Thermodynamics of Linear Systems

Jean-Charles Delvenne, Henrik Sandberg, and John C. Doyle

***Abstract*— We rigorously derive the main results of thermodynamics, including Carnot's theorem, in the framework of time-varying linear systems.**

I. Introduction

Classical thermodynamics, since the work of Carnot and his followers, has been very successful to describe the relations between heat, energy and mechanical work, the quantity of work that can be extracted from heat sources, and to quantify the irreversibility observed in Nature. The discoveries of classical thermodynamics are summarised by four Laws, whose validity is based upon the fact that their consequences have been successfully verified experimentally.

We now quickly review the basics of classical thermodynamics. For a more detailed discussion see for instance [1]. The Universe is partitioned into one or several systems and the environment. A physical system is supposed to be at any moment completely characterized by a small list of 'relevant' state variables, such as internal energy, temperature, entropy, volume, pressure, etc. The first three are in fact defined by the laws themselves. These state variables are not necessarily independent. The systems are supposed to be always 'at equilibrium', meaning that if isolated from the environment, neither the system nor any of its subsystems would undergo any change of state variables. If we are to deal with systems 'out of equilibrium', then we must seek a decomposition of the system into subsystems that are constantly at equilibrium. The four laws give some constraints on the type of evolution of the state variables that any physical systems must respect. We quickly review those laws.

The Zeroth Law states that if two physical systems are in thermal equilibrium (i.e., exchange no heat when put into contact) with a third, they are also in thermal equilibrium between them; hence 'being in thermal equilibrium' is an equivalence relation. It is argued that this allows the introduction of temperature: To every system is associated a real number called temperature such that two systems are in thermal equilibrium if and only if they have the same temperature. We can now legitimately include temperature as a state variable. If a system is in a state such that it is unable to provide any heat to the environment or any other system, then we may fix its temperature to zero. This defines an absolute scale of temperature.

The First Law states that heat is merely a form of energy, like mechanical energy. Those two forms of energies can be converted into each other. The First Law also states that we can associate to any system a state variable called the internal energy, and another called mechanical energy, whose sum, the total energy, changes only through exchange of heat and work with the environment. Definitions of mechanical energy and work are borrowed from classical physics. This is written mathematically by the following:

$$\dot{E} = \dot{U} + \dot{E}_{mech} = \text{heat} + \text{work},$$

where E denotes the total energy of a system, U the internal energy, E_{mech} the mechanical energy, 'heat' is the flow of heat provided by the environment, and 'work' is the work exerted by the environment per unit of time. An example of work is the one exerted by forces of pressure to change the volume of the system. To clarify conventions, we say work is being supplied (by the environment to the system) when work is positive and work is being extracted (from the system to the environment) when work is negative. Similarly for heat.

The First Law provides a way to measure heat with the same unit as mechanical energy, rather than in calories (one calorie is the amount of heat needed to increase the temperature of one litre of water by one Celsius degree). Another consequence is that a system isolated from the environment has a constant total energy.

The Second Law has perhaps the richest consequences and admits several formulations. The Kelvin-Planck statement asserts that a system that exchanges heat with one single heat bath is unable to provide a positive work to the environment if it undergoes a cyclic transformation. A heat bath is a system purely characterized by its temperature, that is, so huge that any exchange of heat in reasonable quantities with another system will not affect its temperature. A cyclic transformation is one in which the state variables of the system assume the same values at the end and at the beginning of the process.

The Clausius formulation of the Second Law states that to any system can be associated with a state variable called entropy and denoted S, whose evolution is given by:

$$\dot{S} = \frac{q}{T},$$

where T is the temperature of the system. It also states that the total entropy of Universe never decreases. From this, it

J.-C. Delvenne is with Imperial College, Institute for Mathematical Sciences, 53 Prince's Gate, South Kensington, London, SW7 2PG, UK. jc.delvenne@imperial.ac.uk

H. Sandberg and J.C. Doyle are with California Institute of Technology, Control and Dynamical Systems, M/C 107-81, Pasadena, CA 91125, USA. {henriks,doyle}@cds.caltech.edu

H. Sandberg is supported by the Hans Werthén foundation and a postdoctoral grant from the Swedish Research Council.

Part of this work was developed while J.-C. Delvenne was with California Institute of Technology and Université catholique de Louvain, and supported by FNRS (Belgian Fund for Scientific Research).

can be proven that heat never flows from a cold system to a hot system spontaneously, and that

$$\dot{S} \geq \frac{q}{T_e},$$

where T_e is the temperature of the system that supplies heat (for instance a heat bath). The Kelvin-Planck statement easily follows from there. We also see that there is no negative temperature on an absolute scale.

The Third Law, or Nernst principle, states that the entropy of any crystalline body at zero temperature can be taken as zero. As a consequence, it is impossible for such a system to reach a zero temperature in finite time.

An important consequence of the Second Law is Carnot's theorem. It states that the conversion of heat into work is possible if and only if we are able to exchange energy with two baths of different temperatures. More precisely, if one or several systems undergo a cyclic transformation, then the total work exerted by these systems on the environment during the cycle is at most $Q_{hot}(1 - \frac{T_{cold}}{T_{hot}})$, where Q_{hot} is the total amount of heat supplied by the hot bath. The heat not converted into work has been transferred to the cold bath. This optimal quantity is attained for a Carnot cycle, in which the entropy of the Universe is preserved. Preservation of entropy implies that all heat transfers are made between systems with same temperature. As systems with same temperature do not exchange heat, we have to suppose that all transfers of heat are 'infinitely slow'. A typical Carnot cycle goes through the following phases:

1) A system is connected to a hot bath of temperature T_{hot}, from which it receives infinitely slowly Q_{hot};
2) the system is connected to the environment, to which it supplies work without exchange of heat, and its temperature drops from T_{hot} to T_{cold};
3) the system is connected to a cold bath of temperature T_{cold}, to which it gives some heat infinitely slowly;
4) the system is connected to the environment without exchange of heat, from which it receives work, and its temperature rises from T_{cold} to T_{hot};

Here ends our review of classical thermodynamics. This theory, although consistent with experiments and of great convenience for engineers, suffers from the fact that its first principles are postulated independently from other fundamental laws of physics: namely, Newton's laws or quantum mechanics.

Statistical mechanics attempts to derive all four Laws of thermodynamics from the fundamental laws of physics. For a detailed account, see for instance [2]. This approach was pioneered by Daniel Bernoulli in the 18th century and developed by Clausius, Maxwell, Boltzmann and their followers in the 19th century. The basic idea is to consider thermodynamics as a theory of large systems, whose full state is described by a number of state variables of the order at least 10^{23}. Since it is in general not feasible to measure and handle this many variables, we settle for a handful of macroscopic variables. The values of the other variables are unknown and endowed with a probability distribution. There are many ways to formalize this idea, depending on the choice of models for physical microscopic reality. We will assume that microscopic physics are the same as macroscopic electro-mechanics, but energy conserving, that is, without dissipation, and with continuous time and state space.

For an isolated system at equilibrium, this microscopic probability distribution is generally assumed to be uniform among all microscopic states that are compatible with the value of macroscopic variables. This fundamental assumption can be used to derive other distributions for non isolated systems, such as the Boltzmann distribution for systems in contact with a heat bath.

In this context, internal energy U is interpreted as the kinetic and potential energy of the many degrees of freedom composing the system, around their mean positions, and is no different in nature from the (macroscopic) mechanical energy E_{mech}. This allows us to derive the First Law as a consequence of the fact that all fundamental interactions between particles are conservative.

For isolated systems at equilibrium, entropy is interpreted as the logarithm of the volume of the microscopic state space compatible with the value of macroscopic variables (up to a physical constant). For non isolated systems, entropy can be generalised as Shannon differential entropy of the distribution of probability (although Shannon defined his entropy long after Boltzmann). Sometimes the state space is discretised into small cells, and the entropy is then defined as the Shannon discrete entropy of the discretised state. If internal energy is chosen as a macroscopic state variable, the concept of temperature is then defined as the inverse of the rate of increase of entropy when the internal energy increases:

$$T^{-1} = dS/dU,$$

up to Boltzmann's constant. From this relation, we can recover Clausius formula. However the statement that entropy always increases remains a deep issue. Statistical mechanics also allows to discover new facts, such as the so-called equipartition of energy: Every degree of freedom that contributes quadratically to the total energy carries on average the same energy $\frac{1}{2}T$ (up to Boltzmann's constant).

Several of the many possible formulations of statistical mechanics can be used to derive a rigorous mathematical formulation of the Second Law. However most of them are based on assumptions that are not known to hold for physical systems. For instance, the fact that the equilibrium distribution of an isolated system is uniform should be justified. The notion of heat bath is equally difficult to model rigorously. Moreover, the very notion of probability distribution is problematic. If probabilities are given a frequentist meaning, then the uniqueness of a distribution given the macroscopic variables cannot be proved. If we understand them in a bayesian meaning, then we are left with the impression, displeasing to many, that thermodynamics is a theory of human observation rather than physical systems. It is fair to say that there is at this time no derivation of

the laws of thermodynamics from the fundamental laws of physics, that would be perfectly rigorous and embrace the generality claimed by classical thermodynamics.

Linear systems theory has been used in statistical mechanics, for instance to derive the so-called fluctuation-dissipation theorem [3]. Conversely, several concepts of thermodynamics have been fruitfully implemented in systems theory. Let us cite a few examples — we apologize to the reader for the probable lack of exhaustivity. Dissipative systems by Willems [4], [5] are now classical. More recently, the exchange of heat and entropy in interconnected dynamical systems has been thoroughly analysed by Haddad et al. [6], from a classical thermodynamics point of view. Mitter and Newton [7] have analysed the balance of entropy and energy in Kalman-Bucy filters. Finally, Brockett and Willems [8] have provided a stochastic formulation of the problem of extraction of work from heat baths in linear systems with a time-varying capacitor, proving Carnot's theorem in this context.

It nevertheless seems that no attempt has been made to unify and synthesise the principal arguments of statistical thermodynamics in the well-understood framework of stochastic linear systems and recover as much as possible of classical thermodynamics. This is precisely our goal. The first motivation is to clarify how different physical concepts interrelate in a well-defined framework. The second is to show how thermodynamics and statistical physics can enrich the theory of stochastic linear systems with new problems, emphasizing in particular physical realizability in the design of controllers. The third is the hope that new results, both in linear systems and in thermodynamics might stem from this framework. In this paper, we meet some of this goals through a generalisation of [8] combined with the results in [9], that provide a microscopic lossless model of heat bath and dissipative systems.

The paper is organised as follows. First we describe the class of time-varying lossless, strictly causal linear systems, which we argue to be the natural class of models to consider. We define total energy, work and entropy in this context. Then we introduce a model of heat bath and dissipative system, introducing heat, dissipation and fluctuation; the First Law is then stated. We then prove the Kelvin-Planck statement and a restricted version of Clausius formula, along with Carnot's theorem.

II. Lossless strictly causal time-varying systems

We consider that all open physical systems, as predicted by classical physics, are lossless (energy-conserving) and strictly causal (the effect of an input cannot be felt immediately in the output). As classical thermodynamics considers systems changing in time, e.g., via a moving wall, piston, connection/disconnection to a heat bath, etc., we use time-varying systems in our study. We model the interaction of the system with the environment in an input-output fashion. A behavioural approach [10] would probably be the most natural. We believe however that the results are not essentially affected by this choice. Note that the environment is not here explicitly modelled as itself another strictly causal, lossless system but as an abstract entity able to interact with the system in any manner compatible with the causality and losslessness of the system.

Hence we consider systems of the kind:

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t), \\ y(t) &= C(t)x(t), \end{aligned} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state. The total energy of the system is

$$E(t) = \frac{1}{2} x(t)^T \Sigma(t) x(t).$$

The matrices $A(t), B(t), C(t), \Sigma(t)$ are thought of as controlled through a vector of inputs $v(t)$. The inputs may be the values of capacitances, inductances, resistances in a circuit. It may also be a discontinuous signal modeling an interrupter. To keep the notation simple, however, we write $A(t)$ instead of $A(v(t))$, etc. We assume any time evolution of these matrices to be possible, as long as conservation of energy is respected, as detailed below.

The system is controlled in open-loop by the signals $u(t)$ and $v(t)$. In the following, we call 'linear control' the control exerted by $u(t)$, and 'nonlinear control' the control exerted by $v(t)$.

Note that several evolutions $A(t), B(t), C(t), \Sigma(t)$ describe the same time-varying system up to a change of coordinates. Indeed, if we consider the variable $z(t) = R(t)x(t)$, then Equation (1) becomes

$$\begin{aligned} \dot{z}(t) &= (RAR^{-1} + \dot{R}R^{-1})z(t) + RBu(t), \\ y(t) &= CR^{-1}z(t), \end{aligned} \tag{2}$$

with energy

$$E(t) = \frac{1}{2} z(t)^T R^{-T} \Sigma R^{-1} z(t),$$

where the dependency on time is dropped to simplify the notation.

Some systems of coordinates are certainly more natural than others. For example, the equation of a time-varying capacitor $C(t)$, with current $i(t)$ as input and voltage $V(t)$ as output, can be written in the three following ways, according to whether the charge $x(t) = q(t)$, the voltage $V(t)$ or $z(t) = q(t)/\sqrt{C(t)}$ is chosen as state variable

$$\begin{aligned} \dot{x} &= 0x + i, & V &= x/C & \text{with} \quad E &= \frac{1}{2C}x^2; \\ \dot{V} &= -\frac{\dot{C}}{C}V + \frac{1}{C}i, & V &= V & \text{with} \quad E &= \frac{1}{2}CV^2; \\ \dot{z} &= -\frac{\dot{C}}{C}z + \frac{1}{\sqrt{C}}i, & V &= \frac{z}{\sqrt{C}} & \text{with} \quad E &= \frac{1}{2}z^2. \end{aligned} \tag{3}$$

The first equation has a zero matrix A, and the third equation has a constant matrix Σ. The second equation has none of these advantages. The coordinates x and z are both used in the following. Note that while we can take the nonlinear

input $v(t) = C(t)$ using the x-coordinates, we need to take, for instance, $v(t) = (C(t), \dot{C}(t))$ when using the other coordinates. Hence if we fix the set of nonlinear inputs, then this restricts the possible changes of coordinates.

In general, let us write the variation of energy:

$$\begin{aligned}\dot{E}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}x(t)^T\Sigma x(t)\\ &= \frac{1}{2}x(t)^T(A^T\Sigma + \Sigma A + \dot{\Sigma})x(t) + u^T(t)B^T\Sigma x(t).\end{aligned}$$

The last term involves u and represents the power provided by the linear input. In many physical systems, this power takes the form $u^T(t)y(t)$, e.g., the product of current and voltage. The first term is the power provided to the system due to the nonlinear control. For instance, changing a capacitance, e.g., by changing the distance between two parallel charged plates, in an electrical circuit, will change the energy stored in the capacitor through an exchange of mechanical work with the environment. Similarly, modifying the shape of a mechanical system may provide work through pressure forces (but we do not have a precise linear system to exemplify this case).

Let us now assume that we fix a vector of nonlinear inputs $v(t)$, such that at any moment, we can instantaneously freeze their values. For instance, if the input v is the distance between two plates, we consider that we can at any moment suddenly stop moving the plates. If we choose coordinates in which only $v(t)$ appears, not its derivatives for instance, then freezing v means freezing $A(t)$, $\Sigma(t)$, $B(t)$ and $C(t)$ to their current value, in which case we have a linear time-invariant system. Then, from dissipativity theory for linear time-invariant systems, we know that the system is lossless if and only if

$$\begin{aligned}A^T(t)\Sigma(t) + \Sigma(t)A(t) &= 0\\ \Sigma(t)B(t) &= C^T(t)\end{aligned} \tag{4}$$

for every time t. Since we could freeze the parameters at any time, (4) must hold instantaneously. We will take such coordinates as our coordinates of reference and denote them by $x(t)$, although another set of coordinates, denoted by $z(t)$, will prove useful in the following. Compare this with the time-varying capacitor example (3). Using (4), the variation of energy in x-coordinates is written:

$$\begin{aligned}\dot{E}(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}x(t)^T\Sigma x(t)\\ &= \frac{1}{2}x(t)^T(A^T\Sigma + \Sigma A + \dot{\Sigma})x(t) + u^T(t)B^T\Sigma x(t)\\ &= \frac{1}{2}x(t)^T\dot{\Sigma}x(t) + u^T(t)y(t).\end{aligned} \tag{5}$$

The second term is the power supplied by the linear input $u(t)$, while the first is the power supplied through the nonlinear control $v(t)$.

If the system has many degrees of freedom or is left unmeasured, then it is reasonable to attribute a probability distribution on $x(t)$. Here we do not have to decide if the probability distribution should have a bayesian or frequentist meaning; see [9] and references within for a short discussion.

If the distribution of the state $x(t)$ is random, then we call $X(t) = \mathbb{E}(x(t) - \mathbb{E}x(t))(x(t) - \mathbb{E}x(t))^T$ the covariance matrix of the state, and it is supposed to be invertible. We define the entropy of the system to be

$$S(t) = \frac{1}{2}\log\det X(t).$$

In the case where the distribution is Gaussian, this is precisely the Shannon entropy, up to the additive constant $\frac{n}{2}\log(2\pi e)$. In any other case, the entropy is higher than the Shannon entropy.

Jacobi's formula for an invertible matrix $X(t)$ yields $\frac{\mathrm{d}}{\mathrm{d}t}\det X(t) = \det X(t)\mathrm{Tr}(X^{-1}(t)\dot{X}(t))$, leading to:

$$\dot{S}(t) = \frac{1}{2}\mathrm{Tr}(X^{-1}\dot{X}).$$

As discussed earlier, we can choose coordinates $z(t) = R(t)x(t)$ such that the new energy matrix $R^{-T}(t)\Sigma(t)R^{-1}(t)$ is the identity, see the last line in (3). We will also use this system of coordinates, in which some computations are easier.

The equation of evolution in z-coordinates is written:

$$\begin{aligned}\dot{z}(t) &= (J + M)z(t) + RBu(t)\\ y(t) &= B^TR^Tz(t).\end{aligned}$$

where $J(t)$ and $M(t)$ are the skew-symmetric and symmetric parts of $R(t)A(t)R^{-1}(t) + \dot{R}(t)R^{-1}(t)$. Then the energy is written $E(t) = \frac{1}{2}z(t)^Tz(t)$ and its variation is

$$\dot{E}(t) = z^T(t)M(t)z(t) + u^T(t)y(t). \tag{6}$$

Now, the work provided through the nonlinear control $v(t)$ is represented by M. Comparing (5) with (6), we find that $\frac{1}{2}\dot{\Sigma} = R^TMR$. The entropy can be written $S = \frac{1}{2}\log\det R^{-1}ZR^{-T} = \frac{1}{2}\log\det Z\Sigma^{-1}$, where $Z(t)$ is the covariance matrix of $z(t)$. The variation of entropy in z-coordinates becomes:

$$\begin{aligned}\dot{S}(t) &= \frac{1}{2}\mathrm{Tr}Z^{-1}(t)\dot{Z}(t) - \frac{1}{2}\mathrm{Tr}\Sigma^{-1}(t)\dot{\Sigma}(t)\\ &= \frac{1}{2}\mathrm{Tr}Z^{-1}(t)\dot{Z}(t) - \mathrm{Tr}(R^TR)^{-1}(t)R^T(t)M(t)R(t)\\ &= \frac{1}{2}\mathrm{Tr}Z^{-1}(t)\dot{Z}(t) - \mathrm{Tr}M(t).\end{aligned}$$

Note that the uncertainty in the state can come from a random initial condition and/or from a random linear input $u(t)$. If the linear input is deterministic, then it is easy to see that $\dot{S}(t) = 0$. Hence the notion of entropy becomes interesting for random $u(t)$, as we shall see in the next section.

III. Dissipative time-varying systems

As our goal is to understand how heat is transformed into work, we will suppose that some input/output pairs (u_i, y_i) are connected to heat baths of temperature T_i. A heat bath is intuitively a very large system whose temperature remains constant for a very long period of time if exchanges

of energies with other systems are moderate. It is shown in [9] how to construct a lossless strictly causal SISO linear time-invariant system with many degrees of freedom whose behaviour approximates arbitrarily well the following equation

$$y(t) = \frac{1}{2}k^2 u(t) + k\sqrt{T}n(t)$$

over an arbitrarily long time horizon, where $n(t)$ is white noise of unit intensity, and u, y are the input/output of the bath. A typical example is a resistor affected by Nyquist-Johnson thermal noise of temperature T.

Now a system is connected to such a heat bath, say, through the connection $u_i = y$ and $y_i = -u$. The minus sign comes from the fact that a lossless connection must satisfy $uy = -u_i y_i$: all the power leaving the system is entering the heat bath.

If one or several input/output pairs u_i, y_i are related to such a heat bath, the system is governed by the equation of the form:

$$\begin{aligned} \dot{z}(t) &= (J + M - \frac{1}{2}\sum_i F_i F_i^T)z(t) + RBu(t) + \sum_i \sqrt{T_i}F_i n_i(t), \\ y(t) &= B^T R^T z(t), \end{aligned} \tag{7}$$

where the n_i are independent Gaussian white noise processes, F_i describe the interconnection with the heat baths and u, y are the input/output pairs not connected to any heat bath; see [9]. Such a system is called 'dissipative'.

We now have to choose in open-loop the evolution of $J(t), M(t), F_i(t), B(t), u(t)$. The case of closed-loop control is discussed in Section VII.

IV. Heat, work, and Clausius formula

Several forms of energies are to be distinguished next. The expected total energy of the system is also denoted $E(t)$, with a slight abuse of notation. We can then write $E(t) = \frac{1}{2}\mathrm{Tr}Z(t) + \frac{1}{2}\mathbb{E}z^T(t)\mathbb{E}z(t)$. The first term $U \doteq \frac{1}{2}\mathrm{Tr}Z(t)$ can be called internal energy, because it is related to the random deviation of variables around their mean, while the second term can be interpreted as (macroscopic) mechanical energy E_{mech}. For instance, a spring-mass system can have a mechanical potential energy proportional to the square of the average length of the spring, and an internal energy due to small random movements of the mass around its average.

The variation of mechanical energy can be expanded to

$$\begin{aligned} \dot{E}_{mech} &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}z^T\mathbb{E}z \\ &= \mathbb{E}z^T(M - \frac{1}{2}\sum_i F_i F_i^T)\mathbb{E}z + \mathbb{E}y^T\mathbb{E}u. \end{aligned}$$

Hence the mechanical energy can be increased or decreased by a supply or extraction of work through both the nonlinear and linear inputs; it can also be dissipated under the form of heat. As far as as the extraction of work is concerned, the best way to manage mechanical energy is to drive the mean $\mathbb{E}z(t)$ to zero as soon as possible, thus extracting the corresponding quantity of work. This can be done for instance by applying an appropriate linear input $u(t)$. If we wait longer, then the mean will converge to zero by the effect of the dissipation term, which means the loss of valuable energy to the heat bath. This phenomenon is explored quantitatively in [11], both in open-loop and feedback schemes. Note that while the linear input drives the mean and has no effect on the covariance matrix, the white noise fluctuation acts on the covariance but not on the mean. That is why linear control is unable to extract any work from a supply of heat, and this justifies a posteriori the introduction of a nonlinear control. From now on, we suppose that the mean has been driven to zero by dissipation or extraction of work, and the mechanical energy is zero. We therefore focus on the sole internal energy.

The variation of covariance matrix $Z(t)$ is written as:

$$\begin{aligned} \dot{Z} = (J+M-\frac{1}{2}\sum_i F_i F_i^T)Z + Z(J+M-\frac{1}{2}\sum_i F_i F_i^T)^T \\ + \sum_i T_i F_i F_i^T. \end{aligned} \tag{8}$$

Hence the variation of internal energy is written:

$$\dot{U}(t) = \mathrm{Tr}(M - \frac{1}{2}\sum_i F_i F_i^T)Z + \frac{1}{2}\sum_i \mathrm{Tr}T_i F_i F_i^T. \tag{9}$$

The term

$$w \doteq \mathrm{Tr}MZ$$

is interpreted as the rate of work supplied to the system by the environment through the nonlinear control. The term $\frac{1}{2}\sum_i \mathrm{Tr}F_i F_i^T Z$ is the amount of power given by the system to the heat baths, i.e., it is the amount of heat flowing out of the system. The term $\frac{1}{2}\sum_i T_i \mathrm{Tr}F_i F_i^T$ is the power provided by the heat baths to the system, i.e., it is the heat flowing into the system. The net heat flow supplied by bath i is

$$q_i \doteq \frac{1}{2}T_i \mathrm{Tr}F_i F_i^T - \mathrm{Tr}F_i F_i^T Z.$$

Now we can write the First Law for internal energy:

$$\dot{U} = w + \sum_i q_i,$$

and prove the following version of the Second Law:

Theorem 1 (Clausius formula): For a dissipative system in contact with several heat baths of temperature T_i, the variation of entropy $\dot{S}$ is related to the heat flows q_i as follows:

$$\dot{S} \geq \sum_i \frac{q_i}{T_i}. \tag{10}$$

The equality is obtained if and only if at every time, the system is connected to only one temperature (i.e., such that all baths i for which $F_i \neq 0$ have the same temperature), and the covariance matrix $Z(t)$ is $T_i I$.

Proof: We have that

$$\begin{aligned}\dot{S} - \sum_i \frac{q_i}{T_i} &= \frac{1}{2}\mathrm{Tr}(Z^{-1}\dot{Z}) - \mathrm{Tr}M - \sum_i \frac{q_i}{T_i}, \\ &= \frac{1}{2}\sum_i T_i \mathrm{Tr} F_i F_i^T Z^{-1} - \frac{1}{2}\sum_i \mathrm{Tr} F_i F_i^T \\ &\quad - \sum_i \frac{q_i}{T_i} \\ &= \frac{1}{2}\sum_i \mathrm{Tr}(T_i Z^{-1} - I) F_i F_i^T - \frac{1}{2}\sum_i \mathrm{Tr} F_i F_i^T \\ &\quad + \frac{1}{2} T_i^{-1} \mathrm{Tr} F_i F_i^T Z \\ &= \frac{1}{2}\sum_i \mathrm{Tr}(T_i Z^{-1} + T_i^{-1} Z - 2I) F_i F_i^T \\ &= \frac{1}{2}\sum_i \mathrm{Tr} F_i^T (T_i Z^{-1} + T_i^{-1} Z - 2I) F_i.\end{aligned}$$

Now the quantity $T_i Z^{-1} + T_i^{-1} Z - 2I$ has an eigenvalue $\lambda + \lambda^{-1} - 2$ for every eigenvalue λ of $T_i^{-1} Z$. As $\lambda > 0$, we have $\lambda + \lambda^{-1} - 2 \geq 0$, with equality if and only if $\lambda = 1$. Hence $T_i Z^{-1} + T_i^{-1} Z - 2I$ is nonnegative definite, and $\dot{S} - \sum_i \frac{q_i}{T_i} \geq 0$. We have equality if and only if at every moment only one temperature is accessible to the system, say T_i, and $Z = T_i I$. ∎

This a generalisation of the corresponding theorem in [8]. To maintain equality in the Clausius inequality, we need a constant equipartition $Z = TI$, which means that the internal energy U is constant and the heat flow q is zero. Hence the work extracted must be zero as well, and M must be zero (unless $T = 0$). Any attempt to exchange non-zero work by nonlinear control on a system connected to a heat bath must result in an entropy production in excess to the right-hand side of 10. In a somewhat flexible way, physicists attribute such an excess of entropy to 'irreversibilities'.

However, if M is nonzero but small compared to all nonzero $\frac{1}{2} F_i F_i^T$, this allows a slow exchange of work with the environment, while the deviation $Z - TI$ from the equilibrium remains negligible.

If we write $Z = T(I + D)$, where D is a small deviation from equipartition, then

$$\begin{aligned}\dot{S} - \frac{q}{T} &= \frac{1}{2}\sum_i \mathrm{Tr} F_i^T (T Z^{-1} + T^{-1} Z - 2I) F_i, \\ &= \frac{1}{2}\sum_i \mathrm{Tr} F_i^T (D + (I + D)^{-1} - I) F_i, \qquad (11) \\ &\cong \sum_i \frac{1}{2} \mathrm{Tr} F_i^T D^2 F_i.\end{aligned}$$

On the other hand, D and M are related through the following equation:

$$\begin{aligned}\dot{D} = 2M + JD - DJ + MD + DM \\ - \frac{1}{2}\sum_i F_i F_i^T D - \frac{1}{2}\sum_i D F_i F_i^T.\end{aligned}$$

It appears that it takes a small M (slow exchange of work) to have a small D. Thus in the limit of small D, the terms $MD + DM$ are even smaller compared to the others. If we neglect them, we observe the equation is linear in M and D. If we replace M by kM, for any $k \geq 0$, then D becomes approximately kD.

In conclusion, if we speed up the extraction or nonlinear supply of work by a factor k, the excess production rate of entropy due to irreversibilities is multiplied by k^2. As the total time needed to exchange a given amount of work is divided by k, the total excess of entropy is multiplied by k. This means that by slowing down the transformation of a time-varying system, we can reduce arbitrarily the excess amount of entropy. If t is the time of the transformation needed to exchange a given amount of energy, then the excessive entropy generated scales as $1/t$, in the limit of large t. This is to be compared to results in finite-time thermodynamics; see, e.g., [12].

V. One temperature bath: Kelvin-Planck statement of the Second Law

Suppose now that only one heat bath, of temperature T, is available (or equivalently, all heat baths have the same temperature T). Then, from Clausius formula, we have that for any evolution from time 0 to time t, the increase of entropy is

$$\Delta S = \int_0^t \dot{S} \geq \frac{1}{T}\int_0^t q.$$

But the right-hand side is equal to $\int_0^t \dot{U} - w = \Delta U - \mathrm{Work}_{0\to t}$. Hence,

$$\mathrm{Work}_{0\to t} \geq \Delta(U - TS),$$

which is independent of the path taken by the system.

If we suppose that only one heat bath is available, then the drop of $U - TS$ (called Helmholtz free energy) gives the maximum amount of work extractable during a transformation. The system is said to describe a cycle if $A(t) = A(0)$, $B(t) = B(0)$, $C(t) = C(0)$, $\Sigma(t) = \Sigma(0)$, $X(t) = X(0)$. As a result, no work can ever be extracted from a single temperature source by a system describing a cycle, since the change of Helmholtz free energy is zero. This is the Kelvin-Planck's statement of the Second Law.

VI. Two temperature baths: Carnot's theorem

When two heat baths or more are available, we expect to prove Carnot's theorem, namely that if the system describes a cycle, the work extracted divided by the heat entering the system is at most $1 - T_{cold}/T_{hot}$. The difficulty is to define 'heat entering the system'. We will discuss two definitions, and Carnot's theorem is true for both. For one of them, this bound is attained by the 'Carnot cycle'.

A. Hot bath vs. cold bath

Following closely classical thermodynamics, we can separate the heat $\int_0^t q$ as the sum of heat exchanged with the hot bath and heat exchanged with the cold bath (if only two baths are available). Every heat bath i exchanges heat

$$\int_0^t q_i = \frac{1}{2}\int_0^t -\mathrm{Tr}F_iF_i^TZ + \mathrm{Tr}T_iF_iF_i^T.$$

If the system, connected to two baths, describes a cycle, then from Clausius formula, $\frac{1}{T_{hot}}\int q_{hot} + \frac{1}{T_{cold}}\int q_{cold} \leq \Delta S = 0$. From elementary algebraic manipulation, we can show that the total quantity of work extracted is equal to $\int q_{hot} + \int q_{cold}$ and not larger than $(1 - \frac{T_{cold}}{T_{hot}})\int q_{hot}$. Thus, if the work extracted is positive, then $\int q_{hot} > 0$. Since $\frac{1}{T_{hot}}\int q_{hot} + \frac{1}{T_{cold}}\int q_{cold} \leq 0$, we see that $\int q_{cold} < 0$.

Hence, if work is extracted during the cycle, the hot bath is globally a source of heat, while the cold source is globally a sink of heat. The efficiency can hence be defined as the work extracted divided by the heat $\int q_{hot}$. The efficiency is at most $1 - T_{cold}/T_{hot}$.

This efficiency can be attained with equality if Clausius formula is true with equality at all times. This is the case if at every moment, the system is either connected to the hot bath only, with covariance matrix $T_{hot}I$, or it is connected to the cold bath only, with covariance matrix $T_{cold}I$, or it is connected to neither. A cycle respecting those conditions is called a 'Carnot cycle'. As explained in Section IV, it is not strictly possible to achieve a Carnot cycle, except in the limit of large times. The simplest example of ideal Carnot cycle is the following:

1) $F_{cold} = 0$ and M has small constant nonnegative eigenvalues. Work is extracted, $Z = T_{hot}I$ is constant.
2) $F_{cold} = 0$ and $F_{hot} = 0$. M has nonpositive eigenvalues (possibly large), Z is decreased to $T_{cold}I$.
3) $F_{hot} = 0$ and M has small constant nonpositive eigenvalues. Work is supplied, $Z = T_{cold}I$ is constant.
4) $F_{cold} = 0$ and $F_{hot} = 0$. M has nonnegative eigenvalues (possibly large), Z is increased to $T_{hot}I$.

If more than two baths are available, then it is best to connect the system to the hottest and coldest baths.

B. In-coming vs. out-going heat

Every heat bath provides a net heat flow to the system equal to $q_i = -\frac{1}{2}\mathrm{Tr}F_iF_i^TZ + \frac{1}{2}\mathrm{Tr}T_iF_iF_i^T$. The first term, accounting for dissipation, is always nonpositive and the second term, accounting for fluctuation, is always nonnegative. Hence it is natural to consider that the total supply of heat to the system is $q_+ \doteq \frac{1}{2}\sum_i \mathrm{Tr}T_iF_iF_i^T$, while the total loss of heat is $-q_- \doteq \frac{1}{2}\mathrm{Tr}\sum_i F_iF_i^TZ$.

Now the efficiency is the total work extracted $\int_0^t w$ divided by the total supply of heat $\int_0^t q_+$. This is essentially the definition used in [11].

Clausius inequality can then be written

$$\int_0^t \mathrm{Tr}(F_{hot}F_{hot}^T + F_{cold}F_{cold}^T) \leq \int_0^t \mathrm{Tr}(T_{hot}^{-1}F_{hot}F_{hot}^TZ + T_{cold}^{-1}F_{cold}F_{cold}^TZ). \quad (12)$$

On the other hand, $T_{cold} \leq T_{hot}$, $q_+ \leq \frac{1}{2}\mathrm{Tr}T_{hot}\sum_i F_iF_i^T$ and $-q_- = \frac{1}{2}\mathrm{Tr}\sum_i F_iF_i^TZ$. Plugging those inequalities into Clausius inequality leads to $\frac{\int_0^t |q_-|}{\int_0^t q_+} \geq \frac{T_{cold}}{T_{hot}}$. Hence the efficiency is bounded by $1 - \frac{T_{cold}}{T_{hot}}$.

However, we easily see that, unless trivial case, the inequality can never be reached. In particular, it is clear that a very slow cycle when connected to a bath cannot be optimal, since if we act slowly a lot of the energy brought in by the fluctuations will be immediately given back to the bath by dissipation without producing work, thus deteriorating the efficiency. This notion of efficiency therefore seems more challenging to study.

VII. Closed-loop control and Maxwell's demon

So far we have only considered open-loop control. We are allowed to know the initial value of macroscopic state variables such as energy, and we never measure the output. This follows the setting of classical thermodynamics, but does not use the full power of control theory. Can we break the Second Law with feedback control? It seems that by measuring the output, we reduce the uncertainty, thus decreasing the covariance matrix and increasing the expected value. In other terms, measurement reduces entropy and converts internal energy into mechanical energy. This energy can then be easily retrieved as work by linear control, apparently for free. This is essentially what physicists call Maxwell's demon paradox.

The generic solution to the paradox is to explain why such a controller, if physically implemented, should dissipate enough energy and generate enough entropy to keep the Second Law unviolated; see [13]. In our case, the solution goes as follows: Whatever linear time-varying feedback controller (possibly itself controlled in open-loop by nonlinear inputs) we choose, the closed-loop system must take the form (7) again, thus breaks neither the Second Law nor Carnot's theorem, as proved above.

To illustrate this general result, we can for instance think of a controller composed of a measuring device, a Kalman filter, and an actuator, where every of these elements must be modeled by an equation of the form (7). It seems obvious that an estimator like the Kalman filter is to converge to a good estimate of the state, hence any physical linear realization must dissipate energy and be disturbed by thermal noise, leading to imperfections. The impossibility of perfect measurement is discussed quantitatively in [9], and is used in [11] to analyse in depth the performance of heat engines with imperfect measurement, over finite and infinite time intervals.

The following general question remains open: Given a linear system of the form (7), what is the optimal feedback controller of the same form? Here optimal is understood in

relation to the extraction of work. We can optimise, e.g., the power extracted or the efficiency in the one of the meanings of Section VI.

VIII. Conclusion

We recovered the main results of thermodynamics, especially concerning a system connected to one or several heat baths. Future work can be devoted to generalising this to any interconnection between any kind of physical systems, and to explore more precisely the impact of finite-time transformation.

IX. Acknowledgements

The authors would like to thank Ben Recht for many stimulating discussions.

References

[1] E. Fermi, *Thermodynamics*. New York: Dover, 1956.

[2] G. H. Wannier, *Statistical Physics*. New York: Dover Publications, 1987.

[3] R. Kubo, "The fluctuation-dissipation theorem," *Reports on Progress in Physics*, vol. 29, no. 1, pp. 255–284, 1966.

[4] J. C. Willems, "Dissipative dynamical systems part I: General theory," *Archive for Rational Mechanics and Analysis*, vol. 45, pp. 321–351, 1972.

[5] ——, "Dissipative dynamical systems part II: Linear systems with quadratic supply rates," *Archive for Rational Mechanics and Analysis*, vol. 45, pp. 352–393, 1972.

[6] W. M. Haddad, V. S. Chellaboina, and S. G. Nersesov, *Thermodynamics: A Dynamical Systems Approach*. Princeton University Press, 2005.

[7] S. K. Mitter and N. J. Newton, "Information and entropy flow in the Kalman-Bucy filter," *Journal of Statistical Physics*, vol. 118, pp. 145–176, 2005.

[8] R. W. Brockett and J. C. Willems, "Stochastic control and the second law of thermodynamics," in *Proceedings of the IEEE Conference on Decision and Control*, San Diego, California, 1978, pp. 1007–1011.

[9] H. Sandberg, J.-C. Delvenne, and J. C. Doyle, "The statistical mechanics of fluctuation-dissipation and measurement back action," in *Proceedings of the American Control Conference 2007*, New-York, Accepted for publication.

[10] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory — A Behavioral Approach*. Springer, 1997.

[11] H. Sandberg, J.-C. Delvenne, and J. C. Doyle, "Linear-quadratic-gaussian heat engines," *submitted*, 2007.

[12] B. Andresen, R. S. Berry, A. Nitzan, and P. Salamon, "Thermodynamics in finite time. i. the step-carnot cycle," *Physical Review A*, vol. 15, pp. 2086–2093, 1977.

[13] H. S. Leff and A. F. Rex, Eds., *Maxwell's Demon 2: Entropy, Classical and Quantum, Information, Computing*. Institute of Physics Publishing, 2003.