Quantum Tunneling Overview

Introduction

At the heart of quantum mechanics is the idea that matter behaves as both a wave and a particle. Experimental evidence from approximately 1900 through present day shows that as we look more closely at the behavior of very small things, such as molecules, atoms, and fundamental particles, the intuitive classical predictions of how matter should behave are not adequate in predicting the results of experiments. Perhaps the most fundamental reason for this is that these small particles were clasically treated as point-like objects whose position and momentum can be predicted with absolute certainty. It was not until the leading physicists of the early 20th century started toying with the idea that these very small objects may behave as waves at certain times and particles other times that they began predicting the results of experiments.

One of the interesting findings of quantum mechanics is that, due to the wave-like nature of matter, small particles can be found in places that would classically be forbidden. This phenomenon is called "quantum tunneling", and it has allowed for new technologies to be developed throughout the 20th century. Such applications are: the scanning-tunneling microscope, tunneling diodes, tunneling field-effect transistors, and the understanding of radioactive decay (which, for example, powers any nuclear power plant). This phenomenon not only demonstrates the 'strangeness' of quantum mechanics, but also plays a fundamental role in society, and is therefore an important subject in any quantum mechanics course. In the rest of this paper, we describe in greater detail what tunneling is, and how it can be treated mathematically.

Basics of Quantum Mechanics

Energy barriers are ubiquitous in physics. A skate board ramp, an electronic circuit element, a material's emission properties, and much more, can be described with the concept of an energy barrier. For example, a skateboard ramp provides a gravitational energy barrier for the skater, such that while the skater is on the ramp, his/her energy is constrained by the ramp. Similarly, a circuit element can provide an energy barrier for electrons, such that only electrons with a certain energy may cross the circuit element.

Quantum tunneling is a problem that involves an energy barrier. Specifically, this barrier tells us about how a quantum particle (such as an electron or proton or small molecule), can be spatially and temporally located within a region of space. One barrier to consider is the one shown in Figure 1. In this figure, three 'regions' exist in one dimension of space (imagine a very small wire, separated by another wire a distance L away from each other).

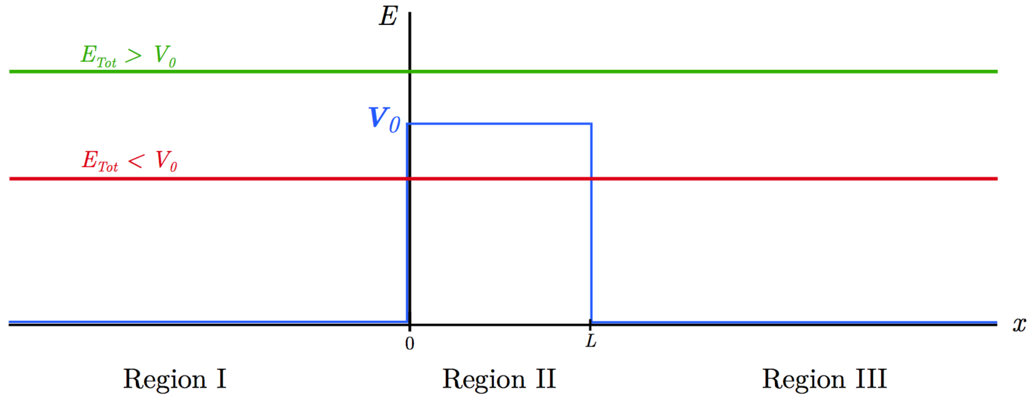

Figure 1: Energy Barrier for a Quantum Particle

In considering how the particle behaves near this barrier, we use the most fundamental equation in quantum mechanics: the Schrödinger equation. The Schrödinger equation describes a quantum particle's 'wave function', analagous to how Maxwell's equations describe electric and magnetic fields.

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(x) + V(x)\psi(x) = E\psi(x) \tag{1}$$

In the Schrödinger equation, $\hbar$ is a constant, m is the mass of the particle under consideration, $V(x)$ is the energy barrier that the particle will see, E is the energy of the particle, and $\psi(x)$ is the wave function of the particle. The wave function is really what we want to figure out from this equation. If we know the mass of the particle, the energy of the particle, and the energy barrier that the particle will encounter, we can solve for $\psi(x)$, and find what we are looking for. Once we know $\psi(x)$, we can calculate many other properties, and essentially know 'everything' there is to know about the system.

Physically, the wave function tells us something about the probability of finding the particle in different locations of space. We know that the particle must be located somewhere in space, and this can be represented in the following way:

$$\int_{-\infty}^{+\infty} |\psi(x)|^2 \, dx = 1 \tag{2}$$

This equation states that by taking the absolute value of the wave function, squaring it, and summing that quantity over all space (or in this case, over an infinitely long one-dimensional line), we should find that quantity to be exactly equal to 1. In other words, because the square of the wave function tells us the probability of finding a particle within a given region of space, if we look for the particle in the entirety of space, our probability of finding it is 100%. On the face of it, this equation might not seem to tell us much. However, this property is frequently exploited in quantum mechanics. Since we can never know the position of a particle with 100% accuracy, we are forced to use this relationship and say that the particle will certainly be located somewhere in a select region of space.

Basics of Tunneling

In studying quantum tunneling, we can solve the Schrödinger equation for the energy barrier shown in Figure 1. We know the energy of the barrier throughout all space ($V(x)$), we can pick an energy (E) for the particle, and we can assume that the particle has some mass that is known (m). With all of this, we can solve for $\psi(x)$ in each of the three regions.

For Region I and III, $V(x) = 0$, while in Region II, $V(x)$ is a constant value (we will call it V_0). Mathematically, this is written:

$$V(x) = \begin{cases} 0, & \text{for } x < 0 \text{ and } x > L \\ V_0, & \text{for } 0 \leq x \leq L \end{cases} \tag{3}$$

At this point, we need to know the energy of the electron (E). There are really three choices for the energy that we care about: the case of $E > V_0$, the case of $E = V_0$, and the case of $E < V_0$.[1] The case of $E = 0$ is one that the tutorial does not cover, so we do not consider it here.

In the case of both $E > V_0$ and $E < V_0$, we first solve for $\psi(x)$ in each of the three regions separately. The Schrödinger equation is a second-order equation, so every time we solve it, we will have 2 unknown quantities. We have to solve this equation for three different regions in space, so we should expect to have 6 unknown quantities, for which, we will need boundary conditions to pin down. We can start figuring out the boundary conditions by think about two constraints on the wave function. The first of these is: $\psi(x = 0)_{R1} = \psi(x = 0)_{R2}$ and $\psi(x = L)_{R2} = \psi(x = L)_{R3}$, meaning that the wave function must be continuous across the two boundaries (so that there can be no discontinuities in $\psi(x)$. The second constraint is that the slope of the wave function across these boundaries must also be the same: $\psi(x = 0)_{R1} = \psi(x = 0)_{R2}$

[1] Also note that E can never be less than 0, since in such a case, there is no solution to the Schrödinger equation. We assume that $E > 0$ for this problem.

and $\psi'(x = L)_{R2} = \psi'(x = L)_{R3}$.[2] These two conditions provide 'boundary conditions' for solving the Schrödinger equation given in (1).

In the case of both $E > V_0$ and $E < V_0$, the solutions for $\psi(x)$ take the same form in Region I and Region III. We skip over the details of solving this, and simply write them here:

$$\psi_{R1}(x) = Ae^{ikx} + Be^{-ikx} \tag{4}$$
$$\psi_{R3}(x) = Fe^{ikx} + Ge^{-ikx} \tag{5}$$

where $k = \sqrt{2mE/\hbar^2}$, and we give different amplitudes (A, B, F and G) for the two regions to indicate that the amplitudes may be different across the different regions. Note that these solutions are just sine waves, with one sine wave (the $+ikx$ term) traveling towards positive x and the other (the $-ikx$ term) traveling towards negative x. A sum of sine waves simply adds to another sine wave, so the wave function in both Region I and Region III looks like a sinusoidal wave.

If we consider the case of $E > V_0$, the solution to the Schrödinger equation in Region II takes the same form as (4), but the constant k is slightly different. We can write this solution as:

$$\psi_{R2}(x) = Ce^{ik_2x} + De^{-ik_2x} \tag{6}$$

where $k_2 = \sqrt{2m(E - V_0)/\hbar^2}$. The difference between k_2 and k is important, because it tells us something about how the wavelengths in Region I and III compare to the wavelength in region II. Recall that $k = 2\pi/\lambda$, where λ is the wavelength of the sine wave. Because $k_2 < k$, we should expect that $\lambda_2 > \lambda$. Therefore the wave function has a larger wavelength in Region II, so the solution looks slightly different.

In the case of $E < V_0$, the solution to the Schrödinger equation in Region II no longer looks like a sine wave. The reason for this is that because V_0 is larger than E, the only acceptable solution in that region takes the form of real exponentials. Thus, we can write the solution in Region II for $E < V_0$ as:

$$\psi_{R2}(x) = Ce^{\kappa x} + De^{-\kappa x} \tag{7}$$

where $\kappa = \sqrt{2m(V_0 - E)/\hbar^2}$. Here, we have a term with exponential decay in x added to a term with exponential increase in x.

At this point, we have solved the solutions to Schrödinger's equation in all of the three regions for the case of $E > V_0$ and the case of $E < V_0$. Here, we summarize our solutions:

[2] These two constraints are postulates of quantum mechanics; there is really no other way to explain why we use this.

For $E > V_0$:

$$\psi(x) = \begin{cases} Ae^{ikx} + Be^{-ikx}, & \text{for } x < 0 \\ Ce^{ik_2x} + De^{-ik_2x}, & \text{for } 0 \leq x \leq L \\ Fe^{ikx} + Ge^{-ikx}, & \text{for } x > L \end{cases} \tag{8}$$

For $E < V_0$:

$$\psi(x) = \begin{cases} Ae^{ikx} + Be^{-ikx}, & \text{for } x < 0 \\ Ce^{\kappa x} + De^{-\kappa x}, & \text{for } 0 \leq x \leq L \\ Fe^{ikx} + Ge^{-ikx}, & \text{for } x > L \end{cases} \tag{9}$$

As is the process with solving any differential equation, after solving for the general solutions (as we have now done), we must plug in boundary conditions to solve for the unknown quantities. In our situation, we now have 6 unknown quantities (A, B, C, D, F, and G) for both cases of E.[3] So far, we have discussed four boundary conditions: the condition that $\psi(x)$ must be continuous at both $x = 0$ and $x = L$ (this gives us two), and the fact that $\psi'(x)$ must be continuous at both $x = 0$ and $x = L$. However, we have 6 unknown quantities, so it seems we need 2 more boundary conditions.

To determine the remaining boundary conditions, we must refer to the physical situation that we are dealing with. We are interested in a particle approaching the barrier from either the left side (Region I, moving in the direction of +x) or a particle approaching from the right side (Region III, moving in the direction of -x).

In the first case, we can consider what happens to the particle as it approaches the energy barrier. Starting with the change in $V(x)$ at $x = 0$, we can have some transmission of the wave function, as well as some reflection of the wave function. Recall that the top equation in (8) and (9) describes the sum of a right-going wave (corresponding to the coefficient A) and a left-going wave (corresponding to B). By saying that the wave is coming from Region I, we are essentially saying that at $x \to -\infty$, we are setting the amplitude of A at some constant value. On the other hand, if we were talking about case 2 (the left-going wave), we would be saying that at distance $x \to +\infty$, we are fixing G to be a constant value. Therefore, by giving information about where the particle is coming from, we have provided ourselves one more boundary condition, and know everything there is to know about A or G.[4]

If we continue to consider the first case of the wave approaching from Region I, the barrier at $x = 0$ can allow for some reflection of the wave, and some transmission

[3] Note that by unknowns, we are not referring to the energy (E), mass (m), or potential ($V(x)$), nor anything that depends on those quantities (such as k, k_2, or κ). We assume that those quantities are chosen for our given physical situation, and that we now want to watch what happens.

[4] It may seem strange to say that A or G is now known, when we haven't set A or G equal to some quantity. But remember that A and G are just constants; we could arbitrarily rename them something else, but it won't make a difference in our calculation at this point.

of the wave. The transmitted wave can then continue to the barrier at $x = L$, and some of it can be transmitted and some can be reflected back.[5] For the wave that is transmitted to Region III, we can now follow the wave along its path towards $x \to +\infty$. This wave never encounters a barrier, and because of this, we can say that there is no reflected wave in this region. Since there is no reflected wave, we know that G in (8) must be 0 (remember that x and k are non-zero, so the only way to get rid of this term altogether is to eliminate G). If we were looking at the case of the wave approaching from the right, we could apply the exact same reasoning and decide that $A \to 0$, since the left-going wave in Region I never encounters a barrier to reflect off of.

Let us now summarize what the wave function should look like for these different cases:

For $E > V_0$, assume the wave is traveling towards the right. In Region I, we have a sine wave that will be both reflected and transmitted. In Region II, we have another sine wave (this time with a larger wavelength), which will be both transmitted and reflected. In Region III, we have another sine wave with the same wavelength as in Region I, but with no reflection.

For $E < V_0$, again assume the wave travels towards the right. Region I looks qualitatively the same as it did for the case of $E > V_0$. In Region II, the solution becomes a sum of real exponentials, which is dominated by the exponential decay term. Therefore, the wave function is "dying off" in Region II. This means that in Region III, though the wave function again appears sinusoidal, the amplitude in this region is smaller than in Region I.

Also note that this solution to the wave function is not physically real, since it is not normalized. That is to say that we can not integrage the probability density over all space and come up with a value that is anything but infinite. (Imagine trying to integrate a $\sin^2(x)$ function over all space. The area under the curve is infinite.) To make this physically real, we would have to use many different solutions to this problem (i.e. many energy values) so that we could sum those different solutions into a wave 'packet', which can be normalized. The reason we use this non-physical situation is because it is simpler and still gives a great deal physical intition into the nature of tunneling.

[5] It is common to worry about the reflected wave in Region II going back and transmitting back into Region I, and even worse, some of it bouncing back and forth inside the barrier until it decides to leave at some random time. It might seem like we can't say exactly where the reflected and transmitted waves are going to go for all times. Fortunately, we don't have to worry about this, since the four boundary conditions mentioned earlier take care of this automatically by forcing $\psi(x)$ and $\psi'(x)$ to be continuous at $x = 0$ and L.

References

[1] Taylor, J.R., Zafiratos, M.D., and Dubson, M.A. (2003). Modern Physics for Scientists and Engineers. Addison-Wesley. ISBN-10: 013805715X

[2] Krane, K.S. (2012). Modern Physics, 3rd Ed. Wiley. ISBN-10: 1118061144.

[3] Griffiths, D.J. (2004). Introduction to Quantum Mechanics, 2nd Ed. Pearson Prentice Hall. ISBN-10: 0131118927.