# Notes for Lecture 15

# Complex number, SHM, Waves, Grating

So far, in this course, our subject has been only one thing, essentially, a simple harmonic motion (SHM)! Note that the wave phenomenon described in terms of a sinusoidal wave is simply lots of SHMs perfectly coordinated, since each point in space where the sinusoidal wave occurs is simply doing a SHM. The best description of SHM and related motions (resonance, damped SHM, driven SHM, etc.) is in terms of *complex* numbers, *not* real numbers. Thus, we can say that all physics that we have described so far is best described with complex numbers. In fact, the earlier you know complex numbers, the better you will be. So, here we go—we introduce complex numbers, at last—as we are about to finish up the central subjects—SHM and waves—that we learned in this course. This lecture note re-summarizes a lot of things, almost all things, one might say, in the best mathematical formalism.

## 15.1 Complex number

Numbers are very useful and represent not only mundane things, like counting apples or money, but also they represent all kinds of useful abstract or imaginative notions, which are related to physical worlds in sophisticated ways. Complex number is a breath-taking invention. I think it would take extreme efforts not to fall in love with complex numbers; such efforts would be very ill-advised. One might even say that in some Platonic sense that complex numbers are more real than real numbers. Let us see why.

Consider a one dimensional motion, such as a SHM. Let us take the one dimen-

sional axis to be the $x$ axis. We like to extend the space two dimensions, $(x, y)$, where $y$ is a *purely imaginary* axis. In this view, the $xy$ plane is referred to as a **complex plane**, and the $y$ axis does *not* correspond to any physical real *spatial* dimension at all. However, mathematically speaking, an $xy$ plane as a complex plane is just the same as an $xy$ plane as a Cartesian coordinate system that represents a set of two dimensional vectors. It is often useful to go between these two *mathematical equivalent* views of a plane: a complex plane and a two dimensional vector space.

Let us consider a polar coordinate system, $r, \theta$, ($\theta$ is measured from the positive $x$ axis, with the counter clock wise direction defined as positive, using the common convention) using which we can express the Cartesian coordinates

$$x = r\cos\theta \tag{15.1}$$

$$y = r\sin\theta \tag{15.2}$$

Practically all first year students, who take the first year physics course, have been quite familiar with these formulae, and so I trust that you are, too!

Now, consider each point $(x, y)$ as a complex number[1], $z$.

$$z \equiv x + iy = r(\cos\theta + i\sin\theta) \qquad\qquad i \equiv \sqrt{-1} \tag{15.3}$$

By introducing a purely imaginary number, $i$, here we are considering the $xy$ plane as a set of complex numbers. Geometrically, $i$ corresponds to point $(0, 1)$.

Note that $i$ is defined as one square root of -1. The other square root is $-i$. That is, the solutions to the quadratic equation, $z^2 = -1$, are $z = \pm i$. **Any non-zero number that is a real number times $i$ is called a purely imaginary number.** So, all points on the $y$ axis except the origin correspond to purely imaginary numbers, and the $y$ axis is referred to as the imaginary axis.

The importance of the following **Euler's formula** cannot be over-emphasized.

$$e^{i\theta} \equiv \exp(i\theta) = \cos\theta + i\sin\theta \qquad \text{Euler's forumla} \tag{15.4}$$

$$e \equiv \lim_{\delta\to 0}(1+\delta)^{1/\delta} = 2.718... \tag{15.5}$$

Here, $e$ is an irrational number. It can be seen to represent how many million dollars you will have in your bank, if you save one million dollars and wait $N$ years, where $N = 1/\delta$, where $\delta$ is a very small yearly interest rate. For instance, if the interest rate is 1 % per year (arguably not *too* small, but small enough for the purpose of illustration here), then $\delta = 0.01$, and $N = 100$. After 100 years, you will have about 2.7 million dollars. Also, $e$ is the base of the natural logarithm.

[1] This notation of using $z$ for a complex number is standard. It has nothing to do with the $z$ axis of the usual Cartesian coordinate system in three dimensions.

The above Euler formula means that $\cos\theta$ is **the real part** of a simple exponential function, $e^{i\theta}$. And **the imaginary part** of it is $\sin\theta$. In other words, $e^{i\theta}$ **is a point on a unit circle in the complex plane, with angle** $\theta$ **measured from the positive** $x$ **axis**.

Using the above Euler formula, and the fact that $\cos(-\theta) = \cos\theta$ and $\sin(-\theta) = -\sin\theta$, we get the following useful formulae.

$$e^{-i\theta} = \cos\theta - i\sin\theta \qquad e^{i\theta}\text{, reflected} \tag{15.6}$$

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} \qquad \text{the real part of } e^{i\theta} \tag{15.7}$$

$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i} \qquad \text{the imaginary part of } e^{i\theta} \tag{15.8}$$

Now, the coordinate transformation formulae, Eqs. 15.1,15.2, can now be expressed compactly, using Euler formula,

$$z = x + iy = re^{i\theta} \tag{15.9}$$

Here, $r$ is referred to as the **magnitude/modulus** of $z$, $r \equiv |z|$, and $\theta$ is referred to as the **argument/angle/phase** of $z$.

While the exponent is a purely imaginary number[2], $e^{i\theta}$ is the same good old exponential function, algebraically speaking. Exponential functions are very pleasant to deal with, due to, e.g., the following properties, which are valid for any complex numbers $z$, $z_1$, $z_2$, $z'$.

$$e^{z_1+z_2} = e^{z_1}e^{z_2} \tag{15.10}$$

$$\left(e^{z_1}\right)^{z_2} = e^{z_1 z_2} \tag{15.11}$$

$$\frac{d}{dz}e^{z} = e^{z} \tag{15.12}$$

$$\int_{z_0}^{z} dz'\, e^{z'} = e^{z} - e^{z_0} \tag{15.13}$$

In other words, the good old exponential function has all its virtues preserved even in the domain of complex numbers. You should feel free to deal with the exponential function with the same ease that you have been dealing with in your calculus class, even when we extend its domain to the complex plane.

Lastly, note the following trivial, but very important, fact, which basically means that the complex number space corresponds to a two dimensional vector space.

[2]If we assume that $\theta$ is a real number. However, Euler formula remains valid, even if $\theta$ itself is a complex number.

The addition of two complex numbers correspond to adding two corresponding two dimensional vectors. The scaling of a complex number by a real number corresponds to the scaling of a corresponding two dimensional vector.

$$\begin{aligned} A_1 z_1 + A_2 z_2 &= (A_1 x_1 + A_2 x_2) + i(A_1 y_1 + A_2 y_2) \\ &\Leftrightarrow (A_1 x_1 + A_2 x_2, A_1 y_1 + A_2 y_2) \end{aligned} \tag{15.14}$$

if $A_1$ and $A_2$ are arbitrary real numbers, $z_1 = x_1 + iy_1 \Leftrightarrow (x_1, y_1)$, $z_2 = x_2 + iy_2 \Leftrightarrow (x_2, y_2)$, and the notation $(x, y)$ means the components of the two dimensional position vector, $\vec{r} \equiv x\hat{x} + y\hat{y}$.

Note that the scaling of a complex number by a negative real number, such as -1, corresponds to flipping the vector. So, even if the above box discusses adding only, it implicitly includes subtracting also.

## 15.2 Simple harmonic motion, revisited

Let us see why all of this is useful for dealing with a SHM. Consider the equation of motion (EOM)

$$\ddot{x} = -\omega^2 x, \qquad \text{SHM EOM} \tag{1.12}$$

$$\ddot{z} = -\omega^2 z. \qquad \text{SHM EOM, a complex number version} \tag{15.15}$$

The second equation is a complex number version of Eq. 2.11, where we used $z = x + iy$ (Eq. 15.3). Note that Eq. 15.15 means that $z$ is a function of $t$ such that its functional form does not change when we take its second derivative. Since an exponential function is just such a function (Eq. 15.12), i.e., since $e^{\alpha}$ is invariant upon differentiation with respect to $\alpha$, we can try a function of the form, $C_1 e^{C_0 t}$, as possible solution to Eq. 15.15, where $C_1$ and $C_0$ are *complex* numbers. Upon differentiating twice with respect to $t$, making use of the chain rule, we get

$$\ddot{z} = C_1 C_0^2 e^{C_0 t} = -\omega^2 z = -\omega^2 C_1 e^{C_0 t}.$$

If this equality is to hold for any non-trivial solution ($C_1 \neq 0$) at any arbitrary time, the following condition must be satisfied.

$$C_0^2 = -\omega^2.$$

Recall that, by convention, $\omega \geq 0$. In any case, $\omega^2$ is positive for a non-zero value of $\omega$. Thus, this equation for $C_0$ can be satisfied for nonzero $\omega$, only if $C_0$ is a purely

imaginary number. In fact, there are two solutions:

$$C_0 \;=\; \pm i\omega.$$

Therefore, we got *two* solutions to Eq. 15.15. Why did we? It is because the equation of motion is the second order differential equation[3]. Any linear combination of the two solutions is also a solution of Eq. 15.15, due to the linearity of the SHM equation of motion above. So, the *general solution* to the complex SHM equation of motion, Eq. 15.15, is

$$z(t) \;=\; C_1 e^{i\omega t} + C_2 e^{-i\omega t}. \qquad\qquad C_1, C_2\text{: complex constants}$$

Now that we solved the problem completely in the complex plane, we must go back to the real axis, since that is, after all, our ultimate goal; i.e. our goal is to obtain solutions to Eq. 2.11, *not* Eq. 15.15. Let us write $C = C_r + iC_i$ for both $C_1$ and $C_2$. Since $C_r$ (real part) and $C_i$ (imaginary part) are both real numbers, we get four real constants out of two complex constants. To get the general solution for $x$, the real part of $z$, we must compute the real part of

$$z(t) \;=\; (C_{1,r} + iC_{1,i})(\cos(\omega t) + i\sin(\omega t)) + (C_{2,r} + iC_{2,i})(\cos(\omega t) - i\sin(\omega t))$$

Expanding all terms, using $i^2 = -1$, we can collect all real terms to obtain $x(t)$. The result is

$$x(t) \;=\; (C_{1,r} + C_{2,r})\cos(\omega t) + (C_{2,i} - C_{1,i})\sin(\omega t)$$

Considering that all of these four constants are arbitrary real constants, we really have only *two* arbitrary real constants[4] in the general solution for $x(t)$:

$$x(t) \;=\; A_1\cos(\omega t) + A_2\sin(\omega t) \qquad\qquad A_1,\, A_2\text{: real constants} \tag{15.16}$$

Indeed, note that we just re-produced Eq. 1.8, the general solution for SHM!

## 15.3 Phasor

The phase of a wave is often visualized with an arrow. The arrow is taken to be a certain length (like unit length), while its angle of orientation relative to a reference direction (e.g., positive $x$ direction) is the key information for a phasor. Indeed, that angle is called "phase," which is the reason why such an arrow is called a phasor.

[3]By the same token any $n$-th order *linear* differential equation will give you $n$ frequencies, while some of them can turn out to be the same (a double root or a triple root, etc.).

[4]Also, the imaginary part $y(t)$, where $z = x + iy$, is governed by two arbitrary real constants. So, the total number of constants remain four.

Here, we can define phasor more precisely.

Looking back at the previous section, we realize the following point. Our goal was to obtain the general solution for $x(t)$. This has been accomplished by going to the complex plane and solving the problem there. However, was it really necessary to keep both complex solutions ($e^{i\omega t}$ and $e^{-i\omega t}$)? The answer is no. We could have kept only one of these two terms, multiplied by an arbitrary complex number, and we would have done just as well to obtain Eq. 15.16. For this reason, it is customary to choose either of the two complex functions, but not both, when we solve SHM problems in the complex space. Here, let us keep the $C_1 e^{i\omega t}$ term, and express $C_1$ as $C_1 = Ae^{i\phi}$, where $A > 0$ is the magnitude of $C_1$, and $\phi$ is the angle of $C_1$ measured from the positive $x$ axis (so we are applying Eq. 15.9 to $C_1$ with $A$ for $r$ and $\phi$ for $\theta$).

$$z(t) \;=\; Ae^{i\phi}e^{i\omega t} \;=\; A\,e^{i(\omega t+\phi)} \qquad \text{SHM described in the complex plane} \qquad (15.17)$$

A phasor is simply the visual representation of $z(t)$! $A$ is the length of the arrow. $\omega t + \phi$ is the phase angle. Note that when viewed in the complex plane, a SHM corresponds to a **uniform circular motion**, since the phase $\omega t + \phi$ changes at a constant rate $\omega$. So, the angular frequency is the angular velocity in the complex plane[5].

Taking the real part of the above expression for $z(t)$, we get

$$x(t) \;=\; A\cos\left(\omega t + \phi\right) \qquad\qquad A \geq 0 \qquad\qquad (1.6)$$

where $A$ and $\phi$ are identified as two real constants, required for the general solution to the second order real differential equation. We just re-produced Eq. 1.6!

## 15.4 Sinusoidal waves, revisited

Based on the Fourier theorem, we have been quite content with dealing almost exclusively with sinusoidal waves, since any other wave form can be expressed as a linear combination of sinusoidal waves of different frequencies and wavelengths.

Continuing to employ this useful practice, here, we will focus on plane waves propagating in the $x$ direction. In order to turn such a wave into a spherical wave, all that is necessary is to change $x$ to $r$, and $A$ to $A/r$ (or simply realizing that $A$ has the $1/r$ dependence). Lastly, what we consider is a traveling sinusoidal wave. A

[5]You might say that this complex plane may be called a "phase space" in this example. The concept of phase space as defined in advanced classical mechanics and statistical mechanics is related to such possible terminology.

standing sinusoidal wave can be formed by superposing two sinusoidal waves with opposite wave vectors ($k \to -k$; wave length stays the same $\lambda = 2\pi/|k|$, but the sense of direction is reversed when we change $k$ to $-k$; in general, $|k|$ is wave number, and $k$ is wave vector; cf., page 5 of LN 4).

Some thoughts about the notation may be helpful at this point. So far we have been using the expression $z = x + iy$, where $x$ and $y$ are coordinates in the complex plane, and $x$ describes the real SHM. For wave phenomena, we have been using $D$ as the displacement field, and it is this $D$ that describes the real SHM for a sinusoidal wave. So, in this section, we shall use the symbol $Z$ for the complex field, whose real part is $D$: $\mathrm{Re}\, Z(x,t) = D(x,t)$. Keeping in mind that $D(x,t)$ describes the motion of each particle in the wave medium at a fixed position $x$, it is really important not to mix up $x$ and $D$. In wave phenomena, $x$ is *not* a dynamical variable, but merely a label for a point in space (see previous lecture notes, esp. LNs 5 and 6). The complex plane is now the set of $Z(x,t)$ values, whose real values are $D(x,t)$.

For a plane wave, we can write

$$Z(x,t) \;=\; Ae^{i(kx-\omega t+\phi)} \tag{15.18}$$

which gives

$$D(x,t) \;\equiv\; \mathrm{Re}\, Z(x,t) \;=\; A\cos(kx-\omega t+\phi) \tag{15.19}$$

Note that this is a bit different from what we have been doing, following the textbook, since we have been using sine instead of cosine[6]. But, by adjusting $\phi$ ($\phi \to \phi - \pi/2$), we can turn cosine into sine (i.e., $\cos(X-\pi/2) = \sin X$, where $X = kx-\omega t+\phi$), and the two ways of writing down the general plane wave are therefore equivalent. Physically, the adjustment of the phase constant $\phi$ corresponds to shifting the origin of time. In any case, this equivalence was to be expected, given the equivalence of all three forms of SHM solutions, Eqs. 1.6, 1.7, and 1.8.

In *this* lecture note, we will use the cosine form for $D(x,t)$, as given above.

As the complex number representation of a plane wave, $Z(x,t)$ represents a **uniform circular motion**, also. This is not surprising, since $Z$ must represent a SHM. Note, however, that the angular velocity is now negative, since the time dependent term in phase is $-i\omega t$, not $i\omega t$. Thus, we have a clock wise motion, rather than a counter clock wise motion. This difference is just a matter of convention. Why use this convention, which is opposite to the sense of rotation of a single SHM (as

[6]Indeed, in the actual lecture, I used sine functions and $D = \mathrm{Im}\, Z$. The fact is that you can solve the problem in the complex domain, and when you need to find the real displacement you can project your solution onto *any* axis, such as the real axis, as we are doing here, or the imaginary axis, as we did in the lecture, or any other axis.

discussed in the previous section)? The answer lies in the fact that the total phase $kx - \omega t + \phi$ is most often considered at a fixed value of time, as a function of $x$, and $kx - \omega t + \phi$ is a nicer function to deal with than $-kx + \omega t + \phi_2$ for the wave that moves in the same direction. In fact, you will realize that for all interference and diffraction phenomena, we have been fixing time at the time value when we measure the intensity at the screen and considering the phases from various sources as a function of $x$ (or $r$ or $\vec{x}$, in general).

Namely, when the phase, $kx - \omega t + \phi$ is viewed as a function of $x$, we get larger phase angle, for a longer path ($x$). This is the reason why the convention is chosen such that as a function of time, we get a clockwise uniform circular motion (negative angular velocity), but as a function of path length, we get a anti-clockwise rotation, i.e., a positive rotation, of $Z(x,t)$, which we again call **phasor**.

Lastly, note a few points. As we have been pointing out in previous lectures, the waves that are relevant for describing interference or diffraction are spherical waves, not plane waves. So, we must understand that $A$ has a $1/r$ dependence, and is *not* a constant in those experiments. Also, in those experiments, $x$ must be replaced by $r$, in the phase ($kx \to kr$).

In addition, we generally assume that $A$ is the same for all spherical waves that emanate from any relevant source point (e.g., the two points corresponding the two slits, in the case of two very narrow slits, or any points within a single slit, in the case of a single slit). We also typically assume that $\phi$ is a constant that applies to any spherical wave. This would be true for a coherent light. If the $\phi$ value for each different point source is uncorrelated (random), that defines an incoherent light. Note that if two sources of light have a fixed difference in phase ($\phi_1 - \phi_2 = \pi/2$, e.g.), then the two sources are still perfectly coherent. Such a difference in phase constants can be translated to the *effective* difference in path lengths, $\Delta x_{eff}$, by the formula

$$k\Delta x_{eff} = \Delta\phi \tag{15.20}$$

$$\Delta x_{eff} = \frac{\Delta\phi}{k} = \frac{\Delta\phi}{2\pi}\lambda \tag{15.21}$$

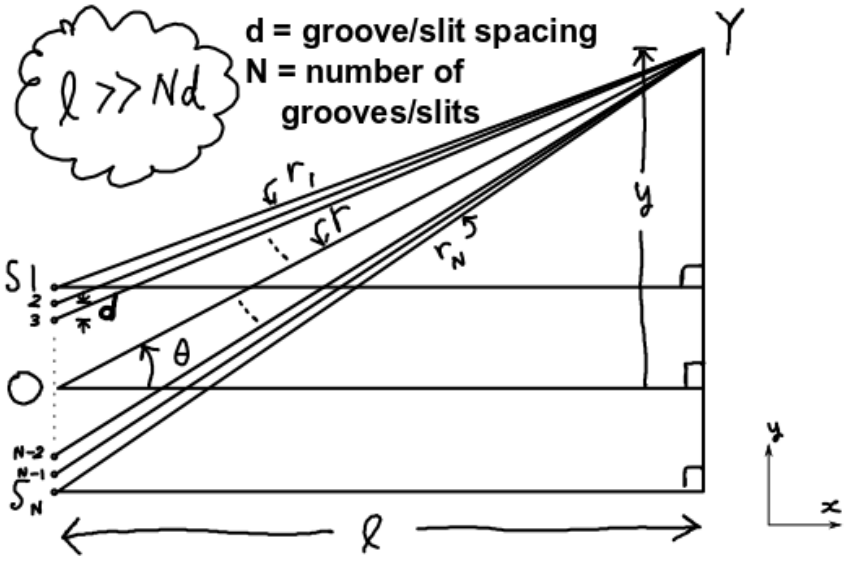
where $\Delta\phi = \phi_2 - \phi_1$. That is, if for some reason the two wave components have different phase constants then we can translate this difference to effective difference in path lengths, and forget about $\phi_1$ and $\phi_2$ as we try to figure out whether the interference is constructive or destructive. This is what has been going on with the thin film interference, where a $\pi$ phase shift causes $\phi$ to increase (or equivalently, decrease) by $\pi$ on reflection off of a medium with a higher index of refraction. Note that $\Delta\phi = 2\pi m$ with any integer $m$ can be considered as no change at all, while $\Delta\phi = \pi$ would correspond to a shift by half wave length. Likewise, $\Delta\phi = \pi/2$ corresponds to a shift of the wave pattern by a quarter wave length.

## 15.5 Diffraction grating

Let us consider a diffracting grating that is made by scratching many ($N$) "rulings" or "grooves" on a flat slab of highly reflective material, e.g. a gold coated glass. However, the theory in this section is valid for any value of positive integer value of $N$.

While the normal slab of materials will simply reflect, and refract, light, a diffraction grating is totally another matter. Each groove is very thin, and will scatter light, acting as a source of scattered light. Except for couple of differences, the physics of a diffraction grating is the same as the physics of interference of light from multiple slits. The differences include the following. First, a diffraction grating is a reflective device, not a transmissive device. Second, the zero order light for a diffraction grating includes not only light scattered from grooves, but also light reflected from the flat portion of the grating (normal specular reflection).

With these differences in mind, we will now discuss the interference of lights scattered by grooves. So, this mathematics is applicable to multiple slit problems, without any modification. Here is a diagram, where sources of light ($S_1$, $S_2$, $S_3$ etc.) and the observation point of light $y$ are clearly marked.




The main thing here is of course that there are now $N$ sources of light ($S_1$ through $S_N$). Whether these are grooves or slits, we assume that they are perpendicular to the paper, and are either very long (like for a grating groove) or very short. The $xy$ plane is assumed to bisect all the grooves or the slits, and our concern is only within the $xy$ plane.

Note above that the angle $\theta$ is measured from the origin, which is defined as the center of the grating groove pattern. We assume that $l \gg Nd$, so that all lights from $N$ sources are all parallel to the one another, to a good approximation. Noting Eq. 15.14, we see that **the superposition principle of $D(x,t)$ implies the superposition principle of $Z(x,t)$, and vice versa**. Thus, the total complex wave displacement field observed at position $y$ on screen is given by

$$\begin{aligned} Z(y,t) &= A\sum_{j=1}^{N} e^{i(kr_j-\omega t)} && \text{assume } \phi = 0 && (15.22)\\ &= Ae^{i(kr_1-\omega t)}\sum_{j=1}^{N} e^{ik(j-1)\Delta r} && \text{path length difference increment is } \Delta r \\ &= Ae^{i(kr_1-\omega t)}\,\frac{e^{ikN\Delta r}-1}{e^{ik\Delta r}-1} \\ \Delta r &\equiv d\sin\theta && \text{length difference of adjacent paths} && (15.23)\\ \delta &\equiv k\Delta r = kd\sin\theta && \text{phase shift for adjacent paths} && (15.24) \end{aligned}$$

$A$ (a real number) is the amplitude per groove/slit. In computing the sum of the second line, the geometric sum has been effected using the standard method. In terms of $\delta$, we get

$$\begin{aligned} Z(y,t) &= Ae^{i(kr_1-\omega t)}\,\frac{e^{iN\delta}-1}{e^{i\delta}-1} \\ &= Ae^{i(kr_1-\omega t)}\,\frac{e^{i\frac{N\delta}{2}}}{e^{i\frac{\delta}{2}}}\cdot\frac{e^{i\frac{N\delta}{2}}-e^{-i\frac{N\delta}{2}}}{e^{i\frac{\delta}{2}}-e^{-i\frac{\delta}{2}}} && \text{factor out } e^{i\frac{N\delta}{2}} \text{ and } e^{i\frac{\delta}{2}} \\ &= Ae^{i(kr_1-\omega t)}\,e^{i\frac{(N-1)\delta}{2}}\cdot\frac{\sin\frac{N\delta}{2}}{\sin\frac{\delta}{2}} && \text{using Eq. 15.8} \\ &= Ae^{i(kr-\omega t)}\cdot\frac{\sin\frac{N\delta}{2}}{\sin\frac{\delta}{2}} && r = r_1 + \frac{(N-1)\delta}{2k} \text{ (see diagram above)} && (15.25) \end{aligned}$$

Therefore, we can now obtain a grand result for $D(y,t)$, by taking the real part of this last result.

$$D(y,t) = A\cdot\frac{\sin\frac{N\delta}{2}}{\sin\frac{\delta}{2}}\cdot\cos(kr-\omega t) \qquad (15.26)$$

Here, we can see that the total wave $D$ has a very simple interpretation. It is a locally plane wave (and globally a spherical wave, $A \propto 1/r$, with amplitude given by $A\sin\frac{N\delta}{2}/\sin\frac{\delta}{2}$ instead of $A$. As the intensity is proportional to amplitude squared (LN 6), we get

$$I = I_R\cdot\frac{\sin^2\frac{N\delta}{2}}{\sin^2\frac{\delta}{2}} \qquad (15.27)$$

where $I_R$ is a constant, independent of $N$ in particular.

The function $\frac{\sin^2\frac{N\delta}{2}}{\sin^2\frac{\delta}{2}}$ vanishes when $N\delta = 2n\pi$, where $n$ is any integer except those values that make the denominator vanishes ($\delta = 2m\pi$; $n = Nm$ is an integer multiple of $N$). On the other hand, when $n$ *is* an integer multiple of $m$, the function $\frac{\sin^2\frac{N\delta}{2}}{\sin^2\frac{\delta}{2}}$ becomes $N^2$ (using L'Hospital's rule). Using Eq. 15.24, we get $\delta = kd\sin\theta = 2\pi d\sin\theta/\lambda$, and so this last condition can be written as

$$d\sin\theta = m\lambda \quad m = 0, \pm 1, \pm 2, ... \text{ (principal maxima, diffraction grating)} \qquad (15.28)$$

The physical origin of this diffraction grating principal maxima condition is very easy to figure out. Since the lights from adjacent sources are all in phase, if $d\sin\theta = m\lambda$, we get the maximum possible intensity. The maximum possible intensity, i.e. the intensity at $\theta$ values where the principle maxima condition is satisfied, is thus proportional to $A^2N^2$:

$$I = I_R N^2 (\equiv I_0). \qquad \text{intensity at principle maxima} \qquad (15.29)$$

From the above discussion, the intensity minimum (zero) condition above ($N\delta = 2n\pi$ except $n$ cannot be an integer multiple of $N$) can be written down also. It is

$$d\sin\theta = \left(m + \frac{n}{N}\right)\lambda \quad n = 1, ..., N-1; m = \text{integer (minimum intensity)} \qquad (15.30)$$

Therefore, each principal maximum point "owns" $N-1$ of these intensity minimum points. This also means that, in addition to the above principle maxima points, there must be local maxima points. These occur approximately when $N\delta$ is an odd integer times $\pi$, so that $\frac{\sin^2\frac{N\delta}{2}}{\sin^2\frac{\delta}{2}}$ becomes $\frac{1}{\sin^2\frac{\delta}{2}}$. This gives $(N-2)$ (low intensity) local maxima points, between two principal maxima points, as shown in Figure T35-18. These local maximum points pale in intensity in comparison to the principal maximum points, even when $N$ is as small as 3, but especially when $N$ becomes very large. Therefore, the width of a principle maximum peak is determined by the angular distance of the above minimum point for $n = 1$ and the principle maximum point, and thus this width scales as $1/N$. This crucial fact ensures that the more grooves one has

on a grating, the sharper the output light will be, in its spatial/angular profile[7]. . Diffraction grating is a main method to filter out a non-monochromatic light to generate a nearly monochromatic light, and this sharpness of the output beam is an important factor that determines the quality of a grating.

The fact that the intensity for principal maxima scales as $N^2$ and the width scales as $1/N$ is consistent with the energy conservation. The total energy per time for scattered light must scale as $N$. As this quantity is proportional to the area under the intensity curve, we see that it does indeed scale as $N$, when we multiply the maximum intensity (peak height) by the width.

**Phasors and interference conditions**

For the $N$ groove/slit problem, the principal maxima of intensity occur when all phasors corresponding to light paths emanating all grooves/slits are identical. The minimum intensity condition occurs when all $N$ phasors are spaced at equal non-zero angle intervals. I will leave it up to you to figure out the phasor conditions for weaker local maxima.

According to Eq. 15.28, different wavelength ($\lambda$) of light will correspond to different angle ($\theta$). Typically, for a research grade monochromator, only a very small angular window portion of the output beam is used to collect an output light that is nearly monochromatic. A large $N$ value would mean that each diffraction peak is very sharp in the angular distribution, and so a large intensity of light can be collected even within a very small angle window, $\Delta\theta$. This means a better light in terms of its monochromatic nature, since Eq. 15.28 means (by differentiating with respect to $\theta$)

$$\Delta\lambda \approx \frac{d\cos\theta}{m}\Delta\theta \qquad\qquad m = \pm 1, \pm 2, ... \qquad\qquad (15.31)$$

Note that $m = 0$ (zero-th order light) is useless for a diffraction grating not only because there is no way to resolve different wave lengths there, but also because of the intense specular reflections from the flat regions of the grating material.

Looking at the above equation, one can ask the following condition. What is the minimum wavelength difference that can be discerned at the $m$-th order if we apply

[7]Mathematically, one can show that the function $I/I_R$ becomes the "Dirac-delta function" times $2\pi N$, in the limit of large $N$. The Dirac-delta function is a function whose value is zero everywhere except at one point, where it is infinite, and whose integral is 1.

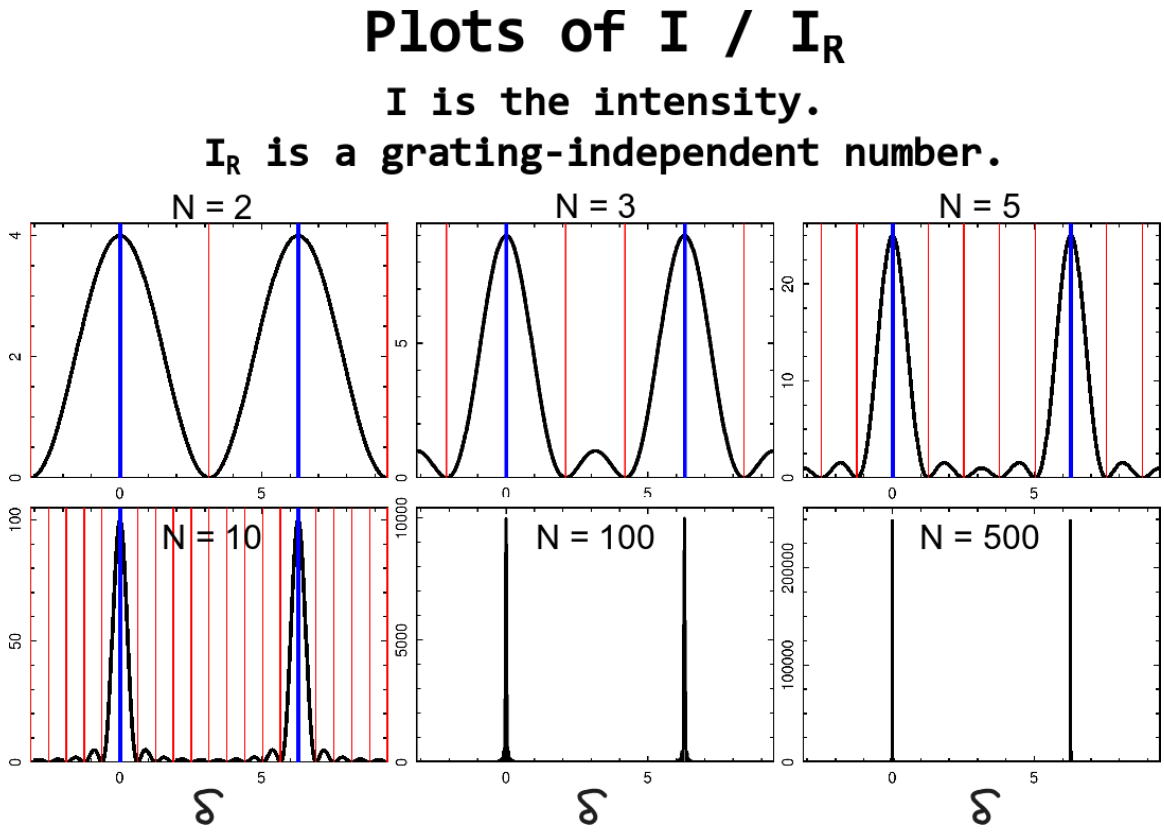



Figure 15.1: Plots of $I/I_R$, where $I$ is given in Eq. 15.27 as a function of $\delta$, the phase shift for adjacent slits/grooves. $N = 2$ corresponds to the double slit pattern. Note: (1) $I/I_R$ is a function of $\delta$ with periodicity $2\pi$, (2) there are $N - 1$ intensity minima (marked by red vertical lines for the first few plots) equally spaced between any two successive principal maxima (marked by blue vertical lines for the first few plots), (3) the intensities at local maxima that appear between principal maxima are very small compared to the intensity at the principal maxima, (4) the intensity at principal maxima is given by $N^2$, and (5) the peaks for principal maxima become very sharp (width $\propto 1/N$) as $N$ becomes large.

the Rayleigh condition[8], assuming that $N \gg 1$? Since the first minimum next to the principal maximum occurs when $d\sin\theta = (m + 1/N)\lambda$ (Eq. 15.30), we see that $d\cos\theta\Delta\theta = (1/N)\lambda$. And, therefore, we obtain

$$\Delta\lambda = \frac{\lambda}{Nm} \qquad \text{resolvable wave length difference at } m\text{-th order } (m > 0) \qquad (15.32)$$

## 15.6 Crystallography

Crystal is a phase of materials, where molecules are arranged with perfect periodicity. Bragg scattering shows that a crystal can be thought of as an infinite set of diffraction gratings. As the diagram below shows, a set of lattice planes cause a kind of regular reflections from each lattice plane, much like those that happen by grooves in a diffraction grating (except that in this crystalline case, the reflection occurs with the law of reflection being obeyed for each lattice plane). In a crystal, an infinite number of such diffraction gratings can be identified, because an infinite crystal can be thought of as an infinitely different sets of lattice planes. A set of lattice planes is defined as an equal-spacing stack of the identical planes of molecules, to which the entire crystal can be reduced. In the diagram below, only two different sets of lattice planes are shown, but a bit of thinking and plotting should convince you that there are indeed infinite possible sets of lattice planes. Indeed, when X-ray light is shone on a crystal, a spectacular diffraction pattern is observed due to the diffractions off of all possible diffraction gratings. Such a pattern provides crucial information used routinely to actually solve for the structure of a crystal, if a crystal is discovered for the first time.

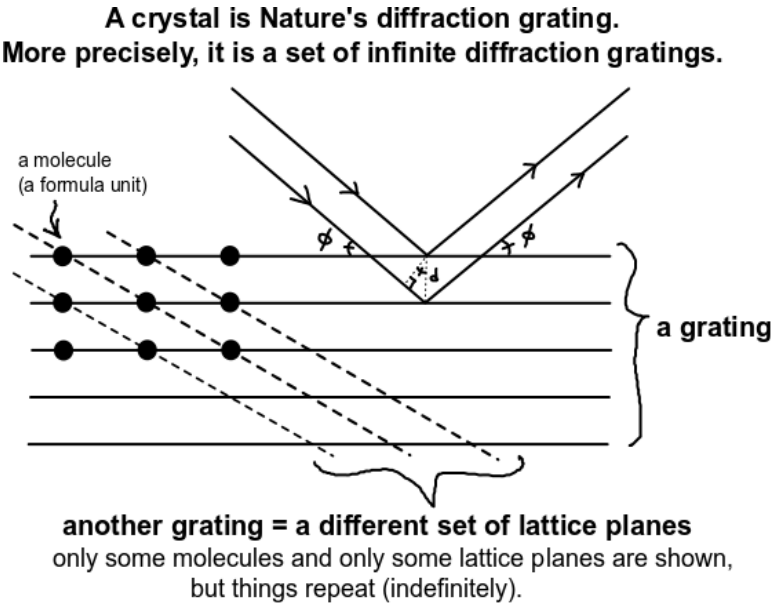
The Bragg diffraction condition is given by

$$2d\sin\phi = m\lambda \qquad\qquad m = 0, \pm1, \pm2, ... \qquad (15.33)$$

as can be deduced from the phase shift of the adjacent light paths depicted below (see diagram).

For a natural crystal, $d$ is on the order of 0.1 nm or 1 nm. And so, only hard X-ray light will diffract significantly. Therefore, in order to solve for a crystal structure of a "mystery" crystal, hard X-ray spectrometer must be used. Conversely, a very high quality crystal of a known matter (such as Si) can be used to produce a superb quality monochromatic hard X-ray beam, at synchrotron laboratories such as APS (advanced photon source, near Chicago), SSRL (Stanford synchrotron lightsource) and ALS (advanced light source; Berkeley).

[8]The first minimum of the image of one source coincides with the maximum of the image of another source, when the two images are barely resolvable.




## 15.7 Slit interference, general, including single slit

As already noted, the theory that we developed in Section 15.5 is not only good for diffraction gratings and crystals, but it is good for the general problem of $N$ slits.

Note that in developing the theory there, we assume that the width of each slit is very small so that we can assume that each slit is a source of a spherical wave. As such, if $N = 1$, the single slit case, then $d = 0$ by definition, and the theory of that section gives a trivial result—no diffraction—for a single slit.

However, this is not all. The theory of that section can be cleverly used for the single slit diffraction case with a *finite* slit width! How is this possible?

Let us see. If we have a single slit of finite width $D$, then we can divide up the length of the slit width $D$ into $N$ segments. If we do that, then, in the language of Section 15.5, $d = D/N$, as this is the distance between the two adjacent such segments. Then, by taking the limit $N \to \infty$, we should be able to obtain a convergent result that is valid for a single slit. (Note that here $D$ is used as the slit width. We will not discuss the displacement field in this section, so we can just avoid the confusion of the notation.)

As we found above, the intensity is proportional to $\frac{\sin^2 \frac{N\delta}{2}}{\sin^2 \frac{\delta}{2}}$, where $\delta = kd\sin\theta = kD\sin\theta/N$. Notice that as $N \to \infty$, $\delta \to 0$. However, $N\delta$ will remain finite. And so it

makes the most sense that we define

$$\beta \equiv kD\sin\theta \tag{15.34}$$

Noting that $\delta = \beta/N$ becomes very small, we get $\frac{\sin^2\frac{N\delta}{2}}{\sin^2\frac{\delta}{2}} \approx \frac{\sin^2\frac{\beta}{2}}{(\frac{\delta}{2})^2} = N^2\frac{\sin^2\frac{\beta}{2}}{\left(\frac{\beta}{2}\right)^2}$. Therefore, the intensity from a single slit is given by[9]

$$I = I_0\left(\frac{\sin\frac{\beta}{2}}{\frac{\beta}{2}}\right)^2 \qquad \text{intensity, single slit, } \beta = kD\sin\theta \tag{15.35}$$

where $I_0$ is the intensity for the zero-th order ($I_0 \equiv I(\beta = 0)$). This is the origin of the formula that is given as Eq. T35-8 in the book!

How about the more mundane double slit problem with negligible slit width? In this case, $I \propto \frac{\sin^2\frac{N\delta}{2}}{\sin^2\frac{\delta}{2}} = \frac{\sin^2\delta}{\sin^2\frac{\delta}{2}}$. Using $\sin\delta = 2\sin\frac{\delta}{2}\cos\frac{\delta}{2}$, we get $I \propto \cos^2\frac{\delta}{2}$:

$$I = I_0\cos^2\frac{\delta}{2} \qquad \text{intensity, double slit, } \delta \equiv kd\sin\theta \tag{15.36}$$

where $I_0$ is the maximum intensity.

What if in a multiple slit experiment, the width of each slit is $D$, which is non-negligible? Then, we must multiply the two intensity modulation factors together. For instance, for a double slit experiment where the width of each slit, $D$, is not negligible, we must use the following formula.

$$I = I_0\left(\frac{\sin\frac{\beta}{2}}{\frac{\beta}{2}}\right)^2\cos^2\frac{\delta}{2}. \qquad \text{double slit, each slit has width} = D \tag{15.37}$$

[9]The $N^2$ factor in the previous sentence is due to the coherence of light used. If you are worried about this $N^2$ dependence of the intensity, which diverges as $N \to \infty$, then you are quite right to do so. The following discussion is in order. Light from each segment has amplitude ($A$) that is proportional to $1/N$ (this factor is again due to the coherence of the beam), and so $A^2$ from each segment has a $1/N^2$ dependence, which cancels out the divergence due to $N^2$.