Classical theory of shot noise in resonant tunneling

John H. Davies,* Per Hyldgaard, Selman Hershfield, and John W. Wilkins
Department of Physics, The Ohio State University, Columbus, Ohio 43210-1168
(Received 22 April 1991)

We show that shot noise for electrons can be suppressed in resonant tunneling through a double barrier, using a classical description based on the rate equation for "sequential" tunneling. The suppression is greatest when the escape rates through the two barriers are equal, in agreement with experiment and with the quantum-mechanical "coherent" model of resonant tunneling. A master equation is needed to calculate the noise, but cannot be uniquely derived from the rate equation; choices differ in the way that they describe transport between the emitter and the resonant state. Our choice for the rates, which are consistent with the exclusion principle, gives a suppression of the shot noise. We briefly discuss the results of choosing rates that are consistent with classical or Bose statistics instead of Fermi statistics. Finally, we apply our results to the two-state regime of the classical Coulomb blockade.

I. INTRODUCTION

Resonant tunneling through a double barrier in a semiconducting heterostructure has continued to attract interest since its observation[1] in 1974. A typical resonant-tunneling diode comprises contacts of n-doped GaAs outside two undoped barriers of $Al_xGa_{1-x}As$, with an undoped well of GaAs between the barriers. A resonant state is trapped between the barriers and causes a peak in the transmission coefficient. Its energy lies above the Fermi sea when the structure is in equilibrium, but a bias between the contacts bends the bands so that the resonant state becomes accessible to electrons in the left-hand contact (emitter), as shown in Fig. 1. The current through the device rises until the energy of the resonant state falls below the Fermi sea in the emitter, when the current drops abruptly and produces a region of negative differential resistance in the current-voltage characteristic $I(V)$. This is the important practical feature of resonant tunneling, on which amplifiers, mixers, and other devices intended for use at extremely high frequencies have been based.

The modeling of transport in a resonant-tunneling diode raises fundamental problems. The above description has been based on the analog with an optical Fabry-Perot etalon. Each energy is treated separately with simple wave mechanics. The transmission coefficient $T(E)$ can be found by several methods that solve the Schrödinger equation; no further scattering is included. The contributions from different energies are then added by integrating over the range of incoming energies from both sides of the device, with an appropriate weighting for the density of states. Note that a resonant-tunneling diode is usually used with a bias larger than the Fermi energy, and current is carried (unequally) by electrons at all energies; transport is not restricted to the Fermi energy as in linear response. This is called the "coherent" model, although it is only the wave function at each energy that is considered to be coherent — no coherence between electrons with different energies is assumed.

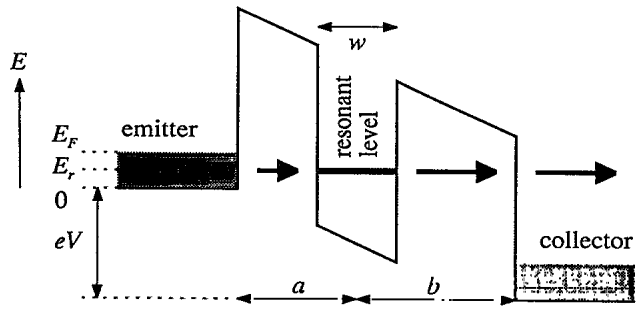

FIG. 1. Band profile through a double-barrier resonant-tunneling diode under typical operating conditions. All electrons in the emitter can contribute to the current, but none from the collector.

The difficulty with this approach is that electrons spend a long time in the resonant state if the width of the resonant peak in $T(E)$ is small, corresponding to a long dwell time. It seems inevitable that electrons in the resonant state will scatter either from phonons or each other. Luryi[2] proposed an alternative model of "sequential" tunneling, in which transport is modeled by a classical rate equation for the density of electrons in the resonant state. One rate describes "hopping" from the emitter to the resonant state, and another describes hopping from the resonant state to the collector. The rate constants are related to the transmission coefficients of the coherent model, but there is no explicit assumption of phase coherence. It was originally suggested that $I(V)$ would be different in the coherent and sequential models, but it has been shown[3,4] that the current is the same in both cases provided that one integrates over the whole width of the resonance in the coherent picture.

If the average current cannot tell us about scattering in the resonant state, what can a more searching experiment such as the shot noise reveal? This was recently measured by Li *et al.*,[5] who studied a range of devices with different ratios of the transmission coefficients of the emitter and

collector barriers. They found that the shot noise took its classical value $2eI$ if this ratio was far from unity, but that shot noise was suppressed, to a minimum of half its classical value, if the transmission coefficients were similar.

There have also been theoretical studies[6–9] of shot noise in purely coherent transport. These show that shot noise is proportional to $T(1-T)$. It is reduced if the transmission coefficient approaches unity, while taking its classical value for weak transmission. A recent study[10] of a resonant-tunneling diode in the coherent limit found suppressed shot noise when the two barriers had similar transmission coefficients. These theoretical results agree with the experimental observations, but raise the question: does this imply that transport in the experimental devices is best described by the coherent model or is the noise, like the current, insensitive to scattering?

We have attempted to answer this question by studying the noise starting from the classical rate equation for sequential tunneling. Unfortunately this is too averaged a description of the system to calculate the noise: we need the greater detail of a master equation. An infinite number of master equations are consistent with the rate equation. We make the simplest choice which is consistent with both the rate equation and with Fermi statistics, which limit the number of states in the well and prevent electrons from tunneling back to the emitter in the limit of large bias, $eV \gg k_BT$. We find that the shot noise is suppressed: the results are identical to those of the coherent model. It therefore appears that the shot noise is insensitive to the degree of coherence *provided* that the master equation holds. We have used nonequilibrium quantum statistical mechanics in a separate paper[11] to discuss resonant tunneling when the rate equation, and consequently our choice for the master equation, does not apply.

The same model can also be applied to the Coulomb blockade in a double tunnel junction, which has a similar master equation.[12] We have also investigated briefly the different master equations that arise if the particles obey classical or Bose statistics. Classical statistics give no suppression of shot noise, while Bose statistics raise the shot noise above its classical value.

The outline of this paper is as follows. In the remainder of the Introduction we define noise and provide a classical derivation of the main general result treated in this paper, the suppression of shot noise for Fermions. The quantum-mechanical theory of shot noise is briefly reviewed in the next section. We set up the rate equation for sequential tunneling in Sec. III, and derive the master equation and the steady-state distribution for the number of electrons in the resonant state. Section IV contains the calculation of the noise from this master equation, and different models are discussed briefly in Sec. V.

A. Definition of noise

We are concerned with noise in the current, defined through the autocorrelation function

$$c_{II}(t) = \langle i(t+t')i(t')\rangle_{t'} - \langle I\rangle^2 , \tag{1.1}$$

where $\langle I\rangle$ is the absolute average value of the current $i(t)$. The averaging is performed over a finite (but long) time T. Taking the Fourier transform, defined by

$$C_{II}(\omega) = \int c_{II}(t)e^{i\omega t}dt, \tag{1.2}$$

gives the two-sided power spectrum of the fluctuations,

$$C_{II}(\omega) = \frac{1}{T}|I(\omega)|^2 - 2\pi\langle I\rangle^2\delta(\omega), \tag{1.3}$$

where $I(\omega)$ is the Fourier transform of $i(t)$. The quantity measured experimentally is the one-sided power spectrum, defined by $S(\omega) = 2C_{II}(\omega)$ for $\omega \geq 0$. We shall usually refer to this simply as the noise.

There are several important sources of noise in a resonant-tunneling diode. First, there is Johnson noise which is related to the linear conductance through the fluctuation-dissipation theorem, and is present even at equilibrium. Other sources give "excess" noise when a current is passed through the device. The second source, our main concern, is shot noise, due to the discrete nature of the charges (electrons) that carry the current. This is reflected in the classical result $S(\omega) = 2e\langle I\rangle$, where e is the magnitude of the electronic charge. The power spectrum is flat at low frequencies, and the noise increases linearly with current provided that the bias V is large, $eV \gg k_BT$. The third important source is $1/f$ noise, which is widely believed to be due to fluctuations in resistance within the device, perhaps by the charging of traps or scattering centers. Its power spectrum should be quadratic in the average current, and of course goes roughly as $1/\omega$. All of these sources of noise are seen in the experiments,[5] and the measurements in units of shot noise had to be made at frequencies above those where $1/f$ noise was significant (of order 1 kHz). This is still in the zero-frequency limit as far as shot noise is concerned, since its dependence on frequency is governed by microscopic rates whose frequencies are typically measured in units of THz. We are therefore justified in taking the limit $\omega \to 0$ in the calculations of shot noise.

B. Classical suppression of shot noise

As an amusing aside, we next provide a classical derivation of the "$T(1-T)$" suppression of shot noise, whose quantum-mechanical analog we shall discuss later [Eq. (2.2)].

Consider a stream of particles incident on a barrier. Transmission through the barrier is treated as a stochastic process with probability T of success; a deterministic approach was taken by Beenakker and van Houten.[13] We assume that T is constant over the range of energies present in the input stream; the range should be subdivided if this is not true, and the noise from each subdivision can then be added.

Divide the incoming stream of particles into time slices of duration τ. Slice i contains N_i incident particles, which varies from slice to slice. If the particles behave independently, the number of electrons transmitted during each slice is given by a binomial distribution. Then the prob-

ability that N particles are transmitted during slice i is given by

$$p_{N_i}(N) = \binom{N_i}{N} T^N (1-T)^{N_i - N} . \tag{1.4}$$

This can be averaged over time slices in two stages. First, consider all slices with the same value of N_i. Averaging the binomial distribution (1.4) gives

$$\langle N \rangle_{N_i} = N_i T \tag{1.5}$$

for the mean and

$$\mathrm{var}(N)_{N_i} = N_i T (1-T) \tag{1.6}$$

for the variance. These expressions can now be averaged over the distribution of N_i. As both are linear, we just replace N_i by its mean $\bar{N}$. The mean transmitted current is

$$\langle I \rangle = e \langle N \rangle / \tau = e\bar{N}T/\tau = e\nu T \tag{1.7}$$

as expected, where $\nu = \bar{N}/T$ is the mean rate of incidence of particles. The shot noise in the limit $\omega \to 0$ is given by Milatz's theorem,[14]

$$S = \lim_{\tau \to \infty} \left[2\tau \,\mathrm{var}\left(\bar{I}_\tau\right) \right], \tag{1.8}$$

where $\bar{I}_\tau$ is the current averaged over a period τ and "var" stands for the variance. In our case $\bar{I}_\tau = eN/\tau$ and

$$\mathrm{var}\left(\bar{I}_\tau\right) = (e/\tau)^2 \,\mathrm{var}(N) = (e/\tau)^2 \bar{N} T (1-T), \tag{1.9}$$

so the shot noise is

$$S = 2e^2 \nu T (1-T) . \tag{1.10}$$

This is suppressed as the probability T of transmission approaches unity, and agrees with the quantum-mechanical result (2.2). The shot noise measures the extra randomness introduced into the flow of particles by the transmission process; there is no randomness, and therefore no shot noise, if $T = 1$. In the opposite limit of small T, transmission through the barrier tends to a Poisson process and full shot noise results.

The simple derivation fails for small biases, $eV < k_B T$, because we have not included thermal fluctuations; it assumes that all the noise is due to the current flowing through the sample. In the limiting case of no current flowing, $\langle N \rangle = 0$, N still has a finite variance. Thus, the above may be thought of as the limit for large voltages, $eV \gg k_B T$, where thermal fluctuations are not important. It is also important that the distribution of N_i, the number of incoming particles, is sufficiently well behaved for the averaging to be done.

Equation (1.10) is far from new; the method is essentially that of Burgess's theorem and the result is closely related to "partition noise;"[14] this is seen, for example, in vacuum tubes, where T would be the probability of *passing* through a grid to the anode. The analogy with classical partition noise has also been pointed out by van der Roer, Heyker, and Kwaspen,[15] and a complementary quantum-mechanical theory can be based on wave packets of electrons.[9,16]

We now move on the corresponding quantum-mechanical theory.

II. QUANTUM-MECHANICAL THEORY

In this section we shall briefly review the coherent theory of the current and noise in resonant tunneling. "Coherent" simply means that scattering (other than by the double barrier) is neglected: simple wave mechanics is used for each energy, and the "intensity" is then summed over energy.

Current and noise in coherent transport

The quantum-mechanical theory of shot noise in quasi-one-dimensional systems was first given by Lesovik[6] and by Yurke and Kochanski.[7] We shall take the limit of zero temperature throughout; the scale is set by the Fermi temperature in the emitter for a resonant-tunneling diode under typical operating conditions, which is usually much larger than room temperature (this contrasts with the linear conductance, where the scale is set by the width of the resonance). The average current in a one-dimensional channel is given by

$$I = e \int \tfrac{1}{2} n_{1\mathrm{D}}(E) v(E) T(E) dE = 2\frac{e}{h} \int T(E) dE; \tag{2.1}$$

the one-dimensional density of states $n_{1\mathrm{D}}(E)$ and the velocity $v(E)$ cancel. The corresponding shot noise in the current at low frequency is[6,7]

$$S = 4\frac{e^2}{h} \int T(E)\left[1 - T(E)\right] dE. \tag{2.2}$$

These include a factor of 2 for spin, and the range of integration is set by the difference in the chemical potentials or by the whole range of incoming energies; the latter holds for a resonant-tunneling diode under the usual large bias, as shown in Fig. 1.

This is clearly the same as the classical expression Eq. (1.10), provided that $T(E)$ is constant over the range of interest, because the rate at which electrons impinge on the barrier (from one side) is given by

$$\nu = \int \tfrac{1}{2} n_{1\mathrm{D}}(E) v(E) = \frac{2}{h} \int dE, \tag{2.3}$$

integrated over the appropriate range. Perhaps Eq. (2.2) is no more than a quantum-mechanical extension of classical partition noise.

We are interested in three-dimensional systems, and need a further sum over transverse wave vectors. For parabolic bands of mass m, and a large bias, the results are

$$\frac{I}{A} = \frac{me}{2\pi^2 \hbar^3} \int_0^{E_F} (E_F - E)\, T(E) dE \tag{2.4}$$

and

$$\frac{S}{A} = \frac{me^2}{\pi^2\hbar^3}\int_0^{E_F} (E_F - E)\,T(E)\,[1-T(E)]\,dE, \qquad (2.5)$$

where A is the cross-sectional area of the device. Equation (2.4) is the well-known Tsu-Esaki formula[17] and (2.5) is its analog for shot noise.

Assume that there is a narrow Lorentzian resonance centered on E_r,

$$T(E) = T_{\rm pk}\left[1+\left(\frac{E-E_r}{\Gamma/2}\right)^2\right]^{-1}. \qquad (2.6)$$

The peak transmission $T_{\rm pk}$ is given in terms of the transmission coefficients of the emitter and collector barriers, T_e and T_c, by

$$T_{\rm pk} = \frac{4T_eT_c}{(T_e+T_c)^2}. \qquad (2.7)$$

The full width Γ gives the total escape rate from the well, $1/\tau$. It can be split into contributions from each barrier, which are given in terms of escape rates and transmission coefficients by

$$\frac{\hbar}{\tau} = \Gamma = \Gamma_e + \Gamma_c = \hbar\left(\frac{1}{\tau_e}+\frac{1}{\tau_c}\right) \approx \hbar\frac{v_r}{2w}(T_e+T_c). \qquad (2.8)$$

Here v_r is the velocity of an electron inside the well at the resonant energy and w is the width of the well, so $v_r/(2w)$ is the frequency at which the electron hits each barrier.

Provided that the resonance is much narrower than the range of incoming energies and is well inside this range, we may approximate the integrals (2.4) and (2.5) by treating $T(E)$ as a δ function. The resulting current density is

$$\frac{I}{A} \approx \frac{me}{2\pi^2\hbar^3}(E_F-E_r)\frac{\pi}{2}\Gamma T_{\rm pk} \qquad (2.9)$$

and the shot noise is

$$\frac{S}{A} \approx \frac{me^2}{\pi^2\hbar^3}(E_F-E_r)\frac{\pi}{2}\Gamma T_{\rm pk}\left(1-\tfrac{1}{2}T_{\rm pk}\right). \qquad (2.10)$$

The dimensionless ratio $\hat{s}$ of the shot noise normalized to its full classical value $2eI$ does not depend on E_r, nor on the width of the resonance:

$$\hat{s} = \frac{S}{2eI} = 1-\tfrac{1}{2}T_{\rm pk} = 1-\frac{1}{2}\frac{4T_eT_c}{(T_e+T_c)^2}. \qquad (2.11)$$

This result was first derived by Chen and Ting[10] using path integrals. Shot noise is suppressed in resonant tunneling, with the maximum effect for a symmetric structure with $T_e = T_c$ giving $T_{\rm pk} = 1$. This result is consistent with the experiments of Li *et al.*[5]

Equations (2.4) and (2.5) can easily be evaluated analytically for a Lorentzian resonance without further approximation. Typical results, for a symmetric structure with $T_{\rm pk} = 1$ and $\Gamma/E_F = 0.02$, are plotted in Fig. 2. The current is normalized to the maximum value predicted by Eq. (2.9), with $E_r = 0$:

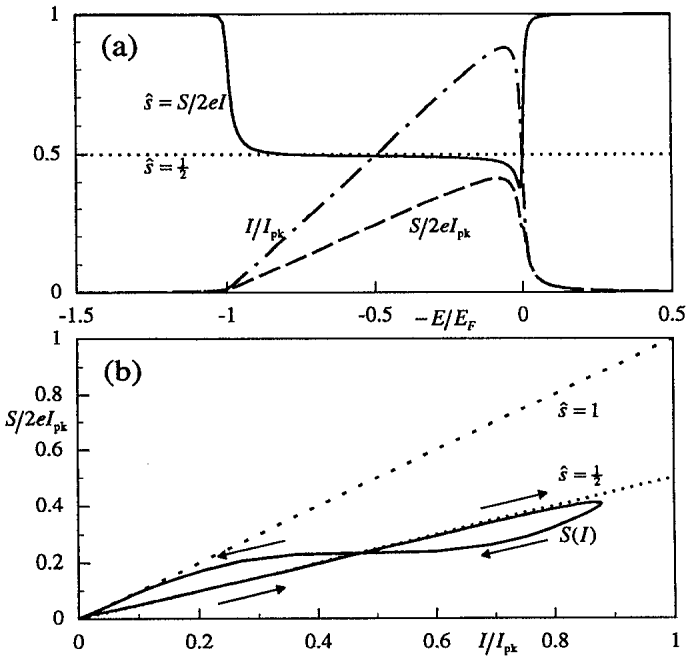

FIG. 2. Current I, shot noise S, and ratio of the noise to its classical value $\hat{s} = S/(2eI)$ for a symmetric resonant-tunneling diode calculated within the coherent picture, assuming a Lorentzian resonance of full width $\Gamma = 0.02E_F$ and peak $T_{\rm pk} = 1$. The dotted lines show the prediction $\hat{s} = 1-\frac{1}{2}T_{\rm pk} = \frac{1}{2}$. The abscissa of (a) is reversed in energy to correspond with increasing bias. The dip near $E = 0$ in (a), where the resonance falls below the range of incoming energies, produces the loop in (b) where the arrows show the direction of increasing bias.

$$I_{\rm pk} = \frac{me}{2\pi^2\hbar^3}E_F\frac{\pi}{2}\Gamma T_{\rm pk}. \qquad (2.12)$$

Both the current and noise have the familiar triangular shape for a three-dimensional resonant-tunneling diode. The ratio $\hat{s}$ of the noise to its full classical value is unity when the resonance is outside the range of incident electrons, and almost constant at $\frac{1}{2}$ when it is within the range. There is a dip in $\hat{s}$ as the resonance passes below the range of incident electrons ($E_r \approx 0$), which is reflected as a loop in the parametric plot $S(I)$ shown in Fig. 2(b). A loop is also seen in the experiments (Li *et al.*,[5] Fig. 2), but there the noise goes *above* $2eI$ which our treatment can never give. Comparison is difficult in this region because the experimental devices show bistability caused by a buildup of charge in the resonant state. This is not seen in our calculations because they are not self-consistent.

The simple coherent picture explains the experiments rather well, and it is tempting to conclude that the experimental devices must be close to the coherent limit. However, we know that the current is insensitive to incoherence, and this may be true of the noise too. A calculation of the effect of incoherence is therefore vital. It is known[18–20] that the effect of structureless scattering is to lower and broaden the peak in $T(E)$, while preserving its

area. The current [Eq. (2.5)] is therefore unchanged, but Eq. (2.11) shows that the noise should increase because it depends only on $T_{\rm pk}$. Unfortunately the derivation of the expressions for I and S holds only in the coherent limit, and it is dangerous to try to draw conclusions from them about the effect of incoherence.

In the next sections we shall calculate the noise within a classical description of a resonant-tunneling diode based on rate equations, and compare it with the quantum-mechanical result.

III. CLASSICAL THEORY

We shall now set up the rate equation for the "sequential" picture of tunneling and develop a master equation, in preparation for calculating the noise in the following section.

A. Rate equation

The sequential model of resonant tunneling was introduced by Luryi[2] as an alternative to the coherent picture using wave mechanics in the preceding section. The idea is that an electron spends so long in the resonant state that it will inevitably be scattered by phonons or other electrons. The processes of entering and leaving the resonance would then not be phase coherent, and a classical description of independent rates is more appropriate.

We shall take the sequential model to be defined by the following rate equation for the ensemble-averaged density of electrons per unit area in the resonant state, n:

$$\frac{dn}{dt}=\frac{n_0-n}{\tau_e}-\frac{n}{\tau_c}. \tag{3.1}$$

The time constants τ_e and τ_c were given in terms of the quantum-mechanical transmission coefficients in Eq. (2.8); their reciprocals are the escape rates for tunneling through the individual barriers to the emitter and to the collector. The resonant state would fill to a density n_0 if it were disconnected from the collector. It is a two-dimensional electron gas, and Fig. 1 shows that the effective Fermi energy in the resonance is (E_F-E_r) (or zero if this is negative), so

$$n_0=\frac{m}{\pi\hbar^2}\left(E_F-E_r\right). \tag{3.2}$$

It is assumed that no electrons enter from the collector, which is appropriate under normal operating conditions (Fig. 1). The average density in a steady state is

$$\langle n\rangle=\frac{\tau_c}{\tau_e+\tau_c}n_0, \tag{3.3}$$

and the average current density is

$$\frac{I}{A}=e\frac{n_0-\langle n\rangle}{\tau_e}=e\frac{\langle n\rangle}{\tau_c}=e\frac{n_0}{\tau_e+\tau_c}. \tag{3.4}$$

Substituting Eq. (2.8) into this gives

$$\begin{aligned}\frac{I}{A}&=e\frac{1}{\hbar\left(\Gamma_e^{-1}+\Gamma_c^{-1}\right)}\frac{m}{\pi\hbar^2}\left(E_F-E_r\right)\\&=\frac{me}{4\pi\hbar^3}\left(E_F-E_r\right)\Gamma T_{\rm pk},\end{aligned} \tag{3.5}$$

which agrees with Eq. (2.9). Thus the "coherent" and "sequential" pictures give the same average current if the coherent result is integrated over the whole resonance,[3,4] a result confirmed by a more searching analysis that includes the effect of scattering.[11] The rate equation is a much less detailed description of the system than the quantum-mechanical treatment used in Sec. II, but contains sufficient information for the current. As we shall see, this is not true for the noise.

The fluctuations in the current are related to those in the density of electrons in the resonant state. The autocorrelation function for the density is defined by

$$c_{nn}(t)=\langle n(t)n(0)\rangle-\langle n\rangle^2. \tag{3.6}$$

Substituting into the rate equation (3.1) shows that $c_{nn}(t)$ obeys the simple equation

$$\frac{d}{dt}c_{nn}(t)=-\frac{1}{\tau}c_{nn}(t) \tag{3.7}$$

for $t>0$. Integrating gives

$$c_{nn}(t)=\mathrm{var}(n)\exp\left(-\frac{|t|}{\tau}\right). \tag{3.8}$$

Fluctuations in density relax with a simple exponential dependence on time scaled by the total escape rate. Unfortunately the prefactor $\mathrm{var}(n)$ is simply a constant of integration, and is *not* predicted by the rate equation (3.1) alone. A more detailed description of the system is needed to make further progress: a master equation.

B. Master equation

A master equation is the simplest extension of the classical rate equation that can describe the fluctuations in density, and has previously been applied to tunnel diodes.[21] As shot noise depends on the particulate nature of current, it is convenient to use $N=nA$, the (integer) number of electrons in the resonance, rather than the density n. We now seek a master equation for $p(N,t)$, the probability that there are N electrons in the resonant state at time t.

Unfortunately we are going the wrong way in attempting to derive a master equation from the rate equation (3.1); the rate equation should be derived by summing over the master equation, not vice versa. An infinite number of master equations is consistent with the rate equation. All have the same structure, but have different rates. There is a "generation" rate $g(N)$ that increases the number of electrons from N to $N+1$, and a "recombination" rate $r(N)$ that decreases it from N to $N-1$:

$$\begin{aligned}\frac{d}{dt}p(N,t)=&\,g(N-1)p(N-1,t)+r(N+1)p(N+1,t)\\&-\left[g(N)+r(N)\right]p(N,t).\end{aligned} \tag{3.9}$$

The picture of a partially occupied resonant state closely resembles the occupation of impurity levels in a semiconductor, whose nomenclature we have adopted. There is only one "generation" process possible under typical operating conditions (Fig. 1), which is an electron hopping

from the emitter into the resonant state (g hop). However, there are two possible "recombination" processes: an electron can hop out of the resonant state forward to the collector (c hop) or backward to the emitter (e hop). The recombination rate is the sum of the rates for these two processes, $r(N) = c(N) + e(N)$. There is an additional constraint, however, which is that the rates must respect the exclusion principle. This means that e hops cannot occur if the states in the emitter are fully occupied, which we take to be the case; this is the "two-process" model. Then $e(N) = 0$ and the rates follow directly from Eq. (3.1):

$$g(N) = \frac{N_0 - N}{\tau_e}, \quad r(N) = c(N) = \frac{N}{\tau_c}, \tag{3.10}$$

where $N_0 = An_0$ is the maximum number of electrons that can be accommodated in the resonance. We shall consider other choices for the master equation that include e hops in Sec. V; these "three-process" models are applicable to classical or Bose particles or if the states in the emitter are only partially filled. The rate equation (3.1) can be recovered from the master equation (3.9) by multiplying by N and summing over all N.

An important point is that we take N_0 to be an integer. This need not be true in general (although the number of electrons N must be an integer), and the master equation used in the theory of the Coulomb blockade[12] provides an example where the continuous nature of N_0 is vital and our approximation is inappropriate. The number of electrons in a typical resonant-tunneling diode is very large, and the distribution of N is not highly skewed, so we do not expect this assumption to be important. It provides enormous simplifications, as we shall see in the following sections, mainly because the functional form of the rates in equation (3.10) automatically satisfy the following end-point conditions:

$$g(N_0) = 0, \quad r(0) = 0. \tag{3.11}$$

These are the rates for the unphysical processes of generation into a full state and recombination from an empty state. They must be zero physically, but in general the extrapolated functional forms of $g(N)$ and $r(N)$ do not vanish, and these terms must be explicitly excluded from summations.

C. Steady-state distribution

The steady-state distribution $p_0(N)$ is obtained by setting the derivative in the master equation (3.9) to zero. An equivalent method is to use detailed balance on the rates, which gives immediately

$$\frac{p_0(N+1)}{p_0(N)} = \frac{g(N)}{r(N+1)} = \frac{\tau_c}{\tau_e}\frac{N_0 - N}{N+1}. \tag{3.12}$$

The rates obey the condition $r(0) = 0$ and $g(N_0) = 0$, which restricts N to the range $0 \le N \le N_0$. The solution to this recurrence relation is a binomial distribution,

$$p_0(N) = \binom{N_0}{N}\left(\frac{\tau_c}{\tau_e+\tau_c}\right)^N\left(\frac{\tau_e}{\tau_e+\tau_c}\right)^{N_0-N}. \tag{3.13}$$

The mean value is

$$\langle N\rangle = N_0\frac{\tau_c}{\tau_e+\tau_c}, \tag{3.14}$$

in agreement with (3.3), and the variance is

$$\mathrm{var}(N) = N_0\frac{\tau_e\tau_c}{(\tau_e+\tau_c)^2} = \langle N\rangle\left(1 - \frac{\langle N\rangle}{N_0}\right). \tag{3.15}$$

Note that $\mathrm{var}(N) < \langle N\rangle$, showing that fluctuations are suppressed compared with a Poisson distribution for which the mean and variance are equal. We can now complete the autocorrelation function,

$$c_{NN}(t) = N_0\frac{\tau_e\tau_c}{(\tau_e+\tau_c)^2}\exp\left(-\frac{|t|}{\tau}\right), \tag{3.16}$$

which we shall use to calculate the shot noise in the next section.

IV. NOISE IN THE CLASSICAL THEORY

We shall now calculate the noise for the models of resonant tunneling defined by the master equations in the preceding section. First we derive the form of the current, and then a general formula for the noise in terms of cross-correlation functions, which we evaluate in detail for the master equation discussed above. A formula applicable to other models will be considered more briefly, followed by a special one-dimensional case in which the number of electrons in the resonance is restricted to 0 or 1.

A. Current

We first need the form of the current. An obvious way to measure the current is to count the number of electrons that hop out of the emitter or into the collector. While this is satisfactory for the average current, these two currents are not equal for frequencies above zero because charge can accumulate in the resonant state. A more precise definition is needed.

As discussed above, an electron passes through the device in two hops, from the emitter to the resonant state (g hop) and from there to the collector (c hop). Each of these hops generates a pulse of current in the external circuit. The Ramo-Shockley theorem[22,23] shows that a g hop causes a charge αe to flow in the external circuit, where $\alpha = a/(a+b)$ and the thicknesses a and b are shown in Fig. 1. Similarly, a c hop causes a charge βe to flow, where $\beta = b/(a+b)$. Clearly $\alpha + \beta = 1$ to ensure conservation of charge. The current therefore takes the form

$$i(t) = \alpha e\sum_j f(t - t_j^g) + \beta e\sum_j f(t - t_j^c), \tag{4.1}$$

where t_j^g are the times of the g pulses and so on. The function $f(t)$ gives the shape of the pulse that flows in

the circuit. Its form is generally determined by the circuit rather than the device, although in principle it could depend on the elusive "tunneling time." We shall assume that it decays rapidly on the time scales of interest, and its only important property is its normalization,

$$\int_{-\infty}^{\infty} f(t)dt = 1. \tag{4.2}$$

This means that its Fourier transform $F(\omega)$ satisfies $F(0) = 1$, and we shall be able to ignore it at low frequencies.

In a more general model the current consists of a sum of three terms, with the addition of negative contributions from e hops when an electron hops back from the resonant state to the emitter:

$$i(t) = \alpha e \sum_j f(t - t_j^g) - \alpha e \sum_k f(t - t_k^e) + \beta e \sum_l f(t - t_l^c). \tag{4.3}$$

The e process is forbidden for Fermions if the states in the emitter are full, but would be allowed for classical particles or bosons. The average rates of the three processes are not equal, as they are for the two-process model, although the average rate of g hops must be the sum of the rates of e hops and c hops to ensure the conservation of electrons.

Although we have taken some care to define the current precisely, it is rarely important at low frequencies. A real device tends to maintain charge neutrality over some time scale, and any measure of the current is adequate for frequencies slower than this.

B. Noise

We can now substitute the specific form for the current into the general formula (1.3) for the noise. The Fourier transform of the current, Eq. (4.1), is

$$I(\omega) = eF(\omega)\left[\alpha \sum_j \exp\left(i\omega t_j^g\right) + \beta \sum_j \exp\left(i\omega t_j^c\right)\right]. \tag{4.4}$$

The summations are restricted to hops within the finite time of observation T, as in Eq. (1.3). Squaring this gives

$$|I(\omega)|^2 = e^2 |F(\omega)|^2 \left\{\alpha^2 \sum_{j,k} \exp\left[i\omega\left(t_j^g - t_k^g\right)\right] + \beta^2 \sum_{j,k} \exp\left[i\omega\left(t_j^c - t_k^c\right)\right] + \alpha\beta \sum_{j,k} \exp\left[i\omega\left(t_j^g - t_k^c\right)\right] + \beta\alpha \sum_{j,k} \exp\left[i\omega\left(t_j^c - t_k^g\right)\right]\right\}. \tag{4.5}$$

The next step is to reduce each of these double summations to a correlation function. This is done in much the same way as the correlation and structure functions for disordered materials (see, for example, Ziman,[24] p. 124).

Start with the first summation,

$$\sum_{j,k} \exp\left[i\omega\left(t_j^g - t_k^g\right)\right]. \tag{4.6}$$

The terms with $j = k$ each give unity and there are N_T of them, which is the average number of g hops or c hops during the period of observation T. The off-diagonal terms are

$$\sum_k \sum_{j \neq k} \exp\left[i\omega\left(t_j^g - t_k^g\right)\right]. \tag{4.7}$$

Consider the sum over j, for fixed k. On average, this can be replaced by the integral

$$\int_{-\infty}^{\infty} \exp\left[i\omega\left(t - t_k^g\right)\right] h_{gg}\left(t - t_k^g\right) dt, \tag{4.8}$$

where the correlation function $h_{gg}(t)$ is the rate of g hops given that there was one at $t = 0$. It is an even function of time. After this averaging, the result must be the same for each value of k, so we can simply multiply by the number of g hops, N_T, and set $t_k^g = 0$ in the integral. Thus

$$\left\langle \sum_{j,k} \exp\left[i\omega\left(t_j^g - t_k^g\right)\right] \right\rangle = N_T\left[1 + \int_{-\infty}^{\infty} h_{gg}(t) e^{i\omega t} dt\right] = N_T\left[1 + H_{gg}(\omega)\right], \tag{4.9}$$

where $H_{gg}(\omega)$ is the Fourier transform of $h_{gg}(t)$.

The other terms can be reduced in much the same way. The cross-correlation terms [the third and fourth in Eq. (4.5)] have no diagonal parts, only the correlation functions. These functions are defined such that $h_{gc}(t)$ is the average rate of g hops at time t, given that there was a c hop at $t = 0$. They have the symmetry

$$h_{gc}(t) = h_{cg}(-t); \tag{4.10}$$

the rate at which g hops come after a c hop must be the same as the rate at which c hops come before a g hop. Collecting all terms, we obtain

$$|I(\omega)|^2 = e^2 |F(\omega)|^2 N_T \{\alpha^2[1 + H_{gg}(\omega)] + \beta^2[1 + H_{cc}(\omega)] + \alpha\beta H_{gc}(\omega) + \beta\alpha H_{cg}(\omega)\}. \tag{4.11}$$

This can be rewritten using "reduced" functions. Correlation between the hops must vanish at large times, and the rates tend to their average values. For example, $h_{gc}(t)$ must have the limit

$$\lim_{t\to\pm\infty} h_{gc}(t) = \bar{h}_g, \tag{4.12}$$

where $\bar{h}_g$ is the average rate of g hops. The average rates of g and c hops are the same for the two-process model, $\bar{h}_g = \bar{h}_c = \bar{h}$, and are given by the average current, Eq. (3.4),

$$\bar{h} = \frac{N_T}{T} = \frac{\langle I\rangle}{e} = \frac{N_0}{\tau_e + \tau_c}. \tag{4.13}$$

Define a reduced function as the nontrivial part of $h_{gc}(t)$ by

$$g_{gc}(t) = h_{gc}(t) - \bar{h}, \tag{4.14}$$

and similarly for the other functions. The Fourier transform is

$$G_{gc}(\omega) = H_{gc}(\omega) - 2\pi\bar{h}\delta(\omega). \tag{4.15}$$

Equation (4.11) becomes

$$\begin{aligned}|I(\omega)|^2 = {} & e^2 |F(\omega)|^2 N_T \left\{\alpha^2 [1 + G_{ee}(\omega)] + \beta^2 [1 + G_{cc}(\omega)] + \alpha\beta G_{ec}(\omega) + \beta\alpha G_{ce}(\omega)\right\} \\ & + e^2 |F(\omega)|^2 N_T 2\pi\delta(\omega)\bar{h}\left(\alpha^2 + \beta^2 + \alpha\beta + \beta\alpha\right).\end{aligned} \tag{4.16}$$

Equation (4.13) and the relation $\alpha + \beta = 1$ allow the term with the δ function to be simplified to

$$2\pi \langle I\rangle^2 \delta(\omega) T, \tag{4.17}$$

which cancels the subtracted term in the current autocorrelation function (1.3) when (4.16) is substituted. The power spectrum becomes

$$S(\omega) = 2e\langle I\rangle |F(\omega)|^2 \left\{\alpha^2 [1 + G_{gg}(\omega)] + \beta^2 [1 + G_{cc}(\omega)] + \alpha\beta G_{gc}(\omega) + \beta\alpha G_{cg}(\omega)\right\}. \tag{4.18}$$

This is the final result for the noise in terms of the hop-hop correlation functions. It is linear in $\langle I\rangle$, as expected for shot noise. Dividing $S(\omega)$ by $2e\langle I\rangle |F(\omega)|^2$ gives the dimensionless ratio $\hat{s}(\omega)$ of the shot noise to its full classical value, introduced in Sec. II.

This result shows a geometrical suppression of shot noise even if there is no correlation between pulses. Discarding all the $G(\omega)$ functions from Eq. (4.18) leaves

$$\hat{s}(\omega) = \frac{S(\omega)}{2e\langle I\rangle |F(\omega)|^2} = (\alpha^2 + \beta^2) = \left[1 - \frac{1}{2}\frac{4ab}{(a+b)^2}\right]. \tag{4.19}$$

The noise ratio $\hat{s}$ reaches its minimum value for a structure whose geometry is symmetric, $a = b$, when shot noise is halved. It is easy to explain this limit: an electron produces two equal pulses of weight $e/2$ in the external circuit as it passes through the device, so it is as if the current were carried by particles of charge $e/2$ rather than e. The classical formula $S = 2e\langle I\rangle$ shows that the shot noise is proportional to the charge of the carriers, and is therefore halved too. More generally, the suppression arises because we have divided the current into two processes whose noise adds incoherently. The magnitude of each current is some fraction of the total current, but the noise depends quadratically on the fraction and is therefore reduced. This is essentially the same as the suppressed shot noise obtained for two resistors in series, which Li *et al.*[5] suggested might explain their data. There would be arbitrarily large fluctuations in the charge in the resonant state if there were really no correlation between hops, so this result cannot apply at low frequencies. This is why it contradicts the statement earlier that the form of the current, and therefore the dependence on α and β, should be unimportant at low frequency. However, it should hold at high frequencies on a scale set by the device, $\omega \gg 1/\tau$ for resonant tunneling, and we will find it as this limit of the exact result.

The next task is to evaluate the correlation functions.

C. Correlation functions

Start with the correlation function $h_{cc}(t)$ for $t > 0$, which is the rate of c hops given that there was one at $t = 0$. The rate of c hops where there are N electrons in the resonance is given by $c(N)$, Eq. (3.10). Define the conditional probability $p(N, t|M, 0)$ to be the probability of finding N electrons in the resonant state at time t, given that there were M electrons at $t = 0$. This must be found by solving the master equation (3.9) subject to the initial condition $p(N, t = 0) = \delta_{N,M}$. Conveniently it turns out that we can avoid this difficult task, and will only need the autocorrelation function for the density which we already know.

Let the probability distribution of the number of electrons immediately after the c hop at $t = 0$ be $p_c(M)$. This is restricted to the range $0 \le M \le (N_0 - 1)$ because the c hop reduced the number of electrons by 1. The probability distribution of N at t is given by summing $p(N, t|M, 0)$ over all M with a weighting given by $p_c(M)$:

$$\sum_{M=0}^{N_0-1} p(N,t|M,0)p_c(M). \tag{4.20}$$

Then $h_{cc}(t)$, the average rate of c hops at t, is given by summing $c(N)$ with this weighting, excluding the state $N=0$ from which recombination is impossible,

$$h_{cc}(t) = \sum_{N=1}^{N_0} c(N) \sum_{M=0}^{N_0-1} p(N,t|M,0)p_c(M). \tag{4.21}$$

We now need the form of $p_c(M)$. There were $M+1$ electrons before the c hop, so $p_c(M)$ is proportional to the steady-state distribution before the hop, $p_0(M+1)$, weighted by the probability of making a c hop, $c(M+1)$:

$$p_c(M) = p_0(M+1)c(M+1)/D. \tag{4.22}$$

The normalization factor is

$$D = \sum_{M'=1}^{N_0} p_0(M')c(M'). \tag{4.23}$$

This is the average recombination rate, and must be the same as the average generation rate. Both are equal to the average hopping rates, Eq. (4.13), in the two-process model, so $D = \bar{h}$. The distribution can be rewritten using detailed balance, Eq. (3.12), as

$$p_c(M) = p_0(M)g(M)/\bar{h}. \tag{4.24}$$

Substituting into Eq. (4.21) gives

$$h_{cc}(t) = \bar{h}^{-1} \sum_{N=1}^{N_0} c(N) \sum_{M=0}^{N_0-1} p(N,t|M,0)p_0(M)g(M). \tag{4.25}$$

This can be rewritten slightly in terms of the joint probability $p(N,t;M,0)$ of finding M electrons at $t=0$ and N electrons at time t, using the general result

$$p(N,t;M,0) = p(N,t|M,0)p_0(M). \tag{4.26}$$

The result is

$$h_{cc}(t) = \bar{h}^{-1} \sum_{N=1}^{N_0} c(N) \sum_{M=0}^{N_0-1} p(N,t;M,0)g(M). \tag{4.27}$$

At large times when correlations have died away, the joint probability factorizes into the product of two steady-state distributions, so

$$\begin{aligned} h_{cc}(t\to\pm\infty) &= \bar{h}^{-1} \sum_{N=1}^{N_0} c(N)p_0(N) \sum_{M=0}^{N_0-1} g(M)p_0(M) \\ &= \bar{h}_c\bar{h}_g/\bar{h} = \bar{h}. \end{aligned} \tag{4.28}$$

This is consistent with the limit (4.12). We can therefore subtract it to leave the reduced function,

$$g_{cc}(t) = \bar{h}^{-1} \sum_{N=1}^{N_0} \sum_{M=0}^{N_0-1} c(N)g(M)\,[p(N,t;M,0) - p_0(N)p_0(M)]. \tag{4.29}$$

Two vital simplifications can now be made that depend on the specific form of the rates in our model, Eq. (3.10), as discussed in Sec. III B. First, the rates are linear functions of the density of electrons. Second, they automatically obey the end-point condition of Eq. (3.11), which prevents the impossible processes of recombination from an empty state and generation into a full state. In general these unphysical terms must be carefully excluded from the summations, as in Eq. (4.29), but here we can safely extend the summations over the whole range $0 \le M, N \le N_0$. Substituting the rates gives

$$g_{cc}(t) = \frac{1}{\bar{h}\tau_e\tau_c} \sum_{M,N=0}^{N_0} (N_0 - M)N\,[p(N,t;M,0) - p_0(M)p_0(N)]. \tag{4.30}$$

This is the difference of two terms. The first of these, with N_0, cancels out because the sum over M gives

$$\sum_M p(N,t;M,0) = p_0(N). \tag{4.31}$$

This leaves

$$\begin{aligned} g_{cc}(t) &= -\frac{1}{\bar{h}\tau_e\tau_c} \sum_{M,N=0}^{N_0} MN[p(N,t;M,0) - p_0(N)p_0(M)] \\ &= -\frac{1}{\bar{h}\tau_e\tau_c} c_{NN}(t) = -\frac{1}{N_0\tau} c_{NN}(t), \end{aligned} \tag{4.32}$$

using the definition of the number autocorrelation function and that of $\bar{h}$ [Eq. (4.13)]. We would not have been able to make this reduction without removing the restrictions on the summations. Substituting Eq. (3.16) for the correlation function finally gives

$$\begin{aligned} g_{cc}(t) &= -\frac{\mathrm{var}(N)}{N_0}\frac{1}{\tau}\exp\left(-\frac{|t|}{\tau}\right) \\ &= -\frac{1}{\tau_e+\tau_c}\exp\left(-\frac{|t|}{\tau}\right). \end{aligned} \tag{4.33}$$

This is negative, which means that the average rate of c hops is suppressed after a c hop. The reason is that the number of electrons is slightly below average after a c hop. The rate $c(N)$ of c hops decreases as N decreases, and is therefore reduced. It is as if the pulses repel one another in time.

The other correlation functions can be evaluated in the same way. Symmetry requires that the other diagonal function $h_{gg}(t)$ be identical to $h_{cc}(t)$; it only requires interchanging M and N in Eq. (4.29). Thus like pulses always repel one other. The off-diagonal functions at positive times are

$$g_{gc}(t) = \frac{\tau_c}{\tau_e}\frac{1}{\tau_e+\tau_c}\exp\left(-\frac{t}{\tau}\right), \tag{4.34}$$

$$g_{cg}(t) = \frac{\tau_e}{\tau_c}\frac{1}{\tau_e+\tau_c}\exp\left(-\frac{t}{\tau}\right). \tag{4.35}$$

The functions for negative time can be obtained by symmetry, Eq. (4.10); they are discontinuous at $t = 0$. These correlation functions are positive, indicating that the average rate of g hops rises after a c hop, which can be explained in the same way as the repulsion of like hops.

D. Noise

We can now substitute these correlation functions into the general expression (4.18) and obtain the noise. The exponential functions in time give a Lorentzian dependence on frequency:

$$S(\omega) = 2e\langle I\rangle |F(\omega)|^2\left\{(\alpha^2+\beta^2) - \frac{2\tau_e\tau_c}{(\tau_e+\tau_c)^2}\frac{1}{1+(\omega\tau)^2}\left[\alpha^2+\beta^2-\alpha\beta\left(\frac{\tau_e}{\tau_c}+\frac{\tau_c}{\tau_e}\right)\right]\right\}, \tag{4.36}$$

$$= 2e\langle I\rangle |F(\omega)|^2\left\{1 - \frac{1}{2}\frac{4\tau_e\tau_c}{(\tau_e+\tau_c)^2}\frac{1}{1+(\omega\tau)^2} - \frac{1}{2}\frac{4ab}{(a+b)^2}\frac{(\omega\tau)^2}{1+(\omega\tau)^2}\right\}. \tag{4.37}$$

This is our final result for the noise, valid at all frequencies. The third term in curly brackets in Eq. (4.37) vanishes in the important experimentally accessible limit of low frequencies, leaving the ratio

$$\hat{s}(\omega\to 0) = \frac{S}{2e\langle I\rangle} = 1 - \frac{1}{2}\frac{4\tau_e\tau_c}{(\tau_e+\tau_c)^2} = 1 - 2\frac{\mathrm{var}(N)}{N_0} \tag{4.38}$$

or

$$S(\omega\to 0) = 2e\langle I\rangle - \frac{4e^2}{\tau_e+\tau_c}\mathrm{var}(N). \tag{4.39}$$

This shows that shot noise in resonant tunneling is directly related to the variance of the number of electrons in the resonant state, and is suppressed at low frequency. Equation (2.8) shows that the lifetimes $\tau_{e,c}$ are inversely proportional to the transmission probabilities $T_{e,c}$, which in turn means that Eq. (4.38) from the master equation is identical to the quantum-mechanical result, Eq. (2.11). Thus the classical model and pure coherent quantum mechanics give exactly the same result for the shot noise at low frequencies, as well as for the average current. Equation (4.38) is plotted in Fig. 3, with the experimental results of Li *et al.*[5] The ratio T_e/T_c for the experimental devices was derived by self-consistent modeling; this is notoriously difficult, and small errors in the potential are greatly magnified in the transmission coefficients, so we regard the agreement as good. We shall consider the implications further in Sec. V.

Note that the geometrical factors α and β vanish from Eq. (4.38), emphasizing that the precise form of the current is not important at low frequencies as discussed earlier. This can be traced back to the general result (4.18), where all four terms with correlation functions become equal as $\omega\to 0$, and can be derived from the conservation of charge. A device with $a = b$ and $\tau_e = \tau_c$ would have $\hat{s} = \frac{1}{2}$ at all frequencies. This is because g and c hops are indistinguishable in the external circuit if $a = b$, both carrying charge $e/2$, and the *total* rate of hops is independent of N if $\tau_e = \tau_c$. It therefore appears that there is a stream of uncorrelated pulses at a constant rate, each of half the electronic charge, which halves the shot noise.

In the limit of high frequency, the noise ratio is

$$\hat{s}(\omega\to\infty) = \alpha^2+\beta^2 = \left[1 - \frac{1}{2}\frac{4ab}{(a+b)^2}\right]. \tag{4.40}$$

This is the noise that would arise from adding the full shot noise coming from each of the two contributing currents in the two-process model, while asuming no correlation between current pulses. It is exactly the result found previously, Eq. (4.19), when all the reduced correlation functions were discarded.

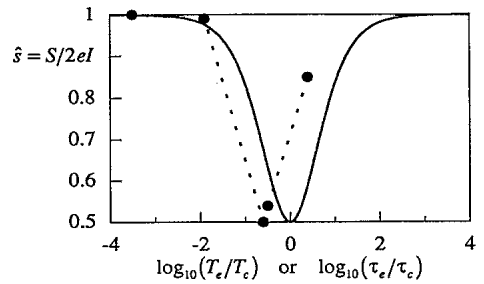

FIG. 3. Calculated ratio $\hat{s}(\omega = 0)$ of shot noise to its full classical value $2eI$ for a resonant-tunneling diode, plotted as a function of the ratio of transmission coefficients T_e/T_c or escape rates τ_e/τ_c for the two barriers. The points are from the experiments of Li *et al.* (Ref. 5). The solid curve is the theoretical result obtained using either the coherent quantum-mechanical approach or the two-process model based on a classical rate equation. The minimum value of $\hat{s} = \frac{1}{2}$ in the theory occurs for a symmetric structure with equal transmission coefficients for the two barriers.

E. Noise in other models

In principle we would need to repeat the above derivation with nine correlation functions involving g, c, and e hops in the more general three-process models where electrons are also allowed to return to the emitter from the resonant state. Fortunately there is a shortcut for $\hat{s}(\omega = 0)$. The precise definition of the current is not important at low frequencies, as noted above, so we can restrict attention to hops from the resonant state to the collector. This is equivalent to setting $\alpha = 0$ and $\beta = 1$, and we only need $h_{cc}(t)$. The correlation function can be calculated in a slightly different way from the preceding section, making no assumption about the processes that transfer electrons between the emitter and the resonant state. This assumption entered when we used detailed balance to replace Eq. (4.22) with (4.24). If we do not take this step, Eqs. (4.21) and (4.22) give

$$h_{cc}(t) = \bar{h}^{-1} \sum_{N=1}^{N_0} \sum_{M=0}^{N_0-1} c(N)c(M+1) \times p(N,t|M,0)p_0(M+1). \quad (4.41)$$

Substituting the form of the rates and extending the summations gives

$$h_{cc}(t) = \frac{1}{\bar{h}\tau_c^2} \sum_{M,N=0}^{N_0} Np(N,t|M,0)\,(M+1)\,p_0(M+1). \quad (4.42)$$

The crucial feature is that $\sum Np(N,t|M,0)$ gives the evolution of the *ensemble average* value of N, given that $N = M$ at $t = 0$. This follows from the rate equation alone, and it is therefore independent of the choice of master equation. The result is exponential relaxation towards the mean, as in the autocorrelation function (3.8),

$$\sum_N Np(N,t|M,0) = \langle N\rangle + (M - \langle N\rangle)\exp(-t/\tau)\,. \quad (4.43)$$

Substituting into Eq. (4.42) gives

$$h_{cc}(t) = \frac{1}{\bar{h}\tau_c^2}\left[\langle N\rangle \sum_M (M+1)\,p_0(M+1) + \exp(-t/\tau)\sum_M (M+1)\,(M - \langle N\rangle) \times p_0(M+1)\right], \quad (4.44)$$

which reduces to

$$h_{cc}(t) = \frac{1}{\bar{h}\tau_c^2}\{\langle N\rangle^2 + \exp(-t/\tau)[\langle N^2\rangle - \langle N\rangle - \langle N\rangle^2]\}. \quad (4.45)$$

The constant part gives $\bar{h}$ as before, and the rest reduces to

$$g_{cc}(t) = -\frac{1}{\tau_c}\left[1 - \frac{\mathrm{var}(N)}{\langle N\rangle}\right]\exp\left(-\frac{|t|}{\tau}\right). \quad (4.46)$$

Taking the limit $G_{cc}(\omega \to 0)$ of the Fourier transform gives the noise ratio,

$$\hat{s}(\omega \to 0) = 1 - \frac{2\tau_e}{\tau_e + \tau_c}\left[1 - \frac{\mathrm{var}(N)}{\langle N\rangle}\right]. \quad (4.47)$$

The processes that transfer electrons between the emitter and resonant state only affect this implicitly through the variance: no assumptions about them have been made. Substituting Eq. (3.15) for the variance shows that the new result agrees with the previous one, Eq. (4.38).

This result is important because it relates the noise directly to the variance of the number of electrons in the resonant state: a smaller variance reduces the shot noise for given values of τ_e and τ_c (and therefore fixed $\langle N\rangle$). It is more general than the previous result, Eq. (4.39), which applies only in the absence of electrons returning to the emitter from the resonant state. We shall apply it to different models in Sec. V where it is shown that different master equations, all consistent with the rate equation and therefore giving the same value of $\langle N\rangle$ and the current, yield different variances and therefore different results for the shot noise.

F. Two-state model

The "two-state" model is a special case with $N_0 = 1$. This restricts the number of electrons N to be 0 or 1, and could be a model of resonant tunneling through a quantum dot or of Coulomb blockade in the limit where only two states of charge are important.[25] The current has the form shown in Fig. 4, with a strict alternation of g and c hops. This picture can usefully be analyzed in terms of the waiting time between hops. The distribution of these times is analogous to the Ornstein-Zernike direct correlation function (see, for example, Ziman,[24] p. 91).

At low frequency we need only consider, say, c hops. Let t_p be the time between successive c hops (Fig. 4). The noise can be written in terms of statistics of t_p as

$$\hat{s}(\omega = 0) = \frac{\mathrm{var}\,(t_p)}{\langle t_p\rangle^2} = \frac{\langle t_p^2\rangle}{\langle t_p\rangle^2} - 1. \quad (4.48)$$

If there were no correlation between pulses they would arrive on average at a constant rate $\bar{h}$ and t_p would have an exponential distribution,

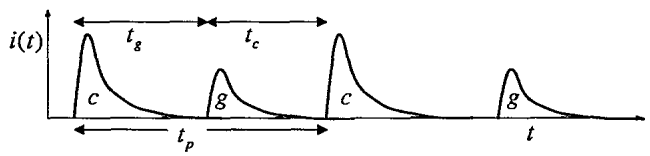

FIG. 4. Current $i(t)$ in the two-state model, consisting of alternating g and c hops of relative weights α and β. The two-state model, which is a special case of the two-process model and can be applied to the Coulomb blockade, can be analyzed in terms of the waiting times between successive pulses t_g and t_c instead of correlation functions.

$$p(t_p) = \bar{h} \exp\left(-\bar{h} t_p\right). \tag{4.49}$$

This has the property that

$$\mathrm{var}\,(t_p) = \langle t_p \rangle^2 = (1/\bar{h})^2, \tag{4.50}$$

for which Eq. (4.48) predicts $\hat{s} = 1$, full shot noise as expected. Shot noise will decrease if the hops repel one another in time. In Fig. 4, with alternating hops, $t_p = t_g + t_c$, where t_g is that time that we have to wait after a c hop for the next g hop, and is exponentially distributed with a time constant τ_e; similarly t_c is exponentially distributed with a time constant τ_c. The distribution of t_p is now given by a convolution of those of t_g and t_c,

$$p(t_p) = \frac{1}{\tau_e - \tau_c}\left[\exp\left(-\frac{t_p}{\tau_e}\right) - \exp\left(-\frac{t_p}{\tau_c}\right)\right]. \tag{4.51}$$

This vanishes as $t_p \to 0$, showing that repulsive correlation has been introduced between successive c hops. The cumulants add when two distributions are convoluted, so Eq. (4.48) becomes

$$\hat{s}(\omega = 0) = \frac{\mathrm{var}\,(t_e) + \mathrm{var}\,(t_c)}{\left(\langle t_e \rangle + \langle t_c \rangle\right)^2} = \frac{\tau_e^2 + \tau_c^2}{\left(\tau_e + \tau_c\right)^2}, \tag{4.52}$$

in agreement with the earlier result, Eq. (4.38). This can be rewritten in terms of the fractional occupation of the state, $f = \tau_c/(\tau_e + \tau_c)$, as $\hat{s} = 1 - 2f(1-f)$, showing that the shot-noise ratio reaches a minimum of $\frac{1}{2}$ when the two states are equally occupied on average.

These results can be applied to the Coulomb blockade in a highly asymmetric double tunnel junction. This is a highly dissipative system, which prevents electrons hopping backwards. In general, the Coulomb blockade is much more difficult to solve because of the different boundary conditions: N_0 is not an integer, and the extrapolated forms of the appropriate rates $g(N)$ and $r(N)$ do not vanish automatically at the extreme values of N, although they are linear. Fortunately, there is a regime of applied biases for which only two charge states are energetically allowed at low temperatures. In this so-called two-state limit[25] the above expression can be used with time constants τ_e and τ_c that depend on voltage. There is a rich structure in the noise because the relative magnitudes of τ_e and τ_c change with voltage. The maximum reduction in the shot noise, to $\hat{s} = \frac{1}{2}$, occurs when $\tau_e = \tau_c$. This case and others are discussed in detail in Ref. 26.

V. APPLICATION TO DIFFERENT MODELS

In the preceding sections we have treated the simplest model of resonant tunneling for electrons, where the states in the emitter are assumed to be full. This suppresses the return to electrons to the emitter from the resonant state, and there are therefore only two processes in the master equation. In general the third process of particles returning to the emitter should be included. This modifies the master equation while remaining consistent with the rate equation (3.1). We shall investigate the effect of this additional process on the shot noise after pursuing the simpler model a little further.

A. Two-process model

Electrons always travel forward in the two-process model, so there is only one physical rate transferring electrons between the emitter and the resonant state. This model gave suppressed shot noise, a result identical to coherent quantum mechanics. Other results agree similarly, such as the variance of the number of electrons in the resonant state. Consider $N_0 = 1$, the two-state model. For the quantum-mechanical case we can use the identity $\hat{n}^2 = \hat{n}$ for Fermion operators, which immediately shows that

$$\langle \mathrm{var}(n) \rangle = \langle n^2 \rangle - \langle n \rangle^2 = \langle n \rangle \left(1 - \langle n \rangle\right). \tag{5.1}$$

Each state with a different transverse wave vector can be treated independently in the coherent case, because there is no scattering between them, so Eq. (5.1) can simply be summed over all these states. The result is exactly the same as Eq. (3.15) for the two-process classical model. It is also possible to derive a master equation for the quantum-mechanical two-state model[27] which is the same as the classical one. Thus the master equation must give the same result as the coherent quantum mechanics even though it has a wider range of validity, and may hold even in the presence of inelastic scattering.

B. More-general models

As discussed above, a general master equation for resonant tunneling should include (at least) three rates, but the return from the resonant state to the emitter can be neglected in the case of electrons (Fermions) if the states in the emitter are fully occupied. This is not true for either classical particles or bosons.

To understand the role of statistics we consider a simple example. Take a one-dimensional system where N_0 is the average occupation of the state in the emitter. In general N_0 will depend on the temperature and density of particles in the emitter; here we simply treat it as a parameter. In another paper[28] we fixed the density in the emitter and allowed the chemical potential to vary with temperature. In the Bose case, for example, there is a tendency torwards Bose condensation at low temperatures, and N_0 can go either to zero or to infinity depending on the position of the resonance. For fermions, on the other hand, N_0 is fixed to lie between 0 and 1, although it still depends on temperature. Electrons can return to the emitter from the resonant state (e hops) at a rate $N(1 - N_0)/\tau_e$ if $N_0 < 1$, while the rate from the emitter to the resonant state (g hops) is $N_0(1 - N)/\tau_e$; the rate of exit to the collector (c hops) remains N/τ_c. The rates for bosons have + signs instead of − signs, while those for classical particles lack the factors containing the occupation of the final state. They are shown in Table I, and are all consistent with the original rate equation (3.1). The master equation now contains three physical rates, but the e and c hops can be absorbed into a composite recombination rate, $r(N) = e(N) + c(N)$. Solving

TABLE I. Summary of the results of the master equations for one-dimensional resonant tunneling of Fermi, classical, and Bose particles. All give the same mean number of particles in the resonant state, $\langle N\rangle = N_0\tau_c/(\tau_e+\tau_c)$. The variance of N increases on going from Fermi to classical to Bose particles. The shot noise $\hat{s}(\omega=0)$ is consequently suppressed for fermions but enhanced for bosons; it is expressed in terms of the peak transmission coefficient for the quantum-mechanical model, $T_{\rm pk} = 4T_eT_c/(T_e+T_c)^2 = 4\tau_e\tau_c/(\tau_e+\tau_c)^2$.

	fermions	classical	bosons
$g(N)$	$\frac{N_L(1-N)}{\tau_e}$	$\frac{N_L}{\tau_e}$	$\frac{N_L(1+N)}{\tau_e}$
$e(N)$	$\frac{N(1-N_L)}{\tau_e}$	$\frac{N}{\tau_e}$	$\frac{N(1+N_L)}{\tau_e}$
$c(N)$	$\frac{N}{\tau_c}$	$\frac{N}{\tau_c}$	$\frac{N}{\tau_c}$
$p(N)$	binomial	Poisson	geometric
var(N)	$\langle N\rangle(1-\langle N\rangle)$	$\langle N\rangle$	$\langle N\rangle(1+\langle N\rangle)$
$\hat{s}(\omega=0)$	$1-\frac{1}{2}N_0T_{\rm pk}$	1	$1+\frac{1}{2}N_0T_{\rm pk}$

the master equation in a steady state gives the distribution function $p(N)$ and the variance. The number of particles N cannot exceed 1 for fermions, but there is no upper limit for classical or Bose particles. This behavior is reflected in their larger variances. The noise at low frequencies can then be found using Eq. (4.47). The results are summarized in Table I. The suppression of shot noise for fermions is reduced if the emitter is not fully occupied, in agreement with the quantum-mechanical results plotted in Fig. 2. The distribution of N is Poisson for classical particles, in which case var$(N) = \langle N\rangle$ and there is full classical shot noise as expected. For bosons, enhanced shot noise is found, as a result of the greater variance.[16,28,29] Note that the shot noise is proportional to N_0 which can be arbitrarily large. This can be understood from the form of the rates, which increase as the occupation of the final state increases; this in turn leads to bunching of bosons unlike the repulsion found in the correlation functions for fermions (Sec. IV C).

We would now like to contrast these results to our earlier calculation[28] of the noise for a structureless transmission barrier. In that case we found that the noise for bosons diverges as one approaches zero temperature because of the incipient Bose condensation. Perhaps more surprising, the noise *decreases* with increasing bias at low temperatures because the condensation is suppressed. Thus, although the noise for a finite bias is larger for bosons than for fermions, it is still smaller than at zero bias (thermal noise). In the present calculation with a narrow resonance the noise for bosons is still larger than the fermion noise, but depends on voltage and temperature in a different way. For a fixed bias larger than the temperature, the noise ratio $\hat{s} = 1+\frac{1}{2}N_0T_{\rm pk}$ can decrease with temperature to the classical value because N_0 may go to 0. For a fixed temperature again small compared with the voltage, increasing the voltage may move the resonance into a region of larger occupation (N_0), *increasing* the noise. On the other hand, the noise may be decreased because the bias may affect the incipient Bose condensation. Thus the two cases, a structureless barrier and a narrow resonace, can show very different behavior for the noise at low temperatures.

VI. CONCLUSIONS

We have shown that suppressed shot noise can be derived within a "classical" description of resonant tunneling based on the rate equation for sequential tunneling. It needs to be derived from a master equation, and the different master equations that are consistent with the rate equation can give significantly different results for the noise, although the average current is the same in all cases. The results from a master equation for fermions are identical to those for "coherent" quantum mechanics. Thus inelastic scattering can leave the noise in resonant tunneling unchanged under some conditions.

There are many other processes that need to be included in a more complete model of the experimental devices. An important step would be to include scattering by phonons explicitly. Polar optic phonons usually provide the strongest electron-phonon scattering in the III-V semiconductors. The Fermi energy was less than the energy of optic phonons $\hbar\Omega_0$ in the devices used by Li *et al.*,[5] so the electrons are unable to emit optic phonons strongly. This would change if the Fermi energy were raised above $\hbar\Omega_0$. Different material might be required, as in the devices based on $In_xGa_{1-x}As$ used by Woodward *et al.*[30] Experiments on a series of such devices would provide a valuable test of the effect of scattering on shot noise in resonant tunneling. Our model omits the effect of scattering in the contacts, the emitter in particular. This is usually a three-dimensional electron gas and might be expected to have faster scattering than the resonant state, which is only two dimensional. The effect of scattering in the contacts has not been investigated for the current, let alone for the noise. There is also space-charge feedback.[31] If the density of electrons in the resonant state rises, self-consistency causes its energy level to rise too. This in turn decreases the effective Fermi energy and thus the inflow of electrons from the emitter. The same effect leads to the bistability seen in the experiments. Clearly it is another mechanism that suppresses fluctuations of density and current. It depends on the capacitance between the resonant state and the contacts, not just on the transmission coefficients. It should be possible to include all of these effects in more sophisticated master equations and hence study their effect on the noise. We have already applied the techniques described here to the classical double junction Coulomb blockade problem.[26]

ACKNOWLEDGMENTS

We would like to thank H. U. Baranger, A. J. F. Levi, and C. J. Stanton for stimulating discussions. This work was supported primarily by the Office of Naval Research and additionally by the NSF through Grant No. PHY89-04035 to the Institute for Theoretical Physics, Santa Barbara, where part of this work was carried out. P.H. gratefully acknowledges the additional support of the Danish Research Academy.

*Permanent address: Department of Electronics and Electrical Engineering, Glasgow University, Glasgow G12 8QQ, U.K.
[1]L. L. Chang, L. Esaki, and R. Tsu, Appl. Phys. Lett. **24**, 593 (1974).
[2]S. Luryi, Appl. Phys. Lett. **47**, 490 (1985).
[3]M. C. Payne, J. Phys. C **19**, 1145 (1986).
[4]T. Weil and B. Vinter, Appl. Phys. Lett. **50**, 1281 (1987).
[5]Y. P. Li, A. Zaslavsky, D. C. Tsui, M. Santos, and M. Shayegan, Phys. Rev. B **41**, 8388 (1990).
[6]G. B. Lesovik, Pis'ma Zh. Eksp. Teor. Fiz. **49**, 513 (1989) [JETP Lett. **49**, 592 (1989)].
[7]B. Yurke and G. P. Kochanski, Phys. Rev. B **41**, 8184 (1990).
[8]M. Büttiker, Phys. Rev. Lett. **65**, 2901 (1990).
[9]R. Landauer and T. Martin, Physica B **175**, 167 (1991).
[10]L. Y. Chen and C. S. Ting, Phys. Rev. B **43**, 4534 (1991).
[11]J. H. Davies, S. Hershfield, P. Hyldgaard, and J. W. Wilkins (unpublished).
[12]I. O. Kulik and R. I. Shekhter, Zh. Eksp. Teor. Fiz. **68**, 623 (1975) [Sov. Phys. JETP **41**, 308 (1975)].
[13]C. W. J. Beenakker and H. van Houten, Phys. Rev. B **43**, 12 066 (1991).
[14]A. van der Ziel, *Noise in Solid State Devices and Circuits* (Wiley, New York, 1986).
[15]T. G. van der Roer, H. C. Heyker, and J. J. M. Kwaspen, Electron. Lett. **27**, 2158 (1991).
[16]T. Martin and R. Landauer, Phys. Rev. B **45**, 1742 (1992).
[17]R. Tsu and L. Esaki, Appl. Phys. Lett. **22**, 562 (1973).
[18]A. D. Stone and P. A. Lee, Phys. Rev. Lett. **54**, 1196 (1985).
[19]M. Jonson and A. Grincwajg, Appl. Phys. Lett. **51**, 1729 (1987).
[20]Y. Hu, J. Phys. C **21**, L23 (1988).
[21]R. Landauer, J. Appl. Phys. **33**, 2209 (1962).
[22]S. Ramo, Proc. IRE **27**, 584 (1939).
[23]W. Shockley, J. Appl. Phys. **9**, 639 (1938).
[24]J. M. Ziman, *Models of Disorder* (Cambridge University Press, Cambridge, 1979).
[25]J.-C. Wan, K. A. McGreer, L. I. Glazman, A. M. Goldman, and R. I. Shekhter, Phys. Rev. B **43**, 9381 (1991).
[26]S. Hershfield, J. H. Davies, P. Hyldgaard, C. J. Stanton, and J. W. Wilkins (unpublished).
[27]L. I. Glazman and K. A. Matveev, Pis'ma Zh. Eksp. Teor. Fiz. **48**, 403 (1988) [JETP Lett. **48**, 445 (1988)].
[28]J. W. Wilkins, S. Hershfield, J. H. Davies, P. Hyldgaard, and C. J. Stanton, Phys. Scr. (to be published).
[29]M. Büttiker, Physica B **175**, 199 (1991).
[30]T. K. Woodward, D. S. Chemla, I. Bar-Joseph, H. U. Baranger, D. L. Sivco, and A. Y. Cho, Phys. Rev. B **44**, 1353 (1991).
[31]J. Han and F. S. Barnes, IEEE Trans. Electron Devices **38**, 237 (1991).