Numerically exact full counting statistics of the nonequilibrium Anderson impurity model
Michael Ridley, Viveka N. Singh, Emanuel Gull, and Guy Cohen
Phys. Rev. B 97, 115109 — Published 6 March 2018
DOI: 10.1103/PhysRevB.97.115109
The time dependent full counting statistics of charge transport through an interacting quantum junction is evaluated from its generating function, controllably computed with the inchworm Monte Carlo method. Exact noninteracting results are reproduced; then, we continue to explore the effect of electron–electron interactions on the time-dependent charge cumulants, first-passage time distributions and \( n \)-electron transfer distributions. We observe a crossover in the noise from Coulomb blockade- to Kondo-dominated physics as the temperature is decreased. In addition, we uncover long-tailed spin distributions in the Kondo regime and analyze queuing behavior caused by correlations between single electron transfer events.

**I. INTRODUCTION**

Mesoscopic quantum dots, molecules in junctions and other small quantum systems coupled to baths are often studied in transport experiments. It is possible to measure not only the electronic current through such systems, but also its noise and higher order moments. The complete set of moments determines the full counting statistics (FCS) of the system, which grants direct access to otherwise concealed properties: for example, the entanglement entropy, the fidelity, and thermoelectric efficiency fluctuations.

Apart from special limits, theoretical modeling of FCS in fermionic systems has so far been restricted to either noninteracting situations, or approximations whose accuracy is difficult to gauge. This difficulty is exacerbated in the experimentally relevant cases where strong electron–electron correlations and nonequilibrium effects are in play. A controlled and fully quantum description of the current and its moments is therefore sorely needed.

In this paper, we provide such a description for a model of interacting fermions in a junction, within a numerically exact inchworm quantum Monte Carlo (iQMC) calculation. By computing the generating function of lead population change (Fig. 1) we gain access to all moments of population transfer, first passage time distributions (FPTDs) and \( n \)-electron transfer distributions. We analyze Fano factors and consider the shapes of the distributions, which provide remarkably detailed information about the dynamics of single tunneling events. We extract indications of queuing effects in the presence of strong interactions.

The paper proceeds as follows. Section II will motivate the full counting statistics, and Section III provide a brief overview of the main definitions and concepts used in FCS. In Section IV we describe how FCS can be obtained within iQMC, specializing to a particular model. Next, we present results for non-interacting (Sec. V) and interacting (Sec. VI) systems. The effect of interactions, temperature and voltage on the first and second charge cumulants is discussed, and we further explore how first passage times and \( n \)-particle passage probabilities depend on the initial preparation. Finally, we discuss our conclusions in Section VII.

**II. BACKGROUND AND MOTIVATION**

The transfer of charges across a small system, such as a molecular electronic junction or a mesoscopic quantum dot, is fundamentally a stochastic process. Like any stochastic process, it is fully characterized by time dependent probability distributions \([2]\. In this case, \( P(\Delta n, t) \) where \( \Delta n \) is the number of charges transferred by time \( t \) \([3]\. Fluctuations in the electron current have long been known to provide experimentally accessible information not contained in the expectation value of the current \([1, 4]\. Such as dwell times \([5\] and transmission probabilities \([6\].

Current noise, the second moment of the current, has been of interest as a way to investigate reflection pro-
cesses and traversal times in molecular junctions [1-10] and propagation of correlated electron–hole pairs in photon-assisted transport [11–13]. A great deal of work has focused on fluctuations in periodically driven systems, including the prediction of noiseless Lorentzian pulses, or “levitons”. Quantum noise has also been used to extract effective quasiparticle charges \( e^* \) of transmitted electrons due to Cooper pair formation in superconducting junctions \( (e^* = 2e) \) [17,19], quantum Hall effect \( (e^* = e/3) \) [20], or two-particle inelastic backscattering processes in the Kondo regime \( (e^* = 5e/3) \) [21,26].

The noise-to-signal ratio, also known as the Fano factor \( F \), can be used to characterize charge transfer statistics. It takes the value \( F = 1 \) when the charge transfer is Poissonian (completely uncorrelated), as generally occurs for small transmission probabilities [27] and weak coupling between the molecule and bath [28]. Super-Poissonian shot noise \( (F > 1) \) has been observed in single [29] and double [30, 31] quantum dot junctions and is generally associated with strong electron bunching and quasiparticle formation [1] that may occur in the transient [32] or steady state [33, 34] regimes. In addition, the Fano factor can be tuned so that it is sub-Poissonian \( (F < 1) \) in quantum dot [35, 36] and graphene nanoribbon [37] structures.

Instead of focusing only on the second moment of the current, it is possible to consider the full counting statistics (FCS) associated with the individual charge transfer events from an underlying moment-generating or characteristic function [2], and recover all higher-order moments and cumulants. This idea was pioneered by Levitov and Lesovik [14, 38], who evaluated the steady state FCS for noninteracting systems.

The FCS formalism for charge can be connected to entanglement entropy [39] and fidelity [40] in noninteracting systems, and was also generalized to a heat and work FCS giving access to efficiency fluctuations in thermoelectric junctions [39, 40]. Parallel with the development of the field of spintronics, the FCS of spin currents has been developed in regular metallic lead junctions [41–44] and magnetic tunnel junctions (MTJs) [45–47], and can be applied, for example, to the detection of spin-singlet pairs in the electron counting statistics [42, 48]. In the steady state, the second cumulant can be used to compute the zero-frequency current power spectrum [49, 50], and in recent years a theory of FCS has been proposed which also gives access to frequency-dependent noise spectra [44, 51, 52]. Theoretical [53–55] and experimental [56] investigations have demonstrated that the generating function satisfies steady state fluctuation–dissipation relations equivalent to the condition for detailed balance, although recent work indicates violation of detailed balance when bound states are established with circular probability currents in the steady state [57] and modification of detailed balance when the system is undergoing transient evolution [58]. It has also been shown that for noninteracting electrons, the zeros of the generating function all lie on the negative real axis [59, 61], so that the position of zeros in the complex plane can shed light on interaction effects in the transport [62].

Recently, the time dependence of higher-order cumulants has become accessible to experiment [63, 64], motivating the study of transient cumulant generating functions and related objects which reveal more information on dynamics than the current or particle number. One such quantity, which is closely related to the generating function, is the waiting time distribution (WTD). The WTD is the probability distribution characterizing the time separating successive charge detection events [65, 67]. Experimentally, WTDs can be used to study the effect of coherence [68] and complex internal molecular structures such as double quantum dot [69, 72] or normal–superconducting junctions exhibiting Andreev reflection [73, 74], as these effects can alter the traversal times for electrons crossing the system. At the level of practical quantum electronics, these dynamical effects determine the maximum operation rate for devices [70].

Despite this flurry of theoretical activity, computing the FCS for generic interacting fermionic systems has so far proven to be challenging. In noninteracting systems, there now exist robust approaches to the study of both steady state and transient FCS, including an appealing coherent state path integral nonequilibrium Green’s function formalism (PI-NEGF) [45, 77, 79] which can be perturbatively extended to interacting cases [79, 81]. Exact results are available for steady state FCS and noise characteristics of the AIM in the Fermi liquid regime and at the Toulouse point [22, 23, 82–85]. The Fermi liquid theory is, however, valid only for low voltages and temperatures [21, 84, 88–94]. Several approximate approaches to FCS in interacting systems have been successfully employed in appropriate regimes: a noncomprehensive list includes various perturbative expansions [29, 48, 50, 53, 83, 88–94] and quantum master equation approaches [51, 54, 57, 66, 68, 85–101]. Nonperturbative numerical approaches also exist, but for models which do not capture the full complexity of interacting fermionic transport. These include hierarchical equations of motion in the spin-boson model [102] at high temperatures, and the density matrix renormalization group for the interacting resonant level model at its self dual point and at zero temperature [103, 105]. Also, an equilibrium determinant QMC method was recently applied to cold atomic gases [106].

To date, no numerically exact method has accessed FCS in a generic nonequilibrium interacting fermionic model; nor is a method currently available which is (even in principle) suitable for arbitrary temperatures, bias voltages and interaction strengths. In fact, this is largely true even for the calculation of the second cumulant, or noise. In this paper, we provide such a numerically exact method for computing the FCS, based upon the recently developed inchworm diagrammatic quantum Monte Carlo (iQMC) method [107].
III. FCS: THEORY AND GENERAL CONSIDERATIONS

In a charge detection experiment, one studies a system in which some regions, labeled \( \ell \), are called “leads”. Lead \( \ell \) is found to contain \( n_\ell(t) \) electrons on any given measurement of its total population at time \( t \). The change in charge \( \Delta n_\ell = n_\ell(t) - n_\ell(t_0) \) on the lead is measured between some initial time \( t_0 \) and the measurement time \( t > t_0 \). The FCS formalism aims to construct a generating function \( Z(\lambda, t) \) for the probability distribution characterizing \( \Delta n_\ell \), determined by the associated operator \( \Delta N_\ell \) (we suppress lead index \( \ell \) from now on). For this purpose, an auxiliary counting field \( \lambda \) is introduced, such that

\[
Z(\lambda, t) \equiv \langle e^{i\lambda \Delta N} \rangle = \sum_{\Delta n \in \mathbb{Z}} P(\Delta n, t) e^{i\lambda \Delta n}. \tag{1}
\]

Here, \( P(\Delta n, t) \) is the probability that the measured number of electrons has changed by the integer \( \Delta n \) after time \( t \). The distributions \( P(\Delta n, t) \) can be obtained from the generating function via

\[
P(\Delta n, t) = \int_{-\pi}^{\pi} d\lambda \, Z(\lambda, t) e^{-i\lambda \Delta n}, \tag{2}
\]

and all moments \( M_k(t) \) and cumulants \( C_k(t) \) can be extracted by taking successive derivatives with respect to \( \lambda \):

\[
M_k(t) \equiv \left\langle \left( \Delta N \right)^k \right\rangle = \lim_{\lambda \to 0} \frac{\partial^k Z(\lambda, t)}{\partial (i\lambda)^k}, \tag{3}
\]

\[
C_k(t) \equiv \lim_{\lambda \to 0} \frac{\partial^k \ln Z(\lambda, t)}{\partial (i\lambda)^k}. \tag{4}
\]

The \( M_k \) and \( C_k \) contain equivalent information; in particular we note that \( C_1 = M_1 \) and \( C_2 = M_2 - M_1^2 \) relate the first and second moments. Of particular interest \( \Pi(t) \) is the probability distribution for no charges passing across the junction, which defines the idle time probability \( \Pi(t) \):

\[
\Pi(t) \equiv P(0, t) = \int_{-\pi}^{\pi} d\lambda \, Z(\lambda, t). \tag{5}
\]

This can be used to extract the first passage time distribution \( W(t) \):

\[
W(t) = -\frac{d\Pi(t)}{dt}, \tag{6}
\]

which describes the probability distribution of the time separating initialization of the counting experiment from the first detection of a charge transfer event \( t_0 \).

The counting statistics of any operator \( \hat{A}(t) \) can be evaluated by way of an effective Hamiltonian transformation originally formulated by Levitov and Lesovik \( \text{[39]} \). We will primarily be interested in the special case \( \hat{A}(t) = \hat{N}_L \). The generating function for the moments of a change \( \Delta \alpha \) in the value associated with \( \Delta \hat{A}(t) \) is given by \( \text{[49]} \)

\[
Z(\lambda, t) = \text{Tr} \left[ \hat{\rho}_0 \hat{U}_t^{-\lambda} (t, t_0) \hat{U}_t^\lambda (t, t_0) \right], \tag{7}
\]

in terms of the modified, operator-specific propagator \( \hat{U}_\gamma(t, t_0) \), which is defined by \( \text{[49]} \)

\[
\hat{U}_\gamma(t, t_0) \equiv e^{i\gamma \hat{A}(t)} \hat{U}(t, t_0) e^{-i\gamma \hat{A}(0)}. \tag{8}
\]

The modified propagator itself can be written as a time-ordered integral with respect to a modified Hamiltonian:

\[
\hat{U}_\gamma(t, t_0) = \text{Tr} \left\{ \exp \left[ -\frac{i}{\hbar} \int_{t_0}^{t} d\tau \hat{H}_\gamma(\tau) \right] \right\}, \tag{9}
\]

where

\[
\hat{H}_\gamma(t) = e^{i\gamma \hat{A}(t)} \hat{H}(t) e^{-i\gamma \hat{A}(0)} - \hbar \gamma \partial_t \hat{A}(t). \tag{10}
\]

This satisfies the relation \( i\hbar \partial_t \hat{U}_\gamma(t, t_0) = \hat{H}_\gamma(t) \hat{U}_\gamma(t, t_0) \). We express the generating function of Eq. \( \text{(7)} \) as a time ordered integral over the full Keldysh contour \( C = (C_+, C_-) \), composed of an “upper” branch \( C_+ \) and a “lower” branch \( C_- \). Using the contour representation, \( Z(\lambda, t) \) is given by the compact form

\[
Z(\lambda, t) = \text{Tr} \left\{ \hat{\rho}_0 \hat{T}_C \left\{ \exp \left[ -\frac{i}{\hbar} \int_C dz \hat{H}_\gamma(z) \right] \right\} \right\}, \tag{11}
\]

where \( \hat{T}_C \) orders times later on the contour to the left.

The auxiliary field \( \gamma \) takes a different value on each branch:

\[
\gamma(z) = \begin{cases} 
\frac{\lambda}{2}, & z \in C_+, \\
-\frac{\lambda}{2}, & z \in C_-.
\end{cases} \tag{12}
\]

When \( \lambda = 0 \), it is immediately apparent that the generating function \( Z(\lambda, t) = 1 \) for all \( t \). In general, \( Z(\lambda, t) \) is periodic in \( \lambda \) with a periodicity of \( 2\pi \). Any value of \( \lambda \neq 2\pi j \) for \( j \in \mathbb{Z} \) means that the Hamiltonian depends on the contour branch, such that time reversal symmetry is explicitly broken.

IV. MODEL AND MONTE CARLO ALGORITHM

A. Model and observables

We now specialize the discussion to the concrete case of the nonequilibrium Anderson impurity model (AIM) \( \text{[109]} \). The model’s Hamiltonian can be written as follows:

\[
\hat{H} = \hat{H}_D + \hat{H}_B + \hat{H}_V. \tag{13}
\]
\( \hat{H}_D \) describes a small, interacting “dot” subsystem; \( \hat{H}_B \) depicts a set of large, noninteracting lead subsystems; and \( \hat{H}_V \) is a hybridization Hamiltonian coupling the dot and the leads. These three terms take the following form:

\[
\begin{align*}
\hat{H}_D &= \sum_{\sigma} \varepsilon_{\sigma} \hat{d}_\sigma^\dagger \hat{d}_\sigma + U \hat{d}_\uparrow^\dagger \hat{d}_\uparrow \hat{d}_\downarrow^\dagger \hat{d}_\downarrow, \\
\hat{H}_B &= \sum_{k,\sigma} \varepsilon_{k,\sigma} \hat{a}_{k,\sigma}^\dagger \hat{a}_{k,\sigma}, \\
\hat{H}_V &= \sum_{k,\sigma,\ell} \left( V_{k,\sigma,\ell} \hat{a}_{k,\sigma}^\dagger \hat{d}_{\sigma} + \text{H.c.} \right).
\end{align*}
\]

Here, the \( \hat{d}_\sigma \) denote dot annihilation operators labeled by a spin index \( \sigma \), while the \( \hat{a}_{k,\sigma} \) operators signify lead annihilation operators labeled by a spatial index \( k \), spin index \( \sigma \) and lead index \( \ell \). We assume two leads, denoted \( \ell = L, R \). The \( \varepsilon_{\sigma} \) represent dot level energies, and the charging energy \( U \) determines the strength of Coulomb repulsion between electrons. In this paper, we set \( \varepsilon_{\sigma} = -U/2 \) (the particle–hole symmetric case, to which our method is not in any way limited). The hybridization parameters \( V_{k,\sigma,\ell} \) couple the lead states to the dot and the \( \varepsilon_{k,\sigma,\ell} \) enumerate the energies of lead states. The system is initially prepared in the state \( |\uparrow\downarrow\rangle \); the singly occupied spin up and spin down down states, \(|\sigma\rangle \), with \( \sigma = \uparrow, \downarrow \); and the doubly occupied dot state \(|\uparrow\downarrow\rangle \). These states will be denoted by the label \( \phi \) in what follows.

The \( \varepsilon_{k,\sigma,\ell} \) and \( V_{k,\sigma,\ell} \) are completely determined by the coupling density

\[
\Gamma_{\ell}(\omega) = \pi \sum_{k} |V_{k,\ell}|^2 \delta(\omega - \varepsilon_{k,\sigma,\ell}).
\]

We choose this to be a flat band with a smooth cut-off [110]:

\[
\Gamma_{\ell}(\omega) = \frac{\Gamma_{\ell}}{\left(1 + e^{\nu(\omega - \Omega_{\ell})}ight) \left(1 + e^{-\nu(\omega + \Omega_{\ell})}\right)}.
\]

In what follows, we set \( \Gamma_{\ell} = \frac{1}{4} \) such that \( \Gamma = \sum_{\ell} \Gamma_{\ell} = 1 \) determines our unit of energy. We take the leads’ band cutoff \( \Omega_{\ell} \) to be 10\( \Gamma \), and their edge width \( \frac{\Gamma}{2} \) to be 0.1\( \Gamma \).

At the initial time \( t_0 = 0 \) of the simulation, a coupling quench is performed and the subsystems are suddenly coupled by the introduction of \( \hat{H}_V \) into the Hamiltonian. This kind of switch-on is often referred to as the partitioned approach, and has been contrasted in the quantum transport literature to partition-free approaches (such as a voltage quench) in which the dot–lead coupling is established before a bias is added to the lead energies [111]. We note in passing that a partition-free voltage quench has also been explored within inchworm QMC [112], by augmenting the finite Keldysh branches with an imaginary time Matsubara branch, thus obtaining the full Konstantinov–Perel’ contour [113].

In the present context, we are interested in the FCS for this model. We therefore set \( A(t) \rightarrow \hat{N}_L \) in the modified Hamiltonian of Eq. (10), where \( \hat{N}_L \equiv \sum_{k,\sigma} \hat{a}_{k,\sigma,L}^\dagger \hat{a}_{k,\sigma,L} \) is the particle number operator in lead \( L \). This counting function describes the moments of population changes (rather than currents), but also provides access to the mean time-dependent current flowing out of the lead, \( I_L(t) = \langle d\hat{N}_L(t)/dt \rangle \), by way of a time derivative. The current noise and higher order cumulants can only be accessed at steady state. For example, the noise

\[
S_{LL}(\omega = 0) \equiv \lim_{t \to \infty} \int d\tau \left\langle \Delta \hat{I}_L(t + \tau) \Delta \hat{I}_L(t) \right\rangle,
\]

where \( \Delta \hat{I}_L = I_L - \langle I_L \rangle \), is related to the long time limit of the second cumulant [49]. We can therefore write:

\[
\begin{align*}
\lim_{t \to \infty} I_L(t) &= \lim_{t \to \infty} \frac{C_2(t)}{t}, \\
S_{LL}(\omega = 0) &= \lim_{t \to \infty} \frac{C_2(t)}{t}.
\end{align*}
\]

From the first and second cumulants, one can define the time-dependent Fano factor for the population [114]

\[
F(t) = \frac{C_2(t)}{C_1(t)}.
\]

In the long-time limit, this converges to the Fano factor (up to a scaling factor of \( 2q \)) [5, 34] for the current,

\[
\frac{S_{LL}}{2qI_L} = \frac{S_{LL}}{S_P} = \lim_{t \to \infty} \frac{F(t)}{2q}.
\]

This is the Fano factor usually observed in steady state transport experiments, where \( q \) is the charge of an individual carrier and \( S_P = 2qI_L \) is the Poisson value of the shot noise [11, 5, 29].

### B. FCS with inchworm Monte Carlo

Within the modified Hamiltonian picture of Sec. [111], the problem of evaluating a generating function can be mapped onto that of evaluating the time dependence of a unit operator propagated by the modified Hamiltonian. For the AIM with \( \hat{N}_L \) as the counting field, using the Baker–Hausdorff lemma and the fact that \( \hat{N}_L \) commutes with \( \hat{H}_D \) one can show [115] that the modified Hamiltonian of Eq. (10) is equivalent to the AIM, but with lead–molecule coupling terms that acquire a branch-dependent phase:

\[
\hat{H}_0' = e^{i\gamma(z)\hat{N}_L} \hat{H} e^{-i\gamma(z)\hat{N}_L} = \hat{H} \left( \{V_{k,\ell}(z; \lambda)\},
\right).
\]

Hence, the modified Hamiltonian is still a Trotter factor of independent AIM Hamiltonians, each of which is given by Eq. (10). For a perceptron Hamiltonian, each of which is given by Eq. (10). For a perceptron Hamiltonian, each of which is given by Eq. (10).
Figure 2. Diagrammatic expansion of the full propagator (upper half of the panel) and the corresponding term in the propagator with contour branch indices exchanged. Thick lines denote a full propagator, thinner lines a bare propagator, wiggly lines a pair of hybridization events, and the ‘X’ the fold in the Keldysh contour. The elements marked in red are those modified by the counting field, resulting in the breaking of time-reversal symmetry for \( \lambda \neq 0 \).

\[
V_{k\sigma,\ell} (z; \lambda) = \begin{cases} 
V_{k\sigma,\ell} e^{i \frac{\pi}{2} \delta_{kL}}, & z \in C_+,
V_{k\sigma,\ell} e^{-i \frac{\pi}{2} \delta_{kL}}, & z \in C_-.
\end{cases}
\]  

(25) The generating function in Eq. (11) may then be evaluated by a standard Keldysh hybridization expansion [110–113, 116]. Whereas usually in such expansions an operator corresponding to some observable is applied at the fold of the Keldysh contour, here no such operator is needed, and the full counting statistics are given directly in terms of a modified propagator:

\[
Z_\phi (\lambda, t) = p_{\phi\phi'} (t^+, t^-; \lambda).
\]

(26) In this expression, \( p_{\phi\phi'} (z_1, z_2; \lambda) \) is analogous to the dressed restricted propagators used in propagator non-crossing approximations [117,119] in bold-line QMC [120,124], and in inchworm QMC [107,112,119,125,126], but modified to include the auxiliary field via Eq. (25):

\[
p_{\phi\phi'} (z_1, z_2; \lambda) \equiv \langle \phi | \mathrm{Tr}_B \left\{ \rho_0 e^{-i \int z_2^1 H_\lambda (z) dz} \right\} | \phi' \rangle,
\]

(27) When \( p_{\phi\phi'} (z_1, z_2; \lambda) \) is expressed in the interaction picture and expanded in powers of the hybridization, it becomes a weighted sum over configurations \( C \), defined by the set of contour times at which hybridization events (including dot and lead operators) occur. Each configuration sampled at a given diagram order \( n \) corresponds to the insertion of \( n \) hybridization lines connecting \( 2n \) “kink” times on the Keldysh contour where a particle of given spin \( \sigma \) is added to or removed from the dot. We illustrate a diagram of order \( n = 0 \) and another of order \( n = 3 \) in Fig. 2 where the full propagator is represented by a thick black line, the bare atomic propagator by a thin black line and paired hybridization events by wiggly lines.

The full propagator can then be rewritten as a sum over diagrams, which can be stochastically sampled by diagrammatic Monte Carlo algorithms [127]:

\[
p_{\phi\phi'} (z_1, z_2; \lambda) = \sum_{n=0}^{\infty} \sum_C w_{\text{loc}}^{(n)} (C) w_{\text{hyb}}^{(n)} (C; \lambda).
\]

(28) Here the first summation is over the expansion order \( n \) and the second summation is over all configurations \( C = \{ z_1, \ldots, z_{2n} \} \) corresponding to the \( 2n \) kink times at a given order. The \( w_{\text{loc}} \) are products of interacting (but purely local) atomic state propagators:

\[
w_{\text{loc}}^{(n)} (C) = (-i)^{(n_+ - n_-)} \times \prod_{i=0}^{2n-1} p_{\phi_{\lambda_i} \phi_{\lambda_{i+1}}} (z_i, z_{i+1}) ,
\]

(29) Their sign depends on the number of kinks on each branch of the contour, \( n^+ \) and \( n^- = 2n - n^+ \). The \( w_{\text{hyb}} \) denote a determinant given by a sum over all possible lead operator pairings \( X \). Each term in this sum is given by the product of hybridization functions \( \Delta_\lambda = \sum \Delta_\lambda \) at the corresponding set of time pairs, with the sign given by the permutation generating that pairing [127]:

\[
w_{\text{hyb}}^{(n)} (C; \lambda) = \sum_X \text{sign} (X) \prod_{i=0}^{n} \Delta_\lambda (z_i, z_{X_i}) .
\]

(30) These hybridization functions, which are dressed by the counting field in a contour-time dependent manner, must be specified on the \( z_1 \) and \( z_2 \) axes:

\[
\Delta_\lambda (z_1, z_2) = e^{-i \lambda (1 - \delta_{\nu\nu'}) \delta_{kL} \theta_C (z_2 - z_1)} \Delta_\lambda^\nu (z_1, z_2) + e^{i \lambda (1 - \delta_{\nu\nu'}) \delta_{kL} \theta_C (z_1 - z_2)} \Delta_\lambda^{\nu'} (z_1, z_2) .
\]

(31) Here, \( \theta_C \) is the Heaviside function ordering times along the Keldysh contour; \( \nu, \nu' \in \{ C_+, C_- \} \) are the branch indices of \( z_1 \) and \( z_2 \), respectively; and \( L \) is the lead for which FCS is being evaluated. Due to the modified coupling in Eq. (25), the phase factors in the dressed hybridization function Eq. (31) cancel when \( z_1 \) and \( z_2 \) lie on the same branch, such that only hybridization lines that cross the folding time \( t_{\text{max}} \) of the Keldysh contour are modified by the counting field (see the red elements in Fig. 2). The undressed lesser and greater hybridization components,

\[
\Delta_\lambda^\nu (t_1, t_2) = -i \int_{-\infty}^{\infty} \frac{d\omega}{\pi} e^{-i \omega (t_1 - t_2)} \times \Gamma_\nu (\omega) f (\omega - \mu_\nu) ,
\]

(32)
\[
\Delta_\lambda^{\nu'} (t_1, t_2) = i \int_{-\infty}^{\infty} \frac{d\omega}{\pi} e^{-i \omega (t_1 - t_2)} \times \Gamma_{\nu'} (\omega) [1 - f (\omega - \mu_{\nu'})] ,
\]
can be expressed in terms of the level width function $\Gamma_{g}(\omega)$ as defined in Eq. (17), and are parametrized only by the physical times $t_1$ and $t_2$ corresponding to the contour times $z_1$ and $z_2$.

Exact numerical approaches to the investigation of dynamics in quantum many-body systems typically suffer from an exponential scaling of computational cost with time. In real time Monte Carlo methods, this manifests in the dynamical sign problem: an exponential growth of stochastic errors with increasing time. However, the inchworm algorithm, as applied to the real time hybridization expansion [110–116, 121–127, 128], has been shown to bypass the dynamical sign problem in a wide variety of parameters. This was not only shown for the AIM [107, 112], but also in the spin–boson model [119, 125] and—within the dynamical mean field approximation—for lattice models [126].

Here, the inchworm algorithm is used to efficiently sum the expansion described above, by reusing data obtained from propagators on shorter time intervals to construct propagators on longer time intervals [107]. In the top panel of Fig. 3, a snapshot of an intermediate step in the inchworm procedure is illustrated for a branch-independent Hamiltonian without a counting field. Each of the $z_1$ and $z_2$ axes shown consists of an “unfolded” Keldysh contour containing times ordered according to the structure $[t_0^+, ..., t_{max}, ..., t_0^-]$. In each inchworm step, propagators are extended along one time direction (this is called “inching”), marked by the green arrows going to the right. In dark gray, we show time points unneeded for the calculation (as we can always select one contour time to appear later on the contour than the other). The white squares correspond to equal time propagators, which can be evaluated analytically. The blue squares correspond to time arguments for which the propagator is already known from previous steps, while the light red squares correspond to arguments at which the propagator may be evaluated immediately, given the currently known propagators. The dark red squares represent values to be computed at a later steps. In light gray, we show time points which can be obtained from reflection about the $z_1 = -z_2$ axis (marked by a white dashed diagonal line), which can be seen as time reversal symmetry: the propagators on one side of this symmetry element can be obtained from those on the other via the relation

$$p_{\phi\phi'}(z_1, z_2) = p_{\phi'\phi}^*(\tilde{z}_2, \tilde{z}_1).$$

Here, a contour time argument marked by a tilde, $\tilde{z}$, denotes the same physical time as that of $z$, but on the opposite contour.

Due to time reversal symmetry, in the original formulation of the inchworm algorithm the total number of propagators computed is $\frac{1}{4}(\Delta^{\max})^2$, as only one quadrant of the two time plane needs to be explicitly evaluated. Extending a propagator to the right in the quadrant used is identical to extending a symmetrically placed propagator up in the mirrored quadrant, so that time reversal symmetry is also maintained in the inching direction.

In diagrammatic terms, time reversal symmetry is illustrated by in Fig. 2. In the upper part of 2, the cross-branch hybridization line corresponds to a factor of $\Delta^>(t_b, t_a)$ in the propagator, which depends only on physical times and not on the branch indices. When the contour branches of all times appearing in this diagram are flipped, one obtains the expansion in the lower part of Fig. 2. For example, the hybridization factor $\Delta^>(t_a, t_b)$ is replaced by $\Delta^>(t_b, t_a) = -[\Delta^>(t_a, t_b)]^*$. All other factors are similarly conjugated, and so Eq. (33) holds.

By contrast, in the presence of the counting field we evaluate the propagator for nonzero $\lambda$, and the two contour-reflected diagrams are modified as illustrated by the red elements in Fig. 2: the dressed cross-branch hybridization line corresponds to a factor of $\Delta^>(t_a, t_b)$.
branch hybridization factor, denoted by the red wiggly line, is replaced by $e^{i\lambda\Delta^+}(t_a, t_b)$ (upper diagram) and $e^{i\lambda\Delta^-(t_b, t_a)}$ (lower diagram). Due to the conjugated prefactor $e^{i\lambda}$, Eq. (33) is no longer valid, and the two quadrants to either side of the $z_1 = -z_2$ line must be evaluated separately, as illustrated in the lower panel of Fig. 3. Since we no longer mirror the data or enforce the symmetry, an additional spurious numerical breaking of time reversal symmetry due to the asymmetry in inching direction can occur. To avoid this, we inch to each time point from the two possible directions and set the result to the average (illustrated by the green arrows, now going both right and up). This increases the number of simulations by an additional factor of two, so that $(\Delta t)^2$ are needed in total, but the overall quadratic scaling with the time step remains. We have found that bidirectional inching generally provides higher quality data for a given amount of computer time, even without the counting field.

V. BENCHMARK AND VALIDATION FOR THE NON-INTERACTING SYSTEM

A. Noninteracting FCS

In the absence of interactions (when $U = 0$), the FCS following the coupling quench can be exactly obtained by the path integral nonequilibrium Green’s function (PI-NEGF) method, which provides an exact solution in the absence of electron–electron interactions \cite{45}. The generating function for a noninteracting AIM is given by the ratio of two Fredholm determinants:

$$Z(\lambda, t) = \det \left( \tilde{G}^{-1}(z_1, z_2) \right) / \det \left( G^{-1}(z_1, z_2) \right). \quad (34)$$

Here, $G$ is the matrix of two-time Green’s functions for the full molecular junction in Keldysh-rotated space \cite{108,129}, such that the times $z_1, z_2$ correspond to pairs of times on the Keldysh contour with a folding point at $t_{\text{max}} = t$. The Green’s function is given by

$$G^{-1}(z_1, z_2) = G^{-1}(1) - \Delta_L (z_1, z_2) - \Delta_R (z_1, z_2), \quad (35)$$

where $g$ denotes the Green’s function of the isolated dot and $\tilde{G} = G - (\Delta_L - \Delta_L)$ is a modified Green’s function that depends upon the counting field $\lambda$ in the left lead. When expressed in the Keldysh-rotated representation, the two hybridizations are related via

$$\dot{\Delta}_L (z_1, z_2) = \exp \left( i\sigma_x \frac{\lambda}{2} \right) \Delta_L (z_1, z_2) \exp \left( -i\sigma_x \frac{\lambda}{2} \right), \quad (36)$$

where $\sigma_x$ is the Pauli spin matrix. It is easy to show that this is equivalent to Eq. (31) when the Keldysh rotation is reversed. Further details on the implementation of Eq. (34) can be found in Refs. \cite{45,77} and \cite{78}. It was also shown in these works that the expression (34) correctly reduces to the Levitov–Lesosvik formula for the generating function in the long-time limit \cite{14,36}, which can be written as

$$Z(\lambda, t) = \exp \left\{ t \int \frac{d\omega}{2\pi} \sum_{k=1}^{\infty} \left( -1 \right)^{k+1} \frac{1}{k} Tr \left[ T(\omega) \right] \right\} \times \left[ (e^{i\lambda} - 1) (1 - f_R(\omega)) f_L(\omega) \right]^{k}.$$ (37)

Here, $f_{\ell}$ denotes the Fermi function of lead $\ell$ and $T(\omega)$ is the transmission probability for electrons to pass through the molecule.

In this work, the PI-NEGF generating function is computed for the noninteracting case without spin degeneracy, as in Ref. \cite{45}. However, we are interested in the AIM. The $U = 0$ generating function for the AIM can be constructed from that of the single-electron model as a product:

$$Z_{\sigma,\sigma'}^{(\text{AIM})}(\lambda, t) = Z_{\sigma}^{(\text{dot})}(\lambda, t) Z_{\sigma'}^{(\text{dot})}(\lambda, t). \quad (38)$$

The $\sigma, \sigma'$ subscripts denote the two spins. In the cases to be shown here, the system is initially decoupled from the leads with the dot in the unoccupied state $|0\rangle$, such that in the noninteracting calculations, each independent model begins with an unoccupied dot. The dot–bath coupling is then turned on at the quench time $t_0 = 0$.

B. QMC results for the non-interacting system

To verify the correctness of the iQMC method, we begin by performing calculations for the noninteracting case $U = 0$ and comparing the results with exact results obtained from the PI-NEGF method. This formalism provides us with an independent verification of the new method: as the iQMC calculation is based on a hybridization (rather than an interaction) expansion,
the noninteracting case is nontrivial and embodies a rigorous test of the method’s accuracy [115]. In Fig. 4 we present the time evolution of $Z(\lambda, t)$. The two panels on the left are generated from iQMC, while the two panels on the right are exact results. As might be expected, $Z(\lambda = 0, t) = 1$ is satisfied to within the numerical errors. The following symmetry relations are also apparent:

$$\begin{align*}
\text{Re}[Z(\lambda, t)] &= \text{Re}[Z(\pi - \lambda, t)], \\
\text{Im}[Z(\lambda, t)] &= -\text{Im}[Z(\pi - \lambda, t)].
\end{align*}$$

We note that the time evolution of both real and imaginary parts of $Z$ exhibit monotonic decay towards zero after a transient timescale on the order of $1/\Gamma$.

In order to explore the correspondence between the numerical data and the exact result in more detail, we plot a series of cuts across the data in Fig 5. Each curve shown here corresponds to a different value of $\lambda$ in the interval $[-\pi, \pi]$. The results from iQMC are shown in solid red, together with exact data in dashed black. The two sets of curves appear to be in very good agreement, with slight numerical deviations (on the order of 1%) in the iQMC results visible at long times.

The introduction of a finite voltage causes the components of the generating function to strongly oscillate as a function of time, as shown in Fig. 6. This indicates oscillations in cumulants of all orders, reflecting a universal phenomenon predicted by Flindt et al. in Ref. [63] and later observed experimentally [64]. Physically, the introduction of a finite bias enhances the short-time “ringing” behavior, as has previously been observed for the lowest cumulant (the current) [118, 130]. As the FCS in the presence of a voltage bias is substantially richer, it is instructive to consider a detailed comparison of specific cuts again (see Fig. 7). Remarkably, it is observable from this figure (which used less computational resources than Fig 5) that at higher voltages it is easier to converge the numerics. This is a notable property of real time iQMC [122, 131], which makes its regime of efficient applicability different from that of, e.g., low-energy wavefunction methods.

Knowledge of the generating functional provides access to all moments and cumulants in principle, but the numerical exercise of obtaining them in practice may not be trivial. One of the main technical difficulties stems from the fact that within iQMC, the relative error in this data is generally larger at small $\lambda$ values. This is for two reasons: first, part of the statistical iQMC er-
ror is independent of the value of the observable. This source of error is strongly mitigated by the inchworm algorithm, but not completely removed, and results in noisy numerical derivatives. Second, there is a small systematic error due to the time discretization. This error is difficult to fully converge, and can sometimes be significant: for example, consider how the imaginary part of $Z(0, t)$ visibly deviates from its exact value of zero in the lower panel of Fig. 5.

Therefore, if we evaluate $\lambda$ derivatives of $Z(\lambda, t)$ in the limit of small $\lambda$ by a finite difference formula

$$\lim_{\lambda \to 0} \frac{dZ(\lambda, t)}{d\lambda} \approx \frac{Z(\lambda, t) - Z(0, t)}{\Delta \lambda}$$

the value of $\Delta \lambda$ cannot easily be taken to zero without investing unjustified amounts of computer time. Furthermore, the oscillatory form of the function hints that an increasingly small zero in the lower panel of Fig. 5.

However, the dependence of the moments and cumulants on $\Delta \lambda$ can be understood at long times by considering the Levitov–Lesovik formula for the current cumulant generating function $S(\lambda) \equiv \lim_{t \to \infty} \frac{\ln Z(\lambda, t)}{t}$, with $Z(\lambda, t)$ given by Eq. (37). If this is inserted into the finite difference derivative with respect to a small shift $\Delta \lambda$, and the integrand is expanded to linear order in $\Delta \lambda$, one arrives at the following expressions for the logarithmic and direct derivatives:

$$\frac{d \ln Z(\lambda, t)}{d (i\lambda)} \bigg|_{\lambda=0} = \lim_{\Delta \lambda \to 0} \frac{1}{\Delta \lambda} \int \frac{d \omega}{2\pi} \left. \right| \left. \begin{array}{l} \times \text{Tr} \left[ T(\omega) (f_L(\omega) - f_R(\omega)) \right] \end{array} \right. ,$$

$$\frac{dZ(\lambda, t)}{d (i\lambda)} \bigg|_{\lambda=0} = \frac{1}{\Delta \lambda} \int \frac{d \omega}{2\pi} \left. \right| \left. \begin{array}{l} \times \exp \left\{ i \Delta \lambda \int \frac{d \omega}{2\pi} \text{Tr} \left[ T(\omega) \right] \right. - f_L(\omega) - f_R(\omega)) \right\} - 1 \right] .$$

Thus, for finite but small $\Delta \lambda$ the first cumulant (which corresponds to Eq. (20)) exhibits no dependence on $\Delta \lambda$, whereas the first moment oscillates at long times as $\exp(\sim i \Delta \lambda t)$. This implies that numerical derivatives taken from $\ln Z$ will converge at a finite $\Delta \lambda$ in the long time limit, while the error in numerical derivatives taken directly from the function will diverge.

To see this in practice, consider Fig. 8. Here, dot populations normalized by time $(C_1(t)/t)$ are plotted. These plots were obtained from both the logarithmic (thick black lines) and direct (dashed black lines) derivatives with a symmetric bias of $V = 1 \Gamma$. Cumulants are obtained from both derivatives for $\Delta \lambda = 0.6$ and compared with the exact result (red crosses) obtained from PI-NEGF at $\Delta \lambda = 0.001$. This is done for high temperatures ($\beta = 0.4\Gamma$, upper panel) and low temperatures ($\beta = 50\Gamma$, lower panel). As suggested by the preceding analysis, the convergence to the $\Delta \lambda \to 0$ limit is substantially faster when logarithmic derivatives are taken. While the direct and logarithmic derivatives are in agreement at short times, the direct derivative diverges from the exact result at long times. Without performing a full error analysis, it is difficult to determine whether the logarithmic derivative differs significantly from the exact result. At low temperatures convergence is slower, and the steady state value (equivalent to the long time limit of the left current) increases.

We repeat this analysis for the normalized second cumulants $C_2(t)/t$ of the population in Fig. 9. The second cumulant $C_2(t)$ increases linearly with time (thick lines) and is related at long times to the current noise in the left lead $S_{\lambda L}$ via Eq. (21). The cumulant extracted from the iQMC data is in excellent agreement with the exact result event for relatively large values of $\Delta \lambda$ when the cumulant is obtained from the logarithmic derivative, but requires convergence to decreasing $\Delta \lambda$ at smaller times when the direct derivative is used. We note that due to the symmetry of Eq. (39), it is in principle only necessary to evaluate $\ln Z(\lambda, t)$ at one value of $\lambda$ to evaluate both the first and second cumulants. In practice we used $\lambda = 0$ and $\lambda = 0.6$ to eliminate some of the systematic errors due to the finite time discretization.

VI. RESULTS IN THE PRESENCE OF INTERACTIONS

We continue to discuss quantities of physical interest, as obtained from our calculations. In what follows, we will show noninteracting results for comparison, which could be compared to the PI-NEGF data as in the previous section. Such comparisons were performed and agreement was found, but for the sake of brevity this is
Figure 9. Second cumulants of the particle number in the noninteracting case \( U = 0 \) are shown, as obtained from iQMC. Cumulants obtained from finite difference derivatives of \( \ln Z \) (thick lines) and of \( Z \) (dashed lines) are plotted against the essentially exact benchmark PI-NEGF logarithmic derivative for \( \Delta \lambda = 0.001 \) (red crosses). The inverse temperatures are \( \beta = 0.4/\Gamma \) (top) and \( \beta = 50/\Gamma \) (bottom).

Figure 10. Evolution of \( C_2(t) \) (upper panel) and the Fano factor \( F(t) = C_2(t)/C_1(t) \) (lower panel) for voltage \( V = 1\Gamma \) and initial condition \( |0\rangle \). To the right extreme of the plot, we show the asymptotic values of these quantities, obtained from linearly fitting the cumulants at long times. The exact values of the noninteracting noise and Fano factor from Levitov–Lesovik theory are shown as filled circles.

no longer shown from here onwards. Furthermore, we do not perform the full (and costly) error analysis needed to rigorously define confidence intervals in the inchworm algorithm [107, 112], but roughly estimate that numerical errors are on the order of several percent.

A. FCS and Shot noise

The FCS in the presence of interactions is shown at two voltages in Fig. 1 at equilibrium and in the presence of a bias voltage. At first glance, this looks only subtly different from the noninteracting results in Figs. 4 and 6. However, these seemingly small changes in the contour plot encode entirely different physics. To see this, we will explore some of the properties that can be derived from the FCS.

First, we study the effect of interactions and temperature on the noise and Fano factor. Fig. 10 shows the evolution of the time-normalized second moment \( C_2(t)/\Gamma \) in the upper panel. In the lower panel, the population Fano factor \( F(t) \) defined in Eq. 22 is shown, at fixed voltage \( V = 1\Gamma \) and for the initially empty dot state. The high (low) temperature plots are shown in red (black), and the interacting (noninteracting) results are distinguished by solid (dashed) lines. To the right extreme of the plot, the asymptotic values of the noise and Fano factor are shown. These were obtained from the coefficient of a linear fit to the first and second cumulants \( C_1(t) \) and \( C_2(t) \), in accordance with the definition of the steady state current and noise defined in Eqs. 20 and 21. In addition, filled circles mark the Levitov–Lesovik values for the noninteracting quantum noise and Fano factor, obtained from the standard Landauer–Büttiker theory [1, 10] via Eqs. 37 and 38. We note that the linear fits (which are not shown here) give reliable values to within 1% in this case, and in the noninteracting case are perfectly consistent with the Levitov–Lesovik values. The quality of the fits improves with the length of time simulated. In addition we note that asymptotic values shown to the right of Fig. 10 are independent of the initial condition.

At high temperatures (red lines), the absolute magnitdude of the noise, and also the Fano factor, is enhanced with respect to the low temperature (black lines) case. This is due to the contribution to the second cumulant of thermal or Nyquist noise, which vanishes at small values of \( k_B T \) [1, 10]. We note that at high temperatures there is an asymptotic divergence between the solid and dashed red curves in the upper panel, as the noise appears to be suppressed by the presence of interactions. In the low temperature case, the interacting shot noise is enhanced with respect to the noninteracting value, thus we observe a crossover from noise suppression to enhancement as the temperature is reduced. At low temperature the quantum noise is mainly attributed to the presence of a finite bias [1, 17]. While the thermal-to-shot noise crossover is well described by scattering theory in the noninteracting case, it is far more complex in the presence of strong interactions.
The low-temperature data corroborates well-known theoretical results by Lesovik [132] and others [133, 135], later confirmed by experiment [136], that the presence of inelastic scattering processes in the junction causes excess noise at low temperatures and finite voltages. In particular, inelastic cotunneling processes have been considered, in which electrons tunnel onto and off the dot simultaneously with the creation of a virtual dot state. These leave the dot in an excited state, and have been shown to enhance the noise [97, 137, 138]. When these virtual tunnel states involve correlated electrons of opposite spin, a Kondo singlet is formed [139]. At high temperatures, the interaction induces Coulomb blockade on the dot, reducing the number of available transport channels. This suppresses both the current and the noise, such that the Fano factor takes a similar value to that in the noninteracting case (lower panel, red lines). The presence of a large temperature suppresses the formation of coherent states required for inelastic spin-flip and cotunneling processes to occur [140–142]. However, as the temperature is lowered, we enter the Kondo regime. The Kondo effect increases the inelastic spin-flip rate and facilitates spin fluctuations on the dot, enhancing the noise and resulting in an overall increase in the Fano factor compared to the noninteracting case (solid black lines in Fig. 10).

B. First passage time distribution

In Figs. 11–12 we plot the FPTD, as defined in Eq. (6), in the unbiased ($V = 0$) and biased ($V = 10\Gamma$) cases. The inverse temperature in all plots is $\beta = 50/\Gamma$, and the dynamics for four different initial states of the dot are shown. Note that the initially half-occupied states $|\sigma\rangle$ are collected into the same line (shown in red), as the particle–hole symmetric parameters ensure their physical equivalence. Within our data, the two half-filled initial conditions are indeed identical to within numerical errors (not shown).

We begin with the noninteracting problem, Fig. 11. Here, a maximum occurs at $\Gamma t$ of order $\sim 1$, corresponding to the most probable first passage time $\tau_{fp}$, i.e. the time at which it is most likely to measure the first change in the left lead’s occupation. The first passage probability decays to zero at long times, as the likelihood that the first particle transfer has been detected at very long times becomes vanishingly small. In the unbiased case (top panel), the effect of initial condition on the first passage time distributions is minimal. In particular, the unoccupied (black) and fully-occupied (green) initial conditions are identical in the presence of particle–hole symmetry. This is not the case for the biased system (bottom panel), where electrons are driven out of the left lead. The bias voltage breaks the symmetry between the left and right leads. Therefore—since we are examining the FPTD of the left lead—the symmetry between the doubly occupied and unoccupied initial states is also broken. As the initial occupation of the dot increases in going from the $|0\rangle$ to the $|\sigma\rangle$ and $|\uparrow\downarrow\rangle$ states, the probability density shifts to longer times. A second peak appears in the distribution, corresponding to the first tunneling event on the left lead occurring after an electron has traveled from the dot onto the right lead. This demonstrates an initial condition dependent “queuing” effect in the FPTD, which is reminiscent of classical queuing [143].

The insets of Fig. 11 display the integral over $W(\tau)$, which by Eq. (3) is equal to the probability $1 - P(0, t)$ that when the total charge of the left lead is measured at time $t$, it has changed from its initial value. In an unbiased system the electron flow across the terminals, as well as the probability that the electron count in the left lead has changed from its initial value, is due entirely to thermal fluctuations and the delocalization dynamics of the wavefunction, and may saturate at a value between one and zero (see upper inset). As the bias $V$ is increased, however, the active driving causes this probability to reach unity rapidly (as in the lower inset).

Fig. 12 is identical to Fig. 11 except for the presence of a finite Coulomb repulsion energy $U = 8\Gamma$. At very short times, when the time dependence is linear, the symmetrized interaction has essentially no effect on the FPTD. However, this changes dramatically at timescales substantially smaller than $1/\Gamma$. Comparing the interacting unbiased case, shown in the upper panel, with the corresponding noninteracting result in Fig. 11, it is clear that the magnitude of the resonant peak and the value of $\tau_{fp}$ are increased for the doubly occupied and unoccupied initial states $|0\rangle$ and $|\uparrow\downarrow\rangle$, whereas both quantities are decreased for the half occu-
plied initial states $|\sigma\rangle$. This can be understood in terms of Coulomb blockade physics: at short times, involving the transfer of the first electron, no difference between the initial conditions is observed. However, in the $|\sigma\rangle$ case, the suppression of the even charge states by the interaction causes the transfer of a second electron in the same direction to be energetically unfavorable. The opposite occurs in the $|0\rangle$ and $|\uparrow\downarrow\rangle$ states.

Another striking difference between the interacting case and its noninteracting counterpart (top panels of Fig. 12 and Fig. 11, respectively) is the long tail appearing in the FPTD of the initially magnetized states $|\sigma\rangle$. The effect of this tail is prevalent in the inset, where very slow relaxation to unity is observed (we show times up to $t = 8/\Gamma$ to emphasize the slow relaxation). This can be attributed to the stabilization of local moments [144] and slow spin dynamics [145] associated with the Kondo regime, to which the system is equilibrating at these parameters. For comparison, the red dashed line in the upper inset of Fig. 12 shows dynamics for the $|\sigma\rangle$ initial states at $\beta = 1/\Gamma$, where the Kondo effect is suppressed by temperature; the relaxation is then faster.

In the lower panel of Fig. 12 we apply a bias voltage of $V = 10\Gamma$ to the interacting system. At such a large voltage, electrons are chiefly injected into the dot from the left lead and ejected into the right lead, simplifying the analysis of the effects of interaction on the queuing. In the case of the half occupied initial conditions $|\sigma\rangle$, the first peak is suppressed relative to the second in the presence of interactions, because the energy penalties discussed above decrease the likelihood of electron injection from the left lead before the original electron escapes into the right lead. Similarly, the first peak for the doubly occupied initial condition becomes larger than the second, because the process where two electrons are ejected on the right before the first electron is injected on the left is energetically suppressed. Remarkably, the slow spin relaxation related to Kondo physics is substantially reduced in the inset for the $|\sigma\rangle$ initial states at this high voltage. This is consistent with claims that a small remnant of Kondo physics can survive at high voltages [124, 146, 147], but requires further investigation.

C. $n$-particle probabilities

The FPTD, related to $P(0, t)$, sheds light on the dynamical response of the system to all possible particle tunneling events. It does not distinguish between events in which different numbers of particles are transported. However, the full counting statistics contain much more information: it is possible to access $P(\Delta n, t)$ for every $\Delta n$. It is of interest to search for manifestations of correlated transport in these detailed distributions. We also consider the most probable time at which the particle number changes by $\Delta n$, given by the maximum of the corresponding distribution. We denote this time by $\tau_{\text{mp}}^{(\Delta n)}$.

In Fig. 13 we begin by considering the probability for the particle number in the left lead to have changed by $\Delta n = 1$ after time $t$, for each possible initial condition and for the interacting (solid lines) and noninteracting (dashed lines) cases. We once again apply a bias volt-
age of $V = 10\Gamma$, so that it is energetically favorable to move charges from the left lead onto the dot, and charges from the dot to the right lead. Regardless of the presence of interactions, the peak of the probability distributions $P(1,t)$ for the different initial conditions is shifted to longer times by an increasing number of electrons initially on the dot. In the noninteracting case, this simply reflects the fact that the empty dot has more transport channels open to electron traversal events than a partially or fully filled dot. If the initial state is fully occupied, transport of the first electron is forbidden by the Fermi rule until at least one dot electron has tunneled to (with high probability) the right lead. The effect of Coulomb charging depends on the initial occupation: when the dot begins empty, interaction does not change the position of the maximum, but somewhat enhances the total probability that a single tunneling event was measured at any time, reflecting a reduction in the probability of detecting higher values of $\Delta n$. In this case there is no initial “queue” for electron tunneling onto the dot, and the interaction makes the first electron tunneling event more energetically favorable, then suppresses the second. For all other initial states, the Coulomb repulsion shifts the distribution to longer times, as electrons on the left lead cannot enter the dot without “queuing” for electrons on the dot to first tunnel to the right lead.

In the lower panel of Fig. 13 we repeat the above analysis for the probability that at time $t$, a change $\Delta n = 2$ in the particle number is measured on the left lead. $P(2,t)$ is shifted to longer times by the interaction for all initial conditions, because the transfer of the second electron must be preceded by that of the first, such that it always encounters some Coulomb repulsion.

The top panel of Fig. 14 shows a sequence of $P(\Delta n, t)$ for $\Delta n = 0,\ldots,6$, focusing on the unoccupied initial state. At short times, we see perfect agreement between the interacting and noninteracting results, as no interaction can take place before some electrons occupy the dot. Increasing the value of $\Delta n$ shifts the $\Delta n > 1$ peaks to longer times, which makes sense physically since a change involving $\Delta n = N$ electrons implies all processes leading to a change of $\Delta n = N-1$ electrons have already occurred. We note that we do not show backscattering events $\Delta n < 1$, which have a small but finite probability even in the presence of a large voltage. In the unbiased system (not shown), backscattering events are of equal importance to the forward scattering events.

The rich, detailed information on population densities shown in Fig. 14 lends itself to a quantitative analysis of electron transfer processes. In the lower panel of the same figure, we study the variation of the distance between maximally probable times $\tau_{mp}^{(\Delta n)} - \tau_{mp}^{(\Delta n-1)}$ as a function of $\Delta n$, for the interacting (black curve) and noninteracting (red curve) cases. As we propagate to longer times, this quantity can be considered a proxy for queuing effects at the single electron level at steady state: it describes the typical waiting time between subsequent tunneling events. As $\Delta n$ increases, the peak-to-peak distance stabilizes to a roughly constant value.

The long-time value in the noninteracting case is of order $1/\Gamma$, as might be expected for a simple rate process where the only relevant timescale is the coupling between the dot and baths. However, the peak-to-peak distance in the interacting case is significantly larger: as each electron enters the dot, it suppresses the next electron from entering by virtue of the Coulomb repulsion. Interestingly, while the point at which each distribution begins to differ significantly from zero is only weakly modified by the interaction, the width of the distributions and the weight at their tail end are almost immediately enhanced. This may indicate that fluctuations play an increasingly important role in the population transfer dynamics as the strength of the interaction is increased.

VII. CONCLUSIONS

We presented the first numerically exact calculation of full counting statistics for a non-integrable model of interacting fermions, in this case the nonequilibrium An-
derson impurity model. Using the inchworm quantum Monte Carlo method, we obtained the generating function \( Z(\lambda, t) \) at a variety of physical parameters ranging from the noninteracting case to the strongly correlated Kondo regime. This provides access to currents, which have been accessed before by numerically exact methods; but also to other experimentally measurable quantities which were not. This includes the current noise at steady state, the Fano factor, all higher moments and cumulants of population transfer event, and the complete time-dependent probability distributions for \( n \)-electron transfer events.

The method was applied to a coupling quench, where a dot is suddenly attached to the leads at time zero and allowed to evolve to equilibrium or to a nonequilibrium steady state. After performing benchmark comparisons in the noninteracting limit to verify the accuracy of our results, we explored the effects of electron–electron interactions and nonequilibrium bias voltage on the full counting statistics and the properties derived thereof. We observed the signatures of the Coulomb blockade and Kondo effects in the noise and Fano factor, and found transient queuing effects depending on the choice of initial condition in the FPTD. By considering the individual, time-dependent probability distributions for the occurrence of \( n \) tunneling events, we were further able to argue that such queuing effects persist at steady state. Rather than simply being associated with a lower effective tunneling rate, the rapid widening of the probability densities in the presence of interactions suggests that fluctuations play a significant role in the interacting dynamics.

With the availability of a reliable scheme for calculating the full counting statistics of generic interacting impurity models, a variety of important research questions can now be addressed. Some immediately relevant examples include quantum thermodynamic topics, such as the verification of quantum fluctuation–dissipation theorems and the calculation of efficiency fluctuations in quantum devices; and the evaluation of noise and Fano factors beyond the Fermi liquid regime. We observed queuing effects, but it would also be interesting to look for bunching effects in the presence of an effectively attractive interaction. As the noninteracting physics enters entirely through the hybridization function, it is also feasible to consider the counting statistics of impurities embedded in more realistic noninteracting models of materials and nanosystems, one interesting example being magnetic impurities in graphene nanoribbons.

Spin-dependent FCS in quantum junctions has only recently been treated in the noninteracting case and, and spintronic applications are anticipated. The method we have presented can easily be generalized to access spin-dependent counting statistics and multi-lead moments: one could then go beyond the analysis of Fig. 14 to consider the time-dependent probability that the left lead lost (\( e.g. \)) one spin up electron, while simultaneously the right lead gained two spin down electrons. In this context, the long-term survival of spin fluctuations at timescales where the mean magnetization has died out would embody a remarkably clear signature of strong correlation physics. Generalization to include bosonic degrees of freedom is also possible, enabling treatment of thermolectric systems or coupling to an optical continuum. The method could further be extended to frequency dependent power spectra and related conductance spectra, which can be used to study photon absorption and emission processes. Additionally, noise in periodically-driven systems with interactions could now be investigated, enabling performance tuning of correlated nanoelectronic devices.

To summarize, the full counting statistics of strongly correlated nonequilibrium quantum impurity problems can now be obtained numerically. This provides unprecedented insight into the stochastic nature of electronic transport at the resolution of single tunneling events, and allows simulation of a variety of experiments for which no systematic theory was previously available.

Acknowledgements We are grateful to Michael Galperin for directing our attention to this problem. G.C. acknowledges support by the Israel Science Foundation (Grant No. 1604/16). M.R. was supported by the Raymond and Beverly Sackler Center for Computational Molecular and Materials Science, Tel Aviv University. E.G. was funded by DOE ER 46932. This research was supported by grant No. 2016087 from the United States-Israel Binational Science Foundation (BSF).

[1] Y. M. Blanter and M. Büttiker, Physics reports 336, 1 (2000).
[2] N. V. Kampen, Stochastic processes in physics and chemistry (North Holland, 2007).
[3] A. Braggio, J. König, and R. Fazio, Physical review letters 96, 026805 (2006).
[4] R. Landauer, Nature 392, 658 (1998).
[5] C. Beenakker, C. Schönenberger, et al., Physics Today 56, 37 (2003).
[6] R. Vardimon, M. Klionsky, and O. Tal, Physical Review B 88, 161404 (2013).
[7] Y. Selzer and U. Peskin, The Journal of Physical Chemistry C 117, 23269 (2013).
[8] P.-Y. Yang, C.-Y. Lin, and W.-M. Zhang, Physical Review B 89, 115411 (2014).
[9] M. A. Ochoa, Y. Selzer, U. Peskin, and M. Galperin, The journal of physical chemistry letters 6, 470 (2015).
[10] M. Ridley, A. MacKinnon, and L. Kantorovich, Physical Review B 95, 165440 (2017).
[11] V. S. Rychkov, M. L. Polianski, and M. Büttiker, Physical Review B 72, 155326 (2005).
[135] P. Vitushinsky, A. Clerk, and K. Le Hur, Physical review letters 100, 036603 (2008).
[136] G. Iannaccone, G. Lombardi, M. Macucci, and B. Pellegrini, Physical review letters 80, 1054 (1998).
[137] E. V. Sukhorukov, G. Burkard, and D. Loss, Physical Review B 63, 125315 (2001).
[138] A. Thielmann, M. H. Hettler, J. König, and G. Schön, Physical review letters 95, 146806 (2005).
[139] S. M. Cronenwett, T. H. Oosterkamp, and L. P. Kouwenhoven, Science 281, 540 (1998).
[140] A. Furusaki and K. Matveev, Physical Review B 52, 16676 (1995).
[141] J. König, H. Schoeller, and G. Schön, Physical Review B 58, 7882 (1998).
[142] T. Tran, I. Beloborodov, J. Hu, X. Lin, T. Rosenbaum, and H. Jaeger, Physical Review B 78, 075437 (2008).
[143] D. V. Lindley, in Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 48 No. 2 (Cambridge University Press, 1952) pp. 277–289.
[144] A. C. Hewson, The Kondo problem to heavy fermions, Vol. 2 (Cambridge university press, 1997).
[145] U. Weiss, Quantum dissipative systems, Vol. 13 (World scientific, 2012).
[146] N. S. Wingreen and Y. Meir, Physical review B 49, 11040 (1994).
[147] F. B. Anders, Physical review letters 101, 066804 (2008).
[148] C. G. da Rocha, R. Tuovinen, R. van Leeuwen, and P. Koskinen, Nanoscale 7, 8627 (2015).
[149] M. Ridley and R. Tuovinen, Physical Review B 96, 195429 (2017).
[150] E. Rothstein, O. Entin-Wohlman, and A. Aharony, Physical Review B 79, 075307 (2009).
[151] N. Gabdank, E. Rothstein, O. Entin-Wohlman, and A. Aharony, Physical Review B 84, 235435 (2011).
[152] H.-A. Engel and D. Loss, Physical review letters 93, 136602 (2004).
[153] O. Entin-Wohlman, Y. Imry, S. Gurvitz, and A. Aharony, Physical Review B 75, 193308 (2007).
[154] C. P. Orth, D. F. Urban, and A. Komnik, Physical Review B 86, 125324 (2012).
[155] K. Miwa, F. Chen, and M. Galperin, Scientific Reports 7, 9735 (2017).
[156] M. Strass, P. Hänggi, and S. Kohler, Physical review letters 95, 130601 (2005).