Finding the quantum thermoelectric with maximal efficiency and minimal entropy production at given power output

Robert S. Whitney

1 Laboratoire de Physique et Modélisation des Milieux Condensés (UMR 5493), Université Grenoble 1 and CNRS, Maison des Magistères, BP 166, 38042 Grenoble, France.

(Dated: October 30, 2014)

We investigate the nonlinear Landauer-Büttiker scattering theory for quantum systems with strong Seebeck and Peltier effects, and consider their use as heat-engines and refrigerators with finite power outputs. This article gives detailed derivations of the results summarized in Phys. Rev. Lett. 112, 130601 (2014). It shows how to use the scattering theory to find (i) the quantum thermoelectric with maximum possible power output, and (ii) the quantum thermoelectric with maximum efficiency at given power output. The latter corresponds to a minimal entropy production at that power output. These quantities are of quantum origin since they depend on system size over electronic wavelength, and so have no analogue in classical thermodynamics. The maximal efficiency coincides with Carnot efficiency at zero power output, but decreases with increasing power output. This gives a fundamental lower bound on entropy production, which means that reversibility (in the thermodynamic sense) is impossible for finite power output. The suppression of efficiency by (non-linear) phonon and photon effects is addressed in detail; when these effects are strong, maximum efficiency coincides with maximum power. Finally, we show that relaxation within the quantum system does not appear to allow it to exceed the bounds derived for relaxation-free systems.

PACS numbers: 73.63.-b, 05.70.Ln, 72.15.Jf, 84.60.Rb

I. INTRODUCTION

Thermoelectric effects in nanostructures,1–5 and molecules,6,7 are of great current interest. They might enable efficient electricity generation and refrigeration,8–10 and could also lead to a new type of sub-Kelvin refrigeration, cooling electrons in solid-state samples to lower temperatures than with conventional cryostats.11 However, they are also extremely interesting at the level of fundamental physics, since they allow one to construct the simplest possible quantum machine that converts heat flows into useful work (electrical power in this case) or vice versa. This makes them an ideal case study for quantum thermodynamics, i.e. the thermodynamics of quantum systems.11

The simplest heat-engine is a thermocouple circuit, as shown in Fig. 1. It consists of a pair of thermoelectrics with opposite thermoelectric responses (filled and open circles) and a load, connected in a ring. Between each such circuit element is a big reservoir of electrons, the reservoir on the left (L) is hotter than the others, $T_L > T_R$, so heat flows from left to right. One thermoelectric’s response causes an electric current to flow the opposite direction to the heat flow (filled circle), while the other’s causes an electric current to flow in the same direction as the heat flow (open circle). Thus the two thermoelectrics turn heat energy into electrical work; a current flow $I$ through the load. The load is assumed to be a device that turns the electrical work into some other form of work; it could be a battery-charger (turning electrical work into chemical work) or a motor (turning electrical work into mechanical work).

The same thermocouple circuit can be made into a refrigerator simply by replacing the load with a power supply. The power supply does work to establish the current $I$ around the circuit, and this current through the thermoelectrics can “drag” heat out of reservoir L. In other words, the electrical current and heat flow are the same as for the heat-engine, but now the former causes the latter rather than vice versa. Thus the refrigerator cools reservoir L, so $T_L < T_R$.

The laws of classical thermodynamics inform us that entropy production can never be negative, and maximal efficiency occurs when a system operates reversibly (zero entropy production). Thus it places fundamental bounds on heat-engine and refrigerator efficiencies; known as Carnot efficiencies. In both cases the efficiency is defined as the power output divided by the power input. For the heat-engine, the power input is the heat current out of the hotter reservoir (reservoir L), $J_L$, and the power output is the electrical power generated $P_{\text{gen}}$. Thus the heat-engine (eng) efficiency is

$$\eta_{\text{eng}} = P_{\text{gen}} / J_L. \quad (1)$$

This efficiency can never exceed Carnot’s limit,

$$\eta_{\text{eng}}^{\text{Carnot}} = 1 - T_R / T_L, \quad (2)$$

where we recall that we have $T_L > T_R$.

For the refrigerator the situation is reversed, the load is replaced by a power supply, and the power input is the electrical power that the circuit absorbs from the power supply, $P_{\text{abs}}$. The power output is the heat current out of the colder reservoir (reservoir L), $J_L$. This is called the cooling power, because it is the rate at which the circuit removes heat energy from reservoir L. Thus the refrigerator (fri) efficiency is

$$\eta_{\text{fri}} = J_L / P_{\text{abs}}. \quad (3)$$
2

2

FIG. 1: (a) The simplest heat-engine is a thermocouple circuit made of two thermoelectrics (filled and open circles). The filled and open circles are quantum systems with opposite thermoelectric responses, an example could be that in (b). For a heat-engine, we assume \( T_L > T_R \), so heat flows as shown, generating a current \( I \), which provides power to a load (battery charger, motor, etc) that converts the electrical power into some other form of work. The same thermocouple circuit can act as a refrigerator; if one replaces the load with a power supply that generates the current \( I \). This induces the heat flow out of reservoir \( L \), which thereby refrigerates reservoir \( L \), so \( T_L < T_R \). Note that in both cases the circuit works because the two thermoelectrics are electrically in series but thermally in parallel. In (b), \( N \) indicates the number of transverse modes in the narrowest part of the quantum system.

This efficiency is often called the coefficient of performance or COP. This efficiency can never exceed Carnot’s limit,

\[
\eta_{\text{Carnot}} = \left( \frac{T_R}{T_L} - 1 \right)^{-1},
\]

where we recall that \( T_L < T_R \) (opposite of heat-engine).

Strangely, the laws of classical thermodynamics do not appear to place a fundamental bound on the power output associated with reversible (Carnot efficient) operation. Most textbooks say that reversibility requires “small” power output, but rarely define what “small” means. The central objective of Ref. [11] was to find the meaning of “small”, and find a fundamental upper bound on the efficiency of an irreversible system in which the power output was not small.

Ref. [11] used a scattering theory which enables one to straightforwardly treat quantum and thermodynamic effects on an equal footing. It summarized two principle results absent from classical thermodynamics. Firstly, there is a quantum bound (qb) on the power output, and no quantum system can exceed this bound (open circles in Fig. 2). Secondly, there is a upper bound on the efficiency at any given power output less than this bound (thick black curves in Fig. 2). The efficiency at given power output can only reach Carnot efficiency when the power output is very small compared to the quantum bound on power output. The upper bound on efficiency then decays monotonically as one increases the power output towards the quantum bound. The objective of this article is to explain in detail the methods used to derive these results, along with the other results that were summarized in Ref. [11].

A. Contents of this article

This article provides detailed derivations of the results in Ref. [11]. The first part of this article is an extended introduction. Section II is a short review of the relevant literature. Section II discusses how we define temperature, heat and entropy. Section IV recalls the connection between efficiency and entropy production in any thermodynamic machine. Section V reviews the nonlinear scattering theory, which section VI uses to make very simple over-estimates of a quantum system’s maximum power output.

The second part of this article considers how to optimize a system which is free of relaxation and has no phonons or photons. Section VII gives a handwaving explanation of the optimal heat engine, while Section VIII gives the full derivation. Section IX gives a handwaving
explanation of the optimal refrigerator, while Section X gives the full derivation. Section XI proposes a system which could in principle come arbitrarily close to the optimal properties given in sections [VII and X] Section XII considers many quantum thermoelectrics in parallel.

The third part of this article considers certain effects neglected in the above idealized system. Section XIII adds the parasitic effect of phonon or photon carrying heat in parallel to the electrons. Section XIV treats relaxation within the quantum system.

II. COMMENTS ON EXISTING LITERATURE

A. Nonlinear systems and the figure of merit ZT

Engineers commonly state that wide ranging applications for thermoelectrics would require them to have a dimensionless figure of merit, ZT, greater than three. The dimensionless figure of merit is a dimensionless combination of the linear response coefficients \( ZT = TGS^2/\Theta \), for temperature \( T \), Seebeck coefficient \( S \), electrical conductance \( G \), and thermal conductance \( \Theta \). However for our purposes, ZT is simply a convenient way of characterizing the efficiency, via

\[
\eta_{\text{eng}} = \eta_{\text{carnot}} \frac{\sqrt{ZT + 1} - 1}{\sqrt{ZT + 1} + 1},
\]

with a similar relationship for refrigerators. Thus an engineer asking for a thermoelectric with a ZT of more than three, is effectively asking for one with an efficiency of more than one third of Carnot efficiency. This is crucial, because the efficiency is the physically important quantity in both linear and nonlinear situations, while ZT is only meaningful in the linear response regime\(^{12–17}\).

Even when \( T_L \) and \( T_R \) are significantly different, linear-response theory rarely fails for the bulk semiconductors that engineers are used to. Yet is will be inadequate for the quantum systems we consider here. The reason is that linear response theory only works when the temperature drop on the scale of the electron relaxation length \( l_{rel} \) (distance travelled before thermalizing) is much less than the average temperature. In bulk systems, the temperature drop on the scale of \( l_{rel} \) is of order \( (T_L - T_R) \times l_{rel}/L \) where \( l_{rel}/L \ll 1 \), so linear-response works well even when \( (T_L - T_R)/T_L \) is of order one. In contrast, for quantum systems \( (L \ll l_{rel}) \), the linear response theory is inapplicable whenever \( (T_L - T_R)/T_L \) is not small.

An example could be getting electricity from thermoelectrics between a diesel motor’s exhaust system \( (T_L \approx 700K) \) and its surroundings \( (T_R \approx 280K) \). For a bulk semiconductor at such temperatures, the relaxation length (inelastic scattering length) is of order the mean free path; 1-100nm in typical semiconductors. The semiconductor is millimetres thick, so the temperature drop on the scale of the relaxation length is tiny (10,000 times smaller than the temperature drop across the whole thermoelectric). If we replaced the bulk semiconductor by a quantum thermoelectric — a quantum dot a few nanometres across or a molecule a fraction of a nanometre across — the full temperature drop will be across this quantum system, placing the quantum system deep in the nonlinear regime.

B. Carnot efficiency

Proposals exist to achieve Carnot efficiencies in bulk\(^{18}\) or quantum\(^{19,21}\) systems. A system can only be Carnot efficient if it is reversible (creates no entropy). Reversibility occurs for quantum thermoelectrics when electrons can only pass between reservoirs L and R at the energy where the electron occupation probabilities are identical in the two reservoirs\(^{19}\). Thus reversibility in a heat-engine circuit, such as Fig. 1, requires two things. Firstly, the quantum thermoelectric must have a \( \delta \)-function-like transmission\(^{18,21}\), which only lets electrons through at energy \( \epsilon_0 \). Secondly, \( l_{rel} \) the load’s resistance must be such that \( eV = \epsilon_0(1 - T_R/T_L) \), so the reservoirs’ occupations are equal at \( \epsilon_0 \), see Fig. 1.

By definition the current in the above reversible situation is strictly zero, and thus so is the power output. However, an infinitesimal deviation from this reversibility point, enables us to see in what manner the power output is “vanishingly small”. Consider a quantum system which lets electrons through in a tiny energy window \( \Delta \) from \( \epsilon_0 \) to \( \epsilon_0 + \Delta \). Let us formally take \( \Delta/(k_BT_{L,R}) \to 0 \) in Eq. (44). Then one has Carnot efficiency, but scattering theory shows (see leading order term in Eq. (44) below) that the power output is

\[
P_{\text{gen}} \propto \frac{1}{\hbar} \Delta^2,
\]

which vanishes as \( \Delta/(k_BT_{L,R}) \to 0 \).

C. Heat-engine efficiency at finite power output and Curzon-Ahborn efficiency

To increase the power output beyond that of a reversible system, one has to consider irreversible machines which generate a finite amount of entropy per unit of work generated. Curzon and Ahlborn\(^{22}\) popularized the idea of studying the efficiency of a heat-engine running at its maximum power output. For classical pumps, they found this efficiency to be \( \eta_{\text{CA}} = 1 - \sqrt{T_L/T_R} \), which is now called the Curzon-Ahborn efficiency, although it was already given in Refs. 24-26. Intriguingly, such pumps as refrigerators have an efficiency at maximum cooling power of zero, although Refs. 27,30 discuss various ways around this.

The response of some thermoelectric heat-engines are given by the “loops” in Fig. 2. Each loop is the efficiency versus power of a given system as one varies the load on the system\(^{22}\). If one takes a peaked transmission function
with a width $\Delta$ (see for example Fig. 5), then the loop moves to the left as one reduces $\Delta$. In the limit $\Delta \to 0$, the whole loop becomes squashed onto the $P_{\text{gen}} = 0$ axis. Ref. [32] showed that in this $\Delta \to 0$ limit, the efficiency at maximum power can be very close to that of Curzon and Ahlborn (marked by the star in Fig. 2), just as Ref. [19] showed that it’s maximum efficiency can be that of Carnot (see previous section). In the language of linear-response, this machine has $ZT \to \infty$. However, its maximum power output is proportional to $e V \Delta / \hbar$ for small $\Delta$, where $V$ is finite (chosen to ensure maximum power). This is much larger than Eq. (3) for $\Delta \to 0$, but it is still vanishingly small. Fig. 2 shows that a system with larger $\Delta$ (such as the red curve) operating near its maximum efficiency will have both higher efficiency and higher power output than the one with small $\Delta$ (left most grey curve) operating at maximum power.

This article, along with Ref. [11], shows how to derive the thick back curve in Fig. 2, thereby showing that there is a fundamental trade-off between efficiency and power output in optimal thermodynamic machines made from thermoelectrics. One can be misled by considering how to tune a system to maximize its efficiency (or efficiency at maximum power), without paying attention to the power output itself. As such, our work overturns the idea that maximizing efficiency at maximum power is the best route to machines with both high efficiency and high power. It also overturns the idea that systems with the narrowest transmission distributions (the largest $ZT$ in linear response) are automatically the best thermoelectrics.

At this point we mention that other works have studied efficiencies for various systems with finite width transmission functions, for which power outputs can be finite. In particular, Ref. [36] considered a boxcar transmission function, which is the form of transmission function that we have shown can be made optimal.

**D. Pendry’s quantum bound on heat-flow**

An essential ingredient in all that follows in this article is that quantum mechanics places a bound on the heat flow through a quantum system. This was noted by Pendry [37], using a scattering theory of the type discussed in Section V below. He showed there is an upper bound on the heat-flow through a quantum system sitting between two reservoirs of fermions (L and R) at different temperatures but the same chemical potential. Ref. [38] pointed out that the same bound applied in the presence of thermoelectric effects and differing chemical potentials. Physically, the upper bound on the heat flow out of reservoir L is when all the electrons and holes arriving at the quantum system from reservoir L escape into reservoir R without impediment, while there is no back-flow of electrons or holes from reservoir R to L. The easiest way to achieve this bound on heat flow out of reservoir L is to couple it through the quantum system to a reservoir R at zero temperature, and then ensure the quantum system does not reflect any particles. In this case the heat current equals

$$J_{qL}^b = \frac{\pi^2}{6\hbar} N k_B T^2_L,$$

where $N$ is the number of transverse modes in the quantum system. We refer to this as the quantum bound (qb) on heat flow, because it depends on the quantum wave nature of the electrons; it depends on $N$, which is given by the cross-sectional area of the quantum system divided by $\lambda_F^2$, where $\lambda_F$ is the electron’s Fermi wavelength. As such $J_{qL}^b$ is ill-defined within classical thermodynamics.

**III. UNIQUELY DEFINING TEMPERATURE, HEAT AND ENTROPY**

Almost all works using the scattering theory for quantum thermoelectrics (including this one) use Butcher’s definitions of heat flow, temperature, and by extension of entropy flow. However, works on classical thermodynamics have shown that the definition of such quantities can be fraught with difficulties. For example, the rate of change of entropy cannot always be uniquely defined in classical continuum thermodynamics [41].
In quantum thermoelectrics, the situation is even more difficult, since the energy distribution for electrons within the quantum systems (circles in Fig. 1) is not of equilibrium type, and so its temperature cannot be defined. When these electrons escape into a reservoir, they carry their non-equilibrium distribution into that reservoir, pushing that reservoir ever so slightly out of equilibrium. Thus it is important to specify the logic which naturally leads to the definitions of temperature, heat flow and entropy flow in Butcher’s work, used here by us (and by many other works that we cite).

To uniquely define these quantities in a physically appealing manner, we consider the following three step procedure, sketched in Fig. 1. For concreteness let us discuss the case of a heat-engine, the procedure for the refrigerator is completely equivalent.

**Step 1.** We assume that initially reservoir L is decoupled from the rest of the circuit. It is in internal equilibrium at temperature $T_L^{(0)}$ and internal energy $E_L^{(0)}$, while the rest of the circuit is in equilibrium at temperature $T_R^{(0)}$ and internal energy $E_R^{(0)}$. Then the Clausius relation defines the initial entropy of left and right parts of the circuit as $S_i^{(0)} = E_i^{(0)}/T_i^{(0)}$ for $i = L, R$.

**Step 2.** We connect reservoir L to the rest of the circuit and leave it connected for a long time $t_{\text{expt}}$. While we assume $t_{\text{expt}}$ is long, we also assume that the reservoirs are all large enough that the energy distributions within them change very little during time $t_{\text{expt}}$. Upon connecting the circuit elements, we assume a transient response during a time $t_{\text{trans}}$, after which the circuit achieves a steady-state. We leave the system connected in this manner for a time $t_{\text{expt}} \gg t_{\text{trans}}$, so the physics is dominated by this steady-state. Even in the steady-state the flow will be noisy[3] so the quantum systems will be in fluctuating non-equilibrium states, due to the fact electrons are discrete and their dynamics are probabilistic. We assume $t_{\text{expt}}$ is very much longer than the correlation time of this noise, so that the noise-induced fluctuations in the electrical and heat currents into any given reservoir have a negligible effect compared to the average currents.

**Step 3.** At the end of the time $t_{\text{expt}}$, we disconnect reservoir L from the rest of the circuit. Again, there will be a transient response, however we assume there is a weak relaxation mechanism within the reservoirs that will cause each of the two parts of the circuit to relax to internal equilibrium. Then one can unambiguously identify the temperature, $T_i$, internal energy $E_i$ and Clausius entropy $S_i = E_i/T_i$ of the two parts of the circuit (for $i = L, R$). Since we assume the reservoirs are large, we can expect $T_i = T_i^{(0)}$.

This procedure enables us to uniquely define the change in heat and entropy in the circuit due to the heat-engine turning heat into useful work. Thus we can unambiguously say that the heat-current out of reservoir $i$ averaged over the time $t_{\text{expt}}$ is

$$\langle J_i \rangle = (E_i^{(0)} - E_i)/t_{\text{expt}}.$$

In what follows we treat the currents for each quantum thermoelectric system separately, thus we write the heat current out of reservoir L as $J_L + J_{L'}$, where $J_L$ is the heat current from reservoir L into the lower thermoelectric in Fig. 1 (the filled circle coupled to reservoir R), and $J_{L'}$ is the heat current from reservoir L into the upper thermoelectric in Fig. 1 (the open circle coupled to reservoir R'). This treatment of each thermoelectric separately is convenient in what follows, and also allows one to generalize the results to more complicated circuits, such as a “thermopile” which contains tens or hundreds of thermoelectrics arranged such that they are electrically in series, but thermally in parallel.

The above definition of average heat current means that the average rate of change of entropy in the circuit during time $t_{\text{expt}}$ is $\langle S_{\text{circuit}} \rangle = \langle S \rangle + \langle S' \rangle$, where $\langle S \rangle$ is the average rate of entropy change associated with the lower thermoelectric in Fig. 1, while $\langle S' \rangle$ is that associated with the upper thermoelectric. If the thermoelectric systems and load are small compared to the reservoirs, then their initial and final state will be determined entirely by the temperature $T_R$, and will be the same at the beginning and end of the process. Thus their rate of change of entropy is zero on average over the above three step procedure, and can be neglected. Then

$$\langle S \rangle = \langle S_L \rangle + \langle S_R \rangle = -\langle J_L \rangle/T_L - \langle J_R \rangle/T_R ,$$

while $\langle S' \rangle$ is the same with $J_L, J_R, T_R$ replaced by $J_{L'}, J_{R'}, T_{R'}$.

The Landauer-Büttiker scattering theory captures long-time average currents (usually called the DC response in electronics), such as electrical current $\langle I \rangle$ and heat current $\langle J \rangle$. We do not review the derivation of the nonlinear Landauer-Büttiker scattering theory, see references in Section V] but do note that it is believed to be exact for non-interacting particles. It can also be applied to interacting electrons when those interactions can be treated in a mean-field approximation (see again section V]. A crucial aspect of the scattering theory is that we do not need to describe the internal state of the quantum system during step 2, which is in general extremely complicated. Instead, we need that quantum system’s transmission function, defined in section V.

In this article we will only discuss the long-time average rates of flows of charge, heat and entropy (not the noisy instantaneous flows), thus for compactness we will not explicitly indicate this average. Hence $I_i$, $J_i$ or $S_i$ should always be interpreted as $\langle I_i \rangle$, $\langle J_i \rangle$ or $\langle S_i \rangle$, respectively.
IV. ENTROPY PRODUCTION

There are simple, but little known, relationships between efficiency and entropy production, that we derive here from the laws of thermodynamics. Consider the lower thermoelectric in Fig. 1 (filled circle), with \( J_L \) and \( J_R \) being steady-state heat currents into it from reservoir L and R. Then the first law of thermodynamics is

\[
J_R + J_L = P_{\text{gen}}, \tag{9}
\]

where \( P_{\text{gen}} \) is the electrical power generated. The Clausius relation for the rate of change of total entropy averaged over long-times as in Eq. (8), is

\[
\dot{S} = -\frac{J_L}{T_L} + \frac{J_L - P_{\text{gen}}}{T_R}, \tag{10}
\]

where we have used Eq. (9) to eliminate \( J_R \).

For a heat engine, we take \( J_L \) to be positive, which means \( T_L > T_R \) and \( J_R \) is negative. We use Eq. (1) to replace \( J_L \) with \( P_{\text{gen}}/\eta_{\text{eng}} \) in Eq. (10). Then, the rate of entropy production by a heat-engine with efficiency \( \eta_{\text{eng}}(P_{\text{gen}}) \) at power output \( P_{\text{gen}} \) is

\[
\dot{S}(P_{\text{gen}}) = \frac{P_{\text{gen}}}{T_R} \left( \frac{\eta_{\text{carnot}}}{\eta_{\text{eng}}(P_{\text{gen}})} - 1 \right), \tag{11}
\]

where the Carnot efficiency, \( \eta_{\text{carnot}} \), is given in Eq. (2).

Hence, knowing the efficiency at power \( P_{\text{gen}} \), tells us the entropy production at that power. Maximizing the former minimizes the latter.

For refrigeration, the load in Fig. 1 is replaced by a power supply, the thermoelectric thus absorbs a power \( P_{\text{abs}} \) to extract heat from the cold reservoir. We take reservoir L as cold (\( T_L < T_R \)), so \( J_L \) is positive. We replace \( P_{\text{gen}} \) by \( -P_{\text{abs}} \) in Eqs. (9,10). We then use Eq. (3) to replace \( P_{\text{abs}} \) by \( J_L/\eta_{\text{fri}} \). Then the rate of entropy production by a refrigerator at cooling power \( J_L \) is

\[
\dot{S}(J_L) = \frac{J_L}{T_R} \left( \frac{1}{\eta_{\text{fri}}(J_L)} - \frac{1}{\eta_{\text{carnot}}} \right), \tag{12}
\]

where the Carnot efficiency, \( \eta_{\text{carnot}} \), is given in Eq. (4). Hence knowing a refrigerator’s efficiency at cooling power \( J_L \) gives us its entropy production, and we see that maximizing the former minimizes the latter.

Eqs. (11,12) hold for systems modelled by scattering theory, because this theory satisfies the laws of thermodynamic.

The rate of entropy production is zero when the efficiency is that of Carnot, but becomes increasingly positive as the efficiency reduces. In this article, we calculate the maximum efficiency for given power output, and then use Eqs. (11,12) to get the minimum rate of entropy production at that power output.

V. NONLINEAR SCATTERING THEORY

The Landauer-Büttiker scattering theory writes particle and heat flows in terms of the transmission function, \( T_{RL}(\epsilon) \), for electrons to go from left (L) to right (R) at energy \( \epsilon \). It is widely used to find the properties of thermoelectric systems from their transmission function, \( T_{RL}(\epsilon) \). Examples in the linear-response regime include Refs. [38,39,40], while nonlinear responses were considered in Refs. [14,15,33,34,57,58]. For very recent reviews, see Refs. [2–4]. Here we do the reverse, instead of finding the thermoelectric efficiency and power output for a given system with a given transmission function, we find the transmission function which maximizes the efficiency. This is rather like Ref. [18], except that we maximize the efficiency for given power output.

In the nonlinear regime, electron-electron interactions cannot be neglected, and the scattering theory treats them as mean-field charging effects. Then the transmission function, \( T_{RL}(\epsilon) \), is a self-consistently determined function of \( T_L, T_R \) and \( V \). In short, the behavior of electrons injected from the leads generates a new distribution of charge in the quantum system, which in turn affects the behavior of those injected electrons. If the charge distribution in the quantum system is treated in a time-independent mean-field approximation (neglecting single electron effects), then it is feasible to determine the transmission function self-consistently with the charge distribution. We note also that the functional RG method has been used to derive the same nonlinear scattering theory for resonant level models, with single-electron charging effects.

Butcher, building on earlier work, applied the linear scattering theory to give the heat current \( J_L \), based on the observation that removing a particle at energy \( \epsilon \) from reservoir \( i \) carries heat \( \epsilon - \mu_i \) out of the reservoir, where \( \mu_i \) is the reservoir’s chemical potential. Thus one is cooling the reservoir by removing a electron above the Fermi surface, but one is heating the reservoir by removing a electron below the Fermi surface. Butcher’s method for thermoelectric responses was recently extended to the above nonlinear regime.

Within the scattering theory, it is convenient to treat empty states below a reference chemical potential (which we define as \( \epsilon = 0 \)), as “holes”. Then we do not need to keep track of a full Fermi sea of electrons, but only the holes in that Fermi sea. Then the heat-currents out of reservoirs L and R and into the quantum system are

\[
J_L = \frac{1}{h} \sum_\mu \int_0^\infty d\epsilon \left( \epsilon - \mu e^V_L \right) T_{RL}^{\mu \mu}(\epsilon) \left[ f_L(\epsilon) - f_R^L(\epsilon) \right], \tag{13}
\]

\[
J_R = \frac{1}{h} \sum_\mu \int_0^\infty d\epsilon \left( \epsilon - \mu e^V_R \right) T_{RL}^{\mu \mu}(\epsilon) \left[ f_R^R(\epsilon) - f_R^R(\epsilon) \right], \tag{14}
\]

where \( e^\cdot \) is the electron charge (\( e^\cdot < 0 \)), so \( e^V_i \) is the chemical potential of reservoir \( i \) measured from the reference chemical potential (\( \epsilon = 0 \)). The sum is over \( \mu = 1 \) for “electron” states (full states above the reference chemical potential), and \( \mu = -1 \) for “hole” states (empty...
states below that chemical potential). The Fermi function for particles entering from reservoir $j$, is

$$f^\mu_j(\epsilon) = (1 + \exp [(\epsilon - \mu e V_j)/(k_B T_j)])^{-1}. \quad (15)$$

The transmission function, $T^{\mu\nu}_{ij}(\epsilon)$, is the probability that a particle $\mu$ with energy $\epsilon$ entering the quantum system from reservoir $j$ will exit the quantum system into reservoir $i$ as a particle $\nu$ with energy $\epsilon$. We only allow $\nu = \mu$ here, since we do not consider electron to hole scattering within the quantum system (only common when superconductors are present). Interactions mean that $T^{\mu\mu}_{RL}(\epsilon)$, is a self-consistently determined function of $T_L$, $T_R$ $V_L$ and $V_R$.

The system generates power $P_{\text{gen}} = (V_R - V_L)I_L$, so

$$P_{\text{gen}} = \frac{1}{h} \sum_\mu \int_0^\infty d\epsilon \mu e(V_R - V_L) T^{\mu\mu}_{RL}(\epsilon) [f^\mu_L(\epsilon) - f^\mu_R(\epsilon)], \quad (16)$$

It is easy to verify that Eqs. (13-16) satisfy the first law of thermodynamics given in Eq. (9). Without loss of generality, we take the reference chemical potential to be that of reservoir L, with

$$V_L = 0, \quad V_R = V. \quad (17)$$

Then the above equations coincide with Eqs. (8,9) of Ref. \[11\]. The theory assumes the quantum system to be relaxation-free, although decoherence is allowed as it does not change the structure of Eqs. (13-16). Relaxation is considered in Section [XIV].

In what follows the thermoelectric response is taken to be such that electrical current flows in the opposite direction to heat current (filled circle in Fig. 1). This is the case when transport is “electron-dominated”, i.e. dominated by electron states above reservoir L’s Fermi surface ($\mu = 1$). However, everything we say applies equally to cases where the electrical current flows in the same direction as the heat flow (open circle in Fig. 1), except that now the transport is “hole-dominated”, i.e. dominated by empty electron states below reservoir L’s Fermi surface ($\mu = -1$). To get from the electron-dominated case that we will analyze in detail in this work to the hole-dominated case, one performs the transformation $T^{\mu\mu}_{RL}(\epsilon) \rightarrow T^{\mu\mu}_{RL}(-\epsilon)$ with $V \rightarrow -V$. Then we see that $I_L \rightarrow -I_L$ while $J_L$ and $P_{\text{gen}}$ are unchanged.

VI. SIMPLE ESTIMATE OF BOUNDS ON POWER OUTPUT

One of the principal results of Ref. \[11\] is the quantum bounds on the power output of heat-engines and refrigerators. The exact derivation of these bounds are given in Sections [VIII A] and [X A]. Here, we give simple arguments for their basic form based on Pendry’s limit of heat flow discussed in Section [II D] above.

VII. GUESSING THE OPTIMAL TRANSMISSION FOR A HEAT-ENGINE

For a refrigerator, it is natural to argue that the upper bound on cooling power will be closely related to Pendry’s bound, Eq. (6). We will show in Section [X A] that this is the case. A two-lead thermoelectric can extract as much as half of $J_L^{\text{gb}}$. In other words, the cooling power of any refrigerator must obey

$$J_L \leq \frac{1}{2} J_L^{\text{gb}} = \frac{\pi^2}{12 h} Nk_B^2 T_L^2. \quad (18)$$

Now let us turn to a heat-engine operating between a hot reservoir L and cold reservoir R. Following Pendry’s logic, we can expect that the heat current into the quantum system from reservoir L cannot be more than $J_L^{\text{over-estimate}} = \frac{\pi^2}{6 h} Nk_B^2(T_L^2 - T_R^2)$. Similarly, no heat engine can exceed Carnot’s efficiency, Eq. (2). Thus we can safely assume any system’s power output is less than

$$P_{\text{gen}} \leq \frac{\pi^2}{12 h} k_B^2(T_L - T_R)^2. \quad (19)$$

We know this is a significant over-estimate, because maximal heat flow cannot coincide with Carnot efficiency. Maximum heat flow requires $T^{\mu\mu}_{RL}(\epsilon)$ is maximal for all $\epsilon$ and $\mu$, while Carnot efficiency requires a $T^{\mu\mu}_{RL}(\epsilon)$ with a $\delta$-function-like dependence on $\epsilon$ (see Section [XIB]). None the less, the full calculation in Section [VIII A] shows that the true quantum bound on power output is such that

$$P_{\text{gen}} \leq P_{\text{gen}}^{\text{gb}} = A_0 \frac{\pi^2}{h} Nk_B^2 (T_L - T_R)^2, \quad (20)$$

where $A_0 \approx 0.0321$. Thus the simple over-estimate of the bound, $P_{\text{gen}}^{\text{over-estimate}}$, differs from the true bound $P_{\text{gen}}^{\text{gb}}$ by a factor of $(1 + T_R/T_L)/(6A_0)$. In other words it over estimates the quantum bound by a factor between 5.19 and 10.38 (that is 5.19 when $T_R = 0$ and 10.38 when $T_R = T_L$). This is not bad for such a simple estimate.
of electrons from reservoir L to reservoir R (the filled circle Fig. 1a, remembering $e' < 0$, so electron flow is in the opposite direction to $I$). To produce power, the electrical current must flow against a bias, so we require $e' V$ to be positive, with $V$ as in Eq. 17. Inspection of the integrand of Eq. 16 shows that it only gives positive contributions to the power output, $P_{gen}$, when $\mu (f^L_L(e) - f^R_R(e)) > 0$. From Eq. 15, one can show that $f^L_L(e)$ and $f^R_R(e)$ cross at

$$\epsilon_0 = \mu e' V / (1 - T_R/T_L),$$

see Fig. 3. Since $e' V$ is positive, we maximize the power output by blocking the transmission of those electrons ($\mu = 1$) which have $\epsilon < \epsilon_0$, and blocking the transmission all holes ($\mu = -1$). For $\mu = 1$, all energies above $\epsilon_0$ add to the power output. Hence, maximizing transmission for all $\epsilon > \epsilon_0$ will maximize the power output, giving $P_{gen} = P_{gen}^{\mu = 1}$. However a detailed calculation, such as that in Section VIII, is required to find the $V$ which will maximize $P_{gen}$, remembering that $P_{gen}$ depends directly on $V$ as well as indirectly (via the above choice of $\epsilon_0$).

Now we consider maximizing the efficiency at a given power output $P_{gen}$, where $P_{gen} < P_{gen}^{\mu = 1}$. Comparing the integrands in Eqs. (13-16), we see that $J_L$ contains an extra factor of energy $e$ compared to $P_{gen}$. As a result, the transmission of electrons ($\mu = 1$) with large $e$ enhances the heat current much more than it enhances the power output. This means that the higher an electron’s $e$ is, the less efficiently it contributes to power production. Thus, one would guess that it is optimal to have an upper cutoff on transmission, $\epsilon_1$, which would be just high enough to ensure the desired power output $P_{gen}$, but no higher.

Then the transmission function will look like a “bandpass filter” (the “boxcar” form in Fig 5), with $\epsilon_0$ and $\epsilon_1$ further apart for higher power outputs. This guess is correct, however the choice of $V$ affects both $\epsilon_0$ and $\epsilon_1$, so the calculation in Section VIII is necessary to find the $V$, $\epsilon_0$ and $\epsilon_1$ which maximize the efficiency for given $P_{gen}$.

VIII. MAXIMIZING HEAT-ENGINE EFFICIENCY FOR GIVEN POWER OUTPUT

Now it is time to enter into the central calculations of this article, that of finding the maximum possible efficiency of a quantum thermoelectric system with given power output. In this section we consider heat-engines, while Section X addresses refrigerators.

For a heat engine, our objective is to find the transmission function $T_{RL}^{\mu \gamma}(\epsilon)$ that maximizes the efficiency $\eta_{eng}(P_{gen})$ for given power output $P_{gen}$. To do this we treat $T_{RL}^{\mu \gamma}(\epsilon)$ as a set of many slices each of width $\Delta \epsilon \to 0$, see the sketch in Fig. 5a. We define $\tau_\gamma^{\mu \gamma}$ as the height of the $\gamma$th slice, which is at energy $\epsilon_\gamma = \gamma \Delta \epsilon$. Our objective is to find the optimal value of $\tau_\gamma^{\mu \gamma}$ for each $\mu, \gamma$, and optimal values of the bias, $V$; all under the constraint of fixed $P_{gen}$. Often such optimization problems are formidable, however this one is fairly straightforward.

The efficiency increases for a fixed power, $P_{gen}$, if $J_L$ decreases for fixed $P_{gen}$. If we make an infinitesimal increase of $\tau_\gamma^{\mu \gamma}$ for fixed $P_{gen}$, then $J_L$ decreases if

$$\left. \frac{\partial J_L}{\partial \tau_\gamma^{\mu \gamma}} \right|_{P_{gen}} = \left. \frac{\partial J_L}{\partial \tau_\gamma^{\mu \gamma}} \right|_{V} - \left. \frac{J'_L}{P_{gen}} \frac{\partial P_{gen}}{\partial \tau_\gamma^{\mu \gamma}} \right|_{V} < 0,$$

(22)

where $|_\epsilon$ indicates that the derivative is taken at constant $x$. The primed means $\partial / \partial V$ for fixed transmission functions, $T_{RL}^{\mu \gamma}(\epsilon)$. Eq. (22) corresponds to asking how $J_L$ changes when one slightly increases $\tau_\gamma^{\mu \gamma}$, and then adjusts $V$ to return $P_{gen}$ to its initial value. Comparing Eq. (13) and Eq. (16), one sees that

$$\left. \frac{\partial J_L}{\partial \tau_\gamma^{\mu \gamma}} \right|_{V} = \left. \frac{\epsilon_\gamma}{\mu e' V} \frac{\partial P_{gen}}{\partial \tau_\gamma^{\mu \gamma}} \right|_{V}.$$

(23)

where $V$ is given in Eq. (17). Thus $\eta_{eng}(P_{gen})$ grows with a small increase of $\tau_\gamma^{\mu \gamma}$ if

$$\left(\epsilon_\gamma - \mu e' V \frac{J'_L}{P_{gen}}\right) \left. \frac{\partial P_{gen}}{\partial \tau_\gamma^{\mu \gamma}} \right|_{V} < 0,$$

(24)

where $P_{gen}$, $P_{gen}'$, $J_L$, $J'_L$ and $e' V$ are positive.

For what follows, let us define two energies

$$\epsilon_0 = e' V / (1 - T_R/T_L),$$

(25)

$$\epsilon_1 = e' V J'_L / P_{gen}'. $$

(26)

One can see that $\left. \partial (P_{gen}/\partial \tau_\gamma^{\mu \gamma}) \right|_\epsilon > 0$ when both $\mu = 1$ and $\epsilon > \epsilon_0$, and is negative otherwise. Thus, for $\mu = 1$, Eq. (24) is satisfied when $\epsilon_\gamma$ is between $\epsilon_0$ and $\epsilon_1$. For $\mu = -1$, Eq. (24) is never satisfied.
for $\tau_0$, there are no stationary points, which is why side of Eq. (24) is not zero for any $\epsilon$.

In (a) $T_{\delta L}^\mu(\epsilon)$ is considered as infinitely many slices of width $\delta \to 0$, so slice $\gamma$ has energy $\epsilon_\gamma \equiv \gamma \delta$ and height $\tau_0^\mu$. Section VIII shows that the optimization leads to (b) with ugly equations for $\epsilon_0$ and $\epsilon_1$.

A heat-engine is only useful if $P_{\text{gen}} > 0$, and this is only true for $\epsilon_0 < \epsilon_1$. Hence, if $\mu = 1$ and $\epsilon_0 < \epsilon < \epsilon_1$, then $\eta_{\text{eng}}(P_{\text{gen}})$ is maximum for $\tau_0^\mu$ at its maximum value, $\tau_0^\mu = N$. For all other $\mu$ and $\epsilon_0, \eta_{\text{eng}}(P_{\text{gen}})$ is maximum for $\tau_0^\mu$ at its minimum value, $\tau_0^\mu = 0$. Since the left-hand-side of Eq. (24) is not zero for any $\epsilon_0, \epsilon_1$, there are no stationary points, which is why $\tau_0^\mu$ never takes a value between its maximum and minimum values. Thus, the optimal $T_{\delta L}^\mu(\epsilon)$ is a “boxcar” or “top-hat” function,

$$T_{\delta L}^\mu(\epsilon) = \begin{cases} N & \text{for } \mu = 1 & \epsilon_0 < \epsilon < \epsilon_1 \\ 0 & \text{otherwise} \end{cases}$$

(27)

see Fig. 6. It hence acts as a band-pass filter, only allowing flow between L and R for electrons ($\mu = 1$) in the energy window between $\epsilon_0$ to $\epsilon_1$. Here $N$ is the number of transverse modes at the narrowest point in the nanostructure, see Fig. 4.

Substituting a boxcar transmission function with arbitrary $\epsilon_0$ and $\epsilon_1$ into Eqs. (13,16) gives

$$J_L = N \left[ F_L(\epsilon_0) - F_R(\epsilon_0) + F_R(\epsilon_1) \right],$$

(28)

$$P_{\text{gen}} = NeV \left[ G_L(\epsilon_0) - G_R(\epsilon_0) - G_L(\epsilon_1) + G_R(\epsilon_1) \right],$$

(29)

where, we define

$$F_j(\epsilon) = \int_{\epsilon}^{\infty} \frac{x \, dx}{1 + \exp \left[ \frac{(x - eV_j)/(k_B T_j)}{1} \right]}$$

(30)

and

$$G_j(\epsilon) = \int_{\epsilon}^{\infty} \frac{dx}{1 + \exp \left[ \frac{(x - eV_j)/(k_B T_j)}{1} \right]}$$

(31)

which are both positive for any $\epsilon > 0$. Remembering that we took $V_L = 0$ and $V_R = V$, these integrals are

$$F_L(\epsilon) = \epsilon G_L(\epsilon) - \frac{(k_B T_L)^2}{h} \text{Li}_2 \left[ -e^{-\epsilon/(k_B T_L)} \right],$$

(32)

$$F_R(\epsilon) = \epsilon G_R(\epsilon) - \frac{(k_B T_R)^2}{h} \text{Li}_2 \left[ -e^{-(\epsilon - eV)/(k_B T_R)} \right],$$

(33)

$$G_L(\epsilon) = \frac{k_B T_L}{h} \ln \left[ 1 + e^{-\epsilon/(k_B T_L)} \right],$$

(34)

$$G_R(\epsilon) = \frac{k_B T_R}{h} \ln \left[ 1 + e^{-(\epsilon - eV)/(k_B T_R)} \right],$$

(35)

for dilogarithm function, $\text{Li}_2(z) = \int_0^z t \, dt / (e^t - 1)$.

We are only interested in cases where $\epsilon_0$ fulfills the condition in Eq. (25), in this case $(\epsilon_0 - eV)/(k_B T_R) = \epsilon_0/(k_B T_L)$, which means $G_R(\epsilon_0)$ and $F_R(\epsilon_0)$ are related to $G_L(\epsilon_0)$ and $F_L(\epsilon_0)$ by

$$G_R(\epsilon_0) = \frac{T_R}{T_L} G_L(\epsilon_0),$$

(36)

$$F_R(\epsilon_0) - \epsilon_0 G_R(\epsilon_0) = \frac{T_R^2}{T_L^2} \left( F_L(\epsilon_0) - \epsilon_0 G_L(\epsilon_0) \right).$$

(37)

We must find $\epsilon_1$ self-consistently, since Eq. (26) tells us that $\epsilon_1$ depends on $J_L$ and $P_{\text{gen}}$, which depend in-turn on $\epsilon_1$. Substituting Eqs. (28,29) into Eq. (26), one gets a transcendental equation for $\epsilon_1$ as a function of $V$ for given $T_R/T_L$. This equation is too hard to solve analytically (except in the high and low power limits, discussed in Sections VIII A and VIII B respectively). A numerical solution for given $T_R/T_L$ is given by the red-curve in Fig. 7.

Having found $\epsilon_1$ as a function of $V$ for given $T_R/T_L$, we can use Eqs. (28,29) to get $J_L(V)$ and $P_{\text{gen}}(V)$. We can then invert the second relation to get $V(P_{\text{gen}})$. At this point we can find $J_L(P_{\text{gen}})$, and then use Eq. (11) to get the quantity that we desire — the maximum efficiency at given power output, $\eta_{\text{eng}}(P_{\text{gen}})$.

In Section VIII A we do this procedure analytically for high power ($P_{\text{gen}} = P_{\text{th}}^2$), and in Section VIII B we do

![FIG. 6: Finding the $T_{\delta L}^\mu(\epsilon)$ which maximizes the efficiency. In (a) $T_{\delta L}^\mu(\epsilon)$ is considered as infinitely many slices of width $\delta \to 0$, so slice $\gamma$ has energy $\epsilon_\gamma \equiv \gamma \delta$ and height $\tau_0^\mu$. Section VIII shows that the optimization leads to (b) with ugly equations for $\epsilon_0$ and $\epsilon_1$.](image)

![FIG. 7: Solutions of the transcendental equations giving optimal $\epsilon_1$ (heat-engine) or $\epsilon_0$ (refrigerator). In (a), the red curve is the optimal $\epsilon_1(V)$ for $\epsilon_1 > \epsilon_0$, and the thick black line is $\epsilon_0$ in Eq. (25). The red circle and red arrow indicate the low and high power limits discussed in the text. In (b), the red curve is the optimal $\epsilon_0(V)$ for $\epsilon_0 < \epsilon_1$, and the thick black line is $\epsilon_1$ in Eq. (25).](image)
parameterize by the parameter dimensional parameter space (Fig. 7a), which we can that Eq. (26) will then determine a line in this two-

\[ P \]

\[ P \]

\[ \epsilon \]

\[ \epsilon \]

\[ V \]

\[ \epsilon \]

\[ V \]

\[ J \]

\[ J \]

\[ J \]

\[ J \]

\( \epsilon \to \infty \). Thus the transmission function \( T_{RL}^{T_{RL}}(\epsilon) \), taking the form of a Heaviside step function, \( \theta(\epsilon - \epsilon_0) \), where \( \epsilon_0 \) is given in Eq. (25). Taking Eq. (20) combined with Eq. (36) for \( \epsilon_1 \to \infty \), gives

\[ P_{gen}(\epsilon_1 \to \infty) = Ne\epsilon V \left(1 - \frac{R}{T_{RL}}\right) G_L \left(\frac{\epsilon V}{1 - \frac{R}{T_{RL}}}\right). \]

The second consequence of \( P'_{gen} = 0 \), is that the \( V \)-derivative of this expression must be zero. This gives us the condition that

\[ (1 + B_0) \ln(1 + B_0) + B_0 \ln[B_0] = 0 \]

(38)

where we define \( B_0 = \exp[-eV/(k_B T_L - k_B T_R)] = \exp[-\epsilon_0/(k_B T_L)] \). Numerically solving this equation gives \( B_0 \approx 0.318 \). Eq. (25) means that this corresponds to \( eV = -k_B(T_L - T_R) \ln[0.318] = 1.146 k_B(T_L - T_R) \), indicated by the red arrow in Fig. 7. Substituting this back into \( P_{gen}(\epsilon_1 \to \infty) \) gives the maximum achievable value of \( P_{gen} \),

\[ P_{gen}^{\text{opt}} = A_0 \frac{\pi^2}{6} N k_B^2 (T_L - T_R)^2 \]

(39)
with
\[ A_0 \equiv B_0 \ln^2[B_0] / [\pi^2(1 + B_0)] \approx 0.0321. \] (40)
We refer to this as the quantum bound (qb) on power output, because of its origin in the Fermi wavelength of the electrons, \( \lambda_F \). We see this in the fact that \( P_{\text{gen}} \) is proportional to the number of transverse modes in the quantum system, \( N \), which is given by the cross-sectional area of the quantum system divided by \( \lambda_F^2 \). This quantity has no analogue in classical thermodynamics.

The efficiency at this maximum power, \( P_{\text{gen}}^{\text{qb2}} \), is
\[ \eta_{\text{eng}}(P_{\text{gen}}^{\text{qb2}}) = \eta_{\text{eng}}^{\text{Carnot}} / (1 + C_0(1 + T_R/T_L)), \] (41)
with
\[ C_0 = -(1 + B_0) Li_2(-B_0) / (B_0 \ln^2[B_0]) \approx 0.936. \] (42)
As such, it varies with \( T_R/T_L \), but is always more than \( 0.3 \eta_{\text{eng}}^{\text{Carnot}} \). This efficiency is less than Curzon and Ahlborn’s efficiency for all \( T_R/T_L \) (although not much less). However the power output here is infinitely larger than the maximum power output of systems that achieve Curzon and Ahlborn’s efficiency, see Section [11].

The form of Eq. (41) is very different from Curzon and Ahlborn’s efficiency. However, we note in passing that Eq. (41) can easily be written as \( \eta_{\text{eng}}(P_{\text{gen}}^{\text{qb2}}) = \eta_{\text{eng}}^{\text{Carnot}} / [(1 + 2C_0) - C_0 \eta_{\text{eng}}^{\text{Carnot}}] \), which is reminiscent of the efficiency at maximum power found for very different systems (certain classical stochastic heat-engines) in Eq. (31) of Ref. [67].

### B. Optimal heat-engine at low power output

Now we turn to the opposite limit, that of low power output, \( P_{\text{gen}} \ll P_{\text{gen}}^{\text{qb2}} \), where we expect the maximum efficiency to be close to Carnot efficiency. In this limit, \( \epsilon_1 \) is close to \( \epsilon_0 \). Defining \( \Delta = \epsilon_1 - \epsilon_0 \), we expand Eqs. (28) in small \( \Delta \) up to order \( \Delta^3 \). This gives
\[ J_L = \frac{P_{\text{gen}}}{1 - T_R/T_L} + \frac{N \Delta^3 (1 - T_R/T_L)}{3h k_B T_R} g(x_0), \] (43)
\[ P_{\text{gen}} = \frac{N \epsilon_0 \Delta^2 (1 - T_R/T_L)^2}{2h k_B T_R} \times \left[ g(x_0) + \frac{\Delta (1 + T_R/T_L)}{3k_B T_R} \frac{dg(x_0)}{dx_0} \right], \] (44)
where Eq. (25) was used to write \( eV \) in terms of \( \epsilon_0 \), and we defined \( x_0 = \epsilon_0 / (k_B T_L) \), and \( g(x) = e^x / (1 + e^x)^2 \). Thus, for small \( \Delta \) we find that,
\[ \eta_{\text{eng}}(\Delta) = \eta_{\text{eng}}^{\text{Carnot}} \left( 1 - \frac{2\Delta}{3x_0 k_B T_L} + \cdots \right). \] (45)
Eq. (26) gives a transcendental equation for \( x_0 \) and \( \Delta \). However \( \Delta \) drops out when it is small, and the transcendental equation reduces to
\[ x_0 \tanh[x_0/2] = 3, \] (46)
for which \( x_0 \equiv \epsilon_0 / (k_B T_L) = 3.2436 \cdots \). Eq. (25) means that this corresponds to \( eV = 3.2436 k_B (T_L - T_R) \), indicated by the circle in Fig. [4]. Now we can use Eq. (44) to lowest order in \( \Delta \), to rewrite Eq. (45) in terms of \( P_{\text{gen}} \).

This gives the efficiency for small \( P_{\text{gen}} \) as,
\[ \eta_{\text{eng}}(P_{\text{gen}}) = \eta_{\text{eng}}^{\text{Carnot}} \left( 1 - 0.478 \frac{T_R}{T_L} \frac{P_{\text{gen}}}{P_{\text{gen}}^{\text{qb2}}} + \cdots \right), \] (47)
where the dots indicate terms of order \( (P_{\text{gen}} / P_{\text{gen}}^{\text{qb2}})^2 \) or higher. Eq. (11) then gives the minimum rate of entropy production at power output \( P_{\text{gen}} \),
\[ \dot{S}(P_{\text{gen}}) = \frac{P_{\text{gen}}}{T_R} \left( 0.478 \frac{T_R}{T_L} \frac{P_{\text{gen}}}{P_{\text{gen}}^{\text{qb2}}} + \cdots \right). \] (48)
Thus the maximal efficiency at small \( P_{\text{gen}} \) is that of Carnot minus a term that grows like \( P_{\text{gen}}^{1/2} \) (dashed curves in Fig. [9]), and the associated minimal rate of entropy production goes like \( P_{\text{gen}}^{3/2} \).

Note that Eq. (45), shows that Carnot efficiency occurs at any \( x_0 \) (i.e. any \( \epsilon_0 \)) when \( \Delta \) is strictly zero (and so \( P_{\text{gen}} \) is strictly zero). However, for arbitrary \( x_0 \) the factor 0.478 in Eq. (47) is replaced by \( \sqrt{8\pi^2 A_0/\left[9\epsilon_0^2 g(x_0)\right]} \). The value of \( x_0 \) that satisfied Eq. (46) is exactly the one which minimizes this prefactor (its minimum being 0.478), and thus maximizes the efficiency for any small but finite \( P_{\text{gen}} \).

### IX. GUESSING THE OPTIMAL TRANSMISSION FOR A REFRIGERATOR

Here we use simple arguments to guess the transmission function which maximizes a refrigerator’s efficiency for given cooling power. The arguments are similar to those for heat-engines (Section VII), although some crucial differences will appear.

We consider the flow of electrons from reservoir \( L \) to reservoir \( R \) (the filled circle in Fig. [1]), remembering \( \epsilon < 0 \) so electrons flow in the opposite direction to \( I \). To refrigerate, the thermoelectric must absorb power, so the electrical current must be due to a bias, this requires \( eV \) to be negative, with \( V \) as in Eq. (17).

Inspection of the integrand of Eq. (13) shows that it only gives positive contributions to the cooling power output, \( J_L \), when \( \left( f^R_\epsilon - f^L_\epsilon \right) > 0 \). Since \( T_L < T_R \) and \( eV < 0 \), we can use Eq. (15) to show that this is never true for holes (\( \mu = -1 \)), and is only true for electrons (\( \mu = 1 \)) with energies \( \epsilon < \epsilon_1 \), where
\[ \epsilon_1 = -eV / (T_R/T_L - 1). \] (49)
Thus it is counter-productive to allow the transmission of electrons with \( \epsilon > \epsilon_1 \), or the transmission of any holes. Note that this argument gives us an upper cut-off on electron transmission energies, despite the fact it
gave a lower cut-off for the heat engine (see Eq. 21 and the text around it). All electron (µ = 1) energies from zero to ε1 contribute positively to the cooling power J_L. To maximize the cooling power, one needs to maximize \( f^{h}_{\mu}(\epsilon) - f^{R}_{\mu}(\epsilon) \), this is done by taking \( \epsilon V \to -\infty \), for which \( \epsilon_1 \to \infty \). This logic gives the maximum cooling power, which Section X will show equals \( \frac{1}{2} J_{L}^{R} \).

Now we consider maximizing the efficiency at a given cooling output J_L, when \( J_L < \frac{1}{2} J_{L}^{R} \). Comparing the integrands in Eqs. (13,16), we see that the extra factor of \( \epsilon \) in J_L means that allowing the transmission of electrons at low energies has a small effect on cooling power, while costing a similar electrical power as higher energies. Thus, one would guess that it is optimal to have a lower cut-off on transmission, \( \epsilon_0 \), which would be just low enough to ensure the desired cooling power J_L, but no lower. Then the transmission function will look like a “band-pass filter” (the “box-car” in Fig 5), with \( \epsilon_0 \) and \( \epsilon_1 \) further apart for higher cooling power. This guess is correct, however the choice of \( V \) affects both \( \epsilon_0 \) and \( \epsilon_1 \), so the calculation in Section X is necessary to find the \( V \), \( \epsilon_0 \) and \( \epsilon_1 \) which maximize the efficiency for given cooling power J_L.

X. Maximizing Refrigerator Efficiency for Given Cooling Power

Here we find the maximum refrigerator efficiency, also called the coefficient of performance (COP), for given cooling power J_L. The method is very similar to that for heat-engines, and here we mainly summarize the differences. The refrigerator efficiency increases for a fixed cooling power, J_L, if the electrical power absorbed \( P_{abs} = -P_{gen} \) decreases for fixed J_L. This is so if

\[
\frac{\partial P_{abs}}{\partial \tau_{\epsilon}^{\mu}} \bigg|_{J_L} = \frac{\partial P_{abs}}{\partial \tau_{\epsilon}^{\mu}} \bigg|_{V} - \frac{P_{\epsilon}'}{J_L} \frac{\partial J_L}{\partial \tau_{\epsilon}^{\mu}} \bigg|_{V} < 0.
\]

where we recall that the primed means (d/dV). This is nothing but Eq. (22) with \( J_L \to P_{abs} \) and \( P_{gen} \to J_L \). Using Eq. (23), we see that \( \eta_{ri}(J_L) \) grows with \( \tau_{\epsilon}^{\mu} \) for

\[
\left( \frac{-\mu e^{V}}{e_{\gamma} - \frac{P_{\epsilon}'}{J_L}} - \frac{\partial J_L}{\partial \tau_{\epsilon}^{\mu}} \bigg|_{V} \right) < 0,
\]

where \( P_{abs}, P_{\epsilon}', J_L, J_L' \) and \( -eV \) are all positive.

To proceed we define the following energies

\[
\epsilon_0 = -eV J_L/J_L', \quad \epsilon_1 = -eV/(T_R/T_L - 1).
\]

Then one can see that \( \left( \frac{\partial J_L}{\partial \tau_{\epsilon}^{\mu}} \bigg|_{V} \right) \) is positive when both \( \mu = 1 \) and \( \epsilon < \epsilon^{RI}_1 \), and is negative otherwise. Thus, for \( \mu = -1 \), Eq. (51) is never satisfied. For \( \mu = 1 \), Eq. (51) is satisfied when \( \epsilon_1 \) is between \( \epsilon^{RI}_0 \) and \( \epsilon^{RI}_1 \). A refrigerator is only useful if \( J_L > 0 \) (i.e. it removes heat from the cold reservoir), and this is only true for \( \epsilon^{RI}_0 < \epsilon^{RI}_1 \). Hence, if \( \mu = 1 \) and \( \epsilon^{RI}_0 < \epsilon < \epsilon^{RI}_1 \), then \( \eta_{ri}(J_L) \) grows upon increasing \( \tau_{\epsilon}^{\mu} \). Thus the optimum is when such \( \tau_{\epsilon}^{\mu} = N \). For all other \( \mu \) and \( \epsilon_1 \), \( \eta_{ri}(J_L) \) grows upon decreasing \( \tau_{\epsilon}^{\mu} \). Thus the optimum is when such \( \tau_{\epsilon}^{\mu} = 0 \). This gives the boxcar transmission function in Eq. (27), with \( \epsilon_0 \) and \( \epsilon_1 \) given by Eqs. (52,53).

Substituting Eqs. (28,29) into Eq. (52), one gets a transcendental equation for \( \epsilon_0 \) as a function of \( V \) for given \( T_R/T_L \). This equation is too hard to solve analytically (except in the high and low power limits, discussed in Sections XA and XB). A numerical solution for given \( T_R/T_L \) is given by the red-curve in Fig. 5.

Having found \( \epsilon_0 \) as a function of \( V \) for given \( T_R/T_L \), we can use Eqs. (28,29) to get \( J_L(V) \) and \( P_{abs}(V) = -P_{gen}(V) \). We can then invert the first relation to get \( V(J_L) \). Now, we can find \( P_{abs}(J_L) \), and then use Eq. (33) to get the quantity that we desire — the maximum efficiency (or COP), \( \eta_{ri}(J_L) \), at cooling power J_L.

Fig. 5 gives the values of \( \Delta = (\epsilon_1 - \epsilon_0) \) and \( eV \) which result from solving the transcendental equation numerically. As noted, \( \epsilon_1 \) is related to \( eV \) by Eq. (53). The qualitative behaviour of the resulting boxcar transmission function is sketched in Fig. 5. This numerical evaluation enables us to find efficiency as a function of \( J_L \) and \( T_R/T_L \), which we plot in Fig. 9.

A. Quantum bound on refrigerator cooling power

To find the maximum allowed cooling power, J_L, we look for the place where \( J'_L = 0 \). From Eq. (54), we see that this immediately implies \( \epsilon_0 = 0 \). Taking Eq. (28) with \( \epsilon_0 = 0 \), we note by using Eq. (30) that \( F_{L}(0) - F_{R}(0) \) grows monotonically as one takes \(-eV \to \infty \). Similarly, for \( \epsilon_1 \) given by Eq. (29), we note by using Eq. (30) and \( T_R > T_L \) that \( F_{R}(\epsilon_1) - F_{L}(\epsilon_1) \) grows monotonically as one takes \(-eV \to \infty \). Thus we can conclude that \( J_L \) is maximal for \(-eV \to \infty \), which implies \( \epsilon_1 \to \infty \) via Eq. (53). Physically, this limit corresponds to all electrons arriving at the quantum system from reservoir L being transmitted into reservoir R, but all holes arriving from reservoir L are reflected back into reservoir L. At the same time, reservoir R is so strongly biased that it has no electrons with \( \epsilon > 0 \) (i.e. no electrons above the chemical potential of reservoir L) to carry heat from R to L.

In this limit, \( F_{L}(\epsilon_1) = F_{L}(\epsilon_1) = F_{R}(\epsilon_0) = 0 \), so the maximal refrigerator cooling power is

\[
J_L = \frac{\pi^2}{12h} Nk_B T_L^2,
\]

where we used the fact that Li_{2}[1] = \pi^2/12. This is exactly half the quantum bound on heat current that can flow out of reservoir L given in Eq. (4). The quantum bound is achieved by coupling reservoir L to another
reservoir with a temperature of absolute zero, through an 
contact with N transverse mode. By definition a refrig- 
erator is cooling reservoir L below the temperature of the 
other reservoirs around it. In doing so, we show its cool-
ing power is always less than or equal to $J_{lb}^L$. However 
it is intriguing that the maximum cooling power is inde-
dent of the temperature of the environment, $T_R$, of the 
reservoir being cooled (reservoir L). In short, the 
best refrigerator can remove all electrons (or all holes) 
that reach it from reservoir L, but it cannot remove all 
electrons and all holes at the same time.

It is easy to see that the efficiency of the refrigerator 
(COP) at this maximum possible cooling power is zero, 
simply because $|V| \to \infty$, so the power absorbed $P_{abs} \to \infty$. However, one gets exponentially close to this limit 
for $-eV \gg k_BT_R$, for which $P_{abs}$ is large but finite, and 
so $\eta_{fri}(J_L)$ remains finite (see Fig. 5b).

B. Optimal refrigerator at low cooling power

Now we turn to the opposite limit, that of low cooling 
power output, $J_L \ll J_{lb}^L$, where we expect the maximum 
efficiency to be close to Carnot efficiency. In this 
limit, $\epsilon_0$ is close to $\epsilon_1$. Defining $\Delta = \epsilon_1 - \epsilon_0$, we expand Eqs. (28,29) in small $\Delta$ up to order $\Delta^3$. This gives

$$J_L = \frac{P_{abs}}{T_R/T_L - 1} - \frac{N \Delta^3 (T_R/T_L - 1)}{3h k_BT_R} g(x_1),$$

(55)

$$P_{abs} = \frac{N \epsilon_1 \Delta^2 (T_R/T_L - 1)^2}{2h k_BT_R} \times \left[ g(x_1) - \frac{\Delta (T_R/T_L + 1) d(g(x_1))}{3k_BT_R} \right],$$

(56)

where Eq. (55) was used to write $eV$ in terms of $\epsilon_1$, 
and we define $x_1 = \epsilon_1/(k_BT_L)$, and $g(x) = e^x/(1+e^x)^2$. Thus for small $\Delta$ we find that the efficiency is

$$\eta_{fri}(\Delta) = \eta_{Carnot} \left( 1 - \frac{2\Delta}{3x_1 k_BT_L} + \cdots \right).$$

(57)

Note that this is the same Eq. (45) for the heat-engine 
at low power output, except that $x_0$ is replaced by $x_1$, 
and the Carnot efficiency is that of the refrigerator rather 
than that of the heat-engine.

Eq. (52) gives a transcendental equation for $x_1$ and $\Delta$. 
However $\Delta$ drops out when it is small, and the transcen-
dental equation reduces to

$$x_1 \tanh[x_1/2] = 3,$$

(58)

for which $x_1 = \epsilon_1/(k_BT_L) = 3.2436 \cdots$. Again this is the 
same as for a heat-engine, Eq. (46), but with $x_1$ replacing 
$x_0$. Eq. (58) means that this corresponds to $-eV = 3.2436 k_BT_L - T_L$, indicated by the circle in Fig. 3b.

Now we can use Eq. (55) to lowest order in $\Delta$, to rewrite 
Eq. (57) in terms of $J_L^1$. This gives the efficiency (or

coefficient of performance, COP) for small $J_L$ as,

$$\eta_{fri}(J_L) = \eta_{Carnot} \left( 1 - 1.09 \sqrt{\frac{T_R}{T_R - T_L}} \frac{J_L}{J_{lb}^L} + \cdots \right).$$

(59)

where the dots indicate terms of order $(J_L/J_{lb}^L)$ or 
higher. Eq. (12) gives the minimum rate of entropy gen-
eration at cooling power output $J_L$, as

$$\dot{S}(J_L) = \frac{J_L}{T_L} \left( 1 - \frac{T_L}{T_R} \right) \left( 1.09 \sqrt{\frac{T_R}{T_R - T_L}} \frac{J_L}{J_{lb}^L} + \cdots \right).$$

(60)

Thus, we conclude that the maximum efficiency at small $J_L$ 
is that of Carnot minus a term that grows like $J_{lb}^{3/2}$ 
(dashed curves in Fig. 8b), while the associated minimum 
entropy production goes like $J_{lb}^{3/2}$.

Finally, we note that Carnot efficiency occurs at $J_L = 0$ 
at any $x_1 = \epsilon_1/(k_BT_L)$. However for arbitrary $x_1$, the 
1.09 factor in Eq. (59) is replaced by $\sqrt{4\pi^2/[27x_1^2g(x_1)]}$. 
The condition in Eq. (58) minimizes this factor (the min-
umum being 1.09), and thereby maximizes the efficiency 
for given $J_L$.

XI. IMPLEMENTATION WITH A CHAIN OF 
QUANTUM SYSTEMS

The previous sections have shown that for a heat-
engine or refrigerator to have maximum efficiency at 
given power output, the quantum thermoelectric systems 
must have boxcar transmission functions with the right
position and width. Here we propose a manner of making such transmission functions. However, before considering boxcar transmission functions, we note that to achieve maximum power output one requires the transmission function to be a Heaviside step-function. This is easily implemented with point-contact. The transmission function of a point-contact is

$$\mathcal{T}_{RL}(\epsilon) = \sum_i \left| \langle k | \epsilon - \mathcal{H}_{\text{chain}} \rangle^{-1} | 1 \rangle \right|^2 a_i,$$  

(66)

where $\mathcal{H}_{\text{chain}}$ is given by

$$\begin{pmatrix}
-i a_0/2 & t_1 & 0 & 0 & 0 \\
t_1 & 0 & t_2 & 0 & 0 \\
0 & t_2 & 0 & t_3 & 0 \\
0 & 0 & t_3 & 0 & t_4 \\
0 & 0 & 0 & t_4 & -i a_0/2
\end{pmatrix}.$$  

(62)

This is easily generalized to arbitrary chain length, $k$. Here we treat $a_0$ as a phenomenological parameter; however in reality it would be given by $|t_0|^2$ multiplied by the density of states in the reservoir. The fact that particles escape from the chain into the reservoirs, means the wavefunction for any given particle in the chain will decay with time. To model this, the Hamiltonian must be non-Hermitian, with the non-Hermiticity entering in the matrix elements for coupling to the reservoirs (top-left and bottom right matrix elements). These induce an imaginary contribution to each eigenstate’s energy $E_i$, with the wavefunction of any eigenstate decaying at a rate given by the imaginary part of $E_i$. The non-Hermiticity of $\mathcal{H}_{\text{chain}}$ also means that its left and right eigenvectors are different, defining $|\psi_i^{(r)}\rangle$ as the ith right eigenvector of the matrix $\mathcal{H}_{\text{chain}}$, and $\langle \psi_i^{(l)} |$ as the ith left eigenvector, we have

$$\langle \psi_i^{(l)} | \mathcal{H}_{\text{chain}} | \psi_j^{(r)} \rangle = \delta_{ij},$$  

(63)

$$\langle \psi_i^{(l)} | \mathcal{H}_{\text{chain}} | \psi_i^{(r)} \rangle = E_i,$$  

(64)

with the resolution of unity

$$\sum_i |\psi_i^{(r)}\rangle \langle \psi_i^{(l)} | = 1$$  

(65)

where $1$ is the $k$-by-$k$ unit matrix.

We define $|1\rangle$ as the vector whose first element is one while all its other elements are zero, and $|k\rangle$ as the vector whose last element (the $k$th element) is one while all its other elements are zero. Then the transmission probability at energy $\epsilon$ is given by

$$\mathcal{T}_{RL}(\epsilon) = \sum_i \left| \langle k | \epsilon - \mathcal{H}_{\text{chain}} \rangle^{-1} | 1 \rangle \right|^2 a_i,$$  

(66)

where $[\cdots]^{-1}$ is a matrix inverse. To evaluate this matrix inverse, we introduce a resolution of unity, Eq. (65), to the left and right of $[\epsilon - \mathcal{H}_{\text{chain}}]^{-1}$. This gives

$$\mathcal{T}_{RL} = \sum_i \left| \langle k | \psi_i^{(r)} \rangle \langle \psi_i^{(l)} | 1 \rangle \right|^2 \frac{a_i}{\epsilon - E_i}.$$  

(67)

For any given set of hop-pings $a_0, t_1, \cdots t_k$, one can easily use a suitable eigenvector finder (we used Mathematica) to evaluate this equation numerically. While an analytic solution is straight-forward for $k \leq 3$. When all hop-pings in the chain are equal, there is a mismatch between the electron’s hopping dynamics in the chain and their free motion in the reservoirs. This causes resonances in the transmission, giving the Fabry-Perot-type oscillations in Fig. 10b for $k = 5$. However, we can carefully tune the hop-pings (to be smallest in the middle of the chain and increasing towards the ends) to get the smooth transmission functions in Fig. 10c. The $k = 5$ curve in Fig. 10c has $t_1 = t_4 = 0.09 a_0$ and $t_2 = t_3 = 0.28 a_0$, and we choose $a_0 = 1.91$ to normalize the band width to 1. As the number of sites in the chain, $k$, increases, the transmission function tends to the desired boxcar function.

The above logic assumes no electron-electron interactions. When we include interaction effects at the mean-field level, things get more complicated. If the states in the chain are all at the same energy $E_0$ when the chain is unbiased, they will not be aligned when there is a bias between the the reservoirs, because the reservoirs also act as gates on the chain states. To engineer a chain where the energies are aligned at the optimal bias, one must adjust the confinement potential of the dots in the chain (or adjust the chemistry of the molecules in the chain) so that their energies are sufficiently out of alignment at zero bias that they all align at optimal bias. In principle, we have the control to do this. However, in practice it would require a great deal of trial-and-error experimental fine tuning. We do not enter further into such practical issues here. Rather, we use the above example to show that there is no fundamental reason that the bound on efficiency cannot be achieved.

**XII. MANY QUANTUM SYSTEMS IN PARALLEL**

One apparently simple way to increase the efficiency at given power output is to increase the number of transverse modes, $N$. This is due to the fact that the efficiency decays with the power output divided by the
quantum bounds in Eqs. (39-54), and these bounds go like $N$. However, there are few thermoelectric quantum systems with $N > 1$. A strong thermoelectric response requires a transmission function that is highly energy dependent, which typically only occurs when the quantum system (point-contact, quantum dot or molecule) has dimensions of about a wavelength, this implies that the number of transverse modes is of order one. Crucial exceptions to this are systems containing superconductors, either SNS structures or Andreev interferometers (see also Ref. [48] and references therein), where strong thermoelectric effects occur for large $N$. However, such superconducting systems are beyond the scope of this work.

In the absence of a superconductor, the only way to get large $N$ is to construct a device consisting of many $N = 1$ systems in parallel, such as a surface covered with a certain density of such systems. In this case $P_{\text{gen}}$ and $J_{\text{L}}$ in Eqs. (49-54) become bounds on the power per unit area, with $N$ being replaced by the number of transverse modes per unit area. With this one modification, all calculations and results in this article can be applied directly to such a situation. Carnot efficiency is achieved for a large enough surface area that the power per unit area is much less than $P_{\text{gen}}$ and $J_{\text{L}}$.

It is worth noting that the number of modes per unit area cannot exceed $\lambda_F^2$, for Fermi wavelength $\lambda_F$. From this we can get a feeling for the magnitude of the bounds discussed in this article. Take a typical semiconductor thermoelectric (with $\lambda_F \sim 10^{-8}\text{m}$), placed between reservoirs at 700 K and 300 K (typical temperatures for a thermoelectric recovering electricity from the heat in the exhaust gases of a diesel engined car). Eq. (39) tells us that to get 100 W of power output from a semiconductor thermoelectric one needs a cross section of at least 4 mm$^2$. Then Eq. (47) tells us that to get this power at 90% of Carnot efficiency, one needs a cross section of at least 0.4 cm$^2$. Remarkably, it is *quantum mechanics* which gives these bounds, even though the cross sections in question are macroscopic.

### XIII. PHONONS AND PHOTONS CARRYING HEAT IN PARALLEL WITH ELECTRONS

Any charge-less excitation (such as phonons or photons) will carry heat from hot to cold, irrespective of the thermoelectric properties of the system. While some of the phonons and photons will flow through the quantum system being used as a thermoelectric device, many will flow via other routes, such as through the thermal insulation between hot and cold reservoirs, see Fig. [11]. A number of theories for these phonon or photon heat current take the form

$$J_{\text{ph}} = \alpha(T_L^\text{gen} - T_R^\text{gen}),$$

where $J_{\text{ph}}$ is the heat flow out of the L reservoir due to phonons or photons. The textbook example of such a theory is that of black-body radiation between the two reservoirs, then $\kappa = 4$ and $\alpha$ is the Stefan-Boltzmann constant. An example relevant to suspended sub-Kelvin nanostructures is a situation where a finite number $N_{\text{ph}}$ of phonon or photon modes carry heat between the two reservoirs, then $\kappa = 2$ and $\alpha \leq N_{\text{ph}}\pi^2 k_B^2/(6h)$.

One of the biggest practical challenges for quantum thermoelectrics is that phonons and photons will often carry much more heat than the electrons. This is simply because the hot reservoir can typically radiate heat in all directions as phonons or photons, while electrons only carry heat through the few nanostructures connected to that reservoir. Thus, in many cases the phonon or photon heat flow will dominate over the electronic one. However progress is being made in blocking phonon and photon flow, by suspending the nanostructure to minimize phonon flow and engineering the electromagnetic environment to minimize photon flow, and it can be hoped that phonon and phonon effects will be greatly reduced in the future. Hence, here we consider the full range from weak to strong phonon or photon heat flows.

For compactness in what follows we will only refer to phonon heat flows (usually the dominant parasitic effect). However, strictly one should consider $J_{\text{ph}}$ as the sum of the heat flow carried by phonons, photons and any more exotic charge-less excitations that might exist in a given circuit (mechanical oscillations, spin-waves, etc).

#### A. Heat-engine with phonons

For heat-engines, the phonon heat-flow is in parallel with electronic heat-flow. Thus the heat-flow for a given $P_{\text{gen}}$ is $(J_L + J_{\text{ph}})$, rather than just $J_L$ (as it was in the absence of phonons). Thus the efficiency in the presence
the phonon heat flow, $J$, in parallel with the heat carried by the electrons. The curves in (a) are for $T_R/T_L = 0.2$, with $J_{ph} = 0.001, 0.1, 1$ (from top to bottom); the curves come from Eq. (71) with $\eta_{eng}(P_{gen})$ given in Fig. 9b. The curves in (b) are for $T_R/T_L = 1.5$, with $J_{ph} = 0.02, 0.1, 0.4$ (from top to bottom); the curves come from Eq. (72) with $\eta_{eng}(P_{gen})$ given in Fig. 9b. The maximum cooling power (open circles) is $(\frac{1}{\pi} J_{L}^{q_b} = J_{ph})$.

FIG. 12: Plots of the maximum efficiency allowed when there is a phonon heat flow, $J_{ph}$, in parallel with the heat carried by the electrons. The curves in (a) are for $T_R/T_L = 0.2$, with $J_{ph} = 0.001, 0.1, 1$ (from top to bottom); the curves come from Eq. (71) with $\eta_{eng}(P_{gen})$ given in Fig. 9b. The curves in (b) are for $T_R/T_L = 1.5$, with $J_{ph} = 0.02, 0.1, 0.4$ (from top to bottom); the curves come from Eq. (72) with $\eta_{eng}(P_{gen})$ given in Fig. 9b. The maximum cooling power (open circles) is $(\frac{1}{\pi} J_{L}^{q_b} = J_{ph})$.

(b) Refrigerator: maximum efficiency for various phonon heat flows

of the phonons is

$$\eta_{eng}^{e+ph}(P_{gen}) = \frac{P_{gen}}{J_{L}(P_{gen}) + J_{ph}}.$$  \tag{69}$$

Writing this in terms of the efficiency, we get

$$\eta_{eng}^{e+ph}(P_{gen}) = \left[\eta_{eng}^{-1}(P_{gen}) + J_{ph}/P_{gen}\right]^{-1},$$  \tag{70}$$

where $\eta_{eng}(P_{gen})$ is the efficiency for $J_{ph} = 0$. Given the maximum efficiency at given power in the absence of phonons, we can use this result to find the maximum efficiency for a given phonon heat flow, $J_{ph}$. An example of this is shown in Fig. 12b. It shows that for finite $J_{ph}$, Carnot efficiency is not possible at any power output.

Phonons have a huge effect on the efficiency at small power output. Whenever $J_{ph}$ is non-zero, the efficiency vanishes at zero power output, with

$$\eta_{eng}^{e+ph}(P_{gen}) = P_{gen}/J_{ph} \quad \text{for} \quad P_{gen} \ll J_{ph}. \tag{71}$$

As $J_{ph}$ increases, the range of applicability of this small $P_{gen}$ approximation (shown as dashed lines in Fig. 12) grows towards the maximum power $P_{gen}^{q_b}$ (open circles). In contrast, phonon heat flows have little effect on the efficiency close to the maximum power output, $P_{gen} \sim P_{gen}^{q_b}$, until they become strong enough that $J_{ph} \sim P_{gen}^{q_b}$.

For strong phonon flow, where $J_{ph} \gg P_{gen}$, Eq. (71) applies at all powers up to the maximum, $P_{gen}^{q_b}$. Then, the efficiency is maximal when the power is maximal, where maximal power is the quantum bound given in Eq. (69). Thus the system with both maximal power and maximal efficiency is that with a Heaviside step transmission function (see section XI).

B. Refrigerator with phonons

For a refrigerator to extract heat from a reservoir at rate $J$ in the presence of phonons carrying a back flow of heat $J_{ph}$, that refrigerator must extract heat at a rate $J_L = J + J_{ph}$. Note that for clarity, in this section we take $J_{ph}$ to be positive when $T_L < T_R$ (opposite sign of that in Eq. (68)). Thus the efficiency, or COP, in the presence of phonons, is the heat current extracted, $J$, divided by the electrical power required to extract heat at the rate $J_L = (J + J_{ph})$. This means that

$$\eta_{e+ph}^{e+ph}(J) = \frac{J \eta_{e+ph}(J + J_{ph})}{J + J_{ph}}.$$  \tag{72}$$

where $\eta_{e+ph}(J)$ is the efficiency for $J_{ph} = 0$. We can use this result to find the maximum efficiency for a given phonon heat flow, $J_{ph}$. An example is shown in Fig. 12b.

Eq. (72) means that the phonon flow suppresses the maximum cooling power, so $J$ must now obey

$$J \leq \frac{1}{2} J_{L}^{q_b} - J_{ph}$$  \tag{73}$$

with $J_{L}^{q_b}$ given in Eq. (6). Thus the upper bound (open circles) in Fig. 12b move to the left as $J_{ph}$ increases.

When the reservoir being refrigerated (reservoir L) is at ambient temperature, $T_R$, then $J_{ph} = 0$ while $J_{L}^{q_b}$ is finite. However as reservoir L is refrigerated (reducing $T_L$), $J_{ph}$ grows, while $J_{L}^{q_b}$ shrinks. As a result, at some point (before $T_L$ gets to zero) one arrives at $J_{ph} = \frac{1}{2} J_{L}^{q_b}$, and further cooling of reservoir L is impossible. Thus given the $T_L$ of $J_{ph}$ for a given system, one can easily find the lowest temperature that reservoir L can be refrigerated to, by solving the equation $J_{ph} = \frac{1}{2} J_{L}^{q_b}$ for $T_L$. To achieve this temperature, one needs the refrigerator with the maximum cooling power (rather than the most efficient one), this is a system with a Heaviside step transmission function (see section XI). Such a system’s refrigeration capacities were discussed in Ref. 14.

We also note that, as with the heat-engine, phonons have a huge effect on the efficiency at small cooling power, as can be seen in Fig. 12b. Whenever $0 < J_{ph} < \frac{1}{2} J_{L}^{q_b}$, the efficiency vanishes for small cooling power, with

$$\eta_{e+ph}^{e+ph}(J) = J \frac{\eta_{e+ph}(J_{ph})}{J_{ph}} \quad \text{for} \quad J \ll J_{ph}. \tag{74}$$

XIV. RELAXATION IN THE QUANTUM SYSTEM

Elsewhere in this article, we neglected relaxation in the quantum system. In other words we assumed that
electrons traverse the system in a time much less than the time for inelastic scattering from phonons, photons or other electrons. We now turn to quantum systems in which there is such relaxation. Under certain circumstances relaxation effects modify a quantum system’s thermoelectric response, thus it is natural to ask if relaxation could enable a system to exceed the bounds found above for relaxationless systems.

We use a voltage-probe model to model such relaxation. A system with relaxation is modeled by a phase-coherent region, and a region in which relaxation occurs. The rate of the relaxation is controlled by modifying the rate at which particles pass from the phase coherent-region to the region where relaxation occurs (for example by choosing the width of the lead between the two). The region where relaxation occurs then acts as a fictitious reservoir $M$, shown in Fig. 13b. Our first step is to separate the phase-coherent system into scatterers 1, 2 and 3, as shown in Fig. 13b; this can be done without loss of generality. The total electrical and heat currents into reservoir $M$ must be zero, and this constraint determines reservoir $M$’s bias, $V_M$, and temperature, $T_M$. If the relaxation is due to electron-phonon or electron-photon interactions (typically any system which is not sub-Kelvin), the phonons or photons with which the electrons interact usually flow easily between the system and the reservoirs. Thus, these phonons or photons can carry heat current between the fictitious reservoir $M$ and reservoirs $L, R$ (dashed arrows in Fig. 13).

A. Method of over-estimation

The optimal choice of $T_{ML}$ and $T_{RM}$ depends on $T_M$, while $T_M$ depend on the heat current, and thus on $T_{ML}$ and $T_{RM}$. The solution and optimization of this self-consistency problem has been beyond our ability to resolve. Instead, we make a simplification which leads to an over-estimate of the efficiency. We assume $V_M, T_M$ are free parameters (not determined from $T_{ML}$ and $T_{RM}$), with $T_M$ between $T_L$ and $T_R$. If we find the optimal $T_{ML}$ and $T_{RM}$ for given $T_M$, and then find the optimal $T_M$ (irrespective of whether it is consistent with $T_{ML}$ and $T_{RM}$ or not), we have an over-estimate of the maximal efficiency. Even with this simplification, we have only been able to address the low-power and high-power limits. However, we show below that this over-estimate is sufficient to prove the following.

1. At low power, relaxation cannot make the system’s efficiency exceed that of the optimal relaxation-free system with $N_{\text{max}}$ modes.

2. Relaxation cannot make a system’s power exceed that of the maximum possible power of a relaxation-free system with $N_{\text{max}}$ modes.

Here, we define

$$N_{\text{max}} = \max\{N_L, N_R\},$$

where $N_L$ and $N_R$ are the number of transverse modes in the system to the left and right of the region where relaxation occurs.

B. Efficiency of heat-engine with relaxation

To get the efficiency for our model of a quantum system with relaxation, we must find the efficiency for the system in Fig. 13b. This system has two “arms”. One arm contains scatterers 1 and 2, and we define its efficiency as $\eta_{\text{eng}}^{(1+2)}$. The other arm contains scatterer 3, and we define its efficiency as $\eta_{\text{eng}}^{(3)}$. The efficiency of the full system, $\eta_{\text{eng}}^{(\text{full})}(P_{\text{gen}})$, is given by

$$\frac{1}{\eta_{\text{eng}}^{(\text{full})}(P_{\text{gen}})} = \frac{p_{\text{rel}}}{\eta_{\text{eng}}^{(1+2)}(P_{\text{gen}})} + \frac{q_{\text{rel}}}{\eta_{\text{eng}}^{(3)}(P_{\text{gen}})},$$

(76)

Here $p_{\text{rel}}$ is the proportion of transmitted electrons that have passed through the arm containing scatterers 1 and 2, while $q_{\text{rel}} = (1-p_{\text{rel}})$ is the proportion that have passed through the arm containing scatterer 3. Physically, $p_{\text{rel}}$ is the probability that an electron entering the quantum system relaxes before transmitting, while $q_{\text{rel}}$ is the probability that it transmits before relaxing. One sees from...
Eq. (76) that the maximal efficiency for a given $p_{\text{rel}}$ occurs when both $\eta_{\text{eng}}^{(1\&2)}$ and $\eta_{\text{eng}}^{(3)}$ are maximal.

The upper-bound on $\eta_{\text{eng}}^{(3)}$ is that given in section VIII with $\eta_{\text{rel}}N_L$ modes to the left and $\eta_{\text{rel}}N_R$ modes to the right. Our objective now is to find the maximum $\eta_{\text{eng}}$ with $N_1 = p_{\text{rel}}N_L$ modes on the left and $N_2 = p_{\text{rel}}N_R$ modes on the right. More precisely our objective is to find an over-estimate of this maximum. For the heat flows indicated in Fig. 13, the efficiency is

$$\eta_{\text{eng}}^{(1\&2)} = \frac{\sqrt{\frac{P_{\text{gen}}}{\sqrt{T_L-T_R}}}}{\sqrt{P_{\text{gen}}^{(3)}(J_2; T_R, T_M)}} \times \left(1 - 0.478 \sqrt{\frac{T_R}{P_{\text{gen}}^{(1\&2)}}} \right).$$

(82)

Since $P_{\text{gen}}^{(1\&2)} = p_{\text{rel}}P_{\text{gen}}$, we can simplify Eq. (82) by noting that

$$\frac{P_{\text{gen}}^{(1\&2)}}{P_{\text{gen}}^{(3)}(\eta_{\text{rel}}N_{\text{max}})} = \frac{P_{\text{gen}}}{P_{\text{gen}}^{(3)}(\eta_{\text{rel}}N_{\text{max}})}$$

(83)

where $P_{\text{gen}}$ is the total power generated by the combined system made of scatterers 1, 2 and 3. Then substituting the result into Eq. (76), we get an over-estimate of the efficiency at power output $P_{\text{gen}}$, which is equal to the upper bound we found in the absence of relaxation, Eq. (47).

Thus we can conclude that for small power outputs, no quantum system with relaxation within it can exceed the upper-bound on efficiency found for a relaxation-free system with $N_{\text{max}}$ transverse modes. Since the proof is based on an over-estimate of the efficiency for a system with relaxation, we cannot say if a system with finite relaxation can approach the bound in Eq. (47). Unlike in the relaxation-free case, we cannot say what properties the quantum system with relaxation (as given in terms of properties of the effective scatterers 1, 2 and 3) are necessary to maximize the efficiency at given power output. We simply know that it cannot exceed Eq. (47).

C. Refrigerator with relaxation

Our objective is to find an over-estimate of the maximal efficiency of a refrigerator that is made of quantum systems in which relaxation occurs. The efficiency of the system with relaxation, $\eta_{\text{rel}}^{\text{total}}(P_{\text{gen}})$, is given by

$$\eta_{\text{rel}}^{\text{total}}(J_L) = p_{\text{rel}}\eta_{\text{rel}}^{(1\&2)}(p_{\text{rel}}J_L) + \eta_{\text{rel}}^{(3)}(q_{\text{rel}}J_L).$$

(84)

thus we need to find an upper bound on $\eta_{\text{rel}}^{(1\&2)}$. We make an over-estimate of this efficiency by taking $T_M$ to be a free parameter between $T_L$ and $T_R$. For given $T_M$, the efficiency of the combined systems 1 and 2 is

$$\eta_{\text{rel}}^{(1\&2)}(J) = J / \left[ P_{\text{abs}}^{(1)}(J_1) + P_{\text{abs}}^{(2)}(J_2) \right].$$

(85)

where $J_1 = J_1^{ph} + J_2^{ph}$ and $J_2 = J_2^{ph} + J_3^{ph} + P_{\text{abs}}^{(1)}$, see Fig. 13d. This efficiency is maximized when $J_1^{ph}, J_2^{ph}, J_3^{ph} = 0$ (since $T_L < T_M < T_R$ means these currents cannot be negative). Taking this case, a little algebra gives

$$1 + \frac{1}{\eta_{\text{rel}}^{(1\&2)}(J)} = \left[ 1 + \frac{1}{\eta_{\text{rel}}^{(1)}(J)} \right] \left[ 1 + \frac{1}{\eta_{\text{rel}}^{(2)}(J)} \right]^{-1}.$$

(86)
where \( J_2 = J + P_{\text{abs}}^{(1)} = J \left[ 1 + 1/N_{\text{fr}}^{(1)}(J) \right] \). Thus to maximize \( \eta_{\text{fr}}^{(1+2)}(J) \) for given \( T_M \), one must maximize both \( \eta_{\text{fr}}^{(1)} \) and \( \eta_{\text{fr}}^{(2)} \). For low power, this can be done using Eq. (59) (much as for the heat-engine in Section XVIB above) giving

\[
\eta_{\text{fr}}^{(1+2)} \leq \eta_{\text{Carnot}} \left( 1 - 1.09 \frac{T_R}{T_R - T_L} \frac{J_L K_{\text{rel}}}{J_L^\text{rel} (N = 1)} \right),
\]

where \( K_{\text{rel}} \) is given in Eq. (80), and \( J_L^\text{rel} (N = 1) \) is given by Eq. (44) with \( N = 1 \). The over-estimate of \( \eta_{\text{fr}}^{(1+2)} \) is maximal when \( K_{\text{rel}} \) is minimal, see Eq. (81). Substituting this into Eq. (84), we see that the efficiency with relaxation does not exceed the result in Eq. (59) for a relaxation-less system, Eq. (54), with \( N_{\text{max}} \) transverse modes.

**D. Quantum bounds on power with relaxation**

For a heat-engine, the arm with scatterers 1 and 2, has a maximum power,

\[
P_{\text{gen}}^{(1+2)} \leq A_0 \frac{\pi^2}{h} k_B^2 \left[ N_1 (T_L - T_M)^2 + N_2 (T_M - T_R)^2 \right],
\]

Since \((T_L - T_M)^2 + (T_M - T_R)^2 \leq (T_L - T_R)^2\), the power of the full system cannot exceed the maximum power of a relaxation-less system, Eq. (39), with \( N_{\text{max}} \) modes.

For a refrigerator, the arm containing scatterers 1 and 2 has a maximum cooling power,

\[
J \leq \left\{ \pi^2 N_1 k_B^2 T_L^2/(12h) \right\} + \left\{ \pi^2 N_2 k_B^2 T_M^2/(12h) - P_{\text{abs}}^{(1)} \right\},
\]

where \( P_{\text{abs}}^{(1)} \) is the electrical power absorbed by scatter 1. The upper (lower) term is the limit on the heat-flow into scatterer 1 (scatterer 2), noting that the heat-flow into scatterer 2 is \( J + P_{\text{abs}}^{(1)} \). Unless \( N_2 \gg N_1 \), the lower limit is the more restrictive one. In any case, the cooling power of the full system can never exceed the maximum power of a relaxation-less system, Eq. (54), with \( N_{\text{max}} \) modes.

**XV. CONCLUSIONS**

Finally, we make a general comment. The upper bound on efficiency at zero power (i.e. Carnot efficiency) is classical, since it is independent of wave-like nature of the electrons. However, this work on quantum thermoelectrics shows that the upper bound on efficiency at finite power is quantum, depending on the ratio of the thermoelectric’s cross-section to the electrons’ Fermi wavelength. If one thought that electrons were classical (strictly zero transverse modes), quantum mechanics tells us that this is not so. It would be interesting to see if quantum mechanics places the same (or similar) bounds on the efficiency at finite power of other thermodynamics machines, such as the Carnot heat engine.

**XVI. ACKNOWLEDGEMENTS**

I am very grateful to M. Büttiker for the suggestion which led to the implementation in Section XI. I thank P. Hänggi for questions on entropy flow which led to section XIII. I thank L. Correa for questions which led to a great improvement of section IX. I thank C. Grenier for an analytic solution of Eq. (67) for \( k = 3 \).
T.E. Humphrey, and H. Linke, Phys. Rev. Lett. 94, 096601 (2005).
R. Kim, S. Datta, M.S. Lundstrom, J. Appl. Phys. 105, 034506 (2009).
A.N. Jordan, B. Sothmann, R. Sanchez, and M. Büttiker, Phys. Rev. B, 87, 075312 (2013).
B. Sothmann, R. Sanchez, A.N. Jordan, and M. Büttiker, New J. Phys. 15 (2013) 095021.
F.L. Curzon, and B. Ahlborn, Am. J. Phys. 43, 22 (1975).
J. Yvon, Proceedings of the International Conference on Peaceful Uses of Atomic Energy (Vol. 2), 337 (United Nations, New York, 1956).
P. Chambadal, Les Centrales Nucléaires (Armand Colin, 1957), p. 41.
I. I. Novikov, J. Nucl. Energy II 7, 125 (1958) [Atomnaya Energiya 3, 409 (1957)].
S. Velasco, J.M.M. Roco, A. Medina, and A.C. Hernández, Phys. Rev. Lett. 78, 3241 (1997).
C. de Tomás, A.C. Hernández, and J.M.M. Roco, Phys. Rev. E 85, 010104(R) (2012).
Y. Apertet, H. Ouerdane, A. Michot, C. Goupil, P. Lecoeur, EPL, 103, 40001 (2013).
L.A. Correa, J.P. Palao, G. Adesso, and D. Alonso, Eprint [arXiv:1408.0473].
This is point is easily overlooked in plots where the power is in units of each system’s maximum power, such as Fig. 4 of Ref. [2]. From that figure, one might imagine that the system with $ZT \to \infty$ is the most efficient at all powers. However, our Fig. 4 makes it clear that curves with larger maximum efficiency have lower maximum power. Ref. [2]’s large $ZT$ curve (they work in the linear response regime, where $ZT$ has meaning) correspond to the left most grey curve in our Fig. 2, which has tiny maximum power.
M. Esposito, K. Lindenburg and C. Van den Broeck, EPL, 85, 60010 (2009).
N. Nakpathomkun, H.Q. Xu, and H. Linke, Phys. Rev. B, 82, 235428 (2010).
D.M. Kennes, D. Schuricht and V. Meden, EPL 102 (2013) 57003 (2013).
M. Leijnse, M.R. Wegewijs, and K. Flensberg, Phys. Rev. B 82, 045412 (2010).
S. Hershfield, K.A. Muttalib, and B.J. Narlow, Phys. Rev. B, 88, 085426 (2013).
J.B. Pendry, J. Phys. A.: Math. Gen. 16, 2161 (1983).
R.S. Whitney, Phys. Rev. B 87, 115404 (2013).
P.N. Butcher, J. Phys.: Condens. Matter, 2, 4869 (1990).
W. Kern, Z. Physik B20, 215 (1975).
W.A. Day, Acta Mechanica 27, 251 (1977).
F. Battista, F. Haupt, J. Splettstoesser, Eprint [arXiv:1405.4326].
M.A. Sierra, and D. Sánchez, Eprint [arXiv:1408.0181].
T. Christen and M. Büttiker, Europhys. Lett. 35, 523 (1996).
H.-L. Engquist and P.W. Anderson, Phys. Rev. B 24, 1151 (1981).
U. Sivan and Y. Imry, Phys. Rev. B 33, 551 (1986).
The “2” on $P_{\text{ref}}$ indicates that this bound is for quantum system between two reservoirs at different temperatures (as in the thermocouple). We do not have an exact calculation of the bound for quantum systems coupled to three (or more) reservoirs, each at their own temperature, however an over-estimate of this bound is given in section IX of Ref. [38].
T. Schmiedl and U. Seifert, EPL, 81, 20003 (2008).
M. Büttiker, Phys. Rev. B 41, 7906 (1990).
This was an idea proposed by M. Büttiker, private comm.
C. Grenier, private comm.
J. Eom, C.-J. Chien, and V. Chandrasekhar, Phys. Rev. Lett. 81, 437 (1998).
J.S. Heron, T. Fournier, N. Mingo, O. Bourgeois, Nano Lett. 9, 1861 (2009). J.-S. Heron, C. Bera, T. Fournier, N. Mingo, and O. Bourgeois, Phys. Rev. B 82, 155458 (2010).
D.R. Schmidt, R.J. Schoelkopf, and A.N. Cleland, Phys. Rev. Lett. 93, 045901 (2004). L.M.A. Pascal, H. Courtis, and F.W.J. Hekking, Phys. Rev. B 83, 125113 (2011).
O. Entin-Wohlman and A. Aharony, Phys. Rev. B 85, 085401 (2012).
M. Büttiker, IBM J. Res. Dev. 32, 63 (1988).