimental data alone (assuming the data are complete, of course).

As examples, consider the reprocessing of data from the inner box labeled "inelastic coherent mono":

1. Averaging the scattering over all crystal orientations (from $\vec{Q}$ to Q) allows reprocessing into the box above it, "inelastic coherent poly,"

2. Averaging over Q allows transformation to incoherent scattering, to the box in back, "inelastic incoherent mono,"

3. Discarding the inelastic contribution moves to the box at right "elastic coherent mono," which is useful for single-crystal diffraction patterns.

These three statements are discussed further below, as are other transformation paths. Some are less obvious than others – statement 2 above is less obvious than 1 and 3, and it has been discussed in the literature. Nevertheless, it should be almost intuitive that considerable information is lost in going to the remote cube, "elastic incoherent poly." Returning from such data to the origin is impossible.

7.2.2 Coherent – Incoherent

Forward

An energy spectrum can be obtained by integrating over the Q coordinate of $S(Q, E)$. With some care to account for thermal factors, multiple scattering, and multiphonon scattering, often before integrating over Q, this can be converted to a phonon energy spectrum or a density of states. This approach is rigorous

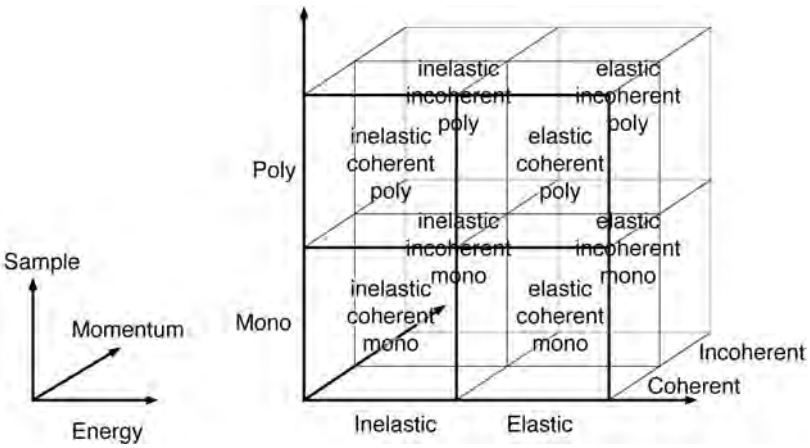


Figure 7.3: Eight possibilities for scattering processes in a neutron experiment. Each of the eight can be calculated independently in a simulation or modeling process. Some can be obtained from another by transformation. In general, information is lost as one moves away from the origin, as explained in the text.

if all values of Q are accounted for, even if the sample is a coherent scatterer with sharp dispersion information. An integration over all Q accounts for all phonons, at least in principle. Some of the issues concerning the incoherent averaging over coherent scattering have been discussed in the literature:

- V.S. Oskotskii, "Measurement of the Phonon Distribution Function in Polycrystalline Mateirals using Coherent Scattering of Slow Neutrons into a Solid Angle" Sov. Phys. Solid State **9**, 420 (1967).
- F. de Wette and A. Rahman, "Inelastic Scattering of Neutrons by Polycrystals" Phys. Rev. **176**, 784 (1968).
- M.M. Bredov, B.A. Kotov, N.M. Okuneva, V.S. Oskotskii and A. L. Shakh-Budagov, "Possibility of Measuring the Thermal Vibration Spectrum $g(\omega)$ using Coherent Inelastic Neutron Scattering from a Polycrystalline Sample," Sov. Phys. Solid State **9**, 214 (1967).

Nevertheless, the forward transformation from "inelastic coherent" to "inelastic incoherent" can be done reliably, especially for data from direct geometry chopper spectrometers that provide a wide range of Q. This transformation is offered by the `DANSE` software.

It should be mentioned that for many years, phonon DOS information has been obtained from phonon dispersion curves measured in special crystallographic directions on a triple-axis spectrometer. These dispersions are not good averages over Q, and energies are not measured for all phonons in the solid. Nevertheless, after fitting the dispersion curves to a lattice dynamics model, the energies of all other phonons can be calculated. This approach has never

been challenged for measurements on pure elements – it seems to work well because the high amount of information in the dispersion curves cause the interatomic force constants to be well determined. This method is not appropriate for disordered alloys, at least when the virtual crystal approximation is employed.

Reverse

In the same sense that it takes a lattice dynamics model to go from coherent inelastic scattering data from a triple-axis instrument to a good sampling of the incoherent inelastic scattering over all phonons, a model is always required for the reverse transformation. The positions of the Van Hove singularities in the phonon DOS are effective in constraining the dispersion curves, but there are only three of them in a phonon DOS of a pure element. It is not yet clear how reliably the reverse transformation can be performed, but it is certainly a useful capability that is offered in the `DANSE` software.

7.2.3 Monocrystal – Polycrystal

Forward

Transforming a complete data set from a monocrystalline sample to a polycrystalline average is simple in principle, and has the most physical interpretation of all the transformations in this section. In essence, the single crystal data need to be averaged over all crystallographic orientations by "rotating" the monocrystalline data to produce a polycrystalline average. This amounts to transforming $\vec{Q}$ to Q. The results can be similar to the transformation of coherent to incoherent scattering, and in fact proves identical when the crystals have isotropic scattering. `DANSE` supports these transformations.

Reverse

Except in special cases, the reverse transformation from polycrystal data to monocrystal data requires a model. The question of importance is how reliably can the polycrystalline average of, for example phonon dispersions, be used to define the single crystal phonon dispersions along specific crystallographic directions. The answer is not fully known today, but in the case of crystals such as tungsten, which are elastically isotropic, it is expected that the amount of information is essentially the same for data from polycrystalline and monocrystalline scattering experiments. In the case of anisotropic crystals it is expected that the dispersive information in $S(Q, E)$ will have intensity variations from anisotropic effects, and perhaps important information can be obtained from the intensity variations within the dispersion curves if the data are of good statistical quality. `DANSE` supports these transformations, using both phenomenological and ab-initio models. It is hoped that the neutron scattering

community will eagerly assess these capabilities of DANSE, because they free the experimenter from acquiring single crystals.

7.2.4 Inelastic – Elastic

Forward

A good diffraction pattern should be possible to acquire using white-beam mode on a direct-geometry chopper spectrometer, and the interpretation of the data would be the same as for a total scattering diffractometer. Here, however, we consider operation with a monochromatic beam as in the normal operation of a direct-geometry chopper spectrometer such as ARCS. The coherent elastic scattering is measured as part of the spectrum, so ignoring the scattering with energy transfers greater than the elastic peak width is essentially the same as performing a diffraction experiment. There are some defects in this approach – the spread of incident wavelengths broadens the diffraction peaks in Q. Nevertheless, useful information about the sample can often be obtained by simply ignoring the inelastic part of the data. Transformations to do this are provided by DANSE.

Reverse

Taking elastic data and transforming it to inelastic data is not impossible. Unfortunately, direct transformations are not robust. Robust approaches require substantial additional information. Phonon effects are evident in pair distribution functions, but care must be taken to separate these from atomic size effect diffuse scattering, which can be of the same order. There have been some efforts to obtain phonon dispersions from diffraction patterns, and the reader is welcome to explore these:

- D.A. Dimitrov, D. Louca, H. Roder, "Phonons from neutron powder diffraction," Phys. Rev. B **60**, 6204 (1999).
- W. Reichardt, L. Pintschovius, "Influence of phonons on the pair distribution function deduced from neutron powder diffraction," Phys. Rev. B **63**, 174302 (2001).
- A.L. Goodwin, M.G. Tucker, M.T. Dove, D.A. Keen, "Phonons from powder diffraction: A quantitative model-independent evaluation," Phys. Rev Lett. **93** 075502 (2004), *Ibid*. **95**, 119901 (2005).

At the inception of the DANSE project, the methods in these reference were placed outside the scope of the effort.

Nevertheless, DANSE does offer methods for reverse transformation. These are robust in that they give stable answers, but they require substantial computational effort. The approach is to obtain the lattice dynamics for the structure. From the lattice dynamics the inelastic scattering can be obtained. If the crystal

structure or molecular structure is obtained from a diffraction measurement, for example, phonon dynamics can be calculated in one of two ways.

- A model of the lattice dynamics for the structure can be used to calculate the phonons. This requires input on interatomic force constants. They may be guessed, or obtained from other sources. From the dynamics, the inelastic scattering can then be calculated.
- Using the unit cell obtained from a diffraction pattern, or a molecular structure obtained by other means, an ab-initio electronic structure calculation can be performed to optimize the positions of the atoms in the structure. These may not be the real atom positions, but relaxed positions are needed for the following computational step. Next, the atom positions are displaced by a controlling software package such as PHON, and the ab-initio code is run again for the distorted unit cell. This permits calculation of tensorial force constants that are used in the lattice dynamics. From the dynamics, the inelastic scattering can be calculated.

Without such theoretical input, however, attempting a reverse transformation from elastic to inelastic would not be robust, and with the `DANSE` software it is an error to do so. Note that the transformation of structural information into inelastic inelastic scattering information requires perhaps the most sophisticated tools offered by `DANSE`. An increasing loss of experimental information requires an compensating sophistication in the theoretical tools to reverse the situation.

7.2.5 All Specific Cases

The previous subsections discussed general issues when transforming data forward or backwards between elastic and inelastic, coherent and incoherent, and monocrystal and polycrystal. For each of these three cases, and six general transformations, there are four specific cases. For example, Fig. 7.3 shows that between the monocrystal and polycrystal samples (up in the graph), there are a total of eight forward and reverse transformations:

- inelastic coherent (mono$\leftrightarrows$poly),
- inelastic incoherent (mono$\leftrightarrows$poly),
- elastic coherent (mono$\leftrightarrows$poly),
- elastic incoherent (mono$\leftrightarrows$poly).

Similar forward and reverse transformations between adjacent cubes are also possible along the elastic$\leftrightarrows$inelastic and coherent$\leftrightarrows$incoherent directions of Fig. 7.3, giving a total of 24 such transformations between adjacent cubes in the figure (along the $\langle 100 \rangle$ directions). In addition, a total of 12 transformations are possible between scattering data in cubes in Fig. 7.3 that are not immediately adjacent (along the $\langle 110 \rangle$ directions), and a total of 8 possibilities for transformations between cubes located diagonally in the figure (along $\langle 111 \rangle$

directions). Tables **??** – **??** list all 44 possible transformations, and identify the specific transformations supported by the DANSE software. As of early 2007, the plan for DANSE supports 24 of these possible transformations.

Sections 7.2.2 - 7.2.4 provided general reasons for the why transformations along the main axes of Fig. 7.3 are not provided by DANSE. Eight transformations are possible along each of these axes, however. Nearly all cases in Table **??** can be understood with the rules of Sects. 7.2.2 - 7.2.4, with the following exceptions:

- Transformations from elastic coherent polycrystal data (powder diffraction patterns) to elastic coherent monocrystal data (single crystal diffraction data) are not supported in the rebinning operation in the first release of DANSE. These capabilities are possible with Rietveld refinement methods (or PDF modeling), which are available as methods outside the scattering kernel.

- Elastic incoherent scattering pertains to the isotopic incoherence of neutron scattering, for example, and not to incoherence from atomic disorder (which is treated in methods for elastic coherent scattering). DANSE therefore does not support well transformations involving elastic incoherent scattering from either monocrystals or polycrystals.

The more complex transformations of two or three data characteristics, listed in Tables **??** and **??**, are generally consistent with the previous rules and individual transformations. More transformations are possible by chaining individual transformations from Table **??** than are listed in these Tables **??** and **??**. These latter tables list algorithms that do both transformations in an integrated way. The emphasis is on data from inelastic scattering, since this field was first to exploit this method of direct experiment simulation.

7.3 Absorption

With an absorbing sample, one observes $\mathrm{d}^2\sigma/\mathrm{d}\Omega\mathrm{d}\omega \times T(\Omega, \omega)$ rather than just $\mathrm{d}^2\sigma/\mathrm{d}\Omega\mathrm{d}\omega$, where $T(\Omega, \omega)$ is the probability of a neutron being transmitted through the sample in a given direction (described by the solid angle Ω) and with a given final energy (final energy = incident energy − energy transfer, or $E_f = E_i - \hbar\omega$). All that needs to be done is to calculate the transmission for each final energy and for each detector, and then divide the observed scattering by the transmission probability. For an experiment in which all the observed neutrons scatter once, the calculation is straightforward, though somewhat computationally demanding.

We suppose that the sample is divided into a number of cells of equal volume; then a neutron observed in any given detector with a given final energy was scattered in any cell with equal probability. We compute the transmission probability for each scattering cell, then average over probabilities for all cells. The reason we have to be so careful is that the transmission probability depends on the distance the neutron travelled through the sample, and this depends on

where in the sample the scattering occurred, and the outgoing direction of the neutron[9].

For a neutron travelling across an absorbing medium, the probability of transmission is [10]

$$T = \frac{N}{N_0} = \exp(-l\sigma_{abs}\rho). \tag{7.15}$$

Here, l is the length of the path through the sample, σ_a is the absorption cross section, and ρ is the number density of absorbers. If the neutron changes direction (scatters) while travelling through the sample, the probability becomes

$$T = \exp(-l_{inc}\sigma_{abs}\rho - l_f\sigma_{abs}\rho) \tag{7.16}$$

where l_{inc} is the path length from where the neutron enters the sample to the scattering spot, and l_f is the path length from the scattering spot to where the neutron exits the sample. There's one more difficulty: the absorption cross section depends on the neutron energy, so if the scattering is inelastic, there will be two absorption cross sections, $\sigma(\omega_{inc})$ and $\sigma(\omega_f)$, and we have the transmission probability as

$$T = \exp(-l_{inc}\sigma_{abs}(\omega_{inc})\rho - l_f\sigma_{abs}(\omega_f)\rho). \tag{7.17}$$

In what follows, we assume that the scattering is purely horizontal. This is an approximation that greatly reduces complexity of the computation. This is justifiable for several existing instruments including LRMECS and PHAROS. In fact, any horizontal scattering lies within about ±5° of the horizontal on an instrument like Pharos,[11] and is therefore negligible.[12] This means each scattering cell needs two labels.

In considering a number of cells, the incoming path length depends on where the cell is located. The outgoing path length depends on where the cell is, and in what direction the neutron is heading. This direction can be determined if we know the detector d that the neutron is hitting. So we need four labels for the transmission from a given scattering cell to a given detector with a given final energy:

$$T(d, \omega_f, i, j) = \exp(-l_{inc}(i,j)\sigma_{abs}(\omega_{inc})\rho - l_f(i,j)\sigma_{abs}(\omega_f)\rho). \tag{7.18}$$

In the last step, we're going to average over all N cells for each detector and energy:

$$T(d, \omega_f, i, j) = \frac{1}{N}\sum_{i,j} \exp\Big(- l_{inc}(i,j)\sigma_{abs}(\omega_{inc})\rho - l_f(i,j)\sigma_{abs}(\omega_f)\rho\Big). \tag{7.19}$$

[9] We assume the incoming neutrons are perfectly monochromatic and collimated.

[10] Think $\mathrm{d}N/\mathrm{d}x = -\sigma\rho$...

[11] the detectors are at most 0.5 m above the scattering plane, while the middle of the detector is 4 m from the sample, so the maximum angle is about ±7°. However, the top and bottom 10 cm or so of the tubes are thrown out...

[12] This assumption has not been sufficiently examined. Unless the effect is very small ($\ll 1\%$), it should be included in the calculation when that becomes computationally affordable. On Pharos, the extra path length is about 0.5%. Is this important? For instruments with larger vertical divergence (ARCS), this will be necessary anyway, so might as well get used to it.

It does not get much easier than that. Now all we need is a program that computes Eq. 7.19 for a number of interesting geometries.

7.4 Neutron Weighting

The phonon DOS obtained by the data reduction methods of this chapter is not the true phonon DOS, except in the case of pure elements. It is instead the "neutron-weighted DOS." The neutron-weighted DOS accounts for the different efficiencies of phonon scattering from different chemical elements or isotopes. (If the sample is chemically homogeneous, this is usually not a problem.) The neutron-weighted phonon DOS is rigorously defined (Sect. 2.3.3)as:

$$g_{\mathrm{NW}}(E) \propto \sum_d g_d(E) \exp(-2W_d) \exp(2W) \frac{\sigma_d}{m_d} \tag{7.20}$$

where $\exp(-2W_d)$, σ_d and m_d are the Debye-Waller factor, total scattering cross-section and mass of atom d. The Debye-Waller factor is an explicit function of $g_d(E)$. The term $\exp(2W)$ is the average Debye-Waller correction; this is calculated from the self-consistent neutron-weighted DOS. The factor $\exp[2(W - W_d)]$ is approximately unity. For the case where σ_d/m_d is the same for all species d, $g_{\mathrm{NW}}(E) \approx g(E)$.

Obtaining the true phonon DOS from the neutron-weighted phonon DOS requires a full analysis of the lattice dynamics. This can be performed by simulational procedures described in a later chapter. The neutron-weight correction as well as other approximations involved in the data analysis can be overcome by fitting a dynamics model to the neutron-scattering data directly. Although this approach is both scientifically and computationally demanding, we foresee no better method for extracting the vibrational dynamics from inelastic neutron scattering measurements.

7.5 Calculation of Multiphonon Scattering

The multiphonon expansion was developed with some rigor in Section 4.2.2. Here we explain how it works in practice. The total measured spectrum, $S(E)$, is the sum of components, $\sum_{n=0}^{\infty} S_n(E)$, from neutrons scattered after creating different numbers, n, of phonons in the sample.[13] Only the 1-phonon scattering is useful for obtaining a phonon DOS, so it is important to have an understanding of the higher-order terms for planning an experiment, or if one seeks quantitative corrections of experimental data. Performing a multiphonon expansion is done in two steps.

- First, the weights of the different n-components are calculated independently with input information on Q and atom displacements u at the temperature T.

[13] Phonon annihilation is handled by extending the range of E to negative numbers for each $S_n(E)$.

- Second, the spectral shape of each component $S_n(E)$ is obtained by sequentially convoluting the 1-phonon profile with itself a total of $n-1$ times.

Finally, for a chopper spectrometer, detectors at different scattering angles ϕ provide energy spectra $S_\phi(E)$ where Q varies with E across the spectrum. The multiphonon corrections must take into account the kinematical relation of Q vs. E at each scattering angle ϕ if data are analyzed from each detector bank. The constant-Q multiphonon analysis is simpler to understand, and is applicable for data from triple-axis spectrometers, or for data from chopper instruments that have been reduced to $S(Q,E)$.

Intensities of n-Phonon Spectral Components

Start with the Debye–Waller factor, which attenuates the elastic scattering S_0, to a fraction of the total scattering S:

$$S_0 = S\exp(-2W)\,, \tag{7.21}$$

which we rearrange and expand:

$$S = S_0\exp(+2W)\,, \tag{7.22}$$

$$S = S_0\sum_{n=0}^{\infty}\frac{(2W)^n}{n!}\,. \tag{7.23}$$

We now have a series of terms in an expansion, but the next step of substituting (7.21) into (7.23) amounts to nothing more than writing $\exp(-2W)\exp(+2W) = 1$:

$$S = S\exp(-2W)\sum_{n=0}^{\infty}\frac{(2W)^n}{n!}\,, \tag{7.24}$$

$$1 = \sum_{n=0}^{\infty}\frac{(2W)^n}{n!}\exp(-2W)\,. \tag{7.25}$$

The terms in (7.25) are associated as 0 for elastic scattering, 1 for 1-phonon scattering, 2 for 2-phonon scattering, etc. What is special about (7.25) is that the n^{th} term in the series is exactly the fraction of the n-phonon spectral component in the total scattering (elastic plus inelastic).

To calculate the fraction of any multiple phonon term (e.g., the amount of 2-phonon scattering), all we need is $2W$. Recall the physical origin of $2W$:

$$2W = \langle Q^2u^2\rangle\,, \tag{7.26}$$

where $\langle u^2\rangle$ is the mean-squared atom displacement. Equation (7.26) shows that $\exp(-2W)$ becomes much less than 1 when the atom displacement becomes comparable to $1/Q \sim \lambda/2\pi$, the number of wavelengths associated with the

scattering angle. This is consistent with $2W$ originating from the destructive interference of scattered wavelets. The Q^2 and u^2 are related to the energy of recoil, and the thermal energy, respectively. The recoil energy, E_R is

$$E_R = \frac{\hbar^2 Q^2}{2M} \ , \tag{7.27}$$

where M is the mass of the atom that is scattering, in units of the neutron mass. (The units of M are very similar to the atomic weight.) The thermal energy is:

$$k_B T = \frac{1}{2} M\omega^2 \langle u^2 \rangle \tag{7.28}$$

for one mode of frequency ω, and is a quantity that depends on the phonon DOS for a real material. Sadly, an exact evaluation of (7.28) is not simple, in part because the phonon states are not fully occupied at modest temperatures, but also because of the commutation relationships discussed in Section 4.1. The result is summarized as γ_0, where:

$$\gamma_0 = \int_0^\infty \coth(E/2k_B T) \frac{g(E)}{E} dE \ . \tag{7.29}$$

In the limit of high temperatures, where $\coth(E/2k_B T) \rightarrow 1$, it is clear from (7.29) that γ_0 increases as $1/E$ (because the amplitudes of atom motions become larger in low-energy modes as predicted by (7.28)). The low-energy modes are even more important at low temperatures, where they have a larger phonon occupancy. The Debye–Waller factor of (7.26) is, rigorously:

$$2W = \gamma_0 E_R \ , \tag{7.30}$$

In performing a multiphonon expansion for one T and one phonon DOS $g(E)$, the value of γ_0 is computed once with (7.29). The $2W$ from (7.30) for appropriate Q is used in (7.38) to get the fractions of all n-phonon scatterings. We therefore rewrite (7.25) as

$$S(Q) = \sum_{n=0}^{\infty} S_n(Q) \ , \quad \text{where :} \tag{7.31}$$

$$S_n(Q) = \frac{\Big(2W(Q)\Big)^n}{n!} \exp\Big(-2W(Q)\Big) \ . \tag{7.32}$$

Shapes of n-Phonon Spectral Components

It remains to get the spectral shape of each order of the multiphonon scattering. The energy spectrum for 1-phonon scattering was discussed in Sections 2.3 and 4.2.2. The spectrum for one-phonon scattering weights more heavily the low-energy modes because they have larger amplitudes of motion, providing a factor of $g(E)/E$. The number of phonons is the Planck distribution

$1/[\exp(E/k_B T) - 1]$, so the two factors provide the shape of the 1-phonon profile, $A_1(E)$:

$$A_1(E) = \frac{g(E)}{E\gamma_0} \frac{1}{e^{E/k_B T} - 1} . \tag{7.33}$$

Each phonon created has the profile $A_1(E)$ of (7.33). When two phonons are created simultaneously, the total spectrum of energies is the convolution of the 1-phonon profile with the 1-phonon profile.[14] The 2-phonon spectrum is:

$$A_2(E) = A_1 * A_1 = \int_{-\infty}^{\infty} A_1(E - E')\, A_1(E')\, dE' , \tag{7.34}$$

and the n-phonon profile is the convolution of another 1-phonon profile with the $(n-1)$-phonon profile:

$$A_n(E) = A_1 * A_{n-1} = \int_{-\infty}^{\infty} A_1(E - E')\, A_{n-1}(E')\, dE' . \tag{7.35}$$

Starting with A_1, we can generate the spectral shapes of all the orders of multiphonon scattering by the systematic application of (7.35). The total scattering is the sum of these spectral profiles, weighted by the corresponding terms of (7.31):

$$S(Q,E) = \sum_{n=0}^{\infty} S_n(Q)\, A_n(E) , \tag{7.36}$$

$$S(Q,E) = \sum_{n=0}^{\infty} \frac{(2W)^n}{n!} \exp(-2W)\, A_n(E) . \tag{7.37}$$

If we define each term in the sum as $S_n(E)$:

$$S(E) = \sum_{n=0}^{\infty} S_n(E) . \tag{7.38}$$

Examples of Multiphonon Scattering from a Chopper Spectrometer

The analysis presented here is for inelastic energy spectra obtained from a group of detectors centered about the scattering angle, ϕ. The energy transfer, E, is obtained simply and reliably from the arrival times of the neutrons at the detector. On the other hand, for each detector bank at ϕ, the value of Q varies with E and the energy of the incident neutron E_{inc} as:

$$Q[\text{Å}^{-1}] = 0.6947 \sqrt{2E_{inc} - E - 2\sqrt{E_{inc}(E_{inc} - E)}\cos\phi} , \tag{7.39}$$

[14] Consider each phonon excitation to be a random variable with a probability distribution of A_1. The sum of two random variables has a distribution that is the convolution of the probability distributions $A_1 * A_1$.

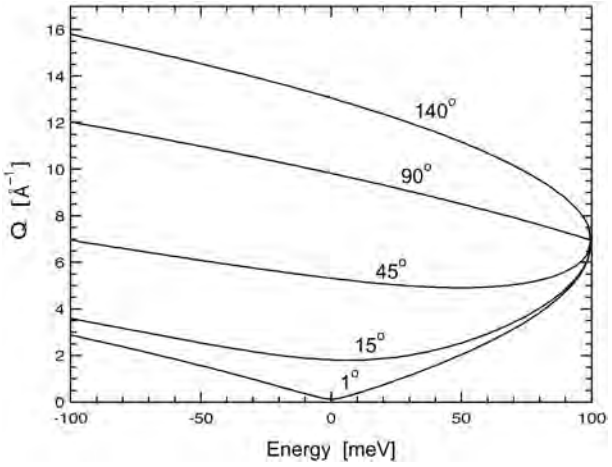


Figure 7.4: Graphs of Q vs. E for a time-of-flight chopper spectrometer, calculated using (7.39) for five values of scattering angle ϕ (labels on curves). Incident energy was $E_{\text{inc}} = 100$ meV. Positive E correspond to phonon creation, negative to phnon annihilation.

where E and E_{inc} have units [meV]. Curves showing relationships $Q(E)$ are presented in Figure 7.4. The curves cannot extend to the right beyond E_{inc} because the incident neutron cannot lose a greater amount of energy to the sample. Notice that the curves are generally asymmetrical in $\pm E$.

For calculating the multiphonon scattering at a particular scattering angle ϕ, it is necessary to evaluate the $2W$ and the Debye–Waller factor at each energy of the spectrum, because Q varies with E as in (7.39) and as shown in Figure 7.4. Examples of such calculations are shown in Figure 7.5. Figure 7.5a shows the phonon DOS curve from fcc nickel metal from which the subsequent curves were calculated. Figure 7.5b shows the total inelastic scattering calculated for a temperature of 500 K.[15] The 1-phonon scattering is confined to the range ±37 meV, since this is the maximum phonon energy. With the excitation of two phonons, the energy range can be extended to $2 \times \pm 37 = \pm 54$ meV, and this is the energy range of the 2-phonon scattering. In Figure 7.5b, the intensity at 40–50 meV is almost entirely 2-phonon scattering.

Notice the general trend in Figure 7.5b showing that the inelastic scattering to increases with the scattering angle, owing to the factor of $2W^n$ in (7.32). On the other hand, the scattering is suppressed by the Debye–Waller factor, $\exp(-2W)$. At high Q or high T the Debye–Waller factor becomes increasingly important, and effects of this are seen in Figure 7.5c. First notice that although there are about twice as many phonons at 1000 K as at 500 K, the inelastic

[15] An experimental spectrum would contain a large peak at $E = 0$ from the elastic scattering. In principle, the area of this elastic peak would be the factor $\exp(-2W)$ of the total integrated scattering (elastic plus inelastic).

scattering of Figure 7.5c is nowhere near twice as large as in Figure 7.5b. Also notice the change in symmetry of the spectra in Figure 7.5c with detector angle. For a detector angle of 60°, Figure 7.5c shows an inelastic scattering that is approximately symmetrical in $\pm E$. This is consistent with the gradual variation of Q with E in the kinematics as shown in Figure 7.4. On the other hand, for the larger detector angles of 90° and 140°, the value of Q is larger for $-E$ than $+E$. This asymmetry causes the Debye–Waller factor to attenuate the scattering at $-E$ more severely than at $+E$. When these effects are significant, it is obviously naive to interpret the asymmetry in E of the inelastic scattering in terms of detailed balance.

7.5.1 Multiphonon Correction – Iterative

The present section describes a correction procedure that can be used to remove higher-order multiphonon scattering from an experimental inelastic scattering spectrum, so the spectrum can the be used to deduce a 1-phonon profile and a phonon DOS. It is based on the incoherent approximation, but this procedure has been used for many years with coherent multiphonon scattering data, and it is often expected to be reliable in such cases, at least for polycrystalline samples. (The idea is that averaging over crystalline orientations provides a broad sampling of reciprocal space, performing a good average over phonon modes. The incoherent approximation is adequate when all modes are sampled this way. It is unlikely that this approach could be reliable for the analysis of coherent scattering from single crystals, of course.)

An example of multiphonon calculations and an extraction of 1-phonon scattering is shown in Figure 7.6. Cerium is a coherent scatterer, and the data were acquired at several values of Q to average over the different phonon modes. An inelastic spectrum for one value of $Q = 3.924$ Å^{-1} is presented in the figure.[16] Steps of an iterative procedure to refine the phonon DOS are shown in Figure 7.6. The first step of the iteration used a very crude approximation for the multiphonon scattering plus random background – a simple constant function of approximately 70 counts. After this constant was subtracted from the experimental data, the DOS curve, $g(E)$ labeled "1" in Figure 7.6b was obtained from the 1-phonon profile from (7.33) with a Debye–Waller factor of (7.32). The 2–5 phonon scattering was calculated from this first iteration of a phonon DOS, scaled in height (plus a constant background was also fit to the data), and a second iteration of the DOS was generated by the same procedure. Notice that the second and third iterations of the phonon DOS in Figure 7.6b are rather similar. Even for this difficult case where the 2–5 phonon scattering is a large fraction of the total inelastic scattering, convergence is fairly quick because the 2–5 phonon scattering does not have much structure.

The multiphonon expansion assumes that the scattering is incoherent. Unfortunately, a better approximation would require a detailed simulation of the

[16] A spectrum at one value of Q can provide a piece of the dynamical structure factor intensity, which we loosely call a "phonon DOS," since the result would be a phonon DOS if the scattering were incoherent. Data were acquired from the triple-axis spectrometer HB3 at HFIR.

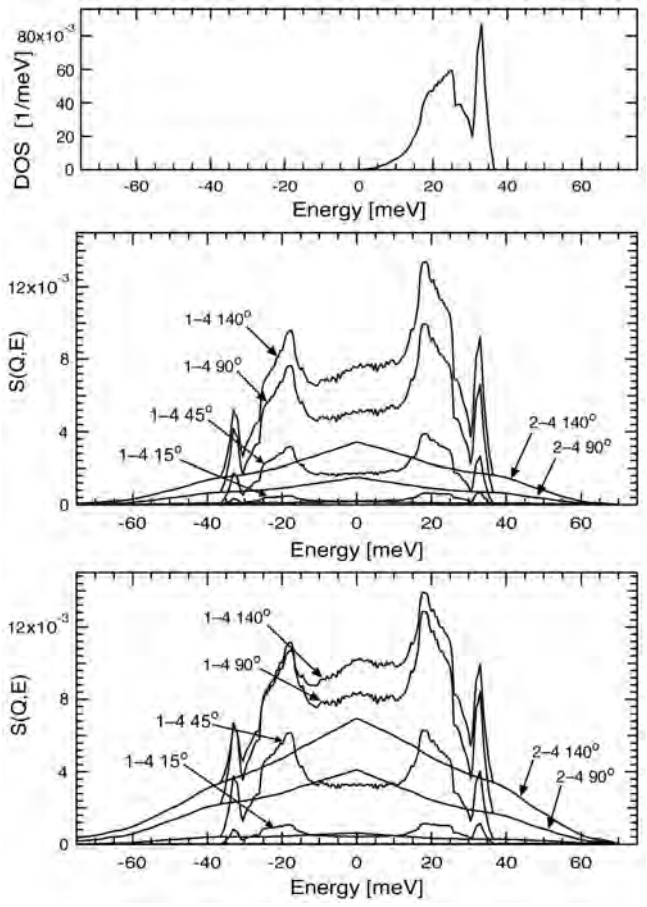


Figure 7.5: (**a**) Phonon DOS of fcc Ni metal. Inelastic scattering from four angle banks of a time-of-flight chopper spectrometer with incident neutron energy of 70 meV are shown in: (**b**) for 500 K and (**c**) for 1000 K. In parts b and c, the four curves with detailed features are the total scattering, summing the 1-phonon through 4-phonon terms. The broad curves, which are the full intensity above 37 meV, are the mulitphonon scattering from 2-phonon through 4-phonon contributions.

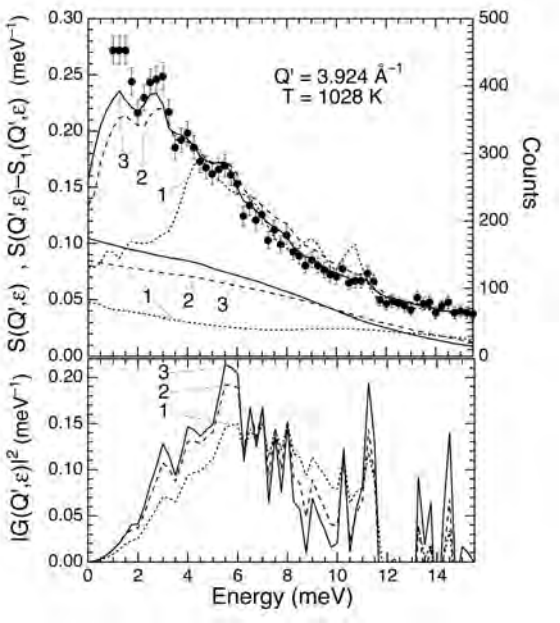


Figure 7.6: (**a**) points: experimental inelastic scattering from cerium metal at 1028 K and $Q = 3.924\,\text{Å}^{-1}$. The elastic peak at low energy extends an order-of-magnitude beyond the top of the graph. Calculations of the 1–5 phonon scattering plus a constant background were scaled to fit these data, and are labeled 1, 2, 3, denoting the order of iteration. The 2–5 phonon scattering is shown with the same scaling, again with labels 1, 2, 3 for the iteration. By subtacting the 2–5 phonon scattering from the experimental data, the difference was converted into a "phonon DOS", $g(E)$, by correcting for the thermal factor of (7.33) and Debye–Waller factor of (7.32). (**b**) The "phonon DOS" obtained from the experimental data of part a at different iterations. The first iteration, labeled 1, was obtained by subtraction of a constant background from the experimental data, and correcting by the thermal factor of (7.33). This "DOS" was used for generating the multiphonon corrections for the next iteration of the procedure.

lattice dynamics to account for the $\vec{Q}$-dependence. To our knowledge, such a calculation has not been done as of 2004. Another case where the assumptions of this procedure are stretched is the application of the incoherent multiphonon correction procedure to inelastic spectra from alloys. One concern about alloys is that the Debye–Waller factors differ for the different atoms in the alloy, causing some phonons to be weighted more than others. Although the multiphonon correction procedure has been used without adapting to the different $\langle U^2 \rangle$ of the atoms in the alloy, comparisons with other measurements have shown that the procedure is often acceptable, perhaps because the multiphonon scattering is essentially featureless.

7.5.2 Multiphonon Correction – Fourier Log

Another way to correct for multiphonon scattering is possible for energy spectra at constant Q. Following the notation of Section 7.5, for fixed values of Q and temperature, the energy dependence of the various phonon contribu-tions, $S_n(Q, E)$, is

$$S(Q, E) = \frac{(2W)^n}{n!} \mathrm{e}^{-2W} A_n(E) \ , \tag{7.40}$$

The measured scattering intensity, $S(Q, E)$, is the sum of intensities from all phonon processes:

$$S(Q, E) = \sum_n S_n(Q, E) \ , \tag{7.41}$$

Our problem is to use the measured $S(Q, E)$ to isolate the single scattering profile, $A_1(E)$.

An approach to extracting the phonon partial DOS from experimental data is called the "Fourier-log deconvolution method." It was used for analysis of plasmon scattering measurements in electron energy-loss spectrometry [Spence]. This method is also used for incoherent inelastic phonon scattering [M. Y. Hu, W. Sturhahn, et al.]. It has the additional feature of being able to correct for the broadening of the elastic line, but only if the energy resolution is constant across the energy spectrum (not the case in TOF spectrometers). The zero-loss peak, $Z(E)$, is convoluted with the scattering from the specimen as:

$$S(Q, E) = Z(E) * \left(\mathrm{e}^{-2W} \delta(E) + \sum_n S_n(Q, E) \right) \ , \tag{7.42}$$

where the first term in (7.42) is the zero-loss peak, reduced by the Debye–Waller factor. Taking the Fourier transform, $F[]$, of (7.42) simplifies the convolutions of (7.35), which become multiplications in Fourier space:

$$F[S(Q, E)] = F[Z(E)] \mathrm{e}^{-2W} \left(1 + \sum_n \frac{(2W)^n}{n!} \left(F[A_1(E)] \right)^n \right) \ . \tag{7.43}$$

The term in parentheses () on the right side of (7.43) is recognized as the expansion of an exponential function:

$$F[S(Q,E)] = F[Z(E)]\mathrm{e}^{-2W}\left(\mathrm{e}^{2WF[A_1(E)]}\right) . \tag{7.44}$$

Taking the logarithm of (7.44) and rearranging:

$$F[A_1(E)] = \frac{1}{2W}\ln\left(\frac{F[S(E)]}{F[Z(E)]}\right) + 2W . \tag{7.45}$$

The inverse Fourier transformation, $F^{-1}[]$, provides the single scattering profile, $A_1(E)$:

$$A_1(E) = \frac{1}{2W}F^{-1}\left[\ln\left(\frac{F[S(E)]}{F[Z(E)]}\right)\right] + 2W\delta(E) . \tag{7.46}$$

If we do not care about the normalization of the single scattering profile, and if we delete the zero-loss peak $Z(E)$ from the data (it is suppressed by the thermal correction anyway), we obtain:

$$A_1(E) = \frac{1}{2W}F^{-1}\left[\ln\left(\frac{F[S(E)]}{F[Z(E)]}\right)\right] . \tag{7.47}$$

7.6 Corrections for Simultaneous Multiple Scattering and Multiphonon Scattering

7.6.1 Best Practices

Recursive computation

Multiple-scattering is naturally computed with a recursive algorithm, as implemented in MCViNE **?**) or MCSCAT **?**). For the purpose of this discussion, we differentiate between two types of multiple scattering: single-scatterer multiple scattering (SSMS) to describe the multiple scattering within one neutron scatterer and multiple-scatterer multiple scattering (MSMS) to describe multiple scattering among different neutron scatterers, such as between the sample and its holder.[17] A neutron may scatter by different processes, and it may undergo these processes in different sequences and in different locations. The set of scattering processes (implemented as "scattering kernels") used in MCViNE in 2014 is

- Incoherent elastic scattering
- Coherent elastic scattering from powder sample

[17] Sample components in some linear-chain-based MC ray tracing packages can support SSMS, but they do not have abstractions similar to composite scatterer or scattering kernel. As a result, the SSMS algorithm must be duplicated in these sample components.

- Incoherent inelastic single-phonon scattering
- Coherent inelastic single-phonon scattering from a powder sample
- Multi-phonon scattering
- Scattering from a dispersion surface where the dispersion relation and the dynamical structure factor are described by analytical functions of momentum transfer vector Q.

Figure 2 shows an example of SSMS in which a neutron is scattered three times before exiting a scatterer. When a neutron is scattered inside a homogeneous scatterer, the original incident neutron is split into two for computational efficiency. One neutron is propagated through the scatterer with its probability lowered by attenuation, while the other is scattered by one of the scattering kernels, chosen by random selection, at a point also randomly selected along the forward path of the incident neutron. This splitting process repeats for several scatterings, as illustrated in Fig. 2, until either the neutron probability is lower than a pre-selected limit, or the maximum order of multiple scattering is reached. At this point, the neutron that is still inside the scatterer is allowed to propagate out, with its probability attenuated appropriately.

Figure 3 shows an example of multiple scattering in a sample assembly with a simple "concentric" arrangement of the sample and the sample environment. Only five of an infinite number of possible multiple scattering paths are illustrated, and the splitting processes are not shown.

Multiple scattering from an aluminum sample and environment

In this example, we present experimental and simulated inelastic spectra from polycrystalline Al metal. Most interesting, perhaps, are the simulations that include subsets of the scattering processes. These can show how to optimize the experiment, or how to do the most important corrections for unwanted intensities. The experiment and simulations used a 60 mm×60 mm×4 mm polycrystalline aluminum (1100 alloy) plate in the ARCS spectrometer. A quick calculation using the total scattering cross section for Al shows that the sample scatters 5% of the incident beam. In the simulation and experiment, the sample was placed at approximately 135° to the beam direction, and the incident energy was tuned to 80.5 meV using a Fermi chopper slit package at 480 Hz. The simulated sample assembly contains only one homogeneous scatterer for the aluminum plate. Different simulations made use of different combinations of scattering kernels. All phonon-related scattering kernels used phonon energies and polarization vectors computed on a regular grid in a Brillouin zone from a BvK model [67].

Figure 7.7 displays different $I(Q, \omega)$ for the aluminum plate. The experimental result is in panel (a) and panels (b)–(e) show simulated data. In (b) only the incoherent elastic and incoherent single phonon scattering are included (The intensity was enhanced to show detail – the values on the colorbar give an

indication of the scaling). In (c) only the coherent elastic (powder diffraction) and the coherent single-phonon inelastic scattering are included. In (d), all of the kernels in (b) and (c) with the addition of a multiphonon kernel using the incoherent approximation. In (e), all of the kernels in (d) are used with multiple scattering turned on.

Overall, the features shown in the experimental data (a) and the simulated data (e) agree very well. Comparison of (b) and (c) shows that coherent scattering gives rise to more features such as diffraction peaks and phonon dispersion curves. It is evident from comparing (c) and (d) that multiphonon scattering increases in intensity at higher Q. The most obvious difference in (d) and (e) is in the elastic line, which shows that multiple scattering of coherent elastic scattering seems to contribute similarly to incoherent scattering in the elastic line. The elastic lines in (a) and (e) suggest that the sample used in the experiment may contain traces of an additional phase, most likely from a surface layer of Al_2O_3. From quantitative comparisons between simulation and experiment, it was found that multi-phonon scattering plays a significant role in the observed intensity, whereas multiple scattering plays little role in this Q range **?**). This is as expected because the sample is only a 5% scatterer.

7.6.2 Approximate Corrections for Simultaneous Multiphonon and Multiple Scattering

Corrections for multiple scattering have been performed in many ways, from subtracting a constant from the data [1], to full Monte-Carlo simulations [2]. At high temperatures, the former does not account for the slope of the scattering past the cutoff energy. The latter can be computationally intensive, and requires detailed information about the shape of the sample. Here we take an approach of intermediate complexity, reported previously [3]. For both multiple scattering and multiphonon scattering, a two-scattering profile involves a convolution of two single-scattering profiles. In either case, the idea is that an n-phonon-scattering profile, $P^n(E)$, is related to the 1-phonon-scattering profile, $P^1(E)$, through the recursion relation:

$$P^n(E) = \int_{-\infty}^{\infty} P^{n-1}(E')\, P^1(E - E')\, \mathrm{d}E' \ . \tag{7.48}$$

For multiple scattering processes, the n-phonon probability function has additional position and momentum dependencies, that do not appear for multiphonon scattering processes. Sears, *et al.* [1], argue that the integrals for multiple scattering are related to those for the multiphonon scattering through slowly varying functions of Q and E. Here we take these functions to be constants, a_n. In essence, we make the approximation that the position and momentum dependencies can be factored out. Thus,

$$I(Q,E) = N' \left[\sum_{n=1}^{\infty} (1 + a_n) S^n(Q,E) \right] , \tag{7.49}$$

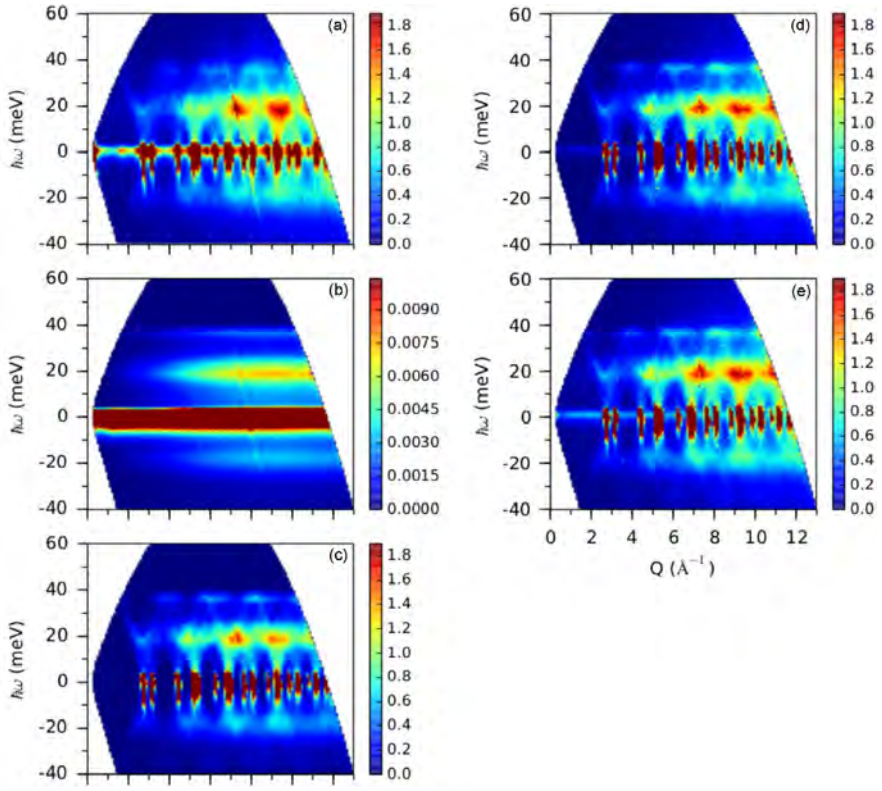


Figure 7.7: $I(Q, \omega)$ plots of aluminum inelastic spectra obtained from experiment (**a**), and simulations (**b**)–(**e**) of an aluminum plate at room temperature in the ARCS instrument. Scattering kernels included in the simulations are (**b**) incoherent elastic and incoherent inelastic single-phonon scattering (see note on intensity scaling in text), (**c**) coherent elastic (powder diffraction) and coherent inelastic single-phonon scattering, (**d**) all kernels included in (b) and (c) plus a multi-phonon kernel using an incoherent approximation and (**e**) all kernels included in (d) plus multiple scattering.

where $I(Q,E)$ is the experimentally-determined total scattering (including multiple scattering), $S^n(Q,E)$ is the n-phonon scattering (both creation and annihilation), and N' is a normalization constant. Note that $I(Q,E)$ is distinct from the scattering function, $S(Q,E)$, which does not include multiple scattering [4]. (When we stripped the elastic peak from the data, the dominant multiple elastic scattering is removed, so the index n in Eq. 7.49 starts at 1 rather than 0.)

Consistent with this factoring of Q and E dependencies, we make the incoherent approximation [5]:

$$S^n_{\rm coh}(Q,E) = \frac{\sigma_{\rm coh}}{\sigma_{\rm inc}} S^n_{\rm inc}(Q,E)\ , \tag{7.50}$$

where we apply this equation to the 1-phonon terms as well as all higher orders. The last step in our procedure will be to assess any error this has introduced into our analysis.

Our next assumption is that $a_n = C'_{\rm ms}$ for all $n \geq 2$, where $C'_{\rm ms}$ is a single constant that relates the multiple scattering to the multiphonon scattering. Since the multiphonon scattering drops off rapidly with increasing n, this approximation will only have a small effect on our results. The final normalization is performed with the total scattering, so the factor $1 + a_1$ is included in the normalization constant. We find:

$$I(Q,E) = N\left[S^1_{\rm inc}(Q,E) + (1 + C_{\rm ms})S^{2+}_{\rm inc}(Q,E)\right]\ , \tag{7.51}$$

where $N = N'(1+a_1)\,(1 + \sigma_{\rm coh}/\sigma_{\rm inc})$ is the normalization constant, and $1 + C_{\rm ms} \equiv (1 + C'_{\rm ms})/(1 + a_1)$. Also, for notational convenience,

$$S^{j+}(Q,E) \equiv \sum_{n=j}^{\infty} S^n(Q,E)\ . \tag{7.52}$$

For a cubic crystal, and a fixed value of $C_{\rm ms}$, we can now find the DOS by solving Eq. 7.51 in the manner described by Sears, *et al.* [1].

Since we do not know the value of $C_{\rm ms}$ a-priori, we generate a list of possible values, and solve for the DOS at each one. In the current study, values of $C_{\rm ms}$ between 0.0 and 2.0 were tested. It then remains to select the "best" DOS from those generated with the different $C_{\rm ms}$. This was done by minimizing a penalty function constructed to find the DOS that produced $S(E)$ that best satisfied the following conditions:

(1)
$$\frac{I(E)}{N} = S_{\rm inc}1(E) + (1 + C_{\rm ms})S^{2+}_{\rm inc}(E)\ , \tag{7.53}$$

where the implied sum over Q allows us to compare the partially coherent scattering on the left with the totally incoherent scattering on the right.

(2) The experimental noise at energy transfers near the incident energy oscillates about $(1 + C_{\rm ms})S^{2+}_{\rm inc}(E)$.

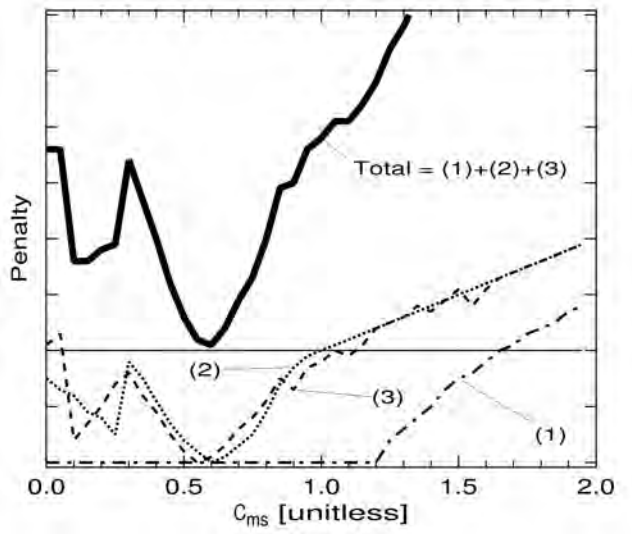


Figure 7.8: Penalty functions for nickel at 300 K, as defined in the text. The dash-dotted line (1) relates to the overall fit, the dotted line (2) relates to the noise near the incident energy, and the dashed line (3) relates to the slope near the incident energy. The solid line is the sum of these three contributions (offset).

(3) At energy transfers near the incident energy, the slope of a linear fit to the experimental noise matches the slope of a linear fit to $(1 + C_{ms})S^{2+}_{inc}(E)$.

These three criteria are correlated, but are not identical. For nickel at 300 K, these three contributions and their sum are shown in Fig. 7.8. Figure 7.9 shows the best fit to the normalized scattering, $I(E)/N$ for nickel at 300 K, which had $C_{ms} = 0.6$.

The DOS curves obtained this way were fit with a Born–von Kármán model, from which all phonon contributions to the scattering, both coherent and incoherent, were calculated. With these results, and with the final value for C_{ms}, the calculation was checked against the measured scattering. It was our experience that this procedure worked well for the present case of nickel, and also works for cases of other BCC and FCC materials.

[1] V.F. Sears, E.C. Svensson, and B.M. Powell, "Phonon density of states in vanadium," Can. J. Phys. **73**, 726 (1995).

[2] E. Johnson, and L. Robinson, "Neutron multiple scattering and absorption factors," Rev. Sci. Instrum. **60**, 3447 (1989).

[3] M. Kresch, O. Delaire, R. Stevens, J.Y.Y. Lin, and B. Fultz "Neutron scattering measurements of phonons in nickel at elevated temperature," Phys. Rev. B **75**, 104301 (2007).

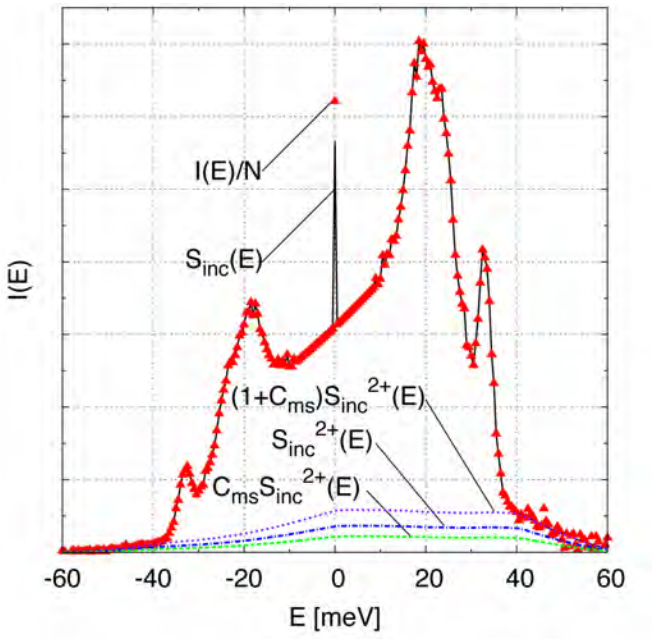


Figure 7.9: Best fit to scattering for nickel at 300 K. The triangles are the normalized experimental scattering, $I(E)/N$. The solid line shows the fit, $S^1_{\rm inc}(E) + (1 + C_{\rm ms})S^{2+}_{\rm inc}(E)$. The dashed line is the multiple scattering, $C_{\rm ms}S^{2+}_{\rm inc}(E)$. The dash-dotted line is the multiphonon scattering, $S^{2+}_{\rm inc}(E)$. The dotted line is the sum, $(1 + C_{\rm ms})S^{2+}_{\rm inc}(E)$.

[4] G. Placzek and L. Van Hove, "Crystal Dynamics and Inelastic Scattering of Neutrons," Phys. Rev. **93**, 1207 (1954).

[5] G.L. Squires, *Introduction to the Theory of Thermal Neutron Scattering*, (Dover Publications, New York, 1997), Chap. 3, Sec. 10, p. 57.

See also: V.F. Sears, "Slow-neutron multiple scattering," Adv. Phys. **24**, 1 (1975).

Further Reading

The contents of the following are described in the Bibliography.

Mark Lutz and David Ascher: *Learning Python* (O'Reilly & Associates, Inc. 1999).

Stephen W. Lovesey: *Theory of Neutron Scattering from Condensed Matter Vol. 1*, (Clarendon Press, Oxford 1984).

A. A. Maradudin, E. W. Montroll, G. H. Weiss and I. P. Ipatova: *The Theory of Lattice Dynamics in the Harmonic Approximation, Second Edn.* (Academic Press, New York, 1971)

Varley F. Sears: *Neutron Optics*, (Oxford University Press, New York and Oxford 1989).

G. Shirane, S. M. Shapiro, and J. M. Tranquada: *Neutron Scattering with a Triple-Axis Spectrometer*, (Cambridge University Press, Cambridge 2002).

G. L. Squires: *Introduction to the Theory of Thermal Neutron Scattering*, (Dover, Mineola NY 1978).

C. G. Windsor: *Pulsed Neutron Scattering*, (Taylor and Francis, London 1981).

Chapter 8

Computational Scattering Science

8.1 Software for Inelastic Scattering

8.1.1 Overview of Capabilities

Data from inelastic chopper spectrometers can reveal the fundamental dynamical processes in materials or condensed matter, but not without substantial data analysis to produce even a basic graph of $S(\vec{Q}, E)$, the intensity as a function of momentum and energy transfers. The elementary excitations in solids are a substantial topic in themselves, and are covered in the earlier chapters of this book. Making comparisons between the experimental $S(\vec{Q}, E)$ and the underlying theory requires another level of software sophistication. This chapter describes the essential types of computational workflows that can be used to go from data to science.

A schematic overview of software for inelastic scattering is presented in Fig. 8.1. This figure, first presented as a roadmap in Sept. 2001, shows three paths for extracting scientific results from the raw data in the upper left corner. They are:

- Data reduction. The raw data are not interpretable by humans. They must be "reduced" to forms that display trends of intensity versus momentum and/or energy transfer between neutrons and the sample. Identifying features in these data may be publishable in their own right.

- Modeling. The intensity as a function of energy and momentum transfer, $I(\vec{Q}, E)$, is a quantity that can be obtained from reduced data, and it can be calculated by theoretical models. The comparison between the models and the measurements can reveal new science.

- Direct Experiment Simulation. Although data reduction procedures make

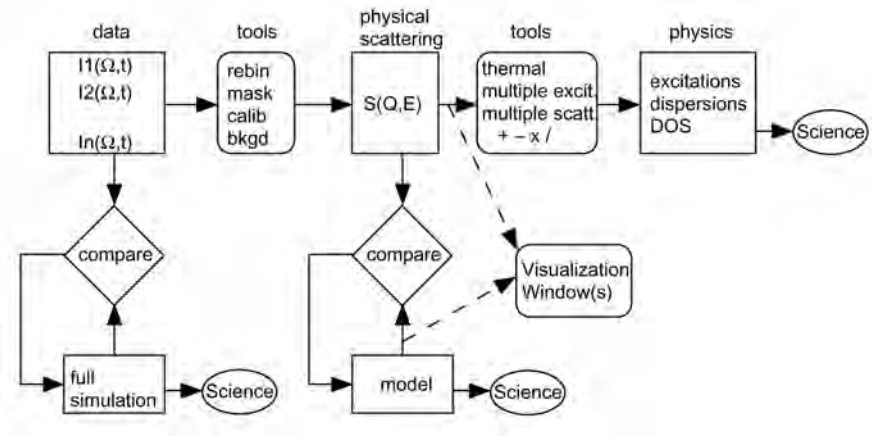


Figure 8.1: Schematic of *DANSE* software for inelastic scattering.

a major effort to remove the quirks of and instrument from the experimental data, this practice is not perfect. For example, it is not always possible to correct accurately for the instrument resolution. Sometimes it is better to simulate the data obtained from an entire experiment, including the effects of the instrument and the sample.

These three paths to science are described in sequence below.

8.1.2 Data Reduction

The first path, horizontally across the top of Fig. 8.1, is the traditional approach to data reduction and visualization. The goal of this analysis is to obtain the intensity as a function of momentum transfer and energy transfer, $S(\vec{Q}, E)$. To do so, the data arrays of counts acquired in terms of instrument parameters such as detector pixel and arrival time must be converted into normalized intensities with physical units such as Å and meV. Instrument backgrounds and other distortions must also be removed. It is often necessary to correct for other distortions caused by, for example, multiple scattering or multiple excitations. By implementing all software components in the interpreted Python language, the architecture can provide a set of components that can be replaced or rearranged to test different processing algorithms. It is also important that data streams can be piped into visualization windows for inspecting the data after the different steps of data processing. An example of a component for energy rebinning of data from a chopper spectrometer is described in the context of the data stream architecture in Sect. **??**. Some of the required corrections were described in Sects. 7.3 – 7.6.

8.1.3 Modeling

The second and third paths for extracting scientific results from experimental data are designed to connect experimental data to theory or analytical models.

The second path, the vertical chain in the center of Fig. 8.1, is for comparing experimental results to models of sample dynamics. This path is especially appropriate for analytical models with adjustable parameters. Consider its use for the "neutron-weighting problem" in phonon dynamics, which originates as follows. A measured $S(Q, E)$ from a polycrystalline sample of a pure element can often be converted into a phonon density-of-states using a thermal correction procedure. On the other hand, a phonon DOS from a binary compound cannot be otained from the measured $S(Q, E)$ because the different elements in the compound do not scatter neutrons with equal efficiencies, causing a "neutron-weighting" of the experimental spectra. Without knowing the lattice dynamics of the compound, it is impossible to know the distortions of an experimental DOS obtained from a measured $S(Q, E)$ after a thermal correction procedure.

The phonon scattering efficiencies of the different atoms are well known, so a lattice dynamics model can be used to calculate an experimental spectrum. We do so with an iterative procedure where the force constants in the dynamics model are varied to obtain the best fit to the experimental data. We therefore can use a lattice dynamics model plus a "neutron weighting" correction to obtain the true phonon density-of-states from the measured $S(Q, E)$ of the compound. The force constants are obtained as parameters that give the best fit between calculation and experiment. This is now a routine procedure, and new experiments can be designed around this capability.

Besides phonon dynamics in ordered compounds, there are many dynamics models that can be compared to experimental data. Applications include data corrections, the determination of parameters such as force constants or exchange stiffnesses, or testing if the model is in fact consistent with the measured $S(Q, E)$. Several types of dynamics models have been included in systems for inelastic scattering software:

1. lattice dynamics with a Born–von Kármán model (periodic structure)
2. spinwave dynamics with a Heisenberg hamiltonian on a periodic structure
3. lattice dynamics on a disordered structure, using a moments analysis of the dynamical matrix
4. lattice dynamics with density functional theory on ordered structures, or disordered structures with supercell calculations
5. ab-initio molecular dynamics calculations with density functional theory on ordered structures, or disordered structures with supercell calculations
6. spin dynamics in a paramagnetic model

7. spin dynamics with density functional theory for chemically ordered structures

8.1.4 Direct Experiment Simulation

The third path from data to science is shown on the left of Fig. 8.1. This approach is a direct simulation of the data measured at the detectors. It is based on Monte Carlo codes that are used in the neutron community for simulation of instrument performance. These codes have been tested and validated against the performance of real instruments. These Monte Carlo simulations have been used with molecular dynamics simulations to perform direct simulations of experimental data. A number of molecular dynamics simulations are available to the theory community. These simulations are complementary to the analytical models of Sect. 8.1.3, and are sometimes advantageous. For example, no approximations are needed when implementing structural models of disordered solids, which are not handled well by the methods of Sect. 8.1.3 (except when the Q information is ignored as in methods 3 and 6).

In these direct experiment simulations, the sample is used as a component in the Monte Carlo computations of neutron trajectories through the instrument, transforming an individual incident neutron into a neutron scattered into the detector array. Typically, simulation results for the primary flight path are stored and used for several simulations. For lattice dynamics simulations, a large four-dimensional (Q_x, Q_y, Q_z, E) data structure of dispersive excitations can be pre-calculated, and the velocity-velocity correlation function of the atoms in the sample can be sampled. This same path could be followed for classical spin dynamics simulations.

8.2 Simulation of Neutron Scattering Experiments with MCViNE

8.2.1 General Features of Monte Carlo Ray Tracing Simulations

Except when interacting with atomic-scale periodicities in the sample, much of the simulation of a neutron scattering experiment can be done with the assumption that neutrons are particles that move in approximately straight lines. Monte Carlo (MC) ray tracing simulations of neutron scattering spectrometers with support of multiple scattering were performed as early as the 1970s using MCS**?**) and MSCAT**?**) to understand the effects of multiple scattering on the measured spectra **?**). In the 1990s, with increasing needs of simulating neutron instruments for the purpose of instrument design and optimization, several MC neutron ray tracing packages emerged, including McStas**??**), Vitess**?**), Ideas**?**), and NISP **?**). Simulations with these packages were intended primarily for instrument design, such as calculating energy resolution under different operating conditions (see, for example, **??????????**)). Unlike MCS and MSCAT, most of these newer MC software packages (with a notable exception in NISP) treat

a neutron instrument as a linear chain of neutron optical components. Each component performs one modification of neutron beam characteristics such as spatial divergence and energy distribution. The approach based on direct acyclic graphs (Sect. **??**) greatly simplifies the coding of instrument simulations. Such an approach is adequate when neutrons are rarely scattered back to the upstream components. These software packages, especially McStas and Vitess, have made important contributions to neutron instrument design.

A Monte Carlo neutron ray tracing simulation follows the neutron from the moderator through the guides and choppers in the primary flight path, to the sample with various possible complexities, and then to a detector or get-lost pipe. There are substantial differences in these components, and natural differences in the structure of the software programs to simulate them.

Detector system

The detector systems in modern neutron scattering instrumentation have a modular and sometimes hierarchical organization, facilitating their construction, maintenance and validation. For example, the four direct geometry time-of-flight spectrometers at Spallation Neutron Source (SNS) **?**), ARCS **?**), SEQUOIA **??**), CNCS**?**), and HYSPEC**?**) have detector arrays that use use the same so-called "8-packs" **?**), each of which is a detector module with eight ^{3}He linear position sensitive detector tubes (LPSD). The 8-packs are arranged in a vertically-oriented cylindrical geometry around the sample position, forming a hierarchical organization of pixels, tubes, packs, and detector rows. A simplified illustration of the detector hierarchy can be found in Figure 8.2(c).

Sample Assembly

Samples, sample holders, and sample environments comprise a "sample assembly." The sample assembly is an aggregation of neutron scatterers near the sample position, as illustrated in Figure 8.2(b). A challenges in simulations of a sample assembly is that neutrons can scatter back and forth among its constituents. Even within a sample of some thickness, multiple scattering is possible. Section 8.3.2 shows simulated and measured effects of multiple scattering in the MICAS furnace, which contains heating elements, heat shields, in addition to the sample **?**).

8.2.2 General capabilities of MCViNE

MCViNE**?**) is a general purpose neutron ray tracing package that combines the two different approaches taken by MSCAT-like packages and McStas-like packages. At the instrument level, it allows users to create a simulation application as a linear chain of neutron components, each fully configurable by its type and corresponding geometrical and physical properties using either command line options or an xml-based configuration file. At the component level, two general components exist with support of hierarchical representation

for the sample assembly and the detector system. A detector system can be specified with an xml file describing its hierarchy, while a sample assembly is specified by a collection of xml files, one for the geometrical organization of the constituent scatterers, and others describing the scattering mechanisms for them. Multiple-scattering among scatterers in a sample assembly can be turned on and off by a command line option.

8.2.3 Software Architecture of MCViNE

MCViNE is implemented in C++ and Python. The two C++ layers contain fundamental mathematics, the neutron ray tracing mechanism, the mechanism to include components from other MC ray tracing packages, and support of multiple scattering and composite scattering kernels. The neutron components from McStas are included in the C++ layer so they can be used individually rather than as a whole compiled instrument. For the sample, the combination of the object oriented features of C++ and the speed of the compiled code make C++ appropriate for implementing the composite kernel and scattering kernels. The "composite scatterer," of MCViNE will be explained in detail in Sect. 8.2.4.

A Python layer on top of the two C++ layers allows construction of a component chain similar to McStas-like neutron ray tracing software packages, provides the interface to the C++ components to include in that chain, and allows for introduction of components that are completely written in Python. This last feature makes it extremely easy to create simple neutron components, and to create prototypes of more sophisticated neutron components.

8.2.4 Composite Scatterer

This section starts with basic concepts of the "composite scatterer," followed by an introduction to the essential object-oriented software designs centered around it, and finishes with examples of ray-tracing algorithms in sample assemblies and detector systems enabled by these designs.

Concepts

In MCViNE, a "composite neutron scatterer" represents a group of physical objects, for example, a powder sample in an aluminum can, a single crystal sample surrounded by a furnace, or (perhaps surprisingly) a detector system. An "elemental scatterer" is a scatterer without constituent scatterers. A "homogeneous scatterer" is one kind of elemental neutron scatterer, whose scattering function is homogeneous within its volume. The scattering properties are modeled using one "scattering kernel" or a combination of several "scattering kernels." In the spirit of Sect. 1.1, each scattering kernel allows one scattering mechanism, such as incoherent one-phonon nuclear scattering or coherent magnetic scattering.

Figure 8.2(a) shows an abstract hierarchy of a composite scatterer in MCViNE. Figure 8.2(b) depicts a sample assembly that is a composite of two-level hierar-

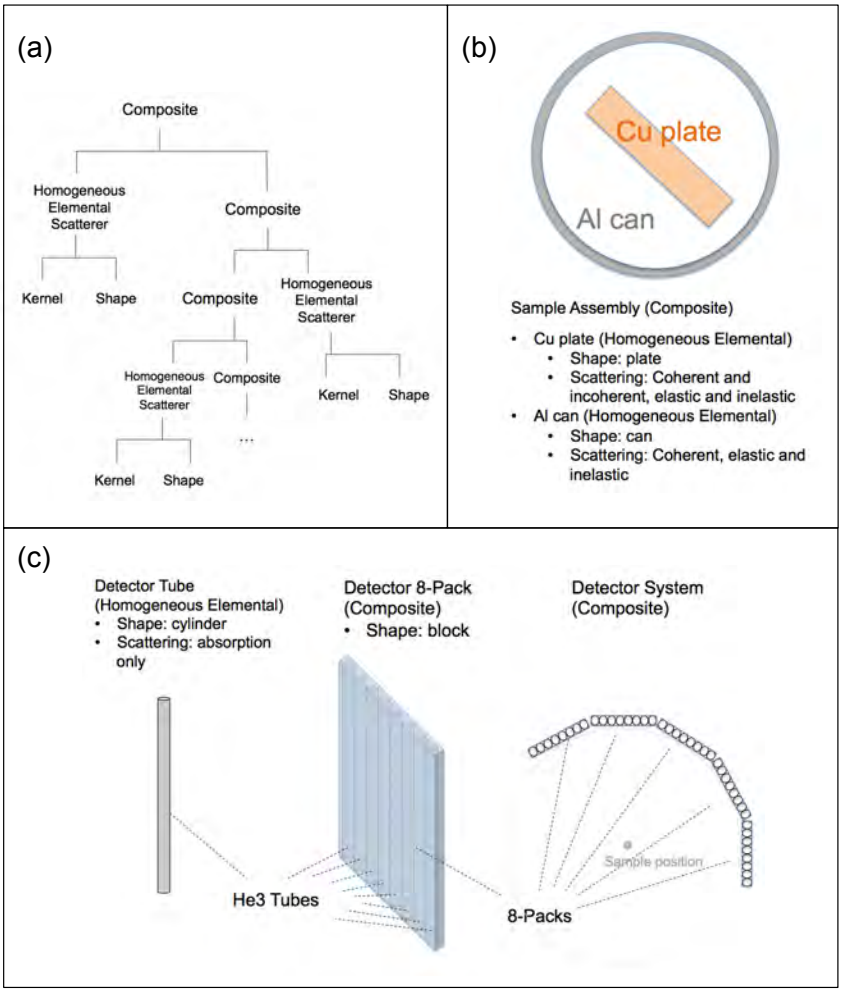


Figure 8.2: Concepts in scattering composites: composite scatterer, homogeneous scatterer, shape, scattering kernels. (a) an abstract hierarchy of an abitrary scattering composite. (b) and (c) are concrete examples of such hierarchies. (b) top view of a sample assembly consisting of an aluminum can and a copper plate. (c) a detector system – how it is constructed from an elemental scatterer (detector tube) in a three-level hierarchy.

chy, in which the bottom level consists of two homogeneous elemental scatterers: one aluminum can and one copper plate sample. Figure 8.2(c) represents a detector system consisting of ^{3}He eight-packs that form roughly a cylindrical arrangement around the sample position. The detector system is represented in MCViNE in a three-level hierarchy: at the bottom level is the ^{3}He detector tube; at the middle level, 8 such tubes construct an 8-pack; at the top-level, the detector system consists of a collection of 8-packs. Such hierarchical representations allow MCViNE to model the closely the physical reality in a way that parallels the experimental work.

Object-Oriented Designs

Design patterns for object-oriented programming are developed in the classic text by Gamma, et al., **?**). Two important object-oriented design patterns used in MCViNE are

- The "composite pattern," which describes the part-whole relationship and to represent a hierarchy. It allows clients to treat composites and their constituents in a uniform way.
- The "visitor pattern", which allows separation of operations from the objects to be operated on, so new operations can be added without touching these objects.

Using these design patterns for scattering composites unifies the programming interfaces to the operations on both the composites and on the individual elemental objects. Composite and visitor patterns are used in three major aspects of the MCViNE neutron scattering model: neutron scatterers, geometric shapes of scatterers, and scattering kernels, as described next.

Neutron scatterers. By using the composite pattern, algorithms for multiple scattering are consolidated in one implementation. Scattering from a composite neutron scatterer starts with a determination of which constituent intersects the incident neutron ray, and then delegates the scattering assessment to that particular constituent. The constituent could be a composite itself that requires another delegation for scattering. The hierarchical representation of neutron scatterers and this recursive algorithm work for both samples and detector systems, and can improve computing efficiency and code maintenance.

Geometric shapes Using constructive solid geometry (CSG) (see for example **?**)), composite shapes are constructed from basic shapes such as cylinders and blocks, and composites by using operations such as union, intersection and difference. Ray-tracing through shapes is therefore simplified as visitor methods of the primitive shapes and the binary shape operations.

Scattering kernels Composite scattering kernels make it easy for the Monte Carlo algorithm to sample a total $S(\mathbf{Q}, \omega)$ consisting of both slowly-varying regions and regions containing sharp features, such as diffraction and coherent phonon scattering, by allowing users to combine different kernels such as incoherent and coherent kernels in a kernel composite. In addition, users can

organize scattering kernels into groups; this makes it easy to apply importance sampling (assign different weights to different kernels or kernel groups) in simulations.

A "scattering kernel" in MCViNE is an abstraction of the scattering mechanisms such as diffraction, nuclear scattering by phonons, or magnetic scattering by spin waves. It does not include the sample geometry, but only the scattering physics. By separating the implementation of "scatterer," "shape," and "scattering kernel," a sample in MCViNE generally consists of a combination of scattering kernels that can be added or deleted easily. Furthermore, the scattering kernel library in the MCViNE framework can be extended without affecting the logic of geometric ray tracing, which is implemented in "shape" and "scatterer." For example, simple scattering kernels taking a histogram form of $S(Q, \omega)$ can be used, or more sophisticated phonon scattering kernels that take phonon energies and polarizations as inputs (these were used in the examples of vanadium and aluminum in Sect. 1.1). Many others are possible.

8.2.5 Algorithms

Ray Tracing

In components such as sample assemblies or detector systems, neutron scatterers are represented by a hierarchy of objects with shapes and scattering mechanisms. Ray tracing of a neutron happens by first determining which neutron scatterer at the top level of the composite hierarchy intercepts the neutron. This is done by computing the intersections of the forward ray of the neutron and all the shapes of the top-level constituents. A random selection might be necessary if multiple top-level objects intercept the neutron. After the top-level neutron scatterer is identified, the neutron is propagated to the front surface of the scatterer if necessary (not necessary if the neutron is already inside the scatterer) with appropriate attenuation, and the ray tracing algorithm then recurses into itself if the scatterer is a composite. Otherwise a point in the forward path of the neutron inside the scatterer is picked at random, and the neutron is propagated to that point with attenuation. At this point a scattering (or an absorption) mechanism of the scatterer is randomly picked, and the neutron is either scattered with its probability adjusted, or is absorbed.

Multiple Scattering

Multiple-scattering (MS) is naturally supported in MCViNE scattering composites, implemented with a recursive algorithm. Here we differentiate between two types of multiple scattering

- single-scatterer multiple scattering (SSMS) to describe the multiple scattering within one neutron scatterer
- multiple-scatterer multiple scattering (MSMS) to describe multiple scattering among neutron scatterers.

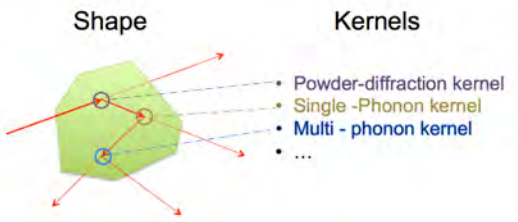


Figure 8.3: An example of multiple scattering within one scatterer. The incident neutron was scattered three times by three different scattering kernels. At each scattering point, the original neutron is also propagated out of the scatterer with proper attenuation. Red arrows are paths of neutron propagation. Circles highlight the location of scattering. Different scattering events are coded using different colors.

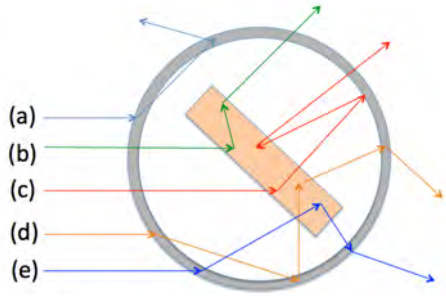


Figure 8.4: An example of multiple scattering in a concentric sample assembly. Five (out of infinite) possible multiple scattering paths are illustrated (see text).

Figure 8.3 shows an SSMS in which a neutron is scattered three times before exiting a scatterer. Each time a neutron is scattered inside a homogeneous scatterer, the original incident neutron packet is split into two neutron packets for computational efficiency. One neutron packet is propagated through the scatterer with its probability lowered by attenuation, while the other is scattered by one of the scattering kernels, chosen by a random selection, at a point also randomly selected along the forward path. This splitting process repeats for the scattered neutron several times, as illustrated in Figure 8.3, until either the neutron probability is lower than a pre-selected limit, or the maximum order of multiple scattering is reached. When either termination condition is met, any neutron that is still inside the scatterer is allowed to propagate out, with its probability attenuated appropriately.

Figure 8.4 shows an example of multiple scattering in a sample assembly consisting of a "concentric" arrangement of a sample and a sample environment. Only five possible multiple scattering paths are illustrated, and the splitting processes are not shown. Sample components in some linear-chain-based

MC ray tracing packages can support SSMS, but they do not have abstractions similar to composite scatterer or scattering kernel. As a result, the SSMS algorithm must be duplicated in these sample components as explained in Sect. **??**. In MCViNE, implementations of new scattering kernels can be added without reimplementing the multiple scattering algorithm. McStas supports MSMS partially for a concentric sample assembly by, for example, adding a second outer cylinder into the simulation component chain **?**). Path (e) of Figure 8.4 is included this way, but path (c) of Figure 8.4 is supported only with more work on the user's part. The multiple scattering algorithm of MCViNE is more comparable to that of MSCAT**?**) in which the sample and the sample environment are treated together in the multiple scattering loop, offering complete treatment of SSMS and MSMS. However, MCViNE allows a straightforward increase in complexity of the sample and sample environment interactions by using the recursive MS algorithm enabled by OOP.

Ray tracing in a detector system

For detector systems, MCViNE takes advantage of the hierarchical representation for neutron scatterers, only in this case the elemental homogeneous neutron scatterer is the ^{3}He detector tube that intercepts neutrons and records them. MCViNE reuses the code for ray tracing in a composite scatterer for simulating ^{3}He detector systems, and the new code needed is a scattering kernel for the ^{3}He that takes into account gas absorption. When a neutron is sent to a detector system shown in Figure 8.2(c), for example, the generic ray tracing algorithm for composite scatterers first checks whether the top level composite scatterer is penetrated by the neutron. If so, all constituents of the composite scatterer, i.e. the detector packs, are examined to determine which of them intercepts the neutron. Unless a neutron traverses a gap between detector packs, the detector pack is identified and then its constituents, the 8 detector tubes, are examined for neutron detection. The path of a neutron through the detector tube is then computed by ray tracing of the neutron through a cylinder, and a MC sampling picks a point in the path for the neutron to be absorbed. (The ray tracing through a cylinder takes care of the parallax effect of the detector tube.) Additionally, the appropriate weighting multiplier for neutron detection probability (computed from absorption probability depending on the ^{3}He pressure and the length of the neutron path through the tube) and the time-of-flight are computed for the neutron to be recorded as a detector event in a "virtual detector electronics device." This hierarchical approach to detectors allows for adding more physical features, including details of the charge cloud around the anode wire of the detector.

8.3 Example of a Neutron Simulation

8.3.1 Relationship to Data Reduction

Inelastic neutron scattering data contains a complex overlap of signals with several origins. Multiple scattering and multi-excitation scattering are obvious sources of signal that are usually unwanted in final data analysis. These contributions will vary depending on factors such as sample temperature, sample thickness, additional scattering components in the beam, such as sample containers and sample environments, and the complex geometrical arrangement of all these components. In addition, the instrument has a complex resolution function. Attempts to describe this resolution function analytically separate contributions to the resolution and include uncertainty in timing from the source and differences in path length due to the finite chopper opening **?**). The data reduction process attempts to strip the measured data of unwanted contributions to the measured signal, leaving only single-excitation inelastic and elastic scattering from the sample. Because the individual signal contributions are not known, this process is done with a series of approximations that are imperfect, and are also incapable of capturing the complex overlap of signals.

As instruments improve, more detailed information is available from experimental measurements. Some trends and phenomena can be ignored when counting statistics are poor, but become more evident in data of higher quality. There is a growing need for more accurate and reproducible approaches to data reduction. Simulations of neutron experiments offer an alternative way to identify and isolate the individual and mixed contributions to measured intensities.

Section 8.2 described some technical features of the Monte Carlo VIrtual Neutron Experiment (MCViNE) software package. In short, MCViNE is a Monte Carlo neutron ray tracing package developed during the commissioning of the ARCS instrument at the SNS **???**). This package has several improvements over previous Monte Carlo neutron ray tracing packages, including the use of a recursive function to inherently handle multiple scattering and a flexible architecture for re-arrangement of components.

8.3.2 Scattering from a Sample in a Furnace

A high-temperature furnace used in the ARCS spectrometer is shown in Fig. 8.10. It has many layers of shielding that cause scattering that appears as an unwanted experimental background. In experimental practice, this background can be measured and then subtracted from data acquired from the sample in the furnace. There are two subtleties about this practice that bring abmiguities to this experimental practice

- When the sample is in the furnace, it removes some of the incident beam, perhaps 10%, by scattering or absorption. The downstream part of the furnace does not see the full incident beam that was seen for the empty

furnace measurement. The typical practice is to assume half of the furnace scattering is from the downstream side, so the full background of an empty furnace is not subtraced from the measured data with the sample. If the sample scatters 10% of the incident neutrons, perhaps 95% of the background from the empty furnace should be removed by subtraction.

- Unfortunately, the parts of the furnace upstream and downstream from the sample are not in the same position, and they give different background signals. Sometimes it is possible to make a third measurement (in addition to the empty furnace, furnace plus sample) with a strongly absorbing material in place of the sample. In this case, the downstream part of the furnace will receive less of the incident beam, and perhaps the scattering contributions from the upstream and downstream parts of the furnace can be isolated by comparing these measurements of a furnace with a black absorber to the measurements of the empty furnace. This can be challenging when the sample and absorber are smaller than the full width of the incident beam, but in principle it can work well.

- A much more difficult problem occurs when there is a first scattering from the furnace to the sample and a second scattering from the sample to the detector (or vice-versa). The intensity from these neutrons that interact with both the sample and the furnace can only be observed when the sample is in the furnace, and an empty furnace measurement is not helpful. Perhaps samples of different thickness can be used to help analyze these multiple scattering effects, but this is not a popular method.

These three problems present an opportunity for simulations. Accurate simulations of the empty furnace, furnace containing the sample, and the sample by itself, can elucidate the contributions of background scattering from the furnace, and between the furnace and sample. These results could show how much of the empty furnace measurement to subtract from the data with the sample, and then show how much multiple scattering between the sample and furnace needs to be removed as well.

8.3.3 Steps of Simulations with MCViNE

A "sample assembly" is constructed from several scatterers, arranged in a specific geometrical configuration to model the physical layout of all components in the path of the beam. Construction of the sample assembly is a primary task before starting the simulation, much in the same way as an experimentalist modifies the sample and its envioronment, rather than the instrument itself. The instrument configuration of the various guides and choppers in the primary flight path upstream of the sample has been previously developed, and neutron beams on the sample for many conditions have already been simulated. The detector configuration is static, and also does not require user modification. The simulation sequence follows four steps illustrated in Fig. 8.5

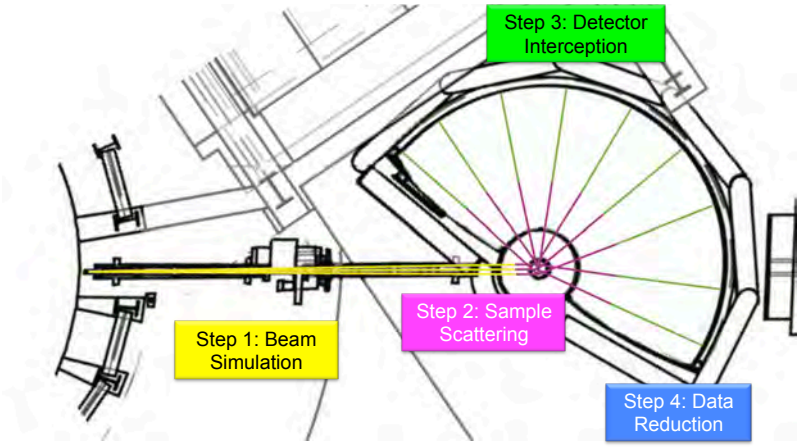


Figure 8.5: The simulation proceeds in four steps, as shown for a schematic of the ARCS instrument. First, the neutrons travel from the moderator to the sample (yellow path). Second, the neutrons are incident on the sample and scatter from the sample (pink path). Third, the neutrons are intercepted by the detector array (green path). Fourth, the event-mode NeXus file is reduced using Mantid.

1. First, the beam is simulated as the neutrons travel from the moderator to the sample. The instrument components before the sample are modeled in a linear chain, and scattering probabilities are propagated and updated by each component in the chain until the neutrons reach the sample. This beam simulation is archived, and does not need to be repeated unless the user selects a new configuration of operating parameters.

2. Second, the neutrons are scattered from the sample assembly. The sample assembly includes the sample and furnace as described below.

3. Third, the scattered neutrons are intercepted by the detector array. This assigns a detector pixel ID and time-of-flight to each neutron that reaches the detector. The collected neutrons are then processed into an event-mode NeXus file. This approach to data collection is the same as is currently in use at the SNS.

4. The final step in the simulation sequence is the reduction of the NeXus file using Mantid. This reduction step uses the routines and procedures that are used for reduction of experimental data.

It is also possible to introduce a radial collimator between the second and third steps. The neutrons are forced to travel in straight lines through the collimator after scattering from the sample and before entering the detectors. The collimator triangulates the neutron paths to converge on the sample, and

Figure 8.6: Photograph of the inner edge of collimator septa around the sample region of the ARCS spectrometer. The perspective of the photo is from a camera slightly above the position of the sample. The detector array lies approximately 2 m beyond the collimator towards the top of the photograph.

eliminates much of the scattering from the furnace components that are at different radii. The radial collimator gave a remarkable improvement in the background from the ARCS instrument. Figure 8.6 shows the actual radial collimator in position around the sample space. For the simulation, the radial septa are configured with the same geometrical positioning and spacing as the real collimator, but are modeled as "ideal" in that they are infinitely thin and have 100% absorption.

8.3.4 Furnace Simulation Template

8.3.5 Furnace Geometry

The MICAS furnace, shown in Fig. 8.7 is capable of heating samples to 1900 K (or more practically, 1500 K). In use, the entire furnace is lowered with a crane into the sample region. The bottom-most portion of the furnace, the many-layered region in the schematic and the crinkled-foil region in the photo, is the part of the furnace in the path of the neutron beam. This is this region that must be accurately described in any Monte Carlo simulation.

The critical region of the furnace in the neutron beam contains the sample

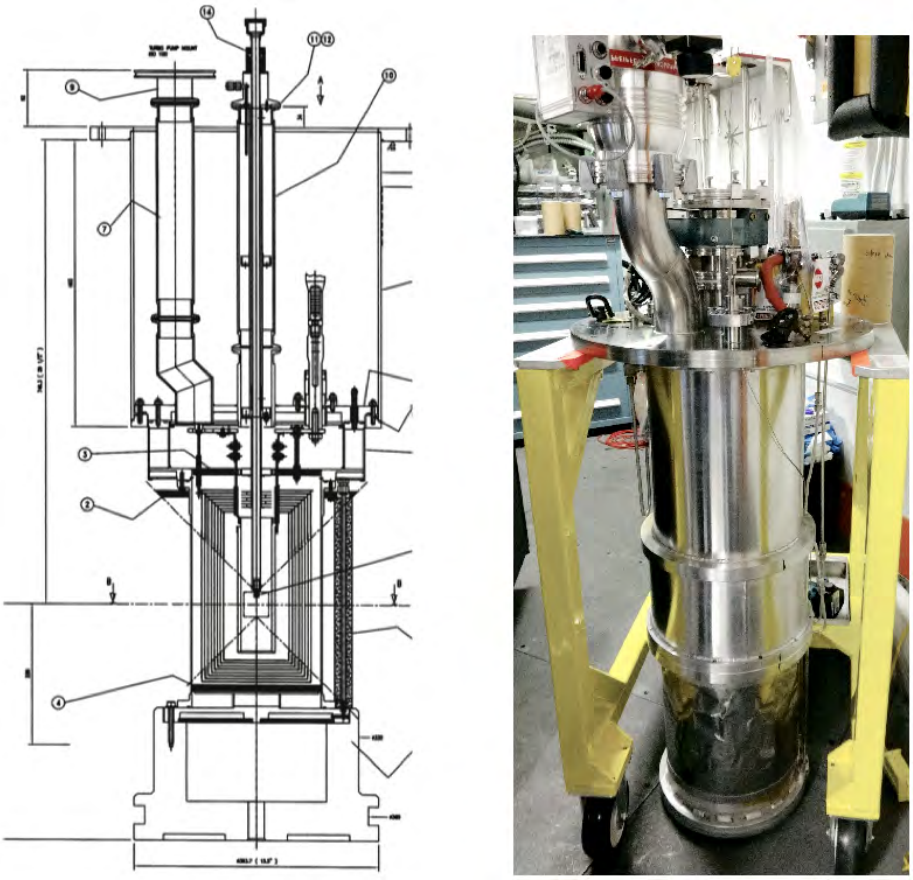


Figure 8.7: Schematic of the MICAS furnace (left) and the furnace itself (right), supported on a yellow cart.

Table 8.1: Outer diameters of the concentric cylinders of heating elements, shielding, and the outer vacuum tank containment for the MICAS furnace. All of these components are made from high-purity Nb foil with a thickness of 0.002 in.

inner heating element	3.13 in.
outer heating element	3.65 in.
heat shield 1	5.10 in.
heat shield 2	5.52 in.
heat shield 3	5.93 in.
heat shield 4	6.35 in.
heat shield 5	6.77 in.
heat shield 6	7.18 in.
heat shield 7	7.60 in.
heat shield 8	8.01 in.
outer tank	11.54 in.

at the center, which is attached to the bottom of a stick and inserted into the furnace. Concentric cylinders around the sample providing heating and thermal radiation shielding. Their dimensions are given in Table 8.1. Figure 8.8 depicts these concentric cylinders. The two inner-most cylinders are the heating element. They are surrounded by between 1 to 8 layers of shielding for thermal radiation that suppress heat loss from the furnace. (Dissipation of heat from the heating element into the sample region of the instrument tank is an issue that must be observed closely, ideally by monitoring the temperature on the outer tank layer. In recent modifications of the MICAS furnace design, this problem is mitigated by water cooling above and below the sample region of the furnace.) It was discovered that the precise diameters of these heating elements and heat shields is crucial for accurately modeling their scattering. The values given in Table 8.1 were obtained from an actual measurement of the heating element and heat shields with a micrometer when the furnace was deconstructed and not in use. Initially, all of the concentric layers were simulated as individual scatterers. This was cumbersome and proved to be unnecessary, however.

Details of Furnace Model

The furnace assembly was modeled as two scatterers: (1) the outer vacuum container, and (2) the two heating elements and eight heat shields. The outer vacuum container is modeled as a hollow cylinder with radius 5.77 in., thickness 0.1 mm, and height 15 in. As is obvious from the photo in Fig. 8.7, this outer layer is not a smooth cylinder because of some mechanical abuse. It also serves as the outer vacuum containment for the furnace, so any pressure imbalance

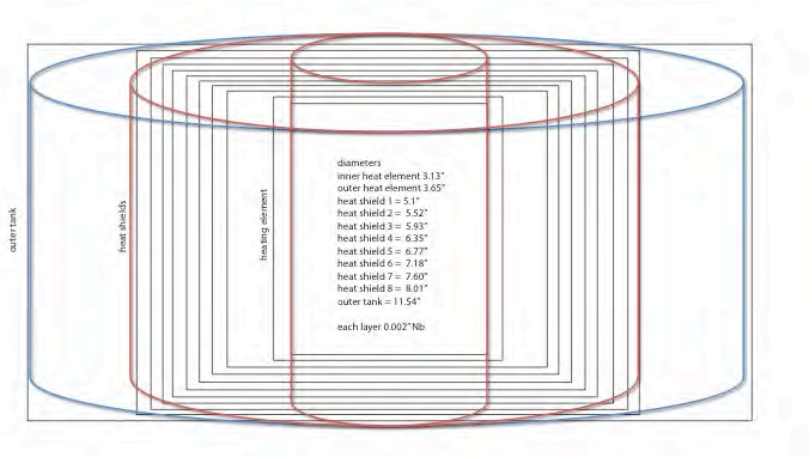


Figure 8.8: Drawing (not to scale) of the heating element and heat shield region of the MICAS furnace. The inner and outer heating elements and the outer tank are fixed. Heat shields can be removed, beginning with the outermost, depending on the maximum temperature of the experiment.

that is created during pump down and venting of the furnace and sample area can cause this thin layer to crinkle. So instead of modeling this region as a perfectly dense cylinder of thickness 0.002 in. (0.05 mm), it is modeled with twice this thickness, and a reduced "packing factor" of 0.5. In Fig. 8.8, it is depicted as the blue cylinder.

The scatterer describing the remaining components also has a hollow cylinder shape with inner radius 1.56 in., outer radius 4 in., and height 15 in. This hollow cylinder has a reduced packing factor of 0.008 to capture the considerable amount of 'empty space' present in this cylinder. In Fig. 8.8, this is depicted by the volume encompassed by the two red cylinders. Figure 8.9 is an example of the furnace template in the sample assembly file. In practice, the sample assembly is an xml file modified by the user to add or remove components of the sample to the simulation. In this example file, the furnace template is in use as indicated by the blue labels. The shape, composition, and geometric conditions described for these components are input variables. A separate "scatterer" file for each of the components provides details including the packing factor, scattering kernels, and scattering probabilities.

Computational Results

Neutron scattering was both measured and computed for an "empty" furnace at room temperature, both with and without the radial collimator. Figure 8.10 compares experimental results (a, b) with simulations (c-f). For these test cases, the empty furnace is not truly empty. The experiment contained an empty Nb

```
<?xml version="1.0"?>

<!DOCTYPE SampleAssembly>

<SampleAssembly
   name="withfurnace"
   max_multiplescattering_loops_among_scatterers="5"
   max_multiplescattering_loops_interactM_path1="1"
   >

<!-- sample -->
  <PowderSample name="sample" type="sample">
    <Shape>
      <block width="6*cm" height="10*cm" thickness="0.2*mm" />
    </Shape>
    <Phase type="crystal">
      <ChemicalFormula>Nb</ChemicalFormula>
      <xyzfile>Nb.xyz</xyzfile>
    </Phase>
  </PowderSample>

<!-- outer most -->
  <PowderSample name="can" type="sample">
    <Shape>
      <hollowCylinder in_radius="5.77*inch" out_radius="5.77*inch+0.1*mm" height="15*inch"/>
    </Shape>
    <Phase type="crystal">
      <ChemicalFormula>Nb</ChemicalFormula>
      <xyzfile>Nb.xyz</xyzfile>
    </Phase>
  </PowderSample>

<!-- Nb heating elements etc. 2+8 -->
  <PowderSample name="heater-shield" type="sample">
    <Shape>
      <hollowCylinder in_radius="1.56*inch" out_radius="4*inch" height="15*inch"/>
    </Shape>
    <Phase type="crystal">
      <ChemicalFormula>Nb</ChemicalFormula>
      <xyzfile>Nb.xyz</xyzfile>
    </Phase>
  </PowderSample>
```

Figure 8.9: The furnace template contributes two components to the sample assembly. The blue labels indicate which component is being described. The "outer most" and "Nb heating elements etc. 2+8" make up the furnace.

sample sachet fixed in a BN absorbing frame at 45 degrees to the incident beam. The simulation was simpler, however, containing two pieces of Nb foil at the sample position, also at 45 degrees to the incident beam.

The experimental data in (a) and (b) are plotted on the same intensity scale, demonstrating the dramatic effect of the radial collimator in reducing unwanted scatter from the sample environment. Prior to the introduction of the radial collimator, the result in (a) shows the considerable background scattering that needed to be removed from the experimental data. Most notable is the high intensity scattering along the elastic line, including a peculiar split in the elastic intensity, especially towards higher Q. Inelastic scattering intensity is also visible, and it is even possible to see some phonon dispersions from Nb.

The broad Bragg diffractions, which split into two at higher Q, originate from two distinct regions of scatterers in the furnace. There is a high concentration of Nb near the sample position, including the Nb foil at the sample position and 10 concentric layers of Nb foil within 4 in. of the beam center. This produces a set of elastic scattering peaks with an intensity somewhat broadened along E. The second set of elastic scattering peaks results from the outermost Nb foil layer, which is at a considerable distance of nearly 6 in. from the sample center. A gap of nearly 2 in. exists between the sample and shielding, and this outer cylinder, causing its own set of diffraction peaks from scattering. In the experimental measurement with the collimator introduced (b), this effect is entirely eliminated. This is as expected because the radial geometry of the collimator is highly effective in removing scattering that occurs at angles not radiating directly from the sample position. The collimator is less effective at low Q, and the lowest-order diffractions from Nb have moderate intensity with the collimator in place.

The simulations in Fig. 8.10 show the furnace without the collimator (c, e) and with the collimator (d, f), and provide a comparison with (e, f) and without (c, d) multiple scattering. The intensity scales are matched in all of the simulated results to demonstrate the high efficiency of the collimator in reducing this unwanted background scattering (so it is easier to analyze the results for the simulation without the collimator). Panel (c) shows the empty furnace with all scattering kernels in use, but without multiple scattering. This reproduces the experiment data in (a) reasonably well, although intensity along the elastic line, especially at low Q, is noticeably absent. This is also evident in comparison of (b) and (d), the experiment with the collimator in place and simulation without multiple scattering. In (e), multiple scattering is introduced, which provides the missing intensity near E=0 and also smears out the scattering along Q. This appears to also be the case in (f), though it is less obvious. Multiple scattering often involves two elastic scattering events, but these tend to be incoherent and contribute intensity between the Bragg peaks from the Nb. A similarly broad elastic line is seen in Fig. 1.1e for multiple scattering from fcc Al, for which the individual scatterings were assumed coherent.

These empty furnace measurements and simulations demonstrate the viability of simulating background intensities in inelastic neutron scattering experiments. Not only are the results interpretable in terms of individual scattering

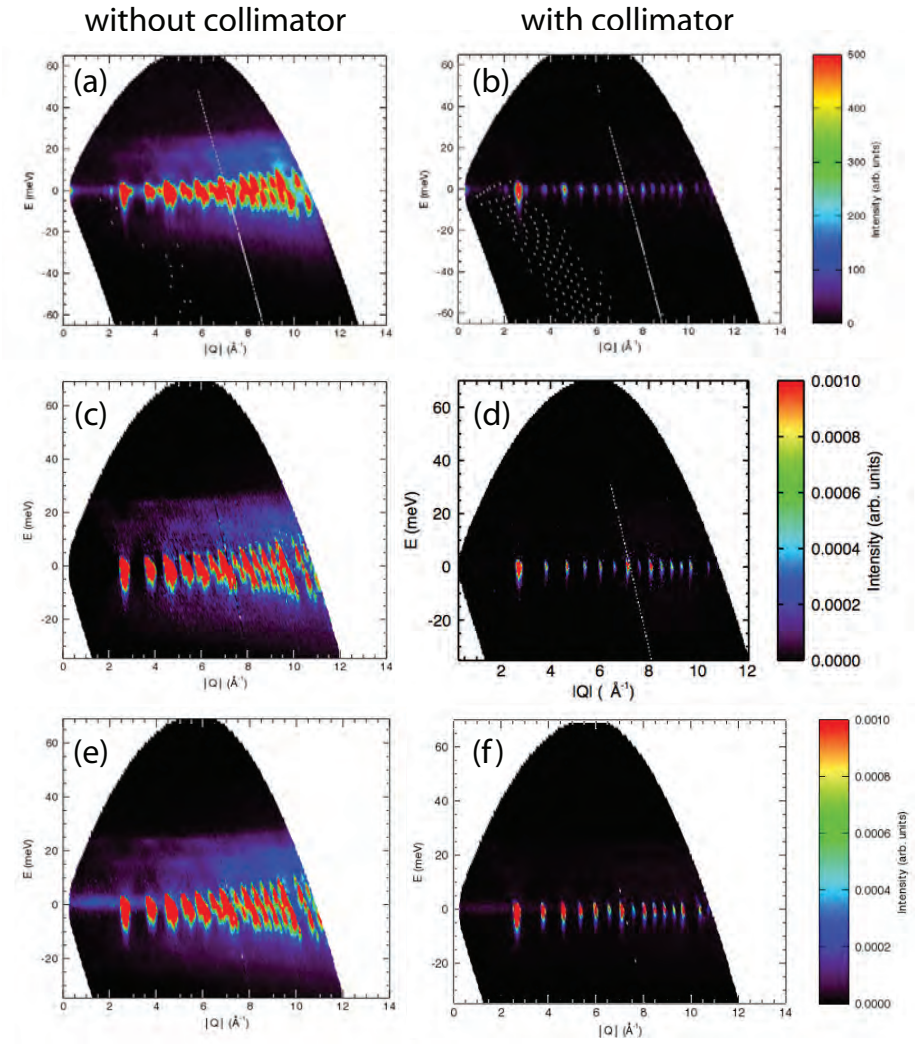


Figure 8.10: The empty furnace provides significant background, as is visible from the experiment without the collimator (a), but the background is considerably reduced with the collimator (b). The simulation without the collimator is shown without multiple scattering (c) and with multiple scattering (e). The simulation with the collimator is also shown without multiple scattering (d) and with multiple scattering (f).

processes, but they are quantitatively accurate. Nevertheless, this approach is not fully proven. It still seems prudent to make experimental measurements of the main contributions to the background, such as the furnace itself, and use the simulations to correct for more difficult effects such as multiple scattering between the sample and the furnace.

8.4 The Future of Computational Scattering Science

The opportunities of computational scattering science, and the effort to achieve them, are not new ideas. The authors had understood many of them around the end of the 1900's. It is somewhat disappointing that they are not more widespread today. Some of this is understandable, since it a project in computational scattering science requires success in both experiment and in computation. Obviously it is easier to address only one of these, and this has been the historical norm. However, "Without deviation from the norm, progress is not possible," as pointed out by Frank Zappa. What follows are a number of observations about the opportunities and the paths needed to achieve them.

8.4.1 Some General Thoughts

Today it requires less effort to use sophisticated tools for computational materials science than it did about 10 years ago. Today a graduate student may routinely use more than a million CPU hours per year, and this will increase steadily. Yet there remains a large learning curve for applying these tools to Scattering Science.

The bigger challenge is to incorporate these computational tools into workflows for doing new science. This is not moving forward very rapidly, and the status quo of an individual scientist/developer is an inefficient cottage industry. The scattering community needs help to make the new computational tools easier to use. There is some need to develop new codes for computational scattering science, but the bigger need is to use the existing codes for calculating the properties of materials, and integrate them into workflows to interpret the results of experimental scattering data. These workflows could be archived and shared, allowing them to be adapted quickly for new scientific endeavors, supporting scattering scientists who need them.

It is of course paramount to quantify the accuracy of quantitative science. When sophisticated features are extracted from a workflow that combines computing and experiment, or when different types of experimental data are combined to develop an underlying materials model, we have little experience with the reliability of the results. Uncertainty quantification needs to be considered when building new scientific workflows. Although underutilized in scattering science, Bayesian methods can be incorporated naturally into efforts that combine computation with experiment, using results from one as prior information for the other, for example. These methods are challenged when the models for obtaining conditional probabilities are not well-known, however.

Nearly all scattering science research with X-ray and neutron facilities involves studies of materials, but the experiments cover an enormous range of different materials and phenomena. Nevertheless, some computational materials science methods are important for broad classes of materials. All atoms obey quantum mechanics, for example. Modern methods for calculating electronic structure, molecular dynamics, and tools for modeling atomic structure and dynamics are well established. Software that simulates materials at the atomic level using quantum mechanics and statistical mechanics has become both commonplace and essential for understanding a large range of phenomena. It is time to develop specific scientific workflows that use these modern tools of computational materials science to assist in interpreting specific scattering experiments.

8.4.2 Some Specific Workflows

Direct experiment simulations have not reached their full potential in 2016. These simulate neutrons sent into a beamline, how the beams are monochromated and focused, and how they are scattered from the sample into detectors. The quirks of the instrument are included in these simulations, and in many cases it is more reliable to compare simulated data to real data, rather than trying to correct experimental data for characteristics of the instrument. Central to these simulation workflows is the scattering from the sample. We know how a neutron is deflected by different interactions with the atoms in the sample, and we now have the tools to calculate the positions of atoms, their vibrational dynamics, their magnetic moments, and how local magnetic moments respond to temperature and magnetic field. These ab-initio codes are a triumph of quantum mechanics in the past two decades. We should be using them to calculate the relevant correlation functions and cross-sections for scattering events, and using these results in the simulations of experimental data.

There is broad agreement that density functional theory (DFT) codes are the standard workhorses for computational materials science today, and there should be no major issues incorporating them into workflows that calculate diffraction patterns and phonon dispersions of regular crystalline material, for example. To date, many studies have already benefited from DFT workflows. Classical molecular dynamics codes could also be made available this way, but using them for accurate simulations of real materials requires expertise in "training" their force fields. This detailed work may be beyond the patience of computational scattering scientists. Although ab-initio molecular dynamics demands much more of the computer, it demands less from the scientist. Finally, there is widespread interest, both in theory and in experiment, in understanding materials when their structures are not in equilibrium. Codes to calculate the structures and dynamics of materials with excited electrons are emerging, and are currently limited to much smaller systems than are typical DFT calculations due to the less favorable scaling with system size of these methods.

Uncertainty quantification is hardly used by the scattering science community today, even though it is almost always desirable to estimate errors

in experimentally-derived quantities. When these quantities are not obtained by direct measurement, but instead the measurements are used to optimize parameters in an underlying atomic model, for example, it is challenging to estimate the uncertainties. There are some families of workflows where classical Bayesian statistics could be adapted to estimate uncertainty. These methods are not implemented in core packages of computational materials science. Software workflows need to give more attention to uncertainty quantification, especially those that combine multiple sets of data and computation. It is unfortunate that there is so little reward for the effort of doing this, but this work is especially important when the models themselves are uncertain or controversial. In such cases the traditional Bayesian approach may not be appropriate, but there are new efforts in the applied mathematics of uncertainty quantification that could help over the next years.

Adapting computational tools to new scientific workflows is most efficient when the software packages are modular, and have consistent interfaces for I/O, for example. The rules of good object-oriented programming go a long way towards satisfying this goal, although compatibility must be designed. Also necessary will be workflow documentation, discussion forum management, and some brokerage of computing resources. Software development with professionalism, and with steady input from the scattering community, is essential for building useful and sustainable software for computational scattering science. We are not doing this so well today, but the promise is there. It seems too obvious to go unrecognized for much longer.

Further Reading

The contents of the following are described in the Bibliography.

H. A. Abelson and G. J. Sussman: *Structure and Interpretation of Computer Programs* (MIT Press, Cambridge Mass, 2001).

Mark Lutz and David Ascher: *Learning Python* (O'Reilly & Associates, Inc. 1999).

Appendix A

Appendix 1: Software Reference

A.1 `reduction`

DANSE `Reduction software` is a set of independent Python packages for reducing neutron scattering data. Those components can be used as building blocks of a new reduction application as explained in Section A.1.10. We define the terms:

- `reduction application`... a complete software solution for reducing experimental data from a particular instrument from archival disk file to graphs of intensity in a space of E, Q, or both,
- `reduction software`... a collection of packages for the purpose of reducing data,
- `reduction package`... one specific package in the `reduction software`.

A.1.1 Introduction

Reduction is a procedure to transform measured raw data to a form that is more easily understandable to scientists. The procedure usually consists of transformations[1] of a histogram measured in a dimension to another dimension that is more physically meaningful (an example would be converting time-of-flight to neutron energy) and/or conversion of multidimensional data to lower dimension (an example is the rebinning of intensity measured in an area detector into "rings" in Q for a powder sample). A reduction procedure consists mainly of transformations from input histograms to output histograms. Our fundamental data structure is a **histogram**. Section A.1.2 has more details about the `histogram` package.

[1] For inelastic neutron scattering, such transformations are described in *Ch. 6*.

In reduction, transformations from input histograms to output histograms require knowledge of the instrument used for the measurements. For example, in direct-geometry time-of-flight spectrometer, to convert time of flight to neutron energy, we need to know the length of the flight path, and the time when the neutron pulse hits the sample. Precise positions of the sample and detectors are other necessary inputs. We therefore need a way to store information about the instrument. The `instrument` package deals with this problem, and details of that package are found in Section A.1.3.

Another major piece of our reduction software is the `measurement` package. The `measurement` package is a layer of classes between the input data files and the reduction engine. Section A.1.4 has more details.

Finally, the transformations from input histograms to output histograms are implemented in the `reduction` package, which queries the `measurement` package for input histograms, calls reduction engines with parameters derived from the `instrument` package, and produces output histograms.

We go deeper into each package in Sections A.1.2, A.1.3, A.1.4, and A.1.5. Some description of how they interact with each other is given in Section A.1.6. In Section A.1.8, details about reduction components are explained. Developers who want to create a reduction application using the available building blocks should read on to Section A.1.10.

A.1.2 Histogram

We need to define clearly what do we mean by "histogram". The result of any measurement is actually a histogram, by which we mean we have data in some bins. For example, if we measure a spectrum with x-axis being time-of-flight, we will get an array of counts, while each element in that array represents the number of counts measured in a predefined time slot (bin). This array of counts can be approximated by

$$\frac{dI}{dx}(x)\Delta x \tag{A.1}$$

where $\frac{dI}{dx}$ is a density function and Δx is bin size. This observation forms the base of our design of histogram classes. There are two pieces of critical information in a histogram: the data, and the axis (or axes). Furthermore, we usually need to know the context of the histogram, and that requires "meta-data."

Here are more rigorous definitions:

- dataset... a dense array of numbers which may have many dimensions. Counts and errors are datasets.
- axis... a one-dimensional dataset whose elements are the bin boundaries of one dimension of a histogram.
- meta-data... data that provide context for other data. Meta-data are data about data.

A histogram is a set of related axes, datasets, and meta-data. A histogram contains (1) at least one dataset whose elements represent the number of counts in some range of axis or axes values; (2) optionally one dataset for error bars [2] (3) a set of associations about a histogram and potentially everything that can be known about it: axes, history, etc. (meta-data).

Design

The design of the histogram package is not too complex. For the numerical operations that are methods of the histogram class at the bottom of Fig. A.1 (e.g., +, − , *, /), we need a way to deal with representations of arrays in low-level language, and an abstract interface NdArray, handles that. By introducing the NdArray layer, we isolate the histogram from any particular C/C++ array implementations. We also need a way to keep meta-data, and this is handled by AttributeContBase class. Figure A.1 shows the class diagram of the histogram package.

Usage

As a fundamental data object, histogram needs functionalities of several categories:

- creation... create histograms from arrays, functions
- slicing... get slice; set slice
- numerical operators... +, −, ×, ÷ normal version and in-place version
- other numerical operations... sum over an axis (projection), transpose
- miscellaneous... retrieve information about an axis, retrive data arrays

Please read `http://drchops.caltech.edu/Docs/docs/reduction/Software-UserGuide/html/index.html` Reduction Software User's guide, Part II for more details.

A.1.3 Instrument

Classes in the `instrument` package are used to describe an instrument, independently of any particular measurement. The instrument package is used in several ways, such as for parsing files derived from engineering drawings of an instrument (as in the `reduction` package), creating a 3d visualization of an instrument, or perhaps in setting up a Monte Carlo simulation.

An instrument consists of several instrument elements in a hierarchial structure. All instrument elements are in the subpackage instrument.elements. In our design, all instrument elements are derived from the class instrument.elements.Element.Element.

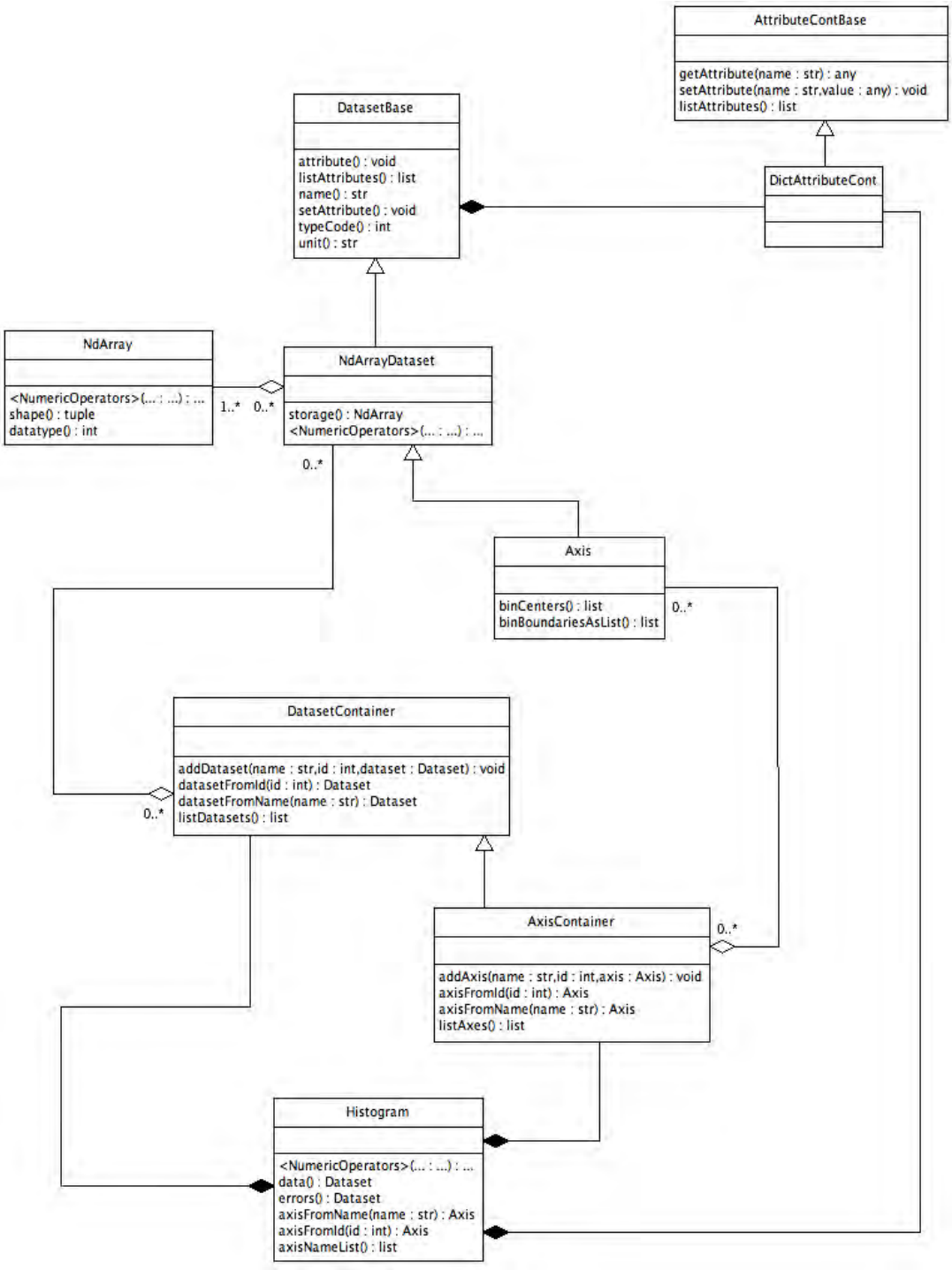


Figure A.1: Class diagram of the histogram package

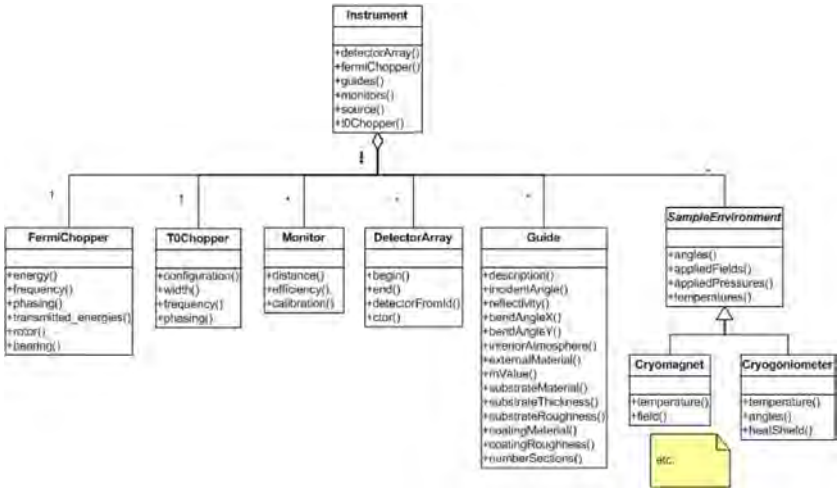


Figure A.2: Hierarchy of instrument elements in a typical instrument

An instrument object is built out of these elements, and follows the real instrument as closely as possible. A diagram of an example instrument is shown in Figure A.2.

Currently this package is mostly concerned with inelastic direct geometry time-of-flight instruments. To support more instruments, we just need to identify those elements that are not yet supported, and add them to the element library.

Geometers

Geometers are responsible for measuring distances, angles, and other geometric quantities. Each instrument should have an Geometer (or several geometers) associated with it. There are some Geometers in the subpackage instrument.geometers.

Instrument Factories

An instrument factory method is needed for each supported instrument. An instrument factory method is a method that creates a hierarchial representation (data object) of an instrument. It contains information about every neutron components in the instrument, including but not limit to a moderator, monitors, a sample, and a detector system.

Three instruments are now supported: LRMECS, Pharos, and ARCS. Factory methods are created for those instruments and a factory method can create a representation of a particular instrument by reading a text configuration file

[2] actually the squares of error bars are stored to improve computation efficiency

(for older instruments not using NeXus) or a NeXus file (for ARCS).[3] More information can be found here:

`http://wiki.cacr.caltech.edu/danse/index.php/Instrument_and_related_classes` wiki page

A.1.4 Measurement

Another important piece of the reduction software is the `measurement` package. It is a layer of classes between input data files and the reduction engine. It converts input data files into input histograms.

In reality, a complete measurement usually consists of several runs:

- a run for the sample you are interested in
- a run for the empty sample can
- a run to calibrate detector efficiency

Each run results in one or more histograms (one main histogram, plus other accompanying histograms, like beam monitor histogram).

The structure of the `measurement` package is simple. It consists of measurment classes for every instrument we are supporting, and their common base class. Each instance of a measurement class is a container of "run"s. Each "run" has methods to extract histograms from input data files.

A.1.5 Reduction Package

The `reduction` package is inside the `reduction software`. The `reduction` package is carefully separated into several layers, giving a clean design.

- The "C/C++" layer is responsible for intensive computations that are only practical in a low-level language. For example, this layer includes a class ERebinAllInOne to rebin data in time-of-flight bins into data in evenly-spaced energy bins.
- The "Python vector compatible" layer vectorCompat is the interface between C++ and Python. All C++ codes are implemented to deal with "vector"-like objects, e.g., energy bins. The vectorCompat Python package accepts vector arguments, and calls the corresponding C++ methods to do the real work. This layer separates other Python layers from C++ codes and Python bindings. It forms a bridge between the reduction operators in the histogram-compatible layer and the low-level implementations of reduction operators.
- The "Python histogram compatible" layer histCompat allows developers to work with objects having more physical meaning. This layer is built

[3] Since the NeXus format for ARCS is not yet settled, the present file format for ARCS may change.

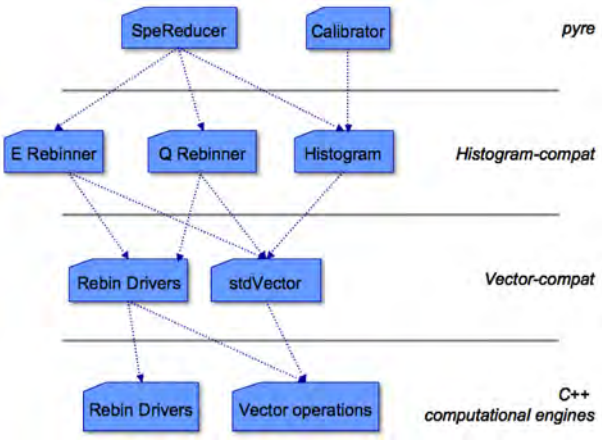


Figure A.3: Layers in reduction package.

on top of the vectorCompat layer. In the histCompat layer, histograms are the focus. Classes in this layer take histograms instead of vectors as arguments, and implementations of these classes decompose histograms into vectors, and call the corresponding methods in the vectorCompat layer.

- The "Pyre vector compatible" layer pyreVC wraps classes in vectorCompat layer as pyre components. There should not be many of these components because we should be dealing with histograms more in the pyre layer
- The "pyre histogram compatible" layer pyreHC wraps classes in histCompat layer to pyre components.
- The "pyre" layer reduction.pyre makes use of pyre components in the pyreVC and pyreHC layers, and implements higher-level classes. The pyreVC and pyreHC layers are more concerned with lower-level operations like "rebin to evenly-spaced energy bins" and "fit a curve to a Gaussian, and find its center". The pyre layer is more concerned with "calculate calibration constants from calibration data" and "reduce I(det, pix, tof) to S(phi,E)."

Figure A.3 is a simplified cartoone of this layered structure. This figure is far from complete, however, and shows only part of the whole structure.

A.1.6 Package Design

As mentioned in Section A.1.1, several packages are used in the reduction software. Four are "central" packages:

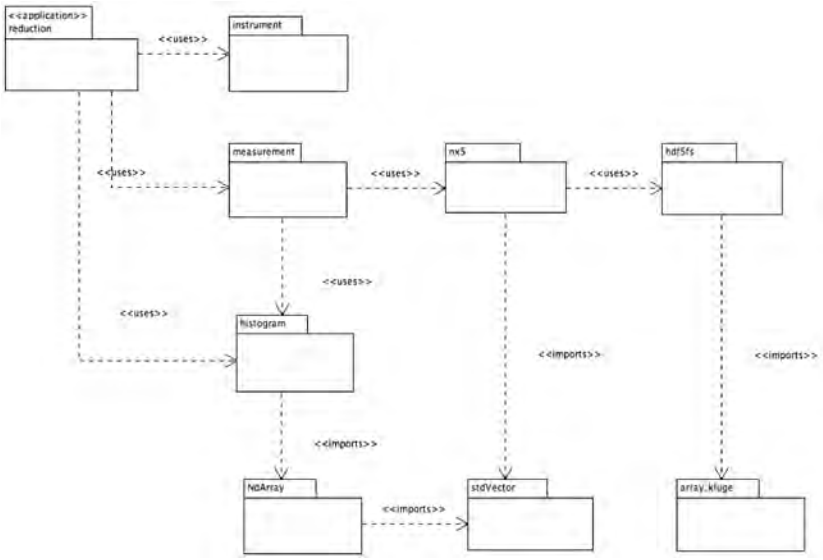


Figure A.4: Reduction software package diagram

- histogram: is the fundamental data structure
- instrument: contains and uses instrument information
- measurement: creates histograms
- reduction: contains the reduction methods (functions)

There are other "supporting" packages:

- nx5: implements "NeXus" readers/(writers) based on hdf5fs
- hdf5fs: is a tool to create/read/write hdf5 files by treating an hdf5 file as a filesystem with trees of directories and files
- stdVector: is a Python binding to C++ std::vector template
- array_kluge: contains tools to manipulate C arrays

Relationship among those packages are shown in Figure A.4.

A.1.7 Miscellaneous Design issues

Error propagation

For all numerical operations (e.g., + − * /) on data histograms, error propagation is performed on the error histogram. Reduction components always propagate errors.

IMPORTANT POINT: The error arrays in histograms are always assumed to contain the squares of errors. The reason is that constantly squaring and square-rooting data is redundant, slow, and diminishes numerical precision. It is up to the user (or programmers) to take square roots as appropriate (for instance, when plotting an intermediate result, or writing a final result to disk).

A.1.8 Reduction Components

This section presents details of those reduction components that tie closely into physics of reduction procedures. Chapter 7, especially Section 7.1, gives important background material for this section.

IncidentEnergySolver

IncidentEnergySolver computes the incident energy of the neutron pulses on the sample for an experimental run. The easiest way to get the incident energy is with the data from the monitors.

$$v = \frac{d_2 - d_1}{tof_2 - tof_1} \tag{A.2}$$

where d_1, tof_1 and d_2, tof_2 are distances and times-of-flight from moderator to monitor 1 and monitor 2. Neutron energy can be calculated from velocity as $E = 1/2\, m_n\, v^2$.

Another way to compute the neutron incident energy is to use the elastic peaks in the measured data. Since neutrons in elastic peaks are scattered elastically by sample, their velocities after scattering are the same as the incident velocity. What we need to do is to find out the time-of-flight of those neutrons, and the distance of travel, and then we can compute incident energy. The difficulty of this method is, for one pixel, the I(tof) spectrum usually does not have good enough counting statistics to identify clearly the elastic peak. One way to solve this problem takes advantage of the fact that in many direct-geometry TOF instruments, detector tubes are arranged as a cylinder around the sample. Most detectors have the same distance from the sample position. We can add together the I(tof) spectra of pixels at the same height for all detectors that have the same distance to the sample position to increate statistics, and also reduce the number of required fittings.[4]

Calibrator and VDataProcessor

It is good practice to run a vanadium calibration sample before making the critical science measurements. This vanadium calibration run allows for correcting for the variation in efficiencies of the detectors so their count rates can be properly normalized. Two important components are:

[4]The implementation of these procedures is called the "Max Incident Energy Solver," in honor of its developer, Max Kresch.

- Calibrator, which is used to calibrate the raw data by given calibration constants.
- VDataProcessor, which is used to compute calibration constants.

Here we use a "white-beam" calibration as an example to explain the procedure.[5] Given a white-beam calibration dataset, the $I(E)$ spectra are first calculated for each detector. We then pick an energy window of interest from $E_{\min} to E_{\max}$ and compute net intensity in that window:

$$cc(\text{det}) = \int_{E_{\min}}^{E_{\max}} I(E)\,\mathrm{d}E \tag{A.3}$$

for each detector. This is a preliminary calibration constant, $cc(\text{det})$.

Next, we generally need to correct the detector efficiency for variations in the ^{3}He gas pressure in different tubes, using an efficiency effect due to pressure $eff(pressure(det))$

$$cc1(det) = \frac{cc(det)}{eff(pressure(det))} \tag{A.4}$$

If the vanadium sample is a plate, we have to correct for the anisotropic self-absorption of the vanadium sample[6]

$$cc2(det) = \frac{cc1(det)}{absorption(angle(det))} \tag{A.5}$$

The calibration constant, $cc2$, is now applied to the experimental data:

$$I_{\text{corrected}}(det) = \frac{I(det)}{cc2(det)} \tag{A.6}$$

Normalizer

Raw datasets need to be normalized by the neutron flux so that different experimental runs can be compared. This is usually done by dividing the measurement counts by the intensity measured with a monitor. In Pharos, there may be no working monitor, so all monitor counts are zero. For Pharos we use the integrated-moderator-current as the normalization constant.

TimeIndependentBackgroundRemover

Section 7.1.2 gives important information about TimeIndependentBackgroundRemover and its sub-components, which remove a time-independent background from the raw data. A user of TimeIndependentBackgroundRemover must pick a

[5] A white-beam run is performed with a wide spectrum of incident neutron energies.

[6] Few neutrons scattered into the plane of the sample will leave the sample and reach the detectors.

window in the time-of-flight axis that is expected to be a constant dominated by noise that is independent of time. TimeIndependentBackgroundRemover will calculate the average counts in those tof channels, and subtract this number from each channel in the full histogram of raw data.

Idpt2Spe

This component reduces $I(det, pix, tof)$ to $S(\phi, E)$. It needs user inputs for the axes ϕ (scattering angle) and E (energy transfer).

Spe2Sqe

This component converts $S(\phi, E)$ to $S(Q, E)$. It needs user input for the axis Q (momentum transfer).

A.1.9 Reduction Applications

Reduction components have been assembled into full reduction applications that are distributed to users as reduction software for particular instruments. User manuals are available online:

`http://drchops.caltech.edu/Docs/docs/reduction/Software-UserGuide/html/index.html`

The user manual is perhaps the best place to get started with the reduction software.

Developer manuals provide additional information, and are available online:

`http://drchops.caltech.edu/Docs/docs/reduction/DeveloperGuide/html/index.html`

, in which you can also find software design information.

A.1.10 How to Create a Reduction Application

This tutorial explains how to create a reduction application using buiding blocks from the DANSE reduction packages and the DANSE reduction framework.[7] A powder reduction application for a direct-geometry time-of-flight spectrometer is used as the example.

Instrument representation

We consider reduction as a transformation between an input-histogram and an output-histogram. Instrumentation information are needed to do those transformations, so the first step is to create a representation of the instrument that was used to acquire the data. An instrument representation should have a hierarchial structure that mimics the structure of the real instrument. It can be done by using instrument elments available in the Python package

[7] We expect that this approach will be important for developers and advanced users. Others may choose to defer reading this section for more detailed reading.

instrument.elements. If an instrument element does not exist in that package, however, a new element will need to be developed. For this tutorial, here we create an instrument from existing instrument elements.

```
>>> from instrument.elements import *
>>> instrument = Instrument.Instrument( 'instr_name', version = '0.1' )
>>> moderator = Moderator.Moderator( instrument.getUniqueID(),
                                     instrument.guid(),
                                     100, 100, 100 )
>>> instrument.addModerator( moderator )
>>> monitor = Monitor( instrument.getUniqueID(), instrument.guid(),
                       xLength=30, yLength=0., zLength=50,
                       monitorNumber=1,
                       name = "monitor1")
>>> instrument.addMonitor( monitor, 1 )
```

There are some items to note in the instrument above:

- The first step is to create an instrument by calling the Instrument constructor.
- A instrument-wide unique identifier is needed for any instrument element except the instrument itself. This unique identifier is generated by calling the method of instrument, "getUniqueID".
- All dimensions are in units of mm.

In this simple example, we did not introduce the geometer. A geometer is always needed before an instrument becomes useful. Positions and orientations of any instrument element must be registered by the instrument's geometer. Here is a more complete example of an instrument, this time with a geometer.

```
>>> from instrument.elements import *
>>> instrument = Instrument.Instrument( "instr_name", version = "0.1" )
>>> from instrument.geometers.ARCSGeometer import Geometer
>>> geometer = Geometer()
>>> moderator = Moderator.Moderator( instrument.getUniqueID(),
                                     instrument.guid(),
                                     100, 100, 100 )
>>> instrument.addModerator( moderator )
>>> position = [20000.0, 90.0, 180.0]
>>> orientation = [0.0, 0.0, 0.0]
>>> geometer.register( moderator, position, orientation)
>>> monitor = Monitor( instrument.getUniqueID(), instrument.guid(),
                       xLength=30, yLength=0., zLength=50,
                       monitorNumber=1,
                       name = "monitor1")
>>> position = [2300.0, 90.0, 180.0]
>>> orientation = [0.0, 0.0, 0.0]
>>> geometer.register( moderator, position, orientation)
```

Things to note here:

- The coordinate system used by ARCSGeometer is a spherical one centered at the sample position. Positions are specified by a 3-tuple (ρ, θ, ϕ).

You might want to implement your own geometer if the ARCSGeometer does not fit your needs. Examples of instrument factories and geometers can be found in The Python packages instrument.factories and instrument.geometers.

Histogram

This section identifies histogram classes that are specific for reduction, and illustrates their use. For direct-geometry TOF spectrometer, these histograms are used, with typical inputs and outputs:

- input histograms
  - I(det, pix, tof)
  - I(det, tof)
  - I(tof) for monitors
- output histograms
  - S(phi,E)
  - S(Q,E)

To implement those histogram classes, we simply inherit a new class from the Histogram base class in the Python package `histogram`, and implement a new interface that is more convenient. Recall that the package `histogram` (Sect. A.1.2) offers a handy set of methods for numerical and boolean operations on histograms, and handles error propagation (Sec. A.1.7).

Measurement

This section discusses how to use measurement classes and measurement pyre components. Section A.1.4 explained that a measurement usually consists of several "run"s. Let's implement a class "PharosRun". The "PharosRun" class is a subclass of the measurement.Run.Run class. There are several methods to extract input histograms (Section A.1.10) from data files. The following interface for PharosRun is appropriate:

- ctor constructor is used to take the data file name as input
- getDetPixTOFData is used to extract I(det,pix,tof) histogram from a data file
- getDetTOFData is used to extract I(det,tof) histogram from a data file
- getMonitorData is used to extract monitor I(tof) histogram from a data file

Now we implement the PharosMeasurement class. It will be a subclass of measurement.Measurement.Measurement. PharosMeasurement should be a container of several runs that are instances of PharosRun. These runs are:

- main data for the sample of interest
- calib data for calibration sample of vanadium
- mt data for empty sample can

The last step is to create a pyre component for the PharosMeasurement class. This pyre component will be very thin because most of the real work is done in the class PharosMeasurement.

Transformation and rebinning

In this section we implement the transformation in low-level language and export those transformations up to Python and pyre layer.

Pyre components

In this section we carefully separate the reduction procedure into small steps and implement pyre components for each step.

Pyre application

In this step we combin pyre components constructed in the previous section and create a reduction pyre application.

User Interface

In this step we use and customize the univeral pyre GUI (prototype) to create GUI for the pyre application created in the previous section

A.1.11 Reduction Software Implementation and Distribution

Status

As of November, 2006, `reduction` has four essential layers: pure C++, and Python-C++ integration (written in C++), pure Python, and pyre. In `reduction`'s precursor, `topline`, all scientific functionality resided in the C++ layer; Python bindings (exposing a few specialized pieces of C++ code) only made it easier to run. `reduction` blurs that relationship: now Python is the "primary" layer, with most of what happens in C++ accessible to Python and a number of high-level services, such as rebin drivers, implemented purely in Python. For simplicity, scientific content is forbidden in the integration layer (Hopefully, this has been enforced). The growth in the code base from `topline` to `reduction`-1.0 reflects these changes: `topline` had about 2000 lines of C++ in a dozen or so classes;

about 900 lines of C++ code provided Python bindings to six entry points for high-level routines. Version 1.0 of `reduction` has about 1500 lines of C++ code, and about 2100 lines of Python bindings providing roughly 60 entry points for much finer-grained access to the C++ layer. The `topline` Python layer was an after thought, weighing in with 100 or so lines of Python; `reduction`'s pure Python layer has over 21000 lines.

Building and Installing

Please download the ARCS reduction source from links reached from:

`http://drchops.caltech.edu/`

and follow the online instructions to build and install.

Performance

Under linux (gcc 4.0, etc.) a 2.0 GHz AMD Opteron box with 4 GB RAM can reduce a dataset to $S(Q, E)$ in about 27 sec—about how long it took to do the same task under these conditions with the previous C++ implementation, `topline`. On a Windows XP box (using Microsoft Visual C++ v7.1), it takes about 60 sec; this is a disappointment, as the `topline` code executed in about 6 sec. Some of the Windows bottleneck seems to be caused by memory swapping. Other factors might include file I/O of hdf5 file, and expensive looping in python layer.

To Do

- `instrument`: More instrument elements should be added to instrument.elements package. Geometers might deserve better implementation. Currently the way geometer deals with coordination system is a little clumsy.
- `reduction`:

  1. The energy rebin driver in the C++ layer has a generic alogrithm to rebin 1-D data, but the implementation is not generic enough. A more generic rebinner could be reused for many other rebin tasks.
  2. The reduction package makes explicit uses of STL vectors. Many of these uses can be replaced by iterators, which would broaden their reusability while insulating them from changes to the underlying containers. For instance, rebinning procedures should be willing and able to operate on any contiguous chunk of data, whether it belongs to an STL vector, an ordinary C-array, NumArray, etc.
  3. We could incorporate Jae Dong Lee's single crystal codes, which implement the methods of Mslice in C++. Alternatively, we could develop a single crystal reduction package that uses a UB matrix formulation.

A.2 Module Documentation

This section provides web access to the online ARCS software documentation. You will need to have an internet connection to make proper use of the present section.

The following include links to the developer's API-level socumentation for all available ARCS 1.0-release packages. You may need to alter your Acrobat preferences to open a web browser or capture the web pages in Acrobat, as you prefer.

As a test of your connection and browser interface, please check some of the top pages of the ARCS and DANSE resources:

DANSE homepage

http://wiki.cacr.caltech.edu/danse/

ARCS homepage

http://www.cacr.caltech.edu/projects/ARCS/

PYRE homepage

http://www.cacr.caltech.edu/projects/pyre/

The top-level web page for the DrChops software, including installation and release information, is:

http://drchops.caltech.edu

The top-level documentation index for the reduction software is:

http://drchops.caltech.edu/Docs

The top-level web page for all software module documentation can be accessed at:

http://docs.danse.us/DrChoops/current/Modules.html

A.2.1 histogram

http://wiki.cacr.caltech.edu/danse/index.php/ARCS_Alpha_subpackage_histogram

The histogram package supplies classes that combine arrays (storage) with attributes to model various complex ideas about datasets for neutron scattering.

http://docs.danse.us/DrChops/current/histogram/histogram/html/index.html

histogram python package

A.2.2 measurement

http://wiki.cacr.caltech.edu/danse/index.php/measurement

The measurement package combines nx5 and histogram to provide a convenient layer for applications programmers through functions such as "getPixelData", "getMonitorData", etc.

http://docs.danse.us/DrChops/current/measurement/measurement/html/index.html

measurement python package

A.2.3 instrument

http://wiki.cacr.caltech.edu/danse/index.php/ARCS_Alpha_subpackage_instrument

The instrument package consists of classes to represent neutron scattering instrument information. An instrument is represented by an instance of the Instrument class, which is the root node of a graph of objects (Elements). The coordinates of the elements of an instrument are maintained by a separate class, Geometer. This creates a generic way to represent and serve information about instrument configuration and properties and sample configuration and properties. Factories construct core representations from both hard-coded sources (such as ARCSBootstrap) and from nx5 file graphs (NX5ToARCS family); additional factories render instrument graphs to file graphs (ARCSToNX5 family) or to visual models (VTKInstrumentViewer).

http://docs.danse.us/DrChops/current/instrument/instrument/html/index.html

instrument python package

A.2.4 Multiphonon

http://wiki.cacr.caltech.edu/danse/index.php/Multiphonon_py

Multiphonon performs a multiphonon expansion to simulate experimental inelastic scattering spectra. Input to the program is a disk file of a phonon DOS in terms of energy, intensity, and error, plus user input on the temperature, atom mass, energy of the incident neutron, request for analysis of data with constant-Q (must supply Q) or TOF angle bank (must supply angle, ϕ). Multiphonon calculates terms in the multiphonon expansion as explained in Section 7.5, using the theory of Section 4.2.2, and sums spectral contributions from the user-specified multiphonon processes. The names of the variables in the code are matched to those of the theory, as developed by Varley Sears and used in Section 7.5. An iterative procedure for performing this correction is explained in Section 7.5.1.

Further Reading

Documentation

The DRCS (DrChops) software release site is here:

`http://drchops.caltech.edu/`

The User's Guide for reduction software is here:

`http://drchops.caltech.edu/Docs/docs/reduction/UserGuide/html/index.html`

The Developer's Guide for reduction software is here:

`http://drchops.caltech.edu/Docs/docs/reduction/DeveloperGuide/html/index.html`

The Doxygen documentation for reduction software, especially useful for developers, is here:

`http://drchops.caltech.edu/docs/reduction/reduction/html/`

Other References

The contents of the following are described in the Bibliography.

Tim Kelley, Mike McKerns, Jiao Lin, Michael Aivazis and Brent Fultz: *DANSE wiki web site*,

`http://wiki.cacr.caltech.edu/danse/index.php/Main\_Page`

Tim Kelley, Mike McKerns, Jiao Lin, Michael Aivazis and Brent Fultz: *ARCS software web site*,

`http://www.cacr.caltech.edu/projects/ARCS/Software.html`

Mark Lutz and David Ascher: *Learning Python* (O'Reilly & Associates, Inc. 1999).

C. G. Windsor: *Pulsed Neutron Scattering*, (Taylor and Francis, London 1981).

Appendix A

Appendix 2: Software Design

A.1 Extending *DANSE*: Writing C++ Extensions to Python

Important sources of information on this topic are available elsewhere: the Python extension and embedding manual, and the Python-C API reference manual. These documents are available at the Python website: `http://python.org`, but better yet, just look in the 'ext' and 'api' subdirectories of your Python distribution's Doc directory for HTML versions.

This section was written to supplement those documents, first by giving an example of dynamically allocating C++ objects and keeping track of them, and second by introducing some of the API functions for working with aggregate types like tuples and lists, for which the easiest Python-C conversion tools don't work. Also, I've concocted some simple, try-this-at-home examples to illustrate the process. Finally, I've added a more real-life example.

Note: In the following, C++ is used to mean both C and C++. The Python-C API is written in C, but for the most part, but that can be called seamlessly from C++. There's only one item that must have C linkage, the init function, described below.

A.1.1 Overview

Why write C++ extensions to Python? To reuse existing code, and to gain better performance. A great deal of software has already been written in C and C++ (not to mention FORTRAN), and, at least at the present, nothing beats compiled languages for performance in the numerically intensive codes that *DANSE* supports. Extending Python allows us to turn all that code into building blocks for solutions. Libraries of extensions package those building blocks into kits that users can adapt to solve their problems.

Of course you can do all that without Python, so why use it at all? Python has an easier learning curve than C++, and it's more flexible and immediate,

making it available to a wider set of users. Once a library of extensions is available on a platform, users can call out to that library, mixing and matching components and testing combinations, without the overhead of compiling and linking. (This immediacy should not be underestimated.) And with all its standard library packages, Python can be manipulated to do some pretty amazing things that would be more difficult to realize in C++, like parsing XML documents, setting up computations, etc.

Writing C++ extensions for Python can be learned in an afternoon, especially if there are some examples to follow.[1] The idea is this: you have a C++ class or function, and you'd like to make it callable from the Python interpreter. You'll need to (I) write some wrapper code in C++, and (II) compile it into a C++ library that the Python interpreter can dynamically load and use. You then typically (III.) write a Python module that mediates between the Python user and this library. This makes life plush for those users who don't want to be bothered with the details of finding out what's in the library and how to call it. More importantly, it gives us the chance to check inputs as soon as possible for errors. When a bad pointer is sent to the C++ level, the results are disastrous (a core dump on a good day), so our software can be made more robust if we handle the pointers ourselves.

It takes more care to create a library that is complete: catches exceptions, checks that preconditions and postconditions have been met, and so on. These are more advanced topics, covered in other parts of the book (*SOMEDAY NAME A FEW??*). Low-level *DANSE* programmers will be expected to incorporate these techniques into their code, but first things first.

A.1.2 A Little More Detail

Here are those three easy steps again, in slightly more detail:

- Write the bindings. There are three essential components:

    1. Wrapper function(s). (One for each function you want callable from Python). The wrapper typically calls a function or a class method, or it creates a heap object. Once you learn how to write one wrapper, you know how to do it, because all wrappers do the same three things.
    2. Method table. (One entry in the table for each function you want callable from Python). The method table tells the Python interpreter which C++ functions it can call from the library.

[1] If it's so easy, why hasn't someone written a program to write the wrappers automatically? They have! Packages like the Simple Wrapper Interface Generator (SWIG, `http://www.swig.org`) will do nearly all of the work for you. There are arguments for and against automatic code-generators like SWIG, and I've worked both ways. At the moment, writing the bindings is such an easy task, and I do so relatively little of it, that I prefer to do it myself. Others no doubt feel differently, and I have no interest in changing their minds. But even if you're going to use SWIG, or a similar library, there's merit in putting in some time writing your own wrappers to learn how and why things get done. Then you can judge well for yourself which approach suits your situation.

3. init function. (One per module) This is the interpreter's entry point into the C++ library.

These three steps are accomplished with generous aid from the Python-C extension API.

- Compile it. This is slightly platform dependent (Michael Aivazis's system for processing source code removes this platform dependence for UNIX flavors, including cygwin; Windows is in the works). The goal is to compile into a shared object library (unix) or a dynamically linked library (the beloved Windows dll).
- Call it from Python. One typically writes a Python module that acts as a layer between the user and the C++ library. By doing things like providing Python classes that mirror the C++ classes, one can make the experience quite similar. Or dissimilar. The choice is yours.

A.1.3 A Lot More Detail: Wrappers

Every wrapper function does three things:

a) Converts a Python object with the arguments to the C++ function into C++ objects,

b) calls the C++ function,

c) converts the output to a Python object with the result and return it.

Let's first run through these steps with numbers and strings, for which there's an immediate connection between Python and C++ types; we can use a function, `PyArg_ParseTuple()`, which is built in to the API. Later examples look at tasks like working with C++ class instances and using aggregate Python types, such as dictionaries and lists.

Simple Example: PyArg_ParseTuple, Py_BuildValue
a) Convert arguments from Python to C++

To convert the Python arguments into C++ objects, first define variables of the appropriate type, one for each C++ argument. Then pass the addresses of these variables into `PyArg_ParseTuple()`. This function takes the args tuple, pulls PyObject's out of it, and converts them to C++ types according to a format string.

Here are some examples:

```
//One integer:
int a;
int ok = PyArg_ParseTuple(args, "i",&a);
if(!ok) return o;
//Two integers
```

```
int a, b;
int ok = PyArg_ParseTuple(args, "i",&a, &b);
if(!ok) return 0;
//One integer, a string, two doubles
int anint;
char * astring;
double dub1, dub2;
int ok = PyArg_ParseTuple(args, "isdd", &anint, &astring,
                          &dub1, &dub2);
if(!ok) return 0;
```

A complete list of what can go into the format string is given in the extension documentation (look in the ext subdirectory of the doc directory in your Python distribution, or look online at `http://python.org/doc/current/ext/ext.html`). In the current (Oct. '02) documentation, you want section 1.7, "Extracting Parameters in Extension Functions".

`PyArg_ParseTuple()` checks the types of the objects in the tuple args against the types given in the format string. If there's a discrepancy, it sets the exception context and returns 0 to our wrapper. If our wrapper detects that, it returns 0 to the Python interpreter, which understands that to mean failure, and raises an exception. So if you're using `PyArg_ParseTuple()`, most of the error checking is done for you! It is not idiot-proof, but it is smart-friendly.

Once `PyArg_ParseTuple` has successfully returned, do any additional processing or checking of the input data that's appropriate. For instance, Python does not have an unsigned integer type. With a C++ function that takes an unsigned int, you'll need to pass an int to `PyArg_ParseTuple`, check that the int is greater than -1, and then convert it to an unsigned int.

b) Call your code

It's your function, call it.

c) Convert output to a Python object, and return it

The API gives a function called `Py_BuildValue`. It returns a pointer to a PyObject; it takes a format string and variables. The format strings are the same as those used in

```
PyObject *py_result = Py_BuildValue("i", result);
```

or—

```
PyObject *py_result = Py_BuildValue("s", astring);
```

etc. What's returned by `Py_BuildValue` is what the wrapper will return. One note: don't return 0; the interpreter will take this as sign of failure. You could, of course, return a Python integer with value zero: `Py_BuildValue("i",0)`. You can specify more than one item to return, in which case `Py_BuildValue` will place the items in a tuple.

Here's a complete example of wrapping a function, "bogus", that takes a double, a string, and an int (in that order) and returns an int. We expect that the arguments will come from Python in the order string, int, double.

```
static PyObject *wrap_bogus(PyObject *, PyObject * args){
 //First, get the arguments from Python
 int anint = 0;
 double adub = 0;
 char * astring = 0;
 int ok = PyArg_ParseTuple(args,"sid",&astring, &anint, &adub);
 if(!ok) return 0;
 //do any checking of arguments here
 //Second, make the function call
 int result = bogus(adub, astring, anint);
 //do any extra stuff you want with the return result here
 //Third, build a Python object to return
 return PyBuildValue("i",result);
}
```

Wrapping classes: PyCObject_FromVoidPtr, PyCObject_AsVoidPtr

a) Creating C++ objects

Wrapping functions is well and good, but what about C++ classes? Python can work with a C++ object by dynamically allocating it and holding onto a pointer. That pointer can be passed back to subsequent wrappers that can then invoke class methods on the object. The pointer is handled in Python by a type PyCObject. To convert a pointer to a PyCObject, use `PyCObject_FromVoidPtr()`. This API function takes two arguments: the void pointer, and a pointer to a function that takes a void pointer and no return. The purpose of the function is to delete the C++ object when nothing in the Python session is paying attention to it any more.

We get the arguments to the constructor from the Python API, create the object using new, and return a pointer to that object to the interpreter. Use the API function `PyCObject_FromVoidPtr()` to create the Python object to return to Python. Here's an example with a real (if dull) class called Numbers:

```
class Numbers
{
public:
 Numbers(int first, double second) : m_first(first), m_second(second){;}
 ~Numbers(void){;}
 double NumMemberMult(void){return (double)m_first*m_second;}
private:
 int m_first;
 double m_second;
};
```

Here's a wrapper that creates a new instance of Numbers:

```
PyObject *wrap_new_Numbers(PyObject *, PyObject* args){
 //First, extract the arguments from a Python tuple
 int arg1;
 double arg2;
 int ok = PyArg_ParseTuple(args,"id",&arg1,&arg2);
 if(!ok) return 0;
 //Second, dynamically allocate a new object
```

```
 Numbers *newnum = new Numbers(arg1, arg2);
 //Third, build a Python object to return
 PyObject * py_newnum = PyCObject_FromVoidPtr( static_cast<void *>(newnum),
 &DelNumbers);
 return py_newnum;
}
```

Look familiar? This wrapper has essentially the same form as `wrap_bogus()`. That's because ALL wrappers have essentially this form.

The pointer to the dynamically allocated object, newnum, goes out of scope as soon as `wrap_new_Numbers()` returns. The only thing keeping this from being the mega-classic memory leak is that the Python interpreter has an object that will keep track of the address of the new object. The interpreter keeps track of that object for us, and when we lose interest in it (when its reference count goes to zero), the interpreter will call a C++ function to delete the C++ object. So, the second slot in `PyCObject_AsVoidPtr()` is a pointer to a function with return type void and one void argument. The signature of `PyCObject_AsVoidPtr()` is:

```
PyObject * PyCObject_FromVoidPtr( void *, void (*DeleteFunction)(void*));
```

You must supply the function pointed to (in this example called DelNumbers). It has the delete corresponding to the new above. Here's it is:

```
static void DelNumbers(void *ptr)
{
    Numbers * oldnum = static_cast<Numbers *>(ptr);
    delete oldnum;
    return;
}
```

This strategy can be used for any dynamically allocated resource, such as file handles or arrays.

Using the object

The user can't actually do anything with the pointer in the Python layer, except send it back to the C++ layer to do something else: call a C++ class method on it, or give it as an argument to another function. Here's an example of wrapping a class method.

```
#include <Python.h>
...
PyObject *wrap_Numbers_MemberMult(PyObject *, PyObject* args)
{
 // First, extract the PyCObject that has the
 // Python version of the address
 // from the args tuple
 PyObject *pynum = 0;
 int ok = PyArg_ParseTuple(args, "O", &pynum);
 //"O" is for Object
 if(!ok) return NULL;
 //Convert the PyCObject to a void *
 void * temp = PyCObject_AsVoidPtr(pynum);
```

```
 //cast void pointer to Numbers pointer
 Numbers * thisnum = static_cast<Numbers *>(temp);
 //Can combine the two lines into one:
 //Numbers *thisNum = static_cast<Numbers *>(
 //                        PyCObject_AsVoidPtr(pynum));
 //Second, make the function call
 double result = thisnum->NumMemberMult();
 //Third, build a Python object with the return value
 return Py_BuildValue("d",result);
}
```

All you have to do is fish the PyCObject out of the tuple, extract the void pointer to a C++ variable, cast it to the appropriate type, and use it; then bundle up the result and send it back to the interpreter.

Working with composite types

What if you want to pass a Python list of numbers to a C++ function? There's no format code to pass to `PyArg_ParseTuple` for lists. The solution is to extract the list from the `args` tuple as a PyObject (format code: "O"). Then use the Python-C API functions for working with lists to load the Python list, item-by-item, into a C++ array. Suppose our target function has the signature

```
double sum_some_numbers(double *numbers, int array_length)
```

Here's some code that could wrap this function. Note that we have to take responsibility for error checking. We can set the exception using `PyErr_SetString()`. We can verify that Python objects are the *type* what we think they are by Py*type*_Check().

```
#include <Python.h>
#include <valarray> // This example uses the std::valarray class.
...
PyObject *py_sum_some_numbers(PyObject *, PyObject* args)
{
    PyObject *pyList;
    int ok = PyArg_ParseTuple(args, "O", &pyList);
    if(!ok) return 0;
    //Did the user send a Python list?
    int isList = PyList_Check(pyList);
    if(!isList)
    {
        //If not, complain to the Python user and raise an exception:
        PyErr_SetString(PyExc_TypeError, "You fool! That's not a list!");
        return 0;
    }
    //How many items are in the list?
    int numNums = PyList_Size(pyList);
    //Maybe you want to do something here if the size of the list is 0.
    // Now transfer the contents of the list to an array
    // valarray: this C++ standard library class is a great
    // way to avoid memory leaks, and much more
    std::valarray<double> nums(0.0, PyList_Size);
    for(int i=0; i<numNums; i++)
    {
        //Extract the next object in the list:
```

```
        PyObject *temp = PyList_GetItem(pyList, i);
        // Was the list item a Python float? If not, quit.
        // Note that we don't need to worry about cleaning
        // up the memory: nums will be automatically destroyed
        // when execution exits the scope of nums.
        if( !PyFloat_Check(temp))
        {
            //Just what was in that list?
            PyErr_SetString(PyExc_TypeError, "You fool! That's not a float";
            return 0;
        }
        // Now convert the Python float to a double, and load
        nums[i] = PyFloat_AsDouble(temp);
    }
    //Step 2: Call the function
    double sum = sum_some_numbers(&nums[0], nums.size());
    //Step 3: return result.
    return Py_BuildValue("d", sum);
}
```

Sometimes you'll want to return a Python object, such as a list. In this case, you'll have created a PyObject pointer at some point. You can return that pointer directly. If you need to return several Python objects, you can use `Py_BuildValue()` with the "O" format code.

The Python-C api is very complete and well-documented. Similar functions exist for inserting objects into lists, and working with dictionaries, tuples, modules, and so on. Consult Chapter 7 of the api reference. Hopefully, these examples have given you the flavor for wrapping C++ functions.

A.1.4 A Lot More Detail: Method Table

The method table is sort of like a table of contents for the Python interpreter. When it loads the library, it reads the method table to find pointers to the functions in the library.

```
static PyMethodDef numbersMethods[] = {
    {"PyNumbers", wrap_new_Numbers, METH_VARARGS,
         "Create new Numbers object"},
    {"PyNumbers_MembMult", wrap_Numbers_MemberMult, METH_VARARGS,
         "Multiply Numbers object's members"},
    {NULL,NULL}
};
```

The name of the table must match the second argument in the init function. Each function in the library gets an entry in the table, and each entry has four components.

1. The string ("PyNumbers" or "PyNumbers_MemberMult") is what you'll call from the interpreter.

2. The name `wrap_whatever` is the name of the corresponding C++ function.

3. `METH_VARARGS` indicates that one is using the "tuple named args" approach.
4. The final string will appear as the docstring for this function in the Python layer.

The `{NULL,NULL}` marks the end of the table for the interpreter.

A.1.5 A Lot More Detail: Init Function

The final component in the bindings is the init function. This function has to be named init*name_of_library*(). If the filename of the library is numbers.dll or numbers.so, this function must be named `initnumbers`; if it's _numbers.dll, then this function is named `init_numbers`. For this example, let's call the extension library _numbers. The init function's return type is void, and it takes no arguments. It calls `Py_InitModule()`, which takes two arguments: a string literal with the name of the library, and the name of the Methods table. It *must* match the name of the Method table. Also, the function must have C linkage, not C++, meaning the function must be declared extern "C" if you're using a C++ compiler. Also, on Windows, the function must be exported by the dll, hence the `__declspec(dllexport)`. To keep some platform independence, wrap this in a pre-processor conditional. This function is executed when the library is loaded, so if there's other initialization steps you need to take, this is the place.

```
extern "C"
#ifdef WIN32 || _WIN32
__declspec(dllexport)
#endif
void init_numbers(void)
{
    (void) Py_InitModule("_numbers", numbersMethods);
}
```

A.1.6 More Detail: Compile

Compiling the Numbers example under Linux

Here's a way to compile the Numbers example under Linux. You'll need files with the Numbers source code (Numbers.cpp and Numbers.h), and the wrapper, Numbers_bindings.cpp. I assume you're using gcc; if not, you'll need to modify the compiler flags appropriately.

To compile, compile each source file:

```
gcc -I/usr/include/python2.2 -I. -c -fpic  Numbers.cpp
gcc -I/usr/include/python2.2 -I. -c -fpic Numbers_bindings.cpp
```

and link them into a shared library:

```
gcc Numbers.o Numbers_bindgins.o -lm -lc -fpic -shared -o _numbers.so
```

In the compile lines, you'll of course need to make sure that you've pointed to the directory where your Python.h file lives. The "-fpic" specifies position independent code, "-shared" a shared library that can be dynamically linked.

Once you've compiled _numbers.so, move it into a place on your system's PYTHON_PATH and go to town.

Compiling the Numbers example under Windows

Well, of course we want all our ARCS modules to run under Linux/Unix, but for all those times Windows needs a helping hand, here's how to do it:

For working in Windows, it may be best to use MS Visual C++. Here's what you'd do to create a project and so on in VC7.

1. From the Start page select New Project.
2. From the Project Types, pick Visual C++ Projects, from the Templates, choose Win32 project. Fill in the name for your library.
3. On the next window, pick Application Settings, and set Application Type to DLL.
4. In the Solution Explorer,
   - get rid of stdafx.cpp
   - right-click on the name of the project, and choose Properties
   - Under the C/C++/General folder, add the additional include directory in which your distribution's Python.h file resides.
   - Under the Linker/General tab, set the output file to _your_project_name.dll. Note that everything up the ".dll" must be the same as what follows "init" in the `init_your_project_name` function in the bindings. In the Numbers example, we called that function init_numbers (it's the last function in the file). So the dll has to be called _numbers.dll. If the name of the file and the init function don't agree, the interpreter will get lost.
5. In Solution Explorer, look in the file your_project_name.cpp– you'll need to get rid of that crap about APIENTRY DllMain. Better yet, just get rid of the automatically generated code.

Now fill add the source files Numbers.cpp, Numbers_bindings.cpp, etc. Build, and when the _your_project_name.dll appears in the output directory, move it into your PYTHON_PATH, and enjoy.

A.1.7 More Detail: Call it from Python

In keeping with the Numbers example used above, here's a Python class called Numbers. It's a "shadow class" for the C++ Numbers class. Pretty much everything you can do with the C++ class can also be done with the Python.

```
import _numbers
class Numbers:
    def __init__(self, an_int, a_float):
        #Check an_int
        if type(an_int) != type(1):
            raise TypeError, "Fool! an_int must be an integer"
        #Check a_float
        if type(a_float) != type(3.14159)
            raise TypeError, "Fool! a_float must be a float"
        self.this = numbers.PyNumbers(an_int, a_float)
    def MemberMult(self):
        return numbers.PyNumbers_MembMult(self.this)
```

If this were saved in a module named numbers.py (stored somewhere on your PYTHON_PATH), then an interpreter command line session might look like

```
>>> import numbers
>>> n = numbers.Numbers(2,3.14)
New Numbers object created
>>> n.MemberMult()
6.280000000002
>>> n = 1
```

The variable n is the only thing keeping track of the Numbers object. When we reassign n, the interpreter calls our bit of code from the C++ library that deletes the object, preventing the resource leak.

So now the Python user has something like "interpreted C++". Pretty cool, eh?

A.1.8 More realistic example

Let's wrap a function from the NeXus API, NXopen, with all the bells and whistles. This wrapper is part of a larger library that wraps the entire NeXus C API. The latter, of course, is a simplified interface to the HDF libraries that the NeXus standard currently uses.

Since this is one of several dozen libraries, we split the bindings up into several files. The wrappers live in pairs of files, one pair for each major functional group: file.h/file.cc for file level operations, group.h/group.cc for group level, etc. Only two file level operations from the original NeXus C API are in the Python NeXus API: NXopen and NXflush. Therefore, file.h looks like the following:

```
#ifndef NeXus_file_h
#define NeXus_file_h
// Python bindings for file level operations:
// NXopen
extern char pyNeXus_NXopen__name__[];
extern char pyNeXus_NXopen__doc__[];
extern "C" PyObject * pyNeXus_NXopen(PyObject *, PyObject *args);
// NXflush
extern char pyNeXus_NXflush__name__[];
extern char pyNeXus_NXflush__doc__[];
```

```
extern "C" PyObject * pyNeXus_NXflush(PyObject *, PyObject *args);
#endif
```

This file will be included into the file that contains the methods table. Note that in addition to the actual wrapper functions, we declare two char arrays for the name and docstring. Defined in file.cc, these variables keep the methods table neat (the docstrings in particular may get lengthy).

How do we implement the wrapper for NXopen? Begin with the signature of NXopen, located in napi.h in the NeXus source distribution:

```
NX_EXTERNAL  NXstatus CALLING_STYLE NXopen(CONSTCHAR * filename,
                NXaccess access_method, NXhandle* pHandle);
```

The various types NXstatus, CONSTCHAR, NXaccess, and NXhandle are defined in the NeXus C API header files; we need to track them down so we can know what the Python user will have to give us in order to satisfy the function call.

Searching through napi.h, we learn that `NXstatus` is a `typedef` for `int`, `CONSTCHAR` is a `typedef` for char,[2] `NXaccess` is an enumeration with members like `NXACC_READ`, and `NXhandle` is a `typedef` for `void *`. We can't map the `NXaccess` enumeration directly into Python types, so we'll expect a string from the user; by comparing values of the string we'll assign the proper value to an `NXaccess` variable. As for the `CONSTCHAR * filename`, we can derive that directly from a Python string.

What about the `NXhandle *pHandle`? This is interesting. Let's open up the NeXus C API source code and find out what exactly is done with that void pointer. Reading through napi.c, we discover that `NXopen` dynamically allocates a structure of type `NexusFunction`, and that the `NXhandle` we pass to `NXopen` becomes a handle to that object. In other words, `pHandle` is an output of `NXopen`, not an input. We don't need to trouble the Python user with giving us a `NXhandle` object; instead, we'll give them one.

Expecting two inputs, both Python strings, we write the first few lines of the wrapper as follows:

```
PyObject * pyNeXus_NXopen(PyObject *, PyObject *args)
{
    char *filename = 0;
    char *acc_method = 0;
    int ok = PyArg_ParseTuple(args, "ss", &filename, &acc_method);
    if(!ok) return 0;
```

and we've got what we needed from the Python user. Or do we? We owe it to ourselves to check the inputs from the user. This could be done either here or in the Python layer. Checking that the inputs are Python strings has already been performed by `PyArg_ParseTuple`. There are two questions about the filename: is it the name of an actual file, and is it the name of an appropriate HDF file?

[2] Prefering to not mislead the readers of our code, we might have chosen CONSTCHAR as typedef const char.

The first question is easily answered in a platform independent way in Python. The second can only be answered by essentially doing what we do anyway in `NXopen`. We need to inspect the value of `acc_method` anyway to convert it to the appropriate member of the NXaccess enumeration. We can do that easily using the C++ standard library class `string`:

```
//Check access_method:
std::string methstring(acc_method);
NXaccess mode;
if     (methstring == "r" ) mode = NXACC_READ;
else if(methstring == "rw") mode = NXACC_RDWR;
else if(methstring == "c" ) mode = NXACC_CREATE;
else if(methstring == "c4") mode = NXACC_CREATE4;
else if(methstring == "c5") mode = NXACC_CREATE5;
else
{
    std::string errstr("NeXus_bindings.cc pyNeXus_NXopen(): ");
    errstr += "unrecognized access_method string.";
    PyErr_SetString( PyExc_ValueError, errstr.c_str() );
    return 0;
}
```

Here we've used `PyErr_SetString` to set the (Python) exception context if we don't recognize what the Python user wants, and then forced the interpreter to raise the exception by returning 0.

At this point, we're nearly ready to call `NXopen`—we only need to declare a variable of type NXhandle:

```
NXhandle handle;
NXstatus status = NXopen(filename, mode, &handle);
if(status != NX_OK)
{
    std::string errstr("NeXus_bindings.cc pyNeXus_NXopen(): ");
    errstr += "NXopen failed.";
    PyErr_SetString(PyExc_IOError, errstr.c_str() );
    return 0;
}
```

If we get to this stage, we're almost ready to return. Since we are not content with any `NXstatus` but `NX_OK`, all we need to do is to return the `NXhandle` initialized by `NXopen`. But there's another issue here: `NXopen` allocates a resource, so we must release that resource when we're finished. As with the discussion of allocating C++ objects (§ A.1.3), we can use `PyCObject_FromVoidPtr` and a helper function to release the resource when the Python user is finished:

```
    return PyCObject_FromVoidPtr(handle, pyNeXus_NXclose);
}
```

The helper function (declared `static` to avoid) wraps `NXclose`, which in turn disposes of the resources allocated in `NXopen`:

```
static void pyNeXus_NXclose(void *file)
{
```

```
    NXhandle oldnxh = static_cast<NXhandle >(file);
    NXclose( &oldnxh);
    return;
}
```

Finally, we define the name and docstring for this function. The name should be something sensible, while the docstring is an opportunity to incorporate a little documentation.

```
char pyNeXus_NXopen__name__[] = "nxopen";
char pyNeXus_NXopen__doc__[] = "Open a nexus file\n"
"2 Arguments: filename, access_method\n"
"Input: \n"
"       filename (Python string)\n"
"       access_method (Python string)\n"
"            allowed values: r (read only)\n"
"                            rw (read/write)\n"
"                            c (create)\n"
"                            c4 (create HDF 4)\n"
"                            c5 (create HDF 5)\n"
"Output: (return)\n"
"        PyCObject holding pointer to NXhandle\n"
"Exceptions: ValueError, IOError\n";
```

Appendix A

Appendix 3: Selected Derivations and Physical Constants

A.1 Convolutions and Correlations

A.1.1 Convolution Theorem

It is easiest to explain convolutions is in terms of a broadening of a sharp peak caused by making a measurement with a blurry instrument. The instrumental broadening function is $f(k)$.[1] We seek the true specimen diffraction profile $g(k)$. What we actually measure with our diffractometer is the convolution of $f(k)$ and $g(k)$, denoted $h(K)$ (where K is the shift of the detector across the diffraction intensity). Deconvolution will require the Fourier transforms of $f(k)$, $g(k)$, $h(K)$:

$$f(k) = \sum_n F(n)\, e^{i2\pi nk/l} \qquad \text{equipment,} \tag{A.1}$$

$$g(k) = \sum_{n'} G(n')\, e^{i2\pi n'k/l} \qquad \text{specimen,} \tag{A.2}$$

$$h(K) = \sum_{n''} H(n'')\, e^{i2\pi n''K/l} \qquad \text{measurement.} \tag{A.3}$$

Note that l has units of inverse distance, so n/l is a real space variable. The range in k of the Fourier series is the interval $-l/2$ to $+l/2$, which includes all features of a diffraction peak.[2] The convolution of f and g is defined as:

$$h(K) = \int_{-\infty}^{\infty} f(K-k)\, g(k)\, dk\ . \tag{A.4}$$

[1] Measurements are typically in scattering angle, which is interpretable as a k-space variable.

[2] We don't care about $f(k)$ and $g(k)$ outside this interval, but with (A.1)–(A.3) these Fourier transforms repeat themselves with a period of l. We confine ourselves to one period, and require that f and g vanish at its ends.

We must choose an interval so that that f and g vanish outside the range $\pm\, l/2$, so we can change the limits of integration from $\pm\,\infty$ to $\pm\, l/2$. Substitute (A.1) and (A.2) into (A.4):

$$h(K) = \int_{-l/2}^{l/2} \sum_n F(n)\, \mathrm{e}^{\mathrm{i}2\pi n(K-k)/l} \sum_{n'} G(n')\, \mathrm{e}^{\mathrm{i}2\pi n' k/l}\, \mathrm{d}k\ . \tag{A.5}$$

We rearrange summations over the independent variables n and n', and remove from the integral all factors independent of k:

$$h(K) = \sum_{n'} \sum_n G(n')F(n)\, \mathrm{e}^{\mathrm{i}2\pi nK/l} \int_{-l/2}^{l/2} \mathrm{e}^{\mathrm{i}2\pi(n'-n)k/l}\, \mathrm{d}k\ . \tag{A.6}$$

Now we employ the orthogonality condition[3]:

$$\int_{-l/2}^{l/2} \mathrm{e}^{\mathrm{i}2\pi(n'-n)k/l} \mathrm{d}k = \left\{ \begin{matrix} l \text{ if } n' = n \\ 0 \text{ if } n' \neq n \end{matrix} \right\}. \tag{A.7}$$

With the orthogonality condition of (A.7), the double sum in (A.6) is reduced to a single sum:

$$h(K) = l \sum_n G(n)F(n)\, \mathrm{e}^{\mathrm{i}2\pi nK/l}\ . \tag{A.8}$$

Compare (A.8) to the definition for $h(K)$ in (A.3). We see that that the Fourier coefficients $H(n'')$ are proportional to the product of $G(n)$ and $F(n)$:

$$l\, G(n)F(n) = H(n)\ . \tag{A.9}$$

By comparing (A.4) and (A.9), we see that a convolution in k-space is equivalent to a multiplication in real space (with variable n/l). The converse is also true; a convolution in real space is equivalent to a multiplication in k-space. This important result is the *convolution theorem*.

A.1.2 Deconvolutions

Equation (A.9) shows how to perform the deconvolution of $f(k)$ from $h(K)$; perform a division in n-space. Specifically, when we have the full sets of Fourier coefficients $\{F(n)\}$ and $\{H(n)\}$, we perform a division in n-space for each Fourier coefficient:

$$G(n) = \frac{1}{l}\, \frac{H(n)}{F(n)}\ . \tag{A.10}$$

[3]Verified by writing the exponential as $\cos(2\pi(n'-n)k/l) + \mathrm{i}\sin(2\pi(n'-n)k/l)$. The sine integration vanishes by symmetry. The cosine integration gives $l[2\pi(n'-n)]^{-1}[\sin(\pi(n'-n)) - \sin(\pi(n'-n))]$, which $= 0$ when $n' - n \neq 0$. In the case when $n' - n = 0$, the integrand in (A.7) equals 1, so the integration gives l.

We obtain each $F(n')$ by multiplying both sides of (A.1) by $\exp(-\mathrm{i}2\pi n'k/l)$ and integrating over k:

$$\int_{-l/2}^{l/2} f(k)\,\mathrm{e}^{-\mathrm{i}2\pi n'k/l}\mathrm{d}k = \sum_n F(n) \int_{-l/2}^{l/2} \mathrm{e}^{\mathrm{i}2\pi(n-n')k/l}\mathrm{d}k\,. \tag{A.11}$$

The orthogonality relationship of (A.7) causes the right-hand-side of (A.11) to equal zero unless $n = n'$. Equation (A.11) therefore becomes:

$$\frac{1}{l}\int_{-l/2}^{l/2} f(k)\,\mathrm{e}^{-\mathrm{i}2\pi n'k/l}\mathrm{d}k = F(n')\,. \tag{A.12}$$

The Fourier coefficients $H(n)$ are obtained the same way. The simple division of Fourier coefficients in (A.10) then provides the set of Fourier coefficients for the true specimen profile, $\{G(n)\}$. If we then use (A.2) to take the Fourier transform of the $\{G(n)\}$ from (A.10), we obtain $g(k)$, the true specimen diffraction profile.

A.2 Fourier Transform of Screened Coulomb Potential

In this subsection we calculate the Fourier transform of a "screened Coulomb" potential, a result that is useful in calculations of form factors of atoms for example. This screened Coulomb potential, $V(r)$, is:

$$V(r) = -\frac{Ze^2}{r}\mathrm{e}^{-r/r_0}\,. \tag{A.13}$$

The exponential factor accounts for the screening of the nuclear charge by the atomic electrons, and r_0 is an effective Bohr radius for the atom. Interestingly, the exponential decay also facilitates the mathematics of working with a potential that is otherwise strong at very large distances.

We now use the first Born approximation, (1.75), to calculate the atomic scattering factor, $f\left(\vec{\Delta k}\right)$, as the Fourier transform of $V(\vec{r})$:

$$f_{\mathrm{el}}\left(\vec{\Delta k}\right) = -\frac{m}{2\pi\hbar^2}\int_{\text{all space}} \mathrm{e}^{-\mathrm{i}\vec{\Delta k}\cdot\vec{r}}\,V(\vec{r})\,\mathrm{d}^3\vec{r}\,. \tag{A.14}$$

Substituting the potential (A.13) into (A.14):

$$f_{\mathrm{el}}(\vec{\Delta k}) = \frac{mZe^2}{2\pi\hbar^2}\int_{\text{all space}} \mathrm{e}^{-\mathrm{i}\vec{\Delta k}\cdot\vec{r}}\frac{\mathrm{e}^{-r/r_0}}{r}\mathrm{d}^3\vec{r}\,. \tag{A.15}$$

The integral, $\mathcal{I}\left(\vec{\Delta k}, r_0\right)$, in (A.15) occurs in other contexts, so we pause to solve it.

$$\mathcal{I}\left(\vec{\Delta k}, r_0\right) \equiv \int\limits_{\text{all space}} \mathrm{e}^{-\mathrm{i}\vec{\Delta k}\cdot\vec{r}} \frac{\mathrm{e}^{-r/r_0}}{r} \mathrm{d}^3\vec{r}\,, \tag{A.16}$$

which is the 3-dimensional Fourier transform of the screened Coulomb potential (A.13). It is natural to use spherical coordinates:

$$\mathcal{I}\left(\vec{\Delta k}, r_0\right) = \int\limits_{r=0}^{\infty}\int\limits_{\theta=0}^{\pi}\int\limits_{\phi=0}^{2\pi} \mathrm{e}^{-\mathrm{i}\vec{\Delta k}\cdot\vec{r}} \frac{\mathrm{e}^{-r/r_0}}{r} r^2 \sin\theta \,\mathrm{d}\theta \,\mathrm{d}\phi \,\mathrm{d}r\,. \tag{A.17}$$

The trick for working with the exponential in (A.17), $\mathrm{e}^{-\mathrm{i}\vec{\Delta k}\cdot\vec{r}}$, is to align the vector $\vec{\Delta k}$ along the z-axis so that $\vec{\Delta k}\cdot\vec{r} = \Delta k z$. Also, since $z = r\cos\theta$:

$$\mathrm{d}z = -r\sin\theta\,\mathrm{d}\theta\,. \tag{A.18}$$

The limits of integration are changed as:

$$\theta = 0 \Longrightarrow \; z \;=\; r\,, \tag{A.19}$$

$$\theta = \pi \Longrightarrow \; z \;=\; -r\,. \tag{A.20}$$

With the substitution of (A.18)–(A.20) into (A.17):

$$\mathcal{I}\left(\vec{\Delta k}, r_0\right) \;=\; \int\limits_{r=0}^{\infty}\int\limits_{z=r}^{-r}\int\limits_{\phi=0}^{2\pi} \mathrm{e}^{-\mathrm{i}\Delta k z}\mathrm{e}^{-r/r_0}\mathrm{d}\phi(-\mathrm{d}z)\mathrm{d}r\,, \tag{A.21}$$

$$\mathcal{I}\left(\vec{\Delta k}, r_0\right) \;=\; 2\pi\int\limits_{r=0}^{\infty}\int\limits_{z=-r}^{r} \mathrm{e}^{-\mathrm{i}\Delta k z}\mathrm{e}^{-r/r_0}\mathrm{d}z\,\mathrm{d}r\,. \tag{A.22}$$

Writing the exponential as $\mathrm{e}^{-\mathrm{i}\Delta k z} = \cos(\Delta k z) - \mathrm{i}\sin(\Delta k z)$, the z-integration of the sine function vanishes by symmetry in the interval $-r$ to $+r$, and the cosine integral is:

$$\int\limits_{z=-r}^{r} \cos(\Delta k z)\,\mathrm{d}z = \frac{+2}{\Delta k}\sin(\Delta k r)\;, \tag{A.23}$$

which does not depend on the direction $\widehat{\vec{\Delta k}}$. Using (A.23) for the z-integration in (A.22), we obtain:

$$\mathcal{I}\left(\Delta k, r_0\right) = \frac{4\pi}{\Delta k}\int\limits_{r=0}^{\infty} \sin(\Delta k r)\,\mathrm{e}^{-r/r_0}\mathrm{d}r\,. \tag{A.24}$$

Equation (A.24) is the Fourier transform of a decaying exponential. This integral can be solved by twice integrating by parts.[4] The result is a Lorentzian function:

$$\int_{r=0}^{\infty} \sin(\Delta k r)\, \mathrm{e}^{-r/r_0}\, \mathrm{d}r = \frac{\Delta k}{\Delta k^2 + \frac{1}{r_0^2}} \ . \tag{A.25}$$

We substitute the result (A.25) into (A.24), completing the evaluation of (A.16):

$$\mathcal{I}(\Delta k, r_0) = \int_{\text{all space}} \mathrm{e}^{-\mathrm{i}\vec{\Delta k}\cdot\vec{r}} \frac{\mathrm{e}^{-r/r_0}}{r} \mathrm{d}^3\vec{r} = \frac{4\pi}{\Delta k^2 + \frac{1}{r_0^2}} \ . \tag{A.26}$$

For later convenience, we now obtain a related result. The use of an exponential screening factor to perform a Fourier transform of the Coulomb potential is a useful mathematical trick. By letting $r_0 \to \infty$, we suppress the screening of the Coulomb potential, so $\mathrm{e}^{-r/r_0} = 1$ in (A.13). The Fourier transform of this bare Coulomb potential, with its mathematical form of $1/r$, is obtained easily from (A.26):

$$\int_{\text{all space}} \mathrm{e}^{-\mathrm{i}\vec{\Delta k}\cdot\vec{r}} \frac{1}{r}\, \mathrm{d}^3\vec{r} = \frac{4\pi}{\Delta k^2} \ . \tag{A.27}$$

[4] Defining $U \equiv \mathrm{e}^{-r/r_0}$ and $\mathrm{d}V \equiv \sin(\Delta k r)\ \mathrm{d}r$, we integrate by parts: $\int U\mathrm{d}V = UV - \int V\mathrm{d}U$. The integral on the right hand side is evaluated as: $(\Delta k r_0)^{-1} \int_{r=0}^{\infty} \cos(\Delta k r)\, \mathrm{e}^{-r/r_0}\, \mathrm{d}r$, which we integrate by parts again to obtain: $-(\Delta k r_0)^{-2} \int_{r=0}^{\infty} \sin(\Delta k r)\, \mathrm{e}^{-r/r_0}\, \mathrm{d}r$. This result can be added to the $\int U\mathrm{d}V$ on the left hand side to obtain (A.25).

A.3 Fundamental and Derived Constants

Fundamental Constants

$\hbar = 1.0546 \times 10^{-27}$ erg·sec = 6.5821×10^{-16} eV·sec
$k_B = 1.3807 \times 10^{-23}$ J/(atom·K) = 8.6174×10^{-5} eV/(atom·K)
R = 0.00198 kcal/(mole·K) = 8.3145 J/(mole·K) (gas constant)
$c = 2.998 \times 10^{10}$ cm/sec (speed of light in vacuum)
$m_e = 0.91094 \times 10^{-27}$ g = 0.5110 MeV·c^{-2} (electron mass)
$m_n = 1.6749 \times 10^{-24}$ g = 939.55 MeV·c^{-2} (neutron mass)
$N_A = 6.02214 \times 10^{23}$ atoms/mole (Avogadro constant)
$e = 4.80 \times 10^{-10}$ esu = 1.6022×10^{-19} coulomb
$\mu_0 = 1.26 \times 10^{-6}$ henry/m
$\varepsilon_0 = 8.85 \times 10^{-12}$ farad/m

$a_0 = \hbar^2/(m_e e^2) = 5.292 \times 10^{-9}$ cm (Bohr radius)
$e^2/(m_e c^2) = 2.81794 \times 10^{-13}$ cm (classical electron radius)
$e^2/(2a_0)$ = R (Rydberg) = 13.606 eV (K-shell energy of hydrogen)
$e\hbar/(2m_e c) = 0.9274 \times 10^{-20}$ erg/oersted (Bohr magneton)
$\hbar^2/(2m_e) = 3.813 \times 10^{-16}$ eV s cm^{-2}

Definitions

1 becquerel (B) = 1 disintegration/second
1 Curie = 3.7×10^{10} disintegrations/second

radiation dose:
1 roentgen (R) = 0.000258 coulomb/kilogram
Gray (Gy) = 1 J/kG

Sievert (Sv) is a unit of "radiation dose equivalent" (meaning that doses of radiation with equal numbers of Sieverts have similar biological effects, even when the types of radiation are different). It includes a dimensionless quality factor, Q (Q~1 for x-rays, 10 for neutrons, and 20 for α-particles), and energy distribution factor, N. The dose in Sv for an energy deposition of D in Grays [J/kG] is:
Sv = Q×N×D [J/kG]
Rad equivalent man (rem) is a unit of radiation dose equivalent approximately equal to 0.01 Sv for hard x-rays.

1 joule = 1 J = 1 W·s = 1 N·m = 1 kg·m^2·s^{-2}
1 joule = 10^7 erg
1 newton = 1 N = 1 kg·m·s^{-2}
1 dyne = 1 g·cm·s^{-2} = 10^{-5} N
1 erg = 1 dyne·cm = 1 g·cm^2·s^{-2}
1 Pascal = 1 Pa = 1 N·m^{-2}
1 coulomb = 1 C = 1 A·s
1 ampere = 1 A = 1 C/s

Element	$K\overline{\alpha}$	$K\alpha_1$	$K\alpha_2$	$K\beta_1$
Cr	2.29092	2.28962	2.29351	2.08480
Co	1.79021	1.78896	1.79278	1.62075
Cu	1.54178	1.54052	1.54433	1.39217
Mo	0.71069	0.70926	0.71354	0.632253
Ag	0.56083	0.55936	0.56377	0.49701

1 volt = 1 V = 1 W·A^{-1} = 1 m^2·kg·A^{-1}·s^{-3}
1 ohm = 1 Ω = 1 V·A^{-1} = 1 m^2·kg·A^{-2}·s^{-3}
1 farad = 1 F = 1 C·V^{-1} = 1 m^{-2}·kg^{-1}·A^2·s^4
1 henry = 1 H = 1 Wb·A^{-1} = 1 m^2·kg·A^{-2}·s^{-2}
1 tesla = 1 T = 10,000 gauss = 1 Wb·m^{-2} = 1 V·s·m^{-2} = 1 kg·s^{-2}·A^{-1}

Conversion Factors

1 Å = 0.1 nm = 10^{-4} μm = 10^{-10} m
1 b (barn) = 10^{-24} cm^2
1 eV = 1.6045×10^{-12} erg
1 eV/atom = 23.0605 kcal/mole = 96.4853 kJ/mole
1 cal = 4.1840 J
1 bar= 10^5 Pa
1 torr = 1 T = 133 Pa
1 kG = 5.6096×10^{29} MeV·c^{-2}

Useful Facts

energy of 1 Å photon = 12.3984 keV
$h\nu$ for 10^{12} Hz = 4.13567 meV
1 meV = 8.0655 cm^{-1}
temperature associated with 1 eV = 11,600 K
lattice parameter of Si (in vacuum at 22.5°C) = 5.431021 Å

Neutron Wavelengths, Energies, Velocities

E_n = 81.81 λ^{-2} (energy-wavelength relation for neutrons [meV, Å])
λ_n = 3955.4/v_n (wavelength-velocity relation for neutrons [Å, m/s])
E_n = 5.2276×10^{-6} v_n^2 (energy-velocity relation for neutrons [meV, m/s])

Relativistic Electron Wavelengths

For an electron of energy E [keV] and wavelength λ [Å]:

$$\lambda = h\Big[2m_e E\Big(1 + \frac{E}{2m_e c^2}\Big)\Big]^{-1/2} = \frac{0.3877}{E^{1/2}\,(1 + 0.9788 \times 10^{-3} E)^{1/2}}$$

kinetic energy $\equiv T = \frac{1}{2} m_e v^2 = \frac{1}{2} E \frac{1+\gamma}{\gamma^2}$

Some X-Ray Wavelengths [Å]

Table A.1: Parameters of high-energy electrons

E [keV]	λ [Å]	γ	v [c]	T [keV]
100	0.03700	1.1957	0.5482	76.79
120	0.03348	1.2348	0.5867	87.94
150	0.02956	1.2935	0.6343	102.8
200	0.02507	1.3914	0.6953	123.6
300	0.01968	1.587	0.7765	154.1
400	0.01643	1.7827	0.8279	175.1
500	0.01421	1.9785	0.8628	190.2
1000	0.008715	2.957	0.9411	226.3

Index