Pramāṇa – J. Phys., Vol. 25, No. 4, October 1985, pp. 431–438. © Printed in India.

Quantum mechanical tunnelling

D K ROY
Department of Physics, Indian Institute of Technology, Hauz Khas, New Delhi 110016, India

Abstract. It was shown earlier that during quantum mechanical tunnelling, a microscopic particle has a distributed probability of emission about its original energy and is not constrained to be field-emitted only at its initial energy. Such an energy distribution process appears obvious on the quantum theory of observation and measurement which relates the energy of a microscopic particle with the time required for its determination through the Heisenberg's uncertainty relation. Here, an account of the tunnelling theory based upon the latter is presented. The consequent analysis gives rise to a spectrum in the energy of the transmitted electrons and also yields a method to estimate the tunnelling time as well as the tunnelling current density across an arbitrary barrier.

Keywords. Quantum mechanical tunnelling; tunnelling theory.

1. Introduction

The foremost experimental evidence of quantum mechanical tunnelling may probably be traced back to the work of Lilienfeld (1922). This stimulated the early physicists to understand the effect assuming that the particle energy must be conserved or would remain exactly the same throughout the process. This idea was subsequently employed by Fowler and Nordheim (1928), Gamow (1928) and Zener (1934) to their respective problems and favourable agreements of some of their predictions with experimental results were reported subsequently. But the experimental observation of the finite limit of resolution of the field-emission microscope (FEM) introduced by Muller (1937) could not be explained on the above premise. The field-ionization of hydrogen atoms was however explained differently by Oppenheimer (1928) by presuming that the tunnelling electrons suffer a time-dependent perturbation on interacting with the barrier.

The physical concepts of the phenomenon became all the more hazy with the discovery of Esaki (1958) and Josephson (1962) effects, as none of them could adequately be explained on the prevailing notion of the tunnelling theory. Esaki's expression (1958) for tunnel diodes could be explained only qualitatively on the perturbation treatment of Bardeen (1961) which was a refinement over the ideas of Oppenheimer (1928). Cohen *et al* (1962) subsequently improved upon this model on the basis of second quantization later to be employed by Josephson (1962) to predict his effects.

The working models of the tunnelling problem which we have referred to above are based upon (i) a time-independent approach relying on the principle of conservation of energy and (ii) a time-dependent treatment depending upon the perturbation theory. The predictions of these two seemingly independent approaches were found to agree reasonably well because in the perturbation treatment a negligible spread in the electron energy is only allowed and when this is very small compared to the original electron

energy, the latter would eventually reduce to the former. Naturally therefore it had been a matter of convenience for one to apply any of the above formalisms to study a particular problem on tunnelling.

It is however needless to elaborate that none of the above formalisms is ideally suited to analyze it in its true perspective. On the principle of quantum measurement the energy of a tunnelling electron can never be estimated unless the process is complete and this requires the lapse of a definite interval of time τ. Consequently, the particle energy at the conclusion of the event cannot continue to remain exactly the same as before but would be spread over an energy interval $\hbar/\tau$. Incidentally, this uncertainty turns out of the order of the barrier height (Roy 1977). Therefore, the ideas of insignificant or no spread in the energy of the tunnelling electrons as mooted earlier do not seem to be tenable. This was noticed by Roy (1970) who also suggested that owing to reasonably larger barrier heights, the energy distribution amongst tunnelling electrons cannot be ignored. Here, we present a latest account of our analysis.

2. Tunnelling theory

The quantum theory of observation and measurement (Scheibe 1973) as contained in the principle of uncertainty of Heisenberg states that a quantum object (e.g. an electron) whose property is to be estimated (or measured) and the agency of observation (e.g. the apparatus) must be treated as an interacting dynamical system. During classical measurements such an interaction is either neglected or compensated for. But while making quantum measurements, the interaction of the object with the apparatus is finite, inevitable, not negligible and not separately accountable owing to the finiteness of the Planck's quantum of action h. Again unlike classical measurements, the surveillance of the quantum measurement process (i.e. determining the energy and time or momentum and position simultaneously and exactly) is impossible. This unavoidable interaction therefore sets an upper limit to the feasibility of expressing the behaviours of atomic objects independent of the means of observation. It should be emphasized here that such an inability on the part of the observer is neither due to faulty apparatus used nor due to any particular observation procedure employed but has been conditioned by the unavoidable interaction between the object and the apparatus as embodied in the basic postulates of wave mechanics. Therefore, any denial of the applicability of the principle of uncertainty to analyze the tunnelling problem is equivalent to rejecting the quantum theory. Consequently, any description of quantum phenomena based on measurements would never be complete in the classical sense because we would then have complete failure in the co-ordination between space and time in our descriptions and the breakdown of the dynamical conservation laws which are so well known for their pictorial and definite predictions during classical measurements.

Now, if the problem of quantum mechanical tunnelling is treated as one of energy measurement of an electron by a potential barrier, one must consider the interaction between the two during the process. Consequently, the electron potential energy during tunnelling must be expressed as (Roy *et al* 1977),

$$V(x, t) = V_1(x) + V_2(t), \tag{1}$$

where $V_1(x)$ denotes the exact potential energy at any point of the barrier while $V_2(t)$ incorporates the effect of the interaction between the barrier and the electron. If the

electron is lodged permanently within the barrier, the interaction term $V_2(t)$ then reduces to zero because infinite time then becomes available for potential energy measurement. The two terms appear in (1) because of the impossibility of exactly measuring the electron potential energy and its transit time simultaneously during tunnelling. The electron motion during the process would however be represented by,

$$H\psi(x, t) = i\hbar \frac{\partial}{\partial t}\psi(x, t) \tag{2}$$

where

$$H = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V(x, t) \tag{3}$$

on regarding tunnelling as one dimensional process. On combining (1), (2) and (3), we get,

$$\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + V_1(x)\right]\psi(x, t) = \left[i\hbar\frac{\partial}{\partial t} - V_2(t)\right]\psi(x, t). \tag{4}$$

Regarding $\psi(x, t)$ to be separable as,

$$\psi(x, t) = X(x)T(t) \tag{5}$$

equation (4) further splits into the following two differential equations:

$$\left[-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V_1(x)\right]X(x) = E_l X(x), \tag{6}$$

and

$$i\hbar\frac{\mathrm{d}}{\mathrm{d}t}T(t) = [E_l + V_2(t)]T(t). \tag{7}$$

The separation constant in (4) turns out to be the total energy E_l of the incident electrons. The solutions belonging to (6) may be represented by $X_l(x)$ and $X_r(x)$. Their exact forms would however depend upon the shape of the barrier. For any arbitrary barrier, one may always obtain two functions $F(x)$ and $G(x)$ from (6), such that,

$$X_l(x) = \alpha F(x) \quad \text{and} \quad X_r(x) = \beta G(x), \tag{8}$$

where α and β are constants to be obtained by matching $F(x)$ with the incident wave and $G(x)$ with the transmitted wave. The situation corresponding to the characteristic decaying and the growing wave forms, if desired, may be obtained by appropriately selecting linear combinations of $F(x)$ and $G(x)$. The solution of (7) on the other hand yields,

$$T(t) = \gamma \exp\left[-\frac{iE_l t}{\hbar} - \frac{i}{\hbar}\int_0^t V_2(t)\mathrm{d}t\right]. \tag{9}$$

The most general solution of the wavefunction may then be obtained by combining (8) and (9) as follows:

$$\psi(x, t) = a_l(t)X_l(x)\exp\left\{-\frac{iE_l t}{\hbar}\right\} + a_r(t)X_r(x)\exp\left\{-\frac{iE_r t}{\hbar}\right\}, \tag{10}$$

where

$$a_l(t) = \gamma_1 \exp\left[-\frac{i}{\hbar}\int_0^t V_2(t)\,\mathrm{d}t\right], \tag{11}$$

and

$$a_r(t) = \gamma_2 \exp\left[-\frac{i}{\hbar}\int_0^t V_2(t)\,\mathrm{d}t\right], \tag{12}$$

$a_l(t)$, $a_r(t)$, γ_1 and γ_2 being arbitrary coefficients. It should be noted that as the two parts of the wavefunction in (10) are being observed at the same instant of time t, their energies have been designated differently as prescribed by the theory of quantum measurement. Next, (10) is substituted in (2) when one gets,

$$\begin{aligned} &i\hbar\dot{a}_l(t)X_l(x) + i\hbar\dot{a}_r(t)X_r(x)\exp(i\omega_{lr}t) \\ &\quad = a_l(t)V_2(t)X_l(x) + a_r(t)V_2(t)X_r(x)\exp(i\omega_{lr}t), \end{aligned} \tag{13}$$

where $$\omega_{lr} = \omega_l - \omega_r = (E_l - E_r)/\hbar. \tag{14}$$

By inspection of (10), one may regard tunnelling as being caused by the gradual transition of electrons from X_l to X_r in response to the time-dependent potential $V_2(t)$. Next in order to evaluate a_l and a_r (13) is multiplied by $X_r^*\,\mathrm{d}x$ and $X_l^*\,\mathrm{d}x$ in succession and integrated between barrier extremities when one gets,

$$i\hbar\dot{a}_l R_{rl} + i\hbar\dot{a}_r R_{rr}\exp(i\omega_{lr}t) = a_l T_{lr} + a_r T_{rr}\exp(i\omega_{lr}t), \tag{15}$$

and

$$i\hbar\dot{a}_l R_{ll} + i\hbar\dot{a}_r R_{lr}\exp(i\omega_{lr}t) = a_l T_{ll} + a_r T_{rl}\exp(i\omega_{lr}t), \tag{16}$$

where

$$R_{lr} = \int_{x_l}^{x_r} X_l^* X_r\,\mathrm{d}x, \quad \text{etc.}, \tag{17}$$

and

$$T_{lr} = \int_{x_l}^{x_r} X_r^* V_2 X_l\,\mathrm{d}x, \quad \text{etc.} \tag{18}$$

Now, on ignoring terms representing self-state modulations (e.g. $T_{ll} = 0 = T_{rr}$) and reverse transition ($T_{rl} = 0$), (15) and (16) further simplify to,

$$i\hbar\dot{a}_l R_{rl} + i\hbar\dot{a}_r R_{rr}\exp(i\omega_{lr}t) = a_l T_{lr} \tag{19}$$

and

$$i\hbar\dot{a}_l R_{ll} + i\hbar\dot{a}_r R_{lr}\exp(i\omega_{lr}t) = 0. \tag{20}$$

From (20), it immediately follows that,

$$\dot{a}_r = -\frac{R_{ll}}{R_{lr}}\dot{a}_l\exp(-i\omega_{lr}t). \tag{21}$$

On substituting for $\dot{a}_r$ in (19), we get,

$$\dot{a}_l(t) = -i\omega_0 a_l(t), \tag{22}$$

where

$$\omega_0 = \frac{R_{lr}T_{lr}}{\hbar[R_{ll}R_{rr} - R_{lr}R_{rl}]}. \tag{23}$$

Now, on integrating (22) with $a_l(0) = 1$, one obtains,

$$a_l(t) = \exp(i\omega_0 t). \tag{24}$$

Next, on combining (21) and (22), one finds,

$$\dot{a}_r(t) = -i\omega_0 \frac{R_{ll}}{R_{lr}} \cdot \exp(-i\omega t), \tag{25}$$

where $$\omega = \omega_{lr} - \omega_0 = \omega_l - \omega_r - \omega_0. \tag{26}$$

Integration of (25) then yields,

$$a_r(t) = -i\omega_0 \frac{R_{ll}}{R_{lr}} \cdot \exp(-i\omega t/2) \frac{\sin(\omega t/2)}{(\omega/2)}, \tag{27}$$

where $a_r(0) = 0$ has been presumed. With a_l and a_r as given by (24) and (27), the barrier wavefunction as expressed by (10) is now fully determined.

The electron tunnelling probability as per its definition now may be written as,

$$P(x, t) = \psi^*(x, t)\psi(x, t) = X + Y + Z \text{ (say)} \tag{28}$$

where

$$X = |a_l(t)|^2 \cdot |X_l(x)|^2 = |X_l(x)|^2, \tag{29}$$

$$Y = |a_r(t)|^2 \cdot |X_r(x)|^2$$

$$= \frac{\omega_0^2 |R_{ll}|^2}{|R_{lr}|^2} |X_r|^2 \cdot \frac{\sin^2(\omega t/2)}{(\omega/2)^2}, \tag{30}$$

and

$$Z = a_l^* a_r X_l^* X_r \exp(i\omega_{lr} t) + a_l a_r^* X_l X_r^* \exp(-i\omega_{lr} t) \tag{31'}$$

$$= \frac{2\omega_0 R_{ll} X_l^* X_r}{R_{lr}} \cdot \frac{\sin(\omega t/2)}{(\omega/2)} \cdot \sin\left(\frac{\omega t}{2} + \phi\right), \tag{31}$$

where ϕ measures any additional phase difference introduced between a_l and a_r due to effects other than time. On combining (28) to (31), the tunnelling probability expression across a rectangular barrier follows to be,

$$P(x, t) = |\alpha|^2 \exp(-2\chi_2 x) + |\beta|^2 \cdot \exp(+2\chi_2 x)$$

$$\times \frac{\omega_0^2 |R_{ll}|^2}{|R_{lr}|^2} \frac{\sin^2(\omega t/2)}{(\omega/2)^2} + \frac{2\omega_0 R_{ll} X_l^* X_r}{R_{lr}}$$

$$\times \frac{\sin(\omega t/2)}{(\omega/2)} \cdot \sin\left(\frac{\omega t}{2} + \phi\right), \tag{32}$$

where

$$\frac{|\beta|^2 \cdot \omega_0^2 \cdot |R_{ll}|^2}{|R_{lr}|^2} = \frac{4|\alpha|^2 \hbar^2 \chi_2^4}{m^2} \exp(-4\chi_2 \omega). \tag{33}$$

Next, on differentiating $P(x, t)$ with respect to x at $x = \omega$ and for $\omega_{lr} = 0$ (so that the variation in time at a definite energy may be recorded), we get,

$$\frac{\partial}{\partial x}[P(x, t)]_{x=\omega} = -2|\alpha|^2 \chi_2 \exp(-2\chi_2 w)$$

$$\left[1 - \frac{4\hbar^2 \chi_2^4}{m^2} \frac{\sin^2(\omega_0 t/2)}{(\omega_0/2)^2}\right]. \tag{34}$$

Equation (34) thus suggests that the flow of the probability current would stop when the right side of (34) reduces to zero implying,

$$\tau = \frac{2}{\omega_0} \sin^{-1}\left[\frac{m\omega_0}{4\hbar\chi_2^2}\right]. \tag{35}$$

If the argument of the inverse sine function is small, (35) then reduces to,

$$\tau = \frac{m}{2\hbar\chi_2^2} = \frac{\hbar}{4(V_0 - E_l)}, \tag{36}$$

where V_0 measures the height of the rectangular barrier.

The resulting barrier current density may however be evaluated by using the equation of continuity (Sai 1984) as follows:

$$J_t = q\int_{x_l}^{x_r} \dot{P}(x,\tau)\,dx. \tag{37}$$

On combining (28) with (37) and upon simplification one obtains,

$$J_t = J_{01}\frac{\sin\gamma}{\gamma} + J_{02}\sin(\gamma+\phi), \tag{38}$$

where

$$J_{01} = \frac{2R_{rr}|R_{ll}|^2\omega_0^2\tau q}{|R_{lr}|^2}, \tag{39}$$

and

$$J_{02} = 2qR_{ll}\omega_0, \tag{40}$$

with $$\gamma = \omega\tau = (\omega_l - \omega_r - \omega_0)\tau = \omega_l(t - \tau/2) \text{ (say)} \tag{41}$$

since time may be measured from any instant chosen conveniently. Equation (38) thus measures the one-electron current density across a potential barrier. It follows from (38) that even the state corresponding to the incident electron is not the most probable one for electron emission.

However, if the incident particle stream is a coherent one with all its constitutents practically belonging to the same phase γ (i.e. they correspond to the same energy E_l but with a negligible spread $\Delta E_l = \hbar\Delta\omega_l$), the form of the net tunnelling current density produced by it would still remain the same as (38), but its magnitude would be obtained on multiplying the latter with the density of the particles. Such a situation is realized in practice when a potential barrier is sandwiched by a pair of identical superconducting electrodes carrying a current due to Cooper particles. The second term in (38) would then account for the well-known Josephson effect (Roy and Sai 1982).

On the other hand, if the incident particles are all incoherent (i.e. they possess random phase differences amongst themselves) but possess the same energy E_l, the net tunnelling current density generated by them may be obtained by summing (38) with due regards to their phase differences. As t of (41) may then lie anywhere between $-\infty$ to $+\infty$, the first term of (38) is then only significant for $0 \leqslant t \leqslant \tau$ (i.e. only for the duration of τ) while the second term adds upto zero. Hence, in this case, the tunnelling current is expected to last only for a definite period equal to the tunnelling time. Such a situation prevails, in practice, when a potential barrier is sandwiched by metallic and/or semiconducting electrodes. Thus, the net differential current density produced by a group of electrons lying within an energy interval dE_l while moving from left to right

would be (Roy *et al* 1982),

$$\mathrm{d}J(E_l) = \rho_l(E_l) f_l(E_l)\,\mathrm{d}E_l \sum_{\gamma=-\infty}^{+\infty} \left[J_{01} \frac{\sin\gamma}{\gamma} + J_{02}\sin(\gamma+\phi) \right] \tag{42}$$

where $\rho_l(E_l) f_l(E_l)\,\mathrm{d}E_l$ measures the density of incident electrons in the left hand conducting system. To convert (42) into an appropriate integral, we note from (41) that,

$$|\mathrm{d}\gamma| = \frac{\tau\,\mathrm{d}E_r}{\hbar} = \frac{\varepsilon_r \tau}{\hbar} \text{ (say)}, \tag{43}$$

where ε_r measures the energy of separation between consecutive levels at the transmitted end. Hence on multiplying (42) by $\hbar\,\mathrm{d}\gamma/\varepsilon_r\tau$ (= 1), we obtain,

$$\mathrm{d}J(E_l) = \frac{\hbar}{\varepsilon_r \tau} \rho_l(E_l) f_l(E_l)\,\mathrm{d}E_l \int_{-\infty}^{+\infty} \left[J_{01} \frac{\sin\gamma}{\gamma} + J_{02}\sin(\gamma+\phi) \right] \mathrm{d}\gamma$$

or

$$\mathrm{d}J(E_l) = \frac{\pi\hbar}{\varepsilon_r \tau} J_{01}\, \rho_l(E_l) f_l(E_l)\,\mathrm{d}E_l, \tag{44}$$

because the second term vanishes upon integration. But while deriving (44) it has explicitly been presumed that an infinite energy continuum of equispaced levels is available at the transmitted end for electrons to tunnel. But in semiconductor devices that is seldom the case. To overcome such a difficulty in the above computation, we may consider first the tunnelling occurring at a particular energy level E_l at the transmitted end due to carriers incident at different instants of time. Such a procedure would still account for the same range of γ from $-\infty$ to $+\infty$ as it was required while obtaining (44). We may however write for ε_r from its definition as,

$$\varepsilon_r = \frac{1}{\Omega_r \rho_r(E_l)\{1 - f_r(E_l)\}}, \tag{45}$$

where Ω_r is the volume of solid electrode at the right hand end. On incorporating (45) in (44), we get,

$$\mathrm{d}J(E_l)_{l\to r} = \frac{\pi\hbar\Omega_r}{\tau} J_{01}\, \rho_l(E_l) f_l(E_l) \rho_r(E_l)\{1 - f_r(E_l)\}\,\mathrm{d}E_l. \tag{46}$$

The reverse tunnelling current density produced at the same energy may then be obtained similarly as,

$$\mathrm{d}J(E_l)_{r\to l} = \frac{\pi\hbar\Omega_l}{\tau} J_{01}\, \rho_r(E_l) f_r(E_l) \rho_l(E_l)\{1 - f_l(E_l)\}\,\mathrm{d}E_l, \tag{47}$$

where Ω_l now measures the volume of the left hand electrode. If $\Omega_l = \Omega_r = \Omega$ the net differential current density from left to right would be,

$$\mathrm{d}J(E_l) = \frac{\pi\hbar\Omega}{\tau} J_{01}[f_l(E_l) - f_r(E_l)]\rho_l(E_l)\rho_r(E_l)\,\mathrm{d}E_l. \tag{48}$$

On integrating (48) over appropriate limits of E_l, one would obtain the required expression for the I – V characteristic of the semiconductor device in question. It may be interesting to note the similarity of (48) with the Esaki integral for tunnel diodes (Esaki 1958).

References

Bardeen J 1961 *Phys. Rev. Lett.* **6** 57
Cohen M H, Falicov L M and Phillips J C 1962 *Phys. Rev. Lett.* **8** 316
Esaki L 1958 *Phys. Rev.* **109** 603
Fowler R and Nordheim L 1928 *Proc. R. Soc.* (*London*) **A119** 173
Gamow G 1928 *Z. Phys.* **51** 204
Josephson B D 1962 *Phys. Lett.* **1** 251
Lilienfeld J E 1922 *Z. Phys.* **23** 505
Muller E W 1937 *Z. Phys.* **106** 541
Oppenheimer J R 1928 *Phys. Rev.* **31** 66
Roy D K 1977 *Tunnelling and negative resistance phenomena in semiconductors* (Oxford: Pergamon)
Roy D K 1970 *Phys. Status Solidi.* **A2** K241
Roy D K and Sai N S T 1982 *Indian J. Pure. Appl. Phys.* **20** 300
Roy P N, Singh P N and Roy D K 1977 *Phys. Lett.* **A63** 81
Roy D K, Sai N S T and Rai K N 1982 *Pramana* **19** 231
Sai N S T 1984 *Quantum mechanical tunnelling in thin solid junctions*, Ph.D. Thesis, I.I.T., Delhi
Scheibe E 1973 *The logical analysis of quantum mechanics* (Oxford: Pergamon)
Zener C 1934 *Proc. R. Soc.* (London) **A145** 523