Theory of Quantum Conduction through a Constriction

Aaron Szafer and A. Douglas Stone
Applied Physics, Yale University, P.O. Box 2157, New Haven, Connecticut 06520
(Received 6 October 1988)

A theory of ballistic conduction through a constriction is developed which predicts a steplike structure with amplitude e^2/h in the conductance as a function of Fermi energy or width. For a sharp (abrupt) constriction the steps are only quantized to 1 part in 10^3 at $T=0$. Disorder destroys the steps and they are modulated by novel resonant structure in certain regimes of shape and temperature.

PACS numbers: 73.20.Dx, 72.20.Dp, 73.40.Lq

Recent experiments[1,2] have observed a striking new effect when measuring the low-temperature conductance of a high-mobility (ballistic) two-dimensional electron system separated into two regions by a gate which creates a constriction of variable width. The conductance as a function of the width of the constriction was found to rise in a long series of rather sharp steps of magnitude e^2/h (per spin). This effect is of general interest for at least two reasons. First, it is reminiscent of the normal quantum Hall effect (but note it occurs even in the *absence* of a B field), thus it naturally raises the question of how precise the "quantization" of the conductance is in the plateau regions. Second, the effect sheds light on questions concerning the physically relevant version of the Landauer formula (relating conductance to the scattering matrix of the disordered conductor) for *two-probe* measurements, and it supports the view[3] that in this case the dimensionless conductance, $g=G/(e^2/h)$, is best described by the formula[4]

$$g=\mathrm{Tr}(\mathbf{t}\mathbf{t}^\dagger)\,, \tag{1}$$

where $\mathbf{t}$ is the transmission matrix through the "sample," evaluated at the Fermi energy.

Although Eq. (1) predicts a finite resistance for a "perfect" conductor, it is now understood that if an experiment measures the ratio of the induced current to *the chemical potential difference between the "reservoirs" serving as current source and current sink,* then Eq. (1) applies.[3] The resistance when the sample is perfectly transmitting can be regarded as an ideal contact resistance occurring at the interface between the sample and the reservoirs, as first pointed out by Imry.[5] In the simplest model, discussed in Ref. 1, $\mathbf{t}$ refers to the transmission matrix of the constriction alone; in the ballistic limit $(\mathbf{t}\mathbf{t}^\dagger)_{ij}=\delta_{ij}$, and so g is equal to the number of propagating channels in the constriction, and increases by unity in perfectly sharp steps each time a new channel opens up. This argument, while useful in understanding the origin of the effect, is inadequate to describe the observed shape of the steps which are *not* perfectly sharp. Moreover, since the real system is phase coherent on a scale much longer than the constriction, to understand the observability of the effect *under experimental conditions,* one must treat correctly the impedance-matching and mode-conversion behaviors at the interfaces between the constriction and 2D regions. In the "adiabatic" limit of an arbitrarily smoothly tapered constriction these effects are almost negligible,[6] and a recent WKB treatment finds only exponentially small corrections to θ-function steps, and no dependence on the length of the constriction.[7] We focus on the limit in which these effects are *most* important, an abrupt interface between wide and narrow regions, and find the striking result that even here one obtains rather sharp steps.

We study a model (equivalent to a scalar waveguide problem) consisting of two wide regions of width W, separated by a constriction of width $W' \ll W$ and length L (the sample geometry is defined by hard-wall bound-

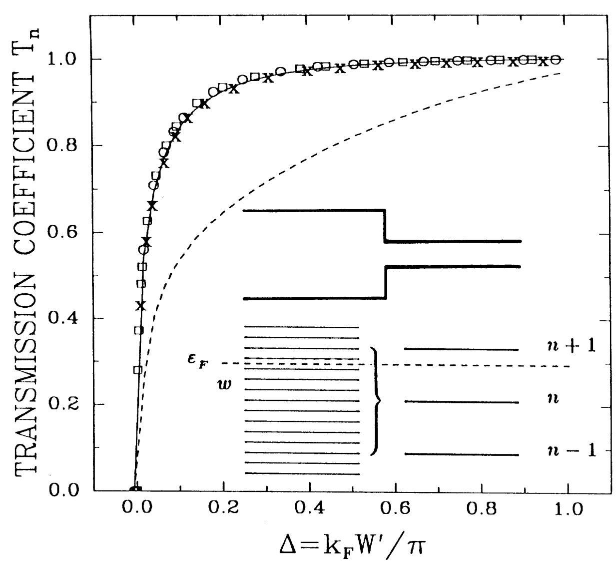

FIG. 1. $T_n(\Delta)$ for wide-narrow (WN) geometry. Solid line is T_3 in the mean-field approximation (MFA) [Eq. (3)], squares are $T(\Delta)$ [Eq. (4)]. Circles and crosses are numerical results for T_3 and T_4, demonstrating scaling property of exact $T_n(\Delta)$. Dashed line is 1D model explained in text. Inset: Schematic of the WN geometry and the threshold energies for the corresponding regions; brace indicates the channels w contributing to T_n in the MFA.

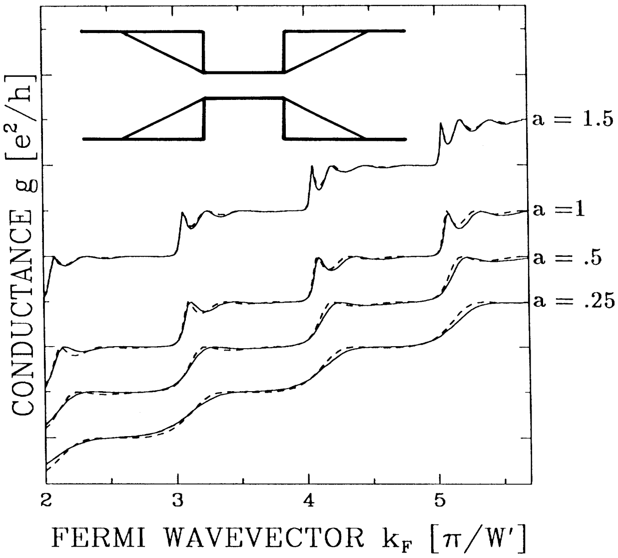

FIG. 2. $g(\epsilon_F)$ for aspect ratios $a \equiv L/W'$ ($W'=12$ lattice units). Solid lines are exact numerical results for abrupt transition, and dashed lines in the three lower curves are MFA results. Dashed line in top curve corresponds to tapering (to $W=4W'$) as shown in the inset. Curves are vertically offset.

ary conditions). Such a linear problem can always be solved exactly by discretization. A particularly flexible technique that we have employed is the recursive Green's function method[8] which allows us to solve for an ordered or disordered constriction of any shape, although initially we will focus on solving the model for the case of an abrupt constriction (Figs. 1 and 2). While the true situation is somewhere in between the adiabatic and abrupt limits, this model (which assumes that the "walls" created by the gate are relatively well defined) is consistent with the approximately equal spacing of the observed steps as a function of W', and the recently observed appearance of resonant structure in the steps.[9]

To understand the mode conversion that occurs at the orifice, we begin by considering the simpler problem of transmission from a wide semi-infinite waveguide (width W), where the transverse modes are dense in energy, abruptly connected to a narrow semi-infinite waveguide (width $W' \ll W$), where they are very sparse in energy (Fig. 1). Crudely one can understand the transmission in this wide-narrow (WN) geometry by analogy to above-barrier transmission in one dimension: As a mode n (with threshold energy ϵ_n) of the narrow region passes through ϵ_F, typically a mode w in the wide region must give up longitudinal energy to propagate through the constriction. When mode n is barely propagating ($\epsilon_F \approx \epsilon_n$) the longitudinal kinetic energy is very small compared to the "barrier" (the mismatch of transverse energies) and there is substantial reflection, whereas when $\epsilon_F \gg \epsilon_n$ the barrier is negligible and the transmission is maximal. More precisely, the unitarity relations and the continuity of wave functions with energy imply the following: (a) The total transmission from all the modes w *into* a given mode n is $T_n = \sum_w |t_{nw}|^2 = \sum_w |t_{wn}|^2 \le 1$. (b) When $\epsilon_F - \epsilon_n \to 0$, then $|r_{nn}|=1$ and $r_{mn} = t_{wn} = 0$ for all w and m, i.e., at threshold there is no transmission or *off-diagonal* reflection. Threshold for propagation of mode n can be crossed by increasing either ϵ_F or W' so that $\epsilon_n(W') \equiv q_n^2 \equiv (n\pi/W')^2$ falls below ϵ_F; thus it is natural to study T_n as a function of the variable $\Delta \equiv (k_F - q_n)W'/\pi \equiv \sqrt{\epsilon_F}W'/\pi - n$, which can be regarded as measuring either the deviation of k_F or W' from threshold, in units of π/W', or of $\lambda_F/2$, respectively. $T_n(\Delta)$ must rise to nearly unity for $\Delta \ll 1$ to obtain sharp steps in g.

To solve for T_n we write a wave function corresponding to a particle with energy $\epsilon_F = q_w^2 + k_w^2$ incident in mode w from the wide side at $x \le 0$ as

$$\Psi_w(x,y) = \chi_w(y)\exp[ik_w x] + \sum_{v=1}^{\infty} \bar{r}_{vw}\chi_v(y)\exp[ik_v x]$$

if $x \le 0$,

$$\Psi_w(x,y) = \theta(W'/2 - |y|)\sum_{n=1}^{\infty} \bar{t}_{nw}\varphi_n(y)\exp[ik_n x]$$

if $x \ge 0$, and match at the interface ($x=0$). The resulting equations can be combined to give

$$\sum_{m=1}^{\infty} A_{nm}\bar{t}_{mw} + k_n \bar{t}_{nw} = 2k_w a_{nw}\,, \tag{2}$$

where $a_{nw} \equiv \int_{-W'/2}^{W'/2} \chi_w(y)\varphi_n(y)dy$ is the overlap of the transverse wave functions. For propagating w and n, $\bar{t}_{nw}$ is related to the unitary transmission coefficient by $t_{nw} = \bar{t}_{nw}(k_n/k_w)^{1/2}$.[3,4] We see that $\bar{t}_{nw}$ is coupled to all other $\bar{t}_{mw}$ by the kernel $A_{nm} \equiv \sum_{v=1}^{\infty} a_{nv}k_v a_{mv}$. We have solved these equations exactly numerically (Fig. 1), and they indeed give a sharp series of steps in $T_n(\Delta)$. However, physical insight and an extremely good analytic approximation for T_n can be obtained by a "mean-field approximation" (MFA), valid for $W \gg W'$.

The MFA is motivated by noting that for fixed n, a_{nw}^2 is "peaked" at w such that its transverse wave vector satisfies $q_w \simeq q_n$. Since A_{nm} involves products of two overlaps peaked at channels n and m, the coupling is small for $n \ne m$. We approximate the true overlaps by a uniform coupling to all modes within one level spacing of ϵ_n,

$$a_{nw}^2 = (W'/W)[\theta(q_w - q_{n-1}) - \theta(q_{n+1} - q_w)]\,,$$

which preserves the completeness relation $\sum_w a_{nw}a_{mw} = \delta_{nm}$. We noted above that $r_{nm}=0$ at $\epsilon_F = \epsilon_n$, and when $\epsilon_F \gg \epsilon_n$, it is easy to show that the MFA is equivalent to setting $r_{nm}=0$ at all energies.

The MFA implies

$$A_{nm} = \delta_{nm}(W'/W)\sum_v^{(n)} k_v \equiv \delta_{nm}(K_n + iJ_n) = \delta_{nm}(W'\epsilon_F/8\pi)[(\sin 2\theta_{n+1} - 2\theta_{n+1}) - (\sin 2\theta_{n-1} - 2\theta_{n-1})]\,,$$

where the last equation is obtained by evaluating the relevant integral. The superscript n indicates that only modes with $q_{n-1} < q_v < q_{n+1}$ are summed, $\cos\theta_n \equiv q_n/k_F$, $\sin\theta_n \equiv k_n/k_F$, and $\mathrm{Im}(\theta_{n+1}) > 0$, when $\epsilon_F < \epsilon_{n+1}$. Thus Eq. (2) decouples giving $g_{\mathrm{WN}} = \sum_{\epsilon_n \le \epsilon_F} T_n$, where

$$T_n = 4K_n k_n / [(K_n + k_n)^2 + J_n^2] . \tag{3}$$

Note that T_n has the form of the above-barrier transmission coefficient for a semi-infinite 1D potential step, except that the wave vector K of the particle before it reaches the step is replaced by a complex wave vector whose real part K_n is an average longitudinal wave vector for all the propagating modes w with $q_{n-1} < q_w \le k_F$, and whose imaginary part J_n is an average of the decaying modes w with $k_F < q_w < q_{n+1}$.

The MFA suggests a remarkable scaling property of the solution to Eq. (2). Since T_n depends only on the modes w in a limited region around threshold, it is reasonable to conjecture that $T_n(\Delta) = T(\Delta)$, *independent of n*. Expanding $T_n(\Delta)$ from Eq. (3) in $1/n$ one indeed finds the asymptotic result

$$T_n(\Delta) \approx T(\Delta) = \frac{12\sqrt{\Delta}(1+\Delta)^{3/2}}{[(1+\Delta)^{3/2} + 3\sqrt{\Delta}]^2 + (1-\Delta)^3} . \tag{4}$$

The results given in Fig. 1 show that both Eqs. (3) and (4) agree very well with the exact solution even for small n (Fig. 1). In addition they show that the exact solution obeys the conjectured scaling relation. We note that g becomes independent of W in the relevant regime $W \gg W'$, both in the exact and MFA solutions.

Summarizing the results for the WN geometry we find that because of (a) the impedance mismatch at threshold, the steps are not arbitrarily sharp at $T=0$. This is expected but was obscured in the discussions in Refs. 1 and 2. The steps are, however, much *sharper* than would be obtained by extrapolation of the 1D behavior, $|t_{nw}|^2 = a_{nw}^2 4k_n k_w/(k_n + k_w)^2$ (Fig. 1). (b) The MFA shows that very many modes (a number of order W/W') in the wide region couple equally to a given mode n; thus each has only a small transmission probability $T_{nw} \sim W'/W$. (c) All steps have the same shape and accuracy of quantization when measured in terms of Δ.

We now consider the experimental wide-narrow-wide (WNW) geometry, in which two new effects occur. First, transmission through the constriction can occur via an evanescent mode $n+1$; such a contribution will decrease exponentially with increasing length of the constriction, becoming negligible when

$$\kappa_{n+1}(\epsilon_F = \epsilon_n) L = (\epsilon_{n+1} - \epsilon_n)^{1/2} L > 1 .$$

Second, resonant structure in g should appear, as a result of alternatively constructive and destructive internal reflection within the constriction. This effect becomes important when

$$k_n(\epsilon_F = \epsilon_{n+1}) L = (\epsilon_{n+1} - \epsilon_n)^{1/2} L > 1 .$$

Thus we see that as soon as the constriction becomes long enough to damp out the evanescent modes, resonances should appear. The occurrence of either effect depends on shape L/W' (for fixed ϵ_F), or $k_F L$ (for fixed W') (Fig. 2).

For the WNW case, the MFA again gives a result similar in form to that for the analogous 1D problem of transmission over a finite potential step. We find $g_{\mathrm{WNW}} = \sum_{\epsilon_{n-1} < \epsilon_F} g_n$, where g_n is associated with the transmission through channel n and is given by

$$g_n = \begin{cases} \dfrac{4K_n\kappa_n}{4K_n\kappa_n + [K_n^2 + (J_n+\kappa_n)^2][K_n^2 + (J_n-\kappa_n)^2]\sinh^2(\kappa_n L + \phi_n)} & \text{if } \epsilon_n > \epsilon_F , \\[2ex] \dfrac{4K_n k_n}{4K_n k_n + [(K_n+k_n)^2 + J_n^2][(K_n-k_n)^2 + J_n^2]\sin^2(k_n L + \phi_n)} & \text{if } \epsilon_n < \epsilon_F , \end{cases} \tag{5}$$

where $\kappa_n = (\epsilon_n - \epsilon_F)^{1/2}$ is the magnitude of the imaginary wave vector of the lowest evanescent mode in the constriction, and the angle ϕ_n is defined by

$$\phi_n = \tanh^{-1}[2J_n\kappa_n/(K_n^2 + J_n^2 + \kappa_n^2)]$$

for that mode and $\phi_n = \tan^{-1}[2J_n k_n/(K_n^2 + J_n^2 - k_n^2)]$ for the propagating modes. The MFA describes g_{WNW} to about (5–10)% accuracy (Figs. 2 and 3) with negligible computational time whereas the corresponding exact calculations require many central-processing-unit hours (IBM 3081).

Note that the resonances have systematic features: It follows from Eq. (5) that they never bring g_n above unity, and the first resonances are always the narrowest. A potential difference between the band bottom within and outside the constriction is known to exist in the experimental system, and will enhance reflection and emphasize the resonances.[10] Conversely, in the experimental system the confining potential changes continuously and so the internal reflection at the ends of the constriction should be smaller than in the sharp geometry we consider. We have studied numerically the effect of tapering the constriction and find that the resonances do not disappear rapidly (Fig. 2), and so we expect some resonant effects to occur in experiments. However, even geometries that lead to large resonances at $T=0$ give smooth steps at higher temperatures [see Fig. 3 where we have used $g(T) = -\int d\epsilon f'(\epsilon,T) g(\epsilon)$], since the derivative of the Fermi function f' averages $g(\epsilon_F)$ over an energy range of order $4k_B T$. It is important to note here the relationship between the results for the WNW

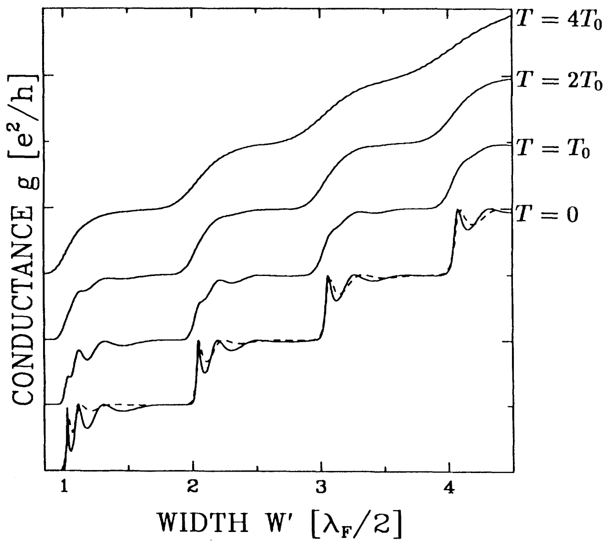

FIG. 3. $g(W')$ for different temperatures, computed as discussed in text, with use of the MFA (solid lines). Dashed line is exact numerical result. $T_0 = 0.02\epsilon_F$. For experimental parameters of Ref. 1, $T_0 \approx 2.8$ K. Curves are vertically offset.

and WN geometries. (a) The WN result, Eq. (4), indicates that the accuracy of quantization of the steps for a sharp geometry is bounded by $T(\Delta=1) \simeq 0.9991$, even disregarding the effects of resonances and evanescent modes. (b) For long constrictions, a small temperature will average the very narrow and dense resonances giving steps precisely described by Eq. (4).

Another possible origin for structure in the steps is disorder. When the elastic and inelastic mean free paths l and L_{in} satisfy $l \ll L < L_{in}$, one expects (and we have verified numerically) that both the steps and the regular structure due to resonances are replaced by fluctuations in g of order e^2/h.[10,11] This emphasizes that the effect is very different from the quantum Hall effect, which is rather insensitive to disorder. The interesting regime when $L \ll l \ll L_{in}$ will be discussed in detail elsewhere.[10] Briefly, we find that there are still fluctuations in g of order e^2/h, but with an amplitude substantially reduced from the analogous homogeneous system. This indicates that such an effect of disorder can appear as temperature is reduced, but is likely to be smaller than the resonant structure.

In summary, the detailed "line shape" of the steps in the abrupt limit can be understood by a simple model which is consistent with experiments.[1,2,9] The existence of sharp steps both in the abrupt and adiabatic limits indicate that this effect is a universal feature of phase-coherent ballistic conduction through a constriction.

We thank B. Van Wees, H. van Houten, G. Timp, H. Baranger, P. McEuen, and O. Shapira for helpful discussions. This work was supported in part by NSF Grant No. DMR-8658135, and by a donation of computer time from IBM.

[1]B. J. van Wees, H. van Houten, C. W. J. Beenakker, J. G. Williamson, L. P. Kouwenhoven, D. van der Marel, and C. T. Foxon, Phys. Rev. Lett. **60**, 848 (1988).

[2]D. A. Wharam, T. J. Thornton, R. Newbury, M. Pepper, H. Ajmed, J. E. F. Frost, D. G. Hasko, D. C. Peacock, D. A. Ritchie, and G. A. C. Jones, J. Phys. C **21**, L209 (1988).

[3]A. D. Stone and A. Szafer, IBM J. Res. Dev. **32**, 384 (1988), and references therein.

[4]D. S. Fisher and P. A. Lee, Phys. Rev. B **23**, 6851 (1981).

[5]Y. Imry, in *Directions in Condensed Matter Physics,* edited by G. Grinstein and G. Mazenko (World Scientific, Singapore, 1986), p. 101.

[6]R. Landauer, Z. Phys. B **68**, 217 (1987).

[7]L. I. Glazman, G. B. Lesovick, D. E. Kmelnitskii, and R. I. Shekhter, Pis'ma Zh. Eksp. Teor. Fiz. **48**, 218 (1988) [JETP Lett. **48**, 238 (1988)].

[8]P. A. Lee and D. S. Fisher, Phys. Rev. Lett. **47**, 882 (1981).

[9]B. J. van Wees *et al.*, unpublished.

[10]A. Szafer and A. D. Stone, unpublished.

[11]P. A. Lee, A. D. Stone, and H. Fukuyama, Phys. Rev. B **35**, 1039 (1987), and references therein.