Chapter 4

19th Century Physics

4.1 Action at a Distance and Field Dynamics

The previous construction of Fresnel/Young/Huygens tell us how to construct an amplitude for light at any point in space given the amplitude at some other point in space. This is the first part of the construction of a field. A field is something, generally a measured quantity, that is defined at every point in space. At each point in space you can measure the entity. In addition, as you move from one point to a nearby point the value of the something changes smoothly; it varies as you change places. There will even be a rule on how the change as you move from point to point is manifest. To appreciate these rather abstract comments let's look at several examples.

There are numerous examples of fields. The temperature in a room is a field. Temperature is measured for instance by a mercury bulb thermometer. As you move the thermometer from point to point, you will get different values for the temperature. If the room is not too drafty, the temperature at nearby points will be similar; the temperature varies smoothly as you move to nearby points. You can even intuit certain rules for how the temperature changes as you move from point to point. For instance, you can guess that a point at the center of a surrounding group of points, the temperature will be the average of the temperatures of the surrounding points. It is because of rules like this that you expect that the temperature varies smoothly as you go among nearby points. Other obvious examples of fields are air pressure in a room, height above or below the normal height of water in a pool, or the transverse displacement of a stretched string. With some amount of smoothing you can make a field from such things as population density on the earth. Any system that is defined over a continuous manifold is a field.

The discussion of the previous examples generally did not deal with the time variation. It is not until we endow something with a time dependence that the something becomes interesting. In fact, as we will see, Section 5.4.4, we cannot really talk about energy until we have temporal evolution. In the Fresnel/Young/Huygens construction of the amplitude for light, we eliminated the effect of the time variation by "seeing" only the brightness, the amplitude squared, and averaging for long times so that the short time oscillations of the phasers cancelled out, Section 3.5.5. Thus although the brightness as a field can be interpreted as slowly varying there is an intrinsic time variation that makes light especially interesting.

In other words, a field is something that is defined over some manifold, usually space, that has a temporal evolution. The rules for the behavior of the field are usually local in the sense that its variation in space and time is determined by what is going on at those points of space at those times. This is the meaning of local causality. It is one of the bedrock principles of modern physics. It ranks with reductionism as one of out formulating rules. The basic idea is that what happens to an entity happens because of what is going on at the place at which the entity is or the immediate neighborhood. This is in sharp contrast to the situation in theories that are based on action at a distance dynamics. Newton's Laws of gravitation are an example of an action at a distance theories. To a large extent, it was the attempt to remove these action at a distance formulation and replace them with locally causal theories that motivated the development of field theories.

4.1.1 Action at a Distance

My former colleague, Johnny Wheeler calls it "spooky" action at a distance. Newton, its inventor, was not comfortable with the concept but could not come up with something better. In a letter to the theologian Robert Bentley, he wrote:

> that gravity should be innate, inherent and essential to Matter, so that one body may act upon another at a Distance thro' a Vacuum, without the Mediation of any thing else, by and through which their Action and Force may be conveyed from one to another, is to me so great an Absurdity that I believe no Man who has in philosophical Matters a competent Faculty of thinking, can ever fall into it. Gravity must be caused by an Agent acting constantly according to certain laws; but whether

> this Agent be material or immaterial, I have left the consideration of my Readers.

Regardless of his own reservations and because of the success of the Newtonian approach, physicists became accepting of the anomalous nature of action at a distance and the early formulations of most laws were all in the pattern of action at a distance. Fortunately, Maxwell could not believe these and, for the case of electricity and magnetism, this lead him to the development of the first first-principle field theory. Prior to Maxwell's work there were field theories but these were derivative of an underlying structure. For example, the rules of fluid flow were formulated in a field theory vocabulary. But this was understood to be a consequence of the underlying structure of the fluid. Maxwell's formulation of the nature of the electric and magnetic systems was actually a statement on the intrinsic properties of these entities. In order to understand this important idea let's review the situation with action at a distance theories and the contrast to field theories.

All the satisfactory theories prior to the 19th century were not what we now call locally causal theories but instead were bases on action at a distance theories, actions resulted from situations that were at a distance from the object of interest. Newton's theory of the gravitational force is a perfect example. In Newton's approach to gravitation, a bodies motion is determined by the separation from a remote other body at the instant under consideration. The moon determined its acceleration from knowledge of the earth's position which is at a distance at that instant. It is hard to accept that, if the earth suddenly ceased to exist that, at instant, the moon would instantaneously react by traveling off in a straight line, no longer in orbit. There are two issues here. First the idea that somehow that moon is influenced not by things going on where it is and the fact that the earths disappearance should be realized by the moon instantaneously; it should take some time. Consider the case that I am standing in the front of the lecture hall and announce that I am going to make the clock at the back of the room run differently. If I could do that, you would infer that I had a wire or used sound waves or some other mechanism to communicate the change to the immediate vicinity of the clock. Whatever ultimately changed the clocks running was at the place of the clock not at a distance.

Coulomb's law and all the other laws of electromagnetism that were formulated before the 19th Century were action at distance laws. A charge here effected a charge there.

The solution to this basic philosophical conundrum is in the idea of strict locality for all phenomena and the vehicle is the concept called the

field. Of course, in physics, a philosophic problem is not a good reason for doing something. The idea must be tested experimentally. The proof of the construction is in the testing. Through his treatment of electromagnetic phenomena as a field theory, Maxwell was lead to predict that light was a disturbance of the electromagnetic field. When this prediction was verified by Heinrich Hertz in 1887, there was a general acceptance of Maxwell's approach. Since that time, we have found that all fundamental theories are field theories; the ultimate modern expression of the nature of matter and energy being through the machinery of quantum field theory. For this reason, it is important to understand the idea of the field. For now we will develop the classical field, we will add the complications of quantum mechanics, see Chaper 18.

4.1.2 Local Field Theory

Maxwell developed a local field theory to describe the phenomena associated with what is called electricity and magnetism. He reduced all the known laws of electricity and magnetism into four reasonably simple equations. In so doing, he unified the electric and magnetic forces and predicted the fundamental nature of light. These are considerable accomplishments in their own right but also he somewhat inadvertently clarified the idea of the field and the idea of causality. His was not the first field theory; it was the first field theory of a fundamental force system. The first local field theory and the easiest to appreciate was the description of fluid flow. It was the success of a field theory of fluid flow that motivated him to attempt to write the rules of the electricity and magnetism in this field theory form.

How fluids move through space is very complex. At any point in the fluid there are several variables that are necessary to describe the state of the fluid. These variables such as density, velocity, and temperature are all fields, defined at each point in space and subject to change by some set of rules that are determined by the values of these variables at that point and nearby points and by the nature of the fluid. For example, if the temperature at a point is higher than its neighbors, that temperature will tend to decrease because of heat flow from the neighbors. Also depending on the nature of the fluid, the density may increase and this will cause flow away from the point. How much effect each variable has on the magnitude of the the other variables and how fast these variables respond will depend on the fluid. The parameters such as the thermal conductivity and compressibility of the fluid which will control the rates at which these effects can take place are measured phenomenologically for each fluid. It is not hard to understand

that the properties of a fluid in motion are controlled by local effects; flow at a point depends on the temperature and pressure and flow at the point and neighboring points not on what is going on some distance away. The rules for the fluid flow are thus local. The difference with the results of Maxwell is that we know there is an underlying structure, the atoms. In the case of the electromagnetic field, it is not made of anything but itself. The inability to associate a reality to the field independent of an underlying structure is the basis for the famous search for an ether, see Section **??**.

In fact, Maxwell suffered from that same problem. He discovered his equations by trying to fill space with a hypothetical something that exhibited reasonable mechanical properties and attributing the electric and magnetic forces to whirling vortices in the pervasive medium. The idea was that charges produced vortices in this medium and that the whirling of the vortices close to the charge then produced other vortices etc. until space was filled with whirling vortices and the amount of whirling at any place was the electric force. In other words, in order to understand his own equations, he needed an ether, the famous ether that Einstein disposed of later. He also needed to have the vortices properties be determined by the charge or the whirlyness locally. To the modern physicist, the idea of an underlying mechanical system seems out of place and a little weird. In fact, several years back, there was a collection of articles published that were "lighthearted" musings by well known scientists, [Weber 1973]. These articles were written as joke. Among the collected articles was the original paper by Maxwell justifying his vortices in the ether as a mechanism for the electromagnetic field. At the time of the writing, there was nothing lighthearted about it.

4.2 The Stretched String

Since the concept of the field and its dynamical rules are rather hard to grasp in the abstract, let's look at a particularly simple mechanical field system – the transverse displacement of a stretched string. I have to emphasis that this is a field with an obvious underlying mechanical structure – the string, a system with mass and an internal force, the tension. This is in contrast with the fields that we will deal with later. These fields are themselves the fundamental entities. The other thing to realize is that the string that we deal with is an idealized element. It has zero thickness and bends with no resistance. Its only possible displacement is transverse to its alignment.

The displacement of the string in a direction transverse to its direction is a field defined on all the points along the string. This field is much

simpler that the electromagnetic field which is a field composed of two vector quantities, the electric and magnetic forces. The string field also obeys a simple mechanical rule for its dynamics.

Like most mechanically based systems the dynamics of the string has two simple sources, energy of the motion of its masses and a potential energy that is due to its configuration. For the case of a string held tightly with a tension T_e and with only transverse displacements, the potential energy is the work associated with making the string longer. The displacement of the string in the transverse direction is the field that we will consider and any non-zero displacement causes the string to be longer and thus changes the potential energy. These are global approaches to the behavior of the string and will be useful to us later when use a more universal approach to dynamics based on a concept called action, see Section 4.4. For now because our goal will be an understanding of the electromagnetic field, we will use a more local approach and find that the electromagnetic field has many of the same properties as this the simplest of fields. In this approach the electromagnetic field is just a more complex field and the complications do not add any to the understanding of the field nature of the system. For example, the stretched string is a one dimensional field defined on a one dimensional manifold, the distance along the direction of the string. The field variable, $y(x,t)$, is also simple in that it is the transverse displacement of the string from its equilibrium position where x is the position along string. Both y and x range over a one dimensional range of values. The electromagnetic field is a pair of vectors in its field variable and it ranges over a three dimensional manifold, space.

You may also be perplexed by the idea of a stretched string under tension. Our experience is that a string has to be fastened to be under tension. If that is the case, think of the string as tightly stretched between fixed walls. The problem with this is that the walls add complications of their own and for the first pass are not necessary. Here we deal with an infinite string under tension. Later, we will deal with the walls, see Section 6.3.

The local statement of the dynamics of the string are easy to understood; the rule is very simple and intuitive: The force on a segment of the string caused by the transverse displacement of that piece of the string is proportional to the negative of the average of the displacement of that segment of the string from the displacement of its neighbors.

In order to implement this algorithm, divide the string into small segments of length Δl and concentrate the mass in the segment at a point, see Figure 4.1. In the example shown, the segment of string labeled i is above the position of the average of its two neighbors. Thus there is a force to

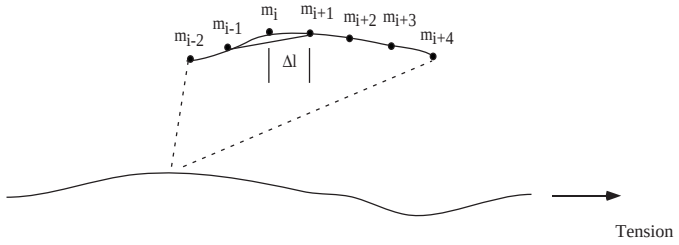

Figure 4.1: **The Stretched String** A string that can move in the transverse direction under tension is a simple example of a local field. In the figure, a section is magnified. In this section, the string is divided into small segments of length Δl and the mass of each segment is concentrated at a point. The dynamic of the string is that the mass at segment at location i has a force on it if its transverse displacement is different from the average of its two neighbors. Thus in the case shown, by drawing a straight line between masses at $i-1$ and $i+1$, we can see that at the place of segment i, the neighbors' average is below i's current position. Thus i has a downward force on it.

bring it to the position of the average. The proportionality constant for this force has the dimensions of a force per unit length and is thus the twice the tension in the string divided by the length of the segment of string; twice since both neighbors pull. ρ is the mass per unit length of the string and thus the mass of each segment is $\rho\Delta l$. Using $\vec{F} = m_i\vec{a}_i$ and using the position along the string x as label for the piece of string, the transverse displacement of the string at x is $y(x,t)$, the average of the two neighbors of x is $\frac{\{y(x+\Delta l,t)+y(x-\Delta l,t)\}}{2}$, the force equation for the segment at x is

$$\rho\Delta l a_{x,t} = -\frac{2T_e}{\Delta l}\left[y(x,t) - \frac{\{y\left(x+\Delta l,t\right)+y\left(x-\Delta l,t\right)\}}{2}\right], \tag{4.1}$$

where T_e is the tension in the string.

Another way to organize the right side of Equation 4.1, is to note that

$$\begin{aligned}\frac{2}{\Delta l^2}\left[y(x,t) - \frac{\{y\left(x+\Delta l,t\right)+y\left(x-\Delta l,t\right)\}}{2}\right] = \\ -\left\{\frac{\Delta y}{\Delta l}\left(x+\frac{\Delta l}{2},t\right) - \frac{\Delta y}{\Delta l}\left(x-\frac{\Delta l}{2},t\right)\right\}.\end{aligned} \tag{4.2}$$

This last term on the right is the negative of the definition of the second derivative of $y(x,t)$. Note also that the acceleration is the second derivative

with respect to time. In the limit that Δl is zero and using partial derivatives because we have both x and t dependence, this force equation becomes

$$\rho \frac{\partial^2 y}{\partial t^2}(x,t) = T_e \frac{\partial^2 y}{\partial x^2}(x,t). \tag{4.3}$$

This is an excellent example of the general form in which the dynamics of fields are expressed. They are generally partial differential equations because we are interested in how the field changes for changes in position and time. Equation 4.3 is second order in the time derivatives because that is how the dynamic operates; it emerged from a mechanical force law. Other orders of time derivatives are possible and it is not uncommon to have laws that are first order in time. In fact, it is preferable because the interpretation of the evolution is simpler. Maxwell's Equations are an example. The stretched sting or any higher order temporal evolution can be reduced to a first order temporal evolution by defining new fields. Defining a new field, $v(x,t) \equiv \frac{\partial y}{\partial t}$, we can get an evolution that has only first time derivatives.

$$\begin{aligned} \frac{\partial y}{\partial t}(x,t) &= v(x,t) \\ \rho \frac{\partial v}{\partial t}(x,t) &= T_e \frac{\partial^2 y}{\partial x^2}(x,t). \end{aligned} \tag{4.4}$$

In a very real sense, you could say the the magnetic part of the electromagnetic system is a manifestation of this kind of substitution. More on this later, see Section 7.3.

The fact that there are only values of the field and spatial derivatives of the field on the right side of the Equation 4.3 is the expression of the locality of the dynamic. How the field evolves at a place depends only on what is going on at that point. Also note that the only parameters in the field equation are ρ and T_e. These express the intrinsic properties of the medium in which the field operates. By dividing Equation 4.3 by ρ, we can reduce the effective number of parameters to one, $\frac{T_e}{\rho} \stackrel{\text{dim}}{=} \frac{L^2}{T^2}$. This has the dimensions of a velocity squared. The fact that there is only this parameter in the dynamic says a great deal about the nature of the evolution of the fields. There are not enough parameters to construct a length or a time. Thus for this field there is no intrinsic size except as it is put in by the starting conditions or put into the problem by boundaries like walls. Thus this particular field system, the stretched rope, is characterized by movement of field configurations. Since the parameter of the medium is a velocity squared, the movement is in both directions with a characteristic

speed, $\pm\sqrt{\frac{T_e}{\rho}}$. It is important to remember that the movement of a piece of string is only in the transverse direction whereas the movement of the field configurations is along the direction in which the string is aligned. This is a difficult situation to describe. If you attribute all reality to the hunk of string the only motion is up and down in the transverse direction. Yet the configuration of the string moves along the string. We will find that there is energy and momentum associated with the configuration of the string and that this thus moves with the configuration along the string. Thus we have the problem of the 'string' only moving up and down but energy and momentum flowing along the string.

The converse of the above result that the parameters of the system are not sufficient to determine a size or time scale is that the medium, in the case of the stretched string are ρ and T_e, implies that the disturbances in the string travel with a speed set by the medium, $\sqrt{\frac{T_e}{\rho}}$ and that this speed is independent of the form of the disturbance. In other words disturbances travel with speed $\pm\sqrt{\frac{T_e}{\rho}}$ without distortion. For this reason, systems with this field dynamic are called wavelike. This is the definition of a wavelike medium. Although many systems are wavelike such as sound and light, other field systems may not be. For instance the dynamic for temperature flow in one spatial dimension is

$$\frac{\partial T_{emp}}{\partial t}(x,t) = a^2 \frac{\partial^2 T_{emp}}{\partial x^2}(x,t). \tag{4.5}$$

where a^2 is called the diffusion constant and is the ratio of the heat conductivity to the heat capacity of the material. Notice that $a^2 \stackrel{\text{dim}}{=} \frac{L^2}{T}$ and thus there is no special speed or length or time that is characteristic of the field.

In order to better understand the operation of field dynamics let's work though the example of the string under tension. Consider our case of a stretched string with mass per unit length ρ and tension T_e. At $t = 0$, we put a distortion in the string as shown in Figure 4.2. Note that at $t = 0$, the string is displaced but no part of the string is moving. It is simplest to interpret the operation of the dynamic in the first order time derivative form, Equation 4.4. In this form, it is clear that a complete description of the initial configuration of the string involves the specification of two fields, the initial velocity field and the initial displacement field. In other words for the case in Figure 4.2 at $t = 0$, the velocity of all parts of the string is zero and there is a simple pulse of displacement in the string. Other starting configurations are possible. You could have the situation in which the string has no displacement and the sting has a distribution of transverse velocity.

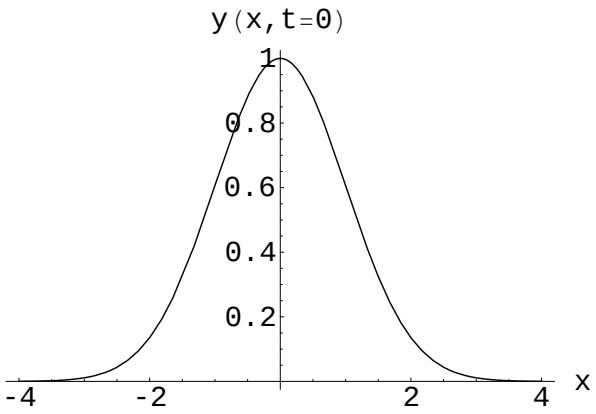

Figure 4.2: **A Simple Displacement Pulse in a String** A simple pulse in a stretched string under tension. At $t = 0$, the string is distorted but no part of the string moving.

The difference in the operation of a harpsichord and piano is the the strings are plucked or distorted in the harpsichord and hammered in a piano. You can also have situations with both an initial displacement and velocity.

The dynamic of the string requires that all points on the string be at the average of its neighbors. An easy way to compute the average is to pick two neighbors, points on the string close to the point of interest and equidistant from it, and connect the points by a straight line. At the point of interest, x, the point on the line is the average of the two neighbors. Thus from Figure 4.3, we see that the center of the string is pulled strongly down and the edges are pulled up. The points of steepest drop are not pulled at all. This last point is interesting to note. The string is not pulled to the neutral position. Each segment is pulled only by its neighbors. If the string where pulled to the neutral position there would be a force for the entire time of descent and then the string would still have a velocity when it reached the neutral position and thus would overshoot and there would be oscillation at each disturbed point on the string. As we know, the disturbance in the stretched string is removed by the dynamic with the string returning gently to its neutral position.

To make this discussion more quantitative, we look at what goes on in a few small time increments. In a small time, Δt, since the velocity field is initially zero everywhere, we find that the string has not moved.

$$y(x, \Delta t) \quad = \quad v(x, 0)\Delta t + y(x, 0)$$

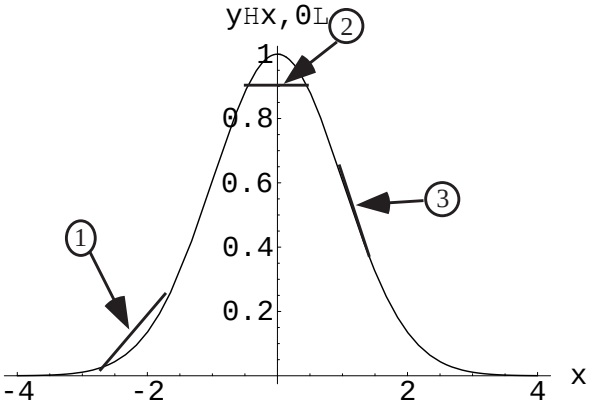

Figure 4.3: **Forces on a Pulsed String** The dynamic of the stretched string require that all points in the string be at the average of its neighbors. A simple rule for finding the force and thus the acceleration of a place on the string is to connect the neighbors with a straight line. If the string at that place is above the line, there will be a downward acceleration with magnitude proportional to the distance above. There are three examples shown. At a point on the edge of the pulse, 1, the string is accelerating upward. At the center, 2, the string is accelerating down. At a point at the midpoint of the side of the pulse, 3, the string has no acceleration.

$$= \quad y(x,0), \tag{4.6}$$

where $v(x,t)$ is the velocity of the string at the point labeled x at time t. At $t=0$, the string is not moving and $y(x,0)$ is known.

We will need the velocity of the string at all times and, even in a small time, because of the forces from Figure 4.3, the velocity changes.

$$\begin{aligned} v(x,\Delta t) &= a_{t=0}(x)\Delta t + v(x,0) \\ &= a_{t=0}(x)\Delta t \end{aligned} \tag{4.7}$$

where we find $a_{t=0}(x)$ from an analysis such as that shown in Figures 4.3 for each point on the string. Thus we see that after a time Δt the velocities will have the same pattern as a function of position as the initial accelerations.

Repeating the process for a second Δt using Equations 4.6 and 4.7 but with the time shifted another increment,

$$\begin{aligned} v(x,2\Delta t) &= a_{t=\Delta t}(x)\Delta t + v(x,\Delta t) \\ &= a_{t=0}(x)\Delta t + a_{t=0}(x)\Delta t \\ &= 2a_{t=0}(x)\Delta t \end{aligned} \tag{4.8}$$

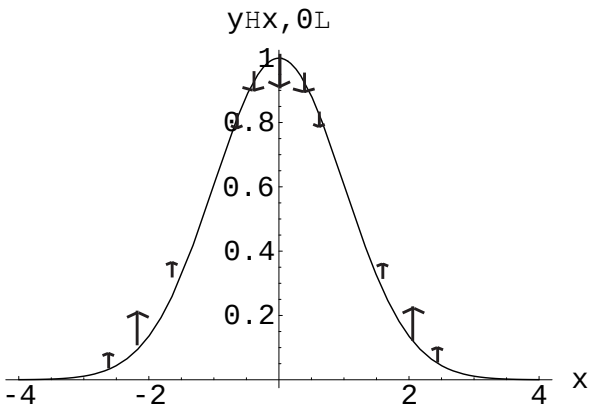

Figure 4.4: **Accelerations on a Pulsed String** Using a technique such as shown in Figure 4.4 for the forces on the string, the algorithm in Equation 4.1 can be applied at each point, x, and find the accelerations shown as arrows above.

where in the second line, I used the fact that since $y(x, \Delta t) = y(x, 0)$ and the accelerations depend only on $y(x, t)$, then $a_{t=\Delta t}(x) = a_{t=0}(x)$.

The second dynamic is handled similarly,

$$\begin{aligned} y(x, 2\Delta t) &= v(x, \Delta t)\Delta t + y(x, \Delta t) \\ &= a_{t=0}(x)\Delta t^2 + y(x, 0). \end{aligned} \tag{4.9}$$

We now begin to see the string moving.

We can intuit that the pattern shown in Figure 4.5 develops. The region where there is a strong bend at the edge is is pulled up and so has an upward velocity and begins to lift. The middle section is unchanged at first. The center is forced down and has a downward velocity. Because of the pattern of the upward velocity at the bends and the downward velocity at the center, the two separating pulses appear to be moving along the string away from each other. We have to remember that the all the motion of the string is transverse to its direction.

The general pattern then develops of two distinct pulses of half the original amplitude one moving to the left and one to the right, see figure 4.6. This transverse velocity is patterned so that the two emergent pulses are one moving to the left with speed $-\sqrt{\frac{T_e}{\rho}}$ and one moving to the right with speed $\sqrt{\frac{T_e}{\rho}}$. Each of these are called traveling waves, one to the left and one to the right. It is the pattern of traveling waves that there is both a transverse

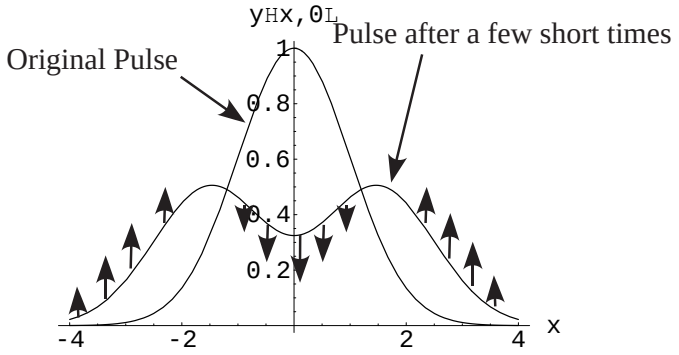

Figure 4.5: **Pulsed String after a Few Short Times** Using appropriate versions of Equations 4.6 and 4.7 to evolve the system, we can see the development of two pulses. Also shown are the velocities by scaled arrows. Remember the parts of the string are only free to move up and down but the pattern of up and down motion conspires to produce the effect that the pulse at negative x is moving toward greater negative x and the pulse at positive x is moving toward greater positive x. The original pulse is shown for comparison.

displacement field and an associated transverse velocity field with the velocity field rising in front of the motion of the traveler and falling behind the traveler. This is a typical pattern for wavelike media. There are two fields that support each other and form the traveling configuration. For sound it is the density of the air and the pressure of the air. For electromagnetic waves, it is the electric and magnetic force fields.

It is worthwhile to also note that our original configuration of the displacement pulse with no velocity, Figure 4.2 can be considered as the sum of two travelers, one going to the left and one going to the right, each of half amplitde. The addition of the displacement field gives the correct shape for the pulse and, at the instant of complete overlap, the initial instant, the two transverse velocity fields add to zero. The ability to treat the original distortion as a sum of two independent distortions is an example of superposition. This will be an important principle in many future discussions, see Section 18.6.2.

In addition. the travelers have an interesting relationship between the displacement field and the velocity field. For a traveler that moves to increasing x, the argument of the displacement field is a single variable, $x - \sqrt{\frac{T_e}{\rho}}t$, instead of x and t as independent variables. This traveler is called a right

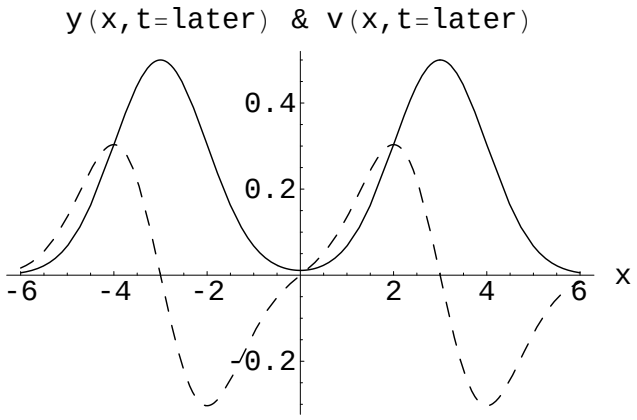

Figure 4.6: **Pulses in String Separating** After a time, the pulse initially placed on a stretched string, see Figure 4.2, separates into two half amplitude pulses. One travels to the left with velocity $v = -\sqrt{\frac{T_e}{\rho}}$ and one travels to the right with velocity $v = \sqrt{\frac{T_e}{\rho}}$. There is also a transverse velocity field that travels along with each pulse shown as the dashed curve instead of using arrows as in Figure 4.5.

traveler. For waves that move to decreasing x, called left travelers, the argument is $x + \sqrt{\frac{T_e}{\rho}}t$. This is what makes them travelers; they move to increasing x or decreasing x uniformly without the shape of the disturbance changing. This is a general result and true for all one dimensional wavelike systems. We worked this out for the particular disturbance of Figure 4.2, a simple pulse. It should be clear that this pattern of two separate travelers superposing to produce an initial distortion with no velocity field will hold for any form of distortion for the displacement field. Figure 4.7, shows a more general initial configuration and the subsequent travelers. Because to the nature of the relationship between the x and t variables in the travelers, $x - \sqrt{\frac{T_e}{\rho}}t$ for the right traveler and $x + \sqrt{\frac{T_e}{\rho}}t$ for the left traveler, the time evolution of the displacement field which is the velocity field in this dynamic is related to the slope of the displacement of the traveler at that point.

$$\frac{\partial y_{rt}}{\partial t}(x,t) \equiv v_{rt}(x,t) = -\sqrt{\frac{T_e}{\rho}}\frac{\partial y_{rt}}{\partial x}(x,t) \tag{4.10}$$

where $y_{rt}(x,t)$ and $v_{rt}(x,t)$ are the right traveling waves displacement field

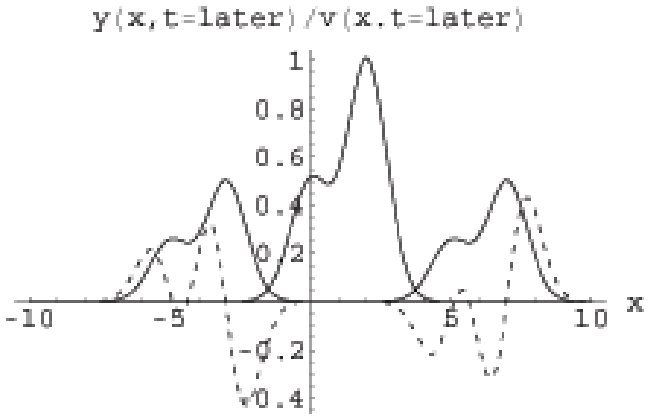

Figure 4.7: **Arbitrary Traveling Waves** Using a more general form for the initial distortion of the string, shown at the center for reference, we see at a later time the two traveling distortions, one moving to increasing x called the right traveler and one moving to decreasing x called the left traveler. The associated velocity profile for each is shown dotted. Because of the special form of the argument of the travelers, the velocity profile for the right traveler is proportional to the negative of the slope of the displacement profile of the right traveler at that instant and the velocity profile of the left traveler is proportional to the slope of the displacement profile of the left traveler at that instant.

and velocity field. The relationship of the velocity field and the displacement field for the left traveler is similarly:

$$\frac{\partial y_{lt}}{\partial t}(x,t) \equiv v_{lt}(x,t) = -\sqrt{\frac{T_e}{\rho}}\frac{\partial y_{lt}}{\partial x}(x,t). \tag{4.11}$$

Another feature of the travelers is that they carry energy and momentum. It takes a certain amount of work to distort the string; it has to becomes longer. This distortion energy is then distributed into the travelers and these then carry it off to remote regions of the string. Similarly there is momentum associated with the travelers that is transported down the string by the travelers. In a later section, we will develop a more nuanced identification for momentum and energy, see Section 4.4 but for now our intuitive ideas will suffice. Notice that this is energy and momentum that moves down the string even though the string itself can only move in a transverse direction. Thus the traveler wave configurations act like a thing that moves along the

string even though nothing moves down the string. Note that a superposition of the travelers constitute the original disturbance. Here we begin to see the development of a thing, something that carries energy and momentum, in the context of a field. The electromagnetic field is a wave field and will have travelers also. These are more complex and constituted differently in the dynamic than these string travelers but they behave similarly. Since they generally operate in three spatial dimensions there is a geometric fall off in strength as they travel but they still carry energy and momentum to remote parts of the system.

In Section 6.3, The Stretched String Revisited, we will return to the dynamics of the string. For now we are content to use it as a simple example of a field system and to have it express the basic ideas of a field theory, a construction that is a local causal dynamical system. In the next section, we will discuss Maxwell's Equations, the first fundamental theory based on a field construction.

4.3 Maxwell's Theory of Electromagnetism

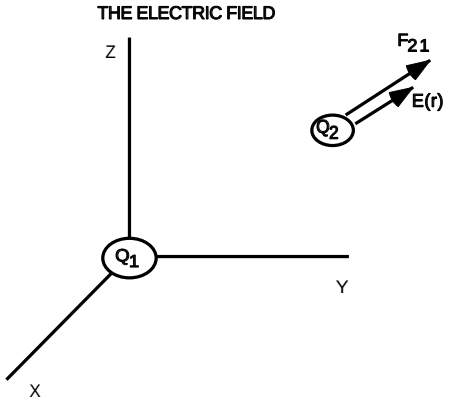

Figure 4.8: **The Electric Field and Electric Forces** Maxwell said that electric and magnetic forces were due to the presence of the electric and magnetic field. In this figure, the electric force on Q_2 is due to the presence of the field at its location, $\vec{F}_{21} = Q_2\vec{E}(\vec{r})$. There is a similar relationship for the magnetic force.

Maxwell was interested in developing a unified description of electric and magnetic phenomena. In his time, many of the basic ideas of the electric and magnetic force systems were known. The law for the electric interaction

between charged particles had been articulated in the period 1785 and 1791 by Coulomb. The force law between magnets and the force between moving charges and magnets was known and even Faraday's Law about the relationship of changing magnetic environments and electric currents was known. In fact, Faraday had already began to describe magnetic and electric phenomena in a field like language. What Maxwell sought was an underlying mechanical basis for all the phenomena associated with electricity and magnetism. Reducing electricity and magnetism to a mechanical basis meant that he was looking for something to push or pull but it had to do so locally. He could not believe that fundamental phenomena could take place as an action at distance phenomena like gravity was thought at the time. In order to have a thing which could push or pull locally, he hypothesized the existence of a rather rich structure for the vacuum of space, whirling vortices in an ether that produced the electric and magnetic force. Thus not only did he seek a mechanical source for electric and magnetic phenomena, he developed a field theory basis for it. His picture of electric and magnetic forces was that they were mediated by fields, the electric, $\vec{E}$, and the magnetic, $\vec{B}$, fields. It was his basic idea that the correct description of electromagnetic phenomena required a locally causal dynamic The idea was that not only did the charges generated the fields but the fields themselves responded to the local environment of the fields themselves. In addition, the forces experienced by the charges were because of the values of the fields at the place occupied by the charges, $\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$, where q is the charge in question and $\vec{v}$ is its velocity.

In order to create the mechanical basis for the fields, Maxwell was forced to endowed the ether with the correct mechanical properties of inertia and size to replicate the success of the earlier laws but now in context of a local mechanical model. The underlying idea was simple. Let's look at the simplest of the cases, Coulomb's Law. The situation is shown in Figures 4.9 and 4.10. A force on a charged particle took place as a two step process. A charge Q_1 is placed in empty unexcited space. This charge excites the ether next to it by creating vortices at its location. These vortices in turn excite neighboring vortices until space is full of whirling vortices. Each vortex is in dynamic equilibrium with its neighbors. There is a 'thing', the whirliness, which is a measure of the electric field at that point.

When a new charge, Q_2, is located at some distance, $\vec{r}$, from the first charge, it detects the level of excitement of the local vortices and thus feels a corresponding force. The force is proportional to the charge Q_2 at that place and the amount of whirliness or electric field at that point.

The mechanical properties of the ether and its vortices determine how

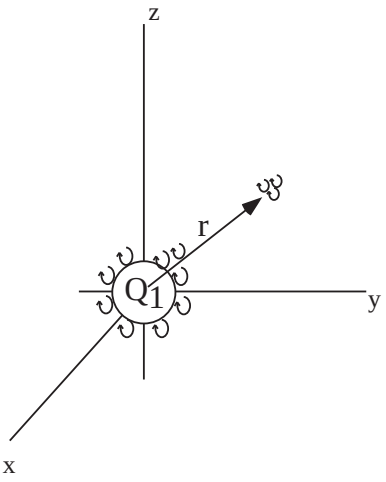

Figure 4.9: **Maxwell's Vortices** Maxwell pictured the electric force as emerging in two steps. First any charged particle would excite vortices in the ether at its location. These vortices would excite other vortices nearby and so forth until all of space would fill with whirling vortices. In a sense, the whirliness of the vortices at any place was a measure of the strength of the electric field at that point.

the whirliness develops. This is set by the vortices inertia and size. These parameters for the mechanical properties of the vortices are then adjusted to accommodate Coulomb's Law.

In other words, Maxwell introduced local fields – a continuous quantity defined at all points in space and for all times – with a rule of dynamics to produce the electromagnetic forces. If an object experiences a force, there must be something at that place, the whirliness. In addition, the whirliness itself must be determined locally in both space and time. Let's go through the example of Coulomb's Law in a little more detail to see how this idea works.

The first problem is to reproduce the well known Coulomb's law of force for static situations. Coulomb's Law is an action at a distance description of interaction,

$$\vec{F}_{21} = \frac{1}{4\pi\epsilon_0}\frac{Q_1 Q_2}{r_{12}^2}\frac{\vec{r}_{21}}{r_{12}} \tag{4.12}$$

where $\vec{r}_{12}$ is the separation between the charges. In order to simplify the discussion, let's place charge Q_1 at the origin. Since the force on the charge Q_2 is supposed to be $Q_2\vec{E}(\vec{r})$, where $\vec{r}$ is now the position at which Q_2 is

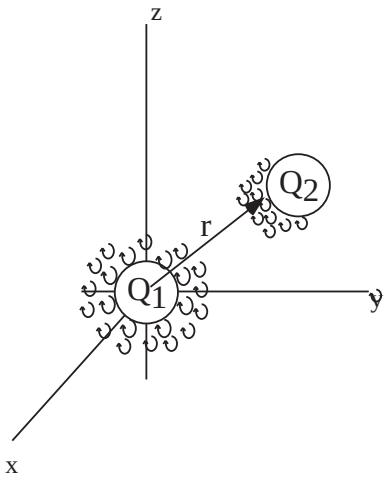

Figure 4.10: **Vortices and the Electric Force** When a charged particle, Q_2, is positioned, the particle detects the local amount of whirliness in the vortices of the ether. This generates the electric force in proportion to the charge and amount of whirliness at its location. The local whirliness is $\vec{E}$ at $\vec{r}$.

located. For this case, we can identify the electric field as

$$\vec{E}(\vec{r}) = \frac{1}{4\pi\epsilon_0}\frac{Q_1}{r^2}\frac{\vec{r}}{r} \tag{4.13}$$

around a spherically symmetric charge placed at the origin. You will reproduce the static Coulomb's Law results with the electric field if you can make a local rule about how $\vec{E}(\vec{r})$ develops that reproduces this result. It should be clear that the hard part will be to reproduce the inverse square fall off with distance in the strength of the field.

In some sense, it is really not correct to say that Q_1 is the source of this field. The field is not attached to the charge. At any point, there is a field only if there is a field or a charge in the neighborhood. The field at some point, like all things, is to be determined locally. Maxwell used his whirling vortices of the ether to discover a rule for whirliness and how whirliness effected whirliness that recovers the characteristic the inverse square fall off with distance of Coulomb's Law. Like the stretched sting, Section 4.2, in which the transverse position of a place on the string is determined by the transverse position of the neighbors to that place, similarly here, the idea is to find the rule on how the field arranges itself and forget about the whirlies.

The following analysis reviews the process and becomes somewhat technical but the struggle to follow it is worth the effort.

Since the electric field is meant to produce a force, it must be a vector field, a directed quantity defined at every point in space and with a local rule for its construction. Basically you ask how much does the field change at a place because of what is there. For now, we are looking at a static case – no time change. But we can still ask about how the field varies as we change positions in space.

For a vector field such as the electric field since it is a vector field, you have a directed strength at each point in space and around each point you have directed strengths. At any point you can ask how much more "out pointy" these directed strengths become as you go from place to place. The analogy for our stretched string is that, at any place on the string, you can ask how "bendy" is the string. On a string, "bendiness" happens when that place differs from its neighbors. The string bends up when the place is lower than its neighbors and it bends down when it is higher. When there is no bend, that place on the string is at the average of its neighbors. In the static string it takes a force to maintain a bend in the string. Our case for the vector field case using "out pointiness" works in the same fashion. You can have "out pointiness" only if there are charges that are placed there, i. e. charge causes an outward directed field. Of course, we have to develop a definition, a measure, of "out pointy" and test it.

The measure of "out pointiness" is called the divergence and it is what you would have thought to define it as if you spent some time playing with the ideas of a vector field. At any point, find out how much the neighboring fields point away from where you are. That should indicate the "out pointiness".

Since fluid flow is also a vector field it is worthwhile to think in terms of it. The vector field in this case is the velocity of the fluid. If at a point all the flow is uniform about you, you would not think of the field as becoming "out pointy". On the other hand, if you were at a place like the drain, you would consider the surrounding flow to be "in pointy", the opposite of "out pointy". To be more quantitative, think of surrounding the place that you are interested in and measuring how much stuff flows in or out. By enclosing the point of interest with a surface, we can measure the incoming fluid by assessing how much stuff comes into any element of area on the surrounding surface and then adding the contribution to each part. In other words, surround the point with a surface. Cover it with elements of area, postage stamps. Each element of area has a normal vector that points either outward or inward, see Figure 4.11. Choosing the outward normal, we are

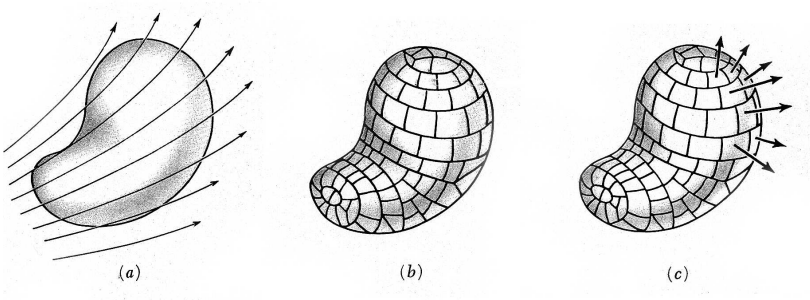

Figure 4.11: **Construction of the Divergence** To find the divergence or "out pointiness" of a vector field at a point, surround the point with a surface, step (a). Cover the surface with small elements of area so that to all intents and purposes they can be considered flat. Each element of surface will now have a normal vector. Find the magnitude of the vector field at the surface to is along the surface. Add these magnitudes for each element of surface and the total is the divergence or "out pointiness" of the vector field at the point surrounded by the surface. Then shrink the volume surrounded to a point. For a fluid, applied to the velocity field, this tells the amount of fluid that goes into a point. This series of steps is encoded in the first part of Equation 4.14 for the case of the electric field.

defining "out pointiness", the amount of the velocity along the normal, is the flow through that area. Now do this for the each element of the entire surface and add up all the contribution from all the pieces. To reduce this analysis to a point, shrink the volume enclosed by the surrounding surface to zero. This same analysis holds for all vector fields. This construction at each point assesses the "out pointiness" of the neighborhood of the point and is called the divergence. Thus,

$$\mathrm{Div}(\vec{\mathbf{E}}(\vec{\mathbf{r}})) = \lim_{\mathrm{V}\to 0} \frac{\sum_{S\supset V} \vec{\mathbf{E}}(\vec{\mathbf{r}}\,') \cdot \Delta^2\vec{\mathbf{S}}}{\mathrm{V}} = \lim_{\mathrm{V}\to 0} \frac{1}{\epsilon_0}\frac{\mathrm{Q}_{\text{inside V}}}{\mathrm{V}} = \frac{1}{\epsilon_0}\rho(\vec{\mathbf{r}}) \quad (4.14)$$

where the first part is a mathematical statement of what is stated above but for the case of the electric field and the subsequent parts are the relationship with charge that is necessary to recover Coulomb's Law, i. e. electric charge is the source of divergence.

Notice that this law, Equation 4.14, says that for a static electric field there is divergence of the field only where there is charge. Yet the picture that we all have of the static electric field around an isolated point charge is a

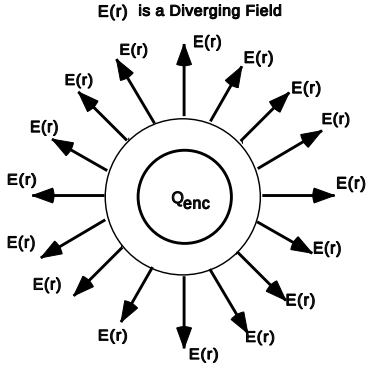

Figure 4.12: **"Outpointiness of the Electric Field"** A characteristic property of the electric field is that charge is the source of "outpointiness". This is the idea that the electric field points away from nearby positive charges and toward nearby negative charges. This last example being negative outpointyness.

diverging field, the electric field points outward from the origin everywhere, see Figure 4.12. How do we reconcile this?

Consider a point away from the isolated point charge. If a surface such as that shown in Figure 4.11 is constructed area at nearer the charge is smaller whereas the area more distant is larger. In fact, the areas are in the ratio of the distances squared. Thus the field strength and the areas combine so that the net "out pointiness", actually in pointiness, of the nearer surface balances the out pointiness of the far surface and the net is zero. Thus it is because the divergence is zero at places other than the charge that the field strength falls off with distance as $\frac{1}{r^2}6$.

Another property that a vector field can manifest is rotation or curl. Again you develop a definition and test it. Here the idea is to follow a closed path around the point and see how much of the vector field follow the path. The electric field does not curl.

$$\text{Curl}(\vec{\mathbf{E}}(\vec{\mathbf{r}})) = \lim_{\text{S}\to 0} \frac{\sum_{p\supset S} \vec{\mathbf{E}}(\vec{\mathbf{r}}\,') \cdot \Delta\vec{\mathbf{r}}\,'}{\text{S}} = 0 \tag{4.15}$$

On the other hand, the magnetic field does curl. The magnetic field is the force experienced by a moving charged particle.

$$\vec{\mathbf{F}}_{\mathbf{mag}} = \text{Q}\vec{\mathbf{v}} \times \vec{\mathbf{B}}(\vec{\mathbf{r}}) \tag{4.16}$$

The magnetic field lines tend to wrap around their sources, the currents.

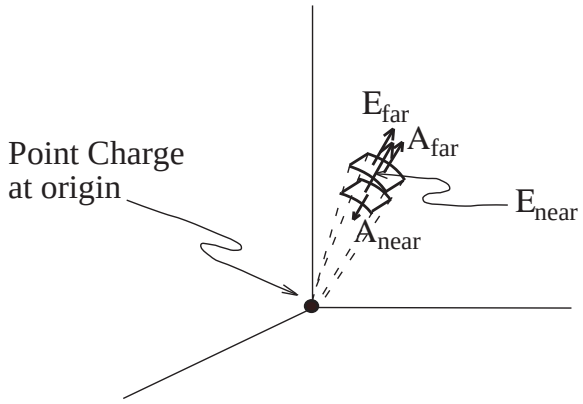

Figure 4.13: **Divergence outside Charge** Barbie A characteristic property of the electric field is that charge is the source of "outpointiness". This is the idea that the electric field points away from nearby positive charges and toward nearby negative charges. This last example being negative outpointyness.

$$\text{Curl}(\vec{\mathbf{B}}(\vec{\mathbf{r}})) = \lim_{\text{S}\to 0} \frac{\sum_{p\supset S} \vec{\mathbf{B}}(\vec{\mathbf{r}}\,') \cdot \Delta\vec{\mathbf{r}}\,'}{\text{S}} = \lim_{\text{S}\to 0} \frac{1}{\mu_0} \frac{\mathbf{i}_{enc\ p}}{\text{S}} = \frac{1}{\mu_0}\vec{\mathbf{j}} \tag{4.17}$$

and does not diverge

$$\text{Div}(\vec{\mathbf{B}}(\vec{\mathbf{r}})) = 0 \tag{4.18}$$

Note that we have not added a time dependence. These are all static situations.

Maxwell insisted that the field was not established everywhere at once. It was made up of whirling vortices that pushed on each other. The rate at which the vortices could push was set by the parameters of the static theory. By endowing these whirling vortices with the correct properties to reproduce the laws of static electricity and magnetism, he found how to add a local set of rules for the time evolution of the fields. These are the full set of Maxwell's equations including time dependence:

$$\text{Div}(\vec{\mathbf{E}}(\vec{\mathbf{r}},t)) = \frac{1}{\epsilon_0}\rho(\vec{\mathbf{r}},t) \tag{4.19}$$

$$\text{Curl}(\vec{\mathbf{E}}(\vec{\mathbf{r}},t)) = \frac{\partial\vec{\mathbf{B}}}{\partial t}(\vec{\mathbf{r}},t) \tag{4.20}$$

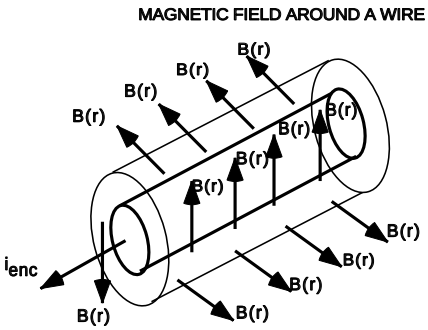

Figure 4.14: **The Curl of the Magnetic Field** In contrast to the electric field, the magnetic field wraps around or curls around its sources, the currents in the problem.

$$\text{Div}(\vec{\mathbf{B}}(\vec{\mathbf{r}},t)) = 0 \tag{4.21}$$

$$\text{Curl}(\vec{\mathbf{B}}(\vec{\mathbf{r}},t)) = \mu_0\vec{\mathbf{j}}(\vec{\mathbf{r}},t) - \mu_0\epsilon_0\frac{\partial\vec{\mathbf{E}}}{\partial t}(\vec{\mathbf{r}},t) \tag{4.22}$$

This is the standard format for these equations. For a discussion of the field dynamics, it is important to realize that only two of these equations are a dynamic, Equations 4.20, and 4.22. The other two equations, Equations 4.19 and 4.21, are what are called constraint equations; they control the pattern of the field but not the temporal evolution. It is apparent that the electromagnetic field is a much more complex field that the stretched string whose dynamic is Equation 4.4. The vector nature of the field, the existence of constraints, and the sources, $\rho(\vec{r},t)$ and $\vec{\mathbf{j}}(\vec{\mathbf{r}},t)$, obviously complicate the situation. We could have added external forces to the dynamic of the string but that would not have clarified the field nature of the string. Similarly, here we can discuss the electromagnetic field without the presence of $\rho(\vec{r},t)$ and $\vec{\mathbf{j}}(\vec{\mathbf{r}},t)$. Rearranging and omitting the sources, the dynamical equations for the evolution of the electromagnetic field become

$$\frac{\partial\vec{\mathbf{E}}}{\partial t}(\vec{\mathbf{r}},t) = -\frac{1}{\mu_0\epsilon_0}\text{Curl}(\vec{\mathbf{B}}(\vec{\mathbf{r}},t)) \tag{4.23}$$

$$\frac{\partial\vec{\mathbf{B}}}{\partial t}(\vec{\mathbf{r}},t) = \text{Curl}(\vec{\mathbf{E}}(\vec{\mathbf{r}},t)) \tag{4.24}$$

Identifying $\vec{\mathbf{E}}(\vec{\mathbf{r}},t)$ with the displacement field of the string, $y(x,t)$, and $\vec{\mathbf{B}}(\vec{\mathbf{r}},t)$ with the velocity field of the string, $v(x,t)$, we see that the electromagnetic dynamic is more complex but similar in structure.

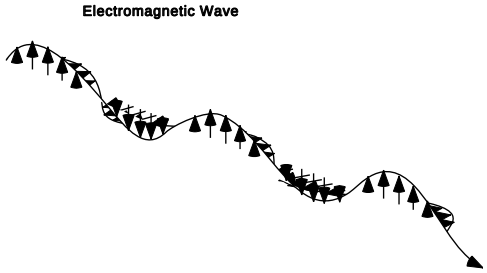

Figure 4.15: **The Field Configuration for Light** Light is a traveling wave solution of Maxwell's Equations and is composed of propagating combination of electric and magnetic fields. The direction of flow of energy and momentum is along the normal to the plane of the oscillating electric and magnetic field vectors. In the figure the upward arrows represent the electric field and the perpendicular arrows are the magnetic field.

An important feature of the electromagnetic field that can be seen from the equations above is that, if you have an electric field in a localized region of space, finite somewhere but zero elsewhere like the pulse in the stretched string, the electric field will have a curl. Thus even if there are no charges or currents, this curl is the source of a developing magnetic field, Equation 4.24. This is like the case in the string of the displacement producing a velocity field. As the new magnetic field grows which will also be localized and thus curled, it produces a reduction in the original electric field, Equation 4.23. Thus the original field will start to reduce and there will be a growing magnetic field. This magnetic field will in turn change and produce a electric field. The relationship of the magnetic and electric fields is much like that of the velocity and displacement of the stretched string which produces traveling pulses, Section 4.2. In fact, using Equations 4.23 and 4.24, in a region without charges or currents, the vacuum, you find that the electric and magnetic fields are a wavelike system and that a field configuration such as that shown in Figure 4.15 produces a traveling wave that travels in the plane perpendicular to the plane of $\vec{\mathbf{E}}(\vec{\mathbf{r}},t)$ and $\vec{\mathbf{B}}(\vec{\mathbf{r}},t)$ with a speed

$$v = \frac{1}{\sqrt{\mu_0 \epsilon_0}}$$

which dimensionally is a speed and the only dimensional factor in the dynamic. This is the same result that Maxwell discovered with his whirlies. Putting the values of μ_0 and ϵ_0 this is the speed of light. If it walks like a duck and quacks like a duck, it is a duck and thus Maxwell concluded that

light is the traveling wave solutions to the equations of electromagnetism.

It is important to realize that like in the stretched string which has only a transverse displacement and transverse velocity, the $\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)$ and $\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)$ fields are not traveling but only the disturbance – changes in the field configuration. It is also important to realize that the velocity of the disturbance does not depend on the field configuration. It only depends on the dynamic of the field. Another way that this is often said is that the velocity of propagation is a function only of the medium. Since the electromagnetic field operates in the vacuum of space, it is the properties of the vacuum that determine the speed with which light propagates. A difference for the electromagnetic travels with the travelers of the field of the stretched string is that in the string any distortion will produce simply related travelers but for the electromagnetic field there are configurations of the field that do not have simply related travelers.

We now understand the amplitude that was invented by Young and Fresnel, see Section 3.5.8. It is the electric field. The Fresnel construction is the general rule for the computation of the propagation of the light and holds for traveling waves of the electric and magnetic fields.

4.4 Dynamics and Action

Dynamics, as mentioned earlier, are the rules for finding the temporal evolution of a system. In Newtonian Physics, this set of rules was succintly summed up in the rule: $\vec{f} = m\vec{a}$, see Section 1.2.3. For a while, we will forget about light and fields and the dynamics of these complex systems and just describe simple point particles that move around freely in a simple space. We will find a new way to formulate the rules of dynamics that are more general but still produce the old $\vec{f} = m\vec{a}$ when it is appropriate. The advantage will be that the new rules will work in circumstances in which Newton's Laws were inappropriate or just did not make sense. With these new rules, we will also find a more powerful understanding of the concepts of symmetry and include systems such as fields all in a single dynamical principle. We will also be able to use this new procedure to form a more solid understanding of the ideas of energy and momentum. One complication will be that in order to formulate the rule, we will need ideas about kinetic and potential energy that we formulated earlier. Before we are done, these same ideas will take on a very different and more useful form. We will be able to understand why the massless photon has momentum but first we need to build the necessary background.

4.4.1 Background on Formulation of Action

It is usually not emphasized that the original formulation of Newton's Laws applied to only a very restricted set of circumstances. In Section 1.2.3, Newton's Laws were described as dealing with the effects of one system on another with the assumption that all the parts of the bodies were basically point objects that could move freely in space. This was fine when talking about the planets but, even for some of the simplest cases, these conditions do not hold.

Consider the problem of the motion of a blackboard eraser tossed into the air in the front of the lecture hall with a twisting spinning motion. Each part of the eraser is subjected to a huge array of forces. For convenience you can think of the parts of the eraser as the atoms but, even without an atomic hypothesis, all the following considerations still hold. Each part of the eraser is subject to the force of gravity and each part is subject to internal forces from the other parts of the eraser. First, there is an absurd number of parts and forces between the parts and between the parts and the world outside the eraser. We simplify this situation somewhat by assuming that the effect of gravity is the same throughout the eraser and thus reduce these many gravitational forces to a single force acting at one point at the mass weighted center of the body. This is a good approximation for the case of a small eraser in the near vicinity of the earth.

More subtly, we know that, as the eraser twists and spins, the different parts of the eraser will effect other parts. In fact, if the eraser was not a reasonably rigid body and held together by cohesive forces, in the spinning twisting motion, the parts would fly apart. Because the eraser is rigid, there are internal forces that act to hold the respective parts in a fixed relationship to each other. These forces are very complicated. They are in a very real sense unknowable; they are what they have to be to maintain the rigid configuration. These are called constraint forces. The eraser is not an exception. A car on the highway has a constraint force from the road called the normal force that is whatever it has to be to stop the car from falling into the road. Actually, with a little thought it becomes clear that almost all systems have constraints. The direct application of Newton's laws to systems that are constrained is wrong or impossible. There are an abundance of forces – too many to handle. Worse yet is the realization that many of them are, in fact, unknowable. The forces hold the eraser as it moves through space are whatever they have to be to maintain the positional relationship between the parts of the eraser. These are generally not known and thus cannot be inserted into a simple Newtonian framework.

In many special cases, fixes were developed that allowed the use of Newton's laws for motion in the presence of constraints and it was well known that this was a problem to both Newton and his immediate followers. The general problem of the motion of systems with algebraically described constraints was solved by Joseph-Louis Lagrange. The procedure that he developed is the modern method for articulating the dynamics of any system and is the one that we will use.

4.4.2 Introduction to Action

The modern approach to dynamics is based on the use of an extremum principle like Fermat's least time theory of light. There is a physical quantity that is called the action. In some sense, this is an unfortunate name for this because we have used the word in another context, see Section 4.1.1, and it has a connotation in the conventional usage. The action is a quantity that we will define in detail later but for now understand that is a quantity evaluated over a trajectory in space and time. Up until now, we have dealt with paths in space. Now, we deal with trajectories but the principles are the same. For instance, the Fermat principle of least time required the time of passage of the light over the entire path between two points in space. Here the action is evaluated for a trajectory on space-time between two events, an initial position and time and a final position and time. Generally, the object moves over the trajectory that has the least action. Obviously, I will need to back up a little to make this clear and to establish the terminology.

We describe the motion of anything as a connected set of events in space-time, a path in space-time called the trajectory of the particle. The events labeled by a place and a time and are the fundamental entities and a trajectory is a catalogue of the places and as time evolves where the object went. Of the infinity of trajectories that can connect two events, the naturally occurring trajectory will turn out to be the one that has the least action.

Consider a piece of chalk tossed up from my hand and returning to my hand some short time later. I am dealing with only one spatial dimension, up. The zero of up is at my hand. The motion of the chalk is a continuous series of events that start with the toss at a time selected to be the zero of time and returns to my hand at a later time T. In between, the chalk has occupied a set of places at specific times between zero and T. If you know the places for all times in that interval you have a trajectory. In Figure 4.16, we show the trajectory in a space-time diagram.

Any trajectory is only one of several that have the same total time interval T and start and stop at the same height. Why did nature chose the

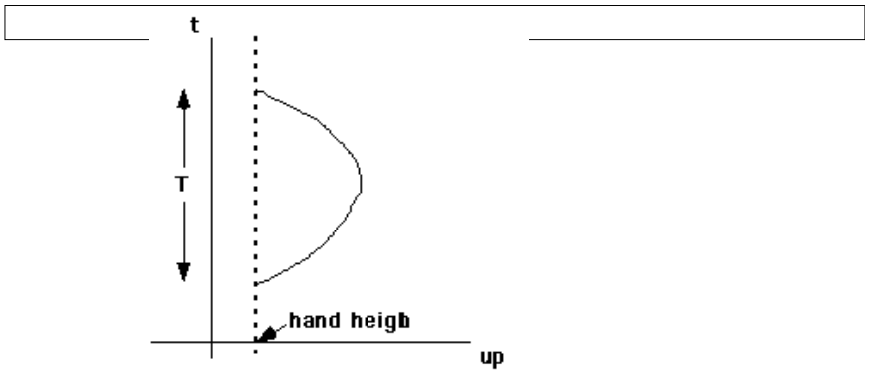

Figure 4.16: **Trajectory of a tossed piece of chalk** Chalk tossed from a height labeled zero rises with decreasing velocity until it reaches a peak and then returns to the hand after a time interval T.

one that she did? Several possible trajectories are shown in Figure 4.17. It will turn out that our rule will be that nature chooses the trajectory from all the possible trajectories that has the least action. Since we have not yet defined the action, this is a little difficult to understand. Not only that but the approach is so different from the Newtonian that we do not have a developed intuition for this way of describing the chosen dynamic .

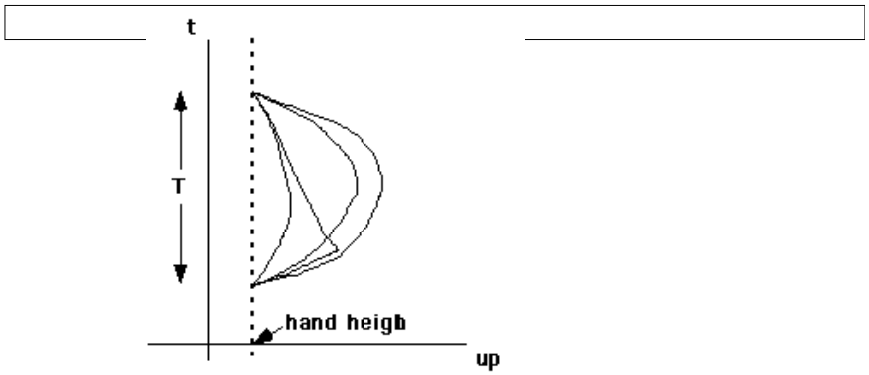

Figure 4.17: **Possible trajectories for a tossed piece of chalk** There are an infinity of trajectories that can connect the event at the start of the toss with the event at the return of the chalk to the hand at a later time T.

If you were approaching this problem from the Newtonian point of view, you would have used $\vec{f} = m\vec{a}$ and said that the chalk starts from a given place and given speed. Because there is a force, the attraction of the earth

for the chalk, there is an acceleration. Since there is an acceleration, the velocity changes. The velocity changes until it is reversed at the maximum height and starts to fall. While all this is happening, the chalk is tracing out a smooth arc in space time. This description is very different than the one that we will be using for action. In the Newtonian formulation, the determination of the trajectory is done at each instant of time at the place at which the chalk is at that time. The action approach on the other hand deals with the action over the entire trajectory. This is a global approach to dynamics. It will be difficult to reconcile these disparate seeming approaches but you have to recover the Newtonian approach for the case in which the chalk can be treated as a point particle and free to move up and down without constraint.

4.4.3 Definition of Action

Instead of $\vec{f} = m\vec{a}$ acting at each point on the body, there is now have a new rule: minimize the action over the trajectory. In other words, nature chooses the least action trajectory from all the trajectories that share the same initial and final event. This is a formulation of motion that is very much like that of Fermat's Least Time formulation for the paths of light in Section 3.2. To determine the trajectory, you pick two events, an initial event, x_0 and t_0, and a final event, x_f and t_f. There is a quantity called the action that is computed for every segment of the trajectory. Choose all possible trajectories and the natural trajectory is the one that has the least action.

The action is defined from a function of the positions and velocities called the Lagrangian. In this approach to dynamics, instead of trying to figure out what forces are causing the motion, you try to find what the correct Lagrangian is. In a real sense, when a modern physicist develops a new fundamental theory of some phenomena, it is by finding the correct Lagrangian so that the trajectory that yields the least action using that Lagrangian is the one that occurs naturally.

There is a slight technical difference in this case and the case of Fermat's least time. In this case, we create our trajectory segments by creating time slices, see Figure 4.18. For Fermat, the segments were sections along the length of the curve. As in the case of least time, the size of the time slices depends on the trajectory and the precision required. This gives a special role to the time variable. Also although we say all possible trajectories, for now, we will only deal with trajectories that advance in time positively. We will be able to lift this condition later, Section **??**.

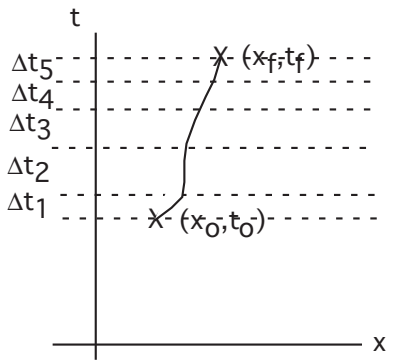

Figure 4.18: **Trajectory for the computation of the action** In order to compute the action for a given trajectory, the trajectory is divided into time slice pieces. For each time slice, the positions and the velocity can be determined. The action is then computed for that time slice and the contributions of each time slice are added to produce the overall action. The sizes of the time slices are determined by the rate of change along the trajectory.

For a simple point object like the piece of chalk moving up and down, the Lagrangian depends on the position and velocity of the object. Given the Lagrangian, the action is

$$S(x_f, t_f, x_0, t_0; trajectory) = \sum_{trajectory, x_0, t_0}^{x_f, t_f} L(x(t), v(t))\Delta t \tag{4.25}$$

Action has the dimensions of an energy times a time. Although this makes the dimensions easy to remember, it is misleading. As we will learn later, the concept of energy is derivative from the action not the other way around, see Section 5.4 . It would be better to say that energy is dimensionally an action divided by a time. In terms of fundamental dimensional units, the units of action are $\frac{\text{mass}\times\text{length}^2}{\text{time}}$. From Equation 4.25, the Lagrangian itself has the dimensions of an energy, $\frac{\text{mass}\times\text{length}^2}{\text{time}^2}$.

The rule that Lagrange found that would reproduce $\vec{f} = m\vec{a}$ for unconstrained systems and also work for more general situations is that the Lagrangian, $L(x(t), v(t))$, should be the difference in the kinetic energy and the potential energy.

$$L(x(t), v(t)) = \frac{mv^2}{2} - V(x) \tag{4.26}$$

where $V(x)$ is the potential energy. Later, Section 4.4.5, we will show how this reproduces Newton's laws. It is important to again point out that although this approach requires that you know the kinetic energy and potential energy that these concepts are actually derived from the actions and not the other way. For now, it seems that you need to know the potential energy before you can write the Lagrangian. This is only for historical and pedagogical reasons. When a modern physicist is struggling with understanding some basic new phenomena, it is the other way around. We start with a Lagrangian and then see what the consequences are. It will also turn out that since the actions become the basis of all dynamics, it is the idea that theories that unify other earlier independent theories are considered unified when all the consequences of the theory arise from a single controlling Lagrangean. In modern language, Maxwell unified the electric and magnetic forces because the entire ensemble of equations is derivable from a single Lagrangian and the least action principle.

4.4.4 Trajectory of a Free Particle

To test our new dynamic, let's look at the simplest situation possible – a free particle. A free particle is one that has no forces acting on it. All places have the same energy value and thus $V(x) = 0$. Using Lagrange's rule to get the solution for the free particle in old fashioned physics, we chose the Lagrangian that is just the kinetic energy or $L(v(t)) = \frac{mv^2}{2}$. To make it even simpler, let's require that the released particle is to return to the original position after a time T. The action is

$$S(0,0,0,T,traj.) \;=\; \sum_{traj.,0,0}^{0,T} \frac{mv^2}{2}\Delta t \tag{4.27}$$

As was stated in the review section, Section 1.2.3, a free particle at rest will remain at rest. Therefore, the natural trajectory for this case is the one that is at the starting place at all times. The is a straight line along the t axis connecting $(0,0)$ and $(0,\mathrm{T})$. How do we obtain this same result using action?

Note that the action is a positive definite quantity for all velocities. Therefore any trajectory that has a non-zero velocity anywhere in the time interval will have a positive action. The trajectory that has $v(t) = 0$ for all t in the interval has an action of zero. This is clearly a minimum of the action since all other trajectories will have a positive action. Thus this is the natural path. Actually any Lagrangian with v^2 in it will accomplish the

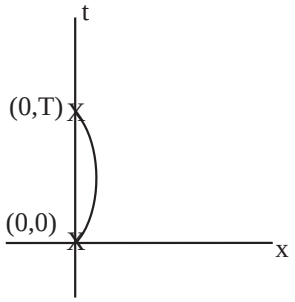

Figure 4.19: **Space-time diagrams for the action for a free particle** A particle with no forces acting on it moves between two events,$(0,0)$ and $(0,\mathrm{T})$. A possible trajectory is shown. Our experience with force free motion is that the straight line trajectory is the one that nature chooses; the particle remains at the point of release.

same thing. The m is in it to give it the correct dimensions and the 2 for historical reasons. In fact, the m that is in the Lagrangian is the definition of mass. More on this later, see SectionSec:Mass.

Using this same result and remembering the material on Galilean invariance in Section 1.2.3, we can solve a more general problem. Suppose we have a free particle that moves through the two events $(0,0)$ and (x_f,t_f). Again, since the particle is free, the natural trajectory is the straight line connecting these events. To an observer moving by us at a speed of $v=\frac{x_f}{t_f}$, the object is a rest during the entire time interval. To that observer it is free and the initial an final events are $(0,0)$ and $(0,t_f)$ and the natural path is the straight line along the t axis as before. Thus to us the natural trajectory will be the straight line with slope $fract_f x_f$. Let's obtain this same result with a direct analysis.

Consider a general trajectory connecting events $(0,0)$ and (x_f,t_f), see Figure 4.20. Our problem is to find all possible trajectories between these events and then, for each trajectory, find the action. As we discussed about paths when dealing with the Fermat's least time approach to optics in Section 3.3.7. path space is a rich mathematical structure. We want to do analysis. To do analysis we have to reduce the complexity of path space to something that can be described by functions. There are all these same difficulties when dealing with trajectories. To simplify our trajectory space, we reduce the trajectories that we consider to those that are "once kinked". Place the kink along the line $t=\frac{t_f}{2}$, see Figure 4.20. In this reduced space, trajectories can be labeled by the distance, a, of the kink from the event

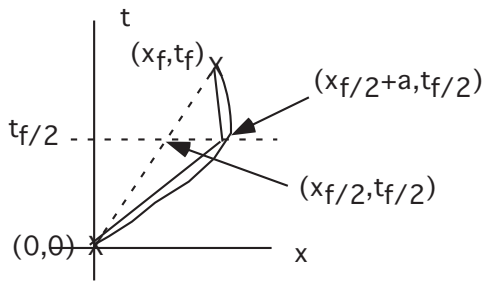

Figure 4.20: **Space-time diagrams for the action for a free particle that changes position** A particle with no forces acting on it moves between two events,$(0,0)$ and (x_f, t_f). A possible trajectory is shown. The general trajectory connecting these events would be very difficult to describe. We will approximate the trajectory with a trajectory that is kinked at the mid-time and straight otherwise.

($\frac{x_f}{2}, \frac{t_f}{2}$ along that line. Using this trajectory in the appropriately modified Equation 4.27 to take account of the new ending event, and the fact that the inverse slope of the line is the velocity in that segment, it is easy to compute the action for the trajectory labeled a. It is

$$S(0,0,x_f,t_f,traj=a) = \frac{m}{2}\left(\frac{\left(\frac{x_f}{2}+a\right)^2}{\frac{t_f}{2}} + \frac{\left(\frac{x_f}{2}-a\right)^2}{\frac{t_f}{2}}\right). \tag{4.28}$$

This is an even function of a and thus has a minimum at $a = 0$. This confirms our result that the natural trajectory, the constant velocity trajectory, is the least action trajectory.

4.4.5 Proof that the Least Action Reproduces Newtonian Physics

See Feynman's famous lecture. It was handed out in class

4.4.6 Examples of action – gravitation near a flat earth

As a simple example that we are all familiar with, consider the case of motion above the surface of the earth. Here the energy of position, the potential energy, is due to the gravitational interaction of a massive body with the earth. For this case, the potential energy at a height h above the earth is $V(\vec{r}) = -\frac{Gm_e m}{R_e+h}$, where m_e is the mass of the earth, m the mass of the body,

and R_e is the radius of the earth. For motion near the surface, a few meters up or down, from "Things Everyone Should Know," Section 1.4.2, we can use $(1+x)^n \approx 1+n\,x$ for $x \ll 1$ to reduce this to

$$V(h) = -m\left(\frac{Gm_e}{R_e}\left(1-\frac{h}{R_e}\right)\right) = V(R_e) + mgh,$$

where we recognize $g = \frac{Gm_e}{R_e^2}$. Since this potential is to be used in an action, as we will see later in Section 5.4, changing the action by a constant does not change the physical results in a significant way, we can drop the $V(R_e)$ term. This reduces the potential energy for objects moving in the near vicinity of the earth to

$$V(h) = mgh. \tag{4.29}$$

Another way to look at this result is to say that for motion restricted to be near the surface of the earth, the earth appears as an infinite plane. In this case, the force of gravity above the plane can not depend on anything, in particular, the height above the plane or the position sideways over the plane. Thus the force also can only be toward or away from the plane. Then realizing from the analysis above in Section 4.4.5 that the change in potential as you change position is the force, the only form for the potential in this case is $mgh + \text{constant}$.

For now let us consider only up and down motion, not any sideways motion. The potential energy is mgh where h is the height. Thus the action for any trajectory between an initial height, h_0 at time t_0 and final height, h_f at time t_f is

$$S(h_0,t_0,h_f,t_f;traj.) = \sum_{traj.,h_0,t_0}^{h_f,t_f}\left(\frac{mv^2}{2}-mgh\right)\Delta t \tag{4.30}$$

where the path is given by $h(t)$. Note that if you know $h(t)$, you also know $v(t)$. You can see from the form of the action that you will lower the action by having $h(t)$ to be at large h for as much time as possible. The problem is that since the initial and final position and time are given, it takes high velocity to get to large h. The high velocity increases the action. $\Longrightarrow$ There is a single least action path. This is the trajectory that the particle follows.

Let's get more specific. This is again the problem of a piece of chalk tossed up in the air. First the simplest case, the chalk is released and returns to the same height after a t time T.

We need to study the action for all trajectories connecting these events. Again, because of the complexity of the idea of all trajectories, we will need

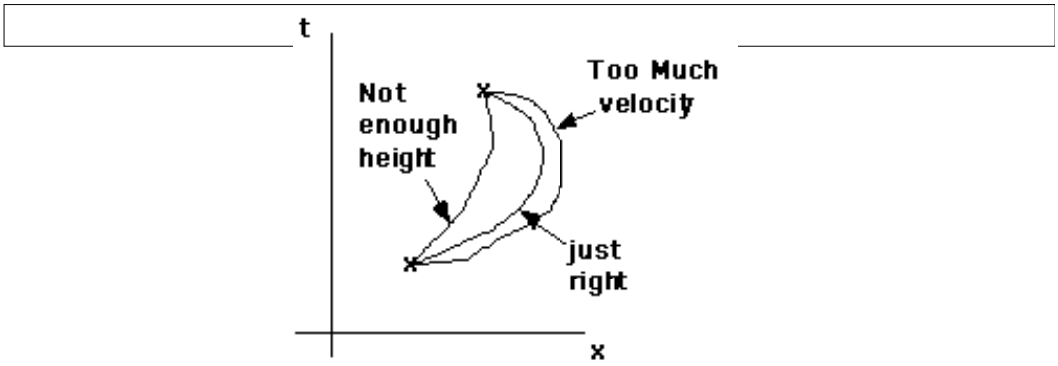

Figure 4.21: **Trajectory for Particle in Uniform Gravitational Field** Space-time diagrams for calculation of the action for a particle in a uniform gravitational field. The least action trajectory is just the right compromise between too much kinetic energy and some potential energy.

to reduce the number of trajectories. A first step is to use our experience to limit ourselves to simple trajectories that rise smoothly to a peak at some height a at which time the velocity is zero and then returns over a trajectory that is a reflection of the one on the rise. Our natural trajectory must be in that family. This is still a very rich family and too rich to do analysis. This is the same problem that we had with the Fermat's Least Time, Section 3.3.7, and the free particle, Section 4.4.4. As in the latter case, the once kinked path can be used to approximate the family of smooth trajectories that have these properties, see Figure 4.22. Here again the variable a is the height of the approximate trajectory but more importantly now it is a label that can be used to specify the particular trajectory from the family with which we are dealing.

Since this approximate trajectory is broken line segments, it is relatively easy to compute the action.

$$S(0,0,0,T;traj.) = \sum_{(0,0)\ \text{traj.}}^{(0,T)} \left(\frac{mv^2}{2} - mgh\right)\Delta t. \tag{4.31}$$

For a straight line path, v is a constant and is the inverse slope of the line, and is $\frac{a}{\frac{T}{2}}$ in magnitude for both segments. The height is a more subtle question since it varies with time from 0 to a. Being reasonable, we can use the average height, $\frac{a}{2}$. For the sophisticates among you, there is the problem that the concept of average is a not trivial, see Section **??**. Thus the action

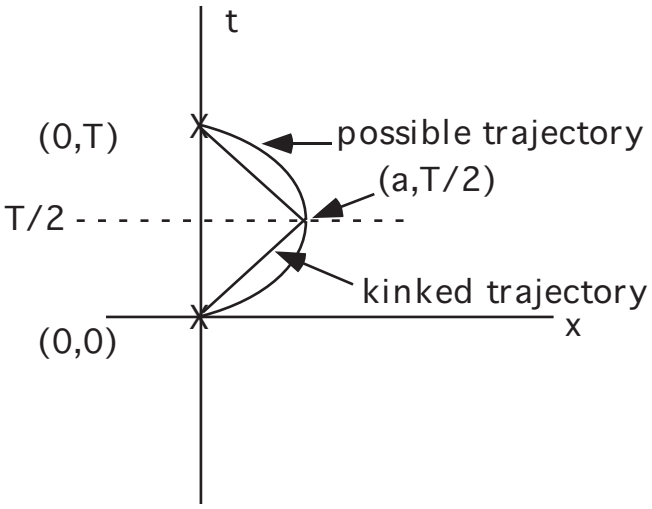

Figure 4.22: **Possible trajectory for the action for a particle in a uniform gravitational field** A piece of chalk is tossed upward and caught later at the the same height. A possible trajectory is shown. The natural trajectory is one from the family of smooth trajectories that rise to a peak at a height a smoothly and then return to a lower height on a reflected trajectory. This is still a large family of trajectories. We can approximate the members of this family with a once kinked trajectory with the same height at the time $\frac{T}{2}$.

for the first segment is

$$S_1(T,a) = \frac{ma^2}{2\left(\frac{T}{2}\right)^2}\frac{T}{2} - \frac{mga}{2}\frac{T}{2}. \tag{4.32}$$

Note that once I have made a mapping of the paths onto the line that S becomes a regular function of the path label, a, instead of a functional. Although the velocity is negative, since only v^2 enters the lagrangian, the action on the second segment is the same and the total action is

$$S(T,a) = 2S_1(T,a) = ma\left(2\,\frac{a}{T} - \frac{g}{2}T\right) \tag{4.33}$$

This has zero's at $a = 0$ and $a = \frac{gT^2}{4}$. The dependence of the action on the path label a is shown in Figure 4.23. I have used dimensions in which $g = T = 1$.

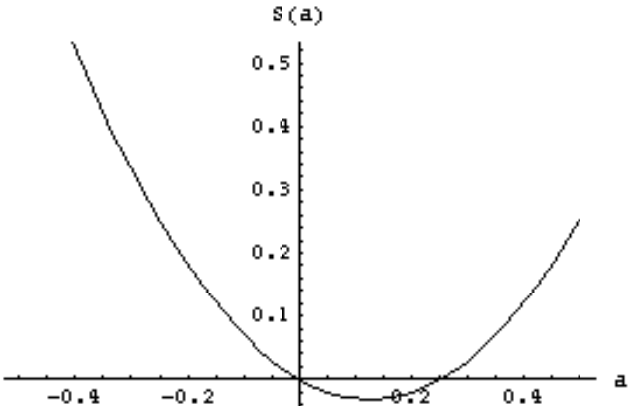

Figure 4.23: **Action as a function of** a The action as a function of of the trajectory label a. This curve is a combination of a parabola, $\frac{2m}{T}a^2$, concave up with its vertex at the origin and a straight line, $-\frac{mgT}{2}a$, with negative slope through the origin.

We can see that there is a minimum half way between the two zero's at $a = 0$ and $a = \frac{gT^2}{4}$. This implies that the trajectory from this set that is the least action trajectory is the one with

$$a_{least\ action} = \frac{gT^2}{8}. \tag{4.34}$$

Since this is not only the path selecting parameter but is also the height, we get that the height is $\frac{gT^2}{8}$.

4.4.7 Same Example done another way

I am going to do some mathematics here that I do not expect that you will be able to reproduce. I do this to show you that it can be done and that the ideas of mathematics are useful. You are not expected to do integrals and take derivatives although you should be able to follow a development using them.

Once again, we want to examine the case of an object of mass m moving in the vicinity of the earth. We can also guess that the correct answer for the height as a function of time is a parabola, all parabolas that fit the time interval are of the form $h(t) = at(t - T) \Rightarrow v(t) = 2at - aT$, where a is label of the path in path space. In this case, a has the dimension of an acceleration, $L \stackrel{\text{dim}}{=} a \times T^2$ or $a \stackrel{\text{dim}}{=} \frac{L}{T^2}$.

The Lagrangian is $L = \frac{1}{2}mv^2 - mgh$ and the action is

$$\begin{aligned} S &= \int_{(x_0,t_0),\text{Path}}^{(x_f,t_f)} \Big(\frac{1}{2}mv^2 - mgh\Big)dt \\ &= m\int_0^T \Big(\frac{1}{2}(2at - aT)^2 \ - gat(t-T)\Big)\,dt \\ &= m\Big(\frac{a^2T^3}{6} + \frac{1}{6}agT^3\Big) \end{aligned}$$

This can be factored to $S = \frac{mT^3}{6}a(a+g)$.

To find the minimum, we can again realize that there are two zeros of S(a). One at $a = 0$ and one at $a = -g$. The minimum is half way between them at $a_{least\ action} = -\frac{g}{2}$

Otherwise, we can take the derivative of S(a) with respect to a and set it equal to zero. Thus

$$\begin{aligned} \frac{dS}{da} &= \frac{d}{da}\left(\frac{mT^3}{6}a(a+g)\right) \\ &= \frac{1}{6}amT^3 + \frac{1}{6}(a+g)mT^3 \\ &= \frac{1}{6}\ (2\ a+g)\ m\ T^3 \end{aligned} \tag{4.35}$$

or $a_{least\ action} = -\frac{g}{2}$ is the natural trajectory. In Figure 4.24, note how the action varies with a. Again I have used units with $g = T = 1$.

4.4.8 Digression on averages and slicing

It should come as no surprise that most people do not think hard about what they mean by averages. This is often exemplified by the puzzle:

> Consider two towns that are one hundred miles apart, for instance Austin and College Station. You want to travel between them with an average speed of fifty miles per hour. You leave Austin but get caught behind a very long funeral procession that you cannot pass that is also going to Hicksville, half way between Austin and College Station. If the funeral procession held your speed to an average of twenty five miles per hour between Austin and Hicksville, how fast do you have to drive in the remainder of the trip to obtain your desired average of fifty miles per hour?

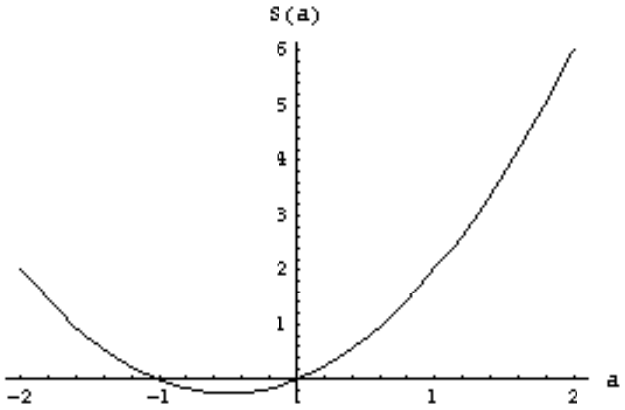

Figure 4.24: **Action as a function of a as an acceleration** Action as a function of a when the parameter a has the dimensions of an acceleration. This example shows that the trajectory label does not have to be a height.

The accepted answer is that you have to go infinitely fast. This is because in the portion of the trip between Austin and Hicksville has taken two hours and, in order to average fifty miles per hour on a one hundred mile trip, you need two hours of travel time. Your time is all used up. Another answer that is often given is seventy five miles per hour in the second segment of the trip. Although not the accepted answer, there is a sense in which this answer is also correct.

How can there be two correct and different answers to the same question? The answer is that, as so often happens, the question is not well posed. The issue is what average is being asked for?

How do you compute an average? What is the average of the set of numbers 1,1,3,1,4,5,7. The rule is that you add up all the numbers, the sum is 22, and divide by the number of numbers which is 7. The result is $\frac{22}{7}$ or a pretty good π. Looking at this process more closely, you realize that what we have is an ordered set of numbers: the first number is 1, the second number is 1, the third number is 3, and so forth. We have a mapping of the set of integers onto our set of numbers, a discrete function. In this language, we can say that to compute the average by sequencing through our ordered set: add the first number to the second, add that sum to the third, add that sum to the fourth, and so forth. You divide by the number of times you take a number. We can display this algorithm for this case in the form

$$Average \equiv \frac{\sum_{i=1}^{n} f(i)}{n} \tag{4.36}$$

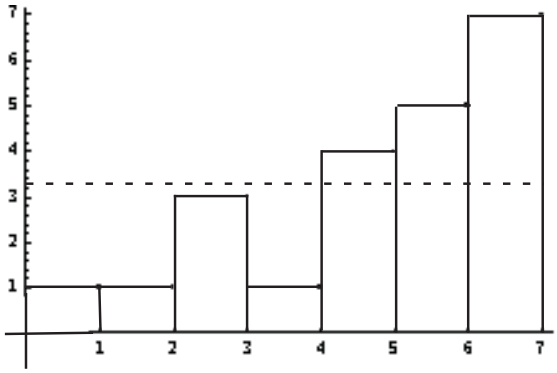

Figure 4.25: **Plot of Discrete Function for Averaging** The set of numbers, 1,1,3,1,4,5,7, are plotted as a discrete function in terms of the postion of the number in the table. In addition, a bar is drawn from the next lowest location at the height of the value. Also the average, $\frac{22}{7}$ is shown as a dotted line. The area under the barred segments and the area under the dotted line are the same. This allows a more general definition of the process of averaging: the average times the interval is equal to the area under the barred plot of the discrete function generated by the set of numbers.

where $f(i)$ is the value of our discrete function for the i element of the table and n is the number of entries or more interesting as a plot of the discrete function that we have generated, see Figure 4.25. In addition to plotting the function as a bar graph, the average is shown as a horizontal dotted line. From the figure, it can be seen that the area under the bars of the bar graph and the area under the dotted line are the the same. This leads to an alternative algorithm for finding the average of a set of numbers: construct the bar graph for the set of numbers and calculate the area under the bar graph divide this area by the number of elements in the set. The advantage of this definition is that it is easy to extend to situations where you want the average over a continuously varying set. An algorithm for this definition is:

$$Average \equiv \frac{\sum_{i=1}^{n} f(i)\Delta i}{\sum_{i=1}^{n} \Delta i} \tag{4.37}$$

where Δi is the width of the elements of the bar graph.

From this construction, the more general definition of the average can be developed that will work for continuous functions. The integral form of

this same definition is

$$Average \equiv < f >_x \equiv \frac{\int_{x_0}^{x_f} f(x)dx}{\int_{x_0}^{x_f} dx} \tag{4.38}$$

where I have introduced a standard notation for taking the average. The subscript x indicates that the average is weighted by the variable x. The important point is that in different circumstances different weighting factors are appropriate and, although the definition looks as if it is independent of the choice of the weighting factor, it is not.

Now let's go back to our problem of the trip from Austin to College Station. To calculate an average, we need a set of numbers. How do we get the numbers? We have to decide what the weighting factor is. There are an infinity of choices but two are particularly obvious, time slicing and space slicing. Were it not for a particular property of time slicing, space slicing is the easier because you will generally know how fast you can go at a given place. Thus to get the average velocity for space slicing choose spatial intervals and find the velocity in each. Applying this method to the Austin-College Station trip would yield the result that a speed of 75 $\frac{\text{m}}{\text{hr}}$ in the second segment would give an average speed of 50 $\frac{\text{m}}{\text{hr}}$.

The more accepted answer is the one that comes from using time slicing. In this case, the average is computed simply for a kinematic quantity like velocity because it is defined in terms of a time derivative. In other words,

$$\begin{aligned} < v >_t &= \frac{\int_{t_0}^{t_f} v(t)dt}{\int_{t_0}^{t_f} dt} \\ &= \frac{\int_{t_0}^{t_f} \frac{dx}{dt}dt}{t_f - t_0} \\ &= \frac{x_f - x_0}{t_f - t_0}, \end{aligned} \tag{4.39}$$

and thus the average velocity is just the displacement divided by the time interval. You loss track of the fact that you time sliced. Unless stated otherwise it is customary to assume that what is wanted is time averaged.

In Section 4.4.6, there was some question regarding the height to use in the Lagrangian since it varied in the segment. We now see that the correct choice is the time average since the action is time sliced. For cases where you replace the curved trajectory with a straight line the two averages always come out the same and thus our substitution was correct. In cases where

you are using a more subtle structure such as in Section 4.4.7, you would get the wrong answer by substituting the mean position.

It also important to note that the action principle always uses time slicing – it is a part of the definitiion. It could turn out that, in some applications, a different slicing is easier to understand, see Section 13.1. In fact, when we did Fermat least time, we did segment slicing. Whatever slicing technique is chosen, the action must always be evaluated using a time slicing.

4.4.9 More Examples of Actions

Scattering

Two particles, one of mass m_1 and the other of mass m_2 collide. After the collision, the particles move away from each other, both still with masses m_1 and m_2. This is a very special problem whose important cannot be over emphasized. In a very real sense, when we probe the nature of the elementary constituents of matter, scattering experiments are the primary source of our knowledge. In addition, the process is so basic that it will allow us to begin to better understand many fundamental issues.

How do we handle this process? First, we have to decide what is meant by two independent particles. Before the particles make contact, they move as if the other particle was not present, i. e. they are independent. It is reasonable therefore to assume that while they are apart or not interacting, the two particles actions add and are the usual free particle action. In other words, there is a free particle action the tells you all the properties of what is meant by a particle and its nature. For our construction of the action of the free particle in Section 4.4.4, we used the Lagrangian $L(x, v) = \frac{mv^2}{2}$. The Lagrangian says the the object identified as a free particle does not treat different places differently and thus there is no x dependence in the Lagrangian. If we want to recover Newton's Law, see Section 4.4.5, we use the usual classical kinetic energy. We will find that in other circumstances, for instance for a rapidly moving particle, Section 13.1, that a different free particle Lagrangian is appropriate. If we wanted to describe something more complicated than a point particle, say a small rod, we would need elements that deal with what a rod is such as moment of inertia and directional variables.

By using as the action the sum of the single particle actions, the properties of the total system will be the sum of the properties of the parts. If we did this though, and this was the end of it, nothing interesting would ever happen; the particles would merely pass through each other unchanged in

their motion. We want them to scatter. Thus in addition, we need to add a part that carries the interaction. The interaction will have a Lagrangian that is made up of relationship variables such as their separation in addition to the particle labels. In other words, the action is made up of the following parts:

$$\begin{aligned} \text{Total Action} &= \text{Free Action(variables particle 1)} \\ &+ \text{Free Action(variables particle 2)} \\ &+ \text{Interaction Action(variables particle 1,} \\ &\quad \text{variables particle 2, relationship variables).} \end{aligned} \tag{4.40}$$

Of course, it is actually redundant to list the relationship variables in the interaction action since they will be composed of the variables of particle 1 and 2 anyway. The importance of displaying the relationship variables separately is to be able to say that, for a scattering situation, the interaction action is zero when the relationship variables such as the separation are large. In a collision, we assume that most of the time the particles travel toward or away from each other and that the interaction terms contribute only for a short time when the particles are in contact and thus this interaction term is small and does not add significantly to the total action of the process. Another point to note is that, since the interaction terms are dominated by the relationship variables, the contribution from the interaction action should be independent of where and when the collision takes place. Thus, we can write the action for this simple one dimensional scattering process as

$$S = \sum_{(x_{10},t_{10}),Path}^{(x_{1f},t_{1f})} m_1 \frac{{v_1}^2}{2}\Delta t + \sum_{(x_{20},t_{20}),Path}^{(x_{2f},t_{2f})} m_2 \frac{{v_2}^2}{2}\Delta t + A, \tag{4.41}$$

where A represents the interaction action. The scattering process is shown in Figure 4.26.

We want to do all paths but we know that the straight path is the least action for a free particle and so all we need to do is use straight paths between the initial and collision and collision and final events. We can immediately write down the action as a function of the position and time of the collision. The coordinates of that event are the only free parameters in the problem.

Note that we are being consistent in our use of action. When you talk about collisions in the general physics class you set the initial velocities. Here we use the initial and final events. Evaluating the free particle actions,

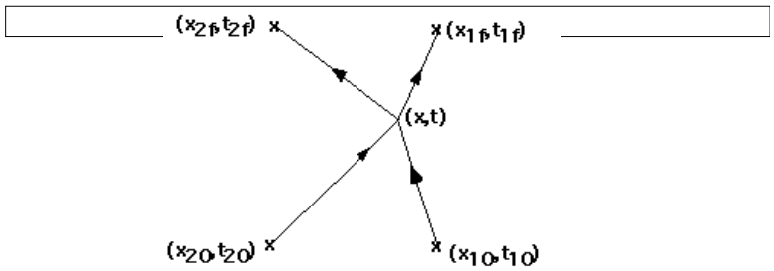

Figure 4.26: **Space-time diagram for a scattering event** Two particles of mass m_1 and m_2 free to move in one spatial dimension are directed at each other and collide at the event (x,t) and then move apart . A space-time diagram for a scattering event with particle one starting at event (x_{1_0}, t_{1_0}) and returning to (x_{1_f}, t_{1_f}) and particle two starting at event (x_{2_0}, t_{2_0}) and returning to (x_{2_f}, t_{2_f}) is shown. Although all trajectories connecting the initial and final events and the collision event should be examined, we know that free particles have a natural trajectory that is a straight line, see Section 4.4.4.

for this system of trajectories, the action is

$$S = \frac{m_1}{2}\frac{(x - x_{1_0})^2}{(t - t_{1_0})} + \frac{m_2}{2}\frac{(x - x_{2_0})^2}{(t - t_{2_0})} + \frac{m_1}{2}\frac{(x_{1_f} - x)^2}{(t_{1_f} - t)} + \frac{m_2}{2}\frac{(x_{2_f} - x)^2}{(t_{2_f} - t)} + A. \tag{4.42}$$

We want to find the trajectory that has the least action and since we have now reduced the world of trajectories to the label of the collision point, x and t. Thus we need to minimize this in what are now the labels, x and t. You could plot this and find the minimum by hand , see Figure 4.27, but, if you allow me to use calculus, I can find a simple analytic expression for the $x = x_{min}$ and $t = t_{min}$ that yields the least action. This means taking the derivatives with respect to x and t and finding the value of x and t that satisfy $\frac{\partial S}{\partial x} = 0$ and $\frac{\partial S}{\partial t} = 0$. This x and t label the naturally occurring trajectory.

Take my word for it. The condition for a minimum in x is

$$m_1\frac{(x_{min} - x_{1_0})}{(t_{min} - t_{1_0})} + m_2\frac{(x_{min} - x_{2_0})}{(t_{min} - t_{2_0})} - m_1\frac{(x_{1_f} - x_{min})}{(t_{1_f} - t_{min})} - m_2\frac{(x_{2_f} - x_{min})}{(t_{2_f} - t_{min})} = 0 \tag{4.43}$$

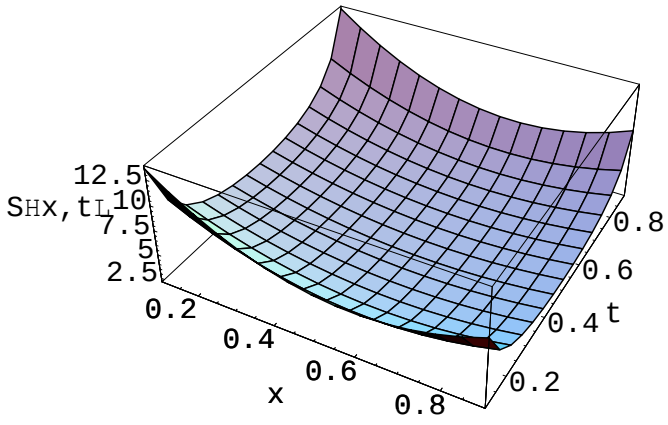

Figure 4.27: **Action for a Scattering Event** Action as a function of x and t for a scattering event shown in Figure 4.26. There is a clear minimum and it occurs at the points at which Equation 4.44 and Equation 4.46 are satisfied.

or

$$m_1\frac{(x_{min}-x_{1_0})}{(t_{min}-t_{1_0})}+m_2\frac{(x_{min}-x_{2_0})}{(t_{min}-t_{2_0})}=m_1\frac{(x_{1_f}-x_{min})}{(t_{1_f}-t_{min})}+m_2\frac{(x_{2_f}-x_{min})}{(t_{2_f}-t_{min})} \tag{4.44}$$

Realizing that momentum is mv in classical physics and that v is the difference in positions divided by the the differences in times, this is the statement that the momentum into the collision is equal to the momentum out of the collision.

The condition that there is a minimum in t gives

$$\frac{m_1}{2}\frac{(x_{min}-x_{1_0})^2}{(t_{min}-t_{1_0})^2}+\frac{m_2}{2}\frac{(x_{min}-x_{2_0})^2}{(t_{min}-t_{2_0})^2}-\frac{m_1}{2}\frac{(x_{1_f}-x_{min})^2}{(t_{1_f}-t_{min})^2}-\frac{m_2}{2}\frac{(x_{2_f}-x_{min})^2}{(t_{2_f}-t_{min})^2}=0 \tag{4.45}$$

or

$$\frac{m_1}{2}\frac{(x_{min}-x_{1_0})^2}{(t_{min}-t_{1_0})^2}+\frac{m_2}{2}\frac{(x_{min}-x_{2_0})^2}{(t_{min}-t_{2_0})^2}=\frac{m_1}{2}\frac{(x_{1_f}-x_{min})^2}{(t_{1_f}-t_{min})^2}+\frac{m_2}{2}\frac{(x_{2_f}-x_{min})^2}{(t_{2_f}-t_{min})^2} \tag{4.46}$$

Which is the same as the statement that the energy into the collision event is equal to the energy out of it.

Figure 4.27 shows the action as a function of the position and time of the collision event. This is for the case that $\frac{m_2}{m_1}$is 1.5 and the original and final events for particle 1 are (0,0) and (0,1) and for particle 2 are (1,0) and(1,1).

This exercise also gives us an interesting insight on what mass is. In an early assignment in this course, you were asked to devise a method for measuring mass that does not relay on gravity. Some of you came up with the idea of using collisions to define a mass scale. You can see that this analysis is directly relevant to that kind of definition. In the construction of the action, for the case of the single particle, mass is an overall factor; it is the thing you put in front of the v^2, in the action. If the world consisted of only one particle, mass would be irrelevant since all it does is multiply the action. The process of finding the natural trajectory is unchanged by the an overall scale factor on the action. Mass becomes interesting only when you have more than one particle. If there is more than one particle, you can not remove all the masses with a single scaling factor. The ratios of the mass remain. Consider a scattering event between two particles with the initial and final positions of the two particles the same before and after the collision. If the particles had equal masses, the position of the collision event is at the center. The trajectories of both particles are equally kinked. On the other hand, the higher the mass ratio of say the second particle, the less the trajectory associated with that particle will kink when it collides with another particle. In the limit of a very large mass second particle, there is no bending of the second trajectory and it looks like the first particle has hit a brick wall. This is the essence of inertia.