# The Double Slit Experiment and Quantum Mechanics*

Richard Rolleigh

2010

**Abstract**

The double slit experiment performed with particles and particle detectors is used to clearly demonstrate the nonclassical behavior of microscopic particles including the delayed choice experiment and causality issues. The realist and orthodox interpretations are presented with an explanation of why most physicists prefer the latter. The nature of a measurement is described precisely. The double slit experiment is extended to provide an experimental basis for the axioms necessary to develop quantum mechanics.

## 1 Introduction.

When we first studied quantum mechanics as college students in the 1960's, my colleagues and I were astounded by strange and weird concepts like wave particle duality, the uncertainty principle, nonexistence of trajectories, and collapse of the wave function. Today, sixty years later, those same concepts have become part of our culture through television shows like Star Trek, Sliders, Quantum Leap, and the NOVA series. However, I suspect that today's students find it almost as difficult as we did to accept a physical theory that contradicts so strongly the Newtonian mechanics that we learned intuitively as children.

We know that moving objects have trajectories because we have played baseball and soccer. We know that inanimate objects like baseballs have a well defined nature and that their behavior is totally

determined by initial conditions and the forces acting on them. All inanimate objects familiar to us obey Newton's laws. Yet the quantum physicists tell us that all these familiar things are made up of microscopic particles that do not obey Newton's laws at all. What rational person would believe this rubbish? In support of their ridiculous claims, the quantum physicists give us convoluted explanations of esoteric experiments and even more convoluted explanations of even more esoteric mathematics.

What is needed is a simple experiment that we can all understand and that unequivocally demonstrates the more disturbing properties of microscopic particles. It would also be nice if the experiment had actually been done and the results corroborated the strange predictions of quantum mechanics. Richard Feynman described just such an experiment in 1963: the double slit interference experiment that you studied in introductory physics.[1, 2, 3]

The double slit experiment (DSE) was first reported to the Royal Society of London by Thomas Young in 1803. Young did the experiment with light waves (photons) and measured the interference bands by observing the brightness of the light. Feynman proposed using modern technology to either do the experiment with electrons or do it with photons and detect individual photons. Clinton Davisson and Lester Germer had demonstrated electron diffraction in 1927, but this is one of those esoteric experiments referred to previously. The Feynman double slit experiment with individual electrons or photons is easier to understand and confronts us with inescapable evidence of the weirdness of microscopic particles. The experiment was not done in the form that Feynman described until 1972.[4] The experiment has since been repeated in a multitude of forms that include all the aspects described here.[5]

The first six sections of this article draw heavily on Reference 2.[6]

---

[1]Richard Feynman, **The Feynman Lectures on Physics**, (Addison wesley 1963), Volume III, Chapter I.

[2]Richard Feynman, **The Character of Physical Law**, (MIT 1965), Chapter 6.

[3]This note is intended for students of introductory Quantum Mechanics. However, if you have had no physics, you should find much of it interesting and comprehensible - you can just ignore the equations.

[4]Am J of Physics, **41**, p 639 - 644, 1972.

[5]The latest was in 2008. For exact references, see http://physicsworld.com/cws/article/indepth/9745 and http://en.wikipedia.org/wiki/Bell_test_experiments#Loopholes.

[6]Reference 2 uses everyday language instead of technical terms, and may be more accessible if you find my article too technical.

Sections 7 and 8 discuss causality issues, Section 9 explains what is meant by "measurement" in quantum mechanics, and Section 10 demonstrates how the axioms of Quantum mechanics follow from the results of the double slit experiment.

# 2 Intrinsic properties of particles that motivate the experiment.

Electrons and photons (and all other microscopic particles) exhibit two important properties that are crucial to the importance of this experiment. The first is that they all obey interference phenomena just like waves. You have probably observed interference of light waves passing through a double slit apparatus. It is firmly established experimentally that electrons behave the same way. In fact, double slit interference has been demonstrated with electrons,[7] neutrons,[8] atoms, [9] and buckyballs.[10]

The second important property that electrons, photons, and all other microscopic particles share is that they are always detected as individual particles, not as waves. When you did the Milikan oil drop experiment, you observed the motion of oil drops (or perhaps spheres made of teflon, plastic, or glass) containing a small discrete number of electrons. If any of those drops behaved as if it contained a fractional number of electrons, you were probably suffering from eyestrain. It is easy to believe that particles like electrons, protons, and neutrons are always detected as a whole particle and never as a piece of a particle. However, you may have imagined that you see light much as you hear sound, and since sound is clearly a wave, light must be too. You would be wrong: you see light very differently from how you hear sound. Your retina is covered with many tiny rods and cones, and when you see anything, individual photons are absorbed by these rods and cones. Each photon causes a discrete electrochemical excitation that is transmitted along the optical nerve. This is a very different process from that of your eardrum which moves as a unit due to air pressure variations spread over the entire eardrum.

[7] American Journal of Physics, Volume 42, pages 4-11, 1974
[8] Reviews of modern Physics, Volume 60, pages 1067 -, 1988
[9] Physical Review Letters, volume 66, page 2689 - , 1991
[10] Letters to Nature, Wave Particle Duality of $C_{60}$ molecules, Markus Arndt, 1999

Let me say this again to emphasize it. Your eyeball is covered with a large number of photon detectors. When you see something, each detector counts the number of photons it received and transmits that number to the brain. Some of the detectors (the cones) can detect the energy of the photons, and they transmit that value to the brain also (thus providing color vision). Your eyeball works much like the detector portion of a digital camera. You have never observed a light wave in your life, but you have added up the numbers of photons striking different places on your retina to create a diffraction pattern.

To me, the most convincing evidence that all particles, including photons, are always detected as individual and whole particles was observing the output of a particle detector on an oscilloscope. The output is a series of pulses. Each pulse represents the passage of one particle (a photon, an electron, or whatever) through the detector. You get the same effect with an old fashioned geiger counter: each click represents the passage of a particle through the detector. If you have never had the opportunity to observe this, you should at least read Wikipedia's article on particle detectors.

All microscopic particles, including photons, exhibit these two properties: they form interference patterns when passed through a double slit apparatus and they are detected individually as whole units. Never is a piece of one detected. The pictures in the referenced articles clearly demonstrate that individual particles are being detected as whole units, and that they form an interference pattern as more and more of them are detected. These experiments have been done with a great variety of microscopic particles, including photons. The results of the experiments have all been the same for all of the various particles. I will henceforth just use the generic word 'particle' and not specify whether I am speaking of an electron, photon, neutron, proton, buckyball, or whatever. They all behave the same in these experiments.

# 3 The double slit experiment with particles.

In the basic experiment, we pass a large number of particles through the double slit apparatus and let them strike detectors attached to the screen as illustrated in Figure 3. The coordinate system that we will use later is illustrated in the figure: the $x$ axis points up, the $y$

axis points out of page, and the $z$ axis points to the right. The origin is between the slits at the vertex of the angle $\theta$ rather than at the coordinate axes illustrated in the figure.

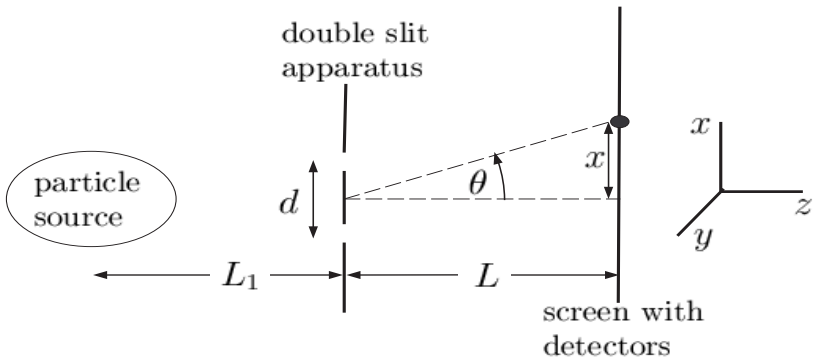


Figure 1: Double Slit Apparatus.

We will have to take care that our particles are all going in the same direction and all have the same wavelength. In other words, we need a columnated beam of particles that all have the same momentum because the de Broglia wavelength for all particles (including photons) is just Planck's constant $h$ divided by momentum $p$,

$$\lambda = h/p.$$

For photons, we can generate the particles with a mercury lamp and various filters and lenses just as you did when you performed the photoelectric experiment. For charged particles, we can use an apparatus similar to the electron gun that you used when you performed the Thompson $e/m$ experiment in introductory physics. The particles are all going in the same direction if $L_1 >> d$.

The screen on the right side of Figure 3 is covered with many closely spaced particle detectors whose positions are indicated by the variable $x$. For each experiment, we will pass a few billion particles through the slit apparatus and record the number of particles striking each detector. We will then make a histogram of the number of particles arriving at each detector as a function of detector position.

First we close the lower slit requiring all the particles to pass through the upper slit. The histogram we observe is illustrated in figure 2. This is the same as the single slit diffraction curve produced

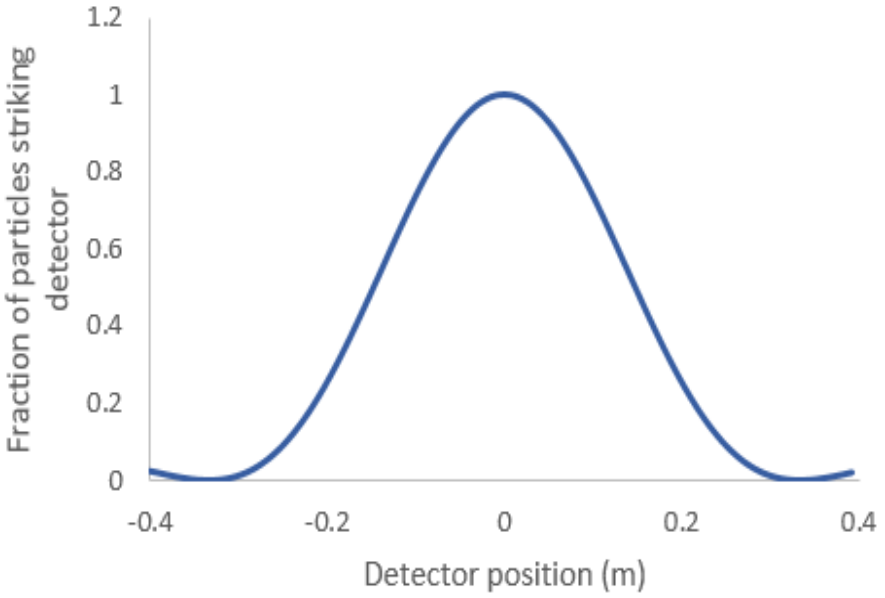


Figure 2: Single slit histogram

by monochromatic waves that pass through a single slit that is 12 times as wide as the wavelength and then strike a screen four meters away. We could obtain this same single slit pattern by either using photons with wavelength 550 nm (green light) and a slit width of 6.6 $\mu$m or electrons accelerated through a potential of 10 Kv and a slit width of .147 nm. The precise expression for single slit diffraction is

$$I(\theta) = I_{\max}\left(\frac{\sin\alpha}{\alpha}\right)^2, \tag{1}$$

where $\theta = \tan^{-1}(z/L)$ is the angle $\theta$ in Figure 3, $I(\theta)$ is the intensity at the angle $\theta$, $I_{\max}$ is the maximum intensity at $\theta = 0$, $\alpha$ is

$$\alpha = \frac{\pi a \sin\theta}{\lambda},$$

$a$ is the width of the slit, $\lambda$ is the wavelength of the monochromatic light or the de Broglie wavelength of the particle (if it has mass), and $L$ is indicated in Figure 3. Derivations and explanations of Equation 1 can be found in most introductory physics texts. Another source is

<http://en.wikipedia.org/wiki/Fraunhofer_diffraction_(mathematics)>.

Of course we could close the upper instead of the lower slit thereby forcing the particles to go through the lower slit. The result is exactly the same except the pattern is displaced downward by the distance between the slits. That distance is less than .1 mm so we can't tell the difference in the curves.

When we open both slits so the particles can go through either slit, we see something entirely new. Figure 3 illustrates the histogram we observe. Monochromatic waves passing through two slits separated

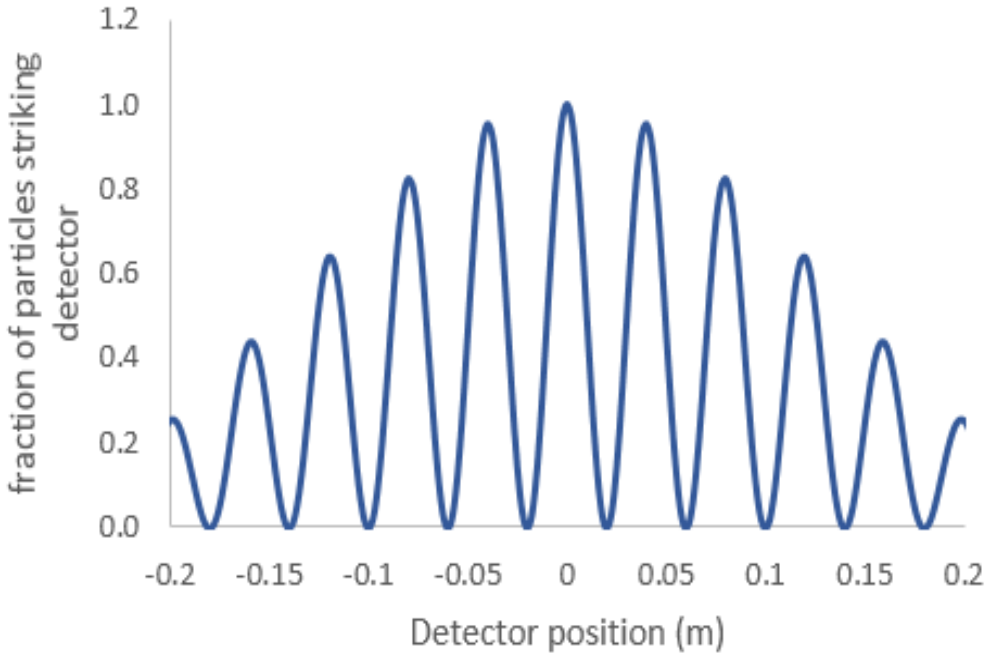


Figure 3: Double Slit histogram

by 100 times their wavelength would produce the same pattern on a screen four meters from the two slits. We could obtain this same double slit pattern by either using photons with wavelength 550 nm (green light) and a slit separation of 55 $\mu$m or electrons accelerated through a potential of 10 Kv and a slit separation of 1.23 nm. The

precise expression for the double slit interference curve is

$$I = I_{\max} \cos^2 \left( \frac{\pi d \sin\theta}{\lambda} \right) \left[ \frac{\sin\left(\pi a \sin\theta/\lambda\right)}{\pi a \sin\theta/\lambda} \right]^2, \qquad (2)$$

where $d$ is the slit separation, $a$ is the slit width, $\theta$ is the angle in Figure 3, $\lambda = h/p$, $h$ is Planck's constant, and $p$ is the momentum of the particle. Derivations and explanations of this expression can be found in most introductory physics texts. Perhaps a more convenient reference is `http://en.wikipedia.org/wiki/Double-slit_experiment`.

We see that if we force the particles to go through only one slit, we obtain a single slit pattern. If we allow the particles to go through both slits, we obtain a double slit pattern.

It is important to note that the shape of the double slit pattern depends on the distance between the slits. If you increase that distance, the interference maxima get closer together. The only rational interpretation of this is that in order for the particles to form a double slit pattern, either each particle must interact with both slits or some particles pass through the upper slit and some pass through the lower slit, and the particles then interact with each other to form the double slit pattern. The second possibility will be discredited by the next experiment.

# 4 The double slit experiment with one particle at a time.

In order to test the conjecture that some of the particles pass through the top slit and some pass through the bottom slit, and then they interact with each other to form the interface pattern, we do the experiment with only one particle at a time passing through the double slit apparatus. If the particles had to interact with each other to produce a double slit pattern, then passing one particle at a time through the apparatus would destroy the pattern. However, we find that even if we pass only one particle at a time through the apparatus, we still get the two slit interference pattern. This was verified by the experiments reported in references 3 and 4.

Up to this point the particles behave just like classical sound waves except for the way they are detected. If you close one slit, each particle goes through the other slit just as sound waves would. If you open both slits, each particle interacts with both slits just like sound waves. With

sufficient time, enough particles will accumulate to form an double slit pattern just like sound waves. The only feature that distinguishes particles from sound waves so far is that only one detector at a time on the screen detects a particle. If we were using sound waves, all the detectors located in bright fringes would fire at the same time. We cannot turn down the amplitude of the sound wave until only one quantum of sound energy at a time passes the slits and strikes the screen because sound wave energy is not quantized.

Since each particle interacts with both slits, each particle's energy must get divided so that some goes through each slit. We try to detect a particle going through both slits at the same time in the next experiment.

# 5 Detect which slit.

The detectors on the screen in Figure 3 probably entirely absorb the particle just like the sensors in your eye absorb photons. For this experiment, we need a detector that will allow the particle to pass through it while recording its passage. In other words, we need a detector that absorbs some but not all of the particle's energy. Actually, most detectors used in the laboratory do just that. If you study the design of particle detectors in Wikipedia, you will understand that by adjusting the length of the detector along the direction of the particle's motion, you can adjust the amount of energy absorbed from zero to 100 per cent. Of course as you reduce the amount of energy absorbed, you decrease the probability that the particle will be detected.

In order to detect how much of each particle goes through each slit, we place detectors after each slit. If we make the slit detectors very sensitive so that they detect everything that goes through their respective slit, we observe that each particle goes through one slit or the other. No particles divide their energy between the slits. Clearly, the particles are not interacting with both slits. How can they then make a double slit pattern? Well, **they don't**! When we turned on the slit detectors and formed a histogram from the outputs of the detectors on the screen, we got the superposition of two single slit patterns. These patterns are so much alike that their sum looks just like the single slit pattern in Figure 2. It seems that detecting which slit they go through forces them to go through one slit or the other and also forces them to produce two single slit patterns instead

of a double slit pattern. Although this experimental result may be intuitively disturbing, it is nice that it agrees with the predictions of quantum mechanics.

This latest particle behavior is quite distinct from that of sound waves. If we measured how much of sound wave energy went through each slit, we would find that the sound wave splits its energy equally between the slits and still forms a double slit pattern. Particles on the other hand, choose one slit or the other (when we measure which slit) and form a single slit pattern.

This experiment has been done with photons[11] and with atoms[12]. The method they used to determine which slit the particle traversed involved an entangled photon and measurements made on it. We may have time to discuss these experiments in more detail after we have studied entangled states. Despite the esoteric nature of these experiments, they fully corroborate the results I have described in this section.

# 6 Weakly detect which slit.

An incorrigible sceptic might argue that in the previous experiment we destroyed the double slit pattern because our slit detectors were too sensitive. They interfered with the particles too much. The obvious solution is to make the detectors absorb less of the particles' energy and thus be less sensitive. If we do this, the slit detectors will miss some of the particles that eventually are detected by the screen detectors. Our data will fall into three classes:

- Particles that are detected traversing the upper slit and then striking the screen,
- particles that are detected traversing the lower slit and then striking the screen, and
- particles that are not detected by either slit detector yet we know they were there because they were detected at the screen.

The percentage of particles in the third group will increase if we decrease the sensitivity of the slit detectors. If we form histograms of each class, the first two classes will form single slit patterns while the third class will form a double slit pattern.

[11] Phys Rev letters, **84**, pp 1 - 5, January, 2000,
[12] Phys Rev Letters, **81**, pp 5705 - 5709, December, 1998

There is no way to escape the conclusion that we determine how the particles traverse the double slit apparatus by what we choose to measure or not measure. If we measure which slit, the particles accommodate and go through one slit or the other and then strike the screen at places that form a single slit pattern. If we do not measure which slit, the particles strike the screen at places that form a double slit pattern. Since the double slit pattern depends on the distance between the slits, the particles must interact with both slits if we do not detect which slit they traverse.

I hope you are not uncomfortable with all this because it will get worse in the next section.

# 7 Delayed Choice Experiment.

The previous experiment tells us that turning on the slit detectors forces the particles to traverse only one slit and turning off the slit detectors forces the particles to interact with both slits. The detectors' settings (on or off) determine how the particle interacts with the slits. What if the decision to turn the slit detectors on or off is made after the particle has already passed through the double slit apparatus? This is not too hard to do with the accurate timing available today and the existence of particle storage devices that can hold a particle isolated from all influence for several ns.

We place a particle storage device between each slit and its corresponding slit detector as illustrated in Figure 4. For photons, the storage device is just an optical fiber loop, and for charged particles it is just a magnetic field that causes the particle to go in circles. Suppose the storage devices will delay the particles for 10 ns, and we randomly change the settings on the detectors every 8 ns. The particles have already interacted with the slits and are in the storage device when the detectors' settings are determined. *But the detectors' settings determine how the particle interacted with the slits, before the detectors were set.* In other words, the act of setting the detectors controls something that happened in the past: how the particle interacted with the slits. Although this delayed choice experiment has not been done exactly as described here, slight variations have been done a number of times,[13] always with the results described here.

[13]Science, **315**, no, 5814, pp 966 - 968, (2007) and references 11 and 12 here.

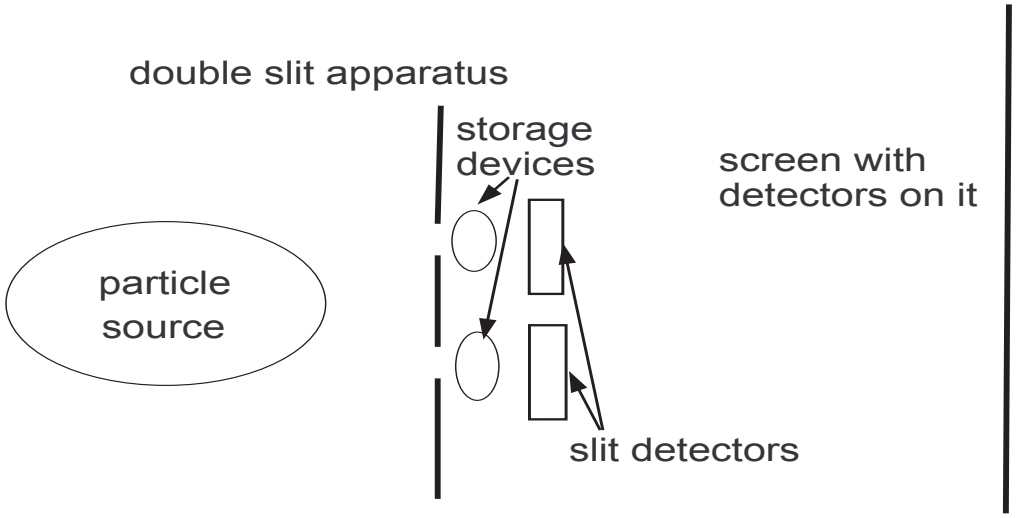


Figure 4: Double Slit with storage devices and slit detectors

# 8 Interpretations.

The only credible interpretation of the experimental results is that the act of measurement seems to influence the behavior of the particle, and that this influence can go backwards in time. There are a plethora of philosophical schemes to explain this strange behavior, but physicists have reduced them to two competing paradigms: realist and orthodox.[14] The old school name for these interpretations are the hidden variables and Copenhagen interpretations respectively, and you will see these names in much of the older literature (Copenhagen equals orthodox and hidden variables equals realist). You should be cautious in your usage of the terms 'realist' and 'realism' because they are widely used in philosophy, art, literature, and politics and they mean different things to different people.

[14]For example, see David Griffiths. *Introduction to Quantm Mechanics*, page 3-4. Prentice Hall, second edition 2005.

How do the realist and orthodox paradigms interpret the experiments we have been discussing? The realist would insist that the path of the particle through the slits (whether it went through only one slit or interacted with both slits) was real and had a precise value before the particle entered the storage device. The realist would also have to conclude that at least for some of the particles, the path through the slits was changed when the particle passed through the slit detector *after passing through the slits*. The action of the slit detectors exerted an action backwards in time that changed the value of a physical property in the past. We physicists have a strong aversion to changing the past. In other words, we have a strong belief in causality. By causality, I mean that if a physical property had a value yesterday, then there is nothing you can do today to change what its value was yesterday.

The orthodox position on the other hand is that the path of the particle through the slits is never real even if the particle is detected by one of the slit detectors. When you detect a particle in the top slit detector, what is real is the localization of that particle in the top slit detector at that time. Although that reality is consistent with the particle having gone through the top slit and not interacting with the bottom slit, to conclude that the particle was really in the top slit at an earlier moment of time is more than most orthodox adherents would claim. They would be more likely to say that until the particle is detected by either of the slit detectors or by a detector on the screen, it has the potential to land anywhere on the screen. If it is detected by a slit detector, then the probabilities of where it will land on the screen are modified by that detection.

The orthodox position is that even though the particle was measured by a slit detector, and the only way it could have gotten to the slit detector was through the slit, this does not require that the particle was ever really in the slit at any time. This position may seem to be evasive, but there are well established experimental results that demonstrate this very thing. I am speaking of the tunneling of particles through potential barriers that require more energy than the particle has. This phenomena has been well known for so long that tunneling diodes and tunneling electron microscopes are based on it. The particle clearly moves from one side of the barrier to the other because it is detected first on one side, then on the other. However, it can't ever be in the barrier without violating conservation of energy.

Some people will argue that the orthodox interpretation claims that the detection of a particle in one of the slit detectors makes

the path of the particle through the slits real retroactively in time. Therefore, they conclude that the orthodox interpretation requires that reality be created in the past and that this is just as large a violation of causality as the realist position that requires that reality be changed in the past. However, the previous two paragraphs expose the fallacy of this argument. The orthodox position neither creates nor changes anything in the past because it claims there is nothing real in the past except what was measured in the past. Although what I have just stated is true, it will have to be clarified when we consider entangled particles and instantaneous creation at a distance (nonlocal creation).[15]

# 9 Measurement.

I have shown experimentally that the process of measurement changes the state of the system, and I have argued that it either changes the past, projects reality into the past, or ultimately defines what is real. Yet I have not provided a precise definition of what constitutes a measurement. That was rather sloppy of me wasn't it. Let me remedy the situation.

I think Niels Bohr said it best when he said that a measurement is an 'irreversible magnification'. You will understand this better if you study the operation of particle detectors. The basic unit of charge is 1.6 $(10)^{-19}$ C (actually one third of that if you consider quarks). We just can't measure this small a charge. However, if any particle that produces an electromagnetic field (this includes uncharged particles like photons and neutrons) passes through some types of matter (semiconductors and gases), then it will transfer small fractions of its energy to many electrons and raise them to the conduction band in a semiconductor or free them from the gas molecules in a gas. If there is a large accelerating potential present, these electrons gain tremendous energy from the external field, and they will liberate more electrons. This produces an avalanche effect. The result is that the single particle being detected produces a pulse of many electrons whose combined charges can be detected. This is obviously an irreversible amplification

[15]It is true that the orthodox interpretation requires nonlocal creation of reality. It follows that different observers will disagree on the order in time in which measurements were made. Consequently, they will disagree on which measurement actually created the reality. But neither observer will observe creation going backwards in time.

just like a snow avalanche is. When a measurement occurs, entropy increases, disorder increases, energy moves from high concentration to low concentration, and the measurement can't be undone.

I like the idea that the only things that are real are those things that can't be undone. If we could go back in time and change reality, it wouldn't be very real would it? I also like the way that the orthodox interpretation of quantum mechanics integrates so well with the second law of thermodynamics.

# 10 Impact on the theory.

How do we use these experiments to guide us in the construction of a theory of microscopic particles? Any useful theory predicts things, so we must first decide what properties of microscopic particles are predictable. For something to be predictable, it must be a consistent measurement result. The positions at which individual particles land on the screen are not consistent: each particle could land in any bright fringe. Positions are not predictable. What is consistent is the probability of each particle's landing at any position, i. e. the probability density function (pdf) of each particle's position. The pdf of position is just the double slit interference pattern illustrated in Figure 3. It is reproduced any time you repeat the experiment and it is predicted by Equation 2. We will find that all observables[16] exhibit this behavior in all experiments with microscopic variables: specific outcomes are not consistent but the probabilities of all possible outcomes are. The only time a specific outcome is predictable is when a measurement is performed, a specific value is obtained, and then the identical measurement is repeated on the same system before it has time to interact with anything.[17] In this case, the same result will be obtained the second time.

The actual value of an observable (position, momentum, etc.) is not predictable because identical measurements of the observable on identically prepared systems produce different results. The fact that

[16] An observable is a measurable physical property such as position or momentum.

[17] Note that position can not qualify for this special case of consistency because you can not obtain a specific value from a position measurement device. All detectors are finite in size, and you can only detect that the particle passed through the detector, not exactly where it passed through. The only observables that can be specifically determined are those that are quantized like energy and angular momentum.

the actual value of an observable is not predictable arises from experiment, not theory. How will this make quantum mechanics different from classical mechanics? In classical mechanics, the values of all observables are predicted as functions of time. Time is the only independent variable and all the observables are dependent variables in classical mechanics. Prediction of the observables as functions of time is the program of classical mechanics. What is the program of quantum mechanics? Quantum mechanics predicts the probability distributions of all the physical variables. In addition to time, position is also an independent variable. The dependent variables are the probability distributions.

If we return to the double slit experiment without slit detectors, we see that the probability of any one particle striking at $x$ on the screen is predicted quite accurately by Equation 2. We must design our quantum theory so that it reproduces Equation 2 as the probability distribution for the positions of where the particles will strike the screen.

It is shown in Appendix A that Equation 2 is just the magnitude squared of the solution $\Psi(\boldsymbol{r}, t)$ to the Classical Wave equation that matches the boundary conditions imposed by the slit and the screen. The Classical Wave equation is written as Equation 10 in Appendix A, and is reproduced here for convenience,

$$\boldsymbol{\nabla}^2\Psi - \frac{1}{c^2}\frac{\partial^2\Psi}{\partial t^2} = 0, \tag{10}$$

where $c$ is the wave's phase speed and $\boldsymbol{\nabla}^2$ is the Laplacian.

Since the square of the solution to Equation 10 predicts the experimentally observed pdf, it seems reasonable to base our theory on the following two ideas:

- for every particle, there exists a wave function that is a solution of the Classical Wave Equation, Equation 10, that also meets the boundary conditions imposed by how the particles were prepared, and
- the probability density function of the particle's position is the magnitude squared of the particle's wave function.

There are three problems with this proposed theory. They are

- Equation 10 does not conserve probability,

$$\frac{d}{dt}\int_{-\infty}^{\infty} (\Psi^*\Psi)\, dx \neq 0.$$

The integral on the left hand side is the probability that the particle is somewhere. We can normalize $\Psi$ so the integral is one today, but since its time derivative is non-zero, it may be two tomorrow. What does this mean? The only reasonable interpretation is that there are two particles tomorrow. There is strong experimental evidence that electrons, protons, and all other fermions are conserved. Any theory that does not conserve probability cannot describe these particles. Incidentally, photons are not conserved so they can be and in fact are described by Equation 10.

- Equation 10 does not include the potential energy of the particle. We know that the potential energy at a point must influence the probability that the particle will be found at that point. For example, we do not expect to find a particle in a region where the potential energy is larger than the total energy of the particle. Also, it should be very likely to find a particle in regions where it moves slowly (because it hangs out there a lot). These will be regions in which the potential is only slightly less than the total energy of the particle.
- The proposed theory is incomplete. It only predicts the probability distribution of position. What about momentum, angular momentum, energy, and all those other interesting physical properties?[18]

Before proceeding, I want to emphasize that photons are not conserved so their wave function does actually satisfy the Classical Wave Equation. Furthermore, the concept of potential energy is meaningless[19] for photons as they have no mass and always travel at the speed of light. The Classical Wave equation is the proper wave equation for photon wave functions. On the other hand, the Classical Wave Equation is not suitable for fermions such as electrons, protons, neutrons, and neutrinos because it does not conserve probability and fermions are conserved. The rest of this section is devoted to obtaining a wave equation for fermions not photons. Appropriately, we will henceforth

[18]I personally would be very interested in the probability distribution of charm when choosing a particle with which to interact.

[19]Some authors use gravitational potential energy of photons to explain gravitational redshift, but it is unnecessary to associate potential energy to the photon to provide a rigorous explanation based on energy conservation. In any event, we are not attempting a quantum theory of gravitation here.

refer to our particles as fermions rather than as just particles.

Let us first tackle the conservation of probability problem. It is easy to see that Equation 10 does not conserve probability by examining the solution for a standing wave on a string of length $a$,

$$\Psi(x,t) = \sin(\pi x/a)\cos(\pi ct/a).$$

Substitution into Equation 10 verifies that this is a solution, and it is obvious that

$$\int_0^a \Psi^*(x,t)\Psi(x,t)\,dx \;= \cos^2(\pi ct/a)\int_0^a \sin^2(\pi x/a)\,dx \;= \frac{a}{2}\cos^2(\pi ct/a)$$

depends on time. The time dependence does not cancel out when we form $\psi^*(x)\psi(x)$ because both $\mathrm{e}^{i\omega t}$ and $\mathrm{e}^{-i\omega t}$ are solutions and the general solution is a superposition of them. In our case,

$$\Psi(x,t) = \sin(\pi x/a)\cos(\pi ct/a) = \sin(\pi x/a)(\mathrm{e}^{i\omega t} + \mathrm{e}^{-i\omega t})/2.$$

If the solution were restricted to

$$\Psi(x,t) = \mathrm{e}^{i\omega t}\sin(\pi x/a),$$

then the time dependence would cancel in $\Psi^*(x,t)\Psi(x,t)$. However, the second time derivative in Equation 10 ensures that if $\mathrm{e}^{-i\omega t}\psi(x)$ is a solution, then $\mathrm{e}^{i\omega t}\psi(x)$ is also a solution, thereby destroying conservation of probability.

We must eliminate the second time derivative to conserve probability. However, we must do this in a way that retains the solution we used in Appendix A to derive the observed pdf in Equation 2. That solution is given by Equations 11 and 12 in Appendix A, which I reproduce here for convenience,

$$\Psi(\boldsymbol{r},t) = \mathrm{e}^{-i\omega t}\psi(\boldsymbol{r}), \tag{11}$$

where $\psi$ satisfies Helmholtz's Equation,

$$\left(\boldsymbol{\nabla}^2 + k^2\right)\psi(\boldsymbol{r}) = 0. \tag{12}$$

The minimum change we can make to Equation 10 that preserves probability is to replace the second time derivative with a first time derivative multiplied by an arbitrary constant,

$$\boldsymbol{\nabla}^2\Psi(\boldsymbol{r},t) - \gamma\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t) = 0.$$

Substituting $\mathrm{e}^{-i\omega t}\psi(\boldsymbol{r})$ into this trial equation produces

$$\boldsymbol{\nabla}^2\psi(\boldsymbol{r}) + i\omega\gamma\psi(\boldsymbol{r}) = 0.$$

Comparison with Helmholtz's equation (Equation 12) reveals that if we choose

$$\gamma = -ik^2/\omega,$$

then our trial wave equation does reduce to Helmholtz's Equation as desired when $\Psi = \mathrm{e}^{-i\omega t}\psi(\boldsymbol{r})$. Replacing $\gamma$ with $-ik^2/\omega$, our trial wave equation becomes

$$\boldsymbol{\nabla}^2\Psi(\boldsymbol{r},t) + \frac{ik^2}{\omega}\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t) = 0. \tag{3}$$

This trial wave equation has limited usefulness until we determine the meaning of $k$ and $\omega$. Substituting the function,

$$\Psi(x,t) = \mathrm{e}^{i(kx-\omega t)},$$

into Equation 3 demonstrates that it is a solution. This function also describes a plane wave moving in the $x$ direction. If we set $t = 0$ and plot the real part of this solution, $\cos kx$, we see that it repeats itself every time that $x$ changes by $2\pi/k$. The wavelength $\lambda$ of a wave is defined as the change in position required to cause a complete cycle of the wave function, Therefore, we see that

$$\lambda = 2\pi/k \;\text{ or }\; k = 2\pi/\lambda.$$

If we set $x = 0$ and plot the real part, $\cos(-\omega t) = \cos\omega t$, we see that the time required for it to repeat itself is $\Delta t = 2\pi/\omega$. This is the period so

$$\omega = 2\pi/\Delta t = 2\pi f,$$

where $f$ is the frequency and is the inverse of the period. Knowing how $k$ and $\omega$ are related to the wavelength $\lambda$ and the frequency $f$ is nice, but it is not useful unless we can measure $\lambda$ and $f$ or relate them to other physical observables.

We can use the double slit experiment to relate $\lambda$ and wave number $k$ to energy. If the particle is charged like an electron or proton, then its energy is just the potential we use to accelerate it times its charge. The wavelength is the same $\lambda$ that appears in Equation 2 and determines

where the interference maxima appear. It is clear from Equation 2 that the change in $\sin\theta$ between interference maxima is

$$\Delta \sin\theta = \lambda/d,$$

so

$$\lambda = d\Delta \sin\theta.$$

The distance $d$ between the slits and the distance $L$ from the slits to the detector are designed into the experiment. The distance $\Delta x$ between interference maxima can be measured directly. In Figure 3, that distance is .04 m. Since the $x$ values are much smaller that $L$ in our experiments, we can use the small angle approximation, $\sin\theta \doteq \tan\theta$, so

$$\Delta \sin\theta \doteq \Delta \tan\theta = \frac{\Delta x}{L} = \frac{.04}{4} = .01,$$

in our experiment. For electrons, we used a slit separation of $d = 1.23$ nm, so the wavelength of the electrons was

$$\lambda = d\frac{\Delta x}{L} = .0123 \text{ nm}.$$

When we vary the energy and measure $\lambda$ as described in the preceding paragraph, we find experimentally that they are related by

$$E = \frac{4\pi^2\hbar^2}{2m\lambda^2} = \frac{\hbar^2 k^2}{2m}. \tag{4}$$

I emphasize that this is an experimental result from the double slit experiment described here. Of course electron microscopes built commercially seventy years ago were based on the same relationship.

We are limiting our theory here to nonrelativistic fermions with nonzero rest mass (like electrons and protons). Incidentally, all known fermions have nonzero rest mass. We know from classical mechanics that in the absence of potential energy, the energy of these particles are related to their momentum $p$ by

$$E = \frac{p^2}{2m}.$$

Comparison to Equation 4 reveals that the momentum $p$ is

$$p = \hbar k. \tag{5}$$

For nonrelativistic fermions, the velocity is just the momentum divided by the mass,

$$v = p/m = \hbar k/m.$$

This can actually be checked with the double slit experiment by measuring the time between when a detector at the slit detects the particle and when a detector at the screen 4 m away detects the particle. The particle's velocity is just 4 m/(the time).

It is shown in Appendix C that the velocity of the particle is the group velocity $v_g$ of the wave packet describing the particle, and this velocity is given in given by Equation 25,

$$v = v_g = \frac{d\omega}{dk}.$$

Combining the last two equations, we have

$$\frac{d\omega}{dk} = \hbar k/m.$$

Integrating this last equation, we get

$$\omega = \hbar k^2/2m + C,$$

where $C$ is an integration constant. Multiplying by $\hbar$ produces

$$\hbar\omega = \frac{\hbar^2 k^2}{2m} + \hbar C = E + \hbar C.$$

We cannot measure absolute energy, we can only measure the change in energy. Consequently, we can choose the zero for energy anywhere we wish, and it is convenient to choose it where $\omega = 0$. This sets $C = 0$, and we have

$$\hbar\omega = E. \tag{6}$$

It is reassuring that the relationship between $E$ and $\omega$ for fermions is the same as the relationship that the photoelectric effect experiment requires for photons. We used the double slit experiment and the classical behavior of wave packets to develop Equation 6 for fermions, and still obtained the same result determined from the photoelectric effect for photons. This relationship is valid for all known particles.

Now that we have $\omega$ and $k$ in terms of measurable quantities, let us return to our trial wave equation for nonrelativistic fermions (Equation 3),

$$\boldsymbol{\nabla}^2\Psi(\boldsymbol{r},t) + \frac{ik^2}{\omega}\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t) = 0. \tag{3}$$

From Equations 6 and 4, we have

$$\frac{k^2}{\omega} = \frac{k^2}{E/\hbar} = \frac{k^2}{\hbar^2 k^2/2m\hbar} = \frac{2m}{\hbar}.$$

Substituting this into Equation 3, multiplying by $-\hbar^2/2m$, and rearranging, produces the Schrodinger Equation for a free particle,

$$-\frac{\hbar^2}{2m}\boldsymbol{\nabla}^2\Psi(\boldsymbol{r},t) = i\hbar\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t). \tag{7}$$

We have developed a wave equations for free particles; how do we add potential energy. Free particles are approximately[20] described by plane wave solutions,

$$\Psi(\boldsymbol{r},t) = \mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{r}-i\omega t}.$$

If we substitute a plane wave solution into Equation 7, we obtain

$$\frac{\hbar^2 k^2}{2m} = E.$$

Using Equation 5, we see that this reduces to

$$E = p^2/2m = \text{ kinetic energy.}$$

This is a well known relationship from classical mechanics for particles with no potential energy. Since large numbers of fermions, such as you would find in a baseball, obey classical mechanics, we should try to include potential energy in a way that is consistent with classical mechanics. Consequentially, we just modify the last expression to agree with classical mechanics for a particle with potential energy $V(\boldsymbol{r})$,

$$E = \text{ kinetic energy plus potential energy} = \frac{p^2}{2m} + V(\boldsymbol{r}).$$

The modification of Equation 7 necessary to obtain this expression for a plane wave is

$$\left(-\frac{\hbar^2}{2m}\boldsymbol{\nabla}^2 + V(\boldsymbol{r})\right)\Psi(\boldsymbol{r},t) = i\hbar\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t). \tag{8}$$

This is Schrodinger's equation. Its solutions are wave functions that accurately describe the behavior of all nonrelativistic fermions. We arrived at it in five steps:

[20]The approximation depends on the envelope function changing much more slowly than the carrier.

1. We recognized that the classical wave equation predicted the interference pattern observed in the double slit experiment with fermions;
2. We modified the classical wave equation to conserve probability because fermions are experimentally conserved;
3. We used experimental evidence from the double slit experiment to determine that momentum $p$ and wave number $k$ are related by $p = \hbar k$;
4. We used the behavior of classical mechanical wave packets to show that energy $E$ and frequency $\omega$ are related by $E = \hbar\omega$.
5. We incorporated potential energy into the Schrodinger Equation in a way that was consistent with classical mechanics.

We need to check the last step by examining the behavior of fermions with potential energy. One way to do this is to apply Schrodinger's equation to electrons bound to an atom. This has been done and Schrodinger's Equation accurately predicts the chemical properties of all elements, and the energy levels of electrons in all elements except for relativistic corrections. Another way is to study the behavior of particles that tunnel through a potential barrier that is greater than their energy. The Schrodinger Equation accurately predicts the behavior of tunneling diodes and tunneling electron microscopes.

Now that we have fixed our theory so that it conserves probability and incorporates potential energy correctly, we will consider the limitation that the only physical variable whose probability distribution our theory predicts is position. Actually, this limitation should be no surprise since we built the theory from the double slit experiment and that experiment only measures position. However, we can get an idea of how the theory will handle other variables if we modify the experiment slightly and if you will allow me to be a little sloppy with normalization.

We need to modify the source of particles so that more particles reach the top slit than the bottom slit. We can do this by placing a very narrow potential barrier in front of the bottom slit. If the potential height of this barrier is slightly greater than the energy of the particles, then the particles that reach the lower slit must tunnel through. The fraction that successfully tunnel through the barrier is determined by the height and width of the barrier, and it can be varied from essentially zero to one.

Let $a^2$ be the fraction of particles that traverses the top slit and $b^2 = 1 - a^2$ be the fraction that traverses the bottom slit. If $a$ and $b$ are not equal, we get a different interference pattern, and we find experimentally that we can predict the new pattern accurately if we replace Equation 20 with

$$\Psi = (a\psi_1 + b\psi_2). \tag{9}$$

We interpret $a\psi_t$ and $b\psi_b$ as the wave functions for the particles that go through the top and bottom slits respectively. It is important to note that Equation 9 is an experimental result. It the simplest modification to Equation 20 that agrees with the experimental results when $a$ and $b$ are not equal. Equation 9 is also suggested by our proposed quantum theory. Since the fraction of particles going through the top slit is $a^2$, our proposed quantum theory requires the square of the wave function for the top slit to be multiplied by $a^2$. Consequently, we must multiply the wave function $\psi_1$ for the top slit by $a$ (and the wave function $\psi_2$ for the bottom slit by $b$).

Now suppose that we place very strong detectors after each slit. Clearly, the top detector will detect $a^*a = a^2$ of the particles and the bottom slit will detect $b^*b = b^2$ of the particles. I choose to use $a^*a$ instead of $a^2$ because that allows $a$ and $b$ to be complex without changing our results. There will be cases in the future in which $a$ and $b$ might be complex. For each particle, the probability of its going through the top slit is $a^*a$. So the possible outcomes of a 'which slit' measurement are top and bottom with probabilities $a^*a$ and $b^*b$ respectively.

The states $\psi_1$ and $\psi_2$ are called pure states for the 'which slit' measurement. If the system is in the state $\psi_1$, we know that a 'which slit' measurement will result in the top slit. We also know that whatever the initial state, if a 'which slit' measurement results in the top slit, then after the measurement the system is in the state $\psi_1$. The initial wave function

$$\Psi = a\psi_1 + b\psi_2$$

is a superposition of pure states. It is called a superposition state, a mixed state, or just the state function. The measurement is described by saying that it causes the initial state function $\Psi$ to collapse instantaneously to a pure state of the measurement. And not just any pure state, it is the pure state corresponding to the value that was measured. Philosophers describe this by saying that before

the measurement the particle has various mutually exclusive potential attributes. The measurement destroys some of those potentials and actualizes only one.

We now have a recipe for predicting the probabilities of all the possible outcomes of any measurement. The recipe is

- Construct the state function $\Psi$. It must satisfy Schrodinger's equation and incorporate all the knowledge we have about the initial state of the system.
- Find the pure states of the measurement. This sounds scary, but actually you have already had much of the math, and the first semester of quantum mechanics is devoted to learning how to find the pure states. The pure states are just the eigenvectors of the operator corresponding to the classical variable being measured.
- Write the state function as a superposition of the pure states.
- The probability of measuring any particular value $\alpha$ is the magnitude squared of the coefficient in the state function superposition of the pure state that corresponds to $\alpha$.

Our basic theory is complete. Now we need to learn how to find the operators that correspond to physical observables and their eigenvectors. These are the pure states. How to do this with examples is presented in the next chapter.

# Appendices

## A Derivation of the Double Slit Interference Pattern.

The double slit interference pattern described by Equation 2 was first obtained for light waves so we will follow that procedure here. The same derivation will work for particle waves such as the waves associated with electrons.

Light waves satisfy the classical wave equation,

$$\boldsymbol{\nabla}^2\Psi - \frac{1}{c^2}\frac{\partial^2\Psi}{\partial t^2} = 0, \tag{10}$$

where $c$ is the wave's phase speed and $\boldsymbol{\nabla}^2$ is the Laplacian. If we use monochromatic light, then all the photons have the same frequency $\omega = 2\pi f$ and wave number $k = 2\pi/\lambda$, where $f$ is the frequency, $\lambda$ is the wavelength, and $c = \lambda f$. With monochromatic light, we can eliminate the time dependence of $\Psi$ by substituting

$$\Psi(\boldsymbol{r}, t) = \mathrm{e}^{-i\omega t}\psi(\boldsymbol{r}) \tag{11}$$

into Equation 10, reducing it to Helmholtz's Equation:

$$\left(\boldsymbol{\nabla}^2 + k^2\right)\psi(\boldsymbol{r}) = 0. \tag{12}$$

Since Helmholtz's Equation is a second order linear homogeneous partial differential equation, there are an infinite number of independent solutions.[21] However, appropriate boundary conditions limit the solution to a single unique solution. The complete solution that satisfies general boundary conditions is provided by the Kirchhoff Integral Theorem,[22, 23]

$$\begin{aligned}\psi(\boldsymbol{r}) &= \frac{1}{4\pi}\oint_S \left[G(\boldsymbol{r}, \boldsymbol{r}_s)\boldsymbol{\nabla}_s\psi(\boldsymbol{r}_s) - \psi(\boldsymbol{r}_s)\boldsymbol{\nabla}_s G(\boldsymbol{r}, \boldsymbol{r}_s)\right]\cdot\hat{\boldsymbol{n}}_s\, dS \\ &\quad + \int_V \rho(\boldsymbol{r})G(\boldsymbol{r}, \boldsymbol{r}_s)dV,\end{aligned} \tag{13}$$

[21] All linear superpositions of functions of the form $\exp(\pm\,(i\boldsymbol{k}\cdot\boldsymbol{r})$ are solutions if $|\boldsymbol{k}| = k$.

[22] See Equation 7.2.7 on page 806 of **Methods of Theoretical Physics**, P. M. Morse and H. Feshbach, McGraw Hill, 1953 or go to the URL in footnote 23.

[23] `https://en.wikipedia.org/wiki/Kirchhoff_integral_theorem`

where the field point $\boldsymbol{r}$ is inside the closed surface $S$, the source point $\boldsymbol{r}_s$ is on the surface $S$ for the surface integral and inside $S$ for the volume integral, $ds$ is the elementary surface element, $\hat{\boldsymbol{n}}_s$ is the outward pointing unit normal to the surface $S$, $\boldsymbol{\nabla}_s$ operates on $\boldsymbol{r}_s$, $dV$ is the elementary volume element, and $G(\boldsymbol{r}, \boldsymbol{r}_s)$ is the Green's function for the Helmholtz's Equation.

In our case, the only source of particles is the one illustrated in Figure 3 and located at $z = -L_1$. As long as we chose our closed surface to exclude negative $z$ values (and we will do that) then $\rho = 0$ inside $S$, and the volume integral is zero. Our Kirchhoff Integral now becomes

$$\psi(\boldsymbol{r}) = \frac{1}{4\pi} \oint_S [G(\boldsymbol{r}, \boldsymbol{r}_s)\boldsymbol{\nabla}_s\psi(\boldsymbol{r}_s) - \psi(\boldsymbol{r}_s)\boldsymbol{\nabla}_s G(\boldsymbol{r}, \boldsymbol{r}_s)] \cdot \hat{\boldsymbol{n}}_s \, dS \,, \quad (14)$$

The Green's function is a solution to the point source Helmholtz equation,

$$\left(\boldsymbol{\nabla}^2 + k^2\right) G(\boldsymbol{r}, \boldsymbol{r}_s) = -4\pi\delta(\boldsymbol{r} - \boldsymbol{r}_s), \quad (15)$$

where $\delta(\boldsymbol{r} - \boldsymbol{r}_s)$ is the three dimensional Dirac Delta function.[24] The free field solution to Equation 15 is

$$g(\boldsymbol{r}, \boldsymbol{r}_s) = \frac{\mathrm{e}^{ikR}}{R}, \quad (16)$$

where $R = |\boldsymbol{r} - \boldsymbol{r}_s|$. This is called the free field Green's Function because it corresponds to a point source at $\boldsymbol{r}_s$ with no boundaries. Any solution to the homogeneous Helmholtz Equation,

$$\left(\boldsymbol{\nabla}^2 + k^2\right) G_h(\boldsymbol{r}, \boldsymbol{r}_s) = 0,$$

can be added to $G(\boldsymbol{r}, \boldsymbol{r}_s)$ without altering its being a solution to Equation 15. By adding homogeneous solutions, we can force $G(\boldsymbol{r}, \boldsymbol{r}_s)$ to fit various desired boundary conditions on the surface $S$.

It is shown in Appendix B that we can use the $xy$ plane containing the barrier and the two slits for the surface $S$. This surface is labeled $S_1$ in Appendix B and in Figure 5, and we will henceforth refer to it as $S_1$, it is just the $xy$ plane.

You will find in the literature two methods of applying Equation 14 to diffraction problems: Fresnel-Kirchhoff (FK)[25]and Rayleigh-Sommerfeld

[24]It is also shown on page 808 of Reference 22 that the Green's Function is symmetric, $G(\boldsymbol{r}, \boldsymbol{r}_s) = G(\boldsymbol{r}_s, \boldsymbol{r})$ if both $\boldsymbol{r}$ and $\boldsymbol{r}_s$ are inside or on the surface $S$.

[25]`https://en.wikipedia.org/wiki/Kirchhoff%27s_diffraction_formula`

(RS).[26, 27] Both methods give the same result if the distances $L_1$ and $L$ in Figure 3 are much larger than the size and spacing of the slits, but the RS method is more mathematically consistent and predicts Poisson's spot (see Reference 26). The FK method assumes that both $\psi$ and $\hat{\boldsymbol{n}} \cdot \boldsymbol{\nabla}\psi$ are zero on the barrier part of $S_1$. Unfortunately, this means that if $\psi$ is an analytic function, then it is zero everywhere. The RS method avoids this issue with a clever choice of $G(\boldsymbol{r}, \boldsymbol{r}_s)$ that requires only $\psi$ to be zero on the barrier.

Since $z > 0$ everywhere inside $S$, we can use the Green's function,

$$G(\boldsymbol{r}, \boldsymbol{r}_s) = \frac{\mathrm{e}^{ikR}}{R} - \frac{\mathrm{e}^{ikR_1}}{R_1}, \tag{17}$$

where

$$\begin{aligned} R &= \sqrt{(x-x_s)^2 + (y-y_s)^2 + (z-z_s)^2}, \text{ and} \\ R_1 &= \sqrt{(x-x_s)^2 + (y-y_s)^2 + (z+z_s)^2}. \end{aligned}$$

This Green's function is zero when $z = 0$ which is the case on the $xy$ plane and $S_1$. With this choice of Green's function, the first term,

$$G(\boldsymbol{r}, \boldsymbol{r}_s)\boldsymbol{\nabla}_s\psi(\boldsymbol{r}_s),$$

in the surface integral over $S_1$ in Equation 14 is zero, and Equation 14 reduces to

$$\psi(\boldsymbol{r}) = -\frac{1}{4\pi}\int_{S_1} [\psi(\boldsymbol{r}_s)\boldsymbol{\nabla}_s G(\boldsymbol{r}, \boldsymbol{r}_s)] \cdot \hat{\boldsymbol{n}}\, dS_1\,. \tag{18}$$

Appropriate boundary conditions on $\psi$ are

$$\begin{aligned} \psi(x, y, 0) &= \frac{1}{\sqrt{2ab}} \text{ in the slits, and} \\ &= 0 \qquad \text{on the barrier where there are no slits,} \end{aligned}$$

where $a$ and $b$ are the width in the $x$ direction and height in the $y$ direction of the slits respectively, and where the factor $1\sqrt{2ab}$ is a normalization factor to require that

$$\int_{S_1} \psi^2\psi\, dS = 1.$$

[26] Robert Lucke, *Rayleigh-Sommerfield Diffraction vs Fresnel-Kirchoff, Fourier Propagation, and Poisson's spot*, Naval Research Laboratory, Report NRL-FR-7218-04-10101, Dec, 30, 2004. This document can be found at `www.dtic.mil/get-tr-doc/pdf?AD=ada429355`

[27] `https://statweb.stanford.edu/~candes/math262/Lectures/Lecture16.pdf`

The normal gradient of the Green's function is

$$\begin{aligned}\hat{\boldsymbol{n}} \cdot \boldsymbol{\nabla}_s G(\boldsymbol{r}, \boldsymbol{r}_s)|_{z=0} &= -\left.\frac{\partial}{\partial z_s} G(\boldsymbol{r}, \boldsymbol{r}_s)\right|_{z_s=0} \\ &= -\left(ik - \frac{1}{R}\right) \frac{\mathrm{e}^{ikR}}{R} \frac{\partial R}{\partial z_s} + \left(ik - \frac{1}{R_1}\right) \frac{\mathrm{e}^{ikR_1}}{R_1} \frac{\partial R_1}{\partial z_s}.\end{aligned}$$

Note that $R = R_1$ when $z_s = 0$, and

$$\left.\frac{\partial R_1}{\partial z_s}\right|_{z_s=0} = -\left.\frac{\partial R}{\partial z_s}\right|_{z_s=0} = \frac{z}{R} = \cos\theta,$$

where $\theta$ is the angle between the $z$ axis and the line to the detector in Figure 3. We can now write the normal gradient of the Green's function as

$$\hat{\boldsymbol{n}} \cdot \boldsymbol{\nabla}_s G(\boldsymbol{r}, \boldsymbol{r}_s)|_{z_s=0} = 2\left(ik - \frac{1}{R}\right) \frac{\mathrm{e}^{ikR}}{R} \cos\theta.$$

With these results, the Kirchhoff Integral becomes

$$\psi(\boldsymbol{r}) = -\frac{1}{2\pi} \int_{S_1} \left[\psi(\boldsymbol{r}_s)\left(ik - \frac{1}{R}\right) \frac{\mathrm{e}^{ikR}}{R} \cos\theta\right] dS_1\,.$$

In all double slit experiments, $R >> \lambda$ so we can neglect terms of order $1/kR = \lambda/(2\pi R)$. Furthermore, the angle $\theta$ is usually very small (less than .05 in out case), so we can replace $\cos\theta$ with unity. We finally arrive at the following integral for $\psi$,

$$\psi(\boldsymbol{r}) = -\frac{ik}{2\pi} \int_{S_1} \psi(\boldsymbol{r}_s) \frac{\mathrm{e}^{ikR}}{R}\, dS_1\,. \tag{19}$$

The only areas on $S_1$ where $\psi$ is not zero are the top and bottom slits. The integral is the sum of the integrals over the top slit and the bottom slit. Labeling these integrals as $\psi_t$ and $\psi_b$ respectively, we have

$$\psi(\boldsymbol{r}) = \psi_t(\boldsymbol{r}) + \psi_b(\boldsymbol{r}). \tag{20}$$

The function $\psi_t(\boldsymbol{r})$ is the contribution to the wave function from the top slit, and $\psi_b(\boldsymbol{r})$ is the contribution from the bottom slit.

First we evaluate $\psi_t$

$$\psi_t(\boldsymbol{r}) = -\frac{ik}{2\pi\sqrt{2ab}} \int_{d/2-a/2}^{d/2+a/2} \int_{-b/2}^{b/2} \frac{\mathrm{e}^{ikR}}{R}\, dx_s\, dy_s\,.$$

If we choose to make the height $b$ of the slits about the same size as their widths $a$, then $|\boldsymbol{r}_s| << |\boldsymbol{r}|$, and we can expand $R$ in powers of $r_s/r$ where $r = |\boldsymbol{r}|$ and $r_s = |\boldsymbol{r}_s|$. In order to determine how many terms we must keep in our expansion, we examine the argument of the exponential, $\mathrm{e}^{ikR}$

$$\begin{aligned} kR &= k\sqrt{r^2 - 2\boldsymbol{r}\cdot\boldsymbol{r}_s + r_s^2} = kr\sqrt{1 - \frac{2\boldsymbol{r}\cdot\boldsymbol{r}_s}{r^2} + \frac{r_s^2}{r^2}} \\ &= kr\left(1 - \frac{\boldsymbol{r}\cdot\boldsymbol{r}_s}{r^2} + \mathcal{O}\left(\frac{r_s}{r}\right)^2 + +\mathcal{O}\left(\frac{r_s}{r}\right)^3 + \cdots\right). \end{aligned}$$

For our parameters, the wave number times the distance $r$ from the slits to the detectors is on the order of $2\times10^7$, and $r_s/r$ is less than $1.6\times 10^{-5}$. We see that we must keep the linear term in the exponential, but can drop all higher order terms. Furthermore, we can drop the linear term in the $R$ that is in the denominator because it is not the argument of an exponential. Dropping these terms is called the Fraunhofer Approximation.

One further simplification for our particular case is that we should set $y = 0$ because our detectors are placed in the plane of the paper in figure 3. With this choice and using the Fraunhofer approximation, the expression for $\psi_t$ becomes

$$\begin{aligned} \psi_t(\boldsymbol{r}) &= -\frac{ik\mathrm{e}^{ikr}}{2\pi r\sqrt{2ab}}\int_{d/2-a/2}^{d/2+a/2}\int_{-b/2}^{b/2}\mathrm{e}^{-ik\boldsymbol{r}\cdot\boldsymbol{r_s}/r}\,dx_s\,dy_s \\ &= -\frac{ik\mathrm{e}^{ikr}}{2\pi r\sqrt{2ab}}\left(\int_{d/2-a/2}^{d/2+a/2}dx_s\,\mathrm{e}^{-ikxx_s/r}\right)\left(\int_{-b/2}^{b/2}\mathrm{e}^{-ikyy_s/r}\,dy_s\right)_{y=0} \\ &= -\frac{ik\sqrt{ab}\,\mathrm{e}^{ikr}}{2\pi r\sqrt{2}}\,\mathrm{e}^{-(ikd\sin\theta)/2}\left(\frac{\sin\alpha}{\alpha}\right), \end{aligned}$$

where

$$\alpha = \frac{ka\sin\theta}{2} = \frac{\pi a\sin\theta}{\lambda}, \quad \text{and} \quad \sin\theta = x/r.$$

The solution for $\psi_b$ follows the same steps and produces

$$\psi_b(\boldsymbol{r}) = -\frac{ik\sqrt{ab}\,\mathrm{e}^{ikr}}{2\pi r\sqrt{2}}\,\mathrm{e}^{(ikd\sin\theta)/2}\left(\frac{\sin\alpha}{\alpha}\right).$$

Adding $\psi_t$ and $\psi_b$, we obtain the time independent solution for both slits open,

$$\psi(\boldsymbol{r}) = \psi_t(\boldsymbol{r}) + \psi_b(\boldsymbol{r}) = -\left(\frac{ik\sqrt{ab}\,\mathrm{e}^{ikr}}{\pi r\sqrt{2}}\right)\cos\beta\,\left(\frac{\sin\alpha}{\alpha}\right), \qquad (21)$$

where

$$\beta = \frac{kd\sin\theta}{2} = \frac{\pi d\sin\theta}{\lambda}.$$

The time dependence is obtained by multiplying by $\mathrm{e}^{-i\omega t}$,

$$\Psi(\boldsymbol{r},t) = \mathrm{e}^{-i\omega t}\psi(\boldsymbol{r}).$$

We are actually interested in the intensity because that is what we can measure,

$$I(\theta) = \Psi^*\Psi = I_o\,\cos^2\beta\,\left(\frac{\sin\alpha}{\alpha}\right)^2, \tag{22}$$

where $I_o$ is the maximum intensity which occurs when $\theta = 0$.

Equation 22 was obtained by squaring the magnitude of the solution to the classical wave equation for the appropriate boundary conditions. It is identical to Equation 2 which is the observed double slit interference pattern.

# B Integration Surface for Kirchhoff Integral.

The Kirchhoff integral requires a closed surface $S$. We will use the union of the two surfaces illustrated with the dashed red curve in Fig 5. Surface $S_1$ is the $xy$ plane containing the barrier with two slits in it. Surface $S_2$ is the hemisphere located to the right of the $xy$ plane with radius $L_2$. The purpose of this Appendix is to show that in the limit as $L_2 \to \infty$, the integral over $S_2$ is zero, so we will only consider that integral here.

There is no requirement that we use the same Green's function for both integrals as long as both are solutions to Equation 15. On surface $S_2$, we will use the free field Green's Function of Equation 16,

$$g(\boldsymbol{r},\boldsymbol{r}_s) = \frac{\mathrm{e}^{ikr}}{R},$$

where $R = |\boldsymbol{r} - \boldsymbol{r}_s|$.

Using this Green's Function, both terms in the surface integral in Equation 14 contribute so we must know the values of both $\psi(\boldsymbol{r}_s)$ and $\hat{\boldsymbol{n}}\cdot\boldsymbol{\nabla}_s\psi(\boldsymbol{r}_s)$ on $S_2$. Since $S_2$ is located at infinity, we can take the limit

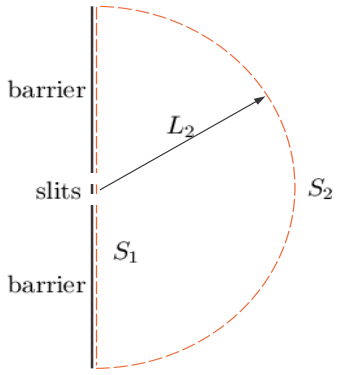


Figure 5: Surface of Integration in Kirchhoff Integral.

as $r_s \to \infty$, and in this limit, the solution will be radially outgoing waves,[28, 29]

$$\lim_{r_s \to \infty} \psi(\boldsymbol{r_s}) = \frac{\mathrm{e}^{ikr_s}}{r_s} f(\xi_s, \phi_s), \tag{23}$$

where $r = \sqrt{x^2 + y^2 + z^2}$, $\xi = \cos^{-1}(z/r)$, and $\phi = \tan^{-1}(y/x)$. It is customary to use the symbol $\theta$ for the polar angle, $\cos^{-1}(z/r)$, rather than $\xi$, but I have already defined $\theta$ differently in Figure 3. Interestingly, $\theta$ and $\xi$ are the same when $y = 0$ which is where our detectors are located.

Now that we have expressions for $G(\boldsymbol{r}, \boldsymbol{r}_s)$ and $\psi(\boldsymbol{r}_s)$ on $S_2$, we can evaluate the integrand of the surface integral over $S_2$ in Equation 14. First we will state the following useful results:

1. The surface $S_2$ is a hemisphere of constant $r_s$ so the unit normal $\hat{\boldsymbol{n}}_s$ is just $\hat{\boldsymbol{r}}_s$ and
$$\boldsymbol{n_s} \cdot \boldsymbol{\nabla}_s = \frac{\partial}{\partial r_s}.$$

2. We can write $R = \sqrt{r_s^2 - 2\boldsymbol{r}_s \cdot \boldsymbol{r} + r^2}$ as
$$\begin{aligned} R &= r_s \sqrt{1 - \frac{2\boldsymbol{r}_s \cdot \boldsymbol{r}}{r_s^2} + \frac{r^2}{r_s^2}} = r_s \left(1 - \frac{\boldsymbol{r}_s \cdot \boldsymbol{r}}{r_s^2} + \mathcal{O}\left(\frac{r}{r_s}\right)^2\right) \\ &= r_s \text{ in the limit as } r_s \to \infty. \end{aligned}$$

[28]The outgoing wave requirement is called the Sommerfeld Radiation condition and is often expressed more generally as $\lim_{r\to\infty} (r\partial\psi(\boldsymbol{r})/\partial r - ikr\psi(\boldsymbol{r})) = 0$.

[29]Substituting the outgoing wave solution into Helmholtz's Equation and taking the limit as $r \to \infty$ shows that it is a solution in this limit.

3. Writing $\boldsymbol{r}_s$ as $\boldsymbol{r}_s = r_s\hat{\boldsymbol{r}}_s$, we see that

$$\frac{\partial}{\partial r_s}\left(\boldsymbol{r}_s \cdot \boldsymbol{r}\right) = \frac{\partial}{\partial r_s}\left(r_s\hat{\boldsymbol{r}}_s \cdot \boldsymbol{r}\right) = \hat{\boldsymbol{r}}_s \cdot \boldsymbol{r}$$

because $\hat{\boldsymbol{r}}_s$ is independent of the magnitude of $\boldsymbol{r}_s$, and $r_s = |\boldsymbol{r}_s|$.

4. It follows that

$$\begin{aligned}\frac{\partial R}{\partial r_s} &= \frac{\partial}{\partial r_s}\sqrt{r_s^2 - 2\boldsymbol{r}_s \cdot \boldsymbol{r} + r^2} = \frac{2r_s - 2\hat{\boldsymbol{r}}_s \cdot \boldsymbol{r}}{2R} \\ &= 1, \text{ in the limit as } r_s \to \infty.\end{aligned}$$

We wish to evaluate the Kirchhoff Integral over surface $S_2$ in the limit as $L_2 \to \infty$ in Figure 5. In other words, we will take the limit of the integrand as $r_s \to \infty$. We must remember that the surface element $dS$ contains a factor of $r_s^2$, so we must multiply the integrand by $r_s^2$ before we evaluate the limit. We can discard terms of order $r/r_s$ and $1/kr_s$ as long as they are not multiplied by quantities that approach infinity as $r_s$ approaches infinity. Labeling the first term of the integrand in Equation 14 as $T_f$, we have

$$\begin{aligned}T_f &= r_s^2 G(\boldsymbol{r}, \boldsymbol{r}_s)\left(\hat{\boldsymbol{n}} \cdot \boldsymbol{\nabla}_s \psi(\boldsymbol{r}_s)\right) = r_s^2 \frac{\mathrm{e}^{ikR}}{R}\frac{\partial}{\partial r_s}\frac{\mathrm{e}^{ikr_s}}{r_s} \\ &= \frac{r_s^2 \mathrm{e}^{ik(R+r_s)}}{Rr_s} k\left(i - \frac{1}{kr_s}\right) = ik\mathrm{e}^{ik(R+r_s)} \text{ in the limit as } r_s \to \infty.\end{aligned}$$

The second term is

$$\begin{aligned}T_s &= -r_s^2 \psi(\boldsymbol{r}_s)\left(\hat{\boldsymbol{n}} \cdot \boldsymbol{\nabla}_s G(\boldsymbol{r}, \boldsymbol{r}_s)\right) = -r_s^2 \frac{\mathrm{e}^{ikr_s}}{r_s}\frac{\partial}{\partial r_s}\frac{\mathrm{e}^{ikR}}{R} \\ &= -\frac{r_s^2 \mathrm{e}^{ik(R+r_s)}}{Rr_s} k\left(i - \frac{1}{kR}\right)\frac{\partial R}{\partial r_s} \\ &= -ik\mathrm{e}^{ik(R+r_s)} = -T_f \text{ in the limit as } r_s \to \infty.\end{aligned}$$

We see that the two terms cancel in the limit so the integral over $S_2$ is zero.

# C Wave velocity and particle velocity.

A simple solution to either the classical Wave Equation (equation 10) or our trial particle wave equation (Equation 3) is a plane wave (PW) traveling in the $x$ direction,

$$\Psi_{\mathrm{pwx}}(\boldsymbol{r}, t) = \mathrm{e}^{i(kx - \omega t)}.$$

However, this function is unsuitable to describe any particle that is confined to our laboratory or even confined to our solar system because the pdf, $\Psi^*\Psi = 1$, is one everywhere. In order to describe a particle that is in our lab, we need a localized disturbance that is nonzero in some finite region and is zero everywhere else. Such a function is called a wave packet and is illustrated in Figure 6.[30] This particle is

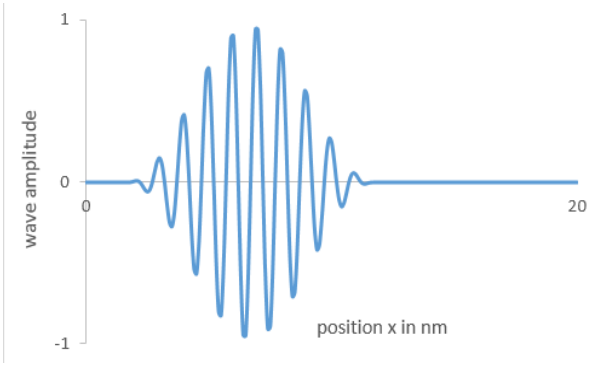


Figure 6: Wave Packet.

located around $x = 7$ nm. It is useful to think of the wave packet as a product of two functions: the carrier which is just the PW solution illustrated in Figure 7, and the envelop function $f(x,t)$ illustrated in Figure 8,

$$\Psi(x,t) = f(x,t)\, \mathrm{e}^{i(k_o x - \omega_o(k_o)t)},$$

where $k_o$ and $\omega_o$ are the wave number and frequency of the carrier respectively, and where I have emphasized that $\omega_o$ depends on $k_o$ by writing $\omega(k_o)$. This separation is useful because the PW solution is simple and because the envelop function changes very slowly compared to the PW solution. Only the envelope function contributes to the pdf,

$$\Psi^*\Psi = f^* f.$$

Since we prepare the particle in a specific way, we know the wave function at $t = 0$,

$$\Psi(x,0) = f(x,0)\mathrm{e}^{ik_o x}.$$

[30] I only plot the real part of the wave functions in these figures; the imaginary part would look very similar.

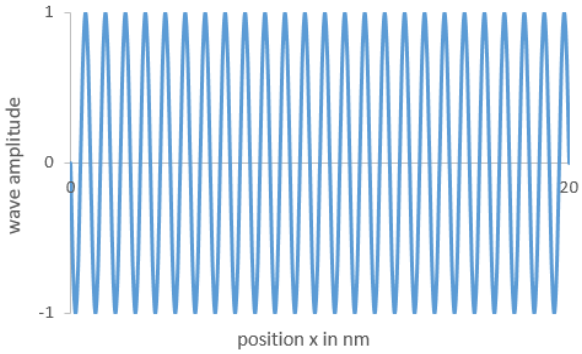


Figure 7: Carrier Wave.

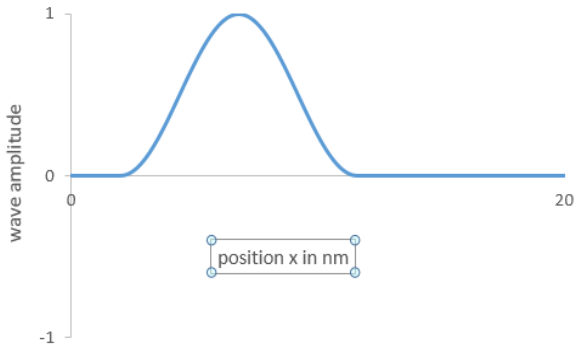


Figure 8: Envelope function.

If this were a plane wave of wave number $k$, we could introduce the time dependence by multiplying by $\mathrm{e}^{-i\omega(k)t}$. Unfortunately, $f(x,t)$ introduces an unknown time dependence. However, we can use a Fourier transform[31, 32, 33] to write $f(x,0)$ as a superposition of plane waves,

$$f(x,0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{e}^{ik'x} g(k')\, dk' \,.$$

Since $f(x)$ changes much more slowly than the carrier (see the figures),

[31] P. M. Morse and H. Feshbach **Methods of Theoretical Physics**, McGraw Hill, 1953. (page 453).

[32] G. Arfken, **mathematical Methods for Physicists**, second edition, Academic Press, 1970. (Equation 6.52, page 314)

[33] `https://en.wikipedia.org/wiki/Fourier_transform`

we expect that its plane wave superposition contains only low wave number plane waves. In other words, $g(k') = 0$ unless $k' << k_o$.

Now the wave function at $t = 0$ is a superposition of plane waves,

$$\Psi(x,0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{e}^{i(k'+k_o)x} g(k')\, dk' \,,$$

and when we introduce the time dependence, we must multiply by

$$\mathrm{e}^{-i\omega(k'+k_o)t} = \mathrm{e}^{-i\omega(k)t},$$

where $k = k' + k_o$ is the overall wave number and therefore is the argument of $\omega(k)$. Introducing the time dependence of $\Psi(x,t)$, we have

$$\Psi(x,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{e}^{ikx - i\omega(k)t} g(k')\, dk' \,. \tag{24}$$

It seems that we cannot proceed much further without knowing how $\omega(k)$ depends on $k$; however, we can use the fact that $k'$ is small to expand $\omega(k_o + k')$ about $k_o$,

$$\omega(k_o + k') = \omega(k_o) + k' \left.\frac{d\omega}{dk}\right|_{k=k_o} + \mathcal{O}((k')^2).$$

Dropping the second order term,[34] defining

$$v_g = \left.\frac{d\omega}{dk}\right|_{k=k_o} \text{ and } \omega_o = \omega(k_o),$$

and substituting into E 24, we have

$$\begin{aligned}
\Psi(x,t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{e}^{i(k'+k_o)x - i(\omega_o + k' v_g)t)} g(k')\, dk' \\
&= \frac{\mathrm{e}^{i(k_o x - \omega_o t)}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{e}^{ik'(x - v_g t)} g(k')\, dk' \\
&= \mathrm{e}^{i(k_o x - \omega_o t)} f(x - v_g t).
\end{aligned}$$

[34] If the detectors are about a mm in size, then the wave packet is about a mm in size and $k' < 10^3$. If an electron is passed through a potential of 10 KV, then its speed is $6 \times 10^7$ m/s, and it will travel across the lab in $10^{-7}$ seconds. Unless the second derivative of $\omega$ is greater than 10, the second order term will not contribute between when the electron is produced and when it is detected. It turns out that the second derivative is $\hbar/m \doteq 10^{-4}$. so the second term is negligible.

This last expression shows that the envelope moves with velocity $v_g$. When time changes by $\Delta t$ then $x$ must change by $\Delta x = v_g \Delta t$ to keep the argument of the envelope function $f(x - v_g t)$ constant. Therefore, the position of the wave packet moves a distance $v_g t$ in a time interval $\Delta t$; $v_g$ is the velocity of the wave packet and the velocity of the particle,

$$v_g = \left. \frac{d\omega}{dk} \right|_{k=k_o} . \tag{25}$$