# Maxwell's Refrigerator: An Exactly Solvable Model

Dibyendu Mandal[1], H. T. Quan[2,3] and Christopher Jarzynski[2,4]
[1] *Department of Physics, University of Maryland, College Park, Maryland 20742, U.S.A.*
[2] *Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742, U.S.A.*
[3] *School of Physics, Peking University, Beijing 100871, China.*
[4] *Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, U.S.A.*

We describe a simple and solvable model of a device that – like the "neat-fingered being" in Maxwell's famous thought experiment – transfers energy from a cold system to a hot system by rectifying thermal fluctuations. In order to accomplish this task, our device requires a memory register to which it can write information: the increase in the Shannon entropy of the memory compensates the decrease in the thermodynamic entropy arising from the flow of heat against a thermal gradient. We construct the nonequilibrium phase diagram for this device, and find that it can alternatively act as an eraser of information. We discuss our model in the context of the second law of thermodynamics.

PACS numbers:



In a thought experiment highlighting the statistical nature of the second law of thermodynamics, Maxwell imagined a tiny creature acting as a gatekeeper between two chambers filled with gases at different temperatures. By preferentially allowing fast-moving molecules to pass from the cold to the hot chamber, and slow ones to pass in the other direction, this creature achieves refrigeration without expending energy. As Maxwell put it: "the hot system has got hotter and the cold colder and yet no work has been done, only the intelligence of a very observant and neat-fingered being has been employed" [1].

In this Letter we propose a simple, solvable model of a physical device that accomplishes the same result as Maxwell's intelligent and observant creature: it creates a flow of energy against a thermal gradient, without the input of external work. Our device is a classical two-state system that interacts with a pair of thermal reservoirs and a memory register, which we model as a stream of bits (Fig. 1(a)). The dynamics consist of stochastic transitions, by means of which the device exchanges energy with the reservoirs and modifies the states of the bits. For appropriate values of the model parameters, these dynamics produce a steady state in which there is a continual flow of energy from the cold reservoir to the hot reservoir, and a record of the system's microscopic evolution is continually written to the stream of bits. Our device is fully autonomous, requiring no intervention by an external agent. Its ability to control the flow of energy between the reservoirs emerges entirely from the microscopic equations of motion.

The term "Maxwell's demon" has come to refer not only to the original setting described by Maxwell, but more generally to any situation in which a rectification of microscopic fluctuations produces a decrease of thermodynamic entropy [2, 3]. A consensus has emerged that a physical device could achieve such a result, without violating the second law, if it were simultaneously to write information to a memory register [4–8]. In this view, the act of writing increases the information entropy of the memory register, thereby compensating the decrease of thermodynamic entropy produced by the device. If the information is later erased from the memory register, then by Landauer's principle [4, 9] there must be an increase in thermodynamic entropy elsewhere. This tidy accounting places the Shannon entropy of a sequence of bits on the same thermodynamic footing as the Clausius entropy, defined in terms of heat and temperature. As long as the sum of these entropies never decreases, the second law remains satisfied. See, however, Refs. [10–13] for dissenting perspectives, which suggest that this consensus is at best an appealing narrative based on the presupposition of the second law, rather than an independent explanation.

Maxwell's demon has recently enjoyed increased attention in a broad range of settings, including artificial molecular machines [14], single photon cooling of atoms [15], biomolecular signal transduction [16], quantum information theory [17] and the feedback control of microscopic fluctuations [18–33]. Maxwell's nineteenth-century thought experiment has become a touchstone for discussing the thermodynamic implications of information processing by physical systems [34–37]. While the consensus described above has identified and clarified these implications, far less effort has been devoted to uncovering precisely *how* a physical device, acting on its own, might accomplish the same result as Maxwell's hypothetical being [38–43]. To the best of our knowledge, the autonomous model we introduce below is the first to generate a flow of energy against a thermal gradient, effectively acting as a refrigerator without a power supply – just as in the setup considered by Maxwell, but with the intelligent creature replaced by a dumb device. This contrasts with an earlier model of a device that acts as an *engine*, supplying work by extracting heat from a single thermal reservoir [40]. Our autonomous framework also differs from that of Refs. [18–33] (including the

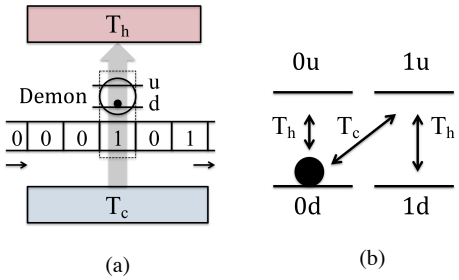


FIG. 1: (a) The device, or "demon", interacts with a sequence of bits, one at a time, while exchanging energy with two thermal reservoirs. (b) The demon makes intrinsic transitions mediated by the hot reservoir (vertical arrows), and the demon and nearest bit make cooperative transitions $0d \leftrightarrow 1u$ mediated by the cold reservoir (diagonal arrows).

experimental realization reported in Ref. [26]), in which external intervention in the form of measurement and feedback is a key element.

In what follows we describe our model and analyze its dynamics. We obtain a nonequilibrium phase diagram for the steady state behavior (Fig. 2), which reveals that our device can act either as a refrigerator, transferring energy from a cold to a hot reservoir, or as an eraser, decreasing the information content of the memory register. Finally, we briefly discuss our model in the context of the second law of thermodynamics.

Our model consists of four components, sketched in Fig. 1(a): a memory register, two thermal reservoirs at temperatures $T_c$ and $T_h > T_c$, and a device that plays the role of Maxwell's demon. The memory register is a sequence of bits (two-state systems) spaced at equal intervals along a tape that slides frictionlessly past the demon. The demon interacts with the nearest bit and with the reservoirs, as we describe in detail in the following paragraphs.

The demon itself is a two-state system, with states $u$ and $d$ characterized by an energy difference $\Delta E = E_u - E_d > 0$. It can make random transitions between these two states by exchanging energy with the hot reservoir, as illustrated by the vertical arrows in Fig. 1(b). We will refer to these as *intrinsic* transitions, to emphasize that they involve the demon but not the bits. The corresponding transition rates satisfy the requirement of detailed balance [1],

$$\frac{R_{d\to u}}{R_{d\leftarrow u}} = e^{-\beta_h \Delta E}, \tag{1}$$

where $\beta_h = 1/kT_h$ and $k$ is Boltzmann's constant. We parametrize these rates as

$$R_{d\to u} = \gamma(1-\sigma), \quad R_{d\leftarrow u} = \gamma(1+\sigma), \quad \sigma = \tanh\frac{\beta_h \Delta E}{2} \tag{2}$$

where $\gamma > 0$ sets a characteristic rate for these transitions, and $0 < \sigma < 1$.

Each bit has two states, 0 and 1, with equal energies. We assume there are no intrinsic transitions between these two states. That is, the state of the bit can change only via interaction with the demon, as we now discuss.

At any instant in time, the demon interacts only with nearest bit. As a result, it interacts sequentially with the bits as they pass by. The duration of interaction with each bit is $\tau = l/v$, where $l$ is the spacing between bits and $v$ is the constant speed of the tape. During one such *interaction interval*, the demon and the nearest bit can make *cooperative* transitions: if the bit is in state 0 and the demon is in state $d$, then they can simultaneously flip to states 1 and $u$, and vice-versa (Fig. 1(b), diagonal arrows). We will use the notation $0d \leftrightarrow 1u$ to denote these transitions, which are accompanied by an exchange of energy with the cold reservoir. The corresponding transition rates again satisfy detailed balance, $R_{0d\to 1u}/R_{0d\leftarrow 1u} = e^{-\beta_c \Delta E}$, where $\beta_c = 1/kT_c$, and we will parametrize them as follows [45]

$$R_{0d\to 1u} = 1-\omega, \quad R_{0d\leftarrow 1u} = 1+\omega, \quad \omega = \tanh\frac{\beta_c \Delta E}{2}, \tag{3}$$

with $0 < \omega < 1$. For later convenience, we also define

$$\epsilon = \frac{\omega - \sigma}{1-\omega\sigma} = \tanh\frac{(\beta_c - \beta_h)\Delta E}{2}, \tag{4}$$

whose value, $0 < \epsilon < 1$, quantifies the temperature difference between the two reservoirs.

Finally, we assume that the incoming bit stream contains a mixture of 0's and 1's, with probabilities $p_0$ and $p_1$, respectively, with no correlations between bits. Let

$$\delta \equiv p_0 - p_1 \tag{5}$$

denote the proportional excess of 0's among incoming bits.

We thus have the following dynamics. When a fresh bit arrives to interact with the demon, its state is 0 or 1. The demon and bit subsequently interact for a time $\tau$, making the transitions shown in Fig. 1(b), thereby exchanging energy with the reservoirs. The state of the bit at the end of the interaction interval is then preserved as the bit joins the outgoing stream, and the next bit in the sequence moves in to have its turn with the demon. The parameters $\gamma$, $\sigma$ and $\omega$ define the intrinsic and cooperative transition rates (Eqs. 2, 3), $\tau$ gives the duration of interaction with each bit, and $\delta$ specifies the statistics of the incoming bits. Under these dynamics, the demon evolves to a periodic steady state, in which its behavior is statistically the same from one interaction interval to the next.

Before proceeding to the solution of these dynamics, we discuss heuristically how our model can achieve the systematic transfer of heat from the cold to the hot reservoir. For this purpose let us assume that each incoming

bit is in state 0, hence $\delta = 1$. At the start of a particular interaction interval, the joint state of the demon and newly arrived bit is either $0u$ or $0d$. The demon and bit then evolve together for a time $\tau$, according to the transitions shown in Fig. 1(b). If the joint state at the end of the interaction interval is $0u$ or $0d$, then it must be the case that every transition $0d \to 1u$ was balanced by a transition $0d \leftarrow 1u$, hence no net energy was absorbed from the cold reservoir. If the final state is $1u$ or $1d$, then we can infer that there was one *net* transition from $0d$ to $1u$, and a quantity of energy $\Delta E$ was absorbed from the cold reservoir. This amounts to *thermal rectification*: over the course of one interaction interval, energy can be withdrawn from the cold reservoir but not delivered to it. Moreover, a record of this process is imprinted in the bit stream, as every outgoing bit in state 1 indicates the absorption of energy $\Delta E$ from the cold reservoir. Since the demon also exchanges energy with the hot reservoir, and since energy cannot accumulate indefinitely within the demon, in the long run we get a net flux of energy from the cold to the hot reservoir, proportional to the rate at which 1's appear in the outgoing bit stream.

More generally, if the incoming bit stream contains a mixture of 0's and 1's, then an excess of 0's (that is, $\delta > 0$) produces a statistical bias that favors the flow of heat from the cold to the hot reservoir, while an excess of 1's ($\delta < 0$) produces the opposite bias. This bias either competes with or enhances the normal thermodynamic bias due to the temperature difference between the two reservoirs. The demon thus affects the flow of energy between the reservoirs, and modifies the states of the bits in the memory register. We now investigate quantitatively the interplay between these two effects.

Once the demon has reached its periodic steady state, let $p'_0$ and $p'_1$ denote the fractions of 0's and 1's in the outgoing bit stream, and let $\delta' = p'_0 - p'_1$ denote the excess of outgoing 0's. Then

$$\Phi \equiv p'_1 - p_1 = \frac{\delta - \delta'}{2} \tag{6}$$

represents the average production of 1's per interaction interval in the outgoing bit stream, relative to the incoming bit stream. Since each transition $0 \to 1$ is accompanied by the absorption of energy $\Delta E$ from the cold reservoir (Fig. 1(b)), the average transfer of energy from the cold to the hot reservoir, per interaction interval, is given by

$$Q_{c\to h} = \Phi \Delta E \,. \tag{7}$$

A positive value of $Q_{c\to h}$ indicates that our device pumps energy against a thermal gradient, like the creature imagined by Maxwell.

To quantify the information-processing capability of the demon, let

$$S(\delta) = -\sum_{i=0}^{1} p_i \ln p_i = -\frac{1-\delta}{2} \ln \frac{1-\delta}{2} - \frac{1+\delta}{2} \ln \frac{1+\delta}{2} \tag{8}$$

denote the information content, per bit, of the incoming bit stream, and define $S(\delta')$ by the same equation, for the outgoing bit stream. Then

$$\Delta S_B \equiv S(\delta') - S(\delta) = S(\delta - 2\Phi) - S(\delta) \tag{9}$$

provides a measure of the extent to which the demon increases the information content of the memory register. We will interpret a positive value of $\Delta S_B$ to indicate that the demon *writes* information to the bit stream, while a negative value indicates *erasure*. (More precisely, since $S(\delta')$ neglects the small correlations that arise between the outgoing bits, $\Delta S_B$ reflects the change in the Shannon information of the *marginal* probability distribution of each outgoing bit.)

From Eqs. 7 and 9 we see that $\Phi$ determines both $Q_{c\to h}$ and $\Delta S_B$. In the Supplemental Material, we show that under the dynamics we have described, the demon reaches a periodic steady state, determined by the model parameters $\Lambda \equiv (\delta, \sigma, \gamma, \omega, \tau)$, in which

$$\Phi(\Lambda) = \frac{\delta - \epsilon}{2}\, \eta(\Lambda) \quad , \quad \eta > 0 \tag{10}$$

and

$$Q_{c\to h}(\beta_h - \beta_c) + \Delta S_B \geq 0 \,. \tag{11}$$

Eq. 11 is a strict inequality when $\delta \neq \epsilon$. An explicit expression for $\eta(\Lambda)$ is given in the Supplemental Material, but for our present purposes the crucial point is that the *sign of* $\Phi$ *is the same as that of* $\delta - \epsilon$. We can think of two effective forces: the bias induced by the incoming bit stream, which favors $\Phi > 0$ when $\delta > 0$ (as discussed above), and the temperature gradient, quantified by $\epsilon$, which favors $\Phi < 0$ (Eq. 7). When these compete, the winner is determined by the difference $\delta - \epsilon$.

Eq. 10 is obtained by solving for the periodic steady state of the demon, using a linear-algebraic approach. Eq. 11 is obtained by constructing a Lyapunov function for the demon and interacting bit. The details of these derivations are provided in the Supplemental Material. Here, we instead use these results to investigate the behavior of our model in the periodic steady state. To that end, we fix $\gamma$ and $\omega$ and construct a phase diagram that illustrates the dependence on $\delta$ and $\epsilon$, for various values of $\tau$, shown in Fig. 2. Let us consider the different regions of this diagram, working our way from right to left.

From Eqs. 7 and 10 it follows that $Q_{c\to h} > 0$ when $\delta > \epsilon$, shown as the most darkly shaded region in Fig. 2. Here, a surplus of incoming 0's prevails over the temperature difference and our demon generates a flow of energy

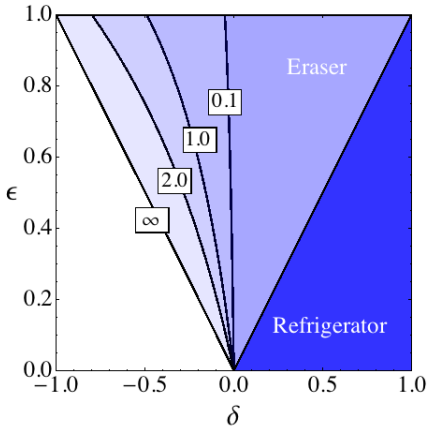


FIG. 2: Phase diagram of our model at fixed $\gamma = 1$ and $\omega = 1/2$. The parameter $\delta$ specifies the incoming bit statistics, and $\epsilon$ is a rescaled temperature difference (Eq. 4). In the most darkly shaded region the demon acts as a refrigerator ($Q_{c\to h} > 0$), while in the lightly shaded regions it acts as an eraser ($\Delta S_B < 0$). The left boundary of the eraser region is shown for $\tau = 0.1$, 1.0, 2.0 and $\infty$. In the blank region at the lower left, our model exhibits neither behavior (see text).

from the cold to the hot reservoir. Moreover, Eq. 11 reveals that $\Delta S_B > 0$ in this region (since $\beta_h < \beta_c$). This agrees with the consensus described earlier: in order for a physical device to act in the manner of Maxwell's demon, it must write information to a physical memory register. In this sense, a bit stream with a low information content can be viewed as a thermodynamic resource, which can be expended (by writing to the available memory) in order to achieve refrigeration.

Now consider the region $\epsilon > \delta > 0$, in which the surplus of 0's in the incoming bit stream is not sufficient to overcome the temperature gradient, and energy flows from the hot to the cold reservoir. Since $\Phi < 0$ we get $\delta' > \delta > 0$ (Eq. 6). This in turn implies $\Delta S_B < 0$, as $S(\delta)$ is a concave function with a maximum at $\delta = 0$. In this region the demon acts as an eraser, lowering the information content of the bit stream, but the price paid for this erasure is the passage of heat from the hot to the cold reservoir.

In the region $\delta < 0$, energy flows from the hot to the cold reservoir (Eqs. 7, 10), but the value of $\Delta S_B$ depends on all the model parameters. In Fig. 2, for four different values of $\tau$, we show the line corresponding to $\Delta S_B = 0$. To the right of this line we have $\Delta S_B < 0$ and to the left we have $\Delta S_B > 0$. In the limit $\tau \to \infty$, the boundary between these two behaviors approaches the line $\epsilon = -\delta$.

Examining the phase diagram as a whole, we see that in the shaded regions our model reaches a steady state in which one thermodynamic resource is replenished at the expense of another. Either energy is pumped against a thermal gradient at the cost of writing information to memory (the refrigerator regime), or else memory is made available, by erasure, at the expense of allowing energy to flow from the hot to the cold reservoir (the eraser regime). The boundary between these two behaviors is the line $\delta = \epsilon$. In the unshaded region at the far left, both resources are consumed, as energy flows down the thermal gradient and information is written to the bit stream.

Finally, to place our model within the context of the second law of thermodynamics, note that the first term on the left side of Eq. 11 is the steady-state change in thermodynamic entropy due to the flow of heat, and the second term is the change in information entropy, per interaction interval. Eq. 11 can be viewed as a modified Clausius inequality, in which the information entropy of a random sequence of data is explicitly assigned the same thermodynamic status as the physical entropy associated with the transfer of heat. (More precisely, Eq. 11 is a weak version of this inequality, as we neglect correlations among the outgoing bits; see Supplemental Material.) Thus our model provides support for the consensus mentioned earlier [4–6], and Eq. 11 is consistent with Landauer's principle [4], which states that a thermodynamic cost must be paid for the erasure of memory. However, in Ref. [4] this cost appears as the dissipation of energy into a single thermal reservoir, whereas in our model it is the transfer of energy from a hot to a cold reservoir.

In summary, we have constructed a simple, solvable model of an autonomous physical system that mimics the behavior of the "neat-fingered being" in Maxwell's thought experiment, generating a systematic flow of energy against a thermal gradient without the input of external work. While Maxwell's creature accomplishes this with intelligence, our inanimate device requires only a memory register to which information can be written. Alternatively, it can harness the flow of energy from hot to cold in order to erase information from the register.

We thank Andy Ballard, Shaon Chakrabarti, Sebastian Deffner, and Zhiyue Lu for useful discussions, and gratefully acknowledge financial support from the National Science Foundation (USA) under grants DMR-0906601, ECCS-0925365, and DMR-1206971, the University of Maryland, College Park, and Peking University.

[1] *Maxwell's Demon 2: Entropy, Classical and Quantum Information, Computing*, H. S. Leff and A. F. Rex, editors, (Institute of Physics Publishing, Bristol, 2003), p. 4.
[2] J. C. Maxwell, *Theory of Heat* (Longmans, London, 1871).
[3] L. Szilard, *Z. Phys.*, **53**, 840 (1929).
[4] R. Landauer, *IBM J. Res. Dev.*, **5**, 183 (1961).
[5] O. Penrose, *Foundations of Statistical Mechanics: A Deductive Treatment* (Pergamon Press, Oxford, 1970).
[6] C. H. Bennett, *Int. J. Theor. Phys.*, **21**, 905 (1982).
[7] C. H. Bennett and R. Landauer, *Sci. Am.*, **253**, 48 (1985).

[8] O. Maroney, in *The Stanford Encyclopedia of Philosophy*, E. N. Zelta, editor, (http://plato.stanford.edu/entries/informationentropy/, 2009) fall 2009 ed.

[9] A. Bérut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz, *Nature*, **483**, 187 (2012).

[10] J. Earman and J. D. Norton, *Stud. Hist. Philos. M. P.*, **29**, 435 (1998).

[11] J. Earman and J. D. Norton, *Stud. Hist. Philos. M. P.*, **30**, 1 (1999).

[12] M. Hemmo and O. Shenker, *J. Philos.*, **107**, 389 (2010).

[13] J. D. Norton, *Stud. Hist. Philos. M. P.*, **42**, 184 (2011).

[14] E. R. Kay, D. A. Leigh, and F. Zerbetto, *Angew. Chem. Int. Ed.*, **46**, 72 (2007), and references therein.

[15] M. G. Raizen, *Sci. Am.*, **304**, 54 (2011).

[16] Y. Tu, *Proc. Natl. Acad. Sci. (USA)*, **105**, 11737 (2008).

[17] L. del Rio, J. Aberg, R. Renner, O. Dahlsten, and V. Vedral, *Nature*, **474**, 61 (2011).

[18] K. H. Kim and H. Qian, *Phys. Rev. E*, **75**, 022102 (2007).

[19] T. Sagawa and M. Ueda, *Phys. Rev. Lett.*, **100**, 080403 (2008).

[20] T. Sagawa and M. Ueda, *Phys. Rev. Lett.*, **102**, 250602 (2009).

[21] K. Jacobs, *Phys. Rev. A*, **80**, 012322 (2009).

[22] F. J. Cao, M. Feito, and H. Touchette, *Physica A*, **388**, 113 (2009).

[23] T. Sagawa and M. Ueda, *Phys. Rev. Lett.*, **104**, 090602 (2010).

[24] M. Ponmurugan, *Phys. Rev. E*, **82**, 031129 (2010).

[25] J. M. Horowitz and S. Vaikuntanathan, *Phys. Rev. E*, **82**, 061120 (2010).

[26] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano, *Nat. Phys.*, **6**, 988 (2010).

[27] D. Abreu and U. Seifert, *Europhys. Lett.*, **94**, 10001 (2011).

[28] S. Vaikuntanathan and C. Jarzynski, *Phys. Rev. E*, **83**, 061120 (2011).

[29] H. Dong, D. Z. Xu, C. Y. Cai, and C. P. Sun, *Phys. Rev. E*, **83**, 061108 (2011).

[30] T. Sagawa, *Prog. Theor. Phys.*, **127**, 1 (2012).

[31] K. Jacobs, *Phys. Rev. E*, **86**, 040106(R) (2012).

[32] D. Abreu and U. Seifert, *Phys. Rev. Lett.*, **108**, 030601 (2012).

[33] T. Sagawa and M. Ueda, *Phys. Rev. Lett.*, **109**, 180602 (2012).

[34] W. H. Zurek, *Nature*, **341**, 119 (1989).

[35] J. Bub, *Stud. Hist. Phil. M. P.*, **32**, 569 (2001).

[36] K. Maruyama, F. Nori, and V. Vedral, *Rev. Mod. Phys.*, **81**, 1 (2009).

[37] A. Hosoya, K. Maruyama, and Y. Shikano, *Phys. Rev. E.*, **84**, 061117 (2011).

[38] H. T. Quan, Y. D. Wang, Y.-X. Liu, C. P. Sun, and F. Nori, *Phys. Rev. Lett.*, **97**, 180402 (2006).

[39] M. Bier and F. J. Cao, *Acta Phys. Pol.*, **43**, 889 (2012).

[40] D. Mandal and C. Jarzynski, *Proc. Natl. Acad. Sci. (USA)*, **109**, 11641 (2012).

[41] P. Strasberg, G. Schaller, T. Brandes, and M. Esposito, *Phys. Rev. Lett.*, **110**, 040601 (2013).

[42] J. M. Horowitz, T. Sagawa, and J. M. R. Parrondo, *Phys. Rev. Lett.* **111**, 010602 (2013).

[43] A. C. Barato and U. Seifert, *Europhys. Lett.* **101**, 60001 (2013).

[44] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 2007), Chap. 5, 3rd ed.

[45] Note the lack of a rate parameter analogous to $\gamma$ in Eq. 2. For the cooperative transition rates, we set this parameter to unity by appropriately choosing the unit of time.


## SUPPLEMENTAL MATERIAL

### Solving for $\Phi(\Lambda)$

Solving for $\Phi$ involves first solving for the periodic steady state of the demon, then using that solution to determine the distribution of the outgoing bits, from which $\Phi$ follows by Eq. 6 of the main text. We will use the notation $\mathbf{p}^D = (p_u, p_d)^T$ (where the superscript $T$ indicates transpose) to denote the statistical state of the demon, $\mathbf{p}^B = (p_0, p_1)^T$ for that of the interacting bit, and $\mathbf{p} = (p_{u0}, p_{d0}, p_{u1}, p_{d1})^T$ to denote their joint probability distribution.

Let $\mathcal{T}$ denote the $2 \times 2$ transition matrix whose element $T_{\mu\nu}$ ($\mu, \nu \in \{u, d\}$) gives the probability for the demon to be in state $\mu$ at the end of an interaction interval, given that it was in state $\nu$ at the start of the interval. As explained below, this matrix can be written as the product

$$\mathcal{T} = \mathcal{P}^D e^{\mathcal{R}\tau} \mathcal{M}, \tag{12}$$

where

$$\mathcal{P}^D = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \quad \mathcal{R} = \begin{pmatrix} \bullet & \gamma(1-\sigma) & 0 & 0 \\ \gamma(1+\sigma) & \bullet & 1+\omega & 0 \\ 0 & 1-\omega & \bullet & \gamma(1-\sigma) \\ 0 & 0 & \gamma(1+\sigma) & \bullet \end{pmatrix}, \quad \mathcal{M} = \begin{pmatrix} p_0 & 0 \\ 0 & p_0 \\ p_1 & 0 \\ 0 & p_1 \end{pmatrix}. \tag{13}$$

Here $\mathcal{R}$ is the transition rate matrix for the demon and the interacting bit. Its off-diagonal elements are given by Eqs. 2 and 3 of the main text, and its diagonal elements are determined by the requirement that the elements in each column sum to zero [1]. To understand Eq. 12, let $\mathbf{p}_0^D$ denote the distribution of the demon at the start of a given interaction interval. Then $\mathbf{p}_0 = \mathcal{M}\mathbf{p}_0^D$ gives the initial joint distribution of the demon and the incoming bit.

From this initial, uncorrelated state the joint distribution evolves under the master equation $\mathrm{d}\mathbf{p}/\mathrm{d}t = \mathcal{R}\mathbf{p}$, therefore $\mathbf{p}_\tau = \exp(\mathcal{R}\tau)\mathcal{M}\mathbf{p}_0^D$ gives the joint distribution at the end of the interaction interval. The matrix $\mathcal{P}^D$ then projects out the state of the bit, thus $\mathbf{p}_\tau^D = \mathcal{P}^D \exp(\mathcal{R}\tau)\mathcal{M}\mathbf{p}_0^D = \mathcal{T}\mathbf{p}_0^D$ gives the final marginal distribution of the demon.

The evolution of the demon over many intervals is given by repeated application of the matrix $\mathcal{T}$. Because $\mathcal{T}$ is a positive transition matrix [2], the demon evolves to a periodic steady state,

$$\lim_{n\to\infty} \mathcal{T}^n \mathbf{p}_0^D = \mathbf{p}_0^{D,ps} \quad , \tag{14}$$

defined uniquely by

$$\mathcal{T}\mathbf{p}_0^{D,ps} = \mathbf{p}_0^{D,ps} \quad . \tag{15}$$

$\mathbf{p}_0^{D,ps}$ gives the marginal distribution of the demon at the start of each interaction interval.

In the periodic steady state, the joint distribution of the demon and the interacting bit, at the end of the interaction interval, is given by $\mathbf{p}_\tau^{ps} = \exp(\mathcal{R}\tau)\mathcal{M}\mathbf{p}_0^{D,ps}$. The marginal distribution of the outgoing bit is then given by projecting out the state of the demon:

$$\mathbf{p}_\tau^{B,ps} = \mathcal{P}^B e^{\mathcal{R}\tau}\mathcal{M}\mathbf{p}_0^{D,ps} \quad , \quad \mathcal{P}^B \equiv \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}. \tag{16}$$

Therefore, to solve for $\Phi$, we first compute the elements of $\mathcal{T}$ using Eq. 12, then find its right eigenstate $\mathbf{p}_0^{D,ps}$ (Eq. 15), then determine $\mathbf{p}_\tau^{B,ps} = (p_0', p_1')^T$ using Eq. 16. $\Phi$ then follows directly from Eq. 6 in the main text: $\Phi = p_1' - p_1$.

We performed these calculations using Mathematica [3], and then simplified the results substantially by hand, finally obtaining

$$\Phi = \frac{\delta - \epsilon}{2}\,\eta(\Lambda) \quad , \quad \eta(\Lambda) = \frac{\nu_2 P + \nu_3 Q}{P + Q}, \tag{17a}$$

$$\begin{array}{lll} P = \mu_2\,(\mu_4\nu_3 + \mu_1\nu_1) & , & Q = \mu_3\,(\mu_4\nu_2 + \mu_1\nu_1), \\ \nu_1 = 1 - e^{-2\gamma\,\tau} & , & \mu_1 = (\delta + \sigma)\,\omega, \\ \nu_2 = 1 - e^{-(1+\gamma-\alpha)\,\tau} & , & \mu_2 = \alpha + \gamma + \sigma\,\omega, \\ \nu_3 = 1 - e^{-(1+\gamma+\alpha)\,\tau} & , & \mu_3 = \alpha - \gamma - \sigma\,\omega, \\ \alpha = \sqrt{1 + \gamma^2 + 2\gamma\sigma\omega} & , & \mu_4 = 1 - \delta\,\omega. \end{array} \tag{17b}$$

If the demon's intrinsic transitions occur rapidly in comparison with the cooperative transitions, $\gamma \to \infty$, then the analysis simplifies substantially: the demon remains in equilibrium with the hot reservoir at all times, and the interacting bit obeys the master equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} p_0^B \\ p_1^B \end{pmatrix} = \begin{pmatrix} -a & b \\ a & -b \end{pmatrix}\begin{pmatrix} p_0^B \\ p_1^B \end{pmatrix}, \tag{18}$$

with $a = (1-\omega)(1+\sigma)/2$ and $b = (1+\omega)(1-\sigma)/2$. Here $p_j^B(t)$ is the probability to find the bit in state $j \in \{0,1\}$ at time $t$ during the interaction interval. Integrating Eq. 18 over one interaction interval, $0 \le t \le \tau$, then setting $p_1 = p_1^B(0)$ and $p_1' = p_1^B(\tau)$ in Eq. 6 of the main text, we obtain

$$\Phi = \frac{\delta - \epsilon}{2}\,\left[1 - e^{-(1-\sigma\omega)\tau}\right]. \tag{19}$$

As a consistency check, we note that this result also follows from our general solution, Eq. 17, with the expression for $\eta(\Lambda)$ evaluated in the limit $\gamma \to \infty$.

Our general expression for $\eta(\delta, \sigma, \gamma, \omega, \tau)$, while exact, is sufficiently complex that we are unable to derive the inequality $\eta > 0$ (which was crucial in our interpretation of the phase diagram in the main text) directly from Eq. 17. Instead we will show in Appendix  that this inequality follows from the modified Clausius inequality, Eq. 11 of the main text, which we now derive.

**Modified Clausius inequality**

During any interaction interval, the joint distribution of the demon and the interacting bit evolves according to the master equation discussed above,

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = \mathcal{R}\mathbf{p}, \tag{20}$$

where $\mathcal{R}$ is given in Eq. 13. For very long interaction intervals ($\tau \to \infty$), the combined system relaxes to the stationary state

$$\overline{\mathbf{p}} = \frac{1}{\mathcal{N}}\left(1, \mu, \mu\nu, \mu^2\nu\right)^T, \quad \mu = \frac{1+\sigma}{1-\sigma}, \quad \nu = \frac{1-\omega}{1+\omega}, \quad \mathcal{N} = (1+\mu)(1+\mu\nu)\,, \tag{21}$$

which satisfies $\mathcal{R}\overline{\mathbf{p}} = \mathbf{0}$. Note that $\overline{\mathbf{p}}$ is actually a product of marginal distributions $\overline{\mathbf{p}}^D$ and $\overline{\mathbf{p}}^B$ for the demon and bit:

$$\overline{p}_{ij} = \overline{p}_i^D\, \overline{p}_j^B, \quad i \in \{u, d\}, \quad j \in \{0, 1\}, \tag{22a}$$

$$\overline{\mathbf{p}}^D = (1, \mu)^T/(1+\mu), \quad \overline{\mathbf{p}}^B = (1, \mu\nu)^T/(1+\mu\nu). \tag{22b}$$

The irreversible approach of $\mathbf{p}(t)$ toward $\overline{\mathbf{p}}$ is described by the *relative entropy* [4],

$$D(\mathbf{p}||\overline{\mathbf{p}}) = \sum_k p_k \ln \frac{p_k}{\overline{p}_k} \geq 0\,. \tag{23}$$

Here and in what follows, we use the index $k$ to indicate a joint state of the demon and the bit, $k \in \{0u, 0d, 1u, 1d\}$, reserving $i$ and $j$ for the demon and the bit, respectively, as in Eq. 22a. A standard calculation [1] shows that $D$ is a Lyapunov function, that is it satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} D(\mathbf{p}||\overline{\mathbf{p}}) \leq 0\,, \tag{24}$$

where the equality holds only when $\mathbf{p} = \overline{\mathbf{p}}$. Thus, as measured by relative entropy, any initial $\mathbf{p} \neq \overline{\mathbf{p}}$ evolves monotonically toward $\overline{\mathbf{p}}$, although for finite interaction intervals this relaxation is interrupted by the arrival of the next bit. We now use these properties to derive the inequality

$$Q_{c\to h}(\beta_h - \beta_c) + \Delta S_B \geq 0\,, \tag{25}$$

which appears as Eq. 11 of the main text.

Let $\mathbf{p}_0$ and $\mathbf{p}_\tau$ denote the joint distributions of the demon and a bit at the beginning and end of a given interaction interval, respectively, and similarly define $\mathbf{p}_0^D$, $\mathbf{p}_\tau^D$, $\mathbf{p}_0^B$ and $\mathbf{p}_\tau^B$ for the marginal distributions of the demon and the bit. Eq. 24 implies

$$D(\mathbf{p}_0||\overline{\mathbf{p}}) - D(\mathbf{p}_\tau||\overline{\mathbf{p}}) \geq 0. \tag{26}$$

Using Eqs. 23 and 22a we rewrite the left side of this equation as

$$S_\tau - S_0 - \sum_{i\in\{u,d\}} \left(p_{\tau,i}^D - p_{0,i}^D\right) \ln \overline{p}_i^D - \sum_{j\in\{0,1\}} \left(p_{\tau,j}^B - p_{0,j}^B\right) \ln \overline{p}_j^B, \tag{27}$$

where $S_0 = -\sum_k p_{0,k} \ln p_{0,k}$ and $S_\tau = -\sum_k p_{\tau,k} \ln p_{\tau,k}$ are the information entropies of the joint distributions of the demon and the bit at the beginning and end of the interaction interval. Let us now evaluate Eq. 27, assuming the demon has reached its periodic steady state.

The joint entropy $S$ can be written as [4]

$$S = S^D + S^B - I(D;B) \quad , \quad I(D;B) \geq 0, \tag{28}$$

where $S^D$ is the marginal entropy of the demon, $S^B$ is the marginal entropy of the bit, and the *mutual information* $I(D;B)$ quantifies the degree of correlation between them. By construction, the demon and bit are uncorrelated at

the start of the interaction interval, hence $I_0(D;B) = 0$. In the periodic steady state we have $S_\tau^D = S_0^D$, because the demon starts and ends in the same distribution. Hence the difference $S_\tau - S_0$ in Eq. 27 can be replaced by $\Delta S_B - I_\tau(D;B)$. We also have $\mathbf{p}_0^D = \mathbf{p}_\tau^D$ in the periodic steady state, so the first sum appearing in Eq. 27 vanishes.

Once the period steady state has been reached, the bit distributions $\mathbf{p}_0^B$ and $\mathbf{p}_\tau^B$ correspond to the statistics of the incoming and outgoing bit streams, defined in the main text:

$$p_{0,j}^B = p_j \quad , \quad p_{\tau,j}^B = p_j' \quad , \quad j \in \{0,1\}\,, \tag{29}$$

hence $p_{\tau,0}^B - p_{0,0}^B = -(p_{\tau,1}^B - p_{0,1}^B) = \Phi$, from the definition of $\Phi$. The last term in Eq. 27 can now be rewritten, using Eq. 21 and Eqs. 2 and 3 of the main text, as

$$-\sum_{j\in\{0,1\}} \left(p_{\tau,j}^B - p_{0,j}^B\right) \ln \overline{p}_j^B = \Phi \ln(\mu\nu) = Q_{c\to h}(\beta_h - \beta_c). \tag{30}$$

Collecting these results, we get

$$D(\mathbf{p}_0||\overline{\mathbf{p}}) - D(\mathbf{p}_\tau||\overline{\mathbf{p}}) = \Delta S_B - I_\tau(D;B) + Q_{c\to h}(\beta_h - \beta_c), \tag{31}$$

which then combines with Eq. 26 to give us

$$Q_{c\to h}(\beta_h - \beta_c) + \Delta S_B \geq I_\tau(D;B) \geq 0. \tag{32}$$

An alternative derivation of this result can be constructed using the integral fluctuation theorem for total entropy production [5].

The first inequality in Eq. 32 is stronger than the modified Clausius statement, Eq. 25. This underscores the fact that Eq. 25 is a weak statement of the second law of thermodynamics (as it applies to our model), since it neglects correlations in the outgoing bits: the quantity $\Delta S_B$ is defined in terms of the marginal distribution of each bit. In reality the bits do develop correlations via their interactions with the demon, as the state of the demon at the end of one interaction interval is also its initial state at the beginning of the next interval. (Explicit numerical simulations indicate that these correlations are small, but not zero.) If these correlations were to be taken into account, then the net change in the Shannon entropy per bit would have a value slightly lower than $\Delta S_B$, and Eq. 25 would be replaced by a somewhat stronger bound. These considerations are reflected, somewhat indirectly, by the term $I_\tau(D;B)$ in Eq. 32.

Finally, note that

$$\frac{\overline{p}_1^B}{\overline{p}_0^B} = \mu\nu = \frac{1-\epsilon}{1+\epsilon} \quad , \quad \frac{p_1}{p_0} = \frac{1-\delta}{1+\delta}\,, \tag{33}$$

using Eqs. 21 and 22b, and the definitions of $\epsilon$ and $\sigma$. Thus, when $\delta = \epsilon$, the incoming bits arrive in the stationary distribution $\overline{\mathbf{p}}$. In this case, no relaxation occurs during the interaction interval; the equality holds in Eqs. 24 and 26; the outgoing bits depart with the same distribution; and $\Phi = 0$. When $\delta \neq \epsilon$, Eqs. 24 and 26 are both strict inequalities, and therefore so is the modified Clausius inequality (Eq. 25 / Eq. 11).

### Positivity of $\eta(\Lambda)$

To investigate the sign of $\eta$, let us take $\delta \neq \epsilon$ [6] and rewrite Eq. 25 in the form

$$f(\delta') > f(\delta)\,, \tag{34}$$

where

$$f(\delta) = K\delta + S(\delta) \quad , \quad K = \frac{1}{2}(\beta_c - \beta_h)\Delta E > 0. \tag{35}$$

Eq. 34 follows by the direct substitution of the relations

$$Q_{c\to h} = \Phi\Delta E \quad , \quad \Phi = \frac{\delta - \delta'}{2} \quad , \quad \Delta S_B = S(\delta') - S(\delta) \tag{36}$$

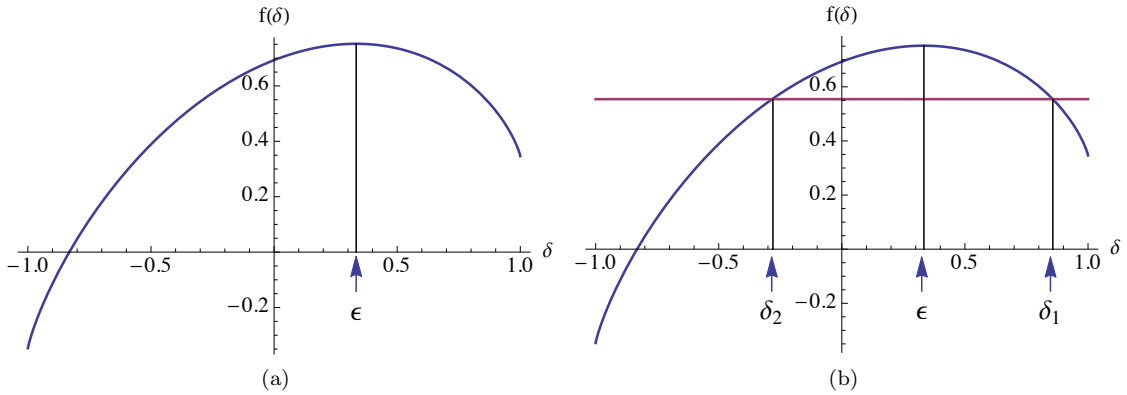


FIG. 3: (a) The concave function $f(\delta)$ has a maximum at $\delta = \epsilon$, as illustrated for $\epsilon = 1/3$. (b) For a given $\delta_1$, we must have $\delta_2 < \delta_1' < \delta_1$ to ensure $f(\delta_1') > f(\delta_1)$. Hence, both $\delta_1'$ and $\epsilon$ lie to the left of $\delta_1$.

into Eq. 25, using a strict inequality since $\delta \neq \epsilon$.

By construction, $\mathrm{d}^2 f/\mathrm{d}\delta^2 < 0$. Setting $\mathrm{d}f/\mathrm{d}\delta = 0$, the unique maximum of $f(\delta)$ is easily shown to occur at $\delta = \epsilon$, as illustrated in Fig. 3(a) for $\epsilon = 1/3$. Now let $\delta_1$ and $\delta_2$ denote two values of $\delta$ that correspond to the same value of $f$, with $\delta_2 < \epsilon < \delta_1$, as shown in Fig. 3(b). Let $\delta_1'$ describe the surplus of 0's in the outgoing bit stream, when the incoming stream is characterized by $\delta_1$. Because the maximum of $f(\delta)$ occurs at $\delta = \epsilon$, Eq. 34 implies that $\delta_2 < \delta_1' < \delta_1$; see Fig. 3(b). If we instead consider incoming and outgoing bit streams described by $\delta_2$ and $\delta_2'$, then the same argument gives us $\delta_2 < \delta_2' < \delta_1$. We therefore conclude that the incoming and outgoing bit streams necessarily satisfy

$$\mathrm{sign}(\delta - \delta') = \mathrm{sign}(\delta - \epsilon)\,, \tag{37}$$

in other words $\delta'$ lies on the same side as $\epsilon$ with respect to $\delta$. Since

$$\frac{\delta - \delta'}{2} = \Phi = \frac{\delta - \epsilon}{2}\eta(\Lambda)\,, \tag{38}$$

we must have $\eta(\Lambda) > 0$.

[1] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 2007), Chap. V, 3rd ed.
[2] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra* (SIAM, Philadelphia, PA, 2000), Chap. 8.
[3] I. Wolfram Research, *Mathematica* (Wolfram Research, Inc., Champaign, Illinois, 2010), 8th ed.
[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, Hoboken, New Jersey, 2006).
[5] U. Seifert, *Phys. Rev. Lett.* **95**, 040602 (2005).
[6] When $\delta = \epsilon$, the value of $\eta$ is inconsequential, by Eq. 17a.