A Maximum Work Theorem for Maxwell's Demons

Henrik Sandberg,[1] Jean-Charles Delvenne,[2] Nigel J. Newton,[3] and Sanjoy K. Mitter[4]

[1] *Automatic Control Lab, KTH Royal Institute of Technology, Stockholm, Sweden*
[2] *ICTEAM and CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium*
[3] *School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK*
[4] *Laboratory for Information and Decision Systems, MIT, Cambridge, Massachusetts, USA*

(Dated: February 6, 2014)

We determine the maximum amount of work extractable in finite time by a demon performing continuous measurements on a quadratic Hamiltonian system subjected to thermal fluctuations, in terms of the information extracted from the system. The optimal demon is found to apply a high-gain continuous feedback from a Kalman-Bucy estimate of the system state. A simple and concrete electrical implementation of the maximum-work protocol is proposed, which allows for analytic expressions of the flows of energy and entropy inside the demon. In particular, we observe that the maximum-power demon dissipates twice as much energy as predicted by Landauer's principle.

PACS numbers: 5.70.Ln, 05.40.-a, 89.70.Cf

Ever since Maxwell [1] put forward the idea of an abstract being (a demon) apparently able to break the second law of thermodynamics, it has served as a great source of inspiration and helped to establish important connections between statistical physics and information theory. See, for example, [2–6]. In the original version, the demon operates a trapdoor between two heat baths, such that a seemingly counterintuitive heat flow is established. Today, more generally, devices that are able to extract work from a single heat bath by rectifying thermal fluctuations are also called Maxwell's demons [7]. Several schemes detailing how the demon could apparently break the second law have been proposed, for example Szilard's heat engine [2]. More recent schemes are presented in [7–10], and [11, 12] where measurement errors are also accounted for.

A classic expression of the second law states that the maximum (average) work extractable from a system during a thermodynamic transformation, $W_{\max}$, cannot exceed the free energy difference, $-\Delta F$, between the system's initial and final equilibrium states of the same temperature T. However, as illustrated by Szilard's heat engine, it is possible to break this bound under the assumption of additional information available to the work-extracting agent. To account for this possibility, the second law under transformations with *discrete feedback* [8, 13–16] takes the form,

$$W_{\max} = kT\mathcal{I} - \Delta F, \tag{1}$$

where k is Boltzmann's constant. The quantity $\mathcal{I}$ is the mutual information [17] between the physical state of the system and measurements made available to the external agent during the work-extraction process. Related generalizations of the second law are stated in [18–20]. It is possible to construct feedback protocols that extract the maximum amount $W_{\max}$ using reversible and quasistatic transformations [16, 21, 22]. Reversible feedback protocols may be optimal in terms of entropy production, but they are also infinitely slow. In [15, 23, 24], the corresponding finite-time problem was addressed.

The main contribution of this letter is to state a *continuous feedback* counterpart to Eq. (1), applicable also for *finite-time* transformations with measurement errors. Assume that the system is in thermal equilibrium at temperature T initially, and we let it relax back to the same equilibrium after the feedback is turned off so that $\Delta F = 0$. We show that the maximum extractable work over a duration t, $W_{\max}(t)$, satisfies the bounds

$$k\int_0^t T_{\min}\dot{\mathcal{I}}\,dt' \le W_{\max}(t) \le kT\mathcal{I}(t), \tag{2}$$

where $T_{\min}(t)$ is the lowest achievable system temperature after t time units of continuous feedback control ($T_{\min}(0) = T$). Therefore, every bit of information, if optimally exploited, allows us to retrieve between $kT_{\min}\ln 2$ and $kT\ln 2$ units of work.

This result holds under the assumption of linear dynamics, e.g. systems with quadratic Hamiltonians in contact with a heat bath. A quadratic Hamiltonian is a common and reasonable assumption for a system excited by thermal fluctuations of moderate temperature around a minimum-energy state. Under continuous work extraction from an overdamped Langevin equation in the time interval $[0, t]$, we show that $W_{\max}(t)$ is equal to the lower bound. In general, $W_{\max}(t)$ asymptotically approaches the lower bound under continuous work extraction, and is equal to it in the non-equilibrium steady state (NESS). The upper bound is reached under short intermittent bursts of control, where the total duration of control t is split into many short, far apart, time intervals, and more generally when the demon operates reversibly.

The mutual information $\mathcal{I}(t)$ under continuous-time feedback is far from trivial to compute [25–27], but has a closed-form solution in our setup. To the authors' best knowledge, most previous studies of feedback in stochastic thermodynamics are discrete in time and/or in state space. As a result, our upper bound is not directly comparable to similar-looking inequalities [8, 13–16, 18–20]. In contrast, traditional controller design in engineering science is often done in a continuous (analog)

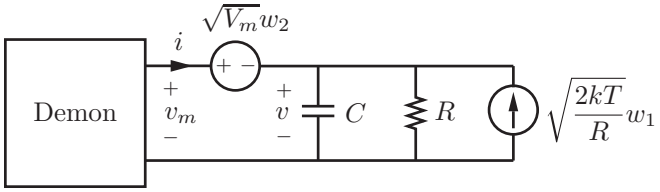

FIG. 1. The demon (the feedback controller) connected to a capacitor, a heat bath of temperature T, and a measurement noise source of intensity V_m. The demon may choose the current i freely, and has access to the noisy voltage measurement v_m.

setting [28, 29]. The papers [30–32] do employ continuous feedback, but exact system state knowledge by the controller is assumed, which significantly simplifies the studied optimal-control problems. The paper [33] does study continuous feedback with measurement errors, but does not characterize information flows and work bounds of the type Eq. (2). The continuous setting allows us to physically implement an optimal demon using easy-to-analyse electric components, rather than abstract or complex digital computing devices, as illustrated at the end of this letter.

System model.— The system we first consider is an electric capacitor C, a resistor R with thermal noise (the heat bath), and a feedback controller (the demon) with access to noisy voltage measurements, see Fig. 1. The resistor is subjected to Johnson-Nyquist noise [34, 35]. The circuit is modeled by an overdamped Langevin equation

$$\begin{aligned} \tau\dot{v} &= -v + Ri + \sqrt{2kTR}w_1, \qquad \langle v(0)\rangle = 0, \\ v_m &= v + \sqrt{V_m}w_2, \qquad\qquad \langle v(0)^2\rangle = \frac{kT}{C}, \end{aligned} \tag{3}$$

with $v(0)$ Gaussian, $w_{1,2}$ uncorrelated Gaussian white noise ($\langle w_{1,2}(t)w_{1,2}(t')\rangle = \delta(t-t')$), V_m the intensity of the measurement noise, and $\tau = RC$ being the time constant of the open circuit. The measurement noise $\sqrt{V_m}w_2$ can be thought of as the Johnson-Nyquist noise of the wire between the capacitor and the demon, whose resistance for simplicity is incorporated in the demon. The heat flow to the capacitor is $\dot{Q}$ and the work-extraction rate of the demon is $\dot{W}$, and satisfy [32]

$$\begin{aligned} \dot{U} &= \dot{Q} - \dot{W}, \qquad & U &= \frac{1}{2}C\langle v^2\rangle = \frac{1}{2}kT_C, \\ \dot{Q} &= \frac{k}{\tau}(T - T_C), \qquad & \dot{W} &= -\langle vi\rangle, \end{aligned} \tag{4}$$

where we denote the *effective* instantaneous temperature of the capacitor by T_C, and its internal energy by U. Furthermore, we assume the capacitor initially is in thermal equilibrium with the heat bath, i.e., $T_C(0) = T$. Just as in [15], we can justify calling $T_C(t)$ a temperature since it appears in a Fourier-like heat conduction law (see $\dot{Q}$). Also, since our applied controls will maintain a Gaussian distribution of v, $T_C(t)$ will be the true temperature of the capacitor if it were to be disconnected from all the other elements at time t. The voltage v_m is the measurement that supplies the demon with information, and can be seen as a noisy measurement of the fluctuating capacitor voltage v. We will show how a demon can optimally control the work extraction by carefully exploiting the measurements v_m and properly choosing the injected current i. Intuitively, the demon can create a positive work rate $\dot{W}$ if it chooses $i < 0$ when it correctly estimates $v > 0$, and vice versa. But how the demon should estimate v, and how to optimally choose i may be less obvious.

For any work-extracting demon it holds that $0 \leq W(t) \leq W_{\max}(t)$, where from Eq. (4) it holds

$$W_{\max}(t) := \int_0^t \frac{k}{\tau}(T - T_{\min})\,dt' + \frac{1}{2}k(T - T_{\min}(t)), \tag{5}$$

and $T_{\min}(t') \leq T_C(t')$ for $0 \leq t' \leq t$ is the lowest achievable effective temperature.

Demon model and optimal continuous feedback.— The *separation principle* [36, 37] implies that $T_{\min}$ is achievable by a demon that optimally and continuously estimates the state $v(t)$ of the capacitance, given the measurements $\{v_m(t'), \quad 0 \leq t' \leq t\}$, and then injects a current $i(t)$ based on this sole estimate. The best possible estimate $\hat{v}(t)$ of $v(t)$, given the measurement trajectory $\{v_m(t'), \quad 0 \leq t' \leq t\}$, can be recursively constructed by the celebrated *Kalman-Bucy filter* [38], which leads to a minimum variance estimation error [37] and exploits as much of the information contained in v_m as is possible [25]. The Kalman-Bucy filter for Eq. (3) is given in Appendix A. Let us assume the demon continuously chooses to inject the current

$$i(t') = -G\hat{v}(t'), \quad 0 \leq t' \leq t, \tag{6}$$

where $0 \leq G < \infty$ is a fixed scalar feedback gain. We may think of the feedback gain G as the 'conductance' of the demon: If the demon believes the voltage of the capacitor to be $\hat{v}$, it will admit the current $G\hat{v}$. If $v \approx \hat{v}$, the demon will indeed look like an electric load of conductance close to G. While $G = 0$ (open circuit) creates a demon that only (optimally) observes, $G \to \infty$ also removes energy from the capacitance at the highest possible rate, achieving the minimum effective temperature $T_{\min}$ for the capacitance, see Appendix A. Furthermore, $T_{\min}$ solves the filter Riccati equation

$$\tau\dot{T}_{\min} = 2(T - T_{\min}) - \frac{T_{\min}^2}{2T\mu}, \quad T_{\min}(0) = T, \tag{7}$$

where $\mu = V_m/(2kTR)$ is a fundamental adimensional characterization of measurement noise, compared to the bath noise. Starting from T, $T_{\min}$ decreases exponentially, and monotonically, to a steady-state value

$$T_{\min}^{\text{NESS}} = 2\mu\left(\sqrt{1 + 1/\mu} - 1\right)T < T. \tag{8}$$

In the noisy measurement limit $\mu \gg 1$, the NESS may reach the effective temperature $(1 - 1/(4\mu))T$, slightly colder than T, while accurate measurements $\mu \ll 1$ allows us to reach a low effective temperature $2\sqrt{\mu}T$.

For a general feedback gain $G \geq 0$ in Eq. (6), the effective temperature of the capacitor will drop exponentially from $T_C(0) = T$ to

$$T_C^{\rm NESS} = \frac{1}{1+GR}T + \frac{GR}{1+GR}T_{\rm min}^{\rm NESS}. \tag{9}$$

The corresponding NESS work-extraction rate can be shown to become

$$\dot{W}^{\rm NESS} = \frac{k}{\tau}(T - T_{\rm min}^{\rm NESS})\frac{GR}{1+GR}. \tag{10}$$

Thus the continuous feedback protocol in Eq. (6) can realize any NESS work rate between 0 and the maximum $\dot{W}_{\rm max}^{\rm NESS} = \frac{k}{\tau}(T - T_{\rm min}^{\rm NESS})$ by proper choice of gain G.

The above optimal controller can be generalized to any system with linear dynamics. Details are given in Appendix B for systems with quadratic Hamiltonians.

Information flow and maximum work theorem.— To establish the maximum work theorem in Eq. (2), we need to quantify the information flow from the voltage v to the measurement v_m, under continuous feedback. For this purpose, let us make a linear decomposition of the dynamics in Eq. (3),

$$v(t) = v_0(t) + v_i(t), \tag{11}$$

such that

$$\begin{aligned} \tau\dot{v}_0 &= -v_0 + \sqrt{2kTR}w_1, \quad & v_0(0) &= v(0) \\ \tau\dot{v}_i &= -v_i + Ri, & v_i(0) &= 0. \end{aligned} \tag{12}$$

The motivation behind this decomposition is that the signal v_i is completely known to the demon, since it controls i, whereas v_0 contains the uncertainty due to the interaction with the thermal bath. Note that if the demon chooses to apply no current, $G = 0$, then $v(t) = v_0(t)$ for all t. A similar decomposition was used in [14, 36]. The information about the uncertain trajectory v_0 contained in v_m is quantified by the following mutual information $\mathcal{I}(t)$ [17, 25],

$$\mathcal{I}(t) = I(v_0(t'),\, t' \in [0,t];\, v_m(t'),\, t' \in [0,t]). \tag{13}$$

Assuming the applied current i is *causally* and *deterministically* dependent on v_m, of which Eq. (6) is an example, we may apply [39, Theorem 16.3] (see also [25, Lemma 3.1]) to the model in Eq. (3) and obtain the closed-form expression

$$\mathcal{I}(t) = \frac{1}{4\mu\tau}\int_0^t \frac{T_{\rm min}}{T}\,dt'. \tag{14}$$

Note that $\mathcal{I}$ *does not* otherwise depend on the details of the demon.

It now follows from Eqs. (5), (7), and (14) that the maximum extracted work must satisfy

$$\begin{aligned} W_{\rm max}(t) &= \int_0^t \frac{k}{\tau}(T - T_{\rm min})\,dt' + \frac{1}{2}k(T - T_{\rm min}(t)) \\ &= \int_0^t \frac{kT_{\rm min}^2}{4\mu\tau T}\,dt' = k\int_0^t T_{\rm min}\dot{\mathcal{I}}\,dt', \end{aligned} \tag{15}$$

which proves the lower bound in Eq. (2) *with equality*. As is shown in Appendix B, for multi-dimensional systems we only obtain an inequality. However, as soon as the system satisfies an equipartition condition (true at all times in the one-dimensional case, and in any dimension when a NESS is reached) the inequality turns into an equality. The upper bound in Eq. (2) follows easily since $T_{\rm min} \leq T$ for all t.

The expressions for $\mathcal{I}$ and $W_{\rm max}$ provide interesting insights concerning information and work flow in the feedback loop. Since $T_{\rm min}$ decreases monotonically, the information rate $\dot{\mathcal{I}}$ is largest just when the measurement and feedback control start, and then decreases until it stabilizes at $\dot{\mathcal{I}}^{\rm NESS} = T_{\rm min}^{\rm NESS}/(4\mu\tau T)$. In NESS, the fresh measurements are no longer able to improve the quality of the estimate, i.e. to decrease the error variance $\langle(v(t) - \hat{v}(t))^2\rangle$ any further. Since $\dot{W}_{\rm max} = kT_{\rm min}\dot{\mathcal{I}}$, the work-extraction rate also decreases until it stabilizes at $\dot{W}_{\rm max}^{\rm NESS}$.

As both the work extraction and information rate are highest when the system is in equilibrium at temperature T, it may be tempting to run the feedback controller only when the system is close to equilibrium. Of course, as soon as the optimal feedback loop is closed, the effective temperature drops along the trajectory $T_{\rm min}$. But if the optimal feedback control is only applied for a very short time, say of duration $\delta t \to 0$, it holds

$$W_{\rm max}(\delta t) \approx kT\mathcal{I}(\delta t) \approx \frac{kT}{4\mu\tau}\delta t, \tag{16}$$

since $T_{\rm min}(0) = T$. The work $\frac{kT}{4\mu\tau}\delta t$ saturates the upper bound in Eq. (2), and has the largest possible efficiency in terms of work per bit of information [24]. On the other hand, the work is also very small since δt is small. Nevertheless, if the system is allowed to relax back to thermal equilibrium again before the next feedback burst, it is possible to operate this feedback controller at the same efficiency as optimal feedback reversible discrete controllers [2, 16, 21], which saturate Eq. (1). It is interesting to compare the optimal controllers. The continuous intermittent controller mimics the discrete ones in that it tries to make a very short (sampled) measurement. It immediately acts on the obtained information and extracts the work it can. If the continuous controller waited with the control application, the value of its information would have decayed due to the constant thermal fluctuations. This result is in contrast to Szilard's engine [2], for example, which extracts work infinitely slowly after the measurement is taken. In the end, it takes all of

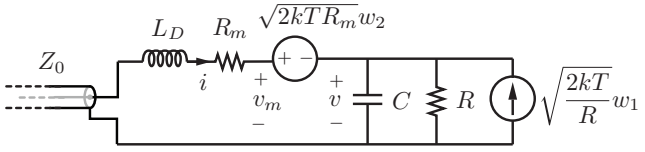

FIG. 2. An exact electric implementation of the demon in NESS. The current i through the inductive element is proportional to the optimal (Kalman-Bucy) estimate $\hat{v}$. The lossless transmission line both serves to cool the circuit and to store extracted work. The wire resistance R_m creates measurement noise and losses.

these maximum-efficiency controllers an infinite amount of time to deliver a fixed positive amount of work.

Physical implementation of the demon.— The demon may be seen as an analog or high-precision, high-frequency digital computer which finds the best estimate $\hat{v}$ of the state, commanding a fast actuator that generates a current $-G\hat{v}$. In the spirit of [7, 10, 11, 22], we find an explicit physical device implementing the demon. To simplify the presentation, we only consider the NESS with arbitrary temperature $T_C^{\rm NESS}$, where it turns out that the demon can be realized by passive circuit elements, i.e., elements with no internal energy generation. We assume the measurement noise arises from a resistance R_m in the wire connecting the system and the demon, in thermal equilibrium with the bath. Thus $V_m = 2kTR_m$ and $\mu = R_m/R$. Only a resistive and an inductive element are now needed to exactly implement the optimal feedback control, as illustrated in Fig. 2. The inductance L_D and the purely resistive characteristic impedance Z_0 of a semi-infinite lossless transmission line should be chosen as

$$L_D = \frac{CR_{\rm MP}}{G}, \quad Z_0 = \frac{1}{G}\left(\frac{R_{\rm MP}}{R}+1\right) + R_{\rm MP} - R_m, \tag{17}$$

where $R_{\rm MP} = V_m/(kT_{\rm min}^{\rm NESS}) = 2R_mT/T_{\rm min}^{\rm NESS}$ is the total resistance in the demon and wire when $G \to \infty$. One may interpret the inductor as both the computer processor and actuator, since the best instantaneous estimate $\hat{v}(t)$ of the capacitor voltage is proportional to the current $i(t) = -G\hat{v}(t)$. The line Z_0 may be interpreted as an ideal infinite memory tape and energy storage in which used information and some of the extracted work is disposed. It can be implemented as an actual close-to-ideal transmission line cooled down to a low temperature, or perhaps more plausibly, as a nonlinear active memory element of equivalent behavior. The lossless transmission line, an electric counterpart of an elastic semi-infinite string, is also electrically equivalent to a resistor of resistance Z_0 [40, 41]. Another implementation, close to Landauer's assumption of a finite memory where information is erased at the rate it is recorded, is a finite line terminated by a resistance $R_{\rm era}$ matched to the line, $R_{\rm era} = Z_0$. In all cases, the line in series with the measurement resistance R_m and thermal noise $\sqrt{2kTR_m}w_2$ together act as a resistance $Z_0 + R_m$ of effective temperature

$$T_D = \frac{TR_m}{R_m + Z_0} = \frac{T_{\rm min}^{\rm NESS}}{2}\frac{GR'}{1+GR'} < T, \tag{18}$$

where $1/R' = 1/R + 1/R_{\rm MP}$. Hence, the line, in whatever implementation, can be interpreted as cooling the circuit to create a thermal gradient, which drives the heat flow from the bath. As $G \to 0$, T_D tends to zero, but the demon impedance tends to infinity to prevent a heat flow. Note that the demon still optimally estimates the voltage v, though. As G increases, T_D increases and at maximum-power extraction, T_D is half the effective capacitor temperature, i.e.,

$$T_D = T_{\rm min}^{\rm NESS}/2. \tag{19}$$

This simple formula is reminiscent of classical maximum-power theorems found in [42, 43]. The maximum-power demon is therefore electrically equivalent to just a line of characteristic impedance $Z_0 = R_{\rm MP} - R_m$ in series with the resistance R_m, and vanishing inductance, $L_D = 0$.

The maximum-power demon extracts work at rate $\dot{W}_{\rm max}^{\rm NESS} = kT_{\rm min}^{\rm NESS}\dot{\mathcal{I}}^{\rm NESS}$ from the capacitance, and ultimately from a single hot temperature T. If the line, and its erasing resistance, are initially in equilibrium with the bath of temperature T, we can still exactly implement the demon by cooling down the demon and the wire to the temperature T_D. Now the power $\dot{W}_{\rm max}^{\rm NESS}$ is ultimately dissipated as heat in the wire and the eraser. The total entropy rate of this setup, $\dot{S}$, can be decomposed as $\dot{S}_{CR} + \dot{S}_D$, where $\dot{S}_{CR} = \dot{W}_{\rm max}^{\rm NESS}(1/T_{\rm min}^{\rm NESS} - 1/T)$ is the entropy generation in the capacitor-bath system, while $\dot{S}_D$ is the entropy generation caused by the demon (and wire). If the demon and wire are reversibly cooled by a Carnot-optimal refrigerator from T to T_D, then the total entropy rate is $\dot{S} = \dot{W}_{\rm max}^{\rm NESS}(1/T_D - 1/T)$, which after a calculation implies

$$\dot{S}_D = k\dot{\mathcal{I}}^{\rm NESS}, \tag{20}$$

an observation apparently compatible with Landauer's principle [3]: every bit stored into the NESS demon overwrites a previous bit and generates a corresponding physical entropy. This implementation requires an external power driving a Carnot refrigerator, $\dot{W}_{\rm ref} = \dot{W}_{\rm max}^{\rm NESS}(T/T_D - 1)$. The total power $\dot{W}_{\rm ref} + \dot{W}_{\rm max}^{\rm NESS}$ supplied to the demon, and dissipated to the bath is therefore at least $2kT\dot{\mathcal{I}}^{\rm NESS}$, twice the amount predicted by Landauer's principle. This is perhaps not surprising as Landauer's principle is supposed to hold with equality only for quasi-static equilibrium processes. We must expect that a maximum-power demon dissipates more work than demons operating at equilibrium. Remarkably, our calculation holds regardless of μ. We conjecture that our implementation in fact is optimal, in the sense that every implementation of NESS memory erasure at maximum power from a linear system should dissipate at least the work $2kT\ln 2$, twice the Landauer bound, for every new bit of information recorded.

Acknowledgements.— The authors would like to thank Jordan Horowitz for helpful discussions and suggestions. H.S. is supported by the Swedish Research Council under grants 2009-4565 and 2013-5523. J.-C. D. is supported by the Interuniversity Attraction Pole 'Dynamical Systems, Control and Optimization (DYSCO)', initiated by the Belgian State, Prime Minister's Office.

Appendix A: Kalman-Bucy Filter for the Overdamped Langevin Equation

As mentioned in the letter, the maximum-work extraction problem for the overdamped Langevin equation, here implemented by a capacitor connected to a resistance at temperature T, is solved by application of the separation principle [36, 37]. Therefore we must first estimate the state v of the capacitance as well as possible given the measurements. This is achieved by the Kalman-Bucy filter for the system in Eq. (3), leading to an estimate $\hat{v}$ that solves

$$\tau\frac{d}{dt}\hat{v} = -\hat{v} + Ri + \frac{T_{\min}}{2\mu T}\left(v_m - \hat{v}\right), \quad \hat{v}(0) = 0, \tag{A1}$$

where $T_{\min}$ solves the filter Riccati equation over the time interval $[0, t]$,

$$\tau\dot{T}_{\min} = 2(T - T_{\min}) - \frac{T_{\min}^2}{2\mu T}, \quad T_{\min}(0) = T, \tag{A2}$$

where $\mu = V_m/(2kTR)$ describes the measurement noise level and $\tau = RC$ is the time constant of the system, as in the letter. The initial conditions $\hat{v}(0) = 0$ and $T_{\min}(0) = T$ reflect the fact that the best unbiased estimate initially is zero, and that the demon knows the temperature of the bath.

Note that the Kalman-Bucy filter can be implemented online in a feedback controller, since it causally depends on the measurement realization v_m, and $T_{\min}$ can be solved for offline. As we will show, $T_{\min}$ is indeed the lowest effective temperature T_C attainable. The filter is variance-optimal, i.e., the variance of the estimation error $\langle\Delta v^2\rangle := \langle(v - \hat{v})^2\rangle = kT_{\min}/C$ is the smallest possible [37]. Furthermore, Eq. (A2) has the closed-form solution

$$T_{\min}(t) = T_{\min}^{\mathrm{NESS}} + \frac{(T - T_{\min}^{\mathrm{NESS}})e^{-2\gamma\tau t}}{1 + (T - T_{\min}^{\mathrm{NESS}})(1 - e^{-2\gamma\tau t})/(4\gamma\mu T)} \tag{A3}$$

where $\gamma = \sqrt{1 + 1/\mu} > 1$ and

$$T_{\min}^{\mathrm{NESS}} = 2\mu\left(\sqrt{1 + 1/\mu} - 1\right)T < T, \tag{A4}$$

is the NESS solution, $T_{\min} \to T_{\min}^{\mathrm{NESS}}$, as $t \to \infty$. From the solution it is seen the steady state is approached monotonically and exponentially fast.

Having computed the optimal estimate $\hat{v}$, let us assume the demon continuously chooses to inject the current

$$i(t') = -G\hat{v}(t'), \quad 0 \le t' \le t, \tag{A5}$$

where $0 \le G < \infty$ is a fixed scalar feedback gain, or 'conductance' of the demon. Inserting Eq. (A5) in Eq. (A1) we can compute the evolution of the variance $\hat{V} = \langle\hat{v}^2\rangle$ of the filter estimate as

$$\tau\frac{d}{dt}\hat{V} = -2\left(1 + GR\right)\hat{V} + \frac{kT_{\min}^2}{2\mu CT}, \quad \hat{V}(0) = 0. \tag{A6}$$

We note that since $T_{\min}$ is bounded, $\hat{V}$ can be made arbitrarily close to zero by increasing the feedback gain G. The estimation error Δv of the Kalman-Bucy filter is orthogonal to the estimate [37], $\langle\hat{v}\Delta v\rangle = 0$, and therefore

$$\frac{kT_C}{C} = \langle v^2\rangle = \langle\hat{v}^2\rangle + \langle\Delta v^2\rangle = \hat{V} + \frac{kT_{\min}}{C}, \tag{A7}$$

where $\langle\Delta v^2\rangle = kT_{\min}/C$ follows from Eqs. (A1)–(A2). Since $T_{\min}$ is independent of G, and $\hat{V}$ can be made arbitrarily close to zero, we realize that the demon through its policy is cooling the capacitor and for all t,

$$T_C(t) \searrow T_{\min}(t) \quad \text{as} \quad G \to \infty. \tag{A8}$$

This shows a demon should implement a Kalman-Bucy filter with a large (infinite) feedback gain G to extract the work $W_{\max}$.

Appendix B: The Hamiltonian Case in Higher Dimension

Let us consider the more general case where the capacitor is replaced by a Hamiltonian system. We assume a quadratic Hamiltonian, $H(x) = \frac{1}{2}x^T Kx$, where $x^T = [q^T \quad p^T] \in \mathbb{R}^{2n}$ is a point in the phase space with generalized positions q and momenta p, and $K \in \mathbb{R}^{2n\times 2n}$ is a symmetric positive-definite matrix. Hamilton's equations under the influence of a generalized external force $B_u u(t)$ (the constant matrix $B_u \in \mathbb{R}^{2n}$ determines which coordinates are directly affected), applied by the demon, now reads

$$\begin{aligned} \dot{x} &= J\nabla H(x) + B_u u \\ y &= B_u^T \nabla H(x), \end{aligned} \tag{B1}$$

where $J = -J^T = \left[\begin{smallmatrix} 0 & I_n \\ -I_n & 0 \end{smallmatrix}\right]$, and y is the generalized velocity conjugate to u. That is, $\dot{H}(t) = y(t)u(t)$ is the rate of work applied to the system. Now, $\nabla H(x) = Kx$, and the Hamiltonian system is a *linear dynamical system.*

We connect the Hamiltonian system to a heat bath of temperature T and with viscous friction coefficient $r > 0$ producing a dissipative force in the direction $B \in \mathbb{R}^{2n}$.

We obtain [32]

$$\begin{aligned} \dot{x} &= (J-D)Kx + B_u u + B\sqrt{2kTr}w_1, \\ \langle x(0)\rangle &= 0, \quad \langle x(0)x(0)^T\rangle = kTK^{-1}, \\ y &= B_u^T K x, \\ y_m &= B^T K x + \sqrt{V_m}w_2, \end{aligned} \tag{B2}$$

where $x(0)$ is Gaussian, $w_{1,2}$ uncorrelated Gaussian white noise, $D = rBB^T$ is the dissipation and $B\sqrt{2kTr}w_1$ models the corresponding thermal fluctuation. We have also assumed a scalar noisy measurement y_m of the generalized velocity conjugate to the dissipative force [44], which is available to the demon. In the following, it is assumed the system in Eq. (B2) is *controllable* and *observable* [29]. That is, in the absence of noise ($w_1 = w_2 = 0$), it is possible to force the system to $x = 0$ in arbitrarily short time from any initial state using some force u, and it is possible to determine the state $x(t)$ exactly given an arbitrarily short measurement trajectory $\{y_m\}_{t-\epsilon}^{t+\epsilon}$, $\epsilon > 0$. If these assumptions do not hold, it means that there are system coordinates that are either invisible to, or beyond the influence of, the demon. Such degrees of freedom can systematically be eliminated to create a *minimal model*, see, for example, [29].

Let us denote the second moment of the phase space coordinate by $X(t) := \langle x(t)x(t)^T\rangle \in \mathbb{R}^{2n\times 2n}$. Then the internal energy can be written as $U(t) = \langle H(t)\rangle = \frac{1}{2}\mathrm{Tr}(KX(t))$. The first law of thermodynamics reads [32]

$$\begin{aligned} \dot{U} &= \dot{Q} - \dot{W} \\ \dot{Q} &= kT\mathrm{Tr}(KD) - \mathrm{Tr}(KDKX) \\ \dot{W} &= -\langle uy\rangle, \end{aligned} \tag{B3}$$

where $\dot{Q}$ is the expected energy exchange rate with the heat bath, and $\dot{W}$ is the expected work extraction rate. We note that in thermal equilibrium ($\dot{Q} = \dot{W} = 0$) we have $X = kTK^{-1}$, and the internal energy is $U = nkT$, in accordance with the equipartition theorem. We say the internal energy is *equipartitioned* when X takes the form kTK^{-1} for some scalar temperature T.

Similarly to the scalar case, we can determine the smallest achievable second-moment of the phase space coordinate, $X_{\min}$, under all possible causal feedback laws $u(t) = f(\{y_m\}_0^t)$. It satisfies the filter Riccati equation

$$\begin{aligned} \dot{X}_{\min} &= (J-D)KX_{\min} + X_{\min}K(J-D)^T \\ &\quad + 2kTD - X_{\min}KBV_m^{-1}B^TKX_{\min}, \\ X_{\min}(0) &= X(0) = kTK^{-1}. \end{aligned} \tag{B4}$$

As before, the internal energy for the controlled system must obey a bound, $U(t) \geq U_{\min}(t) := \frac{1}{2}\mathrm{Tr}(KX_{\min}(t))$. The assumption on controllability and observability ensures that there exists a feedback control that drives the internal energy to the limit $U(t) = U_{\min}(t)$. Just as in the scalar case, one such control is a high-gain feedback from the Kalman-Bucy state estimate $\hat{x}$. For example, one can use $u(t) = -B_u^T G\hat{x}(t)$, for a suitably chosen large positive-definite gain matrix G.

Using the first law of thermodynamics, Eq. (B3), we can quantify the maximum possible amount of extractable work by

$$\begin{aligned} W_{\max}(t) &= \int_0^t -\dot{U}_{\min} + kT\mathrm{Tr}(KD) \\ &\qquad - \mathrm{Tr}(KDKX_{\min})\,dt' \\ &= \frac{1}{2}\int_0^t \mathrm{Tr}(KX_{\min}KBV_m^{-1}B^TKX_{\min})\,dt'. \end{aligned} \tag{B5}$$

The mutual information between the trajectories of the uncontrolled component of Eq. (B2) (analogously to Eqs. (11)–(13)) and the measurement y_m, is [25, 39]

$$\mathcal{I}(t) = \frac{1}{2}\int_0^t \mathrm{Tr}(KBV_m^{-1}B^TKX_{\min})\,dt', \tag{B6}$$

which clearly has many factors in common with $W_{\max}$. Nevertheless, in the matrix case, the integrand in $W_{\max}$ does not generically factorize into a product of the information rate and a scalar temperature, unless $X_{\min}$ is equipartitioned, $X_{\min} = kT_{\min}K^{-1}$ for some scalar $T_{\min}$. However, it is possible to define a useful scalar instantaneous *effective* temperature for arbitrary X as follows. By assuming $\dot{Q} = 0$ instantaneously in Eq. (B3), we *define* the effective temperature in the state $X(t)$ as

$$T_X(t) := \frac{\mathrm{Tr}[KDKX(t)]}{k\mathrm{Tr}(KD)}. \tag{B7}$$

The physical intuition behind the definition is that if the system has covariance $X(t)$ and is connected to a heat bath of temperature $T_X(t)$, along the direction B, then there is no instantaneous heat exchange between the system and the heat bath. This effective temperature does not depend on the friction coefficient r, and transforms Eq. (B3) into a Fourier-like heat conduction equation as in the scalar case (see Eq. (4)):

$$\dot{Q} = k\mathrm{Tr}(KD)(T - T_X). \tag{B8}$$

If the system is equipartitioned at temperature T, then $T_X \equiv T$.

Using the effective temperature and applying the Cauchy-Schwarz inequality ($\mathrm{Tr}(AB)^2 \leq \mathrm{Tr}(AA^T)\mathrm{Tr}(BB^T)$) we obtain the general lower bound in Eq. (2),

$$k\int_0^t T_{\min}\dot{\mathcal{I}}\,dt' \leq W_{\max}(t), \quad T_{\min} := T_{X_{\min}}. \tag{B9}$$

Note that in NESS, $\dot{X}_{\min} = 0$, the solution to Eq. (B4) is given by $X_{\min}^{\mathrm{NESS}} = kT_{\min}^{\mathrm{NESS}}K^{-1}$, where $T_{\min}^{\mathrm{NESS}}$ is given by the same formula as for the overdamped Langevin case, Eq. (A4), using $1/R = r$ in μ. In NESS, it holds that the maximum work extraction rate is exactly given by

$$\dot{W}_{\max}^{\mathrm{NESS}} = kT_{\min}^{\mathrm{NESS}}\dot{\mathcal{I}}^{\mathrm{NESS}}, \tag{B10}$$

and the lower bound in Eq. (2)/Eq. (B9) is reached. Therefore, it is only in an initial transient phase where we expect some slack in the inequality. As $t \to \infty$, the lower bound approaches an equality.

Finally, let us prove the upper bound in Eq. (2) for the multidimensional case. For simplicity, and without loss of generality, let us choose coordinates in the phase space such that $K = I_{2n}$ (the identity matrix). Then $X_{\min}(0) = kTI_{2n}$, and from Eq. (B4) it follows that $X_{\min}(t) - kTI_{2n}$ is symmetric negative semi-definite for all $t \geq 0$. Rewriting the maximum-work formula in Eq. (B5), using that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for matrices of compatible dimensions, we have

$$\begin{aligned} W_{\max}(t) &= \frac{1}{2V_m}\int_0^t B^T X_{\min}^2 B\, dt' \\ &\leq kT\frac{1}{2V_m}\int_0^t B^T X_{\min} B\, dt' \\ &= kT\mathcal{I}(t). \end{aligned} \tag{B11}$$

The inequality follows since $X_{\min}(t) - kTI_{2n}$ is negative semi-definite. This concludes the proof.

[1] J. C. Maxwell, *Theory of Heat* (Longmans, London, 1871).
[2] L. Szilard, Z. Phys. **53**, 840 (1929).
[3] R. Landauer, IBM Journal of Research and Development **5**, 183 (1961).
[4] C. H. Bennett, International Journal of Theoretical Physics **21**, 905 (1982).
[5] O. Penrose, *Foundations of Statistical Mechanics: A Deductive Treatment*, Dover Books on Physics Series (Dover Publications, Incorporated, 2005).
[6] H. Leff and A. Rex, *Maxwell's Demon 2 Entropy, Classical and Quantum Information, Computing*, Maxwell's Demon (CRC Press, 2010).
[7] D. Mandal, H. T. Quan, and C. Jarzynski, Phys. Rev. Lett. **111**, 030602 (2013).
[8] J. M. Horowitz and S. Vaikuntanathan, Phys. Rev. E **82**, 061120 (2010).
[9] T. Sagawa and M. Ueda, Phys. Rev. Lett. **109**, 180602 (2012).
[10] P. Strasberg, G. Schaller, T. Brandes, and M. Esposito, Phys. Rev. Lett. **110**, 040601 (2013).
[11] D. Mandal and C. Jarzynski, PNAS **109**, 11641 (2012).
[12] A. C. Barato and U. Seifert, (2013), arXiv:1308.4598.
[13] T. Sagawa and M. Ueda, Phys. Rev. Lett. **104**, 090602 (2010).
[14] Y. Fujitani and H. Suzuki, Journal of the Physical Society of Japan **79**, 104003 (2010).
[15] D. Abreu and U. Seifert, EPL (Europhysics Letters) **94**, 10001 (2011).
[16] T. Sagawa and M. Ueda, Phys. Rev. E **85**, 021104 (2012).
[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley and Sons, New York, 1991).
[18] H.-H. Hasegawa, J. Ishikawa, K. Takara, and D. Driebe, Physics Letters A **374**, 1001 (2010).
[19] M. Esposito and C. V. den Broeck, EPL (Europhysics Letters) **95**, 40004 (2011).
[20] S. Deffner and C. Jarzynski, Phys. Rev. X **3**, 041003 (2013).
[21] J. M. Horowitz and J. M. R. Parrondo, EPL (Europhysics Letters) **95**, 10005 (2011).
[22] J. M. Horowitz, T. Sagawa, and J. M. R. Parrondo, Phys. Rev. Lett. **111**, 010602 (2013).
[23] T. Schmiedl and U. Seifert, Phys. Rev. Lett. **98**, 108301 (2007).
[24] M. Bauer, D. Abreu, and U. Seifert, Journal of Physics A: Mathematical and Theoretical **45**, 162001 (2012).
[25] S. K. Mitter and N. J. Newton, Journal of Statistical Physics **118**, 145 (2005).
[26] A. C. Barato, D. Hartich, and U. Seifert, Phys. Rev. E **87**, 042104 (2013).
[27] G. Diana and M. Esposito, (2013), arXiv:1307.4728.
[28] J. Bechhoefer, Rev. Mod. Phys. **77**, 783 (2005).
[29] K. J. Åström and R. M. Murray, *Feedback Systems: An Introduction for Scientists and Engineers* (Princeton University Press, 2008).
[30] E. Aurell, C. Mejía-Monasterio, and P. Muratore-Ginanneschi, Phys. Rev. Lett. **106**, 250601 (2011).
[31] E. Aurell, K. Gawdzki, C. Mejá-Monasterio, R. Mohayaee, and P. Muratore-Ginanneschi, Journal of Statistical Physics **147**, 487 (2012).
[32] J.-C. Delvenne and H. Sandberg, Physica D: Nonlinear Phenomena **267**, 123 (2014).
[33] T. Munakata and M. L. Rosinberg, J. Stat. Mech. **2013**, P06014 (2013).
[34] J. B. Johnson, Phys. Rev. **32**, 97 (1928).
[35] H. Nyquist, Phys. Rev. **32**, 110 (1928).
[36] W. Wonham, SIAM Journal on Control **6**, 312 (1968).
[37] K. J. Åström, *Introduction to Stochastic Control Theory*, Dover Books on Electrical Engineering Series (Dover Publications, Incorporated, 2006).
[38] R. S. Bucy and P. D. Joseph, *Filtering for Stochastic Processes with Applications to Guidance* (Interscience Publishers, New York, 1968).
[39] R. Liptser and A. Shiryaev, *Statistics of Random Processes: II. Applications*, Stochastic Modelling and Applied Probability (Springer Berlin Heidelberg, 2001).
[40] H. Lamb, Proc. London Math. Soc. **s1-32**, 208 (1900).
[41] D. K. Cheng, *Field and Wave Electromagnetics*, 2nd ed. (Addison-Wesley, 1989).
[42] C. Van den Broeck, Phys. Rev. Lett. **95**, 190602 (2005).
[43] M. Esposito, K. Lindenberg, and C. Van den Broeck, Phys. Rev. Lett. **102**, 130602 (2009).
[44] The dissipative force does not need to be parallel with the actuation force. This was the case for the overdamped Langevin equation but is not necessary in higher dimension.