Chapter 3

Pre 19th Century Physics

3.1 Introduction

We begin our study of modern physics by examining the phenomena associated with light. Although the phenomena of light is among the oldest examined and there are theories of light that must go back to the first humans, light is particularly interesting from the modern perspective because of the central role that it has played in the development of our ideas, particularly quantum mechanics and relativity. To trace the development of our ideas about light from the ancients to today would take the entire semester and not allow any room for modern physics. For a concise coverage of these ancient ideas, the book by Park [Park 1997] or [Ronchi 1970] are excellent. Instead, we will start with one of two threads of development that emerged in the 17^{th} century and as we will see are relevant to our modern approach to light and related phenomena.

In the 1660's, Fermat[1] proposed that light travels between two points over the path, called a light ray, that has the least travel time of all the possible paths connecting those two points, Section 3.2 on page 57. At the time of its formulation, there was a competing theory: a particle theory often identified with Newton. The particulate theory was the generally accepted description, because it successfully accounted for all the phenomena known at that time to be related to light, basically reflection and refraction. Fermat's approach equally well described the reflection and refraction experiments of the day. In a sense, there was a stalemate with the great prestige of Newton providing the edge to the particulate approach. A significant

[1] Pierre de Fermat, born in 1607 and died in 1662, was responsible for significant contributions to mathematics and optics.

difference between the two theories was that Fermat's Theory required that light traveled slower in a dense media whereas the particulate approach as given by Newton required that light travel faster. At the time of formulation of these competing theories, it was impossible to measure the speed of light in dense media. Once the confirming experiment supported Fermat's theory, it became the accepted approach. We now know that, in some sense, some of the aspects of the particulate theory are correct for a certain range of observation but we will get to this later in Chapter 19 on page 403. With the measurement of the speed of light in dense matter, the particulate theory was then superseded by a Fermat's theory. Fermat's formulation was very successful in describing all the phenomena associated with light that was known at his time and, in fact, most of the common phenomena that we associate with light, see Section 3.3 on page 63. Newton's lasting contribution was the description of the relationship of color to light, see Section 3.4 on page 79. Interestingly, the modern interpretation of the behavior of light comes very close to what Newton developed and some argue that Newton was close to the discovery of quantum mechanics at least as it is applied to light.

As new phenomena, interference and diffraction, associated with light were observed, it became clear that a new construction was needed. Extending and clarifying a construction associated with Huygens, a contemporary of Newton, and Thomas Young, Fresnel formulated a new approach that in the appropriate limits reproduced to all the success of Fermat but incorporated the new phenomena of interference and diffraction, see Section 4 on page 83. Integral to the success of this approach is the idea of an underlying continuous system, the ether, that was the basis for the the phenomena associated with light. Much of the intuition of the new construction was based on the understanding of fluid flows and, in particular, sound that had been developed earlier. These also depended on the properties of an underlying mechanical system, air for sound and water, for surface waves[2], for their interpretation. There was similar material basis for light. Later, in what at the time appeared as an independent investigation, Maxwell was attempting to construct a mechanical model of electric and magnetic forces. In his mechanical model of electric and magnetic forces, Maxwell realized that disturbances traveled at the speed of light and he immediately identi-

[2]It is common practice to call all disturbances that travel in a field system as waves. This is not completely clear since in any field situation that supports traveling solutions any disturbance will travel but is called a wave and the disturbance does not have to take the shape of a sine or cosine. Where appropriate I will try to be more formal and call them travelers.

fied these as light. It also provided a material basis for the propagation of light called the aether. The methods of analysis that emerged are a special case of a local field theory, Section 5.2.2 on page 126, but this carries us well into the 19^{th} century and Chapter 5 on page 123. In the last century, the classical theories of light were superseded by two successful modifications of our understanding of light, the theory of Special Relativity which lead to a new interpretation of the nature of gravity and a full fledged quantum field theory of light, Quantum Electrodynamics or Q. E. D. The exposition of these precursor theories will take most of our time but is necessary to appreciate the nature of our final theories. Although the theories we describe here are not the modern theory, they are interesting predecessors to it, and they provide us with valuable insights into the fundamental concepts of the modern theory.

We will cover the thread of development from Fermat's and Fresnel's approaches to light in some detail here because they will allow us to develop both an intuition about these phenomena but also develop technical tools that are necessary to articulate the modern approaches. These theories also provide a wonderful example of how transition occurs in physical theory and we will see a similar transition to the modern theory. Hopefully, you will also see how the Fresnel approach was required to produce in the appropriate limit the Fermat theory. This is the usual case. The older approach had to have some successes or it would not be accepted. The identification of new phenomena that the older theory could not accommodate are the stimulus for the new approach. Despite its ability to accommodate the newly realized phenomena, the new theory must also fit the old successes. This later issue is often the hardest part in the development of a new theory.

3.2 Least Time Formulation of Light Propagation

. Fermat's 'Least Time Principle' describing how light travels between two points is an excellent example of a theory that agrees with the data and appears to be computationally simple. An interesting feature of this theory is the interaction of the development of the theory with the the concomitant development of new mathematical tools. This theory appears on the surface to have a simple and straight forward computational basis but on careful examination, reveals deep and subtle mathematical complications. This is also typical of all theory development – new mathematical understanding will generally be required for the successful implementation of the theory, see Section 3.3.7 on page 76.

The rule is stated very simply. If you want to find the path that light travels when it moves between any two points. You find all possible paths. On each path, find the time that it takes the light to travel over the path. The light travels between two points in space over the path that has the least travel time. This statement of the rule is so intuitive that two things tend to happen; you think that it is obvious or you tend to say that this is what *light* does.

This process of selecting a path on the basis of some extremum property is very common. You have often selected least time paths or, at least, a least something path. Maybe you want to conserve gasoline, or go the shortest distance, or avoid speed traps. But you have an extremum rule. This is a satisfying way for choosing an action. Similarly, you then feel that it makes sense that light would do this also. There are several problems with this idea. There are lots of choices about what to extremize. Not only that but it implies an anthropomorphic basis for the behavior of inanimate phenomena. But realize that Fermat is not saying that the light calculates the travel time on each path and then selects the least time path. He says instead that, if **you** want to find out the path, you must identify all paths, a prodigious undertaking, see Figure 3.1 on page 59, and actually very subtle issue, know the speed with which the light travels at all points in space, calculate the time for each path, and finally pick the path with the least time. Clearly, *light* does not do all these things. These are activities of people. It is interesting to point out that for light to do this calculation of time on all paths and choosing the right one, we need a natural argument for how light does this. Interestingly, this was accomplished by Fresnel, see section 4.6.5 on page 117. This is an valuable example which we will discuss of how a new theory recovers and clarifies the older theory. Regardless, these least time paths that light travel over are called the rays of the light and they are where the light goes in the Fermat picture. The experimental verification of the predicted path is to place a barrier at a point on the path and see if the light no longer connects the two points that were the end points.

This formulation of the rule raises many interesting conceptual questions beside the anthropic one of how the "light" does it. Note that it is formulated in such a way as to specify where the light goes between two points. This algorithm does not start at one point and in a direction and decide point by point how the light progresses; it does not propagate the light. This rule is not a local rule which is the usual way that we look at how systems develop. This makes it what is called a global rule. You need to determine the time of travel for the total path. You start with two points that are well separated in space. Of course, once you know the path, you can come

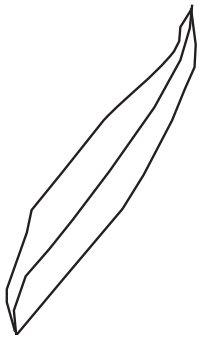

Figure 3.1: **Least Time Path** When light travels between two points, it travels over the path that requires the least time of travel. In order to find that path, simply find the travel time over all paths and choose the minimum of the set.

back and apply it to any pair of points along that path and, in all cases, the segment of path that was obtained as the least time path is also a minimum in family of paths between those points. This holds no matter how closely placed the new points are so that the rule can take on a local character. The only problem is that this local information can be obtained only after the path between the two separated points has been found.

Just how simple, algorithmic, is this rule. This seems algorithmic since it is a prescription that anyone can follow. You have followed it many times when you pick a travel route between two cities. What's the best way to go between Austin and Houston? You take a map with all the roads indicated on it. You classify all routes. On any route, you divide the trip into segments and then estimate your speed in each segment. From the speed and the length of the segment, you can calculate the time for that segment and then you add up the time for each segment to get a total.

$$T(route) = \sum_{\text{segments}} \Delta t_i = \sum_{\text{segments}} \frac{\Delta s_i}{v_i}. \tag{3.1}$$

where Δt_i is the time in each segment labeled i and Δs_i is the length of that segment i and v_i is the speed in that segment. You somehow make an ordered list of routes and repeat this process for all routes. Once you have $T(route)$ for all routes, you look down the list of travel times and select the one with the least time. That is the route that you take if you want the least time. In a similar fashion, if the light goes between two points, with this algorithm, if you know the speed of light at every place, you can find

the routes in space through which the light travels.

Is it really this simple? First, let's take a closer look at the algorithmic nature of this process. Despite its apparent simplicity, is it really well defined? It requires that we look at all paths. How many paths are there? A lot. In contrast to our highway problem, there are an infinity of paths. The problem of making sure that you have all paths is a complex one, and we will reserve a detail discussion for later, Section 3.3.7 on page 76. Just be assured that the requirement for an examination of all paths is not simply met and, in fact, this is one of those cases in which new mathematics had to be developed to meet the needs of the physics. Related is the fact that need to make a table of paths so that we can scan down it to make the choice of the least time, i. e. time as a function of path. This requires that we are able to make an ordered set of paths. Is this always possible or even ever possible? Again new mathematics will be required. It is worth noting that, generally when you do the highway problem, you have so few paths that you can keep track of them in your mind and maintain an order in that fashion.

Another problem is that, for each path in order to calculate the time that it takes the light to travel from end to end, you must know the speed of light at each point on the path and, since you must do this for all paths and since in the family of all paths all the points in the space will be touched. This implies that you will need to know the speed for all points in the space. A great deal of information. Also in a manner similar to the highway problem, you need a speed for each segment. This implies that you must also sensibly rectify the curved parts of the path, see Figure 3.2 on page 61. This is because of both the variation in the speed of light at different points and the curvature of the path. This is what you do when you calculate the time of travel between two cities. You add up the segments with comparable speeds and you make the curved parts out of straight segments that approximate the path.

How do you decide how big to make the straight pieces? You should pick the sizeof the straight intervals so that they follow that path with close precision and so that the speed of light is reasonably constant throughout the segment. Depending on the precision that you need when you calculate the time you may use a more coarse or a more fine grid.

With this done, you can calculate the time Δt_i in any straight line segment where Δs_i is the length of the straight line segment and v_i is the speed of light in that segment:

$$\Delta t_i = \frac{\Delta s_i}{v_i} = \frac{\sqrt{(x_i - x_{i-1})^2 \ + \ (y_i - y_{i-1})^2}}{v_i} \tag{3.2}$$

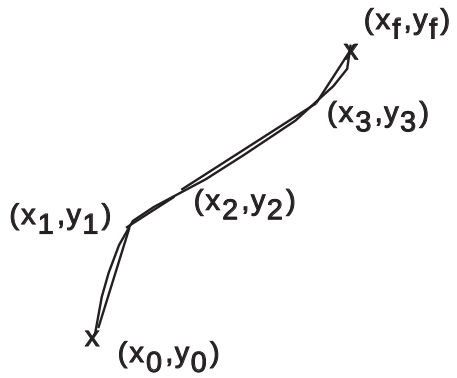

Figure 3.2: **Least Time Path in Inhomogeneous Medium** In order to calculate the time over a curved path in an inhomogeneous space, a space in which the speed of light varies from place to place, you must sensibly rectify the path, i. e. reduce the path to a set of straight line segments. The length of the segments depend on the precision of the calculation and how it is impacted by the variation of the speed and the amount of curvature of the path.

and then add the time for all the segments to get the total travel time, $T(path)$,

$$T(path) = \sum_{\text{segments of path}} \Delta t_i = \sum_{\text{segments of path}} \frac{\Delta s_i}{v_i}. \tag{3.3}$$

where I have now added the phrase "of path" to the right side of the equation to emphasis that throughout the computation we must somehow keep track of the specific path from a large family of paths with which we are dealing. Then, by means of some protocol, we select the path with the least travel time and call it the "least time path" and, according to the theory, this will be the path that the light will travel. In other words, placing a obstacle at a point in this particular path will extinguish the light between the two points.

This hypothesis "explains" reflection, refraction, and many other optical phenomena. Although, in the years since the 17^{th} century, we have developed several layers of superseding theories of light phenomena, the Fermat rules are still the basis for much of the design of optical instruments and the basic explanation for many atmospheric optical phenomena. In the following sections, I will illustrate this. This is not to say that only Fermat's

Theory will explain these phenomena. In fact, a particle based theory would do as well. As stated earlier, the clinching evidence for the Fermat theory was the requirement that to get refraction required that light in the more dense media traveled slower. This is often the case. People will cite examples of phenomena "explained" by a theory that works but a satisfactory explanation is not unique to the theory being cited. There are usually small but compelling differences when different theories are in competition. It has to be this way or there would be no competition.

3.2.1 Speculation on the form of Fermat's Theory

Since this is our first attempt at theory construction, it may be appropriate to speculate on the nature of this construction. Firstly, this is not a Newtonian approach which is local at each place. The path that nature chooses for the light is based on a global measure – the total time of travel of the path. Newtonians would have had the light move from place to place by means of some rule that held at each place at each instant. In Section 6.1.5 on page 162, we show how another global extremum rule, a rule about least action, similar to this one, can recover a local statement about how the system develops. Generally, the idea is that, if we can assume that the extremum is reached smoothly in the very rich path space, then paths which differ slightly have the about the same value. In particular, the requirement that two paths that are the same everywhere except at an isolated point and that the deviation of the path at this point is small then the global measure has almost the same value implies a condition that constrains the effects at that point[3]. This constraint is a local statement on the path development. This result is intuitive from our experience in finding least time paths for travel. The least time path for a trip is always made up of segments that are themselves the least time path between the points at the ends of that segment. In other words, the least time path is always made up of locally, between nearby points, paths that are the least time between those points.

A similar observation is that, although the word time is an important part of the formulation of this rule, there is no real time involved. By this I mean that there is no real evolution of the system. The path is what it is. The time in this approach is just some global measure on path space. This observation is especially relevant when we realize that, at the time

[3]Another way to look at the same result, is to realize that, once you have identified the natural path, the segment of the path between any two points on the natural path is a natural path between those points. If the two points are arbitrarily close you have the above result.

of Fermat's formulation, the speed of light had not been measured. The situation was worse than that. At the time, it was not clear whether or not light even had a velocity. On some occasions, Descartes who was the pre-eminent natural philosopher of the time argued the light was instantaneous and at other times he argued that light had a finite velocity. I have to assume that Fermat choose time as the measure because he knew that there were circumstances in which length did not work and what else could it be. Thus he formulated a global measure which weighted each path segment with the inverse of velocity, $\frac{\text{time}}{\text{length}}$, and then predicted that, if it could be measured, you would find that light traveled slower, a higher inverse velocity, in dense media. Of course, this was his great success. It still leaves the question of what other measures are there. We know from our experience planning travel that all kinds of measures are possible. Instead of least time, there is the most scenic route. In that case, we would develop a measure of scenic, for example hilly, and apply the measure $\frac{\text{hilly}}{\text{length}}$ to each segment and add the contributions. We could even count unpleasant scenery as negative hillyness and develop a measure that can have either sign. It will turn out that, when we expand our study to include dynamics, we will need a new global measure in a path space in space-time called the "action" and we will find that the naturally occurring path in space-time, called the trajectory, is the one that is the least action.

3.3 Applications of Fermat's Principle

3.3.1 Light Travels in Straight Lines

Let's start with the simplest observation. What are the paths of light in a homogeneous medium? A homogeneous medium is one in which every point is the same. In particular, the speed of light must be the same at every point. Thus, in this type of medium, the least time path is the same as the shortest path. By definition, the shortest path is a straight line.

Proof:

$$T(path) \;=\; \sum_{\text{segments}}^{\text{path}} \frac{\Delta s_i}{v_i} \tag{3.4}$$

I have added the path designation to remind you that you must do this for each path. Based on the fact that all the v_i are the same at every point in a homogeneous medium, the $v_i = v$, the common speed for light for allpoints in the medium, and can be factored from the terms in the sum and you will

have:

$$T(path) \ = \frac{1}{v} \sum^{\text{path}} \text{segments } \Delta s_i \tag{3.5}$$

Since $\sum^{\text{path}}_{\text{segments}} \Delta s_i$ is the definition of the length of the path, we see that the time for any path is proportional to the length of the path. Thus the least time path is the shortest-length path which, by definition, is the straight line path.

3.3.2 Refraction & Snell's Law

Refraction is the phenomena that occurs when light passes through a medium that has a varying speed for light. In this case, the ray bends. As the simplest case, chose a system of two media that are themselves homogeneous, separated by a planar interface, and place the two end-points in the different media. Both media are homogeneous, but they have a different speed for light called v_1 in media 1 and v_2 in media 2.

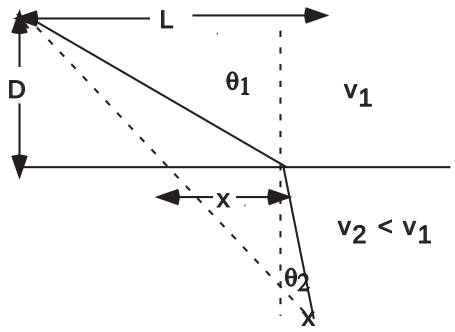

Figure 3.3: **Light Path in Two Homogeneous Media** The light path in two homogeneous media is straight in each part but kinked at the interface. In this example, the starting point of the ray is a distance D from the interface on each side and separated by a distance L measured along the interface direction. The distance along the interface from the point at which a straight light path would strike the interface plane and where the path strikes the interface is x. The angle between the normal to the interface and the path segments in each media are θ. The media are labeled by the speed for light ineach media, v_1 and v_2.

Our first problem is to determine how to discuss the paths that connect the two points. There are an infinity of them, see Section 3.3.7 on page 76. Physical intuition tells us though that the least time paths in a homogeneous

medium must be straight lines and thus the path with the least time overall must be among the paths that are straight within either of the two media and kinked at the interface, see Figure 3.3 on page 64. A path that is curved in one of the media would clearly be a longer time path than the one with the same start point and hitting the other media at the same point and then traveling in the second media. This is an example of how a global rule does have some local content. This ability to reduce the path space to kinked straight line segments is an important reduction in the nature of the problem. With this reduction in the size of the path space, we can label the paths with the distance of the kink position from the place at which the path would meet the interface if the tow media were the same, i. e. the straight line path between the two points, see Figure 3.3 on page 64. Two things have been accomplished. We now have an ordering for the family of paths that we wish to investigate. Even more significantly, we have reduced the path space to one that can be mapped onto the real line. In this case, we are labeling the paths with the parameter x. Remember that functions are mappings of the real line onto the real line. This then gives us access to all the usual tools of mathematics.

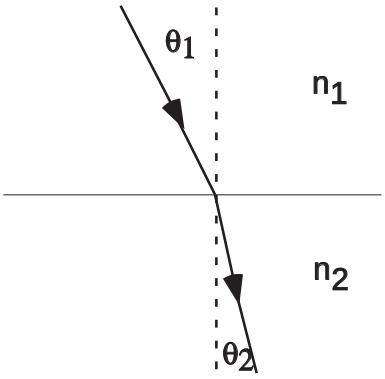

Figure 3.4: **Snell's Law** Snell's Law states that, when light passes from one optical medium to another, the ray of light bends at the interface according to $n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$, where θ_i is the angle of the ray to the normal at the interface and n_i is the index of refraction for the material.

Once the path has been reduced to two straight line segments, it is easy to find the least time path. In this example for simplicity of analysis, I will pick two points that are equidistant from the interface as measured along the normal to the interface and that distance is D. The two points are a distance L apart as measured along the interface, see Figure 3.3 on page 64.

The time for the path with intercept x is

$$T(paths) = T(x) = \frac{\sqrt{D^2 + (\frac{L}{2} + x)^2}}{v_1} + \frac{\sqrt{D^2 + (\frac{L}{2} - x)^2}}{v_2} \tag{3.6}$$

The least time path is the one that has the minimum value for $T(x)$ for all x. This is the$x$ value at which the slope of the T versus x curve is zero. The easiest way to find the slope means taking the derivative of T with respect to x. This is a small bit of calculus which I do not expect you to carry out. You can check my calculus if you like. I just want you to accept that it can be done and agree that the derivative is the slope and that the minimum occurs when the slope is equal to zero.

Taking the derivative, you get

$$\frac{dT}{dx} = \frac{(\frac{L}{2} + x)}{v_1\sqrt{D^2 + (\frac{L}{2} + x)^2}} - \frac{(\frac{L}{2} - x)}{v_2\sqrt{D^2(\frac{L}{2} - x)^2}}. \tag{3.7}$$

Setting the derivative equal to zero, and solving for the path with the least time yields

$$\frac{(\frac{L}{2} + x_0)}{v_1\sqrt{D^2 + (\frac{L}{2} + x_0)^2}} = \frac{(\frac{L}{2} - x_0)}{v_2\sqrt{D^2 + (\frac{L}{2} - x_0)^2}} \tag{3.8}$$

where x_0 is the label of the least time path. Using some simple trigonometry, we can relate the angle of the least time path with the normal at the interface to x_0, see Figure 3.3 on page 64. From the figure, we have that $\sin(\theta_1) = \frac{(\frac{L}{2}+x_0)}{\sqrt{D^2+(\frac{L}{2}+x_0)^2}}$ and $\sin(\theta_2) = \frac{(\frac{L}{2}-x_0)}{\sqrt{D^2+(\frac{L}{2}-x_0)^2}}$ so that

$$\frac{\sin(\theta_1)}{v_1} = \frac{\sin(\theta_2)}{v_2} \tag{3.9}$$

or

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \tag{3.10}$$

where $n_i \equiv \frac{c}{v_i}$ and is called the index of refraction. c is the speed of light in a vacuum. Since $v_i \leq c$, $n_i \geq 1$. This is known as Snell's Law, see Figure 3.4 on page 65.

Following any derivation, it is useful to see if this agrees with our intuition. The light wants to spend the least time traveling between the two points. It is better to have more distance in the faster medium. Think of the lifeguard at the beach. She sees someone off to the side drowning. Although

she is a good swimmer, she can run faster on the beach that she can swim. Therefore, instead of going directly to the person drowning, she runs a little further up the beach past the point on the direct line to get to the victim in the shortest possible time.

It is worthwhile to note that, in the particulate theory of light, the path of the particles is bent toward the normal by the fact that the particles travel faster in the dense medium. Once it was found that light travels slower in the dense medium, the particle theory was not tenable. This is an often cited example of the Popper hypothesis of the use of falsifiability to prove or disprove theory in physics, [Popper 1973].

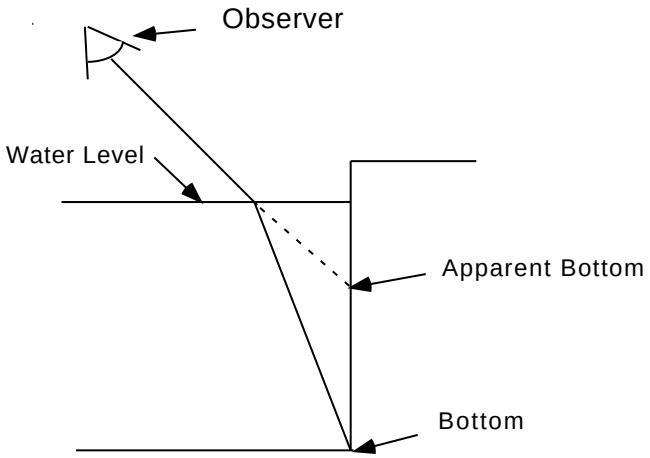

Figure 3.5: **Apparent Depth in Water** To an observer outside and above viewing a pool of water sees it as much shallower than it is. This is because the brain reconstructs the light ray that comes to the eye as a straight line path. Since the density of water is greater than air, the ray from the intersection of the bottom and the side of the pool and the eye of the observer is bent toward the normal in the water.

A direct application of Snell's Law is the observation that when viewed from outside a pool does not appear as deep as it actually is. The ray from the edge at the bottom of the pool to the eye is refracted, see Figure 3.5 on page 67. Since the speed of light is lower in the water than in the air, the ray bends away from the normal in the air. The observers brain assumes that light travels in straight lines and thus places the intersection of the side and bottom of the pool at a much shallower depth.

The discerning reader may protest the interpretation of the apparent depth above does not make sense. A single ray cannot determine a point, in

this case the intersection of the bottom and edge of the pool. To find this point at the bottom of the pool, you need the intersection of at least two rays. As we all know, for humans the trick of depth perception is binocular vision. Even with one eye, the seeing involves the collection of several rays around the one shown and then the lens in the eye provides a measure of distance. Thus there are other rays that are not shown in Figure 3.5 on page 67 and these in combination with the ray shown ultimately determines the depth. This is a general truth. To find an image you require a collection of rays. Again, the discerning reader may note that, even without the binocular vision argument above, there really are many rays from the bottom edge point coming out the pool. These range in angle from the ray we have shown to rays that do not hit the eye but are from the bottom edge. A particularly simple ray from the edge is the one that runs along the side of the pool from the bottom straight up. Since this is parallel to the normal it is not refracted. It is also not detected by the eye. Regardless, there is a whole collection of rays fanning between this ray and the one depicted in Figure 3.5. The eye collects the rays from this fan that are near the ray depicted. The lens in the eye focuses these rays to a point on the retina. Its intersection with the refracted ray determines the position of the point on the bottom.

3.3.3 Total Internal Reflection

There is an interesting case of refraction that can occur when the light exits a dense or slow medium into a less dense or faster medium. Rearranging Snell's Law, Equation 3.9 on page 66, $\sin\theta_2 = \frac{v_2}{v_1}\sin\theta_1$, we can see that there can be cases in which θ_2 cannot be found.

If θ_1 is large and, thus, close to $\frac{\pi}{2}$, $\sin\theta_1$ will be close to one. Since for this case $v_2 > v_1$, $\frac{v_2}{v_1} > 1$. In this case, the product of the two terms in the rearranged Snell's Law could be greater than one and since the sin function is always less than or equal to one, there is no angle θ_2 that can satisfy the law.

In this case, the light does not penetrate the less dense surface but instead reflects from the surface with the surface acting as a very good mirror. As we see in the a later subsection, Section 3.3.6 on page 72, in mirrors the angle of incidence and the angle of reflection are the equal. The angle of incidence in the dense medium that just skims the interface is called the critical angle, θ_c.

This effect is illustrated in the pool problem of the previous section, Section 3.3.2 on page 64. In Figure 3.6 on page 3.6, rays rising from the

bottom/edge, fill all the space above the water. An observer in this space looking down will see the bottom/edge but not at the actual depth. On the other hand an observer in the water sees a more complex situation. Looking down the observer anywhere in the water will see the bottom/edge directly. An observer in the water far enough away from the edge sees the bottom/edge and the sky, maybe more likely the local vegetation, in the same direction. Note that closer to the edge of pool there is no internal reflection and he/she will see only sky or vegetation. In fact, for an observer near the surface, he/she will see almost all the way up to the edge. The situation is an actual pool is even more complex since the surface is rough and changing. The swimmer whose eye is close to the surface sees a rather chaotic mix of sky and bottom.

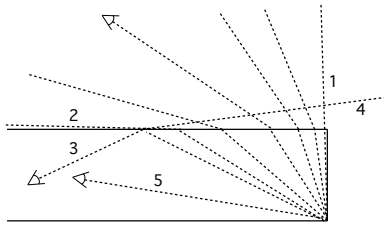

Figure 3.6: **Total Internal Reflection in a Pool** In the pool edge problem of Figure 3.5 on page 67, light leaving the bottom/edge point spreads its fan of rays. Ray 1 is normal to the water air interface and starts the fan. For ray 2, the angle of incidence is just below the critical angle, θ_c and the emergent ray skims the surface. Ray 3 is a ray that has been totally internally reflected. Ray 4 is a ray from the sky that is along the same direction to the underwater observer as ray 3. Ray 5 is from the bottom/edge directly to the observer. An observer above the water looking down can see the bottom/edge from anywhere. An observer in the water looking down can see the bottom/edge from anywhere. An observer in the water, if he/she is far enough away, looking up can see the both the bottom/edge and the sky and coming from the same direction. Which or these is perceived will depend on their relative brightness.

3.3.4 Lenses

We are all familiar with lenses. We have them in our bodies, our eyes. Many of us have them hung on our faces or added to the surface of our eyes to help the working of the biological lens. In all these cases, they are used to bring

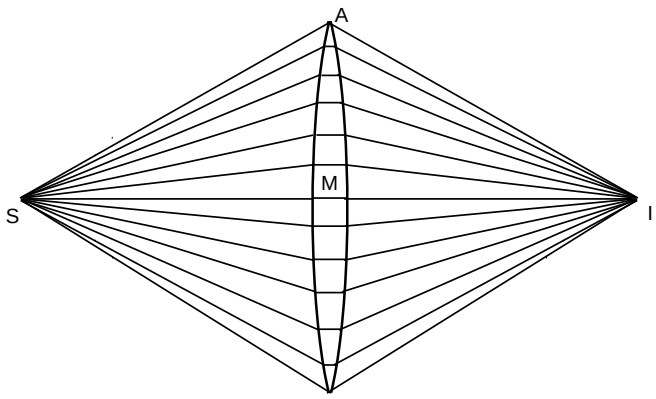

Figure 3.7: **Light in a Lens** The path that goes from S to A to I and the path that goes from S to M to I have the same time. All the rays shown in this figure are between focal points, S and I, and the thickness of the lens is adjusted so that the time for each path is the same.

a spreading beam of light from a source, the object, into a more narrow region and concentrate it at a different place, the image. In the case of our eyes, the deposition of the fan of light starting at a leaf of a tree intercepted by lens is focused on to the retina as an image. Generally all lens are for for imaging or for the concentration of the light energy, focusing, but some lens designs are for the opposite reason, the spreading the light. A lens is another example of a system consisting of two different homogeneous media interacting. The difference with the discussion of the previous section is that here the interface between the two media is curved. The Fermat explanation for focusing or concentrating the light energy is that the glass of the lens is shaped so that all the rays between two points of interest, the object place and image place, that are intercepted by the lens have the same travel travel time and that are on the axis of the lens have the same time of travel, see Feynman [Feynman 1985].

For the configuration shown in Figure 3.7 on page 70, without the lens, the axial ray would be the least time ray and the only one between the points. By placing glass in the path, the time is increased for this ray. In a similar fashion, glass but with a smaller thickness is placed in the way of each of the other paths between the two points in precisely the manner that each path has the same travel time. We will carry out the details of this

computation in a homework assignment. In this case, all the rays that pass through the lens are least time rays. Note how this explains why, when you block a portion of a lens, you do not block a portion of the image but only decrease its brightness. It also explains the concentration of the energy, rays which without the lens would have gone to other points also act at the same point. When we get to the Fresnel/Young/Huygens construction, Section 4 on page 83, we will discover an even more compelling interpretation of the operation of the lens.

3.3.5 Rays in a General Inhomogeneous Space and Mirages.

An inhomogeneous space is one in which, at different places, the light travels with different speeds. In the previous example, we discussed the most trivial example of an inhomogeneous space, two homogeneous media with an interface. In a general inhomogeneous space, the speed of light can vary at each point in the space and you have to calculate the time for the path carefully.

After you select the path, to calculate the time over the path, you must rectify the path and note the different speeds in each segment before adding the times of all the parts, see Figure 3.2 on page 61 and Equation 3.3 on page 61. You select the segments on the basis of the curvature of the path and the rate at which the speed of light is changing. You are working with a certain precision and the length of the segment of path must be the same as that of the straight segment and the speed of light can only vary over the segment within the desired precision.

mirages

Figure 3.8: **Mirages** Due to the bending of light caused by the variation of the density of air and thus variation of the speed of light, to an observer looking down on a hot surface, the ray of light that comes to his/her eyes is not from the road surface but actually comes from the sky.

Mirages are a common experience for Texans. In the summer, the road surface gets extremely hot. A mirage is an example of a phenomena using

the two previous situations, an inhomogeneous space and total internal reflection. When the road is heated to a high temperature from the sun above it, it heats the air immediately over it and that air is thus less dense than air further up. The speed of light in air increases as density of the air decreases. A light ray moving down toward the road surface is moving from a more dense to a less dense medium and is refracted away from the normal. This bends it to larger angles to the normal as it goes closer to the ground and finally reflecting and turning upward. Therefore, for points over a hot road, the least time path is bent upward. This means that when you look down you are actually seeing the sky, and your brain thinks the shimmering blue of the sky is water on the ground, see Figure 3.8 on page 71. The blue spot that you see is shimmering because the less dense warmer air at the bottom is unstable under the dense cool air and the are rising air currents which cause the shimmer.

The opposite effect is associated with looking over a cool surface. An example is with a phenomena known as "ghost ships." In this case, the cool surface is the ocean in the early hours of the morning when the sun has come up to heat the air over the surface. In this case the temperature profile drops as you get closer to the surface and this bends the light ray downwards and the images of a nearby ship seems to hover in the air.

3.3.6 Reflection and Mirrors

Plane Mirror

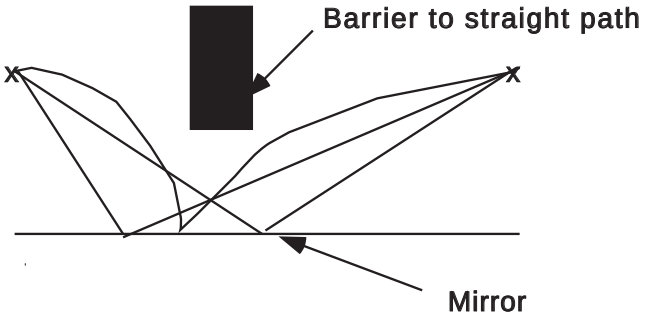

Figure 3.9: **Reflection** Light paths around a reflecting surface. Paths directly connecting the two end points of the paths are blocked by a barrier.

An optical phenomena that appears to be simpler than refraction is reflection. This phenomena is also easily seen to be consistent with the

Fermat's Least Time Principle but, since it was also consistent with the competing particle theory of light, we chose to cover the more complex case first. In the case of reflection, we want to find the light path between two points above a mirrored surface. The trick, in this case, is to realize that we must consider only paths that touch the mirror once. For example, we place a barrier between the points so that the direct path is blocked. The observant student might comment that even with the barrier in place, there are shorter time paths than those obtained by using the mirror. For example those that just graze the edge of the barrier. Why not select these? Later and for different reasons, i. e. diffraction in Section 4.6.3 on page 115, we will. For now though, we will just take it as our definition of reflection that the family of paths under consideration are those that touch the mirror once. Maybe reflection is not that simple after all?

Again, the path is in a region that is homogeneous and, thus, we anticipate that the least time path is the shortest distance path. In this case, you must use the mirror to get past the barrier. What is then the shortest path? Or better said, what is the shortest of all the paths between two points that touch the mirror at only one point? For the simple case of two initial points equidistant from the mirror surface and with a piece of string, it is easy to convince yourself that the path that touches the mirror at the mid point of the interval between the points is the shortest and it, therefore, makes equal angles of incidence and reflection. You should describe how you can use a piece of string to show this.

. Thus:

$$\theta_1 = \theta_2 \tag{3.11}$$

This apparently very simple rule can be used to interpret many interesting situations.

An image is formed when several rays from the same point are brought together by the eye. In addition, the brain extends each set of rays, so that it places the image where the set of rays converge treating them as if they were straight lines. This is similar to the cases that we had above with viewing the bottom of a pool and mirages. Consider the case of two plane mirrors that are perpendicular, using Fermat's Least Time and extending the rays as straight lines, you find three images in addition to the original object, see Figure 3.11 on page 75.

Curved Mirror

Now let's examine the case of a curved mirror. For example, look intoa spoon–the bigger and more polished the better. You see yourself shrunk

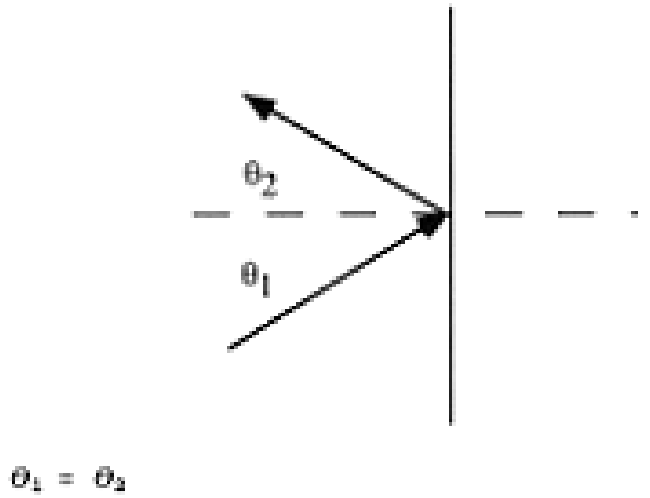

Figure 3.10: **Law of Reflection** For light reflecting from a mirrored surface, the least time ray is the one in which the angle of incidence with the normal is equal to the angle of reflection with the normal.

and upside down. The situation here is the reflection correspondent to the lens. The surface is curved in such a fashion that for the selected points, all rays have the same travel time. Why upside down and shrunk. Look at the light that comes from the tip of the larger arrow in Figure 3.12 on page 76. For this discussion to be strictly correct the arrow though large should be small compared to to the mirror radius. The big arrow is you, the object. The rule is simple. At the mirror the angle of incidence must equal the angle of reflection but, since the direction ofthe normal to the mirror is different at the different points on the surface, different rays reflect in different directions. The three rays that are shown are all least time paths. These three rays are representative and any ray from the large arrow will pass through the point of convergence, the tip of the small arrow. These three are shown because they are particularly simple to describe: using simple trigonometry and the bending of the mirror, it is possible to show that the ray that starts parallel to the axis always reflects so that it passes through the point at $\frac{R}{2}$, where R is the radius of the spherical mirror, from the axis; the ray from the object the goes to the vertex of the mirror reflects so that it is symmetrically located below the axis; and the ray that passes through the point on the axis at $\frac{R}{2}$ from the vertex will emerge on reflection

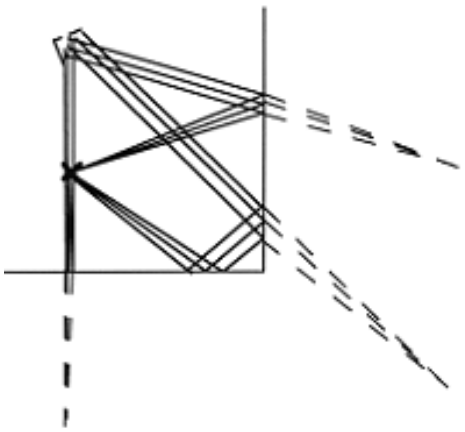

Figure 3.11: **Perpendicular Mirrors** An object viewed from the front of two perpendicular mirrors produces three images. There is the image in each of the two plane mirrors and the image produced by both mirrors.

parallel to the axis. The eye that receives these reflected rays reconstructs the image as the tip of the small arrow, upside down and smaller. Again, the rays that reach your eyes or are all least time rays. This is why when you cover part of the mirror or, in the case of the spoon, really only have a fraction of the mirror, you still see the entire image not a part of it. this is in contrast to the case when you remove part of a plane mirror. In this case, you lose part of the image. We will make more of this later, see Section 4.6.4 on page 117.

The idea of the brain reconstructing the image as the crossing point of the reflected rays reaches its extreme when you move the spoon closer. What happens? Using the same three rays, when the object arrow moves closer to the vertex of the mirror than $\frac{R}{2}$, the image is larger than the object and is upright. You have to get pretty close to the spoon for the image to make much sense and the image is usually too big to interpret as your face. In fact you need a really big spoon for this to work. The interesting thing that emerges from the diagram is the the reflected rays really never cross. Only the extrapolation of the rays beyond the back of the mirror cross. Thus this image is behind the spoon, region where they do not actually go, and is called a virtual image in this case. The name comes from the fact that in the case of the real image there is a point in space where the rays cross. In this case, you can put your finger there and destroy the image. For the virtual image, there is no point at which the rays cross; it was extrapolated

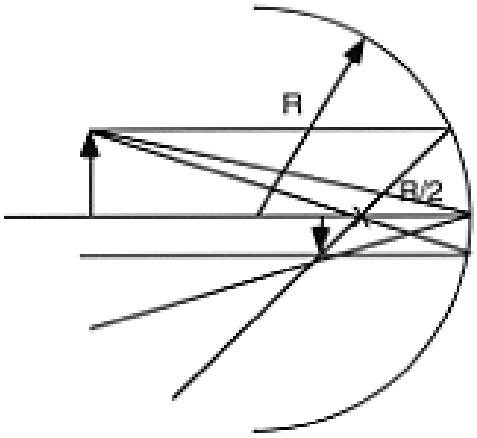

Figure 3.12: **Spherical Mirror** Light rays focusing an image near a spherical mirror. Three rays for which the angle of incidence is the same as the angle of reflection are shown. The axial ray passes through $\frac{R}{2}$ after reflection. A ray through $\frac{R}{2}$ produces an axial ray after reflection. A ray to the vertex produces a reflected ray that is symmetric below the axis. If the object is as shown further than $\frac{R}{2}$ from the vertex the image is smaller and upside down.

in your mind and placing your finger there does nothing to the image.

3.3.7 Mathematical Digression

In our articulation of Fermat's Principle, we casually assumed that it made sense to use the phrase "all possible paths" between two points. In a normal space, that's a lot of paths. To start with does it even make sense to identify "all paths". If you think about it, it means that somehow you produce an ordering so that you can go through the lists to examine all possible cases. An ordering is mapping of the paths onto an ordered set. Without much thought, it should be clear that there are a lot of paths – an infinity, a very large infinity. Are there too many paths to order them like the integers? Two common examples of large sets are the integers for a discrete but infinite set or the points on a line for an infinite but continuous set. The counting of infinite sets is a subtle issue. There are as many integers as there are odd numbers. That's because they can be ordered together – put into a one to one correspondence.

How do you determine the number of paths? You count them or order

them. Counting is a process of matching the elements of two sets, one the set in question, in our case paths, and a given set whose properties are better understood. The smallest of the standard sets of choice are the discrete infinite set that is the number of integers. Sets that have the same number of elements as this are relatively nice to deal with and once an identification with the numbers is established the elements can be manipulated like numbers. Sets of this size are said to be in the class $\aleph_0$. Anytime that you make a table, you are making a mapping between the set of integers and your set of objects that enter the table. If you have an ever larger set of objects, you have a set the size of $\aleph_0$ and you have ordered it with the integers.

In order to use the tools of analysis you need to deal with a system that has the right number of members. Functions are mappings of the real line onto the real line. The real line is, in fact, on example of the next larger infinite set, $\aleph_1$. It is bigger than the number of integers which also happens to be the same size as the number of rationales. The set made of all the points on the real line is the same size as the number of irrationals. A simple ordering argument, we can show that the number of points on a line and the the number of points in a plane are the same. Again, an example of a property of these infinite sets that is not intuitive. In other words, there are as many points on one line as on any countable number of lines.

It is relatively straight forward to convince yourself that the number of paths is larger than the number of points on a line. This makes for a problem. Most of what we can do in analysis is dealt with through functions. By definition, functions are mappings of the real line on to the real line. Thus, our manipulations with paths cannot be considered functions and all the things that we learned about the manipulation of functions does not hold. Mappings of path space onto the real line are called functionals and thus our ambition of finding the least time as a function of path is a functional.

In our first example, refraction, we used our intuition to label the paths as the same as the point of intersection of the path with the interface of the media. This is clearly only a small sample of all the paths. The important point about our selection of the point of intersection was not only for convenience, it was a reduction in the size of the path space to one that allowed it to only the same number of paths as the points on the real line. This choice allows us to write the time T as a function of x, a point on the real line, $T(x)$. Thus although it is nice to think of x as the distance along the interface, its real role is as a label in path space and one that is only $\aleph_1$. This makes T, a real number, into a function in the sense that it provides a mapping of x onto the associated time for the path. We are then free to use usual mathematics to find the minimum.

Reiterating, in general, time over the path is it not a function of the path, because the number of paths is greater than the number of elements in $\aleph_1$; there is no rule for matching paths with points on the real line. The number of paths can be quite large and, using some other information such as our intuition, depending on the restrictions that you put on the family of paths that you consider, the family will be in some class, $\aleph_i$ where $i \geq 2$. In these situations, you cannot call T a function. It is called a "functional" instead. That is to remind us that the ordinary procedures of mathematics are not adequate. Thus this very simple algorithmic looking rule,

$$T = \sum_{path,(x_0,y_0)}^{(x_f,y_f)} \frac{\Delta s_i}{v_i}, \tag{3.12}$$

is actually a complicated mathematic structure. For us, being straight forward people with a simple outlook on life, we will ignore most of these complications and go ahead and, in all our cases find a family that is $\aleph_1$, when we are operating in path space. In other words, we will select some small class of paths and label them one or more intervals on the line. In this way, we reduce the functional to a function.

The other interesting mathematical feature of this supposedly simple algorithm is the need to evaluate a complicated object. These are the problems of sensibly rectifying path either because of curvature or the variation in the speed as the path moves through points in space. These issues were discussed earlier in Section 3.2 on page 57 and Figure 3.2 on page 61. The point is that, although it is often the case that a rule for interpreting a phenomena can be stated simply, there are often subtle issues that require a great deal of mathematical development to disentangle. Much of modern mathematics is devoted to the untangling of what appears to be on the surface very simple physics problems.

Finally, it is important to make clear what we are doing drawing the path the light travels over, the ray. Firstly, the attitude that this is the light gives us the ability to make the theory falsifiable in the Popper sense of falsifiability[4]. Put an obstacle on the path and the light no longer goes to the other point. It is stopped by the obstacle. The path, the ray, is the light. It is the 'reality' of the light. It is what light is made of. If block any point not on the ray we have not effected the light. It is a wonderfully simple thing that is the light. The way we measure light is it brightness.

[4]In his famous book *The Logic of Scientific Discovery*[Popper 2002] Karl Popper argues that science should adopt a methodology based on falsifiability, because no number of positive experiments can prove the truth of a hypothesis.

Brightness is a positive definite quantity and so is the number of rays. The concept of light as rays leads easily to the idea of geometric fall off[5] As we go further into our discussion of light, we will find that although, as a convenience, we draw a line connecting places and call it light. The line will have to become a much richer construct until we no longer call it a line but will call it a propagator which is a thing with much richer embellishments required to account for the observations. A simple example is that a line is either there or not, a one or a zero. We will persist in drawing lines when we need richer constructs.

3.4 Newton and Color

It is a common experience to use a piece of shaped glass, a triangular cut of glass called a prism, to produce a rainbow of color from sun light. This is commonly described in the following way: this is basically a refractive phenomena and a simple extension of Fermat's Least Time Principle can be used to describe it. A narrow beam of white light incident at a non-normal angle on one surface of the glass is refracted; the beam changes direction. The spread of color appears because the different colors in the light have different speeds in the glass with the blue being faster than the red and all colors slower than for light in air, see Section 3.3.2 on page 64. Thus the blue is bent less than the red. The separated rays then emerge from the other interface of the glass spread in this familiar rainbow pattern. This spread of color can be seen by placing a piece of paper after the second interface, see the first part of Figure 3.13 on page 80.

Actually Fermat did not describe the phenomena of color. Among the early studies of color and the best were carried out by Newton in a series of experiments over many decades that he brings together in his treatise called "Opticks", [Newton 1730]. Although his most famous presentation of physics was in his monumental work "Principia" his "Opticks" could be considered his best or even the best physics book ever, Newton developed an interpretation of the nature of light and its relationship to color phenomena. The beauty of the book is that as opposed to the "Principia" it deals with the observed phenomena directly and does not develop an underly structural basis "explaining" light's color behavior. Interestingly, it was Newton, the advocate of a particulate theory, who first articulated the ideas about the white light being composed of the colored components. Prior to Newton's

[5]Light from a local source, a point source, gives less brightness at greater distances. The rays spread out.

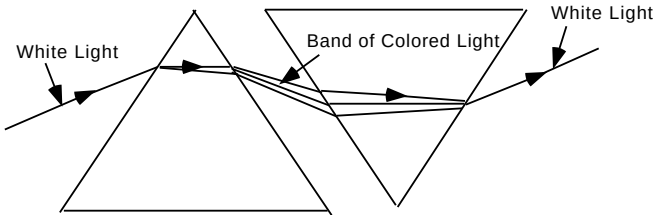

Figure 3.13: **Newton's Experiment with Light and Color** Newton's experiment showing that light is composed of colored components. A narrow beam of light is incident on a prism and produces a broadened and colored band which can be reconstituted back into a narrow white beam of light with a second prism.

interpretation, the idea was that the color in the prism came from the glass and was not an intrinsic property of the light. To show otherwise, Newton placed two prisms, in the path of a narrow beam of sunlight. The beam emerging from the first prism was traveling in a different direction from the original beam, as expected from refraction theory, either by particulate or least time principles. As usual, the beam was spread over a band of angles and, when a piece of paper is located in the beam, after the first prism, a broad smear of light appears and the different parts are a different color, the rainbow alluded to above. Newton then went one step further and inserted the second prism and allowed the spread beam to enter it. When arranged carefully, he found that this reconstituted the original beam in the original direction, see Figure 3.13 on page 80. Newton's interpretation was that the color was intrinsic to the light and; in other words, white light has constituents which we perceive as the colors; and the bending in the glass spread the constituent parts differentially to spread the beam. The same process reversed was then able to reconstitute the beam of white light.

In the Fermat least time approach, the blues travel in the glass at a faster speed than the reds and thus the blue colors are bent less by the prism,. In the particulate theory, the blues would travel in the glass slower than the reds. This is an example of a phenomena the is consistent with both interpretations but for different reasons. This difficulty could not be resolved until it was possible to measure the speed of light in materials.

Regardless, it was Newton who realized that white light was a complex phenomena and that white light was composed of an internal structure – the colors. This realization had immediate and important impact in the interpretation of visual phenomena. You saw different colors because, by

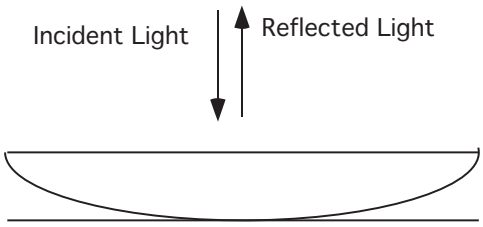

Figure 3.14: **Newton's Rings** Newton's set a curved lens on a plane glass surface and illuminated it from above. When viewing the reflected light from above, there is a series of rainbow colored rings surrounding a central dark spot.

some mechanism, you removed from the white light the other colors or you created color by combining various components such as red and blue to make purple.

It is important to realize that all of this science of color is independent of our modern interpretation of color as the frequency of the oscillations of the light. That came later although the seed had been planted by another observation of Newton. In another experiment, Newton placed a small lens on top of a plane glass surface and illuminated the combination from above, see Figure 3.14 on page 81. Viewing the reflected light from above, there are a series of concentric rainbow colored rings around the central dark spot. This is a direct indication that the color label of the components of the light can be associated with another feature of the light – length. The idea is that because of the curved surface of the lens at different distances from the center, the light, reacts differently depending on the color. Moreover, this phenomena is periodic with the reds repeating as multiples of the gap between the lens and the plate varied. This implied that there was associated with color some sense of length, i. e. the different colors fit the varying places better. We, of course, realize that having a length and speed, the speed of light, is equivalent dimensionally to having a time. The light has components and these are labeled with a time. With the full development of the wave approach of Thomas Young and Christian Huygens, see Section 4 on page 83, color became identified with a very specific interpretation of the time, the time label was the period repeat of the wave at any place, the frequency.

Thus Newton described the basic phenomena of color: White light from the sun or most other luminous bodies is a composite system. There is an internal constituent that is recognized as the color. The different colors can be labeled with a continuously varying parameter, an element of an $\mathbf{R^1}$, that

represents a periodic structure, repeats, as manifest in the Newton's Rings experiment shown in Figure 3.14 on page 81 . The color identification can be characterized as a length peculiar to that color called the wavelength λ or, since light is identified with a particular speed, the characteristic length can also be identified with a time, T, which is now identified more often with its inverse, a frequency, $f = \frac{1}{T}$. Regardless of the interpretation, the length label, λ, and time labels, T or f[6], that designate the color are connected by the speed of light,

$$\frac{\lambda}{T} = v_{light}, \tag{3.13}$$

which could be slightly different for the different colors, labeled by either λ or T, and for different media. We cannot understand this variation in the velocity until we understand better how the light interacts with media through which the light moves. In a vacuum[7], there is no medium and thus mo interaction with and the speed of light is the same for all colors. Air is such a tenuous media that for all but the most precise measurements light shows no effect on the speed making it difficult for people of Newton's time to observe these phenomena in air.

[6]There is an entire vocabulary associated with labeling periodic entities and even idioms in usage. The frequency, f, as defined above is the cycle frequency with units $\frac{cycles}{sec}$ or in SI Hertz. This is the common engineering unit. For physicists, the preferred frequency is the radian frequency, ω or $\frac{2\pi}{T}$, which has units of $\frac{radians}{sec}$. This is preferred since the argument of a harmonic function should be a radian.

[7]In modern interpretation, the vacuum is a rather richly structured entity whose properties effect the behavior of the things passing through the vacuum. Chapter 22 on page 483 and Chapter 23 on page 491 discuss these issues.