Journal of Statistical Physics, Vol. 118, Nos. 1/2, January 2005 (© 2005)
DOI: 10.1007/s10955-004-8781-9

Information and Entropy Flow in the Kalman–Bucy Filter

Sanjoy K. Mitter[1] **and Nigel J. Newton**[2]

Received February 5, 2004; accepted August 3, 2004

We investigate the information theoretic properties of Kalman–Bucy filters in continuous time, developing notions of information supply, storage and dissipation. Introducing a concept of *energy*, we develop a physical analogy in which the unobserved signal describes a statistical mechanical system interacting with a heat bath. The abstract 'universe' comprising the signal and the heat bath obeys a non-increase law of entropy; however, with the introduction of partial observations, this law can be violated. The Kalman–Bucy filter behaves like a Maxwellian demon in this analogy, returning signal energy to the heat bath without causing entropy increase. This is made possible by the steady supply of new information. In a second analogy the signal and filter interact, setting up a stationary non-equilibrium state, in which energy flows between the heat bath, the signal and the filter without causing any overall entropy increase. We introduce a *rate of interactive entropy flow* that isolates the statistical mechanics of this flow from marginal effects. Both analogies provide quantitative examples of Landauer's Principle.

KEY WORDS: Information theory; Landauer's principle; non-equilibrium statistical mechanics; statistical filtering.

1. INTRODUCTION

In this article we study continuous-time Kalman–Bucy filters from information theoretic and statistical mechanical viewpoints. The information flows for such filters are identified and these strongly resemble the

[1]Department of Electrical Engineering and Computer Science, and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, United States of America; e-mail: mitter@mit.edu
[2]Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom; e-mail: njn@essex.ac.uk

entropy flows of non-equilibrium statistical mechanics, which occur when a system is held away from its equilibrium state by an interaction with an exogenous system. (See, for example, ref. 8.)

By introducing a concept of energy, we construct a physical analogy for the Kalman–Bucy filter, in which the partially observed signal interacts with a heat bath. The interaction forces the signal towards a stationary state, which maximises the entropy of the 'universe' comprising the signal and the heat bath. Whatever the initial state might be it is impossible for this entropy to decrease at any stage during convergence, and so our abstract universe obeys a law akin to the Second Law of Thermodynamics. However, this law can be broken in the presence of partial observations: the entropy of the abstract universe can be reduced (at least temporarily) at a rate bounded above by that of the information supply from the observations. We show, in the analogy, that the filter behaves like a Maxwellian *demon*,[20] extracting energy from the signal and returning it to the heat bath, thus 'cooling' the signal. The filter acts as a *heat pump* in this analogy but, unlike those of real physical heat pumps, its operations cause no overall increase in entropy. This is made possible by the steady supply of new observations.

In a second physical analogy, the *joint* system, comprising the signal and filter, interacts with a heat bath. We identify 'conditional signal' and filter subsystems, and show that energy flows around a loop comprising these subsystems and a heat bath. In the stationary state of this system, energy flows with no attendant change in the overall entropy. Thus the system in the second analogy is a type of perpetual motion machine, at the limits of a dynamic theory of non-equilibrium statistical mechanics. We use recent techniques in this theory (see refs. 2, 9 and 17), which are based on dynamic Markov models, to quantify the entropy flows in this system, and introduce a concept of *interactive entropy flow* to isolate the interaction of the components from their internal, autonomous non-equilibrium mechanics.

Information theoretic aspects of filters have received much attention in the literature. For example, a variety of information theoretic measures of optimality such as the mutual information between the observation and the estimation error are considered in refs. 7, 13 and 27. In particular, these articles show that many of these measures are optimised by the Kalman filter in the linear Gaussian filtering problem of this article. More interesting is the fact that they provide a framework for deriving sub-optimal filters in the nonlinear case. The measure of interest in this article is the mutual information between the observation path and the signal value. Our results can be extended to non-linear systems, but then involve the infinite-dimensional non-linear filter. Except in special cases, the *entropic*

efficiency of the optimal filter in the statistical mechanical analogies is not shared by finite-dimensional approximations.

The results most closely connected with our work are to be found in refs. 6 and 26. The first of these develops an expression for the mutual information between the signal and observation *paths* in a fairly general context. This result is the basis of our definition of *information supply* in Section 3. In ref. 26, a comparison is made between Duncan's path information quantity and the mutual information between the observations and signal value. This is intimately connected with our *information dissipation* process. We do not develop any new theorems of filtering (although we do not think that our dynamical view of information theory, namely the way in which information flows from the observations into the conditional distribution, has appeared before). The aim of this paper is, rather, to make connections between two disparate fields (filtering theory and statistical mechanics) and to show that the martingale theoretic techniques of optimal filtering are ideally suited to the study of the interactions between subsystems of a statistical mechanical system, and more widely to the study of non-equilibrium systems. We believe that our definition and study of interactive entropy flow is novel.

Our research is partly inspired by the doctoral thesis of Michael Propp[24], written under the direction of the first author. In this, an input–output view of a thermodynamic system is constructed by associating a Markov process with the system and then defining forces and fluxes for this process. A dissipation inequality, analogous to that of Willems[28] is then derived for this process. There, as in recent developments in non-equilibrium statistical mechanics and the results presented here, time reversibility plays an essential role. These ideas were also applied in ref. 24 to the study of electrical networks involving Nyquist–Johnson resistors. (See, also, the related work in ref. 4.)

Our work is also connected with ideas on the thermodynamics of computation, which have received much attention in recent decades. (See ref. 1 for a review article.) Because they can be investigated in the context of simple abstract universes comprising a few components, our physical analogies for the Kalman–Bucy filter provide precise, quantitative examples of Landauer's Principle.[16] This states that, under certain circumstances, entropy can be increased by the erasure of information. Of course it is not our aim here to investigate the feasibilty (or otherwise) of thermodynamically efficient computing machines.

The analogies in this article concern stable, time-homogeneous systems. Two further articles, in preparation by the second author, develop an *interactive* analogy, which applies to a much wider class of linear systems. They also extend the three analogies to non-linear systems.

The specific problem addressed here is that of the evolution of a linear, partially observed, Gaussian system over the time interval, $[0, \infty)$. The use of a specific start time admits the study of transient effects. The 'steady state' of the system can be investigated by means of appropriate initialisation. (Our model includes an initial observation, which allows the signal and filter to be initialised in any consistent state.)

All random variables and processes will be defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. (In the measure-theoretic tradition of probability, this is a space of *outcomes*, Ω, a σ-field of *events* (measurable subsets of Ω), $\mathcal{F}$, and a *probability measure* $\mathbb{P}: \mathcal{F} \to [0, 1]$. For example, the *event* that a particular scalar random variable takes a value exceeding (say) unity is a member of $\mathcal{F}$ and is assigned a probability of occurence by the map $\mathbb{P}$.) The primary stochastic processes (those appearing in Eqs. (1) and (2)) will be adapted to a *filtration* $(\mathcal{F}_t, t \in [0, \infty))$. This is a non-decreasing (in the sense of inclusion) family of sub-σ-fields of $\mathcal{F}$, and represents the 'growth of randomness' with t. Thus an event may belong to $\mathcal{F}_t$ but not to $\mathcal{F}_s$ for some $s < t$, meaning that whether or not it occurs is determined at least in part by randomness that first manifests itself in the system between times s and t. (An example being the event that the components of X_r are all negative at time r for some $s < r < t$.) The term *filtration* should not be confused with the *filtering* performed by the Kalman–Bucy filter, although there is certainly a connection when the latter is considered in an information-theoretic context. In the dynamic theory of information used in this article (as opposed to that based on ergodic assumptions), measure-theoretic *filtrations* are the elementary repositories of information.

In our partially observed system, the unobservable component (which we shall call the *signal*), X, is an $\mathbb{R}^n$-valued process defined by the following integral equation:

$$X_t = \xi + \int_0^t A X_s \, ds + B V_t \quad \text{for } t \in [0, \infty). \tag{1}$$

Here A and B are $n \times n$ matrices of reals, ξ is an $\mathcal{F}_0$-measurable $\mathbb{R}^n$-valued Gaussian random variable with mean zero and positive-definite covariance matrix P_i, and $(V_t, \mathcal{F}_t, t \in [0, \infty))$ is an n-vector Brownian motion. V can be thought of as being the vector process of time integrals of independent scalar white noises, and so X_t is the state at time t of a finite-dimensional linear system driven by vector white noise. (See the comments in Section 5 regarding infinite-dimensional systems.) The reason (1) is written in integral (rather than differential) form is because of the impossibility of constructing processes with the distributions of white noise that have sample

paths with reasonable analytic properties. (Some concepts from measure-theoretic probability and Itô calculus will occasionally be used in some of the more technical parts of this article, but a full familiarity with them is not essential to an understanding of its primary message. The interested reader is referred to ref. 14 for further information.)

We shall assume that:

(H1) all eigenvalues of A have negative real parts;

(H2) B has full rank.

Of course X, thus defined, is a zero-mean vector Gaussian process with covariance matrix, $P(t) = \mathbf{E}X_t X_t'$, satisfying

$$P(t) = P_i + \int_0^t \left(AP(s) + P(s)A' + \Sigma_V\right) ds,$$

where Σ_V is the positive-definite $n \times n$ matrix BB'.

The observable component of our partially observed system (which we shall call simply the *observation*), Y, is an $\mathbb{R}^d$-valued process (for some $d \leqslant n$) defined by the integral equation

$$Y_t = CX_0 + \zeta + \int_0^t \Gamma X_s \, ds + W_t \quad \text{for } t \in [0, \infty), \tag{2}$$

where C and Γ are $d \times n$ matrices of rank d, ζ is an $\mathcal{F}_0$-measurable $\mathbb{R}^d$-valued Gaussian random variable, independent of ξ, with mean zero and positive-definite covariance matrix M, and $(W_t, \mathcal{F}_t, t \in [0, \infty))$ is a d-vector Brownian motion, independent of V. The observation has an initial value that depends linearly on the initial value of X and the 'noise' term ζ, and a running term $(Y_t - Y_0; t \in [0, \infty))$. For an appropriately chosen C, Y could be thought of as being part of an observation process consisting of the last two terms in Eq. (2) only, but extending over negative as well as positive times. In this interpretation Y_0 would summarise the partial observations of the process X over negative times.

The aim of 'filtering' is to estimate the signal at each time t, making full use of the information furnished by the observations up to that time, $(Y_s, s \in [0, t])$. The Kalman–Bucy filter is a formula (recursive in t) for calculating the conditional distribution of X_t given $(Y_s, s \in [0, t])$, which is Gaussian. The *covariance form* of the filter propagates the mean vector, $\hat{X}$, and covariance matrix, Q, of this conditional distribution. These evolve as

follows

$$\begin{aligned}
\hat{X}_0 &= \left(P_i^{-1} + C'M^{-1}C\right)^{-1} C'M^{-1}Y_0, \\
\hat{X}_t &= \hat{X}_0 + \int_0^t (A - Q(s)\Sigma_W)\,\hat{X}_s\,ds + \int_0^t Q(s)\Gamma'\,dY_s, \\
Q(0) &= \left(P_i^{-1} + C'M^{-1}C\right)^{-1}, \\
\dot{Q}(t) &= AQ(t) + Q(t)A' + \Sigma_V - Q(t)\Sigma_W Q(t),
\end{aligned} \tag{3}$$

where Σ_W is the positive-semi-definite $n \times n$ matrix $\Gamma'\Gamma$. (See, for example, ref. 5 or 10.) Since P_i and M are positive definite the inverse matrices here are well defined, and $P(t)$ and $Q(t)$ are positive definite for all t. Note that the covariance matrix Q is not dependent on the observation, Y, and so all the information in the conditional distribution that is derived from the observation is held in the mean vector, $\hat{X}$.

We denote by $(\mathcal{F}_t^Y,\ t \in [0, \infty))$ the filtration generated by Y, and by ν the associated *innovations process*: for $t \in [0, \infty)$,

$$\begin{aligned}
\mathcal{F}_t^Y &= \sigma(Y_s,\ s \in [0, t]), \\
\nu_t &= Y_t - C\hat{X}_0 - \int_0^t \Gamma \hat{X}_s\,ds.
\end{aligned} \tag{4}$$

For each t, $\mathcal{F}_t^Y$ is the σ-field of events whose occurence or non-occurence is completely determined by the observation path up to time t, $(Y_s, s \in [0, t])$. It is a standard result of filtering theory that $(\nu_t, \mathcal{F}_t^Y)$ is a d-vector Brownian motion with non-zero initial value having the d-variate Gaussian distribution $N(0, (I_d - CQ(0)C'M^{-1})(CP_iC' + M)(I_d - M^{-1}CQ(0)C'))$. (See, for example, ref. 5.) It is also known that the filtration generated by ν coincides with that generated by Y. (Here, and in what follows, we denote the identity matrix of order n by I_n, the multi-variate Gaussian distribution with mean vector μ and covariance matrix Σ by $N(\mu, \Sigma)$, and its density by $n(\mu, \Sigma)$.) The increments of ν convey that part of the information in the corresponding increments of Y that is 'novel' or 'innovative' in the filtering context. The conditional mean vector $\hat{X}$ can be expressed entirely in terms of ν, as follows:

$$\hat{X}_t = P_iC'M^{-1}\nu_0 + \int_0^t A\hat{X}_s\,ds + \int_0^t Q(s)\Gamma'\,d\nu_s. \tag{5}$$

This representation is crucial to the developments in Section 4 since it shows that $\hat{X}$ is Markov in its own right.

2. A PHYSICAL ANALOGY FOR THE SIGNAL

In this section, we explore the notion that the signal, X, of (1) can be thought of as a *mesoscopic* description of an abstract statistical mechanical system. Its evolution is not determined solely by its current value (as would be the case if X were a *microscopic* description (of, say, a Hamiltonian system), nor is it purely deterministic (as would be the case if X were a *macroscopic* (or *thermodynamic*) quantity). The fact that X is Markov is consistent with it corresponding to a 'vanishingly small' component of the phase space variable of a large Hamiltonian system with random initial condition. (See ref. 24.) We follow recent developments in non-equilibrium statistical mechanics (see refs. 2, 9, 17 and 25) in which stationary equilibrium and non-equilibrium states correspond to invariant distributions of Markov processes. Under conditions (H1) and (H2), X has the unique invariant distribution, $N(0, P_{SS})$, with positive-definite covariance matrix satisfying the following algebraic equation

$$A P_{SS} + P_{SS} A' + \Sigma_V = 0. \tag{6}$$

(See, for example, Section 5.6 in ref. 14.)

We consider our statistical mechanical system to be an abstract 'universe', isolated from other systems and energy conserving, and identify two separate energy components: one associated with the degrees of freedom revealed by X, which we shall call the *energy of the signal*, and the other associated with the invisible degrees of freedom. The first of these is determined by the *Hamiltonian*:

$$H_X(x) = \tfrac{1}{2} x' P_{SS}^{-1} x. \tag{7}$$

It turns out that the second component can be thought of as the energy of a unit-temperature heat bath with which the signal interacts.

For a probability measure μ on $(\mathbb{R}^n, \mathcal{B}^n)$, the *average energy*, $\mathcal{E}(\mu)$, *entropy*, $\mathcal{S}(\mu)$, and *free energy*, $\mathcal{F}(\mu)$, of the signal are defined as follows:

$$\begin{aligned} \mathcal{E}(\mu) &= \int H_X(x)\mu(dx), \\ \mathcal{S}(\mu) &= \begin{cases} -\int \log\left(\frac{d\mu}{d\lambda}(x)\right) \mu(dx) & \text{if the integral exists,} \\ -\infty & \text{otherwise,} \end{cases} \\ \mathcal{F}(\mu) &= \mathcal{E}(\mu) - \mathcal{S}(\mu), \end{aligned} \tag{8}$$

where λ is Lebesgue (volume) measure. (The entropy of the signal is, thus, defined in terms of the standard volume element in $\mathbb{R}^n$.) It can easily be

shown by a variational argument that the free energy of the signal, $\mathcal{F}$, is minimised by the invariant distribution $N(0, P_{SS})$.

At time t, the average energy $E_X(t)$, entropy $S_X(t)$ and free energy $F_X(t)$ of the signal are as follows:

$$\begin{aligned} E_X(t) &= \mathcal{E}(N(0, P(t))) = \frac{1}{2}\mathrm{tr}\left(P(t)P_{SS}^{-1}\right), \\ S_X(t) &= \mathcal{S}(N(0, P(t))) = \frac{n}{2}(1+\log(2\pi)) + \frac{1}{2}\log|P(t)|, \\ F_X(t) &= \frac{1}{2}\mathrm{tr}\left(P(t)P_{SS}^{-1}\right) - \frac{n}{2}(1+\log(2\pi)) - \frac{1}{2}\log|P(t)|. \end{aligned} \tag{9}$$

As t increases, energy flows into the signal at an average rate

$$\dot{E}_X(t) = \mathrm{tr}\left(A(P(t) - P_{SS})P_{SS}^{-1}\right), \tag{10}$$

which causes its entropy to increase at rate

$$\dot{S}_X(t) = \mathrm{tr}\left(A(P(t) - P_{SS})P(t)^{-1}\right). \tag{11}$$

(Of course, both of these rates could be negative, corresponding to an average *outflow* of energy.) It then easily follows that

$$\begin{aligned} \dot{F}_X(t) &= -\frac{1}{2}\mathrm{tr}\left(\left(P_{SS}^{-1} - P(t)^{-1}\right)\Sigma_V\left(P_{SS}^{-1} - P(t)^{-1}\right)P(t)\right) \\ &\leqslant 0. \end{aligned}$$

In fact this 'non-increase' property of the free energy is true whatever the distribution of X_0; the positive definiteness of Σ_V ensures that, for every $t > 0$, X_t has a smooth density $p(\cdot, t)$ satisfying the Fokker–Planck equation

$$\frac{\partial p}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}((Ax)_i p) + \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left((\Sigma_V)_{ij} p\right),$$

from which it follows that

$$\begin{aligned} \frac{d}{dt}\mathcal{F}(p(\cdot, t)) &= -\frac{1}{2}\int \left(P_{SS}^{-1}x + \nabla \log p\right)' \Sigma_V \left(P_{SS}^{-1}x + \nabla \log p\right) p(x,t)\, dx \\ &\leqslant 0. \end{aligned}$$

The process X can be thought of as describing the evolution of an abstract statistical mechanical system subject to random exogenous forces, which add or remove energy in order to drive the system towards its invariant distribution, and so minimise its free energy.

In a more general context, stationary states of statistical mechanical systems are minimisers of free energies of the form

$$\mathcal{F}(\mu) = \mathcal{E}(\mu) - T\mathcal{S}(\mu),$$

where T is the *temperature* of the stationary state. (See, for example, ref. 8.) Thus, we can consider the part of the energy of our abstract universe not associated with the signal as being in a *heat bath* at unit temperature that supplies or absorbs heat in order to drive the system towards the stationary state $N(0, P_{\mathrm{SS}})$. During this convergence, the entropy of X may increase or decrease according to the value of the covariance, $P(t)$. Of course, the entropy of the heat bath, $S_H(t)$, is also changed by this interaction. (Our heat bath is *idealised* in the sense that it can supply or absorb any finite amount of energy without suffering a temperature change.)

It follows from the energy conserving property of our abstract universe and the fact that the heat bath has unit temperature that

$$S_H(t) = K - E_X(t)$$

for some constant K, and it easily follows that

$$\frac{d}{dt}\left(S_X(t) + S_H(t)\right) = -\dot{F}_X(t) \geq 0.$$

Thus, the overall rate of increase of entropy of our universe is non-negative, which shows that it obeys a law of non-decrease of entropy similar to the Second Law of Thermodynamics. The fact that the temperature of the equilibrium state is unity is a consequence of the way in which H_X was defined in Eq. (7).

We summarise the foregoing discussion in Proposition 2.1.

Proposition 2.1. Let $\mathcal{U}$ be a closed (energy conserving) system, whose energy and entropy are the sums of those of two sub-systems, $\mathcal{X}$ and $\mathcal{H}$. Let μ (a probability measure on $(\mathbb{R}^n, \mathcal{B}^n)$) be a 'mesoscopic state' of $\mathcal{U}$, for which the average energy and entropy of $\mathcal{X}$, $\mathcal{E}(\mu)$ and $\mathcal{S}(\mu)$, are as defined in Eq. (8), and for which the average energy and entropy of $\mathcal{H}$ are both $K - \mathcal{E}(\mu)$ for some constant K.

(i) The entropy of $\mathcal{U}$ is maximised by the mesoscopic state $N(0, P_{SS})$.

(ii) If μ evolves in time according to the Fokker–Planck equation associated with (1), and if (H1) and (H2) are satisfied, then the entropy of $\mathcal{U}$ is non-decreasing.

When X is in its stationary state there is *on average* no flow of energy between the heat bath and the signal. However, for *individual outcomes* of X there is a continuous exchange of energy back and forth between these components. In fact an application of Itô's formula shows that the energy of the signal evolves according to the following equation:

$$\begin{aligned} H_X(X_t) = H_X(X_0) + \int_0^t \operatorname{tr}\left(A(X_s X_s' - P_{SS}) P_{SS}^{-1}\right) ds \\ + \int_0^t X_s' P_{SS}^{-1} B \, dV_s \end{aligned} \tag{12}$$

It is these *fluctuations* that cause some energy to *loop* in the presence of observations. (See Section 3.) The invariant distribution represents a type of *dynamic equilibrium*.

We follow the nomenclature of ref. 17 by referring to invariant distributions of Markov processes as *stationary equilibrium states* or *stationary non-equilibrium states* according to the net *flow* of entropy at the invariant distribution, an equilibrium state being one for which this flow is zero. In order to quantify entropy flow for the process X we first define its *entropy production*. This involves the time-reversed dynamics of X. For some (large) $T < \infty$ and each $t \in [0, T]$, let

$$\begin{aligned} \bar{X}_t &= X_{T-t}, \\ \bar{\mathcal{F}}_t^X &= \sigma(\bar{X}_s,\ s \in [0, t]), \\ \bar{A}(t) &= -A - \Sigma_V P (T-t)^{-1}, \\ \bar{V}_t &= B^{-1}\left(\bar{X}_t - \bar{X}_0 - \int_0^t \bar{A}(s)\bar{X}_s \, ds\right). \end{aligned} \tag{13}$$

Lemma 2.1. (i) The process $(\bar{V}_t, \bar{\mathcal{F}}_t^X, t \in [0, T])$ is an n-dimensional Brownian motion.

(ii) For each $t \in [0, T]$,

$$\bar{\mathcal{F}}_t^X = \sigma(\bar{X}_0,\ (\bar{V}_s, s \in [0, t])). \tag{14}$$

Proof. It follows from (1) and the definition of $\bar{X}$ that for $0 \leqslant s \leqslant t \leqslant T$

$$\mathbf{E}\bar{X}_s \bar{X}_t' = \exp(-A(t-s))P(T-t).$$

Straightforward calculations now show that, for any $0 \leqslant r \leqslant s \leqslant t \leqslant T$,

$$\mathbf{E}\bar{X}_r(\bar{V}_t - \bar{V}_s)' = 0,$$

and that $\bar{V}$ has the following quadratic covariation:

$$\begin{aligned}[\bar{V}, \bar{V}']_t &:= \lim_{N\uparrow\infty} \sum_{n=1}^{N} \left(\bar{V}_{nt/N} - \bar{V}_{(n-1)t/N}\right)\left(\bar{V}_{nt/N} - \bar{V}_{(n-1)t/N}\right)' \\ &= I_n t \quad \text{for all } t \in [0, T],\end{aligned}$$

and so $\bar{V}$ is an independent-increments Gaussian process, independent of $\bar{X}_0$, and with this quadratic covariation. It is thus a standard n-vector Brownian motion with respect to the filtration it generates. It now follows from the definition of $\bar{V}$ that

$$\bar{X} = \bar{\Phi}\left(\bar{X}_0, \bar{V}\right),$$

where $\bar{\Phi}$ is the strong solution of the following Itô equation:

$$d\bar{X}_t = \bar{A}(t)\bar{X}_t\, dt + B d\bar{V}_t \quad \text{for } t \in [0, T] \tag{15}$$

and this establishes part (ii). This, and the independence of $\bar{X}_0$ and $\bar{V}$ establish part (i). ∎

For each $0 \leqslant t - \epsilon < t + \epsilon \leqslant T$, let $\Pi^X_{t+\epsilon|t}$ and $\Pi^X_{t-\epsilon|t}$ be the X_t-conditional distributions of the processes $(X_u,\ u \in [t, t+\epsilon])$ and $(\bar{X}_u,\ u \in [T-t, T-t+\epsilon])$, respectively. Since the diffusion matrix Σ_V is positive definite it follows from the Cameron–Martin–Girsanov theorem (see, for example, Chapter 6 in ref. 18) that $\Pi^X_{t+\epsilon|t}$ and $\Pi^X_{t-\epsilon|t}$ are mutually absolutely continuous probability measures with Radon Nikodym derivative (relative density)

$$\begin{aligned}\frac{d\Pi^X_{t+\epsilon|t}}{d\Pi^X_{t-\epsilon|t}}(X) = \exp\Bigg(&\int_t^{t+\epsilon} X_s'(A - \bar{A}(\bar{s}(s)))'(B')^{-1}\, dV_s \\ &+ \frac{1}{2}\int_t^{t+\epsilon} X_s'(A - \bar{A}(\bar{s}(s)))'\Sigma_V^{-1}(A - \bar{A}(\bar{s}(s)))X_s\, ds\Bigg),\end{aligned}$$

where $\bar{s}(s) = T + s - 2t$. Thus we may define the *rate of entropy production* of X at time $t \in (0, T)$ as

$$\begin{aligned} R_X(t) &:= \lim_{\epsilon \downarrow 0} \frac{1}{\epsilon} \mathbf{E} \log \left(\frac{d\Pi^X_{t+\epsilon|t}}{d\Pi^X_{t-\epsilon|t}}(X) \right) \\ &= \frac{1}{2} \mathrm{tr} \left((A - \bar{A}(T-t))' \Sigma_V^{-1} (A - \bar{A}(T-t)) P(t) \right). \end{aligned} \tag{16}$$

Remark 2.1. $R_X(t)$ measures the *degree of time-asymmetry* of the process X at time t. Imagine a game in which one player secretly 'cuts out' the small segment of a sample path of X in the interval $(t-\epsilon, t+\epsilon)$, tosses a coin, reversing the time direction of the segment if 'heads' occurs, and then shows the other player the segment, asking whether or not it has been reversed. $R_X(t)$ is a measure of the average degree of ease with which the second player could answer correctly.

Remark 2.2. $R_X(t)$ would be infinite if Σ_V were singular, since it would then be possible, with probability one, for the second player in the game described above to distinguish time directions. For example, consider the case in which

$$A = \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix} \quad \text{and} \quad \Sigma_V = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

The direction of time could be distinguished with probability one, here, from a comparison of the signs of $X_{t,2} - X_{t,1}$ and of the slope of $X_{t,2}$ at time t.

Remark 2.3. The rate of entropy production of the time-reversed process $\bar{X}$ at time t is the same as that of the forward process X at time $T - t$.

Remark 2.4. $R_X(t)$ is non-negative. It is zero if and only if X is in its invariant distribution at time t ($P(t) = P_{\mathrm{SS}}$) and X is *self-adjoint* in the sense that

$$P_{\mathrm{SS}} A' = A P_{\mathrm{SS}}, \tag{17}$$

in which case $\bar{A}(t) = A$ for all t, and the dynamics of X are identical in both time directions.

Remark 2.5. R_X is *homeomorphism invariant*. If f is a continuous, one-to-one mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$, then it induces new probability measures on the space of continuous functions from $[t, t+\epsilon]$ to $\mathbb{R}^n$ corresponding to $\Pi^X_{t+\epsilon|t}$ and $\Pi^X_{t-\epsilon|t}$. These can be used to define the rate of entropy production of the process $f(X_t)$. This is, of course, equal to $R_X(t)$.

We can now define the entropy flow of X (possibly away from its invariant distribution) as the difference between its rate of entropy production and its rate of change of entropy:

$$\Phi_X(t) := R_X(t) - \dot{S}_X(t). \tag{18}$$

Thus the entropy production comprises two parts: one that drives the process towards its stationary (minimum free energy) state, and another that represents net entropy flow. If X is self-adjoint in the sense of Eq. (17), then this flow is zero in the stationary state, and the latter is called an equilibrium state; otherwise it is called a non-equilibrium state.

The following electrical example, involving Nyquist–Johnson resistors, illustrates the foregoing analogy. A Nyquist–Johnson resistor of value R Ohms produces a Gaussian white noise voltage of mean-square value $2TR$, where T is the absolute temperature of the resistor in units for which the Boltzmann constant is unity. (See refs. 11 and 23.) Consider the circuit of Fig. 1, which comprises a linear inductor, a linear capacitor and two Nyquist–Johnson resistors. The latter are supposed to be held at unit temperature by immersion in a (physical) heat bath.

The electrical energy stored in the circuit is determined by the current in the inductor and the voltage between the plates of the capacitor. Taking these to be the two components of an $\mathbb{R}^2$-valued process X, it follows

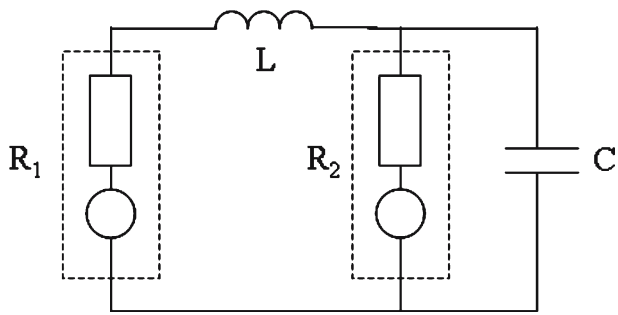

Fig. 1. Example system.

that X satisfies (1) with

$$A = \begin{bmatrix} -R_1/L & -1/L \\ 1/C & -1/(CR_2) \end{bmatrix},$$
$$B = \sqrt{2}\begin{bmatrix} \sqrt{R_1}/L & 0 \\ 0 & 1/(C\sqrt{R_2}) \end{bmatrix}.$$

The invariant distribution of X is $N(0, P_{SS})$, where

$$P_{SS} = \begin{bmatrix} 1/L & 0 \\ 0 & 1/C \end{bmatrix},$$

and so

$$H_X(x) = \frac{1}{2}Lx_1^2 + \frac{1}{2}Cx_2^2.$$

The energy of the signal here corresponds to the electrical energy stored in the circuit. This energy fluctuates as heat passes to and from the physical heat bath through the resistors. The whole 'universe' here has many more degrees of freedom than $X_1(t)$ and $X_2(t)$. However, these are hidden in the inner workings of the heat bath.

This system is not self-adjoint, and has a rate of entropy flow in the stationary state of

$$R_{X,SS} = 1/(R_1C) + R_2/L.$$

This means that, in the spirit of Remark 2.1, it is possible to make informed judgements about the direction of time from observation of small segments of the path of X in steady-state.

In a well known paradox of statistical mechanics due to Maxwell,[20] a *demon* is able to make heat flow from a box containing a low temperature gas into an adjacent box containing a gas at higher temperature, thus (apparently) reducing the entropy of the system and violating the Second Law of Thermodynamics. It does this by observing the molecules of both gases in the vicinity of a (closable) hole connecting the two boxes. When a molecule of the cool gas with unusually high kinetic energy approaches the hole, the demon opens it allowing the molecule through. It does likewise when a molecule of the hot gas with unusually *low* kinetic energy approaches the hole from the other side. In fact it is generally accepted that this does not violate the Second Law since, in carrying out its role,

the demon is not only reducing the entropy of the system of gases but also *erasing* the information held in the system's observational state. According to Landauer's Principle this erasure, in so much as it causes irreversibility, involves entropy increase in other components of the Universe. (See, for example, ref. 16 or 1.) The demon can avoid irreversibility by retaining copies of all the observational state measurements it has used in performing its role. We show in the following section that the existence of the observation process Y of Eq. (2) allows an entropy reduction of the demonic type in the 'universe' comprising the signal and the heat bath.

3. THE ROLE OF OBSERVATIONS

We begin this section by evaluating the information flows that occur in the Kalman–Bucy filter. Let $C(t)$ be the *mutual information* between X_t and $(Y_s,\ s \in [0, t])$:

$$C(t) := I(X_t\,;\ (Y_s, s \in [0, t])),$$

where, for random variables Θ and Φ taking values in Borel spaces and having joint and marginal distributions $\mathbb{P}_{\Theta\Phi}$, $\mathbb{P}_\Theta$ and $\mathbb{P}_\Phi$:

$$I(\Theta\,;\ \Phi) = \int \log\left(\frac{d\mathbb{P}_{\Theta\Phi}}{d(\mathbb{P}_\Theta \otimes \mathbb{P}_\Phi)}\right) d\mathbb{P}_{\Theta\Phi}. \tag{19}$$

$C(t)$ can be thought of as being the observation-derived information on X_t *stored* by the filter at time t. Since it is a *mutual* information, it is not dependent on any underlying reference measure (such as Lebesgue (volume) measure). It is invariant with respect to measurable one-to-one mappings on the underlying Borel spaces, and so has an *absolute* meaning, unlike quantities such as the signal entropy, $S_X(t)$ of (9). (For example, $C(t) = 0$ would imply that the observations up to time t were completely useless for estimating X_t.) Also, since $\hat{X}_t$ is a *sufficient statistic* for the conditional distribution of X_t (i.e., the randomness in the conditional distribution of X_t given $(Y_s, s \in [0, t])$ is completely contained in $\hat{X}_t$), $C(t)$ is also the mutual information between X_t and $\hat{X}_t$. According to the invariance property, it is also the mutual information between $f_1(X_t)$ and $f_2(\hat{X}_t)$ for any measurable one-to-one maps, f_1 and f_2, from $\mathbb{R}^n$ to $\mathbb{R}^n$. It now easily follows that

$$\begin{aligned} C(0) &= \frac{1}{2}\log|P_i + M| - \frac{1}{2}\log|M|, \\ \dot{C}(t) &= \frac{1}{2}\mathrm{tr}\,(\Sigma_W Q(t)) - \frac{1}{2}\mathrm{tr}\left(\Sigma_V (Q(t)^{-1} - P(t)^{-1})\right), \end{aligned} \tag{20}$$

from which it is tempting to think of the information *supply* to the store up to time t as being

$$S(t) = \frac{1}{2}\log|P_i + M| - \frac{1}{2}\log|M| + \frac{1}{2}\int_0^t \mathrm{tr}(\Sigma_W Q(s))\,ds, \tag{21}$$

and the information *dissipated* from the store up to time t as being

$$D(t) = \frac{1}{2}\int_0^t \mathrm{tr}\left(\Sigma_V(Q(s)^{-1} - P(s)^{-1})\right)ds. \tag{22}$$

Lemma 3.1 justifies this interpretation. The novelty here is one of interpretation; the result concerning the mutual information $C_p(0,t)$ was first proved in ref. 6, and that concerning the mutual information $C_p(t,t)$ is proved in ref. 26. (See also Lemma 16.9 in ref. 19.) A sketch proof is included here for the sake of completeness.

Lemma 3.1. Let S and D be as defined in (21) and (22), and, for any $s \leqslant t$ let $C_p(s,t)$ be the mutual information between the *paths* $(X_r,\ r \in [s,t])$ and $(Y_r,\ r \in [0,t])$; then

$$C_p(s,t) = S(t) - D(s). \tag{23}$$

Proof. Let $\mathbb{P}_t^R$ and $\mathbb{P}_t^M$ be the probability measures on $\mathcal{F}$, defined by the following Girsanov transformations. (See, for example, Chapter 6 in ref. 18)

$$\begin{aligned}
\frac{d\mathbb{P}_t^R}{d\mathbb{P}} &= \frac{n(0, I_d)(Y_0)}{n(C\xi, M)(Y_0)}\exp\left(-\int_0^t X_r'\Gamma'\,dY_r + \frac{1}{2}\int_0^t X_r'\Sigma_W X_r\,dr\right),\\
\frac{d\mathbb{P}_t^M}{d\mathbb{P}_t^R} &= \frac{n(0, CP_iC' + M)(Y_0)}{n(0, I_d)(Y_0)}\exp\left(\int_0^t \hat{X}_r'\Gamma'\,dY_r - \frac{1}{2}\int_0^t \hat{X}_r'\Sigma_W\hat{X}_r\,dr\right).
\end{aligned} \tag{24}$$

It easily follows from elementary manipulations of d-variate Gaussian distributions and the Cameron–Martin–Girsanov theorem that neither transformation in (24) alters the distribution of X. However, the first transformation renders $(Y_r, r \in [0,t])$ a Brownian motion, idependent of X, having non-zero initial value with distribution $N(0, I_d)$. ($\mathbb{P}_t^R$ is the *reference* probability of filtering theory.) The second transformation restores the original marginal distribution to $(Y_r, r \in [0,t])$, while retaining its independence from X. (This follows from the innovations representation of Y

in (4).) Thus, under $\mathbb{P}_t^M$, X and $(Y_r, r \in [0,t])$ are independent but have the same marginal distributions as they have under $\mathbb{P}$.

It now follows that

$$\begin{aligned} C_p(s,t) &= C(s) + \mathbf{E}\log\left(\frac{d\mathbb{P}}{d\mathbb{P}_t^M}\left(\frac{d\mathbb{P}}{d\mathbb{P}_s^M}\right)^{-1}\right) \\ &= C(s) + \frac{1}{2}\int_s^t \operatorname{tr}(\Sigma_W Q(r))\, dr \\ &= S(t) - D(s), \end{aligned}$$

as claimed. ∎

Like C, the information quantities S, D and C_p are *mutual* informations, and thus invariant with respect to measurable one-to-one transformations. $S(0)$ is the information gain on the whole process X arising from the initial observation Y_0, and $S(t) - S(0)$ is the information gain arising from the increments of the observation Y between times 0 and t. We can think of $C_p(s,t)$ as being the information stored by a *path* estimator that has access to $(Y_r,\ r \in [0,t])$ but has no interest in the values of X prior to time s. If s increases but t remains constant, the path estimator dissipates this stored information at rate $\dot{D}(s)$; the dissipation process represents observation-derived information that was useful for estimating the past of X, but is of no use in estimating its present or future. It also has the representation

$$D(t) = \frac{1}{2}\int_0^t \mathbf{E}\operatorname{tr}\left(\Sigma_V I_F(X_s, s)\right) ds, \tag{25}$$

where $I_F(x,t)$ is the Fisher information matrix associated with the likelihood function for X_t given $(Y_s, s \in [0,t])$

$$I_F(x,t) = \mathbf{E}\nabla_x \log(\Lambda)\nabla_x \log(\Lambda)'(x,t,Y), \tag{26}$$

where

$$\Lambda(x,t,Y) = \frac{n(\hat{X}_t, Q(t))(x)}{n(0, P(t))(x)}.$$

We now take the view that entropy is simply *unobservable* information. Thus, if the signal process X of (1) were completely unobservable,

its entropy at time t would be $S_X(t)$, as defined in (9). However, this is reduced in the presence of the partial observations $(Y_s, s \in [0,t])$ to

$$\begin{aligned} S_{X|Y}(t) &= \mathbf{E}\mathcal{S}(N(\hat{X}_t, Q(t))) \\ &= S_X(t) - C(t). \end{aligned} \tag{27}$$

We cannot allow *perfect* observations of X_t since these would convert an infinite amount of entropy into stored information and create mathematical difficulties. These difficulties are avoided in (2) by the non-degeneracy of the observation noise terms ζ and W.

With the addition of observations, we could modify our two-component universe, to include a third physical component represented by the observation process Y. This would involve a modified heat bath that accounted for the observation noise, W, as well as that in the signal (V). We defer this approach until Section 4. For the moment we note that the signal energy can be split (at least conceptually) into two components, as follows:

$$H_X(X_t) = \frac{1}{2}(X_t - \hat{X}_t)' P_{\mathrm{SS}}^{-1}(X_t + \hat{X}_t) + \frac{1}{2}\hat{X}_t' P_{\mathrm{SS}}^{-1}\hat{X}_t.$$

Since the second of these is completely determined by the degrees of freedom that are observable through $(Y_s, s \in [0,t])$, it is available to a *demon* having access to Y. Thus the average energy of the signal, $\mathcal{E}(N(0, P(t)))$, can be split into two parts: that available to the demon, $\mathcal{E}(N(0, P(t) - Q(t)))$, which we shall call *work*, and that remaining, $\mathcal{E}(N(0, Q(t)))$, which we shall call *heat*. In this sense the signal is *cooled* by the observations. If the observations were to be turned off at time t (which could be achieved by setting Σ_W to zero) then the heat component of the signal energy would converge towards the steady-state value of $E_X(t)$ in much the same way as $E_X(t)$ itself. The cooled signal has entropy $S_{X|Y}(t)$, and this is less than that of the uncooled signal by the quantity of information stored by the filter, as shown by (27).

The signal now interacts with the heat bath in exactly the way it did in the absence of observations. However, the interaction now sub-divides into two sub-interactions: one between the cooled signal and the heat bath, and one between the demon and the heat bath. During fluctuations, both sub-components of the signal can *lose* energy to the heat bath, but only the cooled signal can *gain* energy from it. This is because energy coming from the heat bath has entropy associated with it. (The signal gains energy from the heat bath through the small fluctuations of the Brownian motion, V, and these are completely unpredictable from $\mathcal{F}_t^Y$.) Of course,

there is also an interaction between the sub-components of the signal: as t increases, the demon continues to 'extract' work from the cooled signal as new observations become available. The combination of these effects causes three energy *flows*, as follows.

Flow 1: Heat Bath to Cooled Signal. The average rate of flow of energy can be found from the rate of change of energy of the cooled signal with the work extraction process 'turned off'. This can be achieved by temporarily setting Σ_W to zero.

$$\begin{aligned}\dot{E}_1(t) &= \frac{d}{dt}\mathcal{E}(N(0, Q(t)))|_{\Sigma_W=0} \\ &= \mathrm{tr}\left(A\,(Q(t) - P_{\mathrm{SS}})\,P_{\mathrm{SS}}^{-1}\right).\end{aligned}$$

Flow 2: Cooled Signal to Demon. The demon continues to receive new information, which allows it to 'extract' work from the cooled signal at an average rate of

$$\begin{aligned}\dot{E}_2(t) &= \dot{E}_1(t) - \frac{d}{dt}\mathcal{E}(N(0, Q(t))) \\ &= \frac{1}{2}\mathrm{tr}\left(Q(t)\Sigma_W Q(t) P_{\mathrm{SS}}^{-1}\right).\end{aligned}$$

Flow 3: Demon to Heat Bath. As described above, the demon loses energy to the heat bath during fluctuations, but gets none back. This results in a net flow of energy from the demon to the heat bath with average rate

$$\begin{aligned}\dot{E}_3(t) &= \dot{E}_2(t) - \frac{d}{dt}\mathcal{E}(N(0, P(t) - Q(t))) \\ &= \mathrm{tr}\left(A(Q(t) - P(t))P_{\mathrm{SS}}^{-1}\right).\end{aligned}$$

The net average rate of outflow of energy from the heat bath is thus

$$\dot{E}_1(t) - \dot{E}_3(t) = \dot{E}_X(t),$$

which is unaltered by the existence of observations. The three energy flows are shown in Fig. 2.

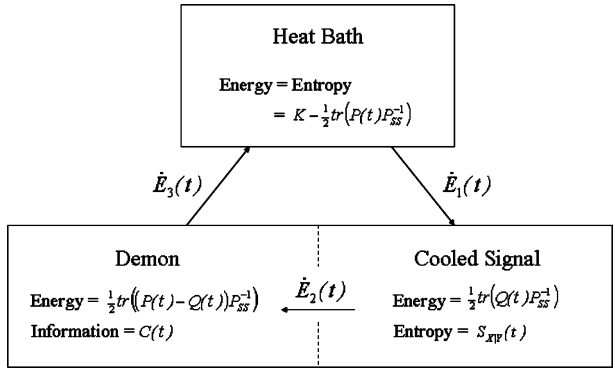

Fig. 2. Energy flows in the observed signal.

The rates of change of entropy and information are as follows.

In the Cooled Signal. The entropy is raised by the inflow of energy (Flow 1) and lowered by the outflow (Flow 2). The net rate of change is

$$\dot{S}_{X|Y}(t) = \dot{S}_X(t) + \dot{D}(t) - \dot{S}(t).$$

In the Demon. The demon has associated with it an amount of *information* $C(t)$, but no entropy. This corresponds with the notion that the energy available to it is work. $C(t)$ is increased at rate $\dot{S}(t)$ by the supply of new information, and reduced at rate $\dot{D}(t)$ by the dissipation of historical information.

In the Heat Bath. Since the net average rate of change of energy in the heat bath is unaltered by the existence of observations, so is its rate of change of entropy.

The term $\dot{D}(t)$ in the equation for $\dot{S}_{X|Y}(t)$ is the extra rate of entropy increase of the cooled signal (as compared with the uncooled signal) and is caused by the increased rate of inflow of energy from the heat bath: $\dot{E}_1(t) - \dot{E}_X(t)$. Thus the filter can be seen to be *entropically efficient* in the sense that it dissipates information at exactly the rate of this (unavoidable) entropy increase. If the filter dissipated at a higher rate, it would cause an additional increase in the entropy of the whole system, illustrating Landauer's Principle; if it dissipated at a lower rate, it would retain more information than strictly needed for estimating the future of X. In order not to cause unnecessary entropy increase, the filter must retain all information that is not held as entropy in other parts of the universe. In the stationary state, this balance means that the filter dissipates information at a rate governed by the fluctuations of $H_X(X_t)$.

The rate of change of entropy of the universe with observations differs from that of the original universe by $\dot{D}(t) - \dot{S}(t)$. If, during convergence towards the invariant distribution, $\dot{S}(t) > \dot{D}(t)$ for some t, then it is possible for the entropy of the universe to decrease at time t. For example, this is the case at $t = 0$ if the signal is initialised in its invariant distribution, $N(0, P_{SS})$, and the filter is initialised with near total 'ignorance', $M \gg C P_{SS} C'$.

In the stationary state the overall rate of change of entropy is the same as that in the original universe (zero), but energy and entropy/information circulate around a loop comprising the heat bath, the cooled signal and the demon. The demon is analogous to a perfect *heat pump* that cools the signal, returning the extracted energy to the heat bath, and doing this with no increase in entropy. It maintains the cooled signal at a temperature lower than that of the heat bath, and this causes an inflow of heat (Flow 1), with a resultant entropy increase. However, the entropy increase is countered by the steady supply of new information, which, arising as it does *within* the universe, constitutes a matching entropy decrease. This illustrates Landauer's Principle in reverse. The entropic efficiency of the Kalman–Bucy filter is a special case of the information conserving properties of Bayesian estimators investigated in ref. 21. These issues are developed further in two sequels to this article by the second author.

In the foregoing analogy the demon is 'passive' in the sense that it converts heat into work simply by observing the associated degrees of freedom. Of course, having made such observations, it could in principle convert the work into any other form of energy. An example of an 'active' demon, which does this, arises in the following *controlled* variant of the partially observed system of Section 1.

$$\begin{aligned} X_t^U &= \xi + \int_0^t \left(A X_s^U + U_s \right) ds + B V_t, \\ Y_t^U &= C X_0^U + \zeta + \int_0^t \Gamma X_s^U \, ds + W_t, \end{aligned} \tag{28}$$

where A, B, C, Γ, ξ, ζ, V and W are as in Section 1, and U is a continuous *causal* control based on the partial observations, Y^U. (This means that, for each t, U_t must be measurable with respect to the σ-field generated by $(Y_s^U, s \in [0, t])$.)

The control is the means by which the active demon extracts work from the controlled signal, its aim being to minimise the energy it leaves behind. Thus the problem the demon faces is to choose the control U to minimise the following *cost functional*

$$J_t(U) = \mathbf{E} H_X(X_t^U) \tag{29}$$

at each time t. This is a special case of the so-called *linear quadratic Gaussian* optimal control problem. (See, for example, ref. 5.) It turns out that the minimum cost is equal to the heat component of the signal energy in the passive analogy, $\mathcal{E}(N(0, Q(t)))$, and that this cost can be approximately realised by controls of the form

$$U_t = -K\hat{X}_t^U \tag{30}$$

for large K, where $\hat{X}_t^U$ is the $(Y_s^U, s \in [0,t])$-conditional mean of X_t^U.

The $(Y_s^U, s \in [0,t])$-conditional distribution of $\hat{X}_t^U$ is the Gaussian distribution $N(\hat{X}_t^U, Q(t))$, where Q is as given in (3), and

$$\begin{aligned} \hat{X}_0^U &= \left(P_i^{-1} + C'M^{-1}C\right)^{-1} C'M^{-1}Y_0^U, \\ \hat{X}_t^U &= \hat{X}_0^U + \int_0^t \left((A - Q(s)\Sigma_W)\,\hat{X}_s^U + U_s\right) ds + \int_0^t Q(s)\Gamma'\, dY_s^U, \end{aligned} \tag{31}$$

This follows from the following decompositions:

$$\begin{aligned} X_t^U &= \int_0^t \exp(A(t-s))U_s\, ds + X_t, \\ Y_t^U &= \int_0^t \int_0^s \Gamma \exp(A(s-r))U_r\, dr\, ds + Y_t, \end{aligned}$$

where X and Y are as in Section 1.

The controlled signal can also be decomposed into the observable and unobservable components $\hat{X}^U$ and $\tilde{X}^U$ $(:= X^U - \hat{X}^U)$. The evolution of the observable component is given by (31), and that of the unobservable component is as follows:

$$\tilde{X}_t^U = \tilde{X}_0^U + \int_0^t (A - Q(s)\Sigma_W)\tilde{X}_s^U\, ds + BV_t - \int_0^t Q(s)\Gamma'\, dW_s.$$

From this it can be seen that the unobservable component is also *uncontrollable*, in the sense that it does not depend on U. It has mean zero and covariance matrix $Q(t)$ at time t. The observable component can be made arbitrarily small by the choice of control (30) with sufficiently large K. Thus the effect of the control (30) is to drive the observable degrees of freedom of X^U towards zero, while not affecting the unobservable degrees

of freedom. The active demon makes immediate use of newly incoming information to extract the associated energy. The fact that it is able to do this with a control that depends on $(Y_s, s \in [0,t])$ only through $\hat{X}_t^U$ illustrates, once again, the redundancy of the dissipated information, $D(t)$.

Because of the circulation of energy in both passive and active analogies, we might expect the stationary state of the universe with observations to be a *non-equilibrium* state, even if the signal process X is self adjoint. To make these ideas precise we introduce, in Section 4, a second statistical mechanical analogy in which the filter is a separate physical component capable of holding energy in its own right.

4. INTERACTIVE STATISTICAL MECHANICS

The physical analogy, developed in Section 2 for the signal alone, can also be applied to the *joint*, $2n$-dimensional process $(X, \hat{X})$. In order to define entropy production for this we require the observation noise covariance matrix, Σ_W, to be *strictly* positive definite. In particular, this requires (unlike the analogies in Section 3) that the dimension of the observation, d, be the same as that of the signal, n. The joint process then has the invariant distribution $N(0, P_J)$, with $2n \times 2n$ covariance matrix

$$P_J = \begin{bmatrix} P_{SS} & P_{SS} - Q_{SS} \\ P_{SS} - Q_{SS} & P_{SS} - Q_{SS} \end{bmatrix}, \tag{32}$$

where, Q_{SS} is the stationary covariance matrix of the filter; this satisfies the algebraic Riccati equation:

$$A Q_{SS} + Q_{SS} A' + \Sigma_V - Q_{SS} \Sigma_W Q_{SS} = 0;$$

the Hamiltonian for $(X, \hat{X})$ is

$$H_J(x, \hat{x}) = \frac{1}{2} \begin{bmatrix} x' & \hat{x}' \end{bmatrix} P_J^{-1} \begin{bmatrix} x \\ \hat{x} \end{bmatrix}. \tag{33}$$

The joint process can be considered as describing a statistical mechanical system interacting with a unit temperature heat bath in the same way as was X in Section 2. This interaction forces the system towards its stationary state thus maximising the entropy of the universe comprising the joint process and the heat bath. The stationary state in this analogy is a *non-equilibrium* state, regardless of whether or not the signal, X, is self-adjoint. This is because of the interaction between the two components, X and

$\hat{X}$. We first investigate this interaction when the system is in its stationary state.

The joint process is a $2n$-vector, Gaussian process with drift coefficient $f(\theta) = A_J\theta$ and diffusion matrix Σ_J, where

$$A_J = \begin{bmatrix} A & 0 \\ Q_{SS}\Sigma_W & A - Q_{SS}\Sigma_W \end{bmatrix} \quad \text{and} \quad \Sigma_J = \begin{bmatrix} \Sigma_V & 0 \\ 0 & Q_{SS}\Sigma_W Q_{SS} \end{bmatrix}.$$

At each time t, $(X_t, \hat{X}_t)$ has mean zero, and the covariance matrix P_J of (32). It can be expressed in time-reversed form as a $2n$-vector Gaussian process with drift coefficient $\bar{A}_J\theta$, and diffusion matrix Σ_J, where

$$\bar{A}_J = -A_J - \Sigma_J P_J^{-1}.$$

(This follows from Lemma 2.1.) The rate of entropy production for the joint process can be found in the same way as was R_X in Section 2. In fact

$$R_J = \frac{1}{2}\mathrm{tr}\left((A_J - \bar{A}_J)'\Sigma_J^{-1}(A_J - \bar{A}_J)P_J\right). \tag{34}$$

Since this is a rate of entropy production in a stationary state, it is also the joint rate of entropy *flow* in this state.

The key to isolating the *interactive* component of this flow is the fact that both X and $\hat{X}$ are *autonomously* Markov. This is clearly true of X, but also true of $\hat{X}$ since the latter can be expressed autonomously as shown in (5). The rate of entropy flow in X alone, R_X, is given by (16) in the stationary state, and that in $\hat{X}$ alone, $R_{\hat{X}}$, is given by the following:

$$R_{\hat{X}} = \frac{1}{2}\mathrm{tr}\left((A - \bar{A}_{\hat{X}})'(Q_{SS}\Sigma_W Q_{SS})^{-1}(A - \bar{A}_{\hat{X}})(P_{SS} - Q_{SS})\right), \tag{35}$$

where

$$\bar{A}_{\hat{X}} = -A - Q_{SS}\Sigma_W Q_{SS}(P_{SS} - Q_{SS})^{-1}.$$

We can now define a rate of *interactive entropy flow*:

$$\begin{aligned} R_I &:= R_J - R_X - R_{\hat{X}} \\ &= \frac{1}{2}\mathrm{tr}(\Sigma_W Q_{SS}) + \frac{1}{2}\mathrm{tr}\left(\Sigma_V\left(Q_{SS}^{-1} - P_{SS}^{-1}\right)\right) \\ &= \dot{S}_{SS} + \dot{D}_{SS}, \end{aligned} \tag{36}$$

where $\dot{S}_{\rm SS}$ and $\dot{D}_{\rm SS}$ are the steady-state values of the information supply and dissipation processes of Section 3. (Of course, these are equal.) The rate of interactive entropy flow is thus the total flow rate of information to and from the information store.

Since X and $\hat{X}$ are Markov processes in their own right, they separately describe statistical mechanical systems interacting with unit temperature heat baths. The *marginal* interactions are governed by the Hamiltonians H_X of (7) and $H_{\hat{X}}$, defined by:

$$H_{\hat{X}}(\hat{x}) = \frac{1}{2}\hat{x}'(P_{\rm SS} - Q_{\rm SS})^{-1}\hat{x}. \tag{37}$$

We can also identify the *conditional* Hamiltonians

$$\begin{aligned} H_{X|\hat{X}}(x,\hat{x}) &= H_J(x,\hat{x}) - H_{\hat{X}}(\hat{x}) \\ H_{\hat{X}|X}(\hat{x},x) &= H_J(x,\hat{x}) - H_X(x). \end{aligned}$$

The Hamiltonian of the joint system can be expressed as the sum of three components:

$$H_J(x,\hat{x}) = H_{X|\hat{X}}(x,\hat{x}) + e_C(x,\hat{x}) + H_{\hat{X}|X}(\hat{x},x), \tag{38}$$

where e_C is a component of energy common to the signal and the filter (defined by (38)). The sum of the first two components in (38) is the Hamiltonian of the signal, H_X, and the sum of the last two components is that of the filter, $H_{\hat{X}}$.

$H_{X|\hat{X}}$ can be expressed in the following form:

$$H_{X|\hat{X}}(x,\hat{x}) = \frac{1}{2}(x-\hat{x})'Q_{\rm SS}^{-1}(x-\hat{x}),$$

and is, therefore, determined by the 'conditional signal', $\tilde{X} := X - \hat{X}$. This is also a Markov process in its own right, and evolves according to the equation

$$\tilde{X}_t = \tilde{X}_0 + \int_0^t (A - Q(s)\Sigma_W)\tilde{X}_s\,ds + BV_t - \int_0^t Q(s)\Gamma'\,dW_s. \tag{39}$$

It describes a statistical mechanical system with Hamiltonian

$$H_{\tilde{X}}(\tilde{x}) = \frac{1}{2}\tilde{x}'Q_{\rm SS}^{-1}\tilde{x},$$

that also interacts with a unit temperature heat bath.

The joint statistical mechanical system, described by $(X, \hat{X})$, can thus be thought of as comprising two 'physically distinct' subsystems: the conditional signal, with associated variable $\tilde{X}_t$ and Hamiltonian $H_{\tilde{X}}$, and the filter, with associated variable $\hat{X}_t$ and Hamiltonian $H_{\hat{X}}$. By 'physically distinct', we mean that the subsystems satisfy three conditions: (i) their state variables are autonomously Markov; (ii) energy is additive—the Hamiltonian of the joint system is the sum of the Hamiltonians of the two subsystems; and (iii) entropy is additive—since $\tilde{X}_t$ and $\hat{X}_t$ are independent, the entropy of the joint system is the sum of the entropies of the subsystems.

The conditional signal has average energy

$$E_{\tilde{X}}(t) = \mathbf{E}H_{\tilde{X}}(\tilde{X}_t) = \frac{1}{2}\mathrm{tr}(Q(t)Q_{\mathrm{SS}}^{-1})$$

and this evolves as follows:

$$\dot{E}_{\tilde{X}}(t) = \frac{1}{2}\mathrm{tr}\left(\Sigma_V Q_{\mathrm{SS}}^{-1}\right) + \mathrm{tr}\left(AQ(t)Q_{\mathrm{SS}}^{-1}\right) - \frac{1}{2}\mathrm{tr}\left(Q(t)\Sigma_W Q(t)Q_{\mathrm{SS}}^{-1}\right). \quad (40)$$

It forms one component of the average signal energy, $E_X(t)$ (as defined in (9)), the other component of which is the average common energy, $\mathbf{E}e_C(X_t, \hat{X}_t)$. The evolution of $E_X(t)$ is not affected by the presence of the filter; in particular, it would not be changed if Σ_W were set to zero, and so we may conclude that the third term on the right-hand side of (40) represents an energy flow from the conditional signal to the common energy, and hence to the filter. This is a delicate point. In trying to identify circular flows of energy between three or more subsystems, we must break one of the connections between subsystems and observe the increase in energy of the 'upstream' component, or the decrease in energy of the 'downstream' component. However, in changing a parameter of the system, we must be careful that we do not alter dynamical aspects of the system other than that intended. Clearly, setting Σ_W to zero not only disconnects the filter from the signal but also alters its interaction with the heat bath. Thus, observing the filter energy with Σ_W set to zero will not reveal the energy inflow from the signal. However, observing the energy of the conditional signal in the same circumstances does. The fact that the marginal statistical mechanics of the signal are not affected by the value of Σ_W is crucial here.

The filter itself has average energy $E_{\hat{X}}(t)$, given by

$$E_{\hat{X}}(t) = \mathbf{E}H_{\hat{X}}(\hat{X}) = \frac{1}{2}\mathrm{tr}\left((P(t) - Q(t))(P_{\mathrm{SS}} - Q_{\mathrm{SS}})^{-1}\right)$$

and this evolves as follows:

$$\dot{E}_{\hat{X}}(t) = \frac{1}{2}\mathrm{tr}\left(Q(t)\Sigma_W Q(t)(P_{\mathrm{SS}} - Q_{\mathrm{SS}})^{-1}\right) + \mathrm{tr}\left(A(P(t) - Q(t))(P_{\mathrm{SS}} - Q_{\mathrm{SS}})^{-1}\right).$$

We can thus identify the following three energy flow rates:

Flow 4: Heat Bath to Conditional Signal.

$$\dot{E}_4(t) = \frac{1}{2}\mathrm{tr}\left(\Sigma_W Q_{\mathrm{SS}}^{-1}\right) + \mathrm{tr}\left(AQ(t)Q_{\mathrm{SS}}^{-1}\right);$$

Flow 5: Conditional Signal to Filter.

$$\dot{E}_5(t) = \frac{1}{2}\mathrm{tr}\left(Q(t)\Sigma_W Q(t)Q_{\mathrm{SS}}^{-1}\right);$$

Flow 6: Filter to Heat Bath.

$$\dot{E}_6(t) = \frac{1}{2}\mathrm{tr}\left(Q(t)\Sigma_W Q(t)\left(Q_{\mathrm{SS}}^{-1} - (P_{\mathrm{SS}} - Q_{\mathrm{SS}})^{-1}\right)\right) - \mathrm{tr}\left(A(P(t) - Q(t))(P_{\mathrm{SS}} - Q_{\mathrm{SS}})^{-1}\right).$$

In the stationary state, all three energy flows have the common rate

$$\dot{E}_{\mathrm{SS}} = \frac{1}{2}\mathrm{tr}(\Sigma_W Q_{\mathrm{SS}}) = \dot{S}_{\mathrm{SS}}.$$

Since all three components of the universe have unit temperature, the energy flows are accompanied by equal entropy flows. In particular, the energy flow from the conditional signal to the filter is associated with an entropy flow of the same rate as that of the information supply, identified in Section 3. The flow of energy in the stationary state is not driven by temperature gradients and does not cause any increase in overall entropy, and so is 'physically reversible'.

The physical analogy here is distinct from that of Section 3 in that the energy flows of the latter are driven by the supply of observation information, whereas the flows here are driven by the nature of the interaction between X and $\hat{X}$, and do not depend on any distinction being made between entropy and observable information.

The joint Hamiltonian can also be expressed as the sum of that of the signal, H_X, and that of the 'conditional filter', $H_{\hat{X}|X}$. The latter can be expressed in the form

$$H_{\hat{X}|X}(\hat{x}, x) = \frac{1}{2}\breve{x}' \left(Q_{\mathrm{SS}}^{-1} - P_{\mathrm{SS}}^{-1} \right)^{-1} \breve{x},$$

where

$$\breve{x} = P_{\mathrm{SS}}^{-1} x - Q_{\mathrm{SS}}^{-1}(x - \hat{x}).$$

This is determined by the process $\breve{X} := P_{\mathrm{SS}}^{-1} X - Q_{\mathrm{SS}}^{-1} \tilde{X}$.

In the stationary state, $\breve{X}$ is a Markov process in its own right, and its value at time t is independent of that of X, and so we can also decompose the joint system into subsystems associated with X and $\breve{X}$. (Note, however, that this is not, in general, true away from the stationary state.) Thus, in the stationary state, we can identify a flow of energy from the signal to the conditional filter in the same way that the flow of energy from the conditional signal to the filter was identified above. (This involves setting Σ_V to zero.) This energy flow has the same rate as the others, $\dot{E}_{\mathrm{SS}}$, and leads to the symmetrical system shown in Fig. 3. The joint system has two 'internal' energy flow points (ie., points of flow not involving the heat bath), each of which has an associated entropy flow of rate $\dot{S}_{\mathrm{SS}}$; the sum of these is equal to the rate of interactive entropy flow R_I, as defined in Eq. (36).

In the presence of observations, we can make the distinction between entropy and information that was made in Section 3. The entropy of the

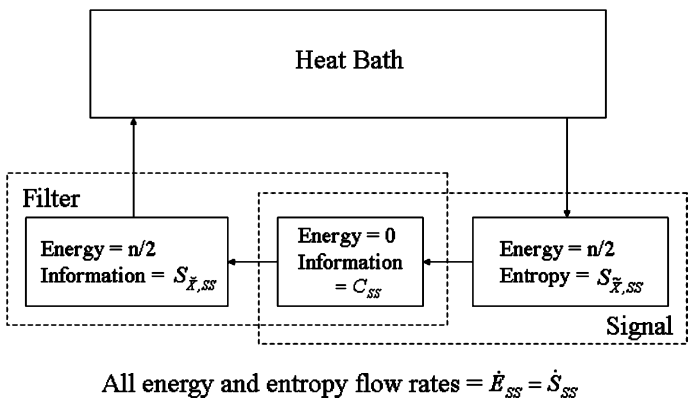

Fig. 3. Energy flows in the joint system.

joint system in the presence of observations then becomes that of the conditional signal,

$$S_{\check{X},\mathrm{SS}} = \frac{n}{2}(1+\log(2\pi)) + \frac{1}{2}\log|Q_{\mathrm{SS}}|,$$

and its information content becomes

$$S_{\hat{X},\mathrm{SS}} = \frac{n}{2}(1+\log(2\pi)) + \frac{1}{2}\log|P_{\mathrm{SS}} - Q_{\mathrm{SS}}|.$$

This is made up of two components: the information stored on the signal

$$C_{\mathrm{SS}} = \frac{1}{2}\log\left|P_{\mathrm{SS}}Q_{\mathrm{SS}}^{-1}\right|,$$

and the 'residual' information

$$S_{\check{X},\mathrm{SS}} = \frac{n}{2}(1+\log(2\pi)) + \frac{1}{2}\log\left|Q_{\mathrm{SS}}\left(Q_{\mathrm{SS}}^{-1} - P_{\mathrm{SS}}^{-1}\right)Q_{\mathrm{SS}}\right|.$$

As in Section 3, we call the energy of the conditional signal *heat*, and that of the filter *work*. Heat is converted into work at rate $\dot{E}_{\mathrm{SS}}$ by the arrival of new observation information. It is converted back into heat when it is returned to the heat bath by the filter. The filter uses its information dissipation process, which has exactly the correct rate, to provide the necessary entropy. This provides a quantitative example of Landauer's Principle, distinct from that in Section 3. There, it was the flow of energy across the temperature gradient between the heat bath and the filter that caused the entropy increase term $\dot{D}_{\mathrm{SS}}$, whereas in the analogy of this section there are no temperature gradients.

Many of the properties of entropy production, discussed in the remarks following its definition in Section 2, are inherited, in the stationary state, by interactive entropy flow. In particular, the interactive entropy flow of the time-reversed joint process $(X, \hat{X})$ is the same as that of the forward-time joint process. It turns out that the components of the time-reversed joint process can be thought of as being the signal and filter processes of a *dual* problem, in which information supply and dissipation exchange roles. The second physical analogy for this dual problem is then the physical reversal of that of the original. These duality ideas are developed elsewhere, for both linear and non-linear filters.

5. CONCLUSIONS AND FURTHER WORK

This article has explored the information flows associated with continuous-time Kalman–Bucy filters, and connected them with the entropy flows occuring in non-equilibrium statistical mechanical systems. It has shown via physical analogies that a law of non-decrease of entropy need not apply to such systems in the presence of observations that continue to supply new information. Furthermore, by introducing a concept of interactive entropy flow, it has provided a framework for the study of interacting statistical mechanical systems.

The results concern systems with finite-dimensional mesoscopic descriptions, but they can be extended to situations where the signal space is an infinite-dimensional Hilbert space, and the observation space is a finite- or infinite-dimensional Hilbert space. The necessary ingredients for such a development are contained, for example, in ref. 22. The generalisation of the Girsanov theorem, which was used to derive the information flow rates, to the infinite-dimensional Hilbert space is contained in the work of Yor, where the martingale problem for certain infinite-dimensional stochastic differential equations is developed. (See, also, ref. 15.) In situations of interest in Physics one may consider the infinite–dimensional Ornstein–Uhlenbeck process evolving, for example, in $H^{-1}(\Lambda)$, where Λ is a bounded set in $\mathbb{R}^2$. (See, for example, refs. 3 and 12.)

Like all Bayesian estimators, the Kalman–Bucy filter is information conserving in the manner described in ref. 21, and, because of this, it is also entropically efficient in the physical analogies: it achieves the maximum possible reduction in entropy from a given supply of observations, and stores no more information than is strictly necessary to do this. This entropic efficiency manifests itself in a second analogy as physically reversible dynamics for the system described by the joint signal-filter process.

The fact that observations can reduce entropy is, essentially, an 'inverse Landauer Principle'. Our analogies provide quantitative examples of both 'regular' and 'inverse' Landauer's Principles.

These ideas can also be further developed for linear, time-inhomogeneous, degenerate systems and to non-linear systems. (See two sequels to this article by the second author, where an interactive analogy is introduced.) The physical analogies are particularly useful in the non-linear case since they can be used as the basis of an information theoretic Lyapunov theory for non-linear filters.

ACKNOWLEDGMENTS

This work was partially supported by Department of Defense MURI Grant: Complex Adaptive Networks for Cooperative Control, Subaward

#03-132, and by ARO-MURI Grant DAAD19-00-1-0466 (Data Fusion in Large Arrays of Microsensors (sensor web)), by NSF Grant #CCF-0325774 and by an INTEL Grant. This publication is also an output from a research project funded by the Cambridge-MIT Institute (CMI). CMI is funded in part by the UK Government. The research was carried out for CMI by the Massachusetts Institute of Technology and the University of Essex. CMI can accept no responsibility for any information provided or views expressed.

REFERENCES

1. C. H. Bennett, The thermodynamics of computation—a review, *Int. J. Theoret. Phys.* **21**:905–940 (1982).
2. L. Bertini, A. De Sole, D. Gabrielli, G. Jona-Lasinio and C. Landim, Macroscopic fluctuation theory for stationary non-equilibrium states, *J. Stat. Phy.* **107**:635–675 (2002).
3. V. S. Borkar, R. T. Chari, and S. K. Mitter, Stochastic quantization of field theory in finite and infinite volume, *J. Func. Anal.* **81** (1988).
4. R. W. Brockett and J. C. Willems, Stochastic control and the second law of thermodynamics, *Proceedings of the 17th IEEE Conference on Decision and Control* (San Diego CA, IEEE, 1979), pp. 1007–1011.
5. M. H. A. Davis, *Linear Estimation and Stochastic Control* (Chapman and Hall, London, 1977).
6. T. E. Duncan, On the calculation of mutual information, *SIAM J. Appl. Math.* **19**:215–220 (1970).
7. X. Feng, K. A. Loparo, and Y. Fang, Optimal state estimation for stochastic systems: and information theoretic approach, *IEEE Trans. Automatic Control* **42**:771–785. (1997).
8. G. Gallavotti, *Statistical Mechanics—A Short Treatise* (Springer-Verlag, Berlin Heidelberg/Newyork, 1999).
9. B. Gaveau and L. S. Schulman, Creation, dissipation and recycling of resources in non-equilibrium systems, *J. Stat. Phys.* **110**: 1317–1367 (2003).
10. A. H. Jazwinski, *Stochastic Processes and Filtering Theory* (Academic Press, London, 1970).
11. J. B. Johnson, Thermal agitation of electricity in conductors, *Phys. Rev.* **32**:97–109 (1928).
12. G. Jona-Lasinio and P. K. Mitter, On the stochastic quantization of field theory, *Comm. Math. Phys.* **101**:409–436 (1985).
13. P. Kalata and R. Priemer, Linear prediction, filtering and smoothing: an information theoretic approach, *Inf. Sci.* **17**:1–14 (1979).
14. I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus* (Springer-Verlag, Berlin/Heidelberg/New York, 1991).
15. T. G. Kurtz and D. L. Ocone, Unique characterization of conditional distributions in non-linear filtering, *Ann. Prob.* **16**:80–107 (1988).
16. R. Landauer, Dissipation and heat generation in the computing process, *IBM J. Res. Dev.* **5**:183–191 (1961).
17. J. L. Lebowitz and H. Spohn, A Gallavotti–Cohen type symmetry in the large deviation functional for stochastic dynamics, *J. Stat. Phys.* **95**:333–366 (1999).
18. R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes 1—General Theory* (Springer-Verlag,Berlin/Heidelberg/New York, 1977).

19. R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes 2—Applications* (Springer-Verlag, Berlin/Heidelberg/New York, 1978).
20. J. C. Maxwell, Theory of Heat (Longmans, London, 1871).
21. S. K. Mitter and N. J. Newton, A Variational approach to non-linear estimation, *SIAM J. Control Optim.* **42**:1813–1833 (2003).
22. S. K. Mitter and R. B. Vinter, Filtering for linear stochastic hereditary differential systems, Lecture Notes in *Economics and Mathematical Systems, 107 Control Theory, Numerical Methods and Computer Modelling* (Springer-Verlag, Berlin/Heidelberg/New York, 1975), pp. 1–21.
23. H. Nyquist, Thermal agitation of electric charge in conductors, *Phys. Rev.* **32**:110–113 (1928).
24. M. B. Propp, *The Thermodynamic Properties of Markov Processes*, Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1985.
25. D. Ruelle, Positivity of entropy production in non-equilibrium statistical mechanics, *J. Stat. Phys.* **85**:1–23 (1996).
26. E. Mayer-Wolf and M. Zakai, On a formula relating the Shannon information to the Fisher information for the filtering problem, in *Filtering and Control of Random Processes*, H. Korezlioglu, G. Mazziotto, and S. Szpirglas, eds., Lecture Notes in Control and Information Sciences 61 (Springer, 1984), pp. 164–171.
27. Y. Tomita, S. Omatu, and T. Soeda, An application of the information theory to filtering problem, *Inf. Sci.* **11**:13–27 (1976).
28. J. C. Willems, Dissipative dynamical systems, part i: general theory, *Arch. Ratinal.* Mech. Anal. **45**:321–351 (1972).