Thermoelectric properties of an interacting quantum dot-based heat engine

Paolo Andrea Erdman,1 Francesco Mazza,1 Riccardo Bosisio,1 Giuliano Benenti,2,3,4 Rosario Fazio,5,1 and Fabio Taddei6

1NEST, Scuola Normale Superiore and Istituto Nanoscienze-CNR, I-56127 Pisa, Italy
2Center for Nonlinear and Complex Systems, Dipartimento di Scienza e Alta Tecnologia, Università degli Studi dell’Insubria, via Valleggio 11, 22100 Como, Italy
3Istituto Nazionale di Fisica Nucleare, Sezione di Milano, via Celoria 16, 20133 Milano, Italy
4NEST, Istituto Nanoscienze-CNR, I-56126 Pisa, Italy
5ICTP, Strada Costiera 11, I-34151 Trieste, Italy
6NEST, Istituto Nanoscienze-CNR and Scuola Normale Superiore, I-56126 Pisa, Italy

We study the thermoelectric properties and heat-to-work conversion performance of an interacting, multi-level quantum dot (QD) weakly coupled to electronic reservoirs. We focus on the sequential tunneling regime. The dynamics of the charge in the QD is studied by means of master equations for the probabilities of occupation. From here we compute the charge and heat currents in the linear response regime. Assuming a generic multi-terminal setup, and for low temperatures (quantum limit), we obtain analytical expressions for the transport coefficients which account for the interplay between interactions (charging energy) and level quantization. In the case of systems with two and three terminals we derive formulas for the power factor Q and the figure of merit ZT for a QD-based heat engine, identifying optimal working conditions which maximize output power and efficiency of heat-to-work conversion. Beyond the linear response we concentrate on the two-terminal setup. We first study the thermoelectric non-linear coefficients assessing the consequences of large temperature and voltage biases, focusing on the breakdown of the Onsager reciprocal relation between thermopower and Peltier coefficient. We then investigate the conditions which optimize the performance of a heat engine, finding that in the quantum limit output power and efficiency at maximum power can almost be simultaneously maximized by choosing appropriate values of electrochemical potential and bias voltage. At last we study how energy level degeneracy can increase the output power.

PACS numbers: 72.20.Pa,73.23.-b

I. INTRODUCTION

The study of thermoelectric effects in nanostructures1–4 is attracting increasing interest. Heat-to-work conversion based on thermoelectricity promises an enhanced efficiency as a consequence of the reduction of the phonon contribution to thermal conductance in disordered nanostructures5 and of the “energy filtering” effect that can result from confinement and quantum effects.6,7 In particular, an increase of the electron contribution to the figure of merit ZT (which controls the maximum efficiency and the efficiency at maximum power) is possible if one can “filter” the electrons participating in the transport to a narrow energy range7. A heat engine composed of a quantum dot (QD) is a paradigmatic example, since it is characterized by a spectrum of discrete levels which maximizes energy filtering. The thermoelectric properties of QD systems8–30 and the performance of QD-based heat engines31–62 has been studied theoretically by a number of authors (see Ref. 63 for a review). The vast majority of the papers dealing with QD-based heat engines consider a single degenerate energy level or two non-degenerate levels.31–37,39,41,42,44–53,55–62. The case of QDs with many levels has been addressed only in few papers38,40,43,54. Moreover, the performance of QD-based heat engine has been mostly studied within the linear response regime31–53,56–57,61,62, where the thermoelectric performance of the system is entirely characterized by ZT. The case of an interacting multi-level QD beyond the linear response has not been addressed so far. On the one hand, the presence of many levels is expected to yield important consequences. Indeed, already in the linear-response regime they have an impact on the thermopower, the thermal conductance and ZT.9,15,17,38. On the other, non-linear effects, relevant when larger temperature and voltage biases are applied, are important as far as power and efficiency are concerned. We emphasize that a number of experiments assessing the thermoelectric properties of QDs have been reported in Refs. 64–84.

In this paper we fill this gap by studying the thermoelectric properties and heat-to-work conversion performance of a multi-level QD in a multi-terminal configuration within the Coulomb blockade regime. We consider the limit of small tunnelling rates (sequential tunnelling regime) and we study both the linear and non-linear response regimes. Coulomb interaction among electrons is accounted for by a finite and small capacitance C associated energy scale is its charging energy (Ne)2/2C, where N is the number of electrons in the QD and e is the electron charge. Moreover, we consider a generic multi-terminal structure, whereby the QD is connected to many (two or more) reservoirs. 

arXiv:1702.06042v2 [cond-mat.mes-hall] 28 Jun 2017
tribution to heat transport. Our results therefore set an upper bound to the thermoelectric efficiency of the QD, approachable only in the limit in which suitable strategies to strongly reduce phonon transport are implemented.

Generalizing Refs. 8 and 9 to the multi-terminal case, by solving a set of kinetic equations one can determine the probability of occupation of the energy levels of the QD in a multi-terminal setup, thus allowing to calculate the charge and heat currents for given values of the electrochemical potentials and temperatures of the reservoirs.

In the linear response regime, where voltage and temperature biases are small, we derive closed-form expressions for the charge and heat currents and specify their limits of validity. We define local and non-local transport coefficients and express them in terms of a generating function under the assumption that the tunneling rates are energy-independent. We then derive, in the low temperature limit, analytical expressions for all transport coefficients as a function of the electrochemical potential $\mu$. Along with the main features of the transport coefficients (located around values of $\mu$ equal to the dominant transition energies required to add or remove an electron from the QD), such expressions also describe a fine structure arising from the interplay between interaction and level quantization (controlled by the two energy scales: charging energy and level spacing). Furthermore, for the calculation of the thermal conductance we find that it is crucial to consider the presence of many levels. Within the linear response, we consider both the two-terminal and the three-terminal system aiming at addressing the performance of heat-to-work conversion. In the former case we obtain analytical expressions for the power factor $Q$ and the figure of merit $ZT$ finding, remarkably, that those quantities are simultaneously maximized for values of the electrochemical potential which differ by about $2.40 k_B T$ with respect to the dominant transition energies. In addition, $ZT$ shows a fine structure of secondary peaks whose height is independent of the system’s parameters and can take values as large as $ZT = 9$. We compare the figure of merit with a non-interacting system, finding that Coulomb interactions dramatically increase $ZT$ by strongly suppressing the thermal conductance. For the case of three terminals with energy-independent tunneling rates, we derive analytic expressions for the maximum power and corresponding efficiency, finding that the addition of a third terminal at an intermediate temperature decreases the efficiency at maximum power, but can increase the power. We also find particular intermediate temperatures where the third terminal increases the maximum power and achieves the same efficiency of a two terminal system.

We analyze the regime beyond the linear response by numerically solving the kinetic equations, focusing on the two-terminal setup. Going beyond the linear response, i.e. considering large temperature and voltage biases, $\Delta T$ and $V$, is interesting for various reasons. On one hand, it allows to increase the Carnot efficiency $\eta_C$ and the power generated by a heat engine (in our case the peak power scales approximately as $(\Delta T)^2$ also beyond the linear response regime). On the other hand, the efficiency at maximum power is not bounded by $\eta_C/2$, as in the linear response, and can even go above the Curzon-Albibn efficiency. In literature, the scattering theory of nonlinear thermoelectric transport in quantum conductors has been developed only recently. The regime beyond linear response for QD-based heat engines has been theoretically addressed in Refs. 34,35,37,41,43,44,47,49,50,54,58–60, but limited to single or double level quantum dots.

In discussing the results, we first focus on the behavior of the non-linear Seebeck and Peltier coefficients aiming at assessing the interplay between charging energy and level spacing on these two quantities and how the Onsager reciprocity relation that connects them is violated beyond linear response. Second, we study the efficiency and output power of a heat engine. In particular, we calculate the maximum efficiency and maximum power by maximizing such quantities with respect to the applied bias voltage, for fixed values of temperature bias and electrochemical potential. The maximum efficiency shows only quantitative changes, with respect to the linear response, by increasing the temperature bias. The efficiency at maximum power instead develops peaks which go beyond the $\eta_C/2$ linear-response limit and approach $\eta_C$ for large temperature differences. Remarkably, the maximum power, normalized to its peak value, only slightly depends on the temperature bias and can be well approximated by the analytic expression obtained for the linear response regime. Moreover, we find that efficiency at maximum power and maximum power take approximately their peak values simultaneously under the same conditions found for the linear response. Finally, we assess the impact of interactions by comparing the efficiency at maximum power in two situations, namely the case of a doubly degenerate level with interaction and the case of two non-degenerate levels without interaction. We find that, especially when charging energy and level spacing are of the order of the thermal energy, the efficiency at maximum power is much higher in the interacting case and goes above the Curzon-Albibn efficiency. We also find that in the doubly degenerate interacting case the maximum power is enhanced by almost a factor 2 with respect to the non-degenerate case.

The paper is organized as follows: in Sec. II we describe the system under investigation and we detail the theoretical model. In Sec. III the model is specified to the linear response regime. Here we report analytic expressions obtained in the quantum limit, Sec. IIIA, and calculate power and efficiency in the cases of a two-terminal system, Sec. IIIB (with the efficiencies of interacting and non-interacting QDs compared in Sec. IIIC), and of a three-terminal system, Sec. IIID. Sec. IV is devoted to the discussion of the regime beyond the linear response in a two-terminal system: in Sec. IV A we study the non-linear Seebeck and Peltier coefficients and in Sec. IV B...
we discuss efficiency and output power of a heat engine. Finally, in Sec. V we draw our conclusions and discuss future developments. In addition, the paper includes four appendices where the details of some calculations and the analytic non linear study of a single energy level QD system are reported.

II. MULTI-LEVEL INTERACTING QD

In this Section we briefly outline the formalism used to describe the thermoelectric properties of a multilevel interacting QD. We will only consider electron transport, neglecting any contribution due to phonons. As shown in Fig. 1 (top), the QD is tunnel-coupled to $N$ electron reservoirs, each characterized by a given temperature $T_\alpha$ and electrochemical potential $\mu_\alpha$, so that the occupation of the electrons within reservoir $\alpha$ follows the Fermi distribution

$$f_\alpha(E) = \frac{1}{1 + \exp\left(\frac{E - \mu_\alpha}{k_B T_\alpha}\right)}, \quad (1)$$

where $k_B$ is Boltzmann’s constant. In Fig. 1 (bottom), $E_p$ (with $p = 1, 2, \ldots$ labeled in ascending order) are the QD single-electron energy levels. These levels can be shifted by means of an applied gate voltage.

The QD is weakly coupled to the reservoirs through large tunneling barriers. More precisely, we assume that thermal energy $k_B T$, level spacing and charging energy are much larger than the coupling energy between reservoirs and QD [$\hbar \sum_\alpha \Gamma_\alpha(p)$, where $\Gamma_\alpha(p)$ is the tunneling rate from level $p$ to reservoir $\alpha$], which we assume independent of the number $N$ of electrons inside the dot. As a consequence, the charge on the QD is quantized, i.e. each energy level $E_p$ can have either zero or one electron, $n_p = 0$ or $n_p = 1$ (any degeneracy, like electron spin, can be taken into account counting each level multiple times), and transport occurs due to single-electron tunneling processes (sequential tunneling regime). The electrostatic energy associated with the electrons within the QD is given by $U(N) = E_C N^2$, where $E_C = e^2/2C$, $N = \sum_i n_i$ is the total number of electrons within the QD, and $C$ is the capacitance of the QD. The QD is described by states characterized by a set of occupation numbers $\{n_i\}$ relative to the energy levels. The QD changes state whenever a single-electron tunneling process takes place. The non-equilibrium probability for a given state $\{n_i\}$ to occur, $P(\{n_i\})$, can be computed8,93 by writing a straightforward set of balance equations for $P(\{n_i\})$. Our aim is to compute in stationary conditions the charge, energy and heat currents out of the electron reservoirs [denoted in Fig. 1 (top) by $j^c_\alpha$, $j^u_\alpha$ and $j^b_\alpha$, respectively], induced by the temperature and electrochemical potential differences.

FIG. 1: (Color online) Top: A quantum dot (QD) is tunnel-coupled to $N$ reservoirs, each kept at a temperature $T_\alpha$ and at an electrochemical potential $\mu_\alpha$, with $\alpha = 1, \ldots, N$. Arrows represent charge, energy and heat currents ($j^c_\alpha$, $j^u_\alpha$, and $j^b_\alpha$, respectively) flowing from the reservoirs $\alpha$ to the QD. Bottom: Schematic energy representation of a multilevel QD. $E_1$, $E_2$, etc. are the single-electron energy levels of the QD, while $\mu_1$ and $\mu_\alpha$ are the electrochemical potentials relative to reservoir 1 and $\alpha$, respectively.

A. Kinetic equations

In what follows we describe a generalization of the method put forward by Beenakker in Refs. 8 and 9. The single-electron tunneling processes that contribute to changing over time the probability $P(\{n_i\})$ are due to electrons that tunnel from the QD to the reservoirs and vice-versa. For an electron exiting the QD, initially with $N$ electrons, from energy level $E_p$ and going into reservoir $\alpha$ at energy $E^{\text{in}}_\alpha$, energy conservation imposes that

$$E_p + U(N) = E^{\text{in}}_\alpha + U(N - 1). \quad (2)$$

On the contrary, for an electron that tunnels from an initial state in reservoir $\alpha$ at energy $E^{\text{in}}_\alpha$ to the level $E_p$ in the QD that initially had $N$ electrons, energy conservation imposes that

$$E^{\text{in}}(N) + U(N) = E_p + U(N + 1). \quad (3)$$

\[\text{Reservoir } 1 \quad \text{QD} \quad \text{Reservoir } \alpha\]
\[ P(\{n_i\}) \] can then be determined by the following set of kinetic equations, one for each configuration \(\{n_i\}\):

\[
\frac{\partial}{\partial t} P(\{n_i\}) = - \sum_{p} \delta_{n_p,0} P(\{n_i\}) \Gamma_\alpha(p) f_\alpha(E^{\text{in}}(N)) - \sum_{p} \delta_{n_p,1} P(\{n_i\}) \Gamma_\alpha(p) \left[ 1 - f_\alpha(E^{\text{fin}}(N)) \right] \\
+ \sum_{p} \delta_{n_p,0} P(\{n_i, n_p = 1\}) \Gamma_\alpha(p) \left[ 1 - f_\alpha(E^{\text{fin}}(N + 1)) \right] + \sum_{p} \delta_{n_p,1} P(\{n_i, n_p = 0\}) \Gamma_\alpha(p) f_\alpha(E^{\text{in}}(N - 1)),
\]

(4)

where we have introduced the notation

\[ P(\{n_i, n_p = 1\}) = P(\{n_1, \ldots, n_{p-1}, 1, n_{p+1}, \ldots\}) \] (5)

and

\[ P(\{n_i, n_p = 0\}) = P(\{n_1, \ldots, n_{p-1}, 0, n_{p+1}, \ldots\}) \] (6)

for the QD states. The first term in Eq. (4) accounts for the decrease of the probability \(P(\{n_i\})\), with the QD initially in the state \(\{n_i\}\), due to an electron coming from a reservoir and occupying an empty level in the QD. The rate of electrons coming from reservoir \(\alpha\) will be given by a sum over all empty levels \(p\) (such that \(n_p = 0\)) of the tunnel rate \(\Gamma_\alpha(p)\), multiplied by the probability of finding the QD in this state, \(P(\{n_i\})\), and multiplied by the reservoir’s occupation \(f_\alpha(E^{\text{in}}(N))\) at the correct energy \(E^{\text{in}}(N)\) to tunnel to level \(p\). The second term accounts for the decrease of the probability \(P(\{n_i\})\), with the QD initially in the state \(\{n_i\}\), due an electron leaving the QD from an occupied level to tunnel into a reservoir. The third term accounts for the increase of the probability \(P(\{n_i\})\) if the QD is in a state with an extra electron in level \(p\) with respect to \(\{n_i\}\), and if this electron leaves the QD, tunneling to the reservoirs. The forth term accounts for the increase of the probability \(P(\{n_i\})\) if the QD is in a state with a missing electron in level \(p\) with respect to \(\{n_i\}\), and if this electron enters the QD in level \(p\), tunneling from the reservoirs. The stationary solution of the kinetic equations, obtained imposing \(\partial P/\partial t = 0\), together with the normalization request

\[ \sum_{\{n_i\}} P(\{n_i\}) = 1 \] (7)

provides a complete set of equations that uniquely defines \(P(\{n_i\})\). The sum over \(\{n_i\}\) means the sum over \(n_i = 0, 1, \ldots\).

**B. Charge, energy, and heat currents**

Charge \(J^c_\alpha\) and energy \(J^e_\alpha\) currents flowing from reservoir \(\alpha\) to the QD can be calculated as the sum of all possible tunneling processes, since the QD can be in any state \(\{n_i\}\) with probability \(P(\{n_i\})\) and an electron can tunnel into or out of any energy level \(E_p\). More precisely, for the charge current we have

\[ J^c_\alpha = e \sum_{p=1}^{\infty} P(\{n_i\}) \Gamma_\alpha(p) \left\{ \delta_{n_p,0} f_\alpha(E^{\text{in}}(N)) - \delta_{n_p,1} [1 - f_\alpha(E^{\text{fin}}(N))] \right\}, \] (8)

\(e\) being the electronic charge, while for the energy current we have

\[ J^e_\alpha = \sum_{p=1}^{\infty} P(\{n_i\}) \Gamma_\alpha(p) \left\{ \delta_{n_p,0} f_\alpha(E^{\text{in}}(N)) [E^{\text{in}}(N) - \mu_\alpha] - \delta_{n_p,1} [1 - f_\alpha(E^{\text{fin}}(N))] [E^{\text{fin}}(N) - \mu_\alpha] \right\}. \] (9)

\(E^{\text{in}}(N)\) \([E^{\text{fin}}(N)]\) being the energy carried by an electron entering (exiting) the QD. The heat currents exiting the reservoirs can be calculated as \(J^h_\alpha = J^c_\alpha - \frac{\partial}{\partial t} J^e_\alpha\). Using Eqs. (8) and (9), we find that

\[ J^h_\alpha = \sum_{p=1}^{\infty} P(\{n_i\}) \Gamma_\alpha(p) \left\{ \delta_{n_p,0} f_\alpha(E^{\text{in}}(N)) \left[ E^{\text{in}}(N) - \mu_\alpha \right] - \delta_{n_p,1} [1 - f_\alpha(E^{\text{fin}}(N))] [E^{\text{fin}}(N) - \mu_\alpha] \right\}. \] (10)

In order to numerically determine the stationary probability distribution \(P(\{n_i\})\) from the kinetic equations,
we will consider a finite number $L$ of energy levels. By organizing the values of $P(\{n_i\})$ into a $2^L$-components vector $\vec{P}$ (two choices $n_i = 0, 1$ for each level), the kinetic equations (4) for the stationary probability distribution, $\frac{d\vec{P}}{dt} = 0$, can be represented as the homogeneous linear system $M\vec{P} = 0$, where $M$ is a $2^L \times 2^L$ matrix. $M$ must have a null space of at least dimension 1, otherwise the only possible solution would be the trivial one ($\vec{P} \equiv 0$). This is demonstrated in App. A by showing that summing together all the stationary kinetic equations yields zero. We can thus find the probabilities by including the normalization condition, Eq. (7).

By defining $\vec{N} = \sum_{i \not\equiv p} n_i$ it is possible to show that the kinetic equations for the stationary probability distribution can be written as

$$\sum_{p} \left( \delta_{n_p,1} - \delta_{n_p,0} \right) \left[ P(\{n_i\}, n_p = 0) A_{\vec{N},p} - P(\{n_i\}, n_p = 1) B_{\vec{N},p} \right] = 0,$$  

where

$$A_{\vec{N},p} = \sum_{\alpha} \Gamma_{\alpha}(p)f_{\alpha} \left( E^{\text{in}}(\vec{N}) \right)$$

and

$$B_{\vec{N},p} = \sum_{\alpha} \Gamma_{\alpha}(p) \left[ 1 - f_{\alpha} \left( E^{\text{in}}(\vec{N}) \right) \right] = \sum_{\alpha} \Gamma_{\alpha}(p) - A_{\vec{N},p}.$$

To derive Eq. (11) we have used the fact that $\vec{N} = N$, for the terms in the kinetic equations proportional to $\delta_{n_p,0}$, and $\vec{N} = N - 1$, for the terms in the kinetic equations proportional to $\delta_{n_p,1}$, and the identity

$$E^{\text{in}}(N + 1) = E^{\text{in}}(N),$$

stemming from Eqs. (2) and (3). It is worth mentioning that using the kinetic equations in the form of Eq. (11), it is possible to prove that the kinetic equations always allow a non-trivial solution (see App. A).

**C. Detailed balance equations**

It is clear that the kinetic equations (11) are automatically satisfied when the following set of equations

$$P(\{n_i\}, n_p = 0) A_{\vec{N},p} - P(\{n_i\}, n_p = 1) B_{\vec{N},p} = 0$$

is fulfilled for all values of $p$ and for all sets of occupation numbers $\{n_i\}$. Following Ref. 8, Eqs. (15) are hereafter referred to as detailed balance equations (DBEs). Eqs. (15) represent a set of $L \cdot 2^{L-1}$ equations, since $p$ can take $L$ values and, at a given $p$, all other occupation numbers ($n_1, \ldots, n_{p-1}, n_{p+1}, \ldots, n_L$) can be chosen in $2^{L-1}$ different ways. Of course, if a solution to the DBEs exists, than it is also a solution of the kinetic equations. We can show, however, that the DBEs are not in general consistent, i.e. no set of $P(\{n_i\})$ exists that can simultaneously satisfy all the DBEs (see App. B). This is also true in the linear response regime. In this case, however, we could prove (see App. C) that the DBEs are consistent and that can be used in the general case to obtain a closed-form expression of the charge current in the linear response regime (see Sec. III).

**D. Level balance equations**

We will now derive a set of equations that is always consistent and that can be used in the general case to obtain a closed-form expression of the charge current in the linear response regime (see Sec. III). We impose that, in stationary conditions, the rate of electrons entering any given QD energy level $p$ must equal the rate of electrons leaving that energy level. For electrons tunneling into the QD, initially with $N$ electrons, from any reservoir, one has to require that level $p$ is empty and must consider all possible states ($\{n_i\}, n_p = 0$), where $\vec{N} = N$. The total rate of electrons entering energy level $E_p$ is given by

$$\sum_{\alpha,\{n_i\} \not\equiv p} P(\{n_i\}, n_p = 0) \Gamma_{\alpha}(p) f_{\alpha} \left( E^{\text{in}}(N) \right).$$

For electrons tunneling out of the QD, initially with $N$ electrons, to any reservoir one has to require that level $p$ is occupied and must consider states with ($\{n_i\}, n_p = 1$), where $\vec{N} = N - 1$. The total rate of electrons leaving level $E_p$ is given by

$$\sum_{\alpha,\{n_i\} \not\equiv p} P(\{n_i\}, n_p = 1) \Gamma_{\alpha}(p) \left[ 1 - f_{\alpha} \left( E^{\text{in}}(N) \right) \right].$$

If we equate the rates of electron entering and leaving level $p$, and use the notation introduced in Eqs. (12) and (13), we obtain

$$\sum_{\{n_i\} \not\equiv p} \left[ P(\{n_i\}, n_p = 0) A_{\vec{N},p} - P(\{n_i\}, n_p = 1) B_{\vec{N},p} \right] = 0.$$  

We will refer to this set of $L$ equations (one for each energy level $p$) as the level balance equations (LBEs). Note that, using an argument similar to that put forward in App. A, it is possible to prove that Eqs. (18) can
be obtained from the kinetic equations, thus the LBEs are always consistent with the kinetic equations. However, the number of LBEs (equal to $L$) is smaller than the number of kinetic equations (equal to $2^L$) and they might not be sufficient to determine the probabilities $P\{\{n_i\}\}$. Note that one can prove that Eqs. (18) yield charge current conservation: $\sum_\alpha J_\alpha^c = 0$.

E. Output power and efficiency

Under steady-state conditions, the output power $P$ of a multi-terminal system is given by the sum of all the heat currents

$$P = \sum_{\alpha=1}^N J_\alpha^h.$$  

(19)

If $P > 0$, the system behaves as a heat engine, i.e. converting heat into work. In this situation the efficiency $\eta$ is defined as the ratio between the output power and the total heat current absorbed by the system

$$\eta = \frac{P}{\sum_\alpha J_\alpha^h},$$  

(20)

where the sum over $\alpha$ runs over all positive heat currents. For a two-terminal system the efficiency cannot exceed the Carnot efficiency defined as $\eta_C = 1 - T_1/T_2$, with $T_2 > T_1$. In addition, for a multi-terminal system $\eta$ cannot go beyond the two-terminal Carnot efficiency calculated using the hottest and coldest temperatures among $T_1, T_2, ..., T_N$.

We define the temperature and electrochemical potential differences as $T_\alpha = T + \Delta T_\alpha$ and $\mu_\alpha = \mu + \Delta \mu_\alpha$, with $\alpha = 1, ..., N$, and choosing reservoir 1 as the reference value, i.e. $\Delta T_1 = \Delta \mu_1 = 0$. In what follows we fix the values of $\Delta T_\alpha$ and calculate the maximum output power, $P_{\text{max}}$, and maximum efficiency, $\eta_{\text{max}}$, by varying $\Delta \mu_\alpha$. We also consider the efficiency at maximum power, $\eta(P_{\text{max}})$, which is the efficiency when the values of $\Delta \mu_\alpha$ are chosen to maximize the power.

For a two-terminal system within the linear response regime, i.e. when the charge and heat currents depend linearly on the temperature and electrochemical potential differences, both the output power and efficiency can be written in terms of the transport coefficients, namely the electrical conductance $G$, the thermopower $S$ and the thermal conductance $K$, which will be defined in Eqs. (34), (35) and (36) by setting $\alpha = \beta = 2$. Defining $\Delta T \equiv \Delta T_2 > 0$, we have the following relations

$$P_{\text{max}} = \frac{1}{4} Q \Delta T^2,$$  

(21)

$$\eta(P_{\text{max}}) = \frac{\eta_C}{2} \frac{ZT}{ZT + 2},$$  

(22)

$$\eta_{\text{max}} = \eta_C \frac{\sqrt{1 + ZT} - 1}{\sqrt{1 + ZT} + 1},$$  

(23)

where $Q = GS^2$ is the power factor and $ZT = GS^2 T/K$ is the (dimensionless) figure of merit. As we can see in Eqs. (22) and (23), both $\eta(P_{\text{max}})$ and $\eta_{\text{max}}$ are monotonous growing functions of $ZT$; the only restriction imposed by thermodynamics is $ZT \geq 0$. When $ZT = 0$ both $\eta_{\text{max}}$ and $\eta(P_{\text{max}})$ vanish, while for $ZT \to \infty$, $\eta_{\text{max}} \to \eta_C$, and $\eta(P_{\text{max}}) \to \eta_{CA}$, where $\eta_{CA} = \eta_C/2$ is the so-called Curzon-Ahlborn efficiency in linear response.

III. LINEAR RESPONSE REGIME

As already mentioned above, in the linear response regime the applied temperature and electrochemical potential biases are small enough so that the currents depend linearly on them. Assuming that $|\Delta T_\alpha| \ll T$ and $|\Delta \mu_\alpha| \ll k_B T$, we follow Refs. 8 and 9 and suppose the probability $P\{\{n_i\}\}$ to differ from its equilibrium distribution $P_{\text{eq}}\{\{n_i\}\}$ in the following way:

$$P\{\{n_i\}\} = P_{\text{eq}}\{\{n_i\}\} [1 + \psi(\{n_i\})],$$  

(24)

where $\psi$ is a “small” function. In Eq. (24)

$$P_{\text{eq}}\{\{n_i\}\} = \frac{1}{Z} \exp \left[ - \frac{1}{k_B T} \left( \sum_{p=1}^\infty E_p n_p + U(N) - \mu N \right) \right]$$  

(25)

is the Gibbs distribution in the grand canonical ensemble, when all reservoirs have the same temperature and electrochemical potential, with grand partition function given by

$$Z = \sum_{\{n_i\}} \exp \left[ - \frac{1}{k_B T} \left( \sum_{p=1}^\infty E_p n_p + U(N) - \mu N \right) \right].$$  

(26)

In our expressions we will consider terms up to first order in $\psi$, $\Delta T_\alpha/T$, and $\Delta \mu_\alpha/k_B T$. By linearizing the LBEs with respect to the above small quantities, Eq. (18), one finds the relation
\[
\sum_{\{n_i\}, \eta_n} P_{\text{eq}}(\{n_i\}, n_p = 0)f \left( E_p + (2N + 1)E_C \right) \\
\times \sum_{\alpha} \Gamma_\alpha(p) \left\{ \psi(\{n_i\}, n_p = 0) - \psi(\{n_i\}, n_p = 1) + \frac{1}{k_BT} \left[ \left( E_p + (2N + 1)E_C - \mu \right) \frac{\Delta T_\alpha}{T} + \Delta \mu_\alpha \right] \right\} = 0, \quad (27)
\]

where \( f(E) \) stands for the Fermi distribution at temperature \( T \) and electrochemical potential \( \mu \). By expressing \( P(\{n_i\}) \) in terms of \( \psi(\{n_i\}) \) and linearizing Eq. (8), we can use Eq. (27) to remove \( \psi(\{n_i\}) \) from the charge current, and we find the following closed-form expression:

\[
J^c_\alpha = \frac{e}{k_BT} \sum_{p} \sum_{N=1}^{\infty} P_{\text{eq}}(N) F_{\text{eq}}(E_p|N) [1 - f(\epsilon(N,p))] \sum_{\beta} \frac{\Gamma_\alpha(p)\Gamma_\beta(p)}{\Gamma_{\text{tot}}(p)} \left[ (\epsilon(N,p) - \mu) \frac{\Delta T_\alpha - \Delta T_\beta}{T} + (\Delta \mu_\alpha - \Delta \mu_\beta) \right],
\]

where \( \Gamma_{\text{tot}}(p) = \sum_{\alpha} \Gamma_\alpha(p) \), and

\[
\epsilon(N,p) = E_p + U(N) - U(N-1) = E_p + (2N-1)E_C
\]

is the energy needed to add to the QD, initially occupied by \( N - 1 \) electrons, the \( N \)-th electron to level \( p \) (and equivalently for the inverse process). In Eq. (28) the quantity

\[
P_{\text{eq}}(N) = \sum_{\{n_i\}} P_{\text{eq}}(\{n_i\}) \delta_{\sum n_i, N}
\]

is the equilibrium probability of finding \( N \) electrons in the QD, and

\[
F(E_p|N) = \frac{P_{\text{eq}}(E_p \cap N)}{P_{\text{eq}}(N)} = \frac{\sum_{\{n_i\}} P_{\text{eq}}(\{n_i\}) \delta_{n_p,1} \delta_{\sum n_i, N}}{P_{\text{eq}}(N)}
\]

is the equilibrium conditional probability of finding level \( p \) occupied, when \( N \) electrons are in the QD. Note that this expression (28) coincides with the one that can be derived using the DBEs. However, the above derivation which uses the LBEs shows that Eq. (28) is always valid within the linear response regime.

Unfortunately, we were able to derive a closed-form expression for the energy current using Eq. (27) only in the absence of interaction \((E_C = 0)\). For \( E_C \neq 0 \), the energy current \( J^h_\alpha \) can be derived using the relation

\[
\sum_{\alpha} \Gamma_\alpha(p) \left\{ \psi(\{n_i\}, n_p = 0) - \psi(\{n_i\}, n_p = 1) + \frac{1}{k_BT} \left[ \left( E_p + (2N + 1)E_C - \mu \right) \frac{\Delta T_\alpha}{T} + \Delta \mu_\alpha \right] \right\} = 0, \quad (32)
\]

obtained by linearizing the DBEs, Eq. (15) (which is equivalent to removing the sum over \( \{n_i\} \) from Eq. (27)). Thus, in the domain of validity of the DBEs, the heat current can be written as

\[
J^h_\alpha = \frac{1}{k_BT} \sum_{p=1}^{\infty} \sum_{N=1}^{\infty} P_{\text{eq}}(N) F_{\text{eq}}(E_p|N) [1 - f(\epsilon(N,p))] [\epsilon(N,p) - \mu] \\
\times \sum_{\beta} \frac{\Gamma_\alpha(p)\Gamma_\beta(p)}{\Gamma_{\text{tot}}(p)} \left[ (\epsilon(N,p) - \mu) \frac{\Delta T_\alpha - \Delta T_\beta}{T} + (\Delta \mu_\alpha - \Delta \mu_\beta) \right]. \quad (33)
\]

We can now define the transport coefficients, namely the electrical conductance \( G_{\alpha\beta} \), the thermopower \( S_{\alpha\beta} \), and the thermal conductance \( K_{\alpha\beta} \) for the multi-terminal case, as

\[
S_{\alpha\beta} = -\left( \frac{\Delta \mu_\beta}{\epsilon \Delta T_\beta} \right) \frac{J^c_\alpha = 0 \forall \gamma,}{\Delta \mu_\gamma = 0 \forall \gamma \neq \beta} \quad (35)
\]

and

\[
G_{\alpha\beta} = \left( \frac{eJ^c_\alpha}{\Delta \mu_\beta} \right) \frac{\Delta T_\gamma = 0 \forall \gamma,}{\Delta \mu_\gamma = 0 \forall \gamma \neq \beta} \quad (34)
\]

and

\[
K_{\alpha\beta} = \left( \frac{J^h_\alpha}{\Delta T_\beta} \right) \frac{J^c_\gamma = 0 \forall \gamma,}{\Delta \mu_\gamma = 0 \forall \gamma \neq \beta} \quad (36)
\]
Note that index $\beta$ takes values in the range $2, ..., N$, since reservoir 1 is chosen as the reference. Local and non-local transport coefficients are distinguished depending on whether the two indices are, respectively, equal or different.

The expressions for the currents [(28) and (33)] have an intuitive interpretation. Indeed, the currents depend on the probability that a given energy level of the QD is occupied, $|P_{\alpha}(N)F_{\alpha}(E_{p}|N)|$ times the probability that there is an empty state with the correct energy in the reservoir $[1 - f(\epsilon(N, p))]$. The sum over all energy levels $p$ and over the total number of electrons $N$ in the QD accounts for all the various tunneling processes that can occur. Moreover, as far as the energy current is concerned, $\epsilon(N, p)$ is the energy carried by an electron that leaves the QD from level $p$ when $N$ electrons are present before the tunneling process, or equivalently, $\epsilon(N, p)$ is the energy carried by a electron that enters the QD into level $p$ increasing the number of total electrons to $N$. We recall that Eq. (28) for the charge current is always valid, while Eq. (33) holds only when the DBEs are valid (see App. C). The expressions for the charge and heat currents, in the case of two terminals, coincide with $\Gamma_{\alpha\beta}$, i.e. $\Gamma_{\alpha\beta}^{\text{loc}}$ and over the total number of electrons $N$.

If we assume that the tunneling rates do not depend on $p$, i.e. $\Gamma_{\alpha}(p) = \Gamma_{\alpha}$, we can rewrite the charge and heat currents in Eqs. (28) and (33) as follows:

$$J_{\alpha} = \frac{e}{k_{B}T} \sum_{\beta} \Gamma_{\alpha} \Gamma_{\beta}^{\text{tot}} \times \mathcal{P} \left[ (\epsilon - \mu) \frac{\Delta T_{\alpha} - \Delta T_{\beta}}{T} + \Delta \mu_{\alpha} - \Delta \mu_{\beta} \right],$$  

and

$$J_{\alpha}^{h} = \frac{1}{k_{B}T} \sum_{\beta} \Gamma_{\alpha} \Gamma_{\beta}^{\text{tot}} \times \mathcal{P} \left[ (\epsilon - \mu) \left( \frac{\Delta T_{\alpha} - \Delta T_{\beta}}{T} + \Delta \mu_{\alpha} - \Delta \mu_{\beta} \right) \right],$$

where $\mathcal{P}$ is the linear functional

$$\mathcal{P}[x] = \sum_{p=1}^{\infty} \sum_{N=1}^{\infty} P_{\alpha\beta}(N, p)x(N, p),$$

with

$$P_{\alpha\beta}(N, p) = P_{\alpha}(N)F_{\beta}(E_{p}|N)[1 - f(\epsilon(N, p))].$$

For a two terminal system it is possible to define analogous equations that do not require the tunneling rates to be energy-independent, but we will not consider this case. We can thus use the definitions of the transport coefficients given in Eqs. (34), (35) and (36) to write them in terms of the functional $\mathcal{P}$ as

$$G_{\alpha\beta} = \frac{e^{2}}{k_{B}T} \left( \delta_{\alpha\beta} \Gamma_{\alpha} - \frac{\Gamma_{\alpha} \Gamma_{\beta}^{\text{tot}}}{\Gamma_{\text{tot}}} \right) \mathcal{P}[1],$$

$$S_{\alpha\beta} = \frac{1}{eT} \delta_{\alpha\beta} \frac{\mathcal{P}[\epsilon - \mu]}{\mathcal{P}[1]},$$

$$K_{\alpha\beta} = \frac{1}{k_{B}T^{2}} \left( \delta_{\alpha\beta} \Gamma_{\alpha} - \frac{\Gamma_{\alpha} \Gamma_{\beta}^{\text{tot}}}{\Gamma_{\text{tot}}} \right) \times \left( \mathcal{P}[\epsilon - \mu^{2}] - \frac{\mathcal{P}^{2}[\epsilon - \mu]}{\mathcal{P}[1]} \right).$$

These expressions only require the calculation of $\mathcal{P}[\epsilon - \mu^{k}]$, with $k = 0, 1, 2$, and make manifest various properties of the transport coefficients. Namely, i) all three transport coefficients are symmetric matrices (as required by the Onsager relations in the presence of time-reversal symmetry); ii) $G_{\alpha\beta}$ and $K_{\alpha\beta}$ have non-local terms, while the thermopower is only local (non-zero non-local $S_{\alpha\beta}$ occur when relaxing the assumption for which the tunneling rates do not depend on the energy levels); iii) $\sum_{\alpha} G_{\alpha\beta} = \sum_{\alpha} K_{\alpha\beta} = 0$, stemming from charge and energy conservation.

By defining the generating function

$$\Omega[\lambda] = \ln \mathcal{P} \left[ e^{\lambda \epsilon - \mu} \right],$$

we can write the transport coefficients as follows:

$$G_{\alpha\beta} = \frac{e^{2}}{k_{B}T} \left( \delta_{\alpha\beta} \Gamma_{\alpha} - \frac{\Gamma_{\alpha} \Gamma_{\beta}^{\text{tot}}}{\Gamma_{\text{tot}}} \right) e^{\Omega[0]},$$

$$S_{\alpha\beta} = \frac{1}{eT} \delta_{\alpha\beta} \frac{\partial \Omega}{\partial \lambda} \bigg|_{\lambda=0},$$

$$K_{\alpha\beta} = \frac{1}{k_{B}T^{2}} \left( \delta_{\alpha\beta} \Gamma_{\alpha} - \frac{\Gamma_{\alpha} \Gamma_{\beta}^{\text{tot}}}{\Gamma_{\text{tot}}} \right) e^{\Omega[0]} \frac{\partial^{2} \Omega}{\partial \lambda^{2}} \bigg|_{\lambda=0}. $$

In the next Subsection, we will compute an analytic expression for $\Omega[\lambda]$ in the quantum limit.

A. Quantum limit

The quantum limit is characterized by having the energy spacing between levels of the QD and the charging energy much bigger than $k_{B}T$ [while $k_{B}T \gg \hbar \Gamma_{\alpha}(p)$]. We start by observing that the sum over $p$ and $N$ in Eq. (39) accounts for the fact that electrons can enter or leave the QD with energy $\epsilon(N, p)$ through, in principle, any energy level $E_{p}$ with the QD being occupied by any number of electrons $N$. The transition energies $\epsilon(N, p)$ are schematically shown in Fig. 2. At low temperatures we expect the lowest energy levels of the QD to be occupied, so that, if there are initially $N - 1$ electrons in the QD, electrons can flow mainly through level $p = N$. Such process gives the dominant contribution to transport and is represented by the dominant transition energy $\tilde{\epsilon}(N) \equiv \epsilon(N, p = N)$ (depicted in bold in Fig. 2). Therefore, in the quantum limit one expects to get the
and for values of $p$ such that $\epsilon(N_{\text{min}}, p)$ is between the electrochemical potential $\mu$ and the dominant transition energy, i.e. such that $\tilde{\epsilon}(N_{\text{min}}) \leq \epsilon(N_{\text{min}}, p) < \mu$ or $\mu < \epsilon(N_{\text{min}}, p) \leq \tilde{\epsilon}(N_{\text{min}})$. In the former case, $p = N_{\text{min}}$, $N_{\text{min}} + 1$, $N_{\text{min}} + 2, \ldots, \bar{p}$, where $\bar{p}$ is the largest integer such that $\epsilon(N_{\text{min}}, p) < \mu$. In the latter case, $p = N_{\text{min}}, N_{\text{min}} - 1, N_{\text{min}} - 2, \ldots, \bar{p}$ where $\bar{p}$ is the smallest integer such that $\epsilon(N_{\text{min}}, p) > \mu$. We then find that Eq. (39) becomes

$$P[x] = \frac{1}{4 \cosh^2 \left( \frac{\Delta_{\text{min}}}{2k_BT} \right)} \sum_{p=N_{\text{min}}}^{\bar{p}} x(N_{\text{min}}, p),$$

where we have defined $\Delta_{\text{min}} \equiv \tilde{\epsilon}(N_{\text{min}}) - \mu$. Eq. (45) only keeps the dominant terms in the low-temperature limit.

This approximation must be improved when $\bar{p} = N_{\text{min}}$, that is, the sum in $P$ in Eq. (45) reduces to the single term $p = N_{\text{min}}$ and transport is provided by the dominant transition energy $\tilde{\epsilon}(N_{\text{min}})$ only. If in this case one imposes that $J^\alpha_0 = 0$, then one obtains $J^h_0 = 0$, since $J^h_\alpha \propto J^\alpha_{\bar{p}}$. As a consequence one gets $K_{\alpha\beta} = 0^{110}$. Thus, when $\bar{p} = N_{\text{min}}$, we improve our approximation of $P$ by extending the sum over $p$ to the two nearest integers, $p = N_{\text{min}} \pm 1$. We have numerically verified that this approximation is valid when $2E_C > \Delta E$.

In order to obtain analytical expressions for the transport coefficients, hereafter we focus on the case of equidistant levels, $E_p - E_{p-1} = \Delta E$. After introducing the parameter

$$\xi \equiv \frac{4 \cosh^2 \left( \frac{\Delta_{\text{min}}}{2k_BT} \right)}{e^{\Delta E/k_BT}}$$

and defining the integer $N_J \equiv \bar{p} - N_{\text{min}}$, we obtain

$$\Omega[\lambda] = -\ln \left[ 4 \cosh^2 \left( \frac{\Delta_{\text{min}}}{2k_BT} \right) \right] + \lambda \Delta_{\text{min}}$$

$$+ \ln \left[ \frac{e^{\Delta E(N_J + 1)\text{sign}(N_J)} - 1}{\text{sign}(N_J)\Delta E - 1} \right] + \xi \delta_{N_J,0} \left( \cosh(\lambda \Delta E) - \tanh \left( \frac{\Delta_{\text{min}}}{2k_BT} \right) \sinh(\lambda \Delta E) \right).$$

Using Eqs. (43) and (47), we finally obtain the multiterminal transport coefficients

$$G_{\alpha\beta} = \left( \delta_{\alpha\beta} \Gamma_\alpha - \frac{\Gamma_\alpha \Gamma_\beta}{\Gamma_{\text{tot}}} \right) \frac{e^2}{4k_BT \cosh^2 \left( \frac{\Delta_{\text{min}}}{2k_BT} \right)} (1 + |N_J|),$$

$$S_{\alpha\beta} = \delta_{\alpha\beta} \frac{1}{e^T} \left( \Delta_{\text{min}} + \frac{\Delta E}{2} N_J \right),$$

$$K_{\alpha\beta} = \left( \delta_{\alpha\beta} \Gamma_\alpha - \frac{\Gamma_\alpha \Gamma_\beta}{\Gamma_{\text{tot}}} \right) k_B \left( \frac{\Delta E}{k_BT} \right)^2 \left\{ \frac{1}{2} e^{-|\Delta_{\text{min}}/k_BT| N_J} \left( |N_J| + 1 \right) \left( |N_J| + 2 \right) \right. \text{ if } N_J \neq 0,$$

$$\left. -\frac{1}{e^{\Delta E/k_BT} + 4 \cosh^2 \left( \frac{\Delta_{\text{min}}}{2k_BT} \right)} \right\} \text{ if } N_J = 0.$$

We have computed $G_{\alpha\beta}$ and $S_{\alpha\beta}$ setting $\xi = 0$, since the term proportional to $\xi$ in Eq. (47) only yields minor corrections that make these quantities more “smooth” as a function of $\mu$; the calculation of $K_{\alpha\beta}$ instead requires a non null value of $\xi$. Eqs. (48) exhibit a number of interesting features. First of all, $G_{\alpha\beta}$ shows peaks as a function of the electrochemical potential $\mu$ every time $\Delta_{\text{min}} = 0$, namely when $\mu = \tilde{\epsilon}(N)$. For example, the $N$-th peak corresponds to $\mu$ equal to the $N$-th dominant transition energy,

$$\mu = \mu_N \equiv (N - 1)\Delta E + (2N - 1)E_C.$$
We set $E_p = (p - 1)\Delta E$, for $p = 1, 2, \ldots$, and therefore the separation between two nearby peaks is given by $\Delta E + 2E_C$. Due to the factor $\cosh^{-2}(\Delta_{\text{min}}/2k_B T)$ in $G_{\alpha\beta}$, these peaks have a bell shape with amplitude of the order of $k_B T$. On the other hand, the thermal conductance $K_{\alpha\beta}$ has plateaus of width $2\Delta E$ around $\mu_N^{15}$, corresponding to the second line of the expression of $K_{\alpha\beta}$ for $N_J = 0$. For $N_J \neq 0$, the thermal conductance $K_{\alpha\beta}$ is then exponentially suppressed due to the term $e^{-\Delta_{\text{min}}/k_B T}$. The local thermopower $S_{\alpha\alpha}$ vanishes at the values $\mu_N$ where the electrical and thermal conductances $G_{\alpha\alpha}$ and $K_{\alpha\alpha}$ exhibit a maximum. $S_{\alpha\alpha}$ has a linear dependence on $\mu$ with slope $dS_{\alpha\alpha}/d\mu = -1/eT$, with jumps when either $N_{\text{min}}$ or $N_J$ change by one. Therefore we have, for $E_C \gg \Delta E$, main oscillations of period $\Delta E + 2E_C$, and a fine structure with spacing $\Delta E^9$. We note that the fine structure is present also for $G_{\alpha\beta}$ and $K_{\alpha\beta}$, but in these cases the amplitude of the fine structure oscillations is exponentially small.

### B. Two-terminal system

In the two-terminal case, the matrices $G_{\alpha\beta}, S_{\alpha\beta},$ and $K_{\alpha\beta}$ reduce to the familiar transport coefficients, namely the electrical conductance $G = G_{22}$, and the thermal conductance $K = K_{22}$. From Eqs. (48) we recover the formulas for $G$ and $S$ well-known in literature$^9$, while our expression for $K$ coincides for $N_J = 0$ with the result of Ref. 15, but also provides the fine structure oscillations for $N_J \neq 0$. Although such oscillations are not appreciable in $K$ as a function of $\mu$, we will see below that they give rise to a visible fine structure in $ZT$.

As we have shown in Sec. II E, within the linear response regime the relevant quantities to characterize the performance of a thermoelectric device are the power factor $Q$ and the figure of merit $ZT$. From the expressions of the transport coefficients in Eqs. (48), specified for the two-terminal case, one can compute $Q$ and $ZT$ analytically within the quantum limit. The obtained expressions are given below, and compared with a numerical calculation performed using the kinetic equations.

#### 1. Power factor

Let us start by studying the power factor $Q$. Within the quantum limit, we find that

$$Q = \frac{\gamma (1 + |N_J|)}{4k_B T^3 \cosh^2(\Delta_{\text{min}}/2k_B T)} \left( \Delta_{\text{min}} + \frac{N_J}{2} \Delta E \right)^2,$$

where we have defined $\gamma = \Gamma_1 \Gamma_2 / (\Gamma_1 + \Gamma_2)$. As for $G$, the power factor $Q$ is dominated by a fast decrease, given by the term $\cosh^{-2}(\Delta_{\text{min}}/2k_B T)$, thus becoming vanishingly small within a few $k_B T$ around $\mu = \mu_N$ [see Fig. 3(a)]. In fact the fine structure, given by the terms with $N_J$, is not visible in Fig. 3(a) due to the rapid suppression given by the $\cosh^{-2}(\Delta_{\text{min}}/2k_B T)$ term in Eq. (50). Differently from $G$, the power factor vanishes at $\mu = \mu_N$, due to the fact that in this point the thermopower $S = 0$. So as $\mu$ moves away from $\mu_N$, $Q$ increases quadratically due to the linear growth of the thermopower with $\mu$, and then it rapidly decreases within a few $k_B T$ due to the $\cosh^{-2}(\Delta_{\text{min}}/2k_B T)$ term. Hence there are two symmetric peaks around $\mu = \mu_N$, within a few $k_B T$. These double peaks are the dominant feature of Fig. 3(a) and identify the optimal values of $\Delta_{\text{min}}$ (and consequently of $\mu$) to obtain the absolute maximum power $P_{\text{peak}}$, namely when the power factor $Q$ is maximum, $Q = Q^*$. From Eq. (50), we obtain that $Q$ is maximum for values $\Delta_{\text{min}}$ of $\Delta_{\text{min}}$ such that

$$\frac{\Delta_{\text{min}}}{2k_B T} = \coth \left( \frac{\Delta_{\text{min}}}{2k_B T} \right).$$

The numerical solution is $\Delta_{\text{min}}^* \approx \pm 2.40k_B T$, which corresponds to $\mu = \mu_N \pm 2.40k_B T$. This result does not depend on any energy scale of the system except for $k_B T$, and coincides with the non-interacting single-level case.
(see App. D). The value $Q^*$ of $Q$ in these points is

$$ Q^* \simeq 0.44 \frac{\gamma k_B}{T}, \quad (52) $$

so that the peaks of the power factor only depend on $\gamma$ and on the reference temperature. In conclusion, if we want to extract maximum power from this system, we must choose $\mu = \mu_N \pm 2.40 k_B T$. We will now show that also $ZT$ reaches a maximum at these same values of the electrochemical potential confirming that these are the optimal values for heat to work conversion in the quantum limit linear response regime.

2. Figure of merit

Let us now study the figure of merit $ZT$ in the quantum limit. To obtain a more manageable analytical expression, we compute $K$ from the function $\Omega[\lambda]$ expanded to the first order in $\xi$ [this corresponds to approximating $K$ with a constant plateau when $N_J = 0$, namely $4 \cosh^2(\Delta_{\text{min}}/2k_B T)$ is neglected with respect to $e^{\Delta E/k_B T}$ in the last line of Eqs. (48)]. We then obtain

$$ ZT = \begin{cases} 
\frac{1}{4} \left( \frac{\Delta_{\text{min}}}{\Delta E} \right)^2 e^{\Delta E / k_B T} \cosh^2 \left( \frac{\Delta_{\text{min}}}{2k_B T} \right) & \text{if } N_J = 0, \\
\frac{3|N_J|}{2 + |N_J|} \left( 1 - 2 \frac{\Delta_{\text{min}}}{\Delta E} |N_J| \right)^2 & \text{if } N_J \neq 0,
\end{cases} \quad (53) $$

which implies that the behavior of $ZT$ is different for the two cases $|\mu - \mu_N| < \Delta E$ ($N_J = 0$) and $|\mu - \mu_N| > \Delta E$ ($N_J \neq 0$). In the former case $K$ exhibits a plateau, so that $ZT$ is directly proportional to $Q$ and therefore it has the same double peak structure at $\mu = \mu_N \pm 2.40 k_B T$. This is clearly shown in Fig. 3(b) where $ZT$ is plotted as a function of $\mu$.

The value of $ZT$ in these points is

$$ ZT^* \approx 0.44 e^{\Delta E / k_B T} (\Delta E / k_B T)^2. \quad (54) $$

This results has been obtained also in Ref. 98. Eq. (54) shows that in the limit $\Delta E / k_B T \to \infty$, we have that $ZT \to \infty$. For example, for $\Delta E = 6k_B T$, we reach $ZT^* \approx 5$; for $\Delta E = 10k_B T$, we reach $ZT^* \approx 97$, and so on. This is consistent with Mahan and Sofo’s observation that a narrow transmission function yields $ZT \to \infty$. Furthermore, these peaks in $ZT$ correspond to peaks in $Q$, so in these points we can maximize $P_{\text{max}}$ and $\eta(P_{\text{max}})$ simultaneously. Instead, when $N_J \neq 0$, $ZT$ has a discontinuity every time $\mu = \epsilon(N,p)$ with $p \neq N$, which means with a $\Delta E$ spacing. This fine structure is the origin of the saw-tooth oscillations of Fig. 3(b). The value of $ZT$ in each $\mu = \epsilon(N,p)$ is given by

$$ ZT_p = \begin{cases} 
\frac{3}{|N - p|} + 1 & \text{if } |N - p| \geq 2, \\
1 & \text{if } |N - p| = 1.
\end{cases} \quad (55) $$

The height of these peaks, as opposed to $ZT^*$, has no dependence on the parameters of the system. The highest peak is obtained for $|N - p| = 2$, where $ZT_{p = N \pm 2} = 9$. For values of $p$ distant from $N$, the height of the peak decreases to an asymptotic value of $ZT_{\infty} = 3$. The comparison is performed within the linear response quantum limit for a two-terminal setup. The generating function in Eq. (47) cannot be used in the case $E_C = 0$ since it requires $2E_C > \Delta E \gg k_B T$. The generating function for $E_C = 0$ will be denoted as $\Omega_{\text{NI}}[\lambda]$ (where NI stands for “non interacting”) and calculated as follows. As we can see from Eq. (42), we must compute $P[\exp \{\lambda(\epsilon(N,p) - \mu)\}]$ using the definition of $P$ given in Eq. (39). When $E_C = 0$, $\epsilon(N,p) = E_p$, so the transition energies correspond to the energy levels of the QD, and they do not depend on the number of electrons in the QD. As a consequence, there is no dependence on $N$ in the argument of $P$, so we can explicitly perform the sum over $N$ in Eq. (39) yielding the following expression:

$$ P_{\text{NI}}[x(p)] = \sum_{p=1}^{\infty} \frac{1}{4 \cosh^2 [(E_p - \mu)/2k_B T]} x(p). \quad (56) $$

Each term in the series in Eq. (56), as a function of $\mu$, is a bell-shaped function centered around $E_p$ of width set by $k_B T$. Therefore, within the quantum limit we can restrict the sum over $p$ to the three bell-shaped functions closest to $\mu$, namely $p = N_{\text{min}}, N_{\text{min}} \pm 1$ $(N_{\text{min}}$, defined as in Eq. (44), is such that $E_{N_{\text{min}}}$ is the closest energy level to $\mu$). This approximation allows us to find

$$ \Omega_{\text{NI}}[\lambda] = - \ln \left[ 4 \cosh^2 \left( \frac{\Delta_{\text{min}}}{2k_B T} \right) \right] + \lambda \Delta_{\text{min}} + \ln \left[ 1 + 2 \xi \cosh \left( \lambda \Delta E - \frac{\Delta_{\text{min}}}{k_B T} \right) \right], \quad (57) $$

where $\Delta_{\text{min}}$ in the non-interacting case reduces to the distance between $\mu$ and the nearest energy level: $\Delta_{\text{min}} = E_{N_{\text{min}}} - \mu$.

Let us now compare Eq. (57) with its interacting counterpart, Eq. (47), setting $N_J = 0$, thus neglecting the fine structure. We notice that the two expressions are identical when $\xi = 0$; this implies that $G_{\alpha\beta}$ and $S_{\alpha\beta}$ are equal in the interacting and non-interacting case [in Eqs. (48) $G_{\alpha\beta}$ and $S_{\alpha\beta}$ are calculated at $\xi = 0$], while $K_{\alpha\beta}$ is different in the two cases, since it is determined by the term
FIG. 4: (Color online) Comparison between $K^{(NI)} = K^{(NI)}_{22}$ and $K = K_{22}$, given respectively by Eqs. (58) and (48), plotted as a function of $\mu$. Both cases have been computed with the parameters used in Fig. 3, except for setting $E_C = 0$ in the non-interacting case. The interacting thermal conductance has been multiplied by a factor 70. In particular, its maximum value is half the minimum of $K^{(NI)}$.

A comparison between the interacting and non-interacting thermal conductances is plotted for a two-terminal system in Fig. 4, using Eqs. (48) and (58) at equal $\Delta E = 10k_B T$. In this figure the dominant transition energy of the interacting system is located, see Eq. (49), at $\mu = \mu_{N=2} = 160k_B T$, while the energy levels of the non-interacting system are chosen as $E_p = (p - 1)10k_B T$ so that in both cases $G$ has a peak in $\mu = 160k_B T$. We have verified that a numerical calculation is in very good agreement with Eq. (58) using the parameters of Fig. 4. As we can see from Fig. 4, $K^{(NI)} = K^{(NI)}_{22}$ and $K = K_{22}$ are very different. While $K$ has a plateau of width $2\Delta E$ centered in $\mu_{N=2}$, $K^{(NI)}$ has a minimum in $\mu_{N=2}$, and reaches a maximum value for $\mu$ between two energy levels. Furthermore, by comparing Eq. (58) with the last line of Eqs. (48), we see that $K^{(NI)} = C \cosh (\Delta_{\min}/k_B T)K$, so the minimum of $K^{(NI)}$, occurring at $\Delta_{\min} = 0$, is twice the maximum of $K$, and $K^{(NI)}$ increases exponentially with respect to $K$ as $\Delta_{\min}$ increases.

Intuitively, the striking difference between the two models can be explained as follows. As discussed in App. D, if we consider a single energy level QD the thermal conductance vanishes ($K = 0$) since $K$ is computed at zero charge current and charge and heat currents are proportional in this case. However, $K$ can be finite when at least two energy levels are available, and gets bigger by increasing the flux of electrons tunneling at different energies. Now, Coulomb interaction produces a correlation between electrons tunneling at different energies. Namely, if one electron enters the QD the electrostatic energy increases by $2E_C$, preventing other electrons from entering the QD at any other energy level. Therefore, until that electron tunnels out of the QD, all other processes are blocked: this is a manifestation of Coulomb blockade. On the contrary, in the non-interacting case all tunneling events are independent. This correlation is thus responsible for suppressing simultaneous tunneling through different energy levels in the interacting case, which results in a suppression of $K$.

So in general $K$ is much smaller than $K^{(NI)}$. As a consequence of these observations, we expect $ZT$ to be suppressed in the non-interacting case.

By setting $\xi = 0$ in Eq. (58), we find

$$ZT_{NI} = \frac{1}{8} \left( \frac{\Delta_{\min}}{\Delta E} \right)^2 e^{\Delta E/k_B T} \frac{e^{\Delta E/k_B T}}{\cosh^2 \left( \frac{\Delta_{\min}}{2k_B T} \right) \cosh \left( \frac{\Delta_{\min}}{k_B T} \right)}.$$  \hspace{5cm} (59)

Comparing Eq. (59) with Eq. (53), we see that, for $N_J = 0$, $ZT = 2 \cosh (\Delta_{\min}/k_B T)ZT_{NI}$, so $ZT_{NI}$ is exponentially suppressed as $\Delta_{\min}$ increases. Given this suppression, the maximum of $ZT_{NI}$ occurs at $\Delta_{\min} \approx \pm 1.36k_B T$, corresponding to

$$ZT_{NI} \approx \frac{1}{13.8} e^{\Delta E/k_B T}.$$  \hspace{5cm} (60)

This value is approximately 6 times smaller than $ZT^*$ for the interacting case, see Eq. (54).

Furthermore, since $K$ is “flat” around the dominant transition energies in the interacting case, the peak power $P_{peak}$ and the figure of merit $ZT^*$ are reached at the same electrochemical potential, $\Delta_{\min} \approx \pm 2.40k_B T$, so $ZT(P_{peak}) = ZT^*$. Instead in the non-interacting case these two quantities are not simultaneously maximized, due to the strong dependance of $K^{(NI)}$ on $\mu$, so we have that

$$ZT_{NI}(P_{peak}) \approx \frac{1}{25.3} e^{\Delta E/k_B T},$$

which is approximately 11 times smaller than $ZT(P_{peak})$, see Eq. (54).

In conclusion, within the linear response quantum limit, an interacting QD (with $2E_C > \Delta E$) has a considerably higher $ZT$ with respect to a non-interacting QD both at peak efficiency, and at peak power, while having the same power factor ($G$ and $S$ being equal).

At last we will study how these two models violate the Wiedemann-Franz law, which states that for macroscopic ordinary metals the Lorenz ratio $L = K/GT$ is a constant equal to the Lorenz number $L_0 = (k_B/e)^2(\pi^2/3)$. Using
Eqs. (48) and (58), we find that
\[ L = \frac{L_0}{\pi^2} \left( \frac{\Delta E}{k_B T} \right)^2 \left\{ |N_J| |N_J + 2| \right\} \text{ if } N_J \neq 0,
12e^{-\frac{\Delta E}{k_B T}} \cosh^2 \frac{\Delta_{\min}}{2k_BT} \text{ if } N_J = 0, \]
and
\[ L_{NI} = \frac{L_0}{\pi^2} \left( \frac{\Delta E}{k_B T} \right)^2 \frac{24}{\pi^2 e^{-\Delta E/k_B T}} \cosh^2 \frac{\Delta_{\min}}{2k_BT} \cosh \frac{\Delta_{\min}}{k_BT}, \]
where \( L \) refers to the interacting case, and the expression of \( K \), for simplicity, has been computed at \( \xi = 0 \). In both cases the Wiedemann-Franz law is strongly violated: at \( \Delta_{\min} = 0 \), the Lorenz ratio is exponentially smaller than \( L_0 \) thanks to \( (\Delta E/k_B T)^2 \exp(-\Delta E/k_B T) \) (this has been noticed in Ref. 98 for the interacting model). In both cases the Lorenz ratio exponentially increases with \( \Delta_{\min} \). In the non-interacting model the exponent is \( 2\Delta_{\min}/(k_B T) \) (twice the interacting case), and the maximum value, achieved at \( \Delta_{\min} = \Delta E/2 \), is of the order of \( L_{NI} \approx (\Delta E/k_B T)^2 L_0 \). Interestingly, in the interacting case [Eq. (60)], when \( |\Delta_{\min}| > \Delta E \), i.e. \( N_J \neq 0 \), we find plateaus whose height increases with \( N_J \).

D. Three-terminal system

In this Section we consider the case of a three-terminal system, which allows to study the non-local transport coefficients and the influence of an additional terminal on the thermoelectric performance of the system. We focus on the simplest case when the couplings to the reservoirs are energy independent, that is, the rates \( \Gamma_\alpha \) do not depend on \( p \), and we will consider an equidistant QD spectrum. Analytical expressions in the quantum limit for the power and the efficiency at maximum power can be obtained also in this case on the basis of the expressions written in Ref. 53 and of Eqs. (48) for the transport coefficients. All considerations made in this Section are valid for both interacting and non-interacting systems.

In a three-terminal setup with time reversal symmetry we have nine independent coefficients: three electrical conductances \( G_{22}, G_{33}, \) and \( G_{23} \), three thermopowers \( S_{22}, S_{33}, \) and \( S_{23} \), and three thermal conductances \( K_{22}, K_{33}, \) and \( K_{23} \). According to the expressions in Eqs. (48), valid in the quantum limit when the couplings to the leads are independent of energy, one finds that the local and non-local electrical (thermal) conductances are characterized by peaks (plateaus) located in the same positions as for the two-terminal case. The two local electrical (thermal) conductances \( G_{22}, G_{33}, (K_{22}, K_{33}) \), can have different heights if the coupling to reservoirs 2 and 3 are different \( (\Gamma_2 \neq \Gamma_3) \), while the non-local conductances, \( G_{23} \) and \( K_{23} \), are negative. On the other hand, the local thermopowers \( S_{22} \) and \( S_{33} \) are equal and exhibit the same fine structure as in the two-terminal case. Moreover, the non-local thermopower \( S_{23} \) vanishes, as a consequence of energy independent tunneling rates.

The power and efficiency of a three-terminal system are defined in Eqs. (19) and (20). For definiteness, let’s consider \( T_2 \geq T_3 \geq T_1 \), so \( \Delta T_2 \geq \Delta T_3 \geq 0 \). In general, Carnot’s efficiency cannot be written only in terms of the temperatures \( T_1, T_2, \) and \( T_3 \), but it depends on the details of the system. Nonetheless, if we fix the temperature of the hottest and coldest reservoir, that is, \( T_2 \) and \( T_3 \), it can be shown that \( \eta_C \leq \eta_C^{(2)} \), where \( \eta_C^{(2)} = 1 - (T_1/T_3) \) is the Carnot efficiency of the two-terminal system; the equal sign can be achieved when two reservoirs have the same temperature. So adding a third terminal at an intermediate temperature cannot increase the maximum efficiency beyond the two terminal Carnot’s efficiency.

Also in the three terminal case the efficiency at maximum power cannot go beyond the linear response Curzon-Ahlborn efficiency \( \eta_{CA} = \eta_C^{(2)}/2 \). Our aim is to maximize the power \( P \) with respect to \( \Delta \mu_2 \) and \( \Delta \mu_3 \), at given temperature differences. Then we will consider a fixed value of \( \Delta T_3 \), and we will study the maximum power \( P_{\text{max}} \) and the efficiency at maximum power \( \eta (P_{\text{max}}) \) varying \( T_2 \) between the fixed \( T_1 \) and \( T_3 \). These calculations can be performed by writing the currents in terms of the temperature and electrochemical potential differences through the Onsager matrix \( L_{ij} \); in turn, \( L_{ij} \) can be related to the transport coefficients. By rewriting Eqs. (43) as
\[ G_{\alpha \beta} = M_{\alpha \beta} G, 
S_{\alpha \beta} = \delta_{\alpha \beta} S, 
K_{\alpha \beta} = M_{\alpha \beta} K, \]
where
\[ M_{\alpha \beta} = \delta_{\alpha \beta} \Gamma_\alpha - \frac{\Gamma_\alpha \Gamma_\beta}{\Gamma_{\text{tot}}}, \]
and by defining
\[ ZT = \frac{G S^2 T}{K}, \]
\[ \Gamma_{ij} = \Gamma_i + \Gamma_j, \]
we can write the currents as
\[ \left( \frac{J^i_2 / \Gamma_2}{J^i_3 / \Gamma_3} \right) = \frac{G}{e \Gamma_{\text{tot}}} \begin{pmatrix} \Gamma_{13} & \Gamma_{13} & -\Gamma_3 & -\Gamma_3 \\ -\Gamma_2 & -\Gamma_2 & \Gamma_{12} & \Gamma_{12} \end{pmatrix} \left( \begin{pmatrix} \Delta \mu_2 \\ \Delta \mu_3 \\ e S \Delta T_2 \\ e S \Delta T_3 \end{pmatrix} \right), \]
\[ \left( \frac{J^i_2 / \Gamma_2}{J^i_3 / \Gamma_3} \right) = \frac{K}{\Gamma_{\text{tot}}} \begin{pmatrix} \Gamma_{13} & \Gamma_{13} & -\Gamma_3 & -\Gamma_3 \\ -\Gamma_2 & -\Gamma_2 & \Gamma_{12} & \Gamma_{12} \end{pmatrix} \left( \begin{pmatrix} 0 \\ \Delta T_2 \\ 0 \\ \Delta T_3 \end{pmatrix} \right) + \frac{ZT}{\Gamma_{\text{tot}}} \left( \begin{pmatrix} J^i_2 / \Gamma_2 \\ J^i_3 / \Gamma_3 \end{pmatrix} \right). \]
Note that the quantities \( G, S, K \) and \( ZT \) only depend on the properties of the QD, which can be interacting or
non-interacting, and on the reference electrochemical potential $\mu$; they do not depend on the number of reservoirs nor on the tunneling rates.

The electrochemical potential differences that maximize $P$ at given reservoir temperatures can be written as

$$\Delta \mu_\alpha = -\frac{1}{2}e\Delta T_\alpha.$$  \hspace{1cm} (67)

Inserting these expressions into $P$ yields

$$P_{\text{max}} = \frac{1}{4} \frac{Q}{\Gamma_{\text{tot}}} \times \left[ \Gamma_1 \Gamma_2 \Delta T_2^3 + \Gamma_1 \Gamma_3 \Delta T_3^2 + \Gamma_2 \Gamma_3 (\Delta T_3 - \Delta T_2)^2 \right],$$ \hspace{1cm} (68)

where $Q = GS^2$. The maximum power is thus an always positive quantity and, in the same manner as in the two-terminal case, see Eq. (21), it is proportional to $Q$ and quadratic in the temperature differences. Furthermore, the properties of the QD and the chosen $\mu$ are all contained in the $Q$ term, while the coupling to the reservoirs and the temperature differences are separately accounted for in the term between square parentheses in Eq. (68). The efficiency at maximum power instead is given by

$$\eta(P_{\text{max}}) = \frac{\eta_C}{2} \frac{ZT}{ZT + 2} \left[ \Gamma_1 \Gamma_2 \Delta T_2^3 + \Gamma_1 \Gamma_3 \Delta T_3^2 + \Gamma_2 \Gamma_3 (\Delta T_3 - \Delta T_2)^2 \right] \times \begin{cases} \frac{1}{\Gamma_1 \Delta T_3 (\Gamma_2 \Delta T_2 + \Gamma_3 \Delta T_3)} & \text{if } \frac{\Gamma_3}{\Gamma_1 + \Gamma_3} \Delta T_3 \leq \Delta T_2 \leq \Delta T_3, \\ \frac{1}{\Gamma_3 \Delta T_3 (\Gamma_1 \Delta T_3 + \Gamma_2 (\Delta T_3 - \Delta T_2))} & \text{if } 0 \leq \Delta T_2 \leq \frac{\Gamma_3}{\Gamma_1 + \Gamma_3} \Delta T_3. \end{cases}$$ \hspace{1cm} (69)

It is interesting to notice that in this equation the term before square parentheses, which does not depend on $\Gamma_\alpha$, is exactly equal to the two terminal efficiency at maximum power, see Eq. (22), since $ZT = ZT$ for a two terminal system. The remaining part of Eq. (69) takes into account the particular temperatures and couplings to the three terminals. Furthermore the efficiency at maximum power, also in this 3 terminal system, only depends on $ZT$.

There are three limiting cases we will first study: $\Delta T_2 = 0$, $\Delta T_2 = \Delta T_3$ and $\Delta T_2 = \Delta T_3 \Gamma_4/(\Gamma_1 + \Gamma_3)$.

If $\Delta T_2 = 0$, $T_1 = T_2$, so we have one hot reservoir at temperature $T_3$ and two cold ones at the same temperature. In this case we obtain

$$P_{\text{max}} = \frac{1}{4} Q \Delta T_3^2 \frac{(\Gamma_1 + \Gamma_2) \Gamma_3}{(\Gamma_1 + \Gamma_2) + \Gamma_3},$$ \hspace{1cm} (70)

$$\eta(P_{\text{max}}) = \eta_C \frac{ZT}{ZT + 2}.$$ \hspace{1cm} (71)

Note that for a two-terminal system the maximum power, obtained by inserting Eq. (62) into Eq. (21), is given by

$$P_{\text{max}}^{(2)} = \frac{1}{4} Q \Delta T_2^2 \frac{\Gamma_1 \Gamma_2}{\Gamma_1 + \Gamma_2}.$$ \hspace{1cm} (72)

Comparing Eq. (72) with Eq. (70), and Eq. (22) with Eq. (71), we notice that the 3 terminal system is formally equivalent to a 2-terminal system with temperature difference $\Delta T_3$, with tunneling rate $\Gamma_1 + \Gamma_2$ instead of $\Gamma_1$, and $\Gamma_3$ instead of $\Gamma_2$.

If $\Delta T_2 = \Delta T_3$, $T_2 = T_3$, so we have two hot reservoir at temperature $T_3$ and one cold reservoir at temperature $T_1$. In this case we obtain

$$P_{\text{max}} = \frac{1}{4} Q \Delta T_3^2 \frac{\Gamma_1 (\Gamma_2 + \Gamma_3)}{(\Gamma_1 + \Gamma_2 + \Gamma_3)},$$ \hspace{1cm} (73)

$$\eta(P_{\text{max}}) = \eta_C \frac{ZT}{ZT + 2}.$$ \hspace{1cm} (74)

As in the previous limiting case, this system behaves like a 2-terminal system with temperature difference $\Delta T_3$ and with tunneling rate $\Gamma_2 + \Gamma_3$ instead of $\Gamma_2$.

If $\Delta T_2 = \Delta T_3 \Gamma_3/(\Gamma_1 + \Gamma_3)$, reservoir 2 has an intermediate temperature such that $J_2^h = 0$. In fact this specific value of $\Delta T_3$ distinguishes the two regimes where $J_2^h > 0$ and $J_2^h < 0$. In this case we obtain

$$P_{\text{max}} = \frac{1}{4} Q \Delta T_3^2 \frac{\Gamma_1 \Gamma_3}{\Gamma_1 + \Gamma_3},$$ \hspace{1cm} (75)

$$\eta(P_{\text{max}}) = \eta_C \frac{ZT}{ZT + 2}.$$ \hspace{1cm} (76)

As in the other two limiting cases, this system behaves like a 2-terminal system, where reservoir 2 has been removed; this is to be expected because at this particular temperature, no heat flows through the second reservoir.

According to Eq. (69), all values of $\Delta T_2$ other than the three cases discussed above decrease the efficiency at maximum power with respect to the two-terminal case. The maximum power instead, at given tunneling rates, is
increased with respect to the two-terminal case. In fact the maximum power is formally equal to that of a two-terminal system, coupled to the same QD, with increased tunneling rates. So if we have fixed values of $\Gamma_1$, $\Gamma_2$ and $\Gamma_3$, we achieve the largest maximum power by choosing $\Delta T_2 = 0$ if $\Gamma_1 < \Gamma_3$ and $\Delta T_2 = \Delta T_3$ if $\Gamma_1 > \Gamma_3$. In Fig. 5, $P_{\text{max}}$ and $\eta(P_{\text{max}})$ from Eqs. (68) and (69) are plotted as a function of $\Delta T_2/\Delta T_3$, choosing $h\Gamma_1 = 0.02k_BT$ and $h\Gamma_2 = h\Gamma_3 = 0.01k_BT$. As we can see in panel (a), the power is maximum when $\Delta T_2 = \Delta T_3$ (this result is expected since $\Gamma_1 > \Gamma_3$). Furthermore the three maxima in panel (b) correspond to the three limiting cases previously studied, where $\eta(P_{\text{max}})$ reaches the two-terminal performance.

IV. NON-LINEAR RESPONSE

The linear response theory describes correctly the thermoelectric properties of bulk materials in most experimental conditions. However, as discussed for instance in Ref. 99, non-linear effects are important in nanoscopic setups, since the temperature difference is applied across very small elements of the order of tens or hundreds of nanometers. As far as heat-to-work conversion is concerned, there is a practical reason to consider the non-linear response, namely efficiency and power output may increase with increasing temperature difference. Furthermore, for systems with time-reversal symmetry the efficiency at maximum power can overcome the limit of $\eta_C/2$ only beyond the linear response\textsuperscript{111}.

In this Section we will consider a two-terminal QD system and discuss the numerical results obtained solving the kinetic equations (4) as discussed in Sec. II. We will focus our discussion on the thermoelectric properties, and on the efficiency and power produced by a QD-based heat engine. Let us define the charge current $J^c = J_2^c = -J_1^c$, thanks to charge current conservation, and the average reservoir temperature $T = (T_1 + T_2)/2$, which determines the typical thermal energy scale of the system beyond linear response (all energies will be given in units of $k_BT$). Furthermore, $\Delta\mu \equiv \Delta\mu_2 = eV$, with $V$ applied voltage, $\Delta T \equiv \Delta T_2$, and assume equidistant energy levels with spacing given by $\Delta E$. In order to describe the potential drop between the QD and the two reservoirs, we will assume that the set of energy levels is shifted as $E_p(V) = E_p + (1 - \theta_0)eV$, where $0 \leq \theta_0 \leq 1$ is the fraction of potential $V$ that drops over the tunnel barrier which couples reservoir 2 to the QD.

Regarding charge transport, we recall that in the linear response, within the quantum limit, the conductance exhibits peaks occurring at the dominant transition energies, i.e. when $\mu = \mu_N$, of width given by $k_BT$. By applying a finite voltage bias $V$, in the absence of a temperature difference ($\Delta T = 0$), the differential conductance, defined as

$$G = \left(\frac{\partial J^c}{\partial V}\right),$$

exhibits the typical Coulomb diamond structure, with visible excited states, as a function of $V$ and $\mu$, see for instance Ref. 105.

A. Non-linear Seebeck and Peltier coefficients

In the non-linear regime the thermopower (Seebeck coefficient) can be defined as follows

$$S = -\frac{V}{\Delta T}igg|_{J^c=0},$$

i.e. as the ratio between the thermovoltage $V$ that develops as a result of a finite $\Delta T$ applied, at open circuit ($J^c = 0$). In Fig. 6, $S$ is plotted as a function of $\mu$ for various values of $\Delta T/T$. The black solid curve (calculated for $\Delta T/T = 10^{-4}$) is the linear-response reference that is well approximated by the expression given in Eqs. (48). As discussed in Sec. III A, the black solid curve presents main oscillations of period $\Delta E + 2E_C$, and a fine structure with a $\Delta E$ spacing\textsuperscript{9}. Since we have chosen an equidistant energy spectrum, all curves share a number of features with the linear-response reference. Namely, i) $S$ crosses zero with positive slope at the main transition energies $\mu_N$ and is periodic with periodicity $\Delta E + 2E_C$ (in Fig. 6...
\( \mu = 290 \, k_B \bar{T} \) corresponds \( \text{[see Eq. (49)]} \) to \( \mu_{N=3} \); ii) in the range of \( \mu \) considered, \( S \) is antisymmetric with respect to \( \mu = 290 \, k_B \bar{T} \); iii) \( S \) vanishes for values of \( \mu \) in the middle points between two dominant transitions \( \mu_N \) and \( \mu_{N+1} \) (in Fig. 6 such points are located at \( \mu = 230 \, k_B \bar{T} \) and \( \mu = 350 \, k_B \bar{T} \)). Moreover, since we set \( \theta_0 = 1/2 \), the linear increase of \( S \) for \( \mu \approx \mu_N \) does not depend on the ratio \( \Delta T/\bar{T} \), i.e. it is well described by the linear-response proportionality coefficient \( 1/(|e|T) \). Interestingly, such features (except for the fine structure oscillations) can be understood in terms of a non-interacting model (see App. D), which also explains the reduction of the negative slope of \( S \) at the middle points as the ratio \( \Delta T/\bar{T} \) increases.

Let us now discuss the behavior of \( S \) when departing from the linear response regime. Fig. 6 shows that for all values of \( \Delta T \) the thermopower deviates from the linear-response curve only for \( \mu > 310 \, k_B \bar{T} \) (or below 270 \( k_B \bar{T} \)). A sharp departure already occurs at \( \Delta T/\bar{T} = 0.5 \) (red curve). This can be understood from the fact that, for \( \mu > 310 \, k_B \bar{T} \), \( S \) is of the order of 15 \( k_B /|e| \) which corresponds to a value of the thermovoltage \( (V = -7.5 \, k_B \bar{T} /|e|) \) such that \( |eV| \gg k_B T \)). Note that \( \mu = 310 \, k_B \bar{T} \) roughly corresponds to the first step of the fine structure in linear response. In particular, while the first step hardly moves by increasing \( \Delta T \) from its position in linear response, the second step, occurring at \( \mu = 330 \, k_B \bar{T} \) in the linear response, shifts to a smaller value for increasing \( \Delta T/\bar{T} \), eventually disappearing or merging with the first step. This behavior may be attributed to the combination of the following two effects. On one hand, the thermovoltage \( V \), which determines the transport energy window, depends on \( \mu \) and increases with \( \Delta T \) according to the definition (78). On the other hand, an increase of \( \Delta T/\bar{T} \) moves the lowest temperature \( (T_1) \) towards absolute zero, thus sharpening the Fermi distribution function \( f_1(E) \). This last effect is also responsible for the sharpening of \( S(\mu) \) as \( \Delta T/\bar{T} \) increases. Note furthermore that the extremal values of \( S \) decrease as \( \Delta T/\bar{T} \) increases.

Let us now consider the non-linear Peltier coefficient defined, for a given voltage \( V \), as

\[
\Pi = \frac{J^h}{Te} \bigg|_{\Delta T=0}.
\]

Our aim is to assess the failure of the Onsager reciprocity relation \( \Pi = TS \), which holds in the linear response regime. Beyond linear response, for a single-level non-interacting QD, one finds a “corrected” reciprocity relation, namely \( \Pi + V/2 = \bar{T}S \) in the case where \( \theta_0 = 1/2 \) (see App. D). To single out the effect of interactions in a multi-level QD, in Fig. 7 we plot the ratio \( r = (\Pi + V/2)/(\bar{T}S) \) as a function of \( \mu \) for various values of \( \Delta T/\bar{T} \) and \( |eV|/(k_B \bar{T}) \) (\( \Delta T/\bar{T} \) is used to compute \( S \), while \( |eV|/(k_B \bar{T}) \) to compute \( \Pi \)). Fig. 7 shows that the ratio \( r \) departs significantly from 1, the linear response result (blue thin line), only far enough from the dominant transition energy \( \mu_{N=3} = 290 \, k_B \bar{T} \). In particular, when \( \Pi \) is in the linear response regime and \( S \) is not (black solid curve) the strong deviations occurring for \( \mu \) around the middle points between dominant transition energies, namely \( \mu = 230 \, k_B \bar{T} \) and \( \mu = 350 \, k_B \bar{T} \), can be explained by a two-level non-interacting model (see App. D). However, the deviations occurring in the range of values of \( \mu \) between 250 \( k_B \bar{T} \) and 330 \( k_B \bar{T} \) can be imputed to interaction effects. In the opposite case, where \( S \) is in the linear response and \( \Pi \) is not (red curve), the deviations of \( r \) from 1 are entirely due to interaction effects and \( r \).
takes values between 0 and 1 in the entire range of values of \( \mu \). When both \( S \) and \( \Pi \) are beyond linear response the two behaviors discussed above coexist giving rise to the black dashed curve\(^{113}\).

\section*{B. Efficiency and output power}

In this Section we consider the efficiency for heat-to-work conversion and output power in a two terminal system. Specifying Eq. (20) to a two terminal system where \( \Delta T > 0 \), we have that

\[ \eta = \frac{P}{J_T^2}, \]  

\[(80)\]

where \( P > 0 \) is the output power, defined in Eq. (19), and \( J_T^2 > 0 \) is the heat current absorbed by the system. Apart from the system’s parameters, \( \eta \) depends on \( V \) and \( \Delta T \).

Let us first consider the maximum efficiency \( \eta_{\text{max}} \), obtained by maximizing the efficiency \( \eta \) with respect to the applied voltage \( V \), at given \( \Delta T \). \( \eta_{\text{max}} \) is plotted, normalized to \( \eta_C \), in Fig. 8(a) as a function of \( \mu \) for different values of \( \Delta T / T \). All plots show pairs of peaks close to \( \mu_N \), whose maximum is very close to \( \eta_C \), and secondary peaks of smaller height. The solid black curve, relative to the linear response regime (\( \Delta T / T = 10^{-4} \)), is related through Eq. (23) to the plot of \( ZT \) [Fig. 3(b)]\(^{114}\) discussed in Sec. III B. For the black curve a pair of maxima approaching \( \eta_C \) occur at \( \mu = \pm 2.40k_B T \), while \( \eta_{\text{max}} \) vanishes at the dominant transition energies (and at the middle points between two dominant transition energies). Moreover, a fine structure of secondary peaks, with spacing \( \Delta E \), appears for intermediate values of \( \mu \). Moving away from the linear response, the main observation is that an increase of \( \Delta T / T > 0.1 \) produces only quantitative changes to the curves. As shown in Fig. 8(a), the main peaks of \( \eta_{\text{max}} \) are still approximately located at \( \mu = \pm 2.40k_B T \) and approaching the Carnot efficiency, while the peaks’ width reduces slightly with increasing \( \Delta T \), and the fine structure of the secondary peaks gets simply distorted.

Another important quantity in heat-to-work conversion is the maximum output power generated \( P_{\text{max}} \), which is obtained by maximizing the output power with respect to the applied voltage \( V \). It turns out that \( P_{\text{max}} \) exhibits pairs of peaks approximately located at \( \mu = \pm 2.40k_B T \) whose height increases approximately quadratically with \( \Delta T \), as long as \( \Delta T / T \) is not close to 2. Interestingly, Fig. 8(b) shows that the maximum output power \( P_{\text{max}} \), when normalized to its peak value \( P_{\text{peak}} \), only very weakly depends on the ratio \( \Delta T / T \). In particular, \( P_{\text{max}} / P_{\text{peak}} \) is well approximated by the linear-response result, whose analytical expression is obtained by substituting Eq. (50) into Eq. (21).

The efficiency at maximum power \( \eta(P_{\text{max}}) \) can now be calculated by taking, for each value of \( \mu \), the value of \( V \) which maximizes the power. \( \eta(P_{\text{max}}) \) is plotted, as a function of \( \mu \), in Fig. 8(c) for various values of the ratio \( \Delta T / T \). By increasing such ratio starting from the linear response [solid black curve, related to the plot of \( ZT \) in Fig. 3(b) through Eq. (22)] one finds that the peak values, again occurring approximately at \( \mu = \pm 2.40k_B T \), increase well above \( \eta_C / 2 \) (the upper limit for the linear response). On the contrary, the efficiency at maximum power for values of \( \mu \) away from \( \mu_N \), relative to the fine structure, decreases with increasing \( \Delta T / T \) beyond the linear response, but is only slightly different moving from \( \Delta T / T = 1 \) (red curve) to \( \Delta T / T = 1.95 \) (black dashed curve).

It is now interesting to compare the peak values of the efficiency at maximum power with various reference values, such as the Curzon-Alhborn (CA) efficiency\(^{88}\) \( \eta_{\text{CA}} = 1 - \sqrt{1 - \eta_C} \) and the Schmiedl-Seifert (SS) efficiency\(^{106}\) \( \eta_{\text{SS}} = \eta_C / (2 - \eta_C) \). To do so we calculate the peak power, i.e. maximizing the power with respect to \( V \) and \( \mu \), and plot the corresponding efficiency as a function of \( \eta_C \) (determined by the temperature difference) in various situations, see Fig. 9. In particular, we
consider the case of a QD with one doubly degenerate level with a finite charging energy (solid thick curves) and the case of a QD with two non-degenerate levels and zero charging energy (dashed curves). The parameters are chosen such that the two situations can be compared, namely $\Delta E = 2E_C$, i.e. the differential conductance consists of two peaks separated by the same electrochemical potential. Figure 9 (inset) shows the following general feature for small $\eta_C$, i.e. in the linear response regime: $\eta(P_{\text{peak}})$ increases linearly with $\eta_C$ with slope determined by the value of $ZT$ [see Eq. (22)]. In particular, the two black curves (relative to $E_C = 10 \, k_B T$ and $\Delta E = 20 \, k_B T$) virtually coincide, and are equal to the one for a single non-interacting level QD [see App. D and the first line of Eqs. (D13)], since the parameters are such that $k_B T \ll E_C, \Delta E$, where the transport is mostly accounted for by a single energy level. On the contrary, the two red curves (relative to $E_C = 1.5 \, k_B T$ and $\Delta E = 3 \, k_B T$) differ by a large extent, with the interacting case (solid red curve) exhibiting larger efficiency at maximum power with respect to the associated non-interacting case (dashed red curve). Note that the efficiency at maximum power relative to the interacting case goes beyond the CA efficiency, when $\eta_C$ is larger than about 0.5, for all values of $E_C$ between $1.5 \, k_B T$ and $10 \, k_B T$. Finally, we find that the SS efficiency is never overcome.

To complete the analysis, we show the correlation between efficiency and output power calculated for a few points of the plots in Fig. 9, both for doubly degenerate interacting case (solid curves), and non-degenerate non-interacting case (dashed curves). Curves obtained by increasing the value of $V$ from zero to the thermovoltage value, corresponding to the open-circuit situation, for various values of $\eta_C$. Panel (a) refers to $\Delta E = 20 \, k_B T$ (with $E_C = 0$, dashed curves) and to $E_C = 10 \, k_B T$ (with $\Delta E = 0$, solid curves), while panel (b) refers to $\Delta E = 3 \, k_B T$ (with $E_C = 0$, dashed curves) and to $E_C = 1.5 \, k_B T$ (with $\Delta E = 0$, solid curves).
FIG. 11: (Color online) Maximum power, normalized to the peak value, plotted as a function of μ for different values of ηC with EC = 10 kB ¯T, ∆E = 0 and considering a single doubly degenerate energy level. Surprisingly, not only the linear-response curve (black) is very well approximated by the analytic expression in Eq. (81) (black dashed curve) valid in the quantum limit. Tunneling rates are ℏΓ1(p) = ℏΓ2(p) = 0.01kB ¯T.

for all the curves in the figure, both maximum power and efficiency at maximum power increase. For large values of ηC (green curves), the efficiency remains close to the Carnot efficiency when V is increased beyond the point of maximum power. The general feature is that in the interacting case (solid curves) the power is much larger than in the associated non-interacting situation (dashed curves). When ∆E and EC are of the same order as kB ¯T [Fig. 10(b)], both the maximum power and the efficiency at maximum power are increased in the interacting case as compared with the non-interacting case. A remarkable property of both regimes discussed in panels (a) and (b) is that in the strongly nonlinear regime, the maximum power is obtained for values where the efficiency is high and close to the maximum efficiency.

We finally discuss a peculiarity of the maximum power output considering a doubly degenerate energy level in the interacting case117 (as we will point out, some considerations are also valid in the non degenerate case). Fig. 11 shows the maximum power, normalized with respect to the peak value, as a function of μ for different values of ηC including the linear-response case (parameter values are the same as for Fig. 9, with EC = 10 kB ¯T). The first remarkable feature is that the two peaks around a dominant transition energy have different heights (here μN=1 = 10 kB ¯T and μN=2 = 30 kB ¯T); in the absence of degeneracy, within the quantum limit all the peaks have the same height (see Fig. 3). More precisely, see Eq. (81), the external peaks, displaced of 2.53kB ¯T from μN, are higher with respect to the internal peaks, displaced of 2.32kB ¯T from μN, whose height is almost equal to the non-degenerate (interacting) case. The second feature is that all curves, apart from the case of very large ηC (blue curve), are well approximated by the linear-response quantum limit expression (black dashed curve in Fig. 11)

\[
P_{\text{max}} = \frac{\sqrt{2}\gamma \Delta E}{16k_B\bar{T}} \left( \frac{\Delta T}{T} \right)^2 \cosh \left( \frac{\Delta_{\text{min}}}{2k_B\bar{T}} \right) \frac{\Delta_{\text{min}}^2}{2}. \tag{81}
\]

The plus (minus) sign in Eq. (81) is to be taken when N_{\text{min}} is even (odd). Eq. (81) allows us to find that level degeneracy gives rise to an enhancement of output power of about 1.77 times respect to the non-degenerate case, independently of the parameter values (see also Ref. 61). Eq. (81) also makes clear that the origin of the asymmetry for \(\Delta_{\text{min}} \rightarrow -\Delta_{\text{min}}\) and of the difference in peaks’ height is the term ± log2/2 occurring in the presence of degeneracy. The case of EC = 1.5 kB ¯T (not shown), which is not within the quantum limit, produces a much more asymmetric behavior. We have verified that also in the non degenerate case the analytic formula derived in the linear response regime [see Eq. (50)] well describes the maximum power also beyond the linear response regime.

V. CONCLUSIONS

We have studied the thermoelectric properties of a multi-level interacting QD in the sequential tunneling regime, in a multi-terminal setup, both in the linear response regime and beyond. In particular, we have

- generalized the sequential tunneling method put forward by Beenakker in Refs. 8 and 9 to a multi-terminal configuration and set the range of validity of the expressions for the charge and heat currents in the linear response regime;
- found simple analytic formulas for the multi-terminal transport coefficients in the low temperature limit;
- found simple analytic formulas for the power factor Q and the figure of merit ZT in the low temperature limit for a two-terminal setup;
- found that Q and ZT can be simultaneously maximized for suitable values of the electrochemical potential;
- found that Coulomb interactions can dramatically enhance ZT by suppressing the thermal conductance;
- found that both the interacting an non-interacting models strongly violate the Wiedemann-Franz law;
- found analytic expressions for the maximum power and for the efficiency at maximum power in a three terminal setup;
investigated the nonlinear Seebeck and Peltier coefficients in a two-terminal setup, identifying features of the breakdown on the Onsager reciprocity relation;

- computed numerically the maximum efficiency, the maximum power, and the efficiency at maximum power in the non-linear regime, finding optimal system parameters for heat-to-work conversion such that the efficiency at maximum power can go beyond Curzon-Alhborn’s efficiency;

- compared the case of a doubly degenerate level with interaction and the case of two non-degenerate levels without interaction finding that the interacting case enhances the power output and, especially when charging energy and level spacing are of the order of the thermal energy, it increases the efficiency at maximum power that can go beyond Curzon-Alhborn’s efficiency;

- found that the non-linear maximum power is well approximated by the analytic linear response expression;

- found that QDs with degenerate energy levels and Coulomb interactions achieve higher efficiency and output power than non-degenerate QDs; in particular the maximum power is enhanced almost of a factor 2;

- calculated the transport coefficients for a non-linear, non-interacting QD with 1 and 2 energy levels (App. D);

- found approximate analytic expressions for the maximum power and efficiency at maximum power for a non-linear, non-interacting QD with 1 energy level (App. D).

The multi-terminal formalism developed in this paper and the expressions we have obtained for charge and heat currents, transport coefficients, power and efficiency could be used to design and analyze experimental data. Extensions of the studies presented in this paper could include level spacings different from the equidistant and regimes beyond the quantum limit. Finally, a comprehensive description of the thermoelectric properties and performance of a QD should assess the role of quantum coherence going beyond the sequential tunneling limit and the relevance of phonon contribution to heat transport.

Acknowledgements

We would like to acknowledge fruitful discussions with Stefano Roddaro. This work has been supported by the SNS internal projects “Thermoelectricity in nanodevices” and “Non-equilibrium dynamics of one-dimensional quantum systems: From synchronization to many-body localization”, by the EU project ThermIQ and by the COST Action MP1209 “Thermodynamics in the quantum regime”. G.B. acknowledges support by INFN through the project “QUANTUM”.

Appendix A: The kinetic equations always allows a non-trivial solution.

To prove this statements, let us consider the kinetic equations as written in Eq. (11). This equation is made up of a sum over $p$ of the following expression:

$$
\begin{align*}
(\delta_{n_p,1} - \delta_{n_p,0}) & \left[ P(\{n_i\}, n_p = 0) A_{\tilde{N},p} \\
- P(\{n_i\}, n_p = 1) B_{\tilde{N},p} \right],
\end{align*}
$$

(A1)

which is a function of a generic configuration $\{n_i\}$. Let us sum this expression for the two particular configurations: $(\{n_i\}, n_p = 1)$ and $(\{n_i\}, n_p = 0)$. When $n_p = 1$, the first term in round parenthesis gives a plus sign, while when $n_p = 0$, it gives a minus sign. The rest of the expression does not depend on $n_p$, since $\tilde{N} = \sum_{i\neq p} n_i$, thus the sum over the above configurations exactly vanishes. Now let us go back to Eq. (11), and let us sum over the two configurations $(\{n_i\}, n_k = 1)$ and $(\{n_i\}, n_k = 0)$, $k$ being a given index. According to the argument given above, the term in the sum where $p = k$ vanishes, and we obtain

$$
\sum_{n_k = 0,1} \sum_{p \neq k} (\delta_{n_p,1} - \delta_{n_p,0}) \left[ P(\{n_i\}, n_p = 0) A_{\tilde{N},p} \\
- P(\{n_i\}, n_p = 1) B_{\tilde{N},p} \right] = 0.
$$

(A2)

Thus by summing over a given occupation number $n_k = 0,1$ we have removed the case $p = k$ in the sum over $p$. If we now sum over all occupation numbers, we will remove all terms from the sum, yielding zero. The sum over all occupation numbers $\{n_i\}$ is the sum of all $2^L$ equations of the kinetic equations: this demonstrates that any equation in the kinetic equations is linearly dependent from the other ones. Thus the matrix $M$ defined by the kinetic equations has a null space of dimension at least 1, since we demonstrated that the rows of $M$ are not linearly independent. Furthermore, if we perform the same sum over all occupation numbers to the time dependent kinetic equations, given in Eq. (4), we find that:

$$
\frac{\partial}{\partial t} \left( \sum_{\{n_i\}} P \right) = 0.
$$

(A3)

This is an obvious but important property that says that the probability normalization does not depend on time.

Appendix B: The DBEs are not consistent in general

In general, if $E_C \neq 0$, the DBEs are not consistent. This means that no set of $P(\{n_i\})$ exists that can simul-
taneously satisfy all the DBEs. In App. (C) we will discuss which conditions guarantee their consistency within the linear response regime.

Here let us demonstrate their inconsistency in the special case \( L = 2 \). Let us start from Eq. (15) and consider non null temperature and electrochemical potential differences. We will show that these equations form an over-complete set for \( P(\{n_1\}) \) that is in general does not allow any non null solution. Since \( L = 2 \), we have \( 2^L = 4 \) unknown probabilities, and the number of DBEs is \( 2^{L-1}2 = 4 \). In this case, the DBEs can be represented in matrix form as follows

\[
M_D \mathbf{P} = \begin{pmatrix}
A_{0,1} & -B_{0,1} & 0 & 0 \\
0 & 0 & A_{1,1} & -B_{1,1} \\
A_{0,2} & 0 & -B_{0,2} & 0 \\
0 & A_{1,2} & 0 & -B_{1,2}
\end{pmatrix}
\begin{pmatrix}
P_{00} \\
P_{01} \\
P_{10} \\
P_{11}
\end{pmatrix} = \begin{pmatrix}
0 \\
0 \\
0 \\
0
\end{pmatrix},
\]

(\text{B1})

where \( P_{n_2,n_1} \equiv P(n_2, n_1) \) and the coefficients are defined in Eqs. (12) and (13). In order to show that this linear algebra problem does not allow a non null solution, we will show that the determinant \( \det(M_D) \) is in general not zero. We have

\[
\det(M_D) = \Gamma_{\text{tot}}(1)\Gamma_{\text{tot}}(2) \left( A_{0,2}A_{1,1} - A_{0,1}A_{1,2} \right)
+ \Gamma_{\text{tot}}(1)A_{0,2}A_{1,2}(A_{0,1} - A_{1,1})
+ \Gamma_{\text{tot}}(2)A_{0,1}A_{1,1}(A_{1,2} - A_{0,2}),
\]

(\text{B2})

where \( \Gamma_{\text{tot}}(p) = \sum_{\alpha} \Gamma_{\alpha}(p) \). It is pretty clear that since \( \Gamma_{\alpha}(p) \), \( E_p \) and \( E_C \) are arbitrary, this determinant cannot be in general zero. For instance, choosing a two terminal system with \( \Gamma_1(p) = \Gamma_2(p) = \Gamma \), \( \Delta \mu_2 = 4k_B T \), \( E_1 = 1/2k_B T \), \( E_2 = 3/2k_B T \), \( \mu = \Delta T_2 = 0 \) we obtain \( \det(M_D)/\Gamma^4 \approx 0.20 \). It is interesting to notice that at equilibrium \( (\Delta \mu_2 = \Delta T_2 = 0) \), the DBEs are all exactly satisfied by the grand canonical distribution. Furthermore, in the non-interacting limit \( E_C = 0 \), the coefficients \( A_{\tilde{N},p} \) and \( B_{\tilde{N},p} \) depend only on \( p \), so that we can drop the \( \tilde{N} \) argument. Thus

\[
\det(M_D) = A_2A_1B_1B_2 - A_1A_2B_2B_1 = 0.
\]

The proof can be extended to any number \( L \) of levels as follows. We rewrite the DBEs, Eq. (15), as

\[
\ln P(\{n_i\}, n_p = 0) - \ln P(\{n_i\}, n_p = 1) = \ln \frac{B_{\tilde{N},p}}{A_{\tilde{N},p}}.
\]

(\text{B3})

This equation has the same form as the LDBEs (C1) which we will consider in App. C, where the unknown probabilities are replaced by the logarithm of the probabilities, and where

\[
\delta_p(\tilde{N}) = \ln \frac{B_{\tilde{N},p}}{A_{\tilde{N},p}}.
\]

(\text{B4})

As we shall show in App. C, these equations are consistent if \( \delta_p(\tilde{N}) \) satisfies property (C3). This is the case if \( E_C = 0 \), since \( A_{\tilde{N},p} \) and \( B_{\tilde{N},p} \) are then independent of \( \tilde{N} \). Then condition (C3) is trivially fulfilled with \( c = 0 \) and the DBEs are consistent.

To summarize, as we will see in Appendix C, the DBEs are not consistent in general, but they are if we set \( E_C = 0 \) or \( \Delta \mu_\alpha = \Delta T_\alpha = 0 \) for all \( \alpha \). Furthermore, within the linear response regime, they are valid in many more cases, for example if the tunneling rates are proportional, i.e., \( \Gamma_\alpha(p) = k_\alpha \Gamma_1(p) \) for \( \alpha = 2, 3, \ldots, N \).

Appendix C: Conditions of validity of the linearized DBEs

In this Appendix we will assess under which conditions the linearized DBEs are consistent. We start by rewriting Eq. (32) in the following way:

\[
\psi(\{n_i\}, n_p = 0) - \psi(\{n_i\}, n_p = 1) = \delta_p(\tilde{N}),
\]

(C1)

where

\[
\delta_p(\tilde{N}) = -\frac{1}{k_B T} \sum_\alpha \frac{\Gamma_\alpha(p)}{\Gamma_{\text{tot}}(p)} \left( E_p + (2\tilde{N} + 1)E_C - \mu \right) \frac{\Delta T_\alpha}{T} + \Delta \mu_\alpha.
\]

(C2)

We will prove that the linearized DBEs, written in Eq. (C1), are consistent if \( \delta_p(\tilde{N}) \) satisfies the property

\[
\delta_p(N) = \delta_p(M) = c(N - M),
\]

(C3)

where \( c \) is a constant that does not depend on \( p, N \) or \( M \). This statement will be explicitly proven for a two energy level system, then it will be extended to \( L \) energy levels by induction.

Property (C3) is in general satisfied if the tunneling rates are proportional, i.e., \( \Gamma_\alpha(p) = k_\alpha \Gamma_1(p) \), or if \( E_C = 0 \), or if \( \Delta T_\alpha = 0 \) for all \( \alpha \). Furthermore, in a three terminal system, property (C3) is satisfied also if \( \Delta T_2 = 0 \) and \( \Gamma_3(p) = k\Gamma_{\text{tot}}(p) \), or if \( \Delta T_3 = 0 \) and \( \Gamma_2(p) = k\Gamma_{\text{tot}}(p) \). These conditions can be generalized to a generic \( N \)-terminal system by requiring that

\[
E_C \sum_\alpha \frac{\Gamma_\alpha(p)}{\Gamma_{\text{tot}}(p)} \Delta T_\alpha
\]

(C4)

is independent of \( p \).

1. Two energy level system (\( L = 2 \))

We will prove that if property (C3) is satisfied, then the linearized DBEs allow a solution. Eq. (C1) represents a linear algebra problem for \( \psi(\{n_i\}) \). For \( L = 2 \), \( \psi(\{n_i\}) = \psi(n_2, n_1) \equiv \psi_{n_2,n_1} \). Let \( \mathbf{\psi} \) be the vector
\( (\psi_{00}, \psi_{01}, \psi_{10}, \psi_{11}) \) and let \( B \) be the corresponding matrix such that Eq. (C1) can be written as

\[
B \vec{\psi} = \begin{pmatrix}
1 & -1 & 0 & 0 \\
0 & 0 & 1 & -1 \\
1 & 0 & -1 & 0 \\
0 & 1 & 0 & -1
\end{pmatrix}
\begin{pmatrix}
\psi_{00} \\
\psi_{01} \\
\psi_{10} \\
\psi_{11}
\end{pmatrix} = \begin{pmatrix}
\delta_1(0) \\
\delta_1(1) \\
\delta_2(0) \\
\delta_2(1)
\end{pmatrix} \equiv \vec{\delta}.
\] (C5)

Matrix \( B \) has a null space of dimension 1, generated by \( \vec{\psi} = (1, 1, 1, 1) \), thus it is not invertible. This vector actually represents the equilibrium distribution: when \( \Delta T_a = \Delta \mu_a = 0 \), \( \delta_p(\bar{N}) = 0 \), so \( \vec{\psi} = (1, 1, 1, 1) \) satisfies Eq. (C5). Since \( P = P_{eq}(1 + \psi) \), \( P = 2P_{eq} \). Normalizing the probabilities yields \( P \equiv P_{eq} \), so we have demonstrated that the equilibrium distribution is in fact given by the grand canonical distribution. Eq. (C5) will allow a solution if and only if vector \( \vec{\delta} \) only belongs to the image of matrix \( B \), so there is a one dimensional space orthogonal to the image of \( B \) that cannot be obtained by linear combinations of \( B \)'s columns. Vector

\[
\vec{v}_0 = \begin{pmatrix}
1 \\
-1 \\
-1 \\
1
\end{pmatrix}
\] (C6)

is orthogonal to the columns of \( B \), so it is a generator of this one dimensional space. Thus a solution \( \vec{\psi} \) will exist if and only if vector \( \vec{\delta} \) only belongs to the image of \( B \), thus it cannot have a projection on \( \vec{v}_0 \). The projection is zero when

\[
0 = \vec{v}_0 \cdot \vec{\delta} = \delta_1(0) - \delta_1(1) - \delta_2(0) + \delta_2(1)
\] (C7)

which is satisfied using property (C3).

As we have shown in Sec. III, Eq. (28) for the linearized charge current is correct in general, instead the linearized heat current given in Eq. (33) is correct only if the linearized DBE are consistent. We verified numerically all these statements computing exactly, for small temperature and electrochemical potential differences, the charge and heat currents using both the DBEs and the kinetic equations. As demonstrated by Fig. 12, which shows the heat current as a function of \( \mu \) for a choice of tunneling rates that are not proportional to each other, a particular subset of DBEs leads to an incorrect result (red dashed curve) which differs from the one obtained using the kinetic equations (black solid curve). On the other hand, there is no difference when plotting the charge current using the same parameters as in Fig. 12.

2. L energy level system

We extend the previous demonstration to a generic system with \( L \) energy levels. We will use the following notation and conventions:

- \( \{n_i\} \) indicates a generic set of occupation numbers, conventionally ordered the following way: \( \{n_L, n_{L-1}, \ldots, n_2, n_1\} \).
- \( \{n_p = 1, \{n_i\}\} \) is a shorthand notation for: \( \{n_L, \ldots, n_{p+1}, 1, n_{p-1}, \ldots, n_1\} \).
- \( \{n_p = 0, \{n_i\}\} \) is a shorthand notation for: \( \{n_L, \ldots, n_{p+1}, 0, n_{p-1}, \ldots, n_1\} \).
- \( \{n_p = 1, \{0\}\} \) is a shorthand notation for: \( \{0, \ldots, 0, 1, 0, \ldots, 0\} \), where the 1 is relative to \( n_p = 1 \).
- \( \{n_p = 0, \{0\}\} \) is equivalent to: \( \{0, \ldots, 0\} \).
- \( \vec{\psi}_i \) is the component of vector \( \vec{\psi} \) with index \( i \).
- Each linearized DBE in Eq. (C1) can be uniquely defined by specifying \( p \) and \( \{n_i\} \), so \( \{p\{n_i\}\} \) is a shorthand notation to indicate that given equation.
- Let \( A \) be a matrix written in block form, for example:

\[
A = \begin{pmatrix}
B & 0 \\
0 & C \\
D & E
\end{pmatrix}
\] (C8)

where \( B, C, D \) and \( E \) are matrices. By “block row” we intend the set of rows in the block. For example, the first block row of \( A \) is \( \{B, 0\} \), the second block row of \( A \) is \( \{0, C\} \), and so on.

![FIG. 12: (Color online) Heat current computed using Eq. (10) and using non equilibrium probabilities found solving exactly the kinetic equations (black solid line) and a subset of independent DBEs (red dashed line), as a function of the electrochemical potential \( \mu \). The parameters used are: \( \Delta T/T = 10^{-4} \), \( \Delta \mu = 0 \), \( E_C = k_B T \), 5 equidistant energy levels with \( \Delta E = 0.2 k_B T \), \( \hbar \Gamma_1(p) = (5)^{p-1} k_B T \), and \( \hbar \Gamma_2(p) = (0.01)^{p-1} k_B T \).](image-url)
In order to represent the linearized DBE, given in Eq. (C1), in matrix form, we will put the \(2^L\) unknowns \(\psi(n_i)\) into a vector as follows: \(\vec{\psi}(n_i) = \psi(n_i)\), where \(I(n_i) = \sum_{i=1}^{L} n_i 2^{i-1}\).

Let us define \(B(L)\) as the \(L2^{L-1} \times 2^L\) matrix such that Eq. (C1) can be represented as follows:

\[
B(L)\vec{\psi} = \vec{\delta}(L), \tag{C9}
\]

where \(\vec{\delta}(L)\) is a vector containing the various values of \(\delta_p(N)\). Each row of \(B(L)\) contains zeros except for a +1 and a -1 in the position corresponding to \(\psi(n_p = 0, \{n_i\})\) and \(\psi(n_p = 1, \{n_i\})\). Let’s notice that \(B(L)\) has the following property:

\[
B(L) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{C10}
\]

This property is trivial since, as we just discussed, each row of \(B(L)\) only has a +1 and a -1.

a. Recursive decomposition

We want to write \(B(L)\) in terms of \(B(L-1)\). We will thus order the rows in \(B(L)\), and accordingly the elements of \(\vec{\delta}(L)\), starting from those DBE \(p(n_i)\) where \(p < L\), then we will add the ones involving level \(L\). In particular, we will first put equations \(p(n_L = 0, \{n_i\})\) with \(p < L\), then we will add equations \(p(n_L = 1, \{n_i\})\) with \(p < L\), and at last we will add equations \([L\{n_i\}]\), which are the ones involving level \(L\). For example, \(B(2)\) will be written as in Eq. (C5).

This ordering allows us to relate \(B(L)\) to \(B(L-1)\). The first set of equations \([p(n_L = 0, \{n_i\})]\), with \(p < L\), represents all possible DBE between elements of \(\vec{\psi}\) with \(n_L = 0\), so they are equivalent to the DBE for \(L-1\) levels. Since \(n_L = 0\), according to our indexing convention, these rows involve all indexes of \(\vec{\psi}\) with \(0 < i < 2^{L-1}\), which corresponds to the first half of vector \(\vec{\psi}\). Instead, the equations \([p(n_L = 1, \{n_i\})]\), with \(p < L\), represents all possible DBE between elements of \(\vec{\psi}\) with \(n_L = 1\), so also these equations are equivalent to the DBE for \(L-1\) levels. Since \(n_L = 1\), according to our indexing convention, these rows involve all indexes of \(\vec{\psi}\) with \(i \geq 2^{L-1}\), which corresponds to the second half of vector \(\vec{\psi}\). At last, equations \([L\{n_i\}]\) relate components of \(\vec{\psi}\) where only the occupation number \(n_L\) is changed. Since \(I(n_L = 1, \{n_i\}) = I(n_L = 0, \{n_i\}) + 2^{L-1}\), these equations relate indexes of \(\vec{\psi}\) that are distant \(2^{L-1}\), thus they mix indexes between the first and second half of \(\vec{\psi}\). We will thus have the following block representation:

\[
B(L) = \begin{pmatrix} B(L-1) & 0 \\ 0 & B(L-1) \end{pmatrix}, \tag{C11}
\]

where \(I_d(L)\) is the \(2^L \times 2^L\) identity matrix. Since we trivially have that \(B(1) = (1, -1)\), Eq. (C11) can be used to define \(B(L)\) recursively, yielding a precise row ordering. For example, as we can see in Eq. (C5), \(B(2)\) can be obtained by applying Eq. (C11) to \(B(1)\).

Also \(\delta(L)\) allows a decomposition in terms of \(\delta(L-1)\). The first block row in Eq. (C11), \((B(L-1), 0)\), corresponds to equations \(|\{n_L = 0, \{n_i\}\}|\) with \(p < L\). Recalling Eq. (C2), we will notice that \(\delta_p(N)\) depends on level \(L\) only through:

\[
\tilde{N}(L) = \sum_{p < L} n_p = \sum_{p < L} n_p = \tilde{N}(L-1), \tag{C12}
\]

since \(n_L = 0\). So in the first block row we have that \(\delta(L) = \delta(L-1)\). The second block row, \((0, B(L-1))\), corresponds to equations \(|\{n_L = 1, \{n_i\}\}|\) with \(p < L\). So this time the presence of level \(L\) will change the values of \(\tilde{N}\) the following way:

\[
\tilde{N}(L) = \sum_{p \neq L} n_p = \sum_{p \neq L} n_p = \tilde{N}(L-1) + \tilde{N}(2) \tag{C13}
\]

since \(n_L = 1\). So in the second block row \(\delta(L)\) is given by \(\delta(L-1)\) replacing \(\tilde{N}\) with \(\tilde{N} + 1\). Using property (C3), we have that:

\[
\delta_p(\tilde{N} + 1) = \delta_p(\tilde{N}) + c, \tag{C14}
\]

where \(c\) is a constant that does not depend on \(p\) or \(\tilde{N}\). So in the second block row we will have that \(\delta(L) = \delta(L-1) + c\). The last row \((I_d(L-1), -I_d(L-1))\) in Eq. (C11) corresponds to equations \([L\{n_i\}]\), which involve level \(L\). So let us define \(\tilde{\delta}(L)\) as the vector that corresponds to \(\delta(L)\) in the third block row.

Using these observations and Eq. (C11), we can rewrite Eq. (C9) using the following decomposition:

\[
\begin{pmatrix} B(L-1) \\ 0 \\ I_d(L-1) \end{pmatrix} \begin{pmatrix} 0 \\ B(L-1) \\ -I_d(L-1) \end{pmatrix} \begin{pmatrix} \vec{\psi}(n_L = 0, \{n_i\}) \\ \vec{\psi}(n_L = 1, \{n_i\}) \end{pmatrix} = \begin{pmatrix} \delta(L-1) \\ \tilde{\delta}(L) + c \end{pmatrix}. \tag{C15}
\]

This decomposition will be the main tool to perform a demonstration by induction.

b. The independent part of \(B(L)\)

Among the \(2^{L-1}\) rows of \(B(L)\), we will explicitly show how to extract \(2^{L-1}\) linearly independent rows; we will later show that all other rows can be obtained by linear combinations. Let’s define \(\bar{B}(L)\) as the matrix containing
only the rows of $B(L)$ corresponding to equations $[p\{n_i\}]$ with $n_i = 0$ for $i < p$; these are $2^L - 1$ rows. We will now prove by induction that all the rows of $\tilde{B}(L)$ are linearly independent.

Since $\tilde{B}(L)$ is made by selecting rows from $B(L)$, we can decompose it in terms of $\tilde{B}(L-1)$ just like we did for $B(L)$ in Eq. (C11). An interesting characteristic of $B(L)$ is that it only has one equation in the block $[L\{n_i\}]$ involving $\psi(0, 0, \ldots, 0)$ and $\psi(1, 0, \ldots, 0)$. So the decomposition becomes:

$$\tilde{B}(L) = \begin{pmatrix} \tilde{B}(L-1) & 0 \\ 0 & \tilde{B}(L-1) \end{pmatrix}.$$  \hspace{1cm} (C16)

We can use Eq. (C16) to define $\tilde{B}(L)$ recursively noticing that $\tilde{B}(1) = B(1) = (1, -1)$.

Now we can prove by induction over $L$ the following statement: all the rows of $\tilde{B}(L)$ are linearly independent. This statement is obvious for $L = 1$ and $L = 2$. We will now assume that our statement is valid for $L - 1$, and we will show that this implies that it is valid for $L$. The first block row in Eq. (C16) contains rows that are linearly independent by hypothesis. The same is true for the second block row. Also putting together the first and second block row yields $2^L - 2$ independent rows, since the space generated by the respective linear combinations are clearly orthogonal thanks to the block decomposition. So we have to show that the last row cannot be written as a linear combination of previous rows. We will make a proof by contradiction. Let’s assume that the last row $\vec{v} = (1, 0, \ldots, 0) - 1, 0, \ldots, 0)$ can be written as a linear combination of a vector belonging to the space generated the first block row, $\vec{w}_0$, and of one belonging to the second block row, $\vec{w}_1$:

$$\vec{v} = \vec{w}_0 + \vec{w}_1.$$  \hspace{1cm} (C17)

If we define $\vec{n} = (1, \ldots, 1)[0, \ldots, 0)$, we will have that:

$$\vec{n} \cdot \vec{w}_0 = 0,$$  \hspace{1cm} (C18)

$$\vec{n} \cdot \vec{w}_1 = 0.$$

The first equation is zero because, as we have shown in Eq. (C10), $B(L)$ has the property that $B(L)(1, \ldots, 1)^T = 0$, so it is true also for $\tilde{B}(L)$. The second equation is zero because $\vec{n}$ has zeroes on all non-null components of $\vec{w}_1$. So multiplying each sides of Eq. (C17) by $\vec{n}$ yields $\vec{n} \cdot \vec{v} = \vec{n} \cdot (\vec{w}_0 + \vec{w}_1) = 0$, which is absurd because $\vec{n} \cdot \vec{v} = 1$. Thus we have completed the proof by induction.

c. Consistency of the linearized DBEs

We will now show that if property (C3) is valid, the linearized DBEs are consistent and allow a solution for $\psi$ given by a one dimensional space; the unique solution can then be found by imposing the normalization condition on the probabilities. We will thus prove by induction over $L$ the following statement: if $\delta_p(N)$ satisfies property (C3), all equation in the linear algebra system in Eq. (C15) can be written as linear combinations of the equations involving the rows of $\tilde{B}(L)$. Since there are $2^L - 1$ independent rows in $\tilde{B}(L)$, by virtue of Rouché-Capelli’s theorem we will have a solution given by a $2^L - (2^L - 1) = 1$ dimensional space. This statement is trivially true for $L = 1$, since $B(1) = \tilde{B}(1)$, and we have proved it explicitly for $L = 2$, so we will show that it’s valid for $L - 1$, then it’s valid for $L$.

Let’s consider the subset of rows of Eq. (C15) where $B(L) = \tilde{B}(L)$. Among these equations, we will denote by “set 0" those belonging to the first block row, and “set 1" those belonging to the second block row. As we can see from Eq. (C16), this leaves out one equation in the last block row, which is:

$$\psi(0, 0, \ldots, 0) - \psi(1, 0, \ldots, 0) = \delta_L(0).$$  \hspace{1cm} (C19)

By induction hypothesis, all equations in the first block row of Eq. (C15) can be written as linear combinations of set 0. Furthermore, if $\delta_p(N)$ satisfies property (C3), also $\delta_p(N) + c$ satisfies the same property, so by induction hypothesis, also all equations in the second block row of Eq. (C15) can be written as linear combinations of set 1. So any row in the first and second block row in Eq. (C15) can be written as linear combinations of those equations involving $\tilde{B}(L)$.

We are left to show that equations in the last block row of Eq. (C15) can be written as linear combinations of equations involving $\tilde{B}(L)$. Thanks to what we have just demonstrated, this is is equivalent to showing that the equations in the last block row are linear combinations of the first two block rows of Eq. (C15), and of Eq. (C19). The equations of the last block row of Eq. (C15) are of the form

$$\psi(n_L = 0, \{n_i\}) - \psi(n_L = 1, \{n_i\}) = \delta_L(N).$$  \hspace{1cm} (C20)

Given any fixed set of $\{n_i\}$, which fixes $\bar{N} = \sum_{i \neq p} n_i$, we will now show how to obtain Eq. (C20) by relating $\psi(n_L = 0, \{n_i\})$ to $\psi(n_L = 1, \{n_i\})$ only using the equations of the first two block rows of Eq. (C15) and Eq. (C19). We can first relate $\psi(n_L = 0, \{n_i\})$ to $\psi(n_L = 0, \{0\})$ using the equations in the first block row $\bar{N}$ times, each time changing a non null $n_i$ to zero. We will obtain the following result:

$$\psi(n_L = 0, \{n_i\}) - \psi(n_L = 0, \{0\}) = - \sum_{i=1}^{\bar{N}} \delta_{n_i}(\bar{N} - i),$$  \hspace{1cm} (C21)

where $p_i$ is the energy level whose occupation number has been changed to zero at step $i$; the argument $\bar{N} - i$ is given by the fact that at each step we use a linearized DBE where $\sum_{i \neq p} n_i$ decreases by one. Using Eq. (C19)
we obtain
\[
\psi(n_L = 0, \{n_i\}) - \psi(n_L = 1, \{n_i\}) = \delta_L(0) - \sum_{i=1}^{\tilde{N}} \delta_{p_\alpha}(\tilde{N} - i).
\] (C22)

Now we can relate \( \psi(n_L = 1, \{0\}) \) to \( \psi(n_L = 1, \{n_i\}) \) using the equations of the second block row to set each non null \( n_i \) back to 1. If we do this in the exact same order as we did in the previous step, we will obtain
\[
\psi(n_L = 0, \{n_i\}) - \psi(n_L = 1, \{n_i\}) = \sum_{i=1}^{\tilde{N}} \delta_{p_\alpha}(i) + \delta_L(0) - \sum_{i=1}^{\tilde{N}} \delta_{p_\alpha}(\tilde{N} - i).
\] (C23)

The “new” term has the same set of \( \{p_i\} \), and the argument \( i \) is given by the fact that \( \sum_{i \neq p_\alpha} n_i \) increases by one at each step, and it starts from 1 since now \( n_L = 1 \). Eq. (C20) is of the same form, therefore the equations on the last block row can be expressed as linear combinations of \( B(L) \) if and only if they represent the same equation, which requires:
\[
\delta_L(\tilde{N}) - \delta_L(0) = \sum_{i=1}^{\tilde{N}} \left[ \delta_{p_\alpha}(i) - \delta_{p_\alpha}(\tilde{N} - i) \right].
\] (C24)

Using property (C3), this last equation reduces to \( \tilde{N} = \tilde{N} \), concluding the proof by induction.

Appendix D: Non-interacting QD: Non-linear effects

In this Appendix we will study the transport coefficients, maximum and peak efficiency and efficiency at peak power of a two-terminal non-interacting QD. We will first consider a single-level QD and then a two-levels QD (with energy-independent tunneling rates).

Let us first consider a QD with a single energy level \( E_1 \). In this case, the kinetic equations with \( E_C = 0 \) reduces to a single simple equation, so it is possible to compute the currents. Using the two-terminal notation introduced in Secs. III B and IV we obtain
\[
J^c = e\gamma(f_2 - f_1),
\]
\[
J^b_2 = (\Delta_{\min} - \theta_0 eV)\gamma(f_2 - f_1),
\] (D1)

where
\[
f_1 = f \left( \frac{\Delta_{\min} + (1 - \theta_0)eV}{k_B T_1} \right),
\]
\[
f_2 = f \left( \frac{\Delta_{\min} - \theta_0 eV}{k_B T_2} \right),
\]
\[
\Delta_{\min} = E_1 - \mu,
\] (D2)

with the Fermi function \( f(x) = [1 + \exp(x)]^{-1} \). Note that the same result could be obtained using the Landauer-Büttiker scattering formalism with a narrow in energy single-level QD transmission probability. Using Eq. (D1), we can compute
\[
G \equiv \frac{\partial J^c}{\partial V} \bigg|_{\Delta T = 0} = \frac{e^2\gamma}{4k_B T} \left[ \frac{\theta_0}{\cosh^2 \left( \frac{\Delta_{\min} - \theta_0 eV}{2k_B T} \right)} + \frac{1 - \theta_0}{\cosh^2 \left( \frac{\Delta_{\min} + (1 - \theta_0)eV}{2k_B T} \right)} \right].
\] (D3)

Note that the value of \( \theta_0 \) can be determined by measuring \( G \).

We can also compute \( S \) using Eq. (D1). The condition \( J^c = 0 \) is satisfied when the arguments of the two Fermi distributions \( f_1 \) and \( f_2 \) are equal, so we obtain
\[
S \equiv - \frac{V}{\Delta T} \bigg|_{J^c = 0} = \frac{1}{eT^*} \Delta_{\min},
\] (D4)

where
\[
T^* = \theta_0 T_1 + (1 - \theta_0) T_2
\] (D5)

is the average reservoir temperature, weighed with \( \theta_0 \). As we can see, the slope of \( S(\mu) \) beyond the linear response regime is strongly determined by \( \theta_0 \). In the linear response regime \( T^* \approx T \), where \( T \) is the average reservoirs’ temperature, so once again we obtain the result we obtained in the quantum limit linear response regime of an interacting QD neglecting the fine structure oscillations, due to other energy levels.

It is now interesting to study the influence of a second energy level on \( S \). In Fig. 13 we show the comparison between the thermopower of a non-interacting two energy level QD (dashed lines) and the multilevel interacting QD considered in Fig. 6 (solid lines), for different values of \( \Delta T \). For the sake of comparison, we choose the value for the two energy levels \( E_1 \) and \( E_2 \) to match the two dominant transition energies of the multilevel interacting QD, in particular \( E_1 = \mu_{N = 3} = 290 \, k_B T \) and \( E_2 = \mu_{N = 4} = 410 \, k_B T \). On one hand, the non-interacting thermopower shows a linear dependence on \( \mu \) when \( \mu \) is close to \( E_1 \) or \( E_2 \), whose slope is independent of \( \Delta T \) (only because \( \theta_0 = 1/2 \)) and is the same as the interacting QD when \( |\Delta_{\min}| < \Delta E \). On the other, when \( \mu \) is close to the middle point \( \mu^* = (E_1 + E_2)/2 \) between \( E_1 \) and \( E_2 \), the thermopower depends linearly on \( \mu \) with a negative slope \( (S \propto -\delta \mu) \), with \( \delta \mu \) being a small displacement with respect to \( \mu^* \) which depends on the temperatures as
\[
\alpha = \frac{1}{e\Delta T \theta_0 T_1} \frac{T_1 \cosh^2 \left( \frac{\Delta \varepsilon}{4k_B T_1} \right) - T_2 \cosh^2 \left( \frac{\Delta \varepsilon}{4k_B T_2} \right)}{T_1 \cosh \left( \frac{\Delta \varepsilon}{4k_B T_1} \right) + T_2 \cosh \left( \frac{\Delta \varepsilon}{4k_B T_2} \right)},
\] (D6)

where \( \Delta \varepsilon = E_2 - E_1 \). Eq. (D6) has been found by expanding the charge current around \( \mu = \mu^* \). Once again, the slope of the interacting and non-interacting QDs are the same in this region. In between instead the fine structure oscillations of the interacting QD create a substantial difference that causes the maximum value of \( S \) to decrease.
At last we can compute the Peltier coefficient for a single energy level QD using Eq. (D1):

$$\Pi = \frac{J^2}{J^c} \bigg|_{\Delta T=0} = \frac{\Delta_{\text{min}}}{e} - \theta_0 V. \tag{D7}$$

Surprisingly, we would have obtained the same result also without setting $\Delta T = 0$. On the other hand, it strongly depends on $V$: increasing the voltage shifts $\Pi$ by $\theta_0 V$, so measuring this shift would allow to measure $\theta_0$. Furthermore, if we compute $\Pi$ in the linear response regime, setting $V = 0$ in Eq. (D7), we can explicitly verify that $\Pi = TS$ is respected. If we set $\theta_0 = 0$, we see that $\Pi = T^* S$, even beyond the linear response regime, for arbitrary temperature and voltage differences. Interestingly, in the presence of two levels the Peltier coefficient beyond linear response still satisfies the relation $(\Pi + \theta_0 V) = TS_{\text{lin}}$, where $S_{\text{lin}}$ is the linear-response thermopower (black dashed curve in Fig. 13), so that $(\Pi + \theta_0 V)$ does not depend on $V$.

The linearization of $G$ and $S$ for a single-level non-interacting QD yields the same results obtained for a multilevel interacting QD, in the quantum limit linear response regime, if we restrict $|\Delta_{\text{min}}| < \Delta E$. We could expect this result since considering a single energy level is intuitively equivalent to sending $\Delta E$ and $E_C$ to infinity in the interacting QD. On the contrary, a single level QD model cannot be used to estimate $K$ since it predicts $K = 0$ [in fact, from Eq. (D1) we have that $J^2 \propto J^c$, so $K$, calculated at $J^c = 0$, vanishes].

We now want to study the power and efficiency of a single-level QD. Considering $\Delta T = T_2 - T_1 > 0$ and inserting Eq. (D1) into the definitions in Eqs. (19) and (20), we can write the power and efficiency as

$$P = -V J^c = -\gamma e V (f_2 - f_1), \tag{D8}$$

$$\eta = \frac{P}{J^c} = \frac{e V}{\theta_0 e V - \Delta_{\text{min}}} \tag{D9}.$$ 

We will consider a fixed temperature difference $\Delta T = T_2 - T_1 > 0$, and a variable $V$ such that the system behaves as a heat engine ($P > 0$). The power is positive when $V \in [0, V_{\text{stop}}]$, where $V_{\text{stop}}$ is the non zero voltage that creates a null charge current. Imposing this condition we find

$$V_{\text{stop}} = \frac{\Delta_{\text{min}}}{e} \eta_C \frac{\eta_C}{1 - \theta_0 \eta_C}. \tag{D10}$$

Without loss of generality, we can specify our analysis to the region where $\mu < E_1$, $E_1$ being the energy of the single energy level; thus $\Delta_{\text{min}} > 0$, so $V_{\text{stop}}$ is positive, and our system will behave as a heat engine when

$$0 \leq V \leq V_{\text{stop}} = \frac{\Delta_{\text{min}}}{e} \eta_C \frac{\eta_C}{1 - \theta_0 \eta_C}. \tag{D11}$$

Let us now discuss the peak efficiency of the system. We have to maximize Eq. (D9) with respect to $\Delta_{\text{min}}$ and $V$, at fixed $\Delta T > 0$, respecting $\Delta_{\text{min}} > 0$ and Eq. (D11). Since $\eta$ is a growing function of $V$ for $\Delta_{\text{min}} > 0$, $\eta$ will be maximum when computed at the highest allowed voltage, $V_{\text{stop}}$. Inserting $V_{\text{stop}}$ in Eq. (D9) yields:

$$\eta_{\text{max}} = \eta_C. \tag{D12}$$

Thus a single level QD in the sequential tunneling regime always achieves $\eta_{\text{max}} = \eta_C$, regardless of the the temperature difference, $\theta_0$ and the distance between $\mu$ and $E_1^{96,97}$. Using Eq. (D8) we can compute the power when $\eta$ is maximum, i.e. when $V = V_{\text{stop}}$: this yields $P = 0$. These results agrees with the expectation that Carnot’s efficiency is reached when the heat exchange is “reversible”, thus when the power is vanishingly small.

Let us now study the efficiency at peak power, $\eta(P_{\text{peak}})$. $P_{\text{peak}}$ is obtained by maximizing the power with respect to $\Delta_{\text{min}}$ and $V$, at fixed $\Delta T > 0$, imposing $\Delta_{\text{min}} > 0$ and Eq. (D11). By imposing this request we obtain two coupled equations that cannot be solved analytically$^{34}$. However, the Fermi function is always evaluated when the argument is positive, and we approximate it with its exponential tail. By doing so, we obtain

$$\eta(P_{\text{peak}}) = \eta_C \frac{\eta_C}{\eta_C - (1 - \eta_C) \ln (1 - \eta_C)},$$

$$P_{\text{peak}} = \gamma e k_B \Delta T \eta_C \frac{\eta_C}{\left[1 + \tilde{e}(1 - \eta_C)^{1-1/\eta_C}\right] \left[\tilde{e} + (1 - \eta_C)^{1/\eta_C}\right]}, \tag{D13}$$

where $\tilde{e}$ is Napier’s constant. These equations provide an approximate expression of $P_{\text{peak}}$ and $\eta(P_{\text{peak}})$ for a single level QD, valid for any reservoir temperatures. Note how $\eta(P_{\text{peak}})$ only depends on $\eta_C$, while $P_{\text{peak}}$ depends on both $T_1$ and $T_2$ through $\eta_C$ and $\Delta T$. Note
\[ \eta(P_{\text{peak}}) \to \eta_C/2 \text{ as } \eta_C \to 0, \text{ as expected from the fact that } ZT \to \infty \text{ for a narrow transmission probability} \]

As we can see in Fig. 14, there is a good agreement between \( \eta(P_{\text{peak}}) \) given in Eq. (D13) and a numerical calculation: the analytic expression slightly underestimates \( \eta(P_{\text{peak}}) \). Furthermore, we can see that the efficiency at peak power goes beyond \( \eta_{CA} \), while it remains under \( \eta_{SS} \). We have also verified that \( P_{\text{peak}} \) given in Eq. (D13) is in good agreement with the numerical calculation.

To further assess the validity of the approximate analytical formulas (D13), we expand them around \( \eta_C = 0 \) and \( \eta_C = 1 \), and compare the obtained results with the exact expansions known in these limiting cases\(^{34}\). An expansion of Eq. (D13) around \( \eta_C = 0 \) yields

\[
\eta(P_{\text{peak}}) = \frac{\eta_C}{2} + \frac{\eta_C^2}{8} + \frac{7}{96} \eta_C^3 + O(\eta_C^4),
\]

where \( T \) is the average temperature in the linear response regime. Our result for \( \eta(P_{\text{peak}}) \) has to be compared with exact expansion of Ref. 34,

\[
\eta(P_{\text{peak}}) = \frac{\eta_C}{2} + \frac{\eta_C^2}{8} + \frac{7}{96} (1 + 0.0627) \eta_C^3 + O(\eta_C^4). \quad (D15)
\]

As we can see the first two orders are exactly the same, while the third order only has \( \approx 6\% \) correction, confirming that Eq. (D13) slightly underestimates the exact result. The expansion of \( P_{\text{peak}} \) instead confirms that in the linear response regime the peak power depends on \( \Delta T^2 \). Furthermore, evaluating the coefficient numerically yields \( P_{\text{peak}} \approx 0.105 \gamma k_B T \) \( (\Delta T/T)^2 \), which is in very good agreement with the result obtained in Eq. (52) for a multilevel QD in the quantum limit linear response regime (note that \( P_{\text{peak}} = Q^* \Delta T^2/4 \)).

FIG. 14: (Color online) Comparison between \( \eta(P_{\text{peak}}) \) computed numerically and using Eq. (D13). Also \( \eta_{CA} \) and \( \eta_{SS} \) are displayed.

\[ 0.0 \quad 0.2 \quad 0.4 \quad 0.6 \quad 0.8 \quad 1.0 \]
\[ \eta_C \]
\[ \text{numerical} \]
\[ \text{analytic} \]
\[ \eta_{CA} \]
\[ \eta_{SS} \]

FIG. 15: (Color online) Maximum power, normalized to the peak value, plotted as a function of \( \mu \) for different values of \( \eta_C \). As opposed to \( \eta_{SS} \), with \( \Delta E = 20 k_B T \) and \( E_C = 0 \) (top panel) and with \( \Delta E = 3 k_B T \) and \( E_C = 0 \) (bottom panel). Tunneling rates are \( \hbar \Gamma_1(p) = \hbar \Gamma_2(p) = 0.01 k_B T \).

An expansion of Eq. (D13) around \( \eta_C = 1 \) yields

\[
\eta(P_{\text{peak}}) \approx 1 + (1 - \eta_C) \ln (1 - \eta_C), \quad (D16)
\]

\[
P_{\text{peak}} \approx \frac{\gamma k_B \Delta T}{1 + \bar{\epsilon}} \left( 1 + \frac{\bar{\epsilon}}{1 + \bar{\epsilon}} (1 - \eta_C) \ln (1 - \eta_C) \right), \quad (D17)
\]

to be compared with the exact expansion\(^{34}\)

\[
\eta(P_{\text{peak}}) \approx 1 + \frac{(1 - \eta_C) \ln (1 - \eta_C)}{1.278}. \quad (D18)
\]

There is a good agreement between the two formulas, and we can see that Eq. (D13) slightly underestimates \( \eta(P_{\text{peak}}) \). We stress that the expression of \( P_{\text{peak}} \) for \( \eta_C \approx 1 \), Eq. (D17), shows that the peak power is proportional to \( \Delta T \), as opposed to \( \Delta T^2 \) as in the linear response regime. Furthermore the peak power approaches its maximum value given by \( P_{\text{peak}} = \gamma k_B \Delta T/(1 + \bar{\epsilon}) \) when \( \eta_C = 1 \).

At last, we have computed numerically the maximum power \( P_{\text{max}} \) in the case of a two-level non-interacting QD.
$P_{\text{max}}$ is computed by maximizing the power only over $V$; the peak of $P_{\text{max}}$ as a function of $\mu$ will thus yield $P_{\text{peak}}$. In Fig. 15 $P_{\text{max}}/P_{\text{peak}}$ is plotted as a function of $\mu$ for $\Delta E = 20 k_B T$ and $\Delta E = 2k_B T$ respectively in panel (a) and (b). Each curve corresponds to a different value of $\eta_C$, starting from the linear-response case (black solid curve) to the extremely non-linear behavior (black dashed curve) at $\eta_C = 0.99$. In panel (a), representing the quantum limit, all curves nearly coincide with the linear-response one (apart for very large values of $\eta_C$). For each value of $\eta_C$ the height of the four peaks is equal, contrary to the case where interaction is present (see Fig 11). This is to be expected, since the two energy levels do not influence each other in the quantum limit in the absence of interaction, so the behavior is essentially dictated by a single level. In the bottom panel instead, representing a case away from the quantum limit, the external peaks are much higher with respect to the internal peaks, especially for small values of $\eta_C$. This is due to the fact that $\Delta E$ is of the order of $k_B T$ and therefore we cannot consider a single energy level at a time, and this produces the peak asymmetry.

* Electronic address: paolo.erdman@sns.it

1 M. S. Dresselhaus, G. Chen, M. Y. Tang, R. G. Yang, H. Lee, D. Z. Wang, Z. F. Ren, J. P. Fleurial, and P. Gogna, Adv. Mater. 19, 1043 (2007).

2 G. J. Snyder and E. S. Toberer, Nat. Mater. 7, 105 (2008).

3 C. J. Vineis, A. Shkouri, A. Majumdar, and M. G. Kanatzidis, Adv. Mater. 22, 3970 (2010).

4 G. Benenti, G. Casati, K. Saito, and R. Whitney, arXiv:1608.05595.

5 A. I. Hochbaum, R. Chen, R. D. Delgado, W. Liang, E. C. Garnett, M. Najariuan, A. Majumdar, and P. Yang, Nature 451, 163 (2008).

6 D. H. Hicks, T. C. Harman, and M. S. Dresselhaus, Appl. Phys. Lett. 63, 3290 (1993).

7 G. D. Mahan and J. O. Sofo, Proc. Natl. Acad. Sci. USA 93, 7436 (1996).

8 C. W. J. Beenakker, Phys. Rev. B 44, 1646 (1991).

9 C. W. J. Beenakker and A. A. M. Staring, Phys. Rev. B 46, 9667 (1992).

10 D. Boese, and R. Fazio, Europhys. Lett. 56, 576 (2001).

11 A. V. Andreev, and K. A. Matveev, Phys. Rev. Lett. 86, 280 (2001).

12 M. Turek, and K. A. Matveev, Phys. Rev. B 65, 115332 (2002).

13 J. Koch, F. von Oppen, Y. Oreg, and E. Sela, Phys. Rev. B 70, 195107 (2004).

14 B. Kubala, and J. König, Phys. Rev. B 73, 195316 (2006).

15 X. Zianni, Phys. Rev. B 75, 045344 (2007).

16 B. Kubala, J. König, and J. Pekola, Phys. Rev. Lett. 100, 066801 (2008).

17 X. Zianni, Phys. Rev. B 78, 165327 (2008).

18 P. A. Jacquet, J. Stat. Phys. 134, 709 (2009).

19 T. A. Costi, and V. Zlatić, Phys. Rev. B 81, 235127 (2010).

20 G. Billings, A. D. Stone, and Y. Alhassid, Phys. Rev. B 81, 205303 (2010).

21 P. Mani, N. Nakaphomkun, E. A. Hoffmann, and H. Linke, Nano Lett. 11, 4679 (2011).

22 O. Entin-Wohlman, and A. Aharony, Phys. Rev. B 85, 085401 (2012).

23 T. Rejec, R. Žitko, J. Mravlje, and Anton Ramšak, Phys. Rev. B 85, 085117 (2012).

24 R. López, and David Sánchez, Phys. Rev. B 88, 045129 (2013).

25 D. M. Kennes, and V. Meden, Phys. Rev. B 87, 075130 (2013).

26 P. Dutt, and K. Le Hur, Phys. Rev. B 88, 235133 (2013).

27 R. Sánchez, B. Sothmann, A. N. Jordan, and M. Böttiker, New J. Phys. 15, 125001 (2013).

28 B. Muralidharan, and M. Grifoni, Phys. Rev. B 88, 045402 (2013).

29 M. A. Sierra and D. Sánchez, Phys. Rev. B 90, 115313 (2014).

30 A. Zimbovskaya, J. Chem. Phys. 142, 244310 (2016).

31 M. Krawiec and K. I. Wysokiński, Phys. Rev. B 73, 075307 (2006).

32 P. Murphy, S. Mukerjee, and J. Moore, Phys. Rev. B 78, 161406 (2008).

33 Y. Dubi and M. Di Ventra, Phys. Rev. B 79, 081302(R) (2009).

34 M. Esposito, K. Lindenberg, and C. Van den Broeck, Europhys. Lett. 85, 60010 (2009).

35 R. Świrkowicz, M. Wierzbicki, and J. Barnaś, Phys. Rev. B 80, 195409 (2009).

36 D. M.-T. Kuo, Jap. J. App. Phys. 48, 125005 (2009).

37 M. Leijnse, M. R. Wegewijs, and K. Flensberg, Phys. Rev. B 82, 045412 (2010).

38 J. Liu, Q. F. Sun, and X. C. Xie, Phys. Rev. B 81, 245323 (2010).

39 O. Entin-Wohlman, Y. Imry, and A. Aharony, Phys. Rev. B 82, 115314 (2010).

40 X. Zianni, Phys. Rev. B 82, 165302 (2010).

41 N. Nakaphomkun, H. Q. Xu, and H. Linke, Phys. Rev. B 82, 235428 (2010).

42 M. Wierzbicki, and R. Świrkowicz, J. Phys.: Condens. Matter 22, 185302 (2010).

43 D. M.-T. Kuo, and Y.-C. Chang, Phys. Rev. B 81, 205321 (2010).

44 R. Sánchez and M. Böttiker, Phys. Rev. B 83, 085428 (2011).

45 O. Karlström, H. Linke, G. Karlström, and A. Wacker, Phys. Rev. B 84, 113415 (2011).

46 Y.-S. Liu, D.-B. Zhang, X.-F. Yang, and J.-F. Feng, Nanotechnology 22, 225201 (2011).

47 B. Sothmann, and M. Böttiker, Europhys. Lett. 99, 27001 (2012).

48 P. Trocha, and J. Barnaś, Phys. Rev. B 85, 085408 (2012).

49 B. Muralidharan and M. Grifoni, Phys. Rev. B 85, 155423 (2012).

50 A. N. Jordan, B. Sothmann, R. Sánchez, and M. Böttiker, Phys. Rev. B 87, 075312 (2013).

51 D. M.-T. Kuo, and Y.-C. Chang, Nanotechnology 24, 175403 (2013).

52 I. Weymann, and J. Barnaś, Phys. Rev. B 88, 085313.
Note that the DBEs (15) could be obtained using the same...
arguments, but assuming that the QD is in a specific configuration $(\{n_i\})_{i \neq p}$, instead of summing over all possible such configurations. To obtain the LBEs, the probability that level $p$ is occupied is given by the marginal probability $\sum_{\{n_i\}_{i \neq p}} P(\{n_i\}, n_p = 1)$.

This approximation for two terminals leads to zero thermal conductance and to the Carnot efficiency, as discussed for single-level quantum dots.\textsuperscript{7,96,97}

Note, however, that the Curzon-Ahlborn limit can be overcome also within linear response when time-reversal symmetry is broken, see Refs. 100–104.

We have also checked that we are well beyond the linear response for the electric conductance.

The black dashed curve is actually equal to the product of the solid black and the red curve. Indeed, due to the Onsager reciprocal relations in this product the linear response Peltier $\Pi$ (which appears in the ratio $r$ for the solid black curve) simplifies with the linear response $\bar{T}S$ (which appears in the ratio $r$ for the red curve).

Notice that in Fig 8 $\Delta E = 20 k_B T$, while in Fig. 3 $\Delta E = 10 k_B T$.

This is actually the upper bound of the Schmiedl-Seifert efficiency, see Refs. 106,107.

For a given value of $\eta_C$, we first find the values $\mu = \bar{\mu}$ and $V = \bar{V}$ leading to the peak value $P_{\text{peak}}$ for power. We then set $\mu = \bar{\mu}$ and vary the voltage $V$ from zero to the thermovoltage.

The effect of degenerate energy levels on the electron and thermal conductance and on thermopower of a QD was considered in Ref. 17.

By summing over all occupation numbers except for a fixed $n_p$, we could demonstrate that the LBEs derive from the kinetic equations.